เอกสารประกอบการสอน
วิชา การทำเหมืองข้อมูล (204423)
(Data Mining)

จักรเมธ บุตรกระจ่าง

ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

2559

คำนำ

เอกสารประกอบการสอนเล่มนี้จัดทำขึ้นเพื่อประกอบการสอนวิชาการทำเหมืองข้อมูล (Data Mining) รหัสวิชา 204423 ในหลักสูตรวิทยาศาสตรบัณฑิต ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

เป้าหมายของการสอนวิชาการทำเหมืองข้อมูล มุ่งหมายให้ผู้เรียนเห็นภาพกว้างของการทำเหมืองข้อมูล เริ่มตั้งแต่การเก็บข้อมูล การเตรียมข้อมูลก่อนการประมวลผล การประมวลผลข้อมูลแบบอัตโนมัติในลักษณะต่างๆ รวมถึงการนำผลลัพธ์ที่ได้จากการประมวลไปประยุกต์ใช้งานจริง โดยการเรียนการสอนเน้นการศึกษาในสองส่วน คือส่วนทฤษฎีที่ซึ่งผู้เรียนจะได้ศึกษาขั้นตอนวิธีในการทำเหมืองข้อมูลในเชิงลึก และส่วนปฏิบัติที่ซึ่งผู้เรียนจะได้รับการสนับสนุนให้ได้ทดลองถ่ายโอนทฤษฎีและขั้นตอนวิธี จากบนกระดาษสู่โปรแกรมคอมพิวเตอร์ที่สามารถทำงานกับข้อมูลจริงได้ผ่านการบ้าน

เนื่องจากความเป็นพหุวิทยาการของศาสตร์ด้านการทำเหมืองข้อมูล ยังมีขั้นตอนวิธีและประเด็นชั้นสูงอีกมาก ที่ไม่ได้ถูกรวมไว้ในแผนการเรียนการสอนและเอกสารประกอบการสอนฉบับนี้ แต่ผู้แต่งเชื่อว่าสิ่งที่ผู้เรียน จะได้รับจากการเรียนวิชาการทำเหมืองข้อมูลนี้จะเป็นพื้นฐานที่เพียงพอสำหรับการศึกษาเทคนิคในชั้นสูงต่อไป

จักรเมธ บุตรกระจ่าง

สารบัญ

สารบัญรูป

สารบัญตาราง

บทที่ 1

แนวคิดพื้นฐานของการทำเหมืองข้อมูล

1.1 การทำเหมืองข้อมูลคืออะไร

การทำเหมืองข้อมูลหมายถึงการสกัดสารสนเทศ (Information) หรือความรู้ (Knowledge) ที่อาจเป็นประโยชน์จากข้อมูลดิบ (Raw Data) โดยสารสนเทศหรือความรู้ดังกล่าวอาจจะอยู่ในรูปของ ความสัมพันธ์ (Relationship) แบบรูป (Pattern) หรือมโนทัศน์ (Concept) ในเชิงลึกที่ไม่สามารถมองออกได้ชัดโดยวิธีการประมวลผลข้อมูลแบบพื้นฐาน (Non-trivial)

คำว่าการทำเหมืองข้อมูลดูจะเป็นชื่อที่ไม่เหมาะสมนักเมื่อพิจารณาตามความหมาย เนื่องจากคำว่าการทำเหมืองใดๆ มักจะหมายถึงกว่าขุดเจาะแสวงหาสิ่งนั้นๆ ยกตัวอย่างเช่น การทำเหมืองทองคำก็คือการแสวงหาทองคำจากใต้พื้นพิภพ ทว่าเป้าหมายของการทำเหมือนข้อมูลไม่ได้อยู่ที่ตัวข้อมูล แต่เป็นความรู้หรือสารสนเทศที่เป็นประโยชน์จากข้อมูลมากกว่า การทำเหมืองข้อมูลยังมีชื่อเรียกอื่นๆอีก โดยชื่อที่ได้รับความนิยมอันดับต้นๆก็ได้แก่ การค้นพบความรู้จากฐานข้อมูล (Knowledge Discovery in Databases) การสกัดความรู้ (Knowledge Extraction) หรือการวิเคราะห์ข้อมูล (Data Analysis) เป็นต้น

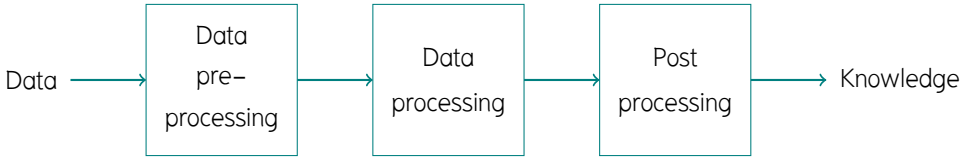
การทำเหมืองข้อมูลถือว่าเป็นศาสตร์ที่กำลังได้รับความสนใจเป็นอย่างมาก เนื่องจากจำนวนข้อมูลในปัจจุบันที่เพิ่มมากขึ้นอย่างรวดเร็วและต่อเนื่อง ยกตัวอย่างเช่น ข้อมูลในลักษณะของข้อความหรือรูปภาพต่างๆบนสื่อออนไลน์ หรือข้อมูลของการซื้อขายและเลือกชมสินค้าบนร้านสินค้าออนไลน์ซึ่งเกิดขึ้นใหม่ทุกๆ วินาที หรือข้อมูลทางวิทยาศาสตร์ที่สามารถเก็บรวบรวมได้เร็วและถี่ขึ้นโดยอาศัยเครื่องมือตรวจวัดแบบอัตโนมัติที่มีประสิทธิภาพ ดังนั้นทั้งภาคธุรกิจและแวดวงวิทยาศาสตร์จึงมีความสนใจที่จะเปลี่ยน สกัด และสรุปข้อมูลมหาศาลเหล่านั้น ให้อยู่ในรูปของสารสนเทศหรือความรู้ ที่เป็นประโยชน์และสามารถนำไปใช้งานต่อได้

ในอดีตเมื่อข้อมูลยังไม่มากและไม่ซับซ้อน การสร้างข้อคำถาม (Query) ด้วยมือ เช่นการใช้คำสั่ง SELECT ในภาษาเอสคิวแอล (SQL) เพื่อดึงข้อมูลจากฐานข้อมูลยังสามารถทำได้ แต่ในปัจจุบันข้อมูลมีความซับซ้อนเกี่ยวพันกันในหลายมิติ ทำให้จำเป็นต้องคิดค้นวิธีการสกัดข้อมูลที่ซับซ้อนให้มีประสิทธิภาพ และเป็นอัตโนมัติโดยลดการยื่นมือเข้าเกี่ยวข้องจากผู้ใช้ให้มากที่สุด อีกทั้งข้อมูลที่เกิดขึ้นในปัจจุบันยังอยู่ในรูปแบบที่ต่างไปจากเดิม เช่นข้อมูลอาจจะอยู่ในรูปของกระแสข้อมูล (Data Stream) อย่างเช่นข้อมูลจากกล้องโทรทัศน์วงจรปิด หรือข้อมูลจากการรับรู้จากระยะไกล (Remote Sensing) [Richards, 1999, Lillesand et al., 2014] หรือข้อมูลอาจจะอยู่ในรูปแบบของกราฟ (Graphical data) เป็นต้น ทำให้เราต้องพัฒนาวิธีการสกัดความรู้จากข้อมูลที่มีประสิทธิภาพ เพื่อรองรับข้อมูลประเภทใหม่ๆดังกล่าวด้วย

คำว่า “อัตโนมัติ” ถือเป็นคำสำคัญที่แบ่งแยกการทำเหมืองข้อมูล ออกจากการดึงข้อมูลจากฐานข้อมูลแบบเดิมๆ โดย การทำเหมืองข้อมูลที่เราสนใจนี้ ต้องการที่จะลดความลำเอียง ความผิดพลาด หรือความไม่แน่นอน ของผู้ใช้หรือผู้เชี่ยวชาญออกให้มากที่สุด แล้วมุ่งค้นหาหรือสกัดความรู้ที่แฝงตัวอยู่ในข้อมูลดิบจริงๆ ด้วยความรวดเร็วและประหยัดเวลากว่าการใช้ผู้เชี่ยวชาญโดยตรง ผลลัพธ์ที่ได้ ไม่ว่าจะอยู่ในรูปของแบบรูป ความรู้ หรือกฎความสัมพันธ์ จะมีส่วนช่วยในการปรับแผนการดำเนินธุรกิจ และสนับสนุนงานวิจัยทางวิทยาศาสตร์และทางการแพทย์ รวมทั้งอาจทำให้เกิดความรู้ความเข้าใจในข้อมูลในมุมมองใหม่ๆได้

1.2 ภาพรวมขั้นตอนการทำเหมืองข้อมูล

ภารกิจสำคัญในการทำเหมืองข้อมูลสามารถแบ่งย่อยออกเป็น 3 ส่วนหลักๆ ได้แก่ ขั้นตอนการเตรียมข้อมูล (Data Preprocessing) ขั้นตอนการประมวลผล (Data Processing) และขั้นตอนหลังการประมวลผล (Post Processing) รูปที่ 1.1 แสดงภาพรวมของการทำเหมืองข้อมูล เริ่มตั้งแต่ข้อมูลดิบจนถึงการได้มาซึ่งความรู้ในขั้นสุดท้าย



รูปภาพ 1.1: ขั้นตอนสำคัญ 3 ประการของการทำเหมืองข้อมูล

สังเกตว่าก่อนที่ข้อมูลดิบจะเข้าไปสู่การประมวลผล จำเป็นจะต้องมีการเตรียมข้อมูลก่อนที่จะลงมือสกัดสารสนเทศหรือความรู้ ขั้นตอนการเตรียมข้อมูลนี้มีจุดประสงค์เพื่อปรับข้อมูลดิบที่เก็บรวบรวมมาได้มีความกลมกลืนและต้องกัน โดยทั่วไปแล้วข้อมูลที่ผ่านการเตรียมแล้วจะสามารถนำไปวิเคราะห์หรือศึกษาต่อได้ง่าย

ขึ้น ขั้นตอนย่อยที่สำคัญในส่วนนี้ได้แก่ การชำระข้อมูล (Data Cleaning) การรวมข้อมูล (Data Integration) การเลือกข้อมูล (Data Selection) การปรับบรรทัดฐานข้อมูล (Data Normalising) รวมไปถึงการลดมิติข้อมูล (Dimensionality Reduction) ด้วย เราจะได้ศึกษาวิธีในการเตรียมข้อมูลแบบต่างๆในเชิงลึกในบทที่ 2 และบทที่ 3 ต่อไป

ส่วนที่สองของการทำเหมืองข้อมูลก็คือส่วนที่เกิดการกระทำกับข้อมูลเพื่อสกัดสารสนเทศหรือความรู้ที่ต้องการออกมา ส่วนนี้นับเป็นส่วนที่สำคัญที่สุดของกระบวนการทั้งหมด เป้าหมายของการสกัดที่อยู่ในความสนใจก็ได้แก่ การประมวลเพื่อหากฎความสัมพันธ์ (Association Rule) การจำแนกข้อมูล (Classification) รวมไปถึงการจัดกลุ่มข้อมูล (Clustering) โดยเครื่องมือและขั้นตอนวิธี (Algorithm) ต่างๆที่นำมาใช้เพื่อสกัดสารสนเทศหรือความรู้ดังกล่าว อาจได้มาจากศาสตร์ที่เรียกว่า การเรียนรู้ของเครื่อง (Machine Learning) รวมไปถึงการประยุกต์ทฤษฎีพื้นฐานทางสถิติและคณิตศาสตร์ด้วย เราจะศึกษาการกฎความสัมพันธ์ การจำแนกข้อมูล และการจัดกลุ่มข้อมูลในบทที่ 4 บทที่ 5 และบทที่ 6 ตามลำดับ

เมื่อได้ผลลัพธ์จากขั้นตอนการประมวลผลแล้ว จำเป็นจะต้องมีการประเมินว่าผลลัพธ์ที่สกัดออกมาได้ มีคุณค่ามากน้อยแค่ไหน ก่อนที่จะสรุปผลลัพธ์ออกมาเป็นความรู้ที่นำไปใช้ประโยชน์ได้ต่อไป

หากพิจารณาในภาพกว้างกว่าส่วนทั้งสามที่กล่าวไปในตอนต้นจะเห็นได้ว่า การทำเหมืองข้อมูลถือเป็นศาสตร์แบบพหุวิทยาการ (Multidisciplinary) ที่นำความรู้จากหลายศาสตร์เข้ามาประยุกต์เพื่อให้บรรลุเป้าหมายที่ต้องการ เริ่มตั้งแต่ขั้นตอนที่เกี่ยวข้องกับการจัดเก็บรวบรวมข้อมูล ซึ่งต้องอาศัยความรู้ของระบบฐานข้อมูล (Database system) เข้ามาช่วย เพื่อให้การเก็บและเข้าถึงข้อมูลมีประสิทธิภาพมากที่สุด ทั้งยังต้องลดความซ้ำซ้อนของข้อมูลที่อาจจะมาจากฐานข้อมูลที่ต่างกัน อีกหนึ่งศาสตร์ที่นำมาประยุกต์ใช้อย่างมากก็คือการเรียนรู้ของเครื่องดังที่กล่าวไปแล้วข้างต้น ที่ถูกนำมาใช้เพื่อสกัดแบบรูปที่น่าสนใจจากข้อมูล เพื่อทำการจำแนก ทำนาย และจัดกลุ่มข้อมูล นอกจากนี้ ส่วนของหลังการประมวลผลอาจจำเป็นจะต้องอาศัย เทคนิคการสร้างมโนภาพจากข้อมูล (Data Visualisation) [Cleveland, 1993, Fayyad et al., 2002] มาเพื่อใช้แสดงผลลัพธ์ในรูปแบบที่เข้าใจและเห็นภาพได้ง่ายด้วย

เมื่อทราบถึงความหมายและขั้นตอนวิธีในการทำเหมืองข้อมูลไปแล้ว ต่อไปเราจะไปทำความรู้จักกับตัวข้อมูลว่ามีกี่ประเภท และแต่ละประเภทแตกต่างกันอย่างไร

1.2.1 ประเภทของข้อมูล

โดยหลักการแล้วเราสามารถทำเหมืองข้อมูลบนข้อมูลทุกประเภท ไม่ว่าจะเป็นข้อมูลที่ไม่ค่อยเปลี่ยนแปลง เช่น ข้อมูลรูปภาพ (Image) ข้อมูลแบบข้อความ (Text) หรือข้อมูลที่มีธรรมชาติเป็นแบบกระแสข้อมูลเช่น ข้อมูลเสียง (Sound) ข้อมูลอนุกรมเวลา (Time Series) เพียงแต่ว่าข้อมูลเหล่านั้น จำเป็นจะต้องผ่านการเตรียมข้อมูลให้เหมาะสมกับเครื่องมือหรือขั้นตอนวิธีที่จะนำมาใช้ โดยทั่วไปแล้วข้อมูลที่มักถูกนำมาสกัดหาความรู้โดยการ

ทำเหมืองข้อมูล ได้แก่ข้อมูลประเภทต่างๆต่อไปนี้

ข้อมูลรายการเปลี่ยนแปลง (Transactional Data)

ข้อมูลประเภทนี้พบเจอได้มากที่สุดในฐานข้อมูลการซื้อขายของบริษัท ซึ่งเก็บประวัติการซื้อสินค้าของลูกค้าไว้ การนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้สำหรับข้อมูลประเภทนี้ อาจเริ่มจากการสรุปข้อมูลแบบเบื้องต้น เช่น การหาสินค้าทั้งหมดที่ขายได้ในวันนี้ หรืออาจจะเพิ่มความซับซ้อนขึ้นไปอีก เช่น การตามหาว่าสินค้าชนิดไหนที่มักจะมีคนซื้อพร้อมๆกัน ซึ่งคำตอบที่ได้อาจนำไปสู่การจัดแผนการวางสินค้าแบบใหม่ หรืออาจนำไปสู่แผนการตลาดใหม่ได้ด้วย

ยกตัวอย่างเช่น จากการวิเคราะห์ข้อมูลรายการซื้อขายสินค้าของห้างสรรพสินค้าแห่งหนึ่งในเชิงลึก พบว่า ผงซักฝอกและผ้าอ้อมเด็ก มักจะได้ขายพร้อมๆกัน ความรู้ดังกล่าวนี้เราอาจจะไม่ทราบได้เลย จากการใช้ข้อคำถามเพื่อสืบค้นจากฐานข้อมูลในแบบเดิมๆ แต่การทำเหมืองข้อมูล สามารถนำมาใช้เพื่อหาสิ่งที่เรียกว่าไอเทมเซตที่พบบ่อย (Frequent Itemset) หรือเซตของสิ่งของที่มักจะพบพร้อมๆกันในฐานข้อมูล เมื่อทราบดังนี้แล้ว ทางห้างสรรพสินค้าก็อาจจะวางแผนเพื่อนำผ้าอ้อมเด็กมาจัดแสดงใกล้ๆกับผลิตภัณฑ์ทำความสะอาดได้ โดยหวังว่าจะทำให้ยอดขายเพิ่มมากขึ้น เราจะศึกษาการค้นหาไอเทมเซตที่พบบ่อยในบทที่ 4

ข้อมูลเชิงเวลา (Temporal Data)

ฐานข้อมูลบางแห่งมีเก็บข้อมูลที่มีความสัมพันธ์กันในเชิงเวลา ยกตัวอย่างได้ง่ายที่สุดก็คือ ฐานข้อมูลที่บันทึกการซื้อขายหุ้นในตลาดหลักทรัพย์ หรือฐานข้อมูลที่เก็บข้อมูลทางอุตุนิยมวิทยา เช่น อุณหภูมิ ระดับน้ำขึ้น–ลง ขั้นตอนวิธีในการทำเหมืองข้อมูลสามารถนำไปประยุกต์ใช้ เพื่อหาแนวโน้มการเปลี่ยนแปลงของข้อมูลเชิงเวลาที่ถูกบันทึกไว้ และอาจนำไปสู่การพยากรณ์หรือคาดการณ์ก็เป็นได้ ความรู้ที่สกัดออกมาได้อาจมีผลดีต่อการวางแผนเพื่อรับมือกับสิ่งที่อาจจะเกิดขึ้นในอนาคต เช่น เมื่อไหร่นักลงทุนควรจะขายหรือซื้อหุ้น หรืออาจจะเป็นการแจ้งเตือนระวังภัยหนาวและพายุฤดูร้อนก็เป็นได้

ข้อมูลเชิงพื้นที่ (Spatial Data)

ฐานข้อมูลบางแห่งก็มีการเก็บข้อมูลที่มีความสัมพันธ์กันในเชิงพื้นที่ ยกตัวอย่างเช่น ข้อมูลทางภูมิศาสตร์ ภาพถ่ายดาวเทียม หรือข้อมูลการรับรู้จากระยะไกล การศึกษาและวิเคราะห์ข้อมูลเหล่านี้นับว่าเป็นประโยชน์มาก ในเชิงสำรวจและป้องกันภัย เช่น การจำแนกการใช้ประโยชน์ของที่ดินภาพถ่ายจากดาวเทียม ออกเป็นผืนป่า พื้นที่อยู่อาศัย พื้นที่การเกษตร โดยอัตโนมัติ การประยุกต์ใช้ทั้งหมดนี้ มีส่วนช่วยในการป้องกันการบุกรุกทำลายป่า เฝ้าระวังไฟป่า (หากบริเวณนั้นเคยเป็นสีเขียวแต่ภาพล่าสุดจับได้ว่าเป็นสีแดงปนดำ) หรือการศึกษาวิเคราะห์ความสัมพันธ์ระหว่างระยะทางจากถนนหลักกับรายได้ประชากรของบริเวณนั้นๆ

นอกจากนี้ ข้อมูลดิบที่ประกอบไปด้วยข้อมูลเชิงพื้นที่และเชิงเวลา ซึ่งเรียกกันว่า Spatio-temporal Data ถูกนำไปใช้ประโยชน์ได้หลายทาง ยกตัวอย่างเช่น การพยากรณ์และเฝ้าระวังการกระจายตัวของโรคระบาดร้ายแรง หรือการคาดการณ์ทิศทางการเคลื่อนที่ของพายุฝน

ข้อมูลประเภทข้อความและสื่อประสม (Textual and Multimedia Data)

การทำเหมืองข้อมูลบนข้อมูลประเภทข้อความ ถือเป็นหัวข้อสำคัญที่สุดหัวข้อหนึ่ง ในการทำเหมืองข้อมูลในปัจจุบัน ในบริบทนี้ ข้อความ (Text) อาจไม่ได้หมายถึงแค่ คำหลัก (Keyword) เท่านั้นแต่ยังหมายรวมถึง ประโยค หรือ ย่อหน้า

ภารกิจ (Task) ที่เป็นพื้นฐานที่สุดสำหรับการดึงหรือสกัดข้อมูลจากข้อความ คือการอธิบายสรุปเอกสารให้อยู่ในรูปแบบที่สั้นและได้ใจความ เช่น การวิเคราะห์หัวเรื่อง (Topic Analysis) [Blei et al., 2003] การตรวจจับความผิดปกติของเอกสาร (Abnormality Detection) เช่น การบอกว่าอีเมลนี้เป็นอีเมลขยะหรือไม่ นอกจากนั้นยังรวมไปถึงการวิเคราะห์ความรู้สึกจากเอกสาร (Sentiment Analysis) ตัวอย่างเช่น การวิเคราะห์เพื่อบอกว่าเนื้อความของเอกสาร เป็นไปทางบวกหรือทางลบ ซึ่งมีประโยชน์ในการวิเคราะห์การรีวิวสินค้าและบริการ [Pang and Lee, 2008]

ข้อมูลสื่อประสม (Multimedia) หมายรวมถึง ข้อมูลรูปภาพ ข้อมูลเสียง และข้อมูลวีดีโอ ก็เป็นเป้าหมายหลักอีกชนิดหนึ่ง ในการสกัดความรู้และสารสนเทศออกมา ยกตัวอย่างเช่น การวิเคราะห์เพื่อบอกแนวดนตรีของเพลง การตรวจจับวัตถุจากภาพถ่าย เป็นต้น ส่วนมากแล้ว การทำเหมืองข้อมูลบนข้อมูลแบบมัลติมีเดียมักจะมีขั้นตอนการแปลงข้อมูลดิบซึ่งอยู่ในรูปเสียงหรือภาพ ให้แสดงอยู่ในเชิงเวคเตอร์ ซึ่งสรุปข้อมูลสำคัญภายในภาพและเสียงนั้นไว้แล้ว

ข้อมูลเชิงกราฟ (Graphical Data)

ข้อมูลเชิงกราฟก็เป็นข้อมูลอีกประเภทหนึ่ง ที่ได้รับความนิยมนำมาสกัดหารูปแบบที่น่าสนใจ ข้อมูลประเภทนี้เราสามารถเข้าถึงได้ง่าย ยกตัวอย่างเช่น เราสามารถมองสิ่งที่อยู่ใกล้ตัวมากอย่าง World-Wide-Web เป็นกราฟ ซึ่งมี จุดต่อ (Node) คือเว็บไซต์และ เส้นเชื่อม (Edge) คือการเชื่อมโยงหลายมิติ (Hyperlink) ที่เชื่อมเข้า หรือออกจากเว็บไซต์นั้น WWW ถือว่าเป็นกราฟที่มีขนาดมหึมา และการจะวิเคราะห์ WWW ถือเป็นงานที่ท้าทายอย่างมาก

การประยุกต์ใช้หนึ่งที่เป็นที่รู้จักกันดี ได้แก่ *AdSense* จาก Google ซึ่งถือเป็นการวิเคราะห์ ประวัติการเข้าชมเว็บไซต์ของผู้ใช้หนึ่งๆ โดยเป้าหมายคือการแนะนำ เว็บไซต์หรือผลิตภัณฑ์ที่ผู้ใช้คนนี้ อาจจะสนใจในอนาคต อีกตัวอย่างหนึ่งของข้อมูลเชิงกราฟได้แก่ กราฟของเพื่อนใน Facebook ในที่นี้ จุดต่อก็คือสมาชิก Facebook หนึ่งคน และจุดต่อสองจุดจะมีเส้นเชื่อมก็ต่อเมื่อสมาชิกสองคนนั้นเป็นเพื่อนกัน โดยอาศัยข้อมูลดัง

กล่าว Facebook สามารถวิเคราะห์ความเชื่อมโยง และนำมาซึ่งการแนะนำเพื่อนให้กับสมาชิกต่อไป

1.2.2 ผลลัพธ์จากการทำเหมืองข้อมูล

เราได้เห็นแล้วว่าขั้นตอนวิธีในการทำเหมืองข้อมูล สามารถนำไปประยุกต์ใช้เพื่อหา แบบรูป สารสนเทศ หรือความรู้ จากข้อมูลหลากหลายประเภท ต่อจากนี้ไปเราจะมาศึกษาว่า มีสารสนเทศหรือความรู้แบบไหนบ้างที่ขั้นตอนวิธีในการทำเหมืองข้อมูลจะสามารถสกัดออกมาได้

โดยทั่วไปแล้ว งานทางด้านการทำเหมืองข้อมูลสามารถแบ่งออกได้เป็น 2 ประเภทคือ งานเชิงอธิบาย (Descriptive) และงานเชิงพยากรณ์ (Predictive)

งานในเชิงอธิบายหมายความถึงการพยายามที่จะสรุปลักษณะสำคัญของข้อมูล ในขณะที่งานในเชิงพยากรณ์นั้น พยายามที่จะทำนายลักษณะของข้อมูลในอนาคต ซึ่งไม่เคยเห็นมาก่อน แน่นอนว่า เราจำเป็นจะต้องมีสมมุติฐานว่าข้อมูลในอนาคตนั้น ถูกสุ่มจากการแจกแจงทางสถิติเดียวกันกับข้อมูลในอดีต

คำอธิบายมโนทัศน์หรือกลุ่มของข้อมูล

ข้อมูลหนึ่งๆอาจสามารถเชื่อมโยงเข้ากับ กลุ่มของข้อมูล (Class) หรือมโนทัศน์ (Concept) บางอย่าง ยกตัวอย่างเช่น รถยนต์สามารถจัดกลุ่มได้เป็น รถกระบะ รถยนต์นั่ง รถยนต์เพื่อการพาณิชย์ หรือสำหรับแวดวงธนาคาร เราก็สามารถเชื่อมโยงลูกค้าเงินกู้ของธนาคาร เข้ากับมโนทัศน์ของลูกค้าชั้นดี หรือลูกค้าชั้นเลว ฉะนั้นแล้วการจะสื่อถึงลักษณะสำคัญของกลุ่มของข้อมูล หรือมโนทัศน์นั้น เราจำเป็นจะต้องหาคำอธิบายสรุปลักษณะสำคัญของกลุ่มต่างๆที่เที่ยงตรงและกระชับ คำอธิบายเหล่านั้นเรียกว่า คำอธิบายมโนทัศน์หรือกลุ่มของข้อมูล (Concept/Class Descriptions)

แนวทางในภาพรวมในการหาคำอธิบายของกลุ่ม อาจเริ่มจากการระบุลักษณะของกลุ่ม (Data Characterisation) ให้ได้ พอได้ลักษณะเฉพาะของกลุ่มแล้ว เราสามารถนำลักษณะของแต่ละกลุ่มมาเปรียบเทียบกันเพื่อแยกลักษณะที่จำเพาะสำหรับกลุ่มนั้น ซึ่งไม่สามารถพบลักษณะดังกล่าวได้ในกลุ่มอื่น วิธีดังกล่าวเรียกว่าการแบ่งแยกข้อมูล (Data Discrimination) ระเบียบวิธีข้างต้น สามารถทำได้ด้วยมือ (Manually) แต่เราจะศึกษาในบทถัดไปถึงระเบียบวิธีที่เป็นอัตโนมัติมากขึ้น ซึ่งประยุกต์ขั้นตอนวิธีจากการเรียนรู้ของเครื่องมาใช้ในการดึงลักษณะพิเศษ ที่รู้จักกันในชื่อว่าการทำ Feature Extraction

แบบรูปที่พบบ่อยและกฎความสัมพันธ์

Frequent Patterns หมายความตรงตัวก็คือ แบบรูปที่เกิดขึ้นและพบได้บ่อยในข้อมูล รูปแบบในที่นี้มีได้หลายประเภท ตั้งแต่เซตของสิ่งของ ลำดับย่อย หรือโครงสร้างย่อย แบบรูปที่พบบ่อยจะนำมาซึ่งกฎความสัมพันธ์

ดังตัวอย่างต่อไปนี้

$$\text{buys}(X,\text{“bike”}) \Rightarrow \text{buys}(X,\text{“helmet”})$$

[support=3%, confidence=60%]

กฎข้างบนใช้สื่อว่า ถ้าลูกค้าซื้อจักรยานแล้วลูกค้าจะซื้อหมวกกันน็อคด้วย แน่นอนว่าเราสามารถสร้างกฎในลักษณะนี้ขึ้นมาได้เรื่อยๆ หากพบว่ามีความสัมพันธ์แบบนี้อยู่ในฐานข้อมูลการซื้อขาย ฉะนั้นการจะแยกแยะว่า กฎไหนเป็นกฎที่มีคุณค่าและน่าสนใจ เราจะวัดด้วยมาตรวัดความน่าสนใจ (Measure of Interestingness) สองตัว ที่เรียกว่า ค่าความเชื่อมั่น (Confidence) และค่าสนับสนุน (Support) ในที่นี้ ค่าความเชื่อมั่นเท่ากับ 60% หมายความว่า หากลูกค้าซื้อจักรยานจะมีโอกาส 60% ที่ลูกค้าจะซื้อหมวกกันน็อคด้วย ส่วนค่าสนับสนุน จะหมายถึงสัดส่วนของจำนวนลูกค้าที่ซื้อจักรยาน ต่อจำนวนการซื้อของทั้งหมดในฐานข้อมูล กฎข้างต้นเรียกว่า กฎความสัมพันธ์แบบมิติเดียว (Single-dimension Association Rule) แต่ในความเป็นจริงเราอาจสามารถสร้างกฎความสัมพันธ์ในมิติที่สูงขึ้นได้ เช่น

$$\text{buys}(X,\text{“bike”}) \wedge \text{age}(X,\text{“30...45”}) \Rightarrow \text{buys}(X,\text{“helmet”})$$

[support=2%, confidence=50%]

โดยทั่วไปแล้ว หากกฎความสัมพันธ์ที่หาได้ มีค่าสนับสนุนและค่าความเชื่อมั่น ต่ำกว่าขีดแบ่งค่าสนับสนุนน้อยที่สุด (Minimum Support Threshold) และต่ำกว่าขีดแบ่งค่าความเชื่อมั่นน้อยที่สุด (Minimum Confidence Threshold) จะถือว่า ก็ดังกล่าวเป็นกฎความสัมพันธ์ที่ไม่น่าสนใจ

การจำแนกข้อมูลและการทำนาย

การจำแนกข้อมูลคือกระบวนการในการหาแบบจำลอง (Model) ที่สามารถอธิบายความสัมพันธ์ระหว่างข้อมูลและกลุ่มหรือมโนทัศน์ของข้อมูลได้อย่างถูกต้อง เป้าหมายของการหาแบบจำลองนี้ ก็เพื่อที่จะใช้แบบจำลองดังกล่าว ในการทำนายกลุ่มหรือมโนทัศน์ ของข้อมูลในอนาคตได้อย่างถูกต้องแม่นยำ แบบจำลองดังกล่าว จะได้จากการวิเคราะห์ข้อมูลฝึกหัดที่มีป้ายบอกกลุ่มกำกับ (Labelled Training Data) แบบจำลองที่ได้จากการวิเคราะห์จะเรียกกันว่า ตัวจำแนก (Classifier) ตัวอย่างของ ตัวจำแนกที่เป็นที่นิยมใช้กันมาก ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree), โครงข่ายประสาทเทียม (Artificial Neural Network), เครื่องกลเวคเตอร์สนับสนุน (Support Vector Machine) และการถดถอยโลจิสติก (Logistic Regression) เราจะกลับมาศึกษาการจำแนกข้อมูลโดยละเอียดในบทที่ 5

การจัดกลุ่มข้อมูล

การจัดกลุ่มข้อมูลจะแตกต่างกับการจำแนกกลุ่มข้อมูลตรงที่ว่า ข้อมูลฝึกหัดจะไม่มีป้ายบอกกลุ่มของข้อมูลมาให้ด้วย แต่เป็นหน้าที่ของขั้นตอนวิธีที่จะทำการจัดกลุ่มให้ข้อมูล การจัดกลุ่มของข้อมูลโดยทั่วไปจะอยู่บนพื้นฐานที่ว่า ข้อมูลที่อยู่ในกลุ่มเดียวกันจะต้องมีความเหมือนกันมากที่สุด และข้อมูลระหว่างกลุ่มจะต้องมีความแตกต่างกันมากที่สุด ปัญหาและความท้าท้ายหลักของการหากลุ่มข้อมูลก็คือ การเลือกใช้มาตรวัดความเหมือน (Similarity Measure) ที่เหมาะสมที่สุดสำหรับข้อมูลชุดนั้น รายละเอียดเนื้อหาของการจัดกลุ่มข้อมูลจะอยู่ในบทที่ 6

การวิเคราะห์ค่าผิดปกติ (Outlier Analysis)

ค่าผิดปกติ (Outlier) คือส่วนหนึ่งของข้อมูลที่ไม่สอดคล้องกับตัวแบบ แนวโน้มหรือพฤติกรรมโดยรวมของข้อมูลส่วนใหญ่ ปกติแล้วค่าผิดปกติมักถูกกำจัดออกไป เพราะถือเป็นสัญญาณรบกวน (Noise) แต่ในบางครั้ง ค่าผิดปกติอาจจะมีประโยชน์ได้ เช่นในการคัดกรองอีเมล [Jindal and Liu, 2007] จะพบว่าค่าผิดปกติในบรรดาอีเมลทั้งหลาย ก็คือเหล่าอีเมลขยะนั่นเอง เช่นเดียวกัน การตรวจจับค่าผิดปกติอาจนำไปใช้ในการวิเคราะห์หารูปแบบการฝอกเงิน โดยดูจากรูปแบบธุรกรรม ที่ผิดปกติไปจากธุรกรรมอื่นอย่างเช่นกรณีศึกษาที่พบใน [Fawcett and Provost, 1997]

การวิเคราะห์วิวัฒนาการ

การวิเคราะห์วิวัฒนาการของข้อมูลก็เป็นอีกสาขาหนึ่งที่ได้รับความสนใจ การวิเคราะห์ลักษณะนี้ มีข้อแตกต่างจากการวิเคราะห์อื่นๆ ตรงที่ว่าจะไม่มีการกำหนดสมมุติฐานว่าข้อมูลที่สำรวจได้ จะมาจากการแจกแจงทางสถิติแบบเดิมตลอด แต่ว่าการแจกแจงหรือลักษณะของข้อมูลอาจจะเปลี่ยนไปได้ตามกาลเวลา ยกตัวอย่างเช่น การวิเคราะห์ค่าเงินที่เราไม่สามารถจะใช้แบบจำลองตัวเดิม ในการพยากรณ์หรือคาดการณ์ค่าเงินในทุกๆช่วงเวลาได้ เพราะลักษณะของข้อมูลค่าเงิน รวมถึงการแจกแจงทางสถิติของค่าเงินมีการเปลี่ยนแปลงเสมอ เนื่องมาจากปัจจัยหลายๆประการ การวิเคราะห์วิวัฒนาการ สามารถนำมาประยุกต์เพื่อตรวจจับการเปลี่ยนแปลงของข้อมูล เพื่อนำไปสู่การสร้างแบบจำลองใหม่ ให้เข้ากับข้อมูลในช่วงเวลานั้นๆ

1.2.3 ความน่าสนใจของผลลัพธ์

ถึงแม้ว่าเราจะหาแบบรูป กฎความสัมพันธ์ กลุ่มข้อมูล หรือสร้างตัวจำแนกข้อมูลได้สำเร็จ ก็มิได้หมายความว่า ผลลัพธ์เหล่านั้นจะน่าสนใจเสมอไป ในทางปฏิบัติแล้วผลลัพธ์ที่ได้จากการทำเหมืองข้อมูล จะมีความน่าสนใจก็ต่อ เมื่อผลลัพธ์ดังกล่าวมีคุณสมบัติดังต่อไปนี้ [Han et al., 2011]

- สามารถเข้าใจและตีความได้ง่าย
- สามารถใช้กับข้อมูลใหม่ที่ไม่เคยเจอมาก่อนได้อย่างถูกต้องในเปอร์เซนต์สูง
- มีศักยภาพที่จะนำไปประยุกต์ใช้ให้เกิดประโยชน์ได้
- มีความใหม่ ไม่เคยมีใครพบมาก่อน

ในบทนี้เราได้ทราบถึงความหมายของการทำเหมืองข้อมูล รวมถึงเป้าหมายหลักๆของการทำเหมืองข้อมูลแล้ว บทต่อไปเราจะศึกษาเครื่องมือพื้นฐานที่สำคัญในการทำเหมืองข้อมูล ซึ่งจะนำไปสู่การศึกษาขั้นตอนการเตรียมข้อมูลต่อไป

แบบฝึกหัด

1. อะไรคือข้อแตกต่างระหว่างการสืบค้นข้อมูลจากฐานข้อมูลและการทำเหมืองข้อมูล
2. กฎความสัมพันธ์ที่มีค่าความไว้วางใจสูง แต่ค่าสนับสนุนต่ำ สามารถสื่ออะไรได้บ้าง
3. ยกตัวอย่างชุดข้อมูลมาหนึ่งชนิด พร้อมเสนอแนวทางในการวิเคราะห์ข้อมูลดังกล่าว

บทที่ 2

เครื่องมือพื้นฐานและการเตรียมข้อมูลก่อนการประมวลผล

การทำเหมืองข้อมูลเป็นศาสตร์ที่เกี่ยวข้องกับศาสตร์อื่นหลายสาขา รวมทั้งมีการประยุกต์เครื่องมือในสาขาอื่นๆมาใช้งาน ในบทนี้เราจะทำความรู้จักกับนิยาม เครื่องมือและแนวคิดสำคัญ ในการทำเหมืองข้อมูลเพื่อเป็นรากฐานสำหรับการเตรียมข้อมูลก่อนการประมวลผล

การทำเหมืองข้อมูลแน่นอนว่าจะต้องเกี่ยวข้องกับข้อมูล ในวิชานี้จะใช้คำว่า “ข้อมูล” (Data) ในการเรียก ตัวอย่างของสิ่งที่เราสนใจจะศึกษา ซึ่งในบางบริบทข้อมูลอาจจะถูกอ้างถึงโดย จุดข้อมูล (Data Point), ตัวอย่าง (Example), กรณีตัวอย่าง (Instance) หรือ เวคเตอร์ขาเข้า (Input Vector) และแทนด้วยสัญลักษณ์ x กลุ่มของข้อมูลหลายๆตัวจะถูกเรียกว่า “ชุดข้อมูล” (Data Set) แทนด้วยสัญลักษณ์ $\mathcal{S}$ ตัวอย่างของชุดข้อมูลได้แก่ ชุดข้อมูลของลูกค้า ชุดข้อมูลของผู้ป่วยโรคมะเร็ง ชุดข้อมูลของสุนัข เป็นต้น

โดยส่วนใหญ่แล้ว ข้อมูลจะได้จากการสังเกตและบันทึกลักษณะทางธรรมชาติของสิ่งที่เราสนใจศึกษา ข้อมูลจะประกอบไปด้วยคุณลักษณะ (Feature) ซึ่งมีหน้าที่อธิบายสภาวะทางธรรมชาติของสิ่งที่เราสนใจในมิตินั้นๆ อย่างเช่น ข้อมูลของลูกค้าอาจประกอบไปด้วยคุณลักษณะ 3 ตัวคือ เพศ ช่วงอายุ และรายได้ คุณลักษณะ 3 มิตินี้ก็คือสภาวะทางธรรมชาติที่ระบุถึงลูกค้าคนนั้น แน่นอนว่าจำนวนหรือมิติของคุณลักษณะก็เป็นส่วนสำคัญที่ช่วยให้เราสามารถระบุตัวตนของลูกค้าได้แม่นยำมากขึ้นตามไปด้วย

ในทางคณิตศาสตร์ เราสามารถมองข้อมูลตัวที่ n จากชุดข้อมูลขนาด N ตัวว่าเป็นเวคเตอร์ตัวหนึ่ง ในปริภูมิแบบยุคลิด (Euclidean Space) ขนาด M มิติ

$$x_n = \{x_n^1, x_n^2, \ldots, x_n^m, \ldots, x_n^M\}$$

ในที่นี้ $n \in N$ จะใช้แสดงลำดับของข้อมูลในชุดข้อมูลดังกล่าว และ $m \in M$ จะหมายถึงคุณลักษณะลำดับที่ m ของข้อมูลนั้น เนื่องจากข้อมูลแต่ละตัวคือเวคเตอร์ เราสามารถแสดงชุดข้อมูลทั้งหมดได้ด้วยเมทริกซ์ขนาด $N \times M$

$$\begin{bmatrix} x_1^1 & \dots & x_1^m & \dots & x_1^M \\ \vdots & \ddots & \vdots & & \vdots \\ x_n^1 & \dots & x_n^m & \dots & x_n^M \\ \vdots & & \vdots & \ddots & \vdots \\ x_N^1 & \dots & x_N^m & \dots & x_N^M \end{bmatrix}$$

โดยที่ ข้อมูลหนึ่งตัวจะแสดงด้วยเวคเตอร์แถว (Row Vector) และคุณลักษณะในมิติหนึ่งๆของข้อมูลทุกตัวจะแสดงด้วยเวคเตอร์แนวตั้ง (Column Vector)

ตัวอย่าง 2.0.1. *กำหนดชุดข้อมูล $\mathcal{S}$ แสดงข้อมูลของผู้ป่วย 5 คน ซึ่งอธิบายด้วยคุณลักษณะคือ อายุ (ปี) น้ำหนัก (กิโลกรัม) ส่วนสูง (เซนติเมตร) ดังต่อไปนี้*

$$\begin{bmatrix} 13 & 39 & 157 \\ 23 & 56 & 157 \\ 15 & 60 & 177 \\ 32 & 72 & 187 \\ 21 & 43 & 162 \end{bmatrix}$$

จากชุดข้อมูลตัวอย่าง เราสามารถตีความได้ว่า ผู้ป่วยคนที่ 1 อายุ 13 ปี หนัก 39 กิโลกรัมและ สูง 157 เซนติเมตร ส่วนผู้ป่วยคนที่ 3 หนัก 60 กิโลกรัม เป็นต้น

2.1 ชนิดของคุณลักษณะ

การบันทึกสภาวะของธรรมชาติสามารถทำได้โดยใช้การอธิบายคุณลักษณะแบบต่างๆ คุณลักษณะหลักที่คนเราใช้แทนหรือบันทึกสภาวะธรรมชาติอาจอยู่ในเชิงตัวเลข ดังทีเราเห็นในตัวอย่างก่อนหน้าไปแล้ว แต่ว่าในบางกรณีการบันทึกสภาวะของธรรมชาติด้วยตัวเลขอาจไม่เหมาะสม เราจะมาศึกษาว่า ยังมีคุณลักษณะแบบไหนอีกบ้างที่เราสามารถใช้ บันทึกสภาวะของธรรมชาติได้บ้าง รวมทั้งคุณลักษณะเหล่านั้นมีชื่อเรียกว่าอะไร

2.1.1 คุณลักษณะเชิงนาม

คุณลักษณะเชิงนาม (Nominal Feature) เป็นคุณลักษณะที่ใช้บ่งบอกถึงสถานะ ประเภท หรือชื่อกลุ่มของคุณลักษณะนั้น มีโดเมน (Domain) เป็นเซตจำกัด (Finite Set) ยกตัวอย่างคุณลักษณะประเภทนี้ได้แก่ คุณลักษณะแสดงสีผม ซึ่งค่าของสีผมอาจมาจากเซตของสีผมต่อไปนี้ {ดำ,ทอง,แดง,น้ำตาล,ขาว,เทา} คุณลักษณะที่ใช้ในการแสดงอาชีพก็ถือเป็นคุณลักษณะเชิงนามเช่นกัน

2.1.2 คุณลักษณะแบบทวิภาค

คุณลักษณะแบบทวิภาค (Binary Feature) เป็นเซตย่อยของคุณลักษณะเชิงนาม คุณลักษณะชนิดนี้ ใช้เพื่อแสดงสภาวะของธรรมชาติที่มีเพียง 2 สถานะ (นั่นคือโดเมนมีขนาดเท่ากับ 2) คุณลักษณะแบบทวิภาค สามารถแบ่งย่อยได้เป็น 2 แบบคือ คุณลักษณะทวิภาคแบบสมมาตร (Symmetric) และแบบไม่สมมาตร (Asymmetric) เรามักเลือกใช้คุณลักษณะทวิภาคแบบสมมาตร เมื่อสถานะทั้งสองมีความน่าจะเป็นในการเกิดขึ้นเท่ากัน เช่น คุณลักษณะที่ใช้แสดงเพศ ชาย หรือหญิง ส่วนคุณลักษณะทวิภาคแบบไม่สมมาตร มักถูกใช้เมื่อสถานะทั้งสองมีความน่าจะเป็นในการเกิดขึ้นไม่เท่ากัน เช่น คุณลักษณะที่ใช้แสดงว่าผู้ป่วยเป็นโรคร้ายแรง หรือไม่เป็นโรค ซึ่งโดยธรรมชาติแล้วโอกาสที่จะพบโรคร้ายแรงดังกล่าวอาจน้อยมาก

2.1.3 คุณลักษณะเชิงลำดับ

คุณลักษณะเชิงลำดับ (Ordinal Feature) คือคุณลักษณะที่สถานะต่างๆมีลำดับในตัวมันเอง หมายความว่าเราจะต้องสามารถบอกได้ว่า สถานะนี้มาก่อนหรือหลังอีกสถานะหนึ่ง อย่างไรก็ตามระยะห่างที่ชัดเจนของสถานะทั้งสองอาจไม่สามารถวัดได้ ยกตัวอย่างคุณลักษณะเชิงลำดับ ได้แก่ คุณลักษณะที่ใช้แทนยศ หรือใช้แทนคะแนนความชอบต่อสิ่งๆหนึ่ง ในกรณีตัวอย่างของการประเมินโรงแรม เรารู้ว่าโรงแรม 5 ดาว น่าจะดีกว่าโรงแรม 4 ดาว แต่เราไม่สามารถบอกในเชิงปริมาณได้ชัดเจนว่าระยะห่าง 1 หน่วยคะแนนต่างกันแค่ไหน

2.1.4 คุณลักษณะเชิงตัวเลข

คุณลักษณะเชิงตัวเลข (Numeric Feature) เป็นคุณลักษณะที่พบและใช้อธิบายสภาวะของธรรมชาติมากที่สุด อาจอยู่ในรูปของจำนวนเต็ม (Integer) หรือจำนวนจริง (Real Number) ตัวอย่างของคุณลักษณะแบบนี้ ได้แก่ น้ำหนัก ส่วนสูง รายได้ อายุ เวลา ระยะทาง ระดับน้ำตาลในเลือด ความจุถังน้ำมัน ปริมาตรกระบอกสูบ ความเร็ว เป็นต้น

2.2 สถิติพื้นฐานของข้อมูล

สถิติพื้นฐานของข้อมูลเป็นสิ่งสำคัญที่จะนำไปสู่ความเข้าใจในลักษณะและพฤติกรรมของข้อมูล สถิติที่จะกล่าวถึงในส่วนนี้ มี 2 ประเภท ได้แก่ ค่ากลางของข้อมูล และการกระจายตัวของข้อมูล การวัดค่ากลางของข้อมูลเบื้องต้น นิยมวัดด้วยสถิติ 3 แบบได้แก่ ค่าเฉลี่ย (Mean) มัธยฐาน (Median) และ ฐานนิยม (Mode)

2.2.1 ค่าเฉลี่ย

โดยทั่วไปเมื่อพูดถึงค่าเฉลี่ย เราจะนึกถึงค่าเฉลี่ยเลขคณิต (Arithmetic Mean) ซึ่งถือได้ว่าเป็นการวัดค่ากลางที่นิยมใช้กันมากที่สุดตัวหนึ่ง เนื่องจากสามารถคำนวณได้ง่าย และพบเห็นการใช้งานในชีวิตประจำวันได้บ่อยครั้ง การคำนวณค่าเฉลี่ยเลขคณิตของข้อมูลที่มี 1 มิติสามารถทำได้โดย

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{2.1}$$

หากข้อมูลที่เราสนใจเป็นเวคเตอร์ M มิติ การหาค่าเฉลี่ยก็คือการหาค่าเฉลี่ยของคุณลักษณะแต่ละตัวแยกกัน

$$\bar{x} = [\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^M] = [\frac{1}{N}\sum_{n=1}^{N} x_n^1, \frac{1}{N}\sum_{n=1}^{N} x_n^1, \ldots, \frac{1}{N}\sum_{n=1}^{N} x_n^M] \tag{2.2}$$

การหาค่าเฉลี่ยเลขคณิตแบบข้างต้นเป็นกรณีพิเศษ ของการหาค่าเฉลี่ยแบบถ่วงน้ำหนัก (Weighted Average) ซึ่งคำนวณโดย

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} w_n x_n \tag{2.3}$$

ในที่นี้น้ำหนัก w_n มีนัยในเชิงของความสำคัญของข้อมูลตัวนั้นๆ หากเรากำหนดให้ w_n ของข้อมูลทุกตัวเท่ากับ 1 การหาค่าเฉลี่ยแบบถ่วงน้ำหนักก็จะลดรูปไปเหมือนการหาค่าเฉลี่ยเลขคณิตในสมการที่ 2.1 หากน้ำหนักของข้อมูลแต่ละตัว แสดงความน่าจะเป็นที่จะเจอข้อมูล x_n โดยที่น้ำหนักของข้อมูลทุกตัวรวมกันได้ 1 นั่นคือ $\sum_{n=1}^{N} w_n = 1$ เราก็จะเรียกค่าเฉลี่ยถ่วงน้ำหนักนั้นว่า ค่าคาดหวัง (Expected Value)

โดยปกติแล้วค่าเฉลี่ยจะมีความอ่อนไหวต่อข้อมูลสุดโต่ง (Extreme Value) และค่าผิดปกติ (Outlier) หากในชุดข้อมูลมีข้อมูลลักษณะดังกล่าวรวมอยู่ด้วย การใช้ค่าเฉลี่ยตัดแต่ง (Trimmed Mean) ซึ่งหมายถึงการตัดค่าที่มากที่สุดและน้อยที่สุด k ตัวออกจากชุดข้อมูล ก่อนที่จะนำมาหาค่าเฉลี่ยก็จะทำให้ค่าเฉลี่ยที่หาได้มีความแม่นยำมากขึ้น

ตัวอย่าง 2.2.1. *การสำรวจรายได้เฉลี่ยของประชากรในจังหวัดลำปาง 6 คนพบว่า คนที่ 1 ถึง 5 มีรายได้ 10000, 15000, 13000, 20000, 30000 บาทต่อเดือนตามลำดับ แต่บังเอิญคนสุดท้ายที่ไปสุ่มสำรวจ กลับเจอเศรษฐีรายใหญ่ของจังหวัดซึ่งมีรายได้ 2000000 บาท ต่อเดือน หากใช้ค่าเฉลี่ยในการคำนวณ จะสามารถสรุปได้ว่าคนจังหวัดลำปางมีรายได้เฉลี่ยเท่ากับ*

$$\frac{10000 + 15000 + 13000 + 20000 + 30000 + 2000000}{6} = 348000$$

ซึ่งถือว่าเกินความจริง ตามหลักสถิติหากทำการสำรวจกลุ่มประชากรมามากกว่านี้ ผลกระทบของค่าสุดโต่งอาจจะน้อยลงทำให้ค่าเฉลี่ยที่ได้ก็น่าจะมีความแม่นยำมากขึ้น แต่ในภาวะที่ข้อมูลมีน้อย ค่าสุดโต่งอาจทำให้ค่าเฉลี่ยผิดเพี้ยนไปได้ ผู้ใช้ควรพิจารณาเลือกใช้การหาค่าเฉลี่ยแบบอื่น เช่น ค่าเฉลี่ยตัดแต่ง เป็นต้น

2.2.2 มัธยฐาน

มัธยฐาน (Median) เป็นอีกหนึ่งวิธีที่ใช้ในการวัดค่ากลางของข้อมูล มัธยฐานจะแสดงข้อมูลที่อยู่ในตำแหน่งกึ่งกลางของชุดข้อมูลซึ่งถูกจัดเรียงลำดับจากน้อยไปมาก การหามัธยฐานสามารถสรุปได้ดังต่อไปนี้

$$\text{median} = \begin{cases} \text{ค่าของข้อมูลตัวที่ } \frac{N}{2} + 1 & \text{เมื่อ N เป็นจำนวนคี่} \\ \text{ค่าเฉลี่ยเลขคณิตของข้อมูลตัวที่ } \frac{N}{2} \text{ และ } \frac{N}{2} + 1 & \text{เมื่อ N เป็นจำนวนคู่} \end{cases} \tag{2.4}$$

2.2.3 ฐานนิยม

ฐานนิยม (Mode) เมื่อพิจารณาตามชื่อก็คือ ค่าของข้อมูลที่ปรากฏตัวถี่มากที่สุด ในจำนวนข้อมูลทั้งหมดในชุดข้อมูล ฐานนิยมสำหรับข้อมูลที่มีการแจกแจง (Distribution) หนึ่งๆไม่จำเป็นจะต้องมีค่าเดียว ในกรณีที่การแจกแจงนั้นมีฐานนิยมตัวเดียว เราจะเรียกการแจกแจงดังกล่าวว่า การแจกแจงที่มีฐานนิยมเดียว (Unimodal) และเรียกการแจกแจงที่มีฐานนิยมมากกว่าหนึ่งตัวว่า การแจกแจงแบบหลายฐานนิยม (Multimodal)

ในการศึกษาหรือสรุปธรรมชาติของข้อมูล การวัดค่ากลางเพียงอย่างเดียวอาจจะไม่เพียงพอ ควรมีการวัดการกระจายตัวของข้อมูลด้วย เช่น คะแนนสอบของวิชาการทำเหมืองข้อมูล และการเรียนรู้ของเครื่อง ที่

มีค่าเฉลี่ยเท่ากันคือ 70 คะแนน แต่การกระจายของคะแนนวิชา การทำเหมืองข้อมูลมีค่าน้อยมากเทียบกับวิชาการเรียนรู้ของเครื่อง จากข้อมูลข้างต้น เราสามารถตั้งข้อสังเกต (โดยไม่ดูคะแนนดิบ) ได้ว่า นักศึกษาส่วนใหญ่ได้คะแนนพอๆกัน ในวิชาการทำเหมืองข้อมูล แต่สำหรับวิชาการเรียนรู้ของเครื่องนั้น มีคนที่เก่งมากและคนอ่อนมาก ปะปนกันอยู่ การวัดการกระจายของข้อมูลสามารถทำได้โดยวิธีพื้นฐานต่อไปนี้

2.2.4 ค่าความแปรปรวน

ค่าความแปรปรวน (Variance) เป็นหน่วยที่ใช้วัดว่าข้อมูลมีการกระจายตัวมากน้อยแค่ไหน ค่าความแปรปรวนสำหรับข้อมูล 1 มิติ สามารถคำนวณได้จากสมการต่อไปนี้

$$\mathrm{Var}(x) = \sigma^2 = \frac{\sum_{n=1}^{N}(x_n - \mu)^2}{N} \tag{2.5}$$

จากสมการจะเห็นว่า การวัดความแปรปรวนก็คือ การวัดค่าเฉลี่ยของระยะห่างกำลังสองของข้อมูลกับค่ากลางของข้อมูลชุดนั้น กรณีที่ค่าความแปรปรวนน้อย แสดงว่าข้อมูลทุกตัวมีค่าใกล้เคียงกับค่ากลางของข้อมูล หรืออีกนัยหนึ่งก็คือ ข้อมูลส่วนใหญ่มีค่าใกล้ๆกันทั้งหมด ไม่มีความแปรปรวนเลย หากค่าความแปรปรวนเพิ่มมากขึ้น แสดงให้เห็นว่าข้อมูลมีการกระจายตัวมาก ในทางสถิติเรานิยามให้ ค่าความแปรปรวน มีค่าเท่ากับค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ยกกำลังสอง

หากข้อมูลอยู่ในมิติที่สูงขึ้น นอกจาก ค่าความแปรปรวนในมิติหนึ่งๆแล้ว เราอาจจะสนใจสิ่งที่เรียกว่า ค่าความแปรปรวนร่วม (Covariance) ระหว่างสองมิติใดๆ ค่าความแปรปรวนร่วมระหว่าง มิติ i, j ใดๆ สามารถคำนวณได้โดย

$$\mathrm{Cov}(x^i, x^j) = \sigma_i \sigma_j = \frac{\sum_{n=1}^{N}(x_n^i - \mu_i)(x_n^j - \mu_j)}{N} \tag{2.6}$$

ในที่นี้ σ_i แทนค่าเบี่ยงเบนมาตรฐานของข้อมูลคิดเฉพาะมิติที่ i ส่วน μ_i แสดงค่ากลางของข้อมูล สำหรับมิติที่ i

เราจะพบว่าหากเราต้องการคำนวณ ค่าความแปรปรวนร่วม ของมิติแบบพบกันหมด สิ่งที่เราได้จะเรียกว่า เมทริกซ์ของค่าความแปรปรวนร่วม (Covariance Matrix) ซึ่งหาได้โดย

$$\mathrm{Cov}(x) = \Sigma = \frac{\sum_{n=1}^{N}(x_n - \mu)^T(x_n - \mu)}{N} \tag{2.7}$$

ในที่นี้ $(x_n - \mu)^T$ แทน การสลับแถวและหลักของเมทริกซ์ (Transposition) ส่วน μ คือ เวคเตอร์ของค่ากลางข้อมูล ผลลัพธ์ที่ได้จะคือ เมทริกซ์ขนาด $M \times M$ แสดงค่าความแปรปรวนร่วมของมิติทุกๆมิติของข้อมูล

$$\begin{bmatrix} \sigma_1^2 & \dots & \sigma_1\sigma_m & \dots & \sigma_1\sigma_M \\ \vdots & \ddots & \vdots & & \vdots \\ \sigma_m\sigma_1 & \dots & \sigma_m^2 & \dots & \sigma_m\sigma_M \\ \vdots & & \vdots & \ddots & \vdots \\ \sigma_M\sigma_1 & \dots & \sigma_M\sigma_m & \dots & \sigma_M^2 \end{bmatrix}$$

สังเกตได้ว่าเมทริกซ์ของค่าความแปรปรวนร่วม จะมีลักษณะสมมาตร (Symmetric) และเป็นจัตุรัส (Square) โดยค่าความแปรปรวนร่วมบนเส้นทแยงมุม (Diagonal) ก็คือค่าความแปรปรวนของตัวแปรในมิตินั้นๆ เรายังพบอีกว่า $\sigma_i\sigma_j = \sigma_j\sigma_i$

2.2.5 ควอนไทล์และระยะระหว่างควอนไทล์

ควอนไทล์ (Quantile) คือเซต Q ของจุดแบ่งที่แบ่งเรนจ์ (Range) ของการแจกแจงทางสถิติออกเป็น n ส่วน โดยความน่าจะเป็นที่จะเจอค่าจากส่วนต่างๆมีค่าเท่ากัน จุดเหล่านั้นเรียกโดยรวมว่า จุดควอนไทล์ สังเกตว่าจะมีจำนวนจุดควอนไทล์ ในเซตดังกล่าวน้อยกว่า จำนวนส่วนอยู่ 1 เสมอ นั่นคือ $|Q| = n - 1$ ทั้งนี้ค่า n ต่างๆกันจะมีชื่อเรียกต่างกัน เช่น

- เมื่อ $n = 2$ จะเรียกจุดนั้นว่า มัธยฐาน
- เมื่อ $n = 4$ จะเรียกจุดนั้นว่า ควอร์ไทล์
- เมื่อ $n = 100$ จะเรียกจุดนั้นว่า เปอร์เซ็นไทล์

จากนิยามของควอนไทล์ ข้างต้น เราสามารถสร้างการวัดการกระจายอีกแบบ คือการวัดระยะห่างระหว่างควอร์ไทล์ที่ 3 (Q3) จาก ควอร์ไทล์ที่ 1 (Q1) ระยะดังกล่าว เรียกกันว่าระยะระหว่างควอนไทล์ (Inter-Quartile Range (IQR)) หากชุดข้อมูลมีการกระจายตัวมาก Q3 และ Q1 จะห่างกันมาก ส่งผลให้ IQR มีค่ามากไปด้วย ตรงกันข้ามชุดข้อมูล ที่มีการกระจายตัวต่ำจะมี IQR น้อย

2.3 การสร้างมโนภาพให้ข้อมูล

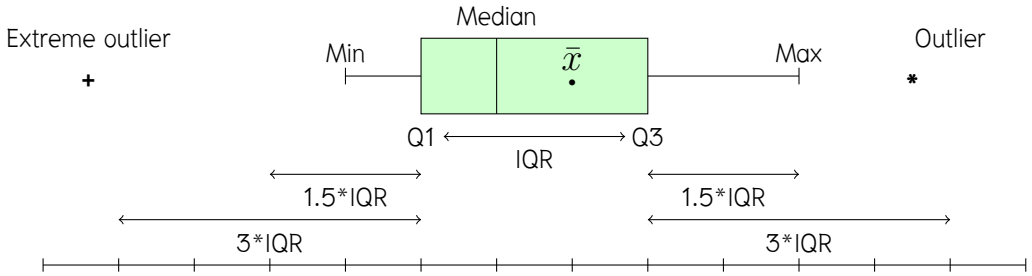
หนึ่งในขั้นตอนสำคัญก่อนการลงมือทำเหมืองข้อมูล คือการแปลงข้อมูลให้ออกมาในรูปภาพที่มองเห็น เข้าใจด้วยตาได้ง่าย วิธีการนี้ เรียกว่า การสร้างมโนภาพ (Visualisation) ประโยชน์ของการสร้างมโนภาพ คือทำให้เข้าใจข้อมูลในมุมมองใหม่ สามารถสรุปชุดข้อมูลที่มีขนาดมหาศาล ให้มองเห็นภาพ เช่น การจัดกลุ่มลูกค้า 100000 คน ว่าจัดกลุ่มได้เป็นกี่กลุ่มสำคัญ นอกจากนั้น การสร้างมโนภาพ ยังอาจทำให้เห็นความผิดปกติของข้อมูล เช่น ค่าผิดปกติ (Outlier) การสร้างมโนภาพของข้อมูลขนาดใหญ่ เป็นศาสตร์ที่แตกแยกออกไปอีกแขนง มีเทคนิควิธีการที่ซับซ้อน ซึ่งอยู่นอกเหนือจาก เอกสารการสอนนี้ ในที่นี้จะกล่าวถึงวิธีการสร้างมโนภาพ อย่างง่ายที่ใช้กันอย่างแพร่หลาย หากผู้อ่านสนใจสามารถอ่าน เพิ่มเติมได้จาก [Hoffman and Grinstein, 2002] ในที่นี้เราจะกล่าวถึงการสร้างมโนภาพ จากข้อมูลพื้นฐานทางด้านสถิติ ซึ่งมีวิธีการที่ใช้กันมากดังนี้

2.3.1 แผนภาพแบบกล่อง

แผนภาพแบบกล่อง (Boxplot) เป็นการแสดงผลเชิงรูปภาพของ สถิติ 5 ตัว (Five Numbers Summary) ซึ่งประกอบไปด้วย

1. ค่าน้อยที่สุด (Minimum)
2. ค่าควอร์ไทล์ ที่ 1
3. ค่ามัธยฐาน
4. ค่าควอร์ไทล์ ที่ 3
5. ค่ามากที่สุด (Maximum)

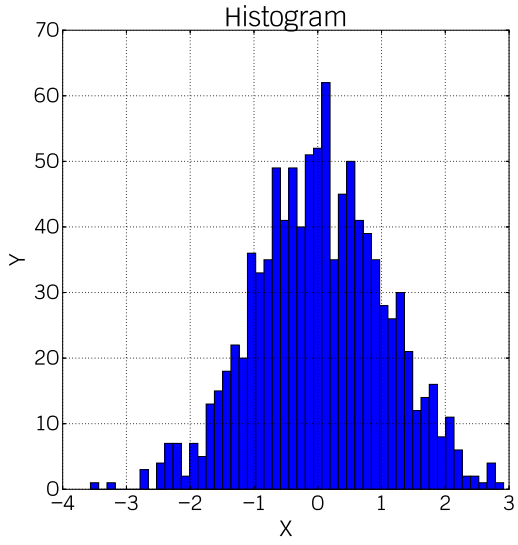
นอกจากจะทำให้เห็นภาพรวมของการกระจายตัวของข้อมูลแล้ว แผนภาพแบบกล่องยังสามารถแสดงข้อมูลที่อาจจะเป็นค่าผิดปกติ ได้ด้วย ตัวอย่างของแผนภาพแบบกล่อง แสดงให้ดูในภาพที่ 2.1 ในที่นี้ ค่าผิดปกติอ่อนๆ (Mild Outlier) อาจ สามารถกำหนด ให้คือข้อมูลที่อยู่เกินจากระยะ Q1 $-\ 1.5\ \times$ IQR หรือเกินระยะ Q3 $+\ 1.5\ \times$ IQR ขึ้นไป ส่วน ค่าผิดปกติแบบสุดโต่ง (Extreme Outlier) สามารถนิยามได้เป็นข้อมูลที่อยู่เกินจากระยะ Q1 $-\ 3\ \times$ IQR หรือเกินระยะ Q3 $+\ 3\ \times$ IQR ขึ้นไป



รูปภาพ 2.1: แผนภาพแบบกล่องและตัวเลขทางสถิติ 5 ตัว รวมถึงค่าผิดปกติต่างๆ (อัตราส่วนอาจไม่ตรง)

2.3.2 ฮิสโทแกรม

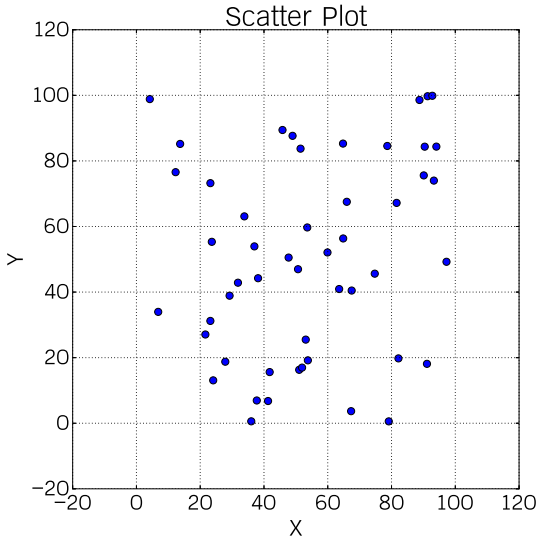
การสร้างมโนภาพที่ได้รับความนิยมอีกวิธีหนึ่งก็คือ การสร้างฮิสโทแกรม (Histogram) โดยการแบ่งข้อมูลออกเป็นช่วง แล้วทำการนับความถี่ที่ข้อมูลตกในช่วงนั้นๆ ทำให้เราทราบถึงการกระจายของข้อมูลคร่าวๆว่าเป็นแบบไหน ปกติแล้วฮิสโทแกรม จะนิยมใช้แสดงภาพของข้อมูลตัวแปรเดียว หรือสองตัวแปร แต่ทางทฤษฎีแล้วสามารถสร้างฮิสโทแกรมของข้อมูล M มิติได้ ทว่าการแสดงออกมาเป็นแผนภาพอาจจะต้อง แสดงทีละ 1 หรือ 2 มิติเรียงกันไปจนครบ M มิติ ตัวอย่างของฮิสโทแกรมของข้อมูลจำลองแสดงไว้ในภาพที่ 2.2



รูปภาพ 2.2: ฮิสโทแกรมของข้อมูลที่สุ่มจากการกระจายตัวแบบปกติ 1000 ตัว

2.3.3 แผนภาพการกระจาย

เนื่องจากข้อมูลแต่ละตัวสามารถมองเป็นจุดหนึ่งจุดบนปริภูมิแบบยูคลิดได้ ด้วยเหตุนี้จึงมีการใช้แผนภาพการกระจาย (Scatter Plot) ในการแสดงจุดเหล่านั้นบนระนาบ เพื่อเห็นภาพรวมของชุดข้อมูล ว่าข้อมูลมีกลุ่มก้อนอย่างไร แผนภาพการกระจายส่วนมากนิยมทำกันในแบบ 2 มิติหรือ 3 มิติ สำหรับข้อมูลมากกว่า 3 มิติ สามารถทำได้โดย ทำแผนภาพการกระจายของมิติเป็นคู่ๆ ตัวอย่างแผนภาพการกระจายใน 2 มิติ แสดงไว้ในภาพที่ 2.3



รูปภาพ 2.3: แผนภาพการกระจายของข้อมูลจำลอง

2.4 วิธีการวัดความคล้าย

บ่อยครั้งที่การทำเหมืองข้อมูลจำเป็นต้องวัดความคล้าย (Similarity) ของข้อมูลสองตัว เพื่อประกอบการประมวลผล ตัวอย่างเช่น ในการจัดกลุ่มข้อมูล เราจำเป็นจะต้องมีมาตรวัดเพื่อบอกว่า ข้อมูลตัวนี้ห่างจากค่ากลางของกลุ่มข้อมูลเท่าไหร่ โดยทั่วไปแล้ว มาตรวัดความคล้าย (Similarity Measure) จะมีค่าเป็นตัวเลขที่ยิ่งมีค่ามากแสดงว่า วัตถุสองชิ้นนั้นมีความเหมือนกันมาก และในทางกลับกันถ้าค่าความเหมือนมีค่าน้อย ก็แสดงว่าข้อมูลสองตัวนั้นแตกต่างกันมาก

ในบางครั้ง แทนที่จะวัดความเหมือนเราอาจจะวัดส่วนกลับของความเหมือน นั่นคือความต่างของข้อมูล (Dissimilarity) สองตัว ในกรณีนี้ ค่าความต่างจะมีค่าน้อยสำหรับข้อมูลที่เหมือนกัน และค่าความต่างจะมีค่ามากสำหรับข้อมูลที่ไม่เหมือนกัน

โดยปกติแล้วการทำเหมืองข้อมูลจะเกี่ยวข้องกับข้อมูลมากกว่า 2 ตัว ดังนั้นจึงนิยมใช้เมทริกซ์ซึ่งเรียกว่าเมทริกซ์ความคล้าย (Similarity Matrix) (หากใช้มาตรวัดความเหมือนในการคำนวณ) หรือ เมทริกซ์ความต่าง (Dissimilarity Matrix) หากใช้มาตรวัดความต่างในการคำนวณ ตัวอย่างของเมทริกซ์ความต่างสามารถแสดงได้ดังข้างล่าง

$$\begin{bmatrix} 0 & & & & \\ d(x_2,x_1) & 0 & & & \\ d(x_3,x_1) & d(x_3,x_2) & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ d(x_N,x_1) & d(x_N,x_2) & \cdots & d(x_N,x_{N-1}) & 0 \end{bmatrix} \tag{2.8}$$

ในที่นี้ $d(x_i, x_j)$ ใช้แสดงฟังก์ชันที่วัดระยะห่างระหว่างจุดข้อมูล x_i และจุดข้อมูล x_j. ซึ่งรูปแบบของฟังก์ชัน $d(\cdot)$ สามารถนิยามได้หลากหลาย และเราจะมาดูกันว่า เราสามารถนิยามฟังก์ชันวัดระยะห่างที่เหมาะสมกับข้อมูล ที่ถูกแสดงด้วยคุณลักษณะประเภทต่างๆกันได้อย่างไรบ้าง

2.4.1 ความคล้ายสำหรับคุณลักษณะเชิงตัวเลข

คุณลักษณะแบบตัวเลข ถือเป็นคุณลักษณะที่ใช้เก็บบันทึกสภาวะของธรรมชาติกันมากที่สุด การวัดระยะทางของข้อมูลที่มีคุณลักษณะแบบตัวเลขสามารถทำได้หลายวิธี เช่น วัดโดย ระยะห่างแมนฮัตตัน (Manhattan Distance) หรือ ระยะห่างแบบยูคลิด (Euclidean Distance) โดยการวัดระยะทางหรือระยะห่าง ดังกล่าวถือเป็นกรณีพิเศษของการวัดระยะห่างที่เรียกว่า ระยะห่างมินโควสกี้ (Minkowski Distance) ระหว่างเวคเตอร์ $x_i, x_j \in \mathcal{R}^M$ ซึ่งนิยามไว้ดังนี้

$$d(x_i, x_j) = (|x_i^1 - x_j^1|^h + |x_i^2 - x_j^2|^h + \cdots + |x_i^M - x_j^M|^h)^{1/h} \tag{2.9}$$

ระยะห่างมินโควสกี้มีคุณสมบัติสำคัญ 3 ประการคือ

- การเป็นบวกแน่นอน (Positive Definiteness): $d(x_i, x_j) > 0$, if $i \neq j$ and $d(x_i, x_i) = 0$
- สมมาตร (Symmetry): $d(x_i, x_j) = d(x_j, x_i)$

- อสมการอิงรูปสามเหลี่ยม (Triangle Inequality): $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$

สำหรับค่า h ที่ต่างกันจะมีชื่อเรียกมาตรวัดระยะห่างต่างกัน

กรณี $h = 1$ เราจะเรียกระยะทางดังกล่าวว่า ระยะห่างแมนฮัตตัน, L1 นอร์ม, หรือ ระยะห่างช่วงตึก (Cityblock Distance)

$$d(x_i, x_j) = (|x_i^1 - x_j^1| + |x_i^2 - x_j^2| + \cdots + |x_i^M - x_j^M|) \tag{2.10}$$

กรณี $h = 2$ เราจะเรียกระยะทางดังกล่าวว่าระยะห่างแบบยูคลิด หรือ L2 นอร์ม

$$d(x_i, x_j) = \sqrt{(x_i^1 - x_j^1)^2 + (x_i^2 - x_j^2)^2 + \cdots + (x_i^m - x_j^m)^2} \tag{2.11}$$

กรณี $h = \infty$ เราจะเรียกระยะทางดังกล่าวว่า ระยะห่างซูพรีมัม (Supremum Distance), Lmax นอร์ม, L_∞-นอร์ม

$$d(x_i, x_j) = \max_f |x_i^f - x_j^f| \tag{2.12}$$

สังเกตว่าการวัดระยะห่างซูพรีมัม ก็คือการวัดระยะทางที่ถือเอาความห่างของคุณลักษณะที่ต่างกันมากที่สุดเป็นตัวตัดสินนั่นเอง

2.4.2 ความคล้ายสำหรับคุณลักษณะแบบทวิภาค

ดังที่ได้อธิบายไปก่อนหน้านี้ คุณลักษณะแบบทวิภาคคือคุณลักษณะที่สามารถเก็บค่าได้เพียง 2 สถานะ เช่น ใช่หรือไม่ใช่ เป็นโรคหรือไม่เป็นโรค โดยทั่วไปแล้วเราอาจแทนสถานะทั้งสองด้วยเลข 0 และเลข 1 ดังนั้นการนำคุณลักษณะแบบทวิภาคของข้อมูลสองตัว มาเปรียบเทียบกันจึงมีผลลัพธ์ได้เพียง 4 แบบคือ เป็นเลข 1 เหมือนกัน เป็นเลข 0 เหมือนกัน หรือคุณลักษณะของข้อมูลอันหนึ่งเป็น 1 ส่วนอีกอันหนึ่งเป็น 0 และกลับกัน ดังนั้นเราจึงสามารถสรุปความเหมือน-ต่าง ออกมาให้อยู่ในรูปของตารางที่เรียกว่าตารางการจร (Contingency Table) ดังนี้

การวัดระยะห่างระหว่าง ข้อมูลที่ถูกบันทึกไว้ด้วยคุณลักษณะแบบทวิภาค จะมีพื้นฐานมาจากข้อมูลบนตารางการจร แต่ก่อนจะนิยามระยะห่าง หากเรานึกย้อนไปถึงเนื้อหาในตอนต้น เราจะพบว่าคุณลักษณะแบบ

	1	0
1	q	r
0	s	t

ตาราง 2.1: ตารางการจรแสดงผลการเปรียบเทียบข้อมูลที่คุณสมบัติเป็นแบบทวิภาค

ทวิภาค ถูกแบ่งออกเป็นประเภทย่อย 2 ประเภท นั่นคือ แบบสมมาตรและแบบอสมมาตร ซึ่งประเภทย่อยของคุณลักษณะนี้ จะมีผล ต่อการนิยาม การคำนวณระยะห่าง

สำหรับการคำนวณระยะห่างสำหรับ คุณลักษณะแบบทวิภาคแบบสมมาตร สามารถทำได้โดย

$$d(x_i, x_j) = \frac{\text{จำนวนคุณลักษณะที่ไม่เหมือนกัน}}{\text{จำนวนคุณลักษณะทั้งหมด}} = \frac{r+s}{q+r+s+t} \tag{2.13}$$

ทว่าหากคุณลักษณะนั้นเป็นแบบอสมมาตร หมายความว่าโอกาสที่เราจะเจอค่า 1 จะน้อยกว่าค่า 0 มากๆ ยกตัวอย่างเช่น หากกำหนดให้ค่า 1 แทนจำนวนคนที่เป็นโรคร้ายแรงที่มีอัตราการเกิดขึ้นน้อย เราจะพบว่าค่า t ในตารางการจรจะมีมากกว่าค่า q,r,s มากๆ หากเรานำค่า t มาใช้ ค่าของ t อาจส่งผลให้ระยะห่างที่คำนวณได้น้อยกว่าความเป็นจริงเกินไป ดังนั้นสำหรับคุณลักษณะแบบอสมมาตร เราจะคำนวณโดย

$$d(x_i, x_j) = \frac{\text{จำนวนคุณลักษณะที่ไม่เหมือนกัน}}{\text{จำนวนคุณลักษณะทั้งหมด-จำนวนคุณลักษณะที่เกิดบ่อย}} = \frac{r+s}{q+r+s} \tag{2.14}$$

ตัวอย่าง 2.4.1. *กำหนดข้อมูลการตรวจโรคของคนสามคนดังนี้ โดยที่ P ย่อมาจาก Positive (แทนว่าตรวจ*

ชื่อ	มีไข้	ไอ	ผลตรวจ 1	ผลตรวจ 2	ผลตรวจ 3	ผลตรวจ 4
สำรวย	P	N	P	N	N	N
ปลาด้า	P	N	P	N	P	N
แองเจลิส	P	P	N	N	N	N

เจอ) และ N ย่อมาจาก Negative (แทนว่าตรวจไม่เจอ) เนื่องจากมีความเป็นไปได้ว่าคุณลักษณะของเราเป็นแบบอสมมาตร เราสามารถคำนวณความแตกต่าง ได้ดังนี้

$$d(\text{สำรวย, ปลาด้า}) = \frac{0+1}{2+0+1} = 0.33$$

d(*สำรวย, แองเจลิส*) $= \frac{1+1}{1+1+1} = 0.67$

d(*ปลาด้า, แองเจลิส*) $= \frac{1+2}{1+1+2} = 0.75$

จากผลลัพธ์ที่ได้ เราพบว่าอาการโรคของปลาด้ามีความใกล้เคียงกับสำรวยมากกว่า (เนื่องจากมีความแตกต่างน้อย) อาจนำไปสู่การจัดกลุ่มเพื่อการเฝ้าระวังและดูแลรักษาไปพร้อมๆกันต่อไป

2.4.3 ความคล้ายสำหรับคุณลักษณะเชิงนาม

ดังที่ได้อธิบายไปก่อนหน้านี้คุณลักษณะเชิงนามคือคุณลักษณะที่ใช้แสดงค่าสภาวะของธรรมชาติที่เป็นไปได้มากกว่า 2 สถานะ เพื่อใช้บอกกลุ่มหรือบอกประเภท เป็นการขยายขีดความสามารถในการแทนสภาวะของคุณลักษณะแบบทวิภาคให้กว้างขึ้น ยกตัวอย่างคุณลักษณะแบบนี้ ได้แก่ คุณลักษณะที่ใช้แสดงสี ซึ่งสามารถมีค่าได้มากกว่าสองแบบ เช่น แดง สีเหลือง ดำ ขาว เป็นต้น การวัดความต่างสำหรับคุณลักษณะชนิดนี้มีด้วยกันสองวิธีหลักๆคือ

วิธีที่ 1 การเทียบความเหมือนโดยตรง

สามารถทำได้โดยการคำนวณต่อไปนี้

$$d(x_i, x_j) = \frac{M - P}{M} \tag{2.15}$$

โดยที่ M คือมิติของข้อมูล ส่วน P คือจำนวนคุณลักษณะที่เหมือนกันของ ข้อมูล x_i และข้อมูล x_j

วิธีที่ 2 ใช้การเข้ารหัสเป็นคุณลักษณะแบบทวิภาค

การเข้ารหัสเป็นคุณลักษณะแบบทวิภาค (Binary Feature Encoding) จะทำการแปลงค่าของ สถานะ (State) ต่างๆ ซึ่งถูกบันทึกไว้ในเชิงนาม (อาจเป็นตัวเลข หรือตัวอักษร) ให้อยู่ในรูปของรหัสฐานสอง (Binary Code) โดยอุดมคติแล้ว ระยะห่างของรหัสฐานสอง คู่ใดๆควรจะมีระยะห่างเท่ากัน เพราะในกรณีของคุณลักษณะเชิงนาม สถานะต่างๆ ถือว่าไม่มีลำดับในตัวมันเอง ยกตัวอย่างเช่น คุณลักษณะที่ใช้เก็บข้อมูลอาชีพ ซึ่งประกอบไปด้วยอาชีพ 3 แบบ คือ หมอ นักวิทยาศาสตร์ นักกฎหมาย อาจถูกเข้ารหัสด้วย 110, 011, 101 ตามลำดับ สังเกตว่า ระยะห่างของ 2 อาชีพใดๆจะเท่ากับ 2 ทั้งหมด เมื่อหารหัสคุณลักษณะแบบทวิภาค (Encoded Binary Feature) ได้แล้ว การเปรียบเทียบก็ใช้วิธีการเปรียบเทียบคุณลักษณะแบบทวิภาคต่อได้เลย

หากว่าสถานะนั้นมีความหมายเชิงลำดับเข้ามาเกี่ยวข้องด้วย เราอาจจะต้องเปลี่ยนไปใช้ตัววัดความห่างของข้อมูลที่เก็บด้วยคุณลักษณะเชิงลำดับ ที่จะกล่าวถึงต่อไป

2.4.4 ความคล้ายสำหรับคุณลักษณะเชิงลำดับ

การวัดความคล้ายของคุณลักษณะที่มีความหมายในเชิงลำดับ สามารถทำได้สองวิธีคือ

วิธีที่ 1 การเทียบความเหมือนโดยตรง

วิธีนี้เป็นวิธีที่ยุ่งยากน้อยที่สุด สามารถทำได้โดยการนำลำดับมาลบกันโดยตรง สำหรับข้อมูลใน 1 มิติ เราจะใช้ค่าสัมบูรณ์ (Absolute Value) ของความแตกต่าง เป็นมาตรวัดความคล้าย และสำหรับข้อมูลที่อยู่ใน 2 มิติขึ้นไป เราอาจใช้ผลรวมของค่าสัมบูรณ์ของความแตกต่างมาเป็นมาตรวัดแทน สังเกตว่าค่าต่ำสุดของมาตรวัดจะมีค่าเท่ากับ 0 (ข้อมูลสองตัวเหมือนกัน) ส่วนค่าสูงสุดนั้นไม่มีขอบเขต (Unbounded) เพราะข้อมูลอาจอยู่ในมิติที่สูงมากๆ

วิธีที่ 2 การปรับบรรทัดฐานของลำดับที่

หากเราต้องการจำกัดช่วงของระยะห่าง ให้อยู่ระหว่าง 0 ถึง 1 เราสามารถปรับบรรทัดฐานของลำดับที่ (Rank) ก่อน ซึ่งการคำนวณนี้ สามารถทำได้โดยใช้สมการ

$$Z = \frac{r-1}{K-1} \tag{2.16}$$

ในที่นี้ r คือค่าลำดับที่ที่ต้องการปรับบรรทัดฐาน และ K คือค่าสูงสุดของลำดับที่ที่เป็นไปได้

2.4.5 ความคล้ายแบบโคไซน์

ความคล้ายแบบโคไซน์ (Cosine Similarity) เป็นอีกวิธีที่นิยมในการวัดความคล้ายของเวคเตอร์ในปริภูมิแบบยูคลิด แต่แทนที่จะวัดความคล้าย ด้วยระยะทางระหว่างเวคเตอร์หนึ่งไปอีกเวคเตอร์หนึ่ง วิธีการวัดความคล้ายแบบโคไซน์จะคำนวณมุมระหว่างเวคเตอร์สองตัว หากเวคเตอร์ทำมุมกัน 0 องศาก็แปลว่าเวคเตอร์นั้นมีทิศทางไปทางเดียวกัน เราจะได้ค่าความคล้ายแบบโคไซน์เป็น 1 (เหมือนกันเชิงมุม) ค่าของความคล้ายแบบโคไซน์จะลดหลั่นไปเรื่อยๆตามค่ามุมที่เพิ่มมากขึ้น จนมีค่าสูงสุดที่ค่าความคล้ายเท่ากับ -1 นั่นคือสำหรับกรณีที่เวคเตอร์ทั้งสองตัวหันไปคนละทิศ หรือทำมุม 180 ต่อกัน อีกนัยหนึ่งก็คือเวคเตอร์ทั้งสองต่างกันมากที่สุดนั่นเอง ทั้งนี้หากเวคเตอร์สองตัวตั้งฉากกัน เราจะวัดค่าความคล้ายแบบโคไซน์ได้เท่ากับ 0 โดยทั่วไปแล้วการคำนวณความคล้ายแบบโคไซน์ทำได้โดย

$$\cos(x_i, x_j) = \frac{x_i^T x_j}{||x_i||||x_j||} \tag{2.17}$$

โดย $||x_i||$ คือ L2 นอร์มของเวคเตอร์ x_i

2.5 การเตรียมข้อมูลก่อนการประมวลผล

ดังที่กล่าวไปในตอนต้น ขั้นตอนสำคัญขั้นตอนหนึ่ง ที่จำเป็นสำหรับการทำเหมืองข้อมูลคือ ขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล ในขั้นตอนนี้ เราสนใจที่จะจัดการ ปรับปรุง ชำระ ข้อมูลที่เก็บมาได้ เพื่อให้ข้อมูลมีคุณภาพมากขึ้น ก่อนที่จะนำเข้ากระบวนการประมวลผลต่อไป คุณภาพของข้อมูลสามารถวัดได้หลายมิติ แต่ที่นิยมกันก็มีอยู่ 4 ตัวชี้วัดคือ

1. ความแม่นยำของข้อมูล (Accuracy)
2. ความครบถ้วนของข้อมูล (Completeness)
3. ความต้องกันของข้อมูล (Consistency)
4. ความเป็นปัจจุบันของข้อมูล (Timeliness)

เพื่อให้ได้มาซึ่งคุณภาพในมิติทั้ง 4 ที่กล่าวไปแล้ว งานที่จะต้องทำในกระบวนการเตรียมข้อมูลก่อนการประมวลผล สามารถ สรุปออกมาได้ 4 ประการดังนี้คือ การรวบรวมข้อมูล การลดจำนวนข้อมูล การชำระข้อมูล และ การแปลงและปรับบรรทัดฐานข้อมูล

2.5.1 การรวมข้อมูล

การรวมข้อมูลที่เก็บไว้ในหลายฐานข้อมูลเข้าด้วยกัน ถือเป็นงานที่มีความจำเป็นงานหนึ่งในการทำเหมืองข้อมูล เนื่องจากข้อมูลที่จะนำมาวิเคราะห์อาจจะมีข้อมูลบางส่วนที่มีเนื้อหาเหมือนกัน การรวมข้อมูลมีจุดประสงค์เพื่อลดความซ้ำซ้อนของข้อมูล ยกตัวอย่างเช่น ฐานข้อมูล A และฐานข้อมูล B เก็บข้อมูลของกลุ่มลูกค้ากลุ่มเดียวกัน แต่ฐานข้อมูล A ใช้แอทริบิวต์ ชื่อ ‘cust.id’ แทนหมายเลขลูกค้า ส่วน ฐานข้อมูล B ใช้แอทริบิวต์ ‘customer.id’ แต่เนื่องจากข้อมูลที่ถูกเก็บในฐานข้อมูลทั้งสองเป็นข้อมูลเดียวกัน การผนวกข้อมูลจากฐานข้อมูล A เข้ากับ B โดยตรง อาจทำให้เกิดฐานข้อมูลใหม่ที่มีแอทริบิวต์ซ้ำกันสองตัว

ในกรณีคล้ายๆกันกับการรวมข้อมูลจากฐานข้อมูลมากกว่าหนึ่งฐานเข้าด้วยกัน การรวมชุดข้อมูลสองชุดเข้าด้วยกันอาจส่งผลให้ชุดข้อมูลหลังการรวมมีคุณลักษณะที่ซ้ำกันอยู่ นั่นหมายถึงมิติของข้อมูลสูงมากกว่าความจำเป็น และนี่อาจจะส่งผลต่อการประมวลผลในขั้นต่อไป ดังที่เราจะได้ศึกษาในบทที่ 3 ว่าด้วยผลกระทบของมิติของข้อมูลต่อประสิทธิภาพของขั้นตอนวิธีในการทำเหมืองข้อมูล ดังนั้นด้วยเหตุผลด้านประสิทธิภาพ

เราจึงควรจะมีวิธีตรวจสอบว่าแอทริบิวต์ หรือคุณลักษณะจากแหล่งข้อมูลต่างกันสื่อถึงสิ่งเดียวกันหรือไม่ เพื่อพยายามที่จะลดความซ้ำซ้อนหลังจากการรวมข้อมูล

การลดความซ้ำซ้อนวิธีหนึ่งที่สามารถทำได้โดยง่าย คือการวิเคราะห์ความสัมพันธ์ระหว่างคุณลักษณะสองตัว หากค่าความสัมพันธ์มีค่าสูง ก็แสดงว่าคุณลักษณะสองตัวนั้นอาจสื่อถึงสารสนเทศเดียวกัน เมื่อทราบเช่นนั้นแล้ว เราก็สามารถที่จะเลือกเก็บคุณลักษณะตัวใดตัวหนึ่งไว้ก็พอ การวิเคราะห์ความสัมพันธ์ที่จะยกตัวอย่างในที่นี้ มีสามแบบคือ การใช้การทดสอบ χ^2 การคำนวณค่าความแปรปรวนร่วม และการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

การทดสอบ χ^2

การทดสอบทางสถิติสามารถนำมาประยุกต์ใช้ในการทดสอบสมมุติฐานว่า ตัวแปรสุ่ม (Random Variable) สองตัวเป็นอิสระต่อกันหรือไม่ ในทางสถิติหากตัวแปรสุ่มสองตัวเป็นอิสระต่อกัน จะพบว่าความน่าจะเป็นที่ตัวแปรสุ่มทั้งคู่จะเกิดขึ้นพร้อมกัน $p(X, Y)$ จะเท่ากับผลคูณของความน่าจะเป็นที่ตัวแปรสุ่มแรก $p(X)$ และความน่าจะเป็นที่ตัวแปรสุ่มที่สอง $p(Y)$ นั่นคือ $p(X, Y) = p(X) \times p(Y)$ หากเรากำหนดให้สมมุติฐานว่าง (Null Hypothesis) แสดงสมมุติฐานที่ตัวแปรสุ่มสองตัวเป็นอิสระต่อกัน เราจะทำการทดสอบทางสถิติว่าเราจะปฏิเสธสมมุติฐานว่างได้หรือไม่ หากเราไม่สามารถปฏิเสธสมมุติฐานว่างได้ ก็แสดงว่าตัวแปรสุ่มสองตัวเป็นอิสระต่อกันจริงๆ การทดสอบ χ^2 นิยมใช้ในการทดสอบตัวแปรที่มีลักษณะเชิงนามซึ่งรวมถึงตัวแปรที่เป็นแบบทวิภาคด้วย เราอาจเข้าใจการทดสอบทางสถิตินี้ได้ง่ายขึ้นหากเริ่มจากตัวอย่าง

สมมุติว่ามีการสุ่มสำรวจความชอบในการดื่มกาแฟในตอนเช้า ของกลุ่มตัวอย่าง ซึ่งเป็นนักศึกษาจำนวน 1000 คน กำหนดให้ตัวแปรนี้แทนด้วย $X \in \{\text{ดื่มกาแฟ,ไม่ดื่มกาแฟ}\}$ และสำรวจว่าชอบกินขนมปังในตอนเช้าหรือไม่ โดยให้ตัวแปรนี้แทนด้วย $Y \in \{\text{กินขนมปัง,ไม่กินขนมปัง}\}$ หากนำข้อมูลดังกล่าวมาใส่ลงในตารางการจรพบว่าจะได้ตารางดังต่อไปนี้

	ดื่มกาแฟ	ไม่ดื่มกาแฟ	รวม
กินขนมปัง	250	200	450
ไม่กินขนมปัง	50	1000	1050
รวม	300	1200	1500

จากตารางการจร เราสามารถคำนวณค่า χ^2 ได้โดยใช้สมการ

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(o_{ij} - e_{ij})^2}{o_{ij}} \tag{2.18}$$

โดยที่ o_{ij} แสดงค่าความถี่ที่พบเห็น (Observed Frequency) ซึ่งเป็นค่าเดียวกันกับค่าที่อยู่ในตาราง ส่วน e_{ij} คือค่าความถี่คาดหวัง (Expected Frequency) ที่จะต้องคำนวณขึ้นมาใหม่โดยยึดตามสมมุติฐานว่าง ที่ว่าตัวแปรสุ่มสองตัวเป็นอิสระต่อกัน

$$e_{ij} = \frac{\sum_{k=1}^{C} o_{ik} \sum_{l=1}^{R} o_{rj}}{N} \tag{2.19}$$

ส่วน R และ C แสดงจำนวนแถวและหลักของตารางการจร ตามลำดับ

หากลองจินตนาการตามสมการ (2.19) จะพบว่า การหาค่าความถี่คาดหวัง ณ ตำแหน่ง e_{ij} ก็คือการหาความน่าจะเป็นที่จะพบเหตุการณ์ที่ $X = i$ และ $Y = j$ จากเหตุการณ์ทั้งหมด หากเชื่อมโยงเข้ากับตารางการจร เราจะพบว่าค่า e_{ij} คือการนำผลคูณของผลรวมของหลักที่ i คูณกับ ผลรวมของแถวที่ $Y = j$ แล้วหารด้วยจำนวนของเหตุการณ์ทั้งหมด

ตัวอย่างเช่น ค่าความถี่คาดหวังของคนที่ดื่มกาแฟ *และ* กินขนมปังในตอนเช้าเป็น $e_{11} = 300 \times 450/1500 = 90$ หากเราหาค่าความถี่คาดหวังให้ครบทุกๆเหตุการณ์ เราจะได้ผลลัพธ์ดังค่าในวงเล็บในตารางข้างล่างนี้

	ดื่มกาแฟ	ไม่ดื่มกาแฟ	รวม
กินขนมปัง	250 (90)	200 (360)	450
ไม่กินขนมปัง	50 (210)	1000 (840)	1050
รวม	300	1200	1500

เมื่อเราทราบค่าของ o_{ij} และ e_{ij} ครบแล้ว เราก็สามารถดำเนินการต่อเพื่อคำนวณค่า χ^2 ได้

$$\begin{aligned}\chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{1000} \\ &= 507.93\end{aligned} \tag{2.20}$$

สังเกตว่า ค่าของ χ^2 แสดงความผิดพลาดยกกำลังสอง ระหว่างความถี่ที่พบเห็น กับความถี่คาดหวังตามสมมุติฐานว่าง และปรับบรรทัดฐานด้วยค่าที่ได้จากการสุ่มสำรวจ ดังนั้นหากผลรวมของแต่ละพจน์ดังกล่าวมีค่ามาก เราสามารถสรุปได้ว่า ค่าความถี่พบเห็นต่างจากค่าความถี่คาดหวัง นั่นคือ ยิ่งความแตกต่างมีค่าสูง เท่าไหร่ เราก็ยิ่งมีความมั่นใจมากขึ้นในการที่จะปฏิเสธสมมุติฐานว่าง การปฏิเสธสมมุติฐานว่าง ก็คือการยอมรับว่าตัวแปรสุ่มทั้งสองตัวมีความสัมพันธ์กันนั่นเอง

ในทางสถิติแล้ว ระดับความมั่นใจที่จะปฏิเสธสมมุติฐานว่าง วัดได้โดยการ คำนวณความน่าจะเป็นที่จะมีค่า χ^{*2} ตัวอื่นที่มีค่ามากกว่า χ^2 ที่คำนวณได้ในข้างต้น หรือในเชิงสัญลักษณ์เราต้องการคำนวณ

$$P(\chi^{*2} > \chi^2)$$

ซึ่งความน่าจะเป็นข้างต้น สามารถหาได้จากฟังก์ชันความหนาแน่นสะสม (Cumulative Distribution Function (CDF)) ของการแจกแจงแบบ χ^2 ซึ่งฟังก์ชันดังกล่าวต้องการพารามิเตอร์สองตัว คือค่า χ^2 ที่คำนวณ ได้ และองศาเสรี (Degree of Freedom) ซึ่งมีค่าเท่ากับ df $= (R-1)(C-1)$ ปัจจุบันมีโปรแกรมบนอินเทอร์เน็ต ที่สามารถคำนวณค่าของ CDF ให้โดยอัตโนมัติ หากค่าความน่าจะเป็นที่ได้จาก CDF มากเท่าไหร่เราก็ยิ่งมั่นใจได้ว่าเราสามารถปฏิเสธสมมุติฐานว่าง ได้มากขึ้นเท่านั้น โดยปกติแล้วหากค่าความน่าจะเป็นจาก CDF มากกว่า 0.95 (เทียบเท่ากับ Significant Value 0.05) เราก็สามารถปฏิเสธสมมุติฐานว่างได้ นั่นคือตัวแปรกสุ่มสองตัวมีความสัมพันธ์กัน

ความแปรปรวนร่วม

อีกวิธีสำหรับการวัดความสัมพันธ์ระหว่างตัวแปรสุ่ม คือการหาค่าความแปรปรวนร่วมระหว่างตัวแปรสุ่มสองตัวใดๆ ซึ่งเราการคำนวณสามารถทำได้โดยอาศัยสมการ 2.6 ซึ่งเราเคยศึกษามาแล้ว ค่าความแปรปรวนร่วมจะเป็นตัวบอกว่า ตัวแปรสุ่มสองตัวเปลี่ยนค่าไปพร้อมกันมากน้อยแค่ไหน สำหรับตัวแปรสุ่ม x^i, x^j สองตัวใดๆที่เป็นอิสระต่อกัน เราจะพบว่าค่า $Cov(x^i, x^j) = 0$ เสมอ แต่ว่าบทกลับไม่เป็นจริงเสมอไป นั่นคือหากเราบังเอิญคำนวณค่า $Cov(x^i, x^j) = 0$ ในกรณีนี้ เราไม่สามารถสรุปได้ว่าตัวแปรสุ่มสองตัวนั้น จะต้องเป็นอิสระต่อกัน บทกลับดังกล่าวจะเป็นจริงก็ต่อเมื่อตัวแปรสุ่มทั้งสองตัวมีการแจกแจงแบบปกติเท่านั้น

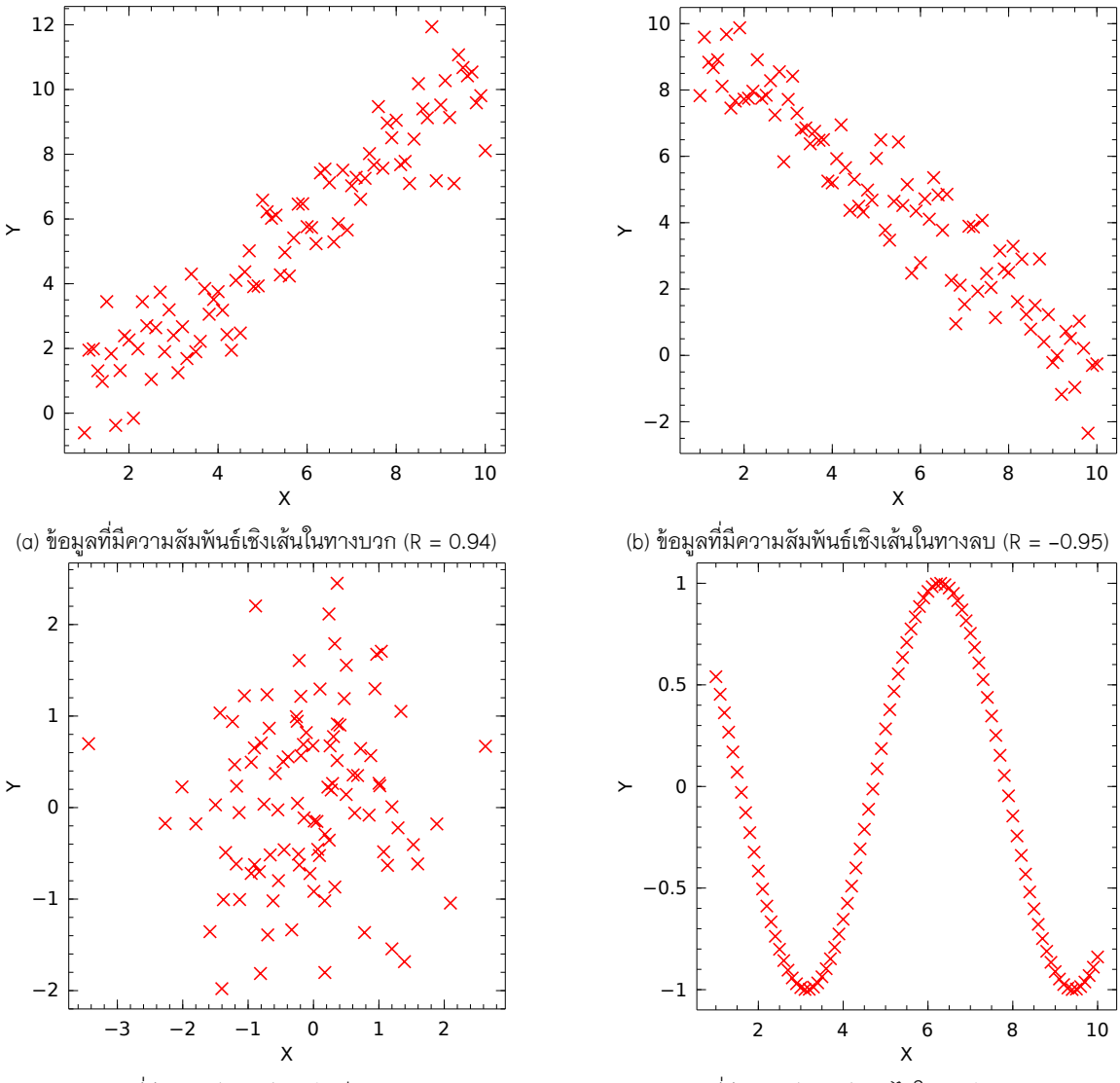
สัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน

การวัดความสัมพันธ์ของตัวแปรสุ่ม สามารถวัดได้โดยการหาสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน วิธีการดังกล่าว ใช้ในการหาความสัมพันธ์เชิงเส้น ของตัวแปรสุ่มสองตัว ซึ่งสามารถคำนวณได้โดย การหารค่าความแปรปรวนร่วมของตัวแปรทั้งสอง ด้วย ส่วนเบี่ยงเบน มาตรฐานของสองตัวแปรดังกล่าว

$$R(x^i, x^j) = \frac{Cov(x^i, x^j)}{\sigma_{x^i}\sigma_{xj}} \tag{2.21}$$

จากสมการข้างต้น เราพบว่า หาก $R(X,Y) > 0$ ตัวแปรทั้งสองตัวจะมีความสัมพันธ์เชิงเส้นในทางบวก หาก $R(X,Y) < 0$ ตัวแปรสองตัวจะมีความสัมพันธ์เชิงเส้นในทางลบ (ตัวแปรหนึ่งเพิ่มค่า อีกตัวแปรหนึ่งจะลดค่า) หาก $R(X,Y) = 0$ จะแสดงว่าตัวแปรทั้งสองเป็นอิสระต่อกัน

แผนภาพการกระจายที่ในรูปที่ 2.4 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลที่มีความสัมพันธ์ในแบบต่างๆ



(a) ข้อมูลที่มีความสัมพันธ์เชิงเส้นในทางบวก (R = 0.94)

(b) ข้อมูลที่มีความสัมพันธ์เชิงเส้นในทางลบ (R = −0.95)

(c) ข้อมูลที่มีความสัมพันธ์เชิงเส้นต่ำ (R = 0.05)

(d) ข้อมูลที่มีความสัมพันธ์แบบไม่ใช่เชิงเส้น (R = −0.02)

รูปภาพ 2.4: ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรในแบบต่างๆ

ข้อควรระวังคือ ตัวแปรที่มีความสัมพันธ์สูง ไม่ได้แปลว่าตัวแปรทั้งสองตัวจะมีผลต่อกันเสมอไป เช่น

มีข้อมูลที่แสดงว่า จำนวนการปล่อยดาวเทียมขึ้นสู่อวกาศ มีความสัมพันธ์อย่างสูงกับจำนวนนักศึกษาที่จบปริญญาเอกสาขาสังคมศาสตร์ แต่เมื่อพิจารณาจากความเป็นจริงแล้ว ตัวแปรสองตัวนี้ยากที่จะส่งผลต่อกันจริงๆ ถึงแม้ว่าค่าความสัมพันธ์จะออกมาสูง หรือที่ในภาษาอังกฤษมักจะกล่าวว่า “correlation doesn’t imply causation”

2.5.2 การลดข้อมูล

การลดข้อมูล (Data reduction) เป็นขั้นตอนหนึ่งที่จำเป็นมากขึ้นเรื่อยๆสำหรับการทำเหมือนข้อมูลยุคที่ชุดข้อมูลมักมีขนาดใหญ่ โดยทั่วไปแล้วการลดข้อมูลอาจสื่อได้สองความหมาย คือการลดจำนวนข้อมูล และการลดมิติของข้อมูลให้น้อยลง เป้าหมายของการลดข้อมูลคือการประมวลผลที่เร็วขึ้น โดยที่คุณภาพของผลลัพธ์ที่ได้ไม่ลดลงไปตามขนาดของข้อมูลที่ลดลง ถึงแม้ว่าเราจะมีคอมพิวเตอร์ที่ทำงานเร็วขึ้นมากและมีหน่วยความจำราคาถูกจำนวนมหาศาล เราก็ยังคงต้องพิจารณาและศึกษาวิธีการลดขนาดข้อมูล เนื่องจากอัตราการเพิ่มของข้อมูลในปัจจุบันนั้น เร็วกว่าความสามารถในการพัฒนาประสิทธิภาพของฮาร์ดแวร์มาก ดังนั้นจึงเป็นที่น่าสนใจไม่น้อย หากเราจะสามารถได้ผลลัพธ์จากการประมวลผลเท่าเดิมโดยใช้เวลาและหน่วยความจำที่น้อยลง ในบทนี้เราจะศึกษาการลดจำนวนข้อมูลในชุดข้อมูลก่อน ส่วนแนวทางการลดมิติของข้อมูลเราจะศึกษาในบทถัดไป

โดยปกติการลดขนาดของชุดข้อมูลแบ่งออกเป็นสองแนวทาง คือการลดข้อมูลเชิงพาราเมทริค (Parametric) และการลดข้อมูลเชิงไม่ใช่พาราเมทริค (Non-parametric)

การลดข้อมูลเชิงพาราเมทริค

การลดข้อมูลเชิงพาราเมทริคตั้งอยู่บนสมมุติฐานที่ว่า ข้อมูลกระจายตัวตามการแจกแจงทางสถิติแบบใดแบบหนึ่ง ดังนั้นเป้าหมายของวิธีการนี้ก็คือการมุ่งหารูปแบบของการแจกแจงทางสถิตินั้น โดยอาศัยการสร้างตัวต้นแบบ (Model) ของการแจกแจงทางสถิติขึ้นเพื่ออธิบายการเกิดขึ้นของข้อมูลให้ใกล้เคียงที่สุด โดยปกติแล้วตัวต้นแบบเหล่านี้จะถูกอธิบายด้วยเซตของพารามิเตอร์ของตัวต้นแบบ (Model’s Parameters) ฉะนั้นการหาตัวต้นแบบในที่นี้ ก็หมายรวมไปถึงการประมาณค่าพารามิเตอร์ของตัวต้นแบบนั้นๆด้วย เมื่อสามารถระบุตัวต้นแบบได้ชัดเจนแล้ว เราก็สามารถเก็บข้อมูลไว้ในรูปของเซตของพารามิเตอร์ของตัวต้นแบบได้ นอกจากนั้นเรายังสามารถลดจำนวนข้อมูลบางส่วนที่ไม่จำเป็นออกไปได้บ้าง เพราะหากจำเป็นที่จะต้องใช้ข้อมูลเพิ่ม เราสามารถสร้างข้อมูลขึ้นมาได้เสมอโดยอาศัยตัวต้นแบบที่หาไว้ได้

ตัวต้นแบบที่นิยมนำมาใช้เพื่อประมาณรูปแบบการกระจายตัวของข้อมูล ชนิดหนึ่งก็คือ Gaussian Mixture Model (GMM) ตัวต้นแบบนี้ตั้งอยู่บนสมมุติฐานที่ว่า ชุดข้อมูลประกอบขึ้นจากกลุ่มข้อมูลย่อยๆซึ่งแต่ละกลุ่มมีการแจกแจงแบบปกติ โดยทั่วไป GMM ที่ประกอบด้วยการแจกแจงแบบปกติ จำนวน k กลุ่ม จะแสดง

ด้วยสมการ

$$GMM(x) = \sum_{i=1}^{k} \lambda_i N(x; \mu_i, \Sigma_i) \tag{2.22}$$

โดยที่ $\lambda_I \geq 0$ และ $\sum_{i=1}^{k} \lambda_i = 1$ เป้าหมายของ GMM คือ การหาค่าพารามิเตอร์ของการแจกแจงแบบปกติ ซึ่งก็คือค่ากลาง μ และค่าความแปรปรวนร่วม Σ พร้อมกับสิ่งที่เรียกว่า ค่าน้ำหนัก (Mixing Weight) λ_i ที่สามารถอธิบายชุดข้อมูลที่มีอยู่ในมือได้ดีที่สุด การประมาณค่าพารามิเตอร์ทั้งหมดนั้น นิยมใช้วิธีที่เรียกว่า ขั้นตอนวิธีแบบ Expectation–Maximisation (EM) ในการประมาณ แต่เนื้อหาเกี่ยวกับขั้นตอนวิธีดังกล่าว อยู่นอกเหนือขอบเขตของเอกสารการสอนนี้ ผู้สนใจสามารถอ่านเพิ่มเติมได้จาก [Dempster et al., 1977] เมื่อได้ค่าพารามิเตอร์ทั้งหมดแล้ว เราสามารถทิ้งข้อมูลส่วนหนึ่งไปได้ เนื่องจากเราสามารถสุ่มสร้างข้อมูลใหม่ ที่มีที่มาเหมือนกับชุดข้อมูลเดิมได้โดยใช้ GMM ที่หาได้ ดังที่เคยกล่าวไว้ตอนต้น

แนวทางพาราเมทริคอาจมีข้อจำกัดตรงที่ หากสมมุติฐานเกี่ยวกับการแจกแจงทางสถิติของข้อมูลที่เราตั้งขึ้นไม่ถูกต้อง จะทำให้เราได้ตัวต้นแบบที่ไร้ประโยชน์ ยกตัวอย่างเช่น เราเชื่อว่าข้อมูลมีการแจกแจงแบบปกติ แต่แท้จริงแล้ว ข้อมูลมีการแจกแจงแบบปัวซง (Poisson Distribution) ตัวต้นแบบที่สร้างขึ้นมาก็จะเป็นตัวต้นแบบที่ผิด และไม่สามารถนำมาใช้งานได้อย่างมีประสิทธิภาพ ข้อจำกัดอีกประการหนึ่งก็คือ สัญญาณรบกวนที่ปะปนในชุดข้อมูล อาจทำให้การประมาณค่าพารามิเตอร์ ของตัวต้นแบบผิดเพี้ยนไป จึงจำเป็นต้องมีแนวทางในการจัดการกับสัญญาณรบกวนที่อาจจะมีผลต่อตัวต้นแบบด้วย

การลดข้อมูลเชิงไม่ใช่พาราเมทริค

การลดขนาดของข้อมูลอีกประเภทหนึ่งคือแบบที่ไม่ตั้งสมมุติฐานเกี่ยวกับการแจกแจงของข้อมูล วิธีนี้อาศัยการรวมข้อมูลที่ใกล้เคียงกันเข้าไว้ในกลุ่มเดียวกัน แล้วใช้ตัวแทนของกลุ่มเพียงตัวเดียว เพื่อนำไปใช้ในการคำนวณ หรือที่เรียกกันโดยทั่วไปว่าการแบ่งนับ วิธีการแบ่งนับวิธีหนึ่งที่นิยมใช้ และนำไปปรับใช้ได้กับข้อมูลหลายประเภทได้แก่ การสร้างฮิสโทแกรมของข้อมูล ซึ่งทำได้สองแบบคือ ฮิสโทแกรมที่มีช่วงเท่ากัน (Equal–width Histogram) และฮิสโทแกรมที่มีความถี่เท่ากัน (Equal–frequency Histogram)

การสร้างฮิสโทแกรมที่มีช่วงเท่ากัน ทำได้โดยแบ่งช่วงข้อมูลออกเป็นช่วงเท่าๆกัน โดยปกติจะเรียกช่วงดังกล่าวว่า ‘ถัง’ แล้วใช้ค่ากลางของข้อมูลในแต่ละถังในการประมวลผลต่อไป ส่วนฮิสโทแกรมที่มีความถี่เท่ากันนั้น จะไม่จำกัดช่วงของถังแต่จะใช้การนับจำนวนข้อมูลที่ตกในถังนั้นให้ครบจำนวน หากเกินจะสร้างถังใหม่ถัดไป วิธีการนี้จะได้จำนวนข้อมูลในแต่ละถังมีค่าเท่ากันหมด

นอกจากวิธีวิเคราะห์โดยใช้ฮิสโทแกรมแล้ว เรายังสามารถใช้การจัดกลุ่มข้อมูลมาช่วยในการลดขนาดของชุดข้อมูล วิธีการ GMM ที่เราได้ทำความรู้จักไปแล้วข้างต้นถือเป็น ขั้นตอนวิธีที่ใช้จัดกลุ่มข้อมูลวิธีหนึ่ง

นอกจากนี้ยังมีขั้นตอนวิธีที่ใช้ในการจัดกลุ่มแบบไม่ใช่พาราเมทริคอย่าง วิธีเคมีนส์ (k-means) หรือการจัดกลุ่มแบบลำดับขั้น ซึ่งเราจะได้ศึกษากันโดยละเอียดในบทที่ 6 ซึ่งวิธีการเหล่านั้นก็สามารถนำมาใช้ในการจัดกลุ่มเพื่อลดข้อมูลได้

โดยใจความสำคัญแล้ว การใช้การจัดกลุ่มข้อมูลมาลดขนาดชุดข้อมูล คือการหากลุ่มก้อนของข้อมูลให้พบ โดยเมื่อพบกลุ่มก้อนแล้ว เราสามารถโยนข้อมูลส่วนใหญ่ทิ้งไป แล้วเก็บไว้แต่ข้อมูลสำคัญที่เพียงพอต่อการ อธิบายลักษณะของกลุ่มก้อนนั้น อาจจะเป็นค่าเฉลี่ยของข้อมูลในกลุ่มก้อนเดียวกัน หรือเลือกข้อมูลบางตัวมาจากกลุ่มก้อนนั้น เพื่อใช้ในการประมวลผลต่อไป

วิธีอีกประเภทหนึ่งในแนวทางแบบไม่ใช่พาราเมทริคที่จะกล่าวถึงก็คือการซักตัวอย่าง (Sampling) ข้อมูลส่วนหนึ่งมาจากชุดข้อมูลที่อาจมีขนาดใหญ่เกินไป การซักตัวอย่างที่ง่ายที่สุดก็คือ การซักตัวอย่างแบบสุ่ม (Random Sampling) โดยการซักตัวอย่างแบบนี้สามารถแบ่ง ได้เป็นอีกสองแบบย่อย คือแบบเลือกซ้ำและแบบไม่เลือกซ้ำ ข้อเสียของการซักตัวอย่างแบบสุ่มคือ วิธีนี้อาจจะไม่เหมาะสมกับข้อมูลที่การแจกแจงตัวแบบบิดเบี้ยว (Skewed Distribution) เพราะหลังจากการซักตัวอย่างแล้ว ชุดข้อมูลที่ได้ อาจมีความบิดเบี้ยวไม่เหมือนเดิม ในกรณีที่การแจกแจงของข้อมูลบิดเบี้ยว สามารถใช้เทคนิคการจัดกลุ่มข้อมูล เข้ามาช่วยในการหากลุ่มของข้อมูลก่อน จากนั้นจึงทำการซักตัวอย่างแบบสุ่มจากแต่ละกลุ่มอีกทีหนึ่ง ซึ่งจะทำให้การกระจายตัวของข้อมูลใกล้เคียงกับการแจกแจงแบบเดิมมากกว่า วิธีที่อาศัยข้อมูลของกลุ่มข้อมูล มาเป็นตัวช่วยในการซักตัวอย่างนี้ มีชื่อเรียกว่า การซักตัวอย่างแบบแบ่งชั้น (Stratified Sampling) โดยสรุปแล้ว การซักตัวอย่างข้อมูลเพื่อลดขนาดข้อมูลมี 4 ประเภทหลักๆดังนี้

- การซักตัวอย่างแบบสุ่ม (Simple Random Sampling)
- การซักตัวอย่างแบบสุ่มแบบเลือกซ้ำ (Random Sampling with Replacement)
- การซักตัวอย่างแบบสุ่มแบบไม่เลือกซ้ำ (Random Sampling without Replacement)
- การซักตัวอย่างแบบแบ่งชั้น (Stratified Sampling)

2.5.3 การชำระข้อมูล

ข้อมูลที่เก็บได้จากธรรมชาติอาจเป็นข้อมูลที่ไม่สมบูรณ์ ซึ่งอาจเกิดจากเหตุปัจจัยหลายประการ เช่น เกิดจากความผิดพลาดจากเครื่องมือตรวจวัดระหว่างเก็บข้อมูล หรือเป็นข้อจำกัดของเครื่องมือตรวจวัดตั้งแต่แรก อาจเกิดจากความผิดพลาดของคนเก็บข้อมูล หรือกระทั่งความผิดพลาดในการส่งข้อมูลทางสายส่งสัญญาณจากที่เก็บข้อมูลไปยังฐานข้อมูล เช่นการเก็บข้อมูลจากดวงจันทร์ส่งผ่านมายังโลก ซึ่งเหตุปัจจัยเหล่านี้ ส่งผลให้ข้อมูลที่เก็บมาได้ มีความผิดพลาดปลอมปนเข้ามา หรือที่มักเรียกกันว่าข้อมูลไม่สะอาด

ความผิดพลาดของข้อมูลสามารถแบ่งออกเป็นสองประเภทใหญ่ๆ ได้แก่

- ความไม่สมบูรณ์ของข้อมูล (Incomplete) มีได้หลายกรณีย่อย เช่น การลืมใส่คำถามที่ต้องการลงไปในแบบฟอร์มสอบถาม ทำให้คุณลักษณะดังกล่าวขาดหายไปสำหรับข้อมูลทุกตัว (Missing Feature) หรือมีคำถามดังกล่าวในแบบฟอร์ม แต่ผู้ตอบแบบสอบถามกรอกข้อมูลไม่ครบ (Missing Value) หรือมีแต่ค่าสถิติแต่ไม่มีข้อมูลดิบ

- ความไม่ถูกต้องของข้อมูล (Incorrect) เช่น ข้อมูลเต็มไปด้วยสัญญาณรบกวน (Noise), มีค่าผิดพลาด (Outlier) มีค่าสุดโต่ง (Extreme Value)

การจัดการกับข้อมูลที่ไม่สมบูรณ์นั้น สามารถทำได้หลายวิธี วิธีที่ง่ายที่สุดอาจทำโดยการตัดข้อมูลตัวนั้น ออกไปจากชุดข้อมูลเลย ข้อเสียของวิธีนี้คือ เราจะอาจสูญเสียข้อมูลที่สำคัญไปได้ อีกทั้งวิธีการนี้อาจไม่เหมาะกับชุดข้อมูลที่เก็บได้ยากและมีราคาแพง เช่น ข้อมูลของหินจากดาวอังคาร ข้อมูลผู้ป่วยโรคร้ายแรงหายาก เป็นต้น หากจะไม่ทำการลบทิ้ง เราสามารถเติมค่าที่คิดว่าเหมาะสม ลงไปในตำแหน่งของคุณลักษณะนั้น ซึ่งวิธีที่นิยม คือ การใช้ค่าบางอย่าง เพื่อแสดงว่าคุณลักษณะในตำแหน่งนั้นบกพร่องไป เช่น การใช้เลข –1 แทนค่าที่หายไป ข้อเสียของวิธีนี้คือ ค่าที่ใส่ลงไปอาจไปรบกวนการประมวลผลในขั้นต่อไป ในแง่มุมที่เราคาดไม่ถึง วิธีที่ดีกว่าอาจเป็น การใช้ค่าเฉลี่ยของข้อมูลทั้งหมดในตำแหน่งของคุณลักษณะนั้น เติมลงไปในช่องที่หายไป อย่างไรก็ตาม วิธีดังกล่าวก็ยังมีข้อจำกัดตรงที่ หากข้อมูลที่เก็บมามีความเป็นกลุ่มก้อนอยู่แล้วเช่น กรณีที่เราเก็บข้อมูลของช้างกับม้าเพื่อทำการศึกษา ต่อมาพบว่าน้ำหนักของช้างตัวหนึ่งขาดไป หากเราทำการเติมข้อมูลที่ขาดหายไปด้วยค่าเฉลี่ยน้ำหนักของชุดข้อมูลดังกล่าว เราจะพบว่า ค่าเฉลี่ยดังกล่าว อาจน้อยกว่าน้ำหนักเฉลี่ยของช้าง เพราะมีการนำน้ำหนักของม้ามาเฉลี่ยด้วย วิธีปรับปรุงแก้ไข คือเราควรเติมค่าที่ขาดหายไปด้วยค่าเฉลี่ยของกลุ่มข้อมูลของช้างเท่านั้น

ข้อมูลที่ไม่สมบูรณ์อีกประเภทหนึ่งก็คือ ข้อมูลที่มีสัญญาณรบกวน สัญญาณรบกวนในที่นี้ จะหมายรวมถึงความผิดพลาดที่เกิดขึ้นแบบสุ่ม (Random Error) หรือความเบี่ยงเบนแบบสุ่ม (Random Deviation) ในคุณลักษณะที่ต้องการวัด สัญญาณรบกวน อาจเกิดได้จากหลายสาเหตุ เช่น เกิดจากความผิดพลาดของเครื่องมือวัด การเห็นไม่ตรงกันของผู้เชี่ยวชาญ ปัญหาในการส่งข้อมูล หรือแม้แต่ข้อจำกัดทางเทคโนโลยี

วิธีที่นิยมใช้ในการลดผลกระทบของสัญญาณรบกวน ในข้อมูล สามารถสรุปออกเป็นวิธีหลัก 3 วิธีดังนี้

การแบ่งนับ

วิธีการแบ่งนับ (Quantisation) เป็นการลดผลกระทบของสัญญาณรบกวนโดยอาศัยข้อมูลจากเพื่อนบ้าน [1] หลายๆตัว มีขั้นตอนคือ รวมข้อมูลที่อยู่ใกล้เคียงกันเข้ามาอยู่ในถัง (Bin) เดียวกัน จากนั้นจะทำการหาค่า

[1] ข้อมูลที่อยู่ใกล้กัน

กลาง (ซึ่งอาจใช้ค่าเฉลี่ยเลขคณิต หรือค่ามัธยฐานก็ได้) ของข้อมูลทั้งถัง แล้วใช้ค่ากลางของถังเป็นตัวแทนของข้อมูลในถังนั้น วิธีการนี้จะได้ผลมากหากสัญญาณรบกวนดังกล่าว มีการกระจายตัว (แบบใดก็ได้) ที่มีค่ากลางเท่ากับ 0 (Zero-mean Noise) เพื่อให้เห็นภาพชัดเจนยิ่งขึ้น สมมุติว่าข้อมูลที่เรามองเห็นหรือที่เราเก็บได้ให้สัญลักษณ์เป็น $\tilde{X}$ คือผลบวกของ ข้อมูลจริงที่ปราศจากสัญญาณรบกวน X กับสัญญาณรบกวน ϵ

$$\tilde{X} = X + \epsilon \tag{2.23}$$

ในที่นี้ หากเราพบว่าการแจกแจงของสัญญาณรบกวนเป็นแบบปกติ ที่มีค่ากลางเท่ากับ 0 ซึ่งในเชิงสัญลักษณ์สามารถเขียนแทนด้วย $\epsilon \sim N(0, \sigma)$. เราจะพบว่า หากเราทำการแบ่งนับ $\tilde{X}$ หลายๆตัวที่มีค่าใกล้เคียงกันเข้าด้วยกัน เราจะได้ว่า

$$\begin{aligned} \frac{\sum_{n=1}^{N} \tilde{X}_N}{N} &= \frac{\sum_{n=1}^{N} X}{N} + \frac{\sum_{n=1}^{N} \epsilon}{N} \\ \frac{\sum_{n=1}^{N} \tilde{X}_N}{N} &= \frac{\sum_{n=1}^{N} X}{N} + 0 \end{aligned} \tag{2.24}$$

เราจะได้ว่า พจน์สุดท้ายซึ่งเป็นค่าเฉลี่ยของ ϵ ที่เรากำหนดไว้แต่ต้นว่ามีค่าเท่ากับ 0 จะหายไป และสิ่งที่ได้ก็คือ ข้อมูลที่เราเก็บมา มีค่าตรงกับข้อมูลจริงที่ปราศจากสัญญาณรบกวน หรือแปลได้ว่าเราสามารถลบสัญญาณรบกวนออกจากข้อมูลได้นั่นเอง

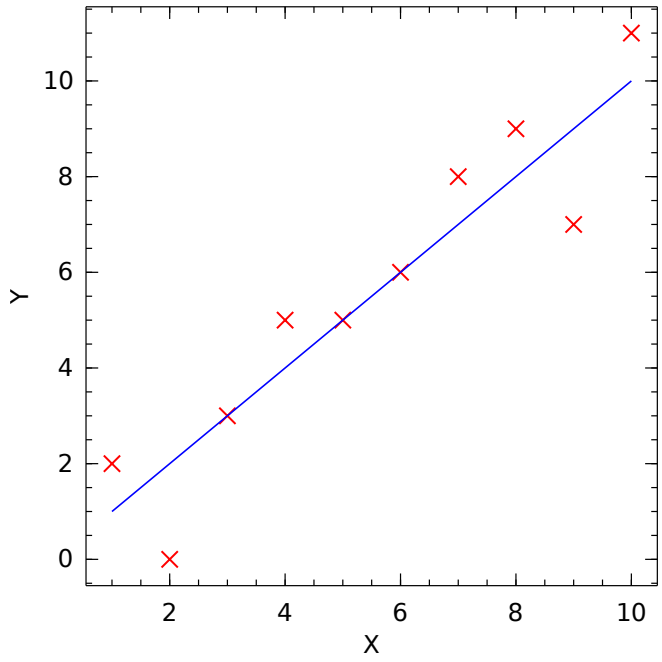
การจัดกลุ่มข้อมูล

วิธีการการจัดกลุ่มข้อมูลสามารถนำมาใช้ในการลดข้อมูลสุดโต่งได้ ปกติแล้วขั้นตอนวิธีสำหรับการจัดกลุ่มมีจุดประสงค์เพื่อรวมข้อมูลที่มีความคล้ายกันเข้าไว้ด้วยกันเป็นกลุ่มก้อน ดังนั้นหากข้อมูลใดไม่สามารถจะจัดเข้าอยู่ในกลุ่มใดได้ ข้อมูลนั้นมีโอกาสที่จะเป็นข้อมูลสุดโต่งที่อาจเกิดจากสัญญาณรบกวน เราจะกลับมาศึกษาการจัดกลุ่มข้อมูล โดยละเอียดในบทที่ 6

วิธีการถดถอย

อีกแนวทางหนึ่งในการลดสัญญาณรบกวน คือการนำการวิเคราะห์การถดถอยมาใช้งาน การวิเคราะห์แบบนี้จะมุ่งหาฟังก์ชัน $f(\cdot)$ ที่แสดงความสัมพันธ์ของตัวแปรต้น a และตัวแปรตาม b ดังสมการ $b = f(a)$ หลักการก็คือ การใช้ชุดข้อมูลขาเข้าเพื่อตามหาฟังก์ชัน $f(\cdot)$ จากนั้น เราจะนำฟังก์ชันที่หาได้ ย้อนกลับมาลองอธิบาย b ในแต่ละตำแหน่งของ a เพื่อตรวจดูว่าค่า b ในตำแหน่งนั้น คลาดเคลื่อนไปจาก สิ่งที่ฟังก์ชันอธิบายหรือทำนายหรือไม่ หากพบว่ามีความคลาดเคลื่อนสูงผิดปกติ นั่นหมายความว่า b ณ ตำแหน่งดังกล่าว

ไม่เป็นไปตามแนวโน้มของข้อมูลส่วนใหญ่ ทำให้สงสัยได้ว่าค่า b ณ ตำแหน่งนั้นอาจจะผิดพลาดได้ เราอาจดำเนินการต่อโดยกำจัด b ออกจากชุดข้อมูล หรือส่งให้ผู้เชี่ยวชาญตรวจวิเคราะห์ต่อไป ตัวอย่างการวิเคราะห์แบบถดถอยแสดงในรูปที่ 2.5 จากตัวอย่าง เส้นสีน้ำเงินแทนความสัมพันธ์เชิงเส้นที่ประมาณค่าได้ โดยในที่นี้เราอาจพิจารณาให้ข้อมูลที่เบี่ยงเบนไปจากเส้นถดถอยมากเกินไปเป็นค่าผิดปกติ (ตัวที่อยู่ต่ำกว่าเส้นทั้งสองตัว)



รูปภาพ 2.5: ตัวอย่างการใช้งานการวิเคราะห์แบบถดถอยเพื่อตรวจจับค่าผิดปกติ

2.5.4 การแปลงและปรับบรรทัดฐานข้อมูล

ในหลายโอกาส ข้อมูลดิบที่เก็บมาได้อาจยังไม่อยู่ในรูปแบบที่เหมาะสมในการนำมาประมวลผลทันที เราจึงจำเป็นต้องทำการแปลงข้อมูลดิบ ให้เป็นข้อมูลที่เหมาะสมสำหรับนำไปผ่านกระบวนการทำเหมืองข้อมูล ในทางคณิตศาสตร์ การแปลงข้อมูลคือการใช้ฟังก์ชันที่ทำการส่ง (Mapping Function) ข้อมูลดิบ z ไปยังข้อมูลปลายทาง x ที่นำไปใช้คำนวณได้

$$x = f(z) \tag{2.25}$$

โดยทั่วไป การแปลงข้อมูลสามารถแบ่งออกเป็น 3 ประเภทคือ

การสร้างคุณลักษณะ

การสร้างคุณลักษณะ (Feature Construction) หมายถึงการสร้างหรือสกัดคุณลักษณะเด่นออกมาจากข้อมูลดิบ ข้อมูลที่จำเป็นต้องนำมาสกัดคุณลักษณะ ส่วนใหญ่แล้วจะเป็นข้อมูลที่ไม่สามารถอธิบายจุดเด่นออกมาในเชิงตัวเลขได้โดยง่าย ตัวอย่างที่เห็นได้ชัดคือ การดึงลักษณะเด่นจากรูปภาพ เป็นที่ทราบกันดีกว่า รูปภาพดิจิตอลสี ประกอบด้วยหน่วยย่อยที่เรียกว่า พิกเซล (Pixel) ซึ่งสามารถมองเป็นเวคเตอร์ขนาด 3 มิติ โดยแต่ละมิติแทนค่าความเข้ม สีแดง สีเขียว และสีน้ำเงินตามลำดับ การจะแปลงภาพดิจิตอลเพื่อนำมาเข้ากระบวนการทำเหมืองข้อมูล เราสามารถนำพิกเซลตัวที่ 1 มาเรียงต่อกันไปเรื่อยๆ จนถึงพิกเซลตัวสุดท้าย หากกำหนดให้ภาพมีขนาด 100 คูณ 100 พิกเซล หลังจากการแปลงข้อมูลให้เป็นรูปเวคเตอร์ เราจะได้เวคเตอร์ขนาด 1 $\times$ 10000 หรือ 10000 มิติ ซึ่งถือว่าสูงมาก สมมุติว่า ภาพดิจิตอลดังกล่าวมี สุนัข 1 ตัวอยู่กลางภาพ ส่วนที่เหลือเป็นพื้นหลัง สมมุติต่อว่า งานของเราต้องการสร้างอัลกอริทึมสำหรับตรวจจับใบหน้าสุนัข เราจะพบว่าข้อมูลของพื้นหลังแทบจะไม่สำคัญสำหรับงานของเราเลย ส่วนสำคัญของภาพคือส่วนหน้าสุนัขที่อยู่ตรงกลางเท่านั้น วิธีการแปลงข้อมูลที่กล่าวไปข้างต้นก็อาจจะไม่เหมาะ เพราะเวคเตอร์ที่ได้หลังจากการแปลง มีข้อมูลที่ไม่สำคัญอยู่มากเกินไป ฉะนั้นแล้ว เพื่อให้การแปลงข้อมูลเพื่อสร้างเป็นเวคเตอร์ของคุณลักษณะ (Feature Vector) ที่มีประสิทธิภาพ เราต้องเลือกใช้วิธีในการสกัดคุณลักษณะที่เหมาะสม สำหรับการสกัดคุณลักษณะจากภาพ ถูกแบ่งออกเป็น 3 แบบตามคุณลักษณะที่ต้องการสกัดคือ

- การสกัดลักษณะเชิงพื้นผิว (Texture Extraction) เป็นการพยายามสกัดลักษณะสำคัญที่อยู่บนพื้นผิวออกมา เช่น ความขรุขระ ลักษณะลายเส้น ตัวอย่างวิธีการที่มุ่งดึงลักษณะเชิงพื้นผิวได้แก่ วิธี Local Binary Pattern ใน [Ojala et al., 2002]

- การสกัดลักษณะเชิงรูปทรง (Shape Extraction) เป็นการพยายามสกัดลักษณะเด่นทางรูปทรงของสิ่งที่อยู่ในภาพ เช่น ทรงกลม เส้นตรง หรือสี่เหลี่ยม ตัวอย่างวิธีการสกัดรูปทรงจากรูปภาพได้แก่ [Ke and Sukthankar, 2004]

- การสกัดลักษณะทางสี (Colour Extraction) เป็นการพยายามสกัดสีออกมาจากรูปภาพ เพื่อใช้ในการรู้จำ หรือจำแนกรูปภาพ ตัวอย่างวิธีการสกัดสีจากรูปภาพได้แก่ [Manjunath et al., 2001]

ในที่นี้เราจะไม่ลงรายละเอียดของการดึงหรือสกัดคุณลักษณะที่กล่าวไป ผู้สนใจสามารถศึกษาเพิ่มเติมได้จากเอกสารเกี่ยวกับ การประมวลผลภาพ (Image Processing) [Sonka et al., 2014] ต่อได้ เช่นเดียวกันกับข้อมูลประเภทรูปภาพ การสร้างหรือสกลัดคุณลักษณะจากข้อมูลดิบประเภทเสียง ก็มีความท้าทายเช่นกัน แนวทางพื้นฐานที่สุดในการดึงคุณลักษณะจากเสียงคือ การแทนคลื่นเสียงด้วยวิธีสุ่มเก็บค่าแอมพลิจูดของคลื่นเสียง ณ ตำแหน่งเวลาต่างๆกัน แล้วสร้างเป็นเวคเตอร์ของแอมปลิจูดของเสียง

การปรับบรรทัดฐาน

การปรับบรรทัดฐาน คือการปรับค่าของคุณลักษณะในมิติหนึ่งๆให้อยู่ในช่วงใหม่ที่ต้องการ การปรับบรรทัดฐานเป็นขั้นตอนหลักที่ใช้ในการปรับช่วงของคุณลักษณะที่มีช่วงค่าต่างกันมากๆ ให้มีช่วงค่าที่ใกล้เคียงกัน ซึ่งมีวิธีการทำ 3 วิธี คือ

1. การปรับบรรทัดฐานแบบ Min-Max จะทำการปรับค่าของคุณลักษณะซึ่งเดิมอยู่ในช่วง $[\min, \max]$ ให้อยู่ในช่วงใหม่ $[\min_n, \max_n]$ มีวิธีการคำนวณคือ

$$v' = \frac{v - \min}{\max - \min}(\max_n - \min_n) + \min_n \tag{2.26}$$

2. การปรับบรรทัดฐานโดยทำให้เป็น Z-score คือการแปลงค่าของคุณลักษณะให้มีค่าเฉลี่ย $\mu = 0$ และส่วนเบี่ยงเบนมาตรฐาน $\sigma = 1$ มีวิธีการคำนวณคือ

$$v' = \frac{v - \mu}{\sigma} \tag{2.27}$$

3. การปรับบรรทัดฐานโดยการปรับทศนิยม (Decimal Scaling) คือการแปลงค่าของคุณลักษณะให้อยู่ในรูปทศนิยมที่น้อยกว่า 1 ที่มากที่สุด มีวิธีการคำนวณคือ

$$v' = \frac{v}{10^j} \quad \text{โดยที่ } j \text{ เป็นจำนวนเต็มที่น้อยที่สุดที่ทำให้ } \max(|v'|) < 1 \tag{2.28}$$

การทำให้เป็นค่าดิสครีต

การทำให้เป็นค่าดิสครีต (Discretisation) คือการแปลงค่าที่ต่อเนื่องให้เป็นดิสครีต หรืออีกนัยหนึ่งคือ การแบ่งช่วงให้กับค่าต่อเนื่อง จากนั้นค่าตัวแทนของช่วงจะถูกใช้เพื่อแทนค่า ของค่าจริงที่เป็นค่าต่อเนื่อง การทำให้เป็นค่าดิสครีตส่วนใหญ่แล้วมีจุดประสงค์เพื่อ ลดจำนวนชุดข้อมูลลง ดังนั้นขั้นตอนวิธีที่จะนำมาใช้ในการแปลงค่าให้เป็นดิสครีต จะคล้ายๆกับเทคนิคการลดจำนวนชุดข้อมูลที่เคยกล่าวไปแล้ว ได้แก่

- การแบ่งนับข้อมูล เป็นการแปลงค่าดิสครีตแบบบนลงล่าง หมายความว่า เริ่มต้นจากชุดข้อมูลทั้งหมด แล้วพยายามแบ่งข้อมูลทั้งหมดให้เป็นช่วงย่อยๆ

- การสร้างฮิสโทแกรม มีลักษณะเป็นแบบบนลงล่าง ซึ่งมีลักษณะคล้ายกับการแบ่งนับข้อมูลข้างต้น

- การสร้างกลุ่มของข้อมูลในชุดข้อมูลนั้น เป็นได้ทั้งแบบบนลงล่าง และแบบล่างขึ้นบน แล้วแต่ขั้นตอนวิธีในการจัดกลุ่มข้อมูลที่จะเลือกใช้ วิธีที่มีลักษณะบนลงล่างได้แก่ ขั้นตอนวิธีแบบเคมีนส์ ส่วนวิธีแบบล่างขึ้นบน ได้แก่ขั้นตอนวิธีแบบลำดับขั้น

- การวิเคราะห์ความสัมพันธ์ มีลักษณะแบบล่างขึ้นบน มีจุดมุ่งหมายเพื่อรวมข้อมูลที่เหมือนกัน เข้าด้วยกันเป็นกลุ่ม

ทั้งนี้เมื่อขั้นตอนวิธีทั้งหมดหากลุ่มหรือกล่องของข้อมูลได้แล้ว อาจจะมีการใช้ค่ากลางของกลุ่มหรือกล่องเป็นตัวแทนของกลุ่มหรือกล่องต่อไป

แบบฝึกหัด

1. เราควรเลือกเก็บค่าต่อไปนี้โดยใช้คุณลักษณะแบบไหน

    - อาชีพ
    - แนวดนตรี
    - รายได้

2. สัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน เป็นการวิเคราะห์ความสัมพันธ์ที่ใช้กันอย่างกว้างขวางวิธีหนึ่ง กระนั้นก็ตามวิธีการนี้ก็ไม่สามารถนำมาใช้หาความสัมพันธ์ได้ในทุกกรณี จงอธิบายว่าในกรณีไหนที่การวิเคราะห์ความสัมพันธ์แบบนี้ใช้ไม่ได้ผล

3. กำหนดชุดข้อมูลหนึ่งมิติประกอบด้วยสมาชิกดังต่อไปนี้

$$2, 3, 4, -1, 11, 10, 7, -5, 2, 3, 6, 0, 4, 2$$

    จงแสดงวิธีหามัธยฐานและฐานนิยมของชุดข้อมูลข้างต้น

4. บางครั้งค่าผิดปกติ หรือค่าสุดโต่งอาจจะมีประโยชน์ต่อการทำเหมืองข้อมูล ลองยกตัวอย่างเหตุการณ์หรือภารกิจที่สามารถนำค่าดังกล่าวมาประยุกต์ใช้งานได้

5. จากการสำรวจนักศึกษาในมหาวิทยาลัยเชียงใหม่จำนวน 1000 คน ในสองประเด็นคือ 1. นักศึกษาคนนั้นชอบกีฬาฟุตบอลหรือไม่ และ 2. ถ้าเลือกได้หนึ่งสีนักศึกษาคนนั้นจะเลือกสีชมพูหรือสีน้ำเงิน ผลปรากฏว่า จำนวนนักศึกษาที่

    - ชอบกีฬาฟุตบอลและชอบสีชมพู มี 50 คน
    - ชอบกีฬาฟุตบอลและชอบสีน้ำเงิน มี 550 คน
    - ไม่ชอบกีฬาฟุตบอลและชอบสีชมพู มี 250 คน
    - ไม่ชอบกีฬาฟุตบอลและชอบสีน้ำเงิน มี 150 คน

    จากข้อมูลข้างต้นจงสร้างตารางการจรและคำนวณค่า χ^2 เพื่อวิเคราะห์ความสัมพันธ์ของตัวแปรสองตัวแปรดังกล่าว

บทที่ 3

การลดมิติข้อมูล

หลายครั้งในการทำเหมืองข้อมูล เราอาจต้องเผชิญกับสถานการณ์ที่ข้อมูลนำเข้ามีมิติที่สูงมาก เป็นที่ทราบกันว่าหากข้อมูลอยู่ในมิติที่สูงมากๆ การที่จะทำเหมืองข้อมูลได้มีประสิทธิภาพ จำเป็นที่จะต้องใช้จำนวนข้อมูลมากขึ้นตามไปด้วย น่าเสียดายที่ว่าอัตราการเพิ่มของจำนวนข้อมูลที่ต้อง การไม่ได้เป็นแบบเชิงเส้น แต่อัตราเพิ่มขึ้นเร็วกว่านั้นมากแบบก้าวกระโดด

นอกจากนั้น เมื่อมิติของข้อมูลเพิ่มขึ้นเรื่อยๆ ระยะทางระหว่างข้อมูลสองตัวใดๆจะลู่เข้าสู่ค่าเดียวกัน ซึ่งอาจส่งผลให้ขั้นตอนวิธีที่อาศัยการวัดระยะทางระหว่างข้อมูล เช่น การจัดกลุ่มข้อมูลแบบเคมีนส์ (k-means clustering) หรือวิธีเพื่อนบ้านใกล้เคียง k ตัว (k-Nearest Neighbours) มีประสิทธิภาพลดน้อยลงตามไปด้วย [Indyk and Motwani, 1998] เนื่องมาจากเหตุการณ์ที่ระยะทางของข้อมูลสองตัวใดๆมีค่าใกล้เคียงกันทั้งหมด ทำให้การแยกแยะข้อมูลที่ถือว่าเพื่อนบ้านออกจากข้อมูลที่ไม่ใช่เพื่อนบ้าน ทำได้ลำบากขึ้น เช่นกันเดียวกับในกรณีของการจำแนกข้อมูลที่มีเป้าหมายเพื่อหาระนาบ (Hyperplane) ที่ใช้แบ่งข้อมูลก็อาจจะทำได้ลำบากขึ้น เพราะว่ามีระนาบที่ต้องเลือกมากขึ้นเนื่องจากปริภูมิของข้อมูลสูงขึ้น ส่งผลให้การหาระนาบที่แบ่งกลุ่มข้อมูลได้ดีที่สุด ยากขึ้นตามไปด้วย ปัญหาที่กล่าวมาทั้งหมดมีต้นเหตุจากการที่ข้อมูลมีมิติที่สูงมากๆ เรามักสรุปรวมเรียกความยุ่งยากที่เกิดขึ้นจากมิติข้อมูลว่า คำสาปของมิติข้อมูล (Curse of Dimensionality) [Friedman, 1997]

การลดมิติของข้อมูล มีเป้าหมายเพื่อลดผลกระทบจากคำสาปของมิติข้อมูล แถมยังมีส่วนช่วยเพิ่มความเร็วในการประมวลผล โดยตัดมิติที่ไม่เกี่ยวข้องออกจากสารบบ ในบางโอกาสการลดมิติของข้อมูล ยังถูกนำมาใช้เพื่ออำนวยความสะดวกในการสร้างมโนภาพของข้อมูลที่มีมิติสูงๆ โดยการลดมิติข้อมูลให้เหลือ 3 มิติหรือน้อยกว่านั้น ทั้งนี้เนื่องมาจากประสาทรับรู้ทางสายตาของคนเราสามารถรับรู้ผ่านการมองเห็นได้สูงสุดแค่ 3 มิติ

เทคนิคการลดมิติของข้อมูลที่จะกล่าวถึงต่อไปนี้ มี 3 วิธีคือ การวิเคราะห์องค์ประกอบหลัก (Principle

Component Analysis) การเลือกเซตย่อยของคุณลักษณะ (Feature Subset Selection) และการทำโปรเจคชันแบบสุ่ม (Random Projection)

3.1 การวิเคราะห์องค์ประกอบหลัก

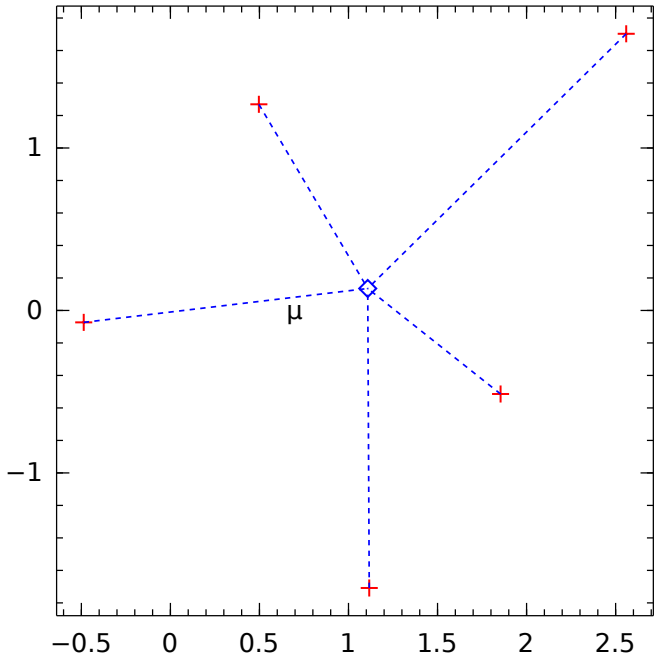
เราจะเริ่มจากการพิจารณาปัญหาการแสดงข้อมูลที่อยู่ใน M มิติ จำนวน N ตัว $x_1, \ldots, x_N$ ด้วยเวคเตอร์ x_0 เพียงตัวเดียว การกระทำดังกล่าวถือเป็นการสรุปชุดข้อมูลให้อยู่ในมิติที่ต่ำลง แน่นอนว่าเราต้องการหา x_0 ที่ดีที่สุดในการแสดงข้อมูลทั้งหมด เพื่อที่จะสามารถบอกได้ว่าเวคเตอร์ตัวหนึ่งดีกว่าอีกตัวอย่างไร จำเป็นจะต้องกำหนดมาตรวัดประสิทธิภาพของ x_0 แต่ละตัวที่เป็นไปได้ ในที่นี้จะกำหนดมาตรวัดประสิทธิภาพดังกล่าวให้คือระยะทางกำลังสองของ x_0 ไปยัง $\{x_n\}_{n=1}^N$ ซึ่งหาก x_0 เป็นเวคเตอร์ที่ดีแล้วระยะทางกำลังสองดังกล่าวจะต้องน้อยที่สุด เรานิยามฟังก์ชันความคลาดเคลื่อนกำลังสองในชื่อ $f_0(x_0)$ ไว้ว่า

$$f_0(x_0) = \sum_{n=1}^{N} ||x_0 - x_n||^2 \tag{3.1}$$

โดยที่ $||x|| = \sqrt{(x^1)^2 + (x^2)^2 + \cdots + (x^M)^2}$ เราสามารถแสดงให้เห็นได้ว่าฟังก์ชันดังกล่าวเป็นฟังก์ชันคอนเวกซ์ (Convex Function) และ x_0 ที่ทำให้ฟังก์ชันคอนเวกซ์มีค่าน้อยที่สุด จะคือ x_0 ที่ทำให้อนุพันธ์ของฟังก์ชันเมื่อเทียบกับ x_0 มีค่าเท่ากับ 0 จากการหาอนุพันธ์เราได้ว่า

$$\begin{aligned} f_0 &= \sum_{n=1}^{N} (x_0 - x_n)^2 \\ \frac{\partial f_0}{\partial x_0} &= \sum_{n=1}^{N} 2(x_0 - x_n) = 0 \qquad (3.2) \\ x_0 &= \frac{1}{N} \sum_{n=1}^{N} x_n = \mu \qquad (3.3) \end{aligned}$$

จากผลของการหาอนุพันธ์เราพบว่า ค่า x_0 ที่ทำให้ฟังก์ชันจุดประสงค์มีค่าน้อยที่สุด ก็คือค่าเฉลี่ยของข้อมูลในชุดข้อมูล ซึ่งในที่นี้แทนด้วย μ นั่นเอง ซึ่งเชื่อมโยงกับเป้าหมายเดิมของเราได้ว่า การที่จะสรุปชุดข้อมูล

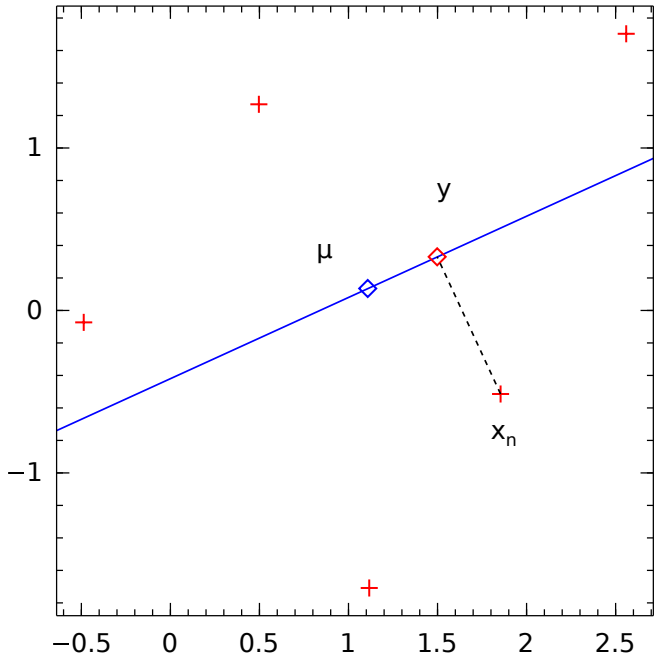


รูปภาพ 3.1: แผนภาพแสดงการสรุปชุดข้อมูลด้วยตัวแทนข้อมูลในศูนย์มิติ (μ) โดยระยะรวมทางจากจุดใดๆ ไปยังค่าเฉลี่ย (จุดสีน้ำเงิน) มีค่าน้อยที่สุด

ทั้งหมดโดยใช้เวคเตอร์ตัวเดียว สามารถทำได้โดยใช้ค่าเฉลี่ยเลขคณิตเป็นตัวสรุป เพราะจะทำให้ระยะทางกำลังสองน้อยที่สุด ตัวอย่างการสรุปชุดข้อมูลโดยอาศัยตัวแทนเป็นค่าเฉลี่ยแสดงในรูปภาพที่ 3.1

ถึงแม้ว่าการสรุปชุดข้อมูลด้วยค่าเฉลี่ยอาจเป็นสิ่งที่ทำได้ง่าย แต่ค่าเฉลี่ยยังขาดสารสนเทศสำคัญเกี่ยวกับข้อมูลอย่างหนึ่ง นั่นคือค่าเฉลี่ยไม่สามารถสรุปการกระจายตัวของชุดข้อมูลได้ อย่างไรก็ตามเราสามารถหาตัวแทนข้อมูลที่มีประโยชน์กว่าค่าเฉลี่ยที่เราหาได้ข้างต้น โดยการเพิ่มมิติของตัวแทนข้อมูลจากจุด (Point) ซึ่งถือว่าเป็นตัวแทนข้อมูลในศูนย์มิติ มาเป็นการหาตัวแทนข้อมูลที่แสดงด้วยเส้นตรง (Line) ซึ่งเป็นตัวแทนข้อมูลในหนึ่งมิติ การสรุปข้อมูลสามารถทำได้โดยฉาย (Project) ข้อมูลลงบนเส้นตรงที่ลากผ่านจุดกลางของข้อมูล ในที่นี้การฉาย (Projection) เวคเตอร์ $x = \left(x^1, x^2, \ldots, x^M\right)$ ลงบนเวคเตอร์ $e = \left(e^1, e^2, \ldots, e^M\right)$ หมายความถึงผลรวมเชิงเส้น (Linear Combination) ระหว่างเวคเตอร์สองตัวซึ่งมีค่าเท่ากับ

$$e^T x = \sum_{i=1}^{M} e^i x^i \tag{3.4}$$



รูปภาพ 3.2: แผนภาพแสดงการสรุปชุดข้อมูลด้วยตัวแทนข้อมูลในหนึ่งมิติ

กำหนดเวคเตอร์ e เป็นเวคเตอร์ขนาดหนึ่งหน่วยซึ่งลากผ่านค่ากลาง μ (เส้นสีน้ำเงินในรูปภาพที่ 3.2) เราสามารถเขียนสมการแสดงจุดใดๆบนเส้นตรงดังกล่าวได้โดย

$$y = \mu + ae \tag{3.5}$$

ในสมการข้างต้น a คือค่าที่แสดงระยะห่างระหว่างค่าเฉลี่ย μ ไปยังจุด x' (อาจแปลความหมายได้ว่าเป็นกี่เท่าของเวคเตอร์หนึ่งหน่วยที่มีทิศทางไปทางเดียวกับเส้นสีน้ำเงิน) หากเราเลือกที่จะสรุปชุดข้อมูลโดยการแทน x_n ด้วยค่าของ y จากสมการ 3.5 เราพบว่า ความคลาดเคลื่อนของตัวแทน $y_n = \mu + a_n e$ จากข้อมูลจริง x_n เมื่อวัดได้โดยฟังก์ชันระยะทางกำลังสองมีค่าเท่ากับ

$$\begin{aligned} f_1(a_1, \ldots, a_N, e) &= \sum_{n=1}^{N} ||(\mu + a_n e) - x_n||^2 \\ &= \sum_{n=1}^{N} a_n^2 ||e||^2 - 2\sum_{n=1}^{N} a_n e^T (x_n - \mu) + \sum_{n=1}^{N} ||x_n - \mu||^2 \end{aligned} \tag{3.6}$$

สังเกตว่าฟังก์ชันจุดประสงค์ขึ้นอยู่กับตัวแปร a_n และ e ก่อนที่จะตามหาทิศทาง e ที่ดีที่สุดในการฉายข้อมูลลงไป เราต้องการหาว่า ณ จุดที่ฟังก์ชันจุดประสงค์มีค่าน้อยที่สุดสัมประสิทธิ์ a_n จะมีค่าเท่าไหร่ จากวิชาแคลคูลัส เราสามารถหาค่าของ a_n ณ จุดที่ฟังก์ชันจุดประสงค์มีค่าน้อยที่สุด ซึ่งก็คือจุดที่ทำให้อนุพันธ์ของ $f_1(\cdot)$ มีค่าเท่ากับศูนย์ เมื่อลองหาอนุพันธ์ย่อย (Partial Derivative) ของ $f_1(\cdot)$ เทียบกับ a_n เราพบว่า

$$\frac{\partial f_1}{\partial a_n} = 2a_n||e||^2 - 2e^T(x_n - \mu) = 0$$

$$a_n = e^T(x_n - \mu) = x'_n \tag{3.7}$$

สมการข้างต้นบอกว่าค่าต่ำสุดของฟังก์ชัน f_1 สามารถได้มาโดยการฉาย x_n ลงบนเวคเตอร์ e โดยผลลัพธ์ของการฉายเราจะให้ชื่อว่า x'_n เมื่อทราบเช่นนี้แล้วเราอยากจะตามหา e ที่ดีที่สุดสำหรับการฉาย x_n ลงไป เมื่อแทนค่า a_n ที่หาได้จากสมการที่ 3.7 ลงในสมการที่ 3.6 เราจะได้ว่า

$$f_1(e) = \sum_{n=1}^{N} a_n^2 - 2\sum_{n=1}^{N} a_n^2 + \sum_{n=1}^{N} ||x_k - \mu||^2 \tag{3.8}$$

$$= -\sum_{n=1}^{N} [e^T(x_n - \mu)]^2 + \sum_{n=1}^{N} ||x_n - \mu||^2 \tag{3.9}$$

$$= -\sum_{n=1}^{N} e^T(x_n - \mu)(x_n - \mu)^T e + \sum_{n=1}^{N} ||x_n - \mu||^2 \tag{3.10}$$

$$= -e^T S e + \sum_{n=1}^{N} ||x_n - \mu||^2 \tag{3.11}$$

โดยเรานิยามเมทริกซ์การกระจาย (Scatter Matrix) ให้มีค่าเท่ากับ $S = \sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T$ เห็นได้ว่าหากต้องการให้ค่า f_1 น้อยที่สุด เราจะต้องทำให้ $e^T S e$ มีค่ามากที่สุดนั่นเอง

การทำค่า $e^T S e$ ให้เพิ่มมากที่สุด สามารถทำได้โดยง่ายโดยเพิ่มขนาดของ e แต่ทว่าการเพิ่มขนาดของเวคเตอร์ e ไม่ได้ทำให้ทิศทางของ e เปลี่ยนไปด้วย เนื่องจากเป้าหมายของเราคือการหา e ที่มีทิศทางที่เมื่อฉายข้อมูลลงมาแล้ว ทำให้ฟังก์ชันระยะทางกำลังสองน้อยที่สุด ดังนั้นหาก e มีทิศทางเดิม (เพิ่มแค่ขนาด) จึงไม่มีประโยชน์ เมื่อเป็นเช่นนี้ เราจะต้องทำการพิจารณาที่จะจำกัดขนาดของเวคเตอร์ e ดังกล่าวไว้ที่ค่าใดค่าหนึ่ง ซึ่งในที่นี้เราจำกัดขนาดของ e ไว้ที่ 1 นั่นคือเราต้องการหา e ที่ทำให้ $e^T S e$ มีค่ามากที่สุดโดยที่

$||e|| = 1$ เพื่อจุดประสงค์นี้เราจำเป็นจะต้องใช้ตัวคูณลากรานจ์ (Lagrange Multiplier) λ เพื่อแสดงข้อจำกัดของการหาค่าที่ดีที่สุดของเรา ซึ่งสามารถแสดงในเชิงคณิตศาสตร์ได้โดย

$$u = e^T S e - \lambda(e^T e - 1) \tag{3.12}$$

เพื่อการหาค่าที่ดีที่สุดเราจะใช้แคลคูลัสอีกครั้งหนึ่ง เพื่อหาอนุพันธ์ของฟังก์ชันจุดประสงค์เทียบกับ e

$$\frac{\partial u}{\partial e} = 2eS - 2\lambda e \tag{3.13}$$

จุดที่ทำให้ฟังก์ชันจุดประสงค์มีค่ามากสุดคือจุดที่อนุพันธ์มีค่าเท่ากับศูนย์ เราพบว่าผลลัพธ์ที่ได้มีรูปแบบตรงกับการแก้ระบบสมการลักษณะเฉพาะ (Eigensystem) นั่นคือ e ที่ทำให้ $e^T S e$ มีค่ามากที่สุด จะต้องเป็นเวคเตอร์ลักษณะเฉพาะ (Eigenvector) ของเมทริกซ์การกระจาย S นั่นคือ

$$Se = \lambda e \tag{3.14}$$

โดยที่ λ คือ ค่าลักษณะเฉพาะ (Eigenvalue) สำหรับเวคเตอร์ลักษณะเฉพาะ e ในบางครั้งเราอาจเรียกเมทริกซ์การกระจายว่า เมทริกซ์ความแปรปรวนร่วมและ จากวิชาพีชคณิตเชิงเส้น (Linear Algebra) เราพบว่าสำหรับเมทริกซ์จำนวนจริง ที่เป็นเมทริกซ์จตุรัสขนาด $M \times M$ แล้วจะมีเวคเตอร์ลักษณะเฉพาะ M ตัว และค่าลักษณะเฉพาะ M ตัวเช่นกัน

เพราะว่า $e^T S e = \lambda e^T e = \lambda$, เราจึงได้ว่า หากต้องการจะหาค่าสูงสุดของ $e^T S e$ เราจะต้องเลือกเวคเตอร์ลักษณะเฉพาะ e ที่คู่กับค่าลักษณะเฉพาะของเมทริกซ์การกระจายที่มีค่ามากที่สุด พูดอีกนัยหนึ่งก็คือ ทิศทางที่เราจะฉายข้อมูลลงไปนั้นคือฉายลงไปบนเวคเตอร์ที่ลากผ่านค่าเฉลี่ยของชุดข้อมูล และมีทิศทางไปทางเดียวกับเวคเตอร์ลักษณะเฉพาะที่คู่กับค่าลักษณะเฉพาะที่มีค่าสูงสุด (e_i ที่ λ_i มากที่สุด)

นอกจากการฉายเวคเตอร์ลงบนเวคเตอร์หนึ่งตัว เรายังสามารถฉายเวคเตอร์ลงบนเวคเตอร์หลายๆตัวที่ประกอบขึ้นเป็นฐานหลัก (Basis) พร้อมกันได้ กำหนดให้ E คือเมทริกซ์ฐานหลักที่ประกอบด้วย เวคเตอร์ k ตัว

$$E = \begin{bmatrix} e_1^1 & e_2^1 & \dots & e_k^1 \\ e_1^2 & e_2^2 & \dots & e_k^2 \\ \vdots & \vdots & \vdots & \vdots \\ e_1^M & e_2^M & \dots & e_k^M \end{bmatrix} \tag{3.15}$$

การฉาย x ลงบนฐานหลักดังกล่าวจะทำให้มิติของ x ลดลงจาก M มิติเป็น k มิติ

$$E^T x = \begin{bmatrix} e_1^1 & e_1^2 & \dots & e_1^M \\ e_2^1 & e_2^2 & \dots & e_2^M \\ \vdots & \vdots & \vdots & \vdots \\ e_k^1 & e_k^2 & \dots & e_k^M \end{bmatrix} \begin{bmatrix} x^1 \\ \vdots \\ x^M \end{bmatrix} = \begin{bmatrix} x'^1 \\ x'^2 \\ \vdots \\ x'^k \end{bmatrix} \tag{3.16}$$

จากนิยามข้างต้นเราสามารถขยายขอบเขตการฉายข้อมูลลงเส้นตรง 1 มิติไปยัง การฉายข้อมูลลงบนฐานหลัก $k \leq M$ มิติได้ด้วย การฉายดังกล่าวสามารถทำได้โดย

$$y = \mu + \sum_{m=1}^{k} a_m e_m \tag{3.17}$$

$$= \mu + a^T E \tag{3.18}$$

$$a = E^T(x - \mu) = x' \tag{3.19}$$

ในที่นี้ a (หรือ x') คือเวคเตอร์ขององค์ประกอบหลัก (Principle Components) k ตัว ของ x ที่เกิดจากการฉาย x ลงบนฐานหลัก E ซึ่งประกอบด้วยเวคเตอร์ลักษณะเฉพาะ e ที่คู่กับค่าลักษณะเฉพาะ λ_i ที่ใหญ่ที่สุด k ตัวของ S

ในทางปฏิบัติแล้ว การวิเคราะห์องค์ประกอบหลักจะเริ่มจาก การหาเวคเตอร์ลักษณะเฉพาะและค่าลักษณะเฉพาะของเมทริกซ์ความแปรปรวนร่วม จากนั้นเราจะเลือกเวคเตอร์ลักษณะเฉพาะ k ตัวแรกซึ่งจัดลำดับโดยค่าลักษณะเฉพาะ เพื่อนำมาสร้างเมทริกซ์ E จากนั้นจะทำการคูณข้างหน้า (Pre-multiply) เมทริกซ์ข้อมูล (Data Matrix) ด้วย E เพื่อที่จะลดมิติของเมทริกซ์ข้อมูล M ลงเหลือ k มิติ ขั้นตอนทั้งหมดสรุปไว้ใน Algorithm 1

Algorithm 1 การวิเคราะห์องค์ประกอบหลัก

1: ปรับบรรทัดฐานของ X ▷ เพื่อให้ค่า $\mu_X = 0$
2: S = cov(X) ▷ คำนวณเมทริกซ์ความแปรปรวนร่วมของข้อมูล
3: T = eig(S) ▷ หาเวคเตอร์ลักษณะเฉพาะทั้งหมดของ S
4: Select k most important principle components and put it in matrix E
5: $X_{pca} = E^T X$ ▷ ฉายข้อมูลลงบน E

3.1.1 การเลือกจำนวนองค์ประกอบหลัก

เมื่อลองพิจาราณาระบบสมการลักษณะเฉพาะ

$$Se_i = \lambda_i e_i \tag{3.20}$$

เราพบว่าค่าลักษณะเฉพาะ λ_i เป็นค่าที่แสดงว่าความแปรปรวนก่อนการฉายซึ่งเก็บอยู่ใน S จะเพิ่มหรือลดหลังจากการฉายนั่นเอง สำหรับการฉายข้อมูลลงบนองค์ประกอบหลัก เรามีความต้องการที่จะให้โครงสร้างของข้อมูลส่วนใหญ่ยังคงอยู่ ซึ่งการคงอยู่ของโครงสร้างข้อมูลแปรผันกับความแปรปรวนของชุดข้อมูลหลังการฉาย ฉะนั้นแล้วหากต้องการรักษาโครงสร้างไว้ เราจำเป็นต้องรักษาความแปรปรวนให้คงอยู่มากสุด ความรู้ดังกล่าวนำมาซึ่งการคัดเลือกองค์ประกอบหลักเพื่อนำมาสร้างเมทริกซ์ E โดยเราจะวัดว่า การตัดองค์ประกอบออกไปจำนวนหนึ่ง ทำให้เกิดการลดลงของความแปรปรวนในชุดข้อมูลเท่าไหร่ ค่าความแปรปรวนที่หายไปถือว่าเป็นค่าความคลาดเคลื่อน (Error) ที่เกิดขึ้นจากการตัดองค์ประกอบหลักทิ้ง ซึ่งสามารถวัดได้โดย

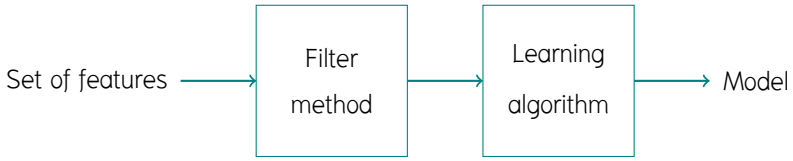
$$\epsilon = \frac{\sum_{i=k+1}^{M} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \tag{3.21}$$

ในที่นี้ k แทนจำนวนองค์ประกอบหลักที่เก็บไว้ หลักการใช้งานกฎนี้ก็คือ เราจะเริ่มจากการเลือกองค์ประกอบหลักมาใส่ในเมทริกซ์ E ทีละตัวแล้วคำนวณค่า ϵ ดังกล่าว เราจะหยุดเลือกองค์ประกอบหลักเพิ่มเมื่อค่า ϵ มีค่าต่ำกว่าค่าขั้นต่ำที่กำหนดไว้

3.2 การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะ (Feature Selection) หมายถึงการเลือกเซตย่อยของคุณลักษณะทั้งหมด เพื่อการนำไปใช้ประมวลผล สาเหตุที่จำเป็นจะต้องทำการคัดเลือกคุณลักษณะ มาจากสมมุติฐานที่ว่าสำหรับงานบางงานคุณลักษณะของข้อมูลที่เก็บ สำรวจ หรือสร้างขึ้นมามีมากเกินไป และมีคุณลักษณะเพียงส่วนหนึ่งเท่านั้นที่เกี่ยวข้องกับงานนั้นๆจริงๆ โดยส่วนใหญ่แล้ว การประยุกต์ใช้การคัดเลือกคุณลักษณะ จะพบได้บ่อยในการทำเหมืองประเภทมุ่งจำแนกหรือทำนายข้อมูล โดยที่คุณลักษณะที่ถูกคัดเลือกมาจากทั้งหมด มักเป็นคุณสมบัติที่เมื่อนำไปสร้างแบบจำลองการทำนายแล้ว ให้แบบจำลองที่ทำนายได้แม่นยำ

การเลือกคุณลักษณะสามารถแบ่งออกเป็น 3 แนวทาง คือ หนึ่งการเลือกคุณลักษณะแบบคัดกรอง (Filter Method) ซึ่งวัดประโยชน์ของคุณลักษณะแต่ละตัวโดยอาศัยเกณฑ์ที่มีพื้นฐานมาจากทฤษฎีสารสนเทศ (In-



รูปภาพ 3.3: การคัดเลือกคุณลักษณะแบบคัดกรอง

formation Theory) [1] สองคือการคัดเลือกคุณลักษณะแบบตัวคลุม (Wrapper Method) ซึ่งจะอาศัยประสิทธิภาพของแบบจำลอง ในการตัดสินว่ากลุ่มของคุณลักษณะที่เลือกมามีประโยชน์หรือไม่ และสุดท้ายคือการคัดเลือกคุณลักษณะแบบฝังตัว (Embedded Method) ซึ่งเกณฑ์ในการเลือกคุณลักษณะ ถูกผนวกเข้าเป็นส่วนเดียวกันกับแบบจำลอง และการคัดเลือกคุณลักษณะจะเกิดขึ้นพร้อมๆกับการประมาณค่า พารามิเตอร์ของแบบจำลอง (การเรียนรู้ของแบบจำลอง) ซึ่งแต่ละวิธีก็จะมีจุดเด่น จุดด้อย แตกต่างกันไป ดังนี้

3.2.1 วิธีคัดกรอง

วิธีการคัดเลือกคุณลักษณะแบบคัดกรอง ทำหน้าที่คล้ายกับการเตรียมข้อมูลก่อนประมวลผล ในเชิงที่ว่าการคัดเลือกคุณลักษณะจะกระทำโดยไม่ขึ้นกับขั้นตอนวิธีหรือแบบจำลองที่จะนำมาวิเคราะห์ข้อมูลต่อ ข้อดีของแนวทางนี้คือ เป็นวิธีที่ค่อนข้างเร็ว และเนื่องจากไม่ขึ้นอยู่กับประเภทของขั้นตอนวิธี ทำให้เซตย่อยของคุณลักษณะที่หาได้ ไม่จำเพาะกับขั้นตอนวิธีหรือแบบจำลอง จึงสามารถนำไปเป็นชุดข้อมูลของขั้นตอนวิธีได้หลากหลาย แต่แน่นอนว่า วิธีนี้จะมีข้อเสียคือ เนื่องจากเซตย่อยของคุณลักษณะ ไม่ได้จำเพาะกับขั้นตอนวิธีใดขั้นตอนวิธีหนึ่ง เซตย่อยของคุณลักษณะที่ได้ จึงอาจไม่ใช่กลุ่มของคุณลักษณะที่ดีที่สุดสำหรับขั้นตอนวิธีที่เราจะต้องใช้ในโอกาสนั้น ทำให้ประสิทธิภาพของผลลัพธ์อาจจะด้อยลงไป เมื่อเทียบกับวิธีที่สร้างมาเฉพาะสำหรับขั้นตอนวิธีดังกล่าว ภาพรวมของการคัดเลือกคุณลักษณะแบบคัดกรองแสดงไว้ในรูปที่ 3.3

สาระสำคัญของการคัดเลือกคุณลักษณะแบบคัดกรอง อยู่ที่การจัดลำดับประโยชน์ หรือความสำคัญของคุณลักษณะแต่ละตัว ต่อเป้าหมายของการทำเหมือง ซึ่งส่วนใหญ่แล้วคือการจำแนกข้อมูลดังที่กล่าวไปตอนต้น เพื่อความเข้าใจวิธีคัดกรองให้ลึกขึ้น เราจะมาศึกษาเกณฑ์ที่สามารถนำมาประยุกต์ใช้ เพื่อวัดประโยชน์ของคุณลักษณะสำหรับการจำแนกข้อมูล ดังต่อไปนี้

อัตราส่วนของสัญญาณต่อสัญญาณรบกวน

เกณฑ์ที่มีประสิทธิภาพเกณฑ์หนึ่ง ที่ใช้ในการพิจารณาการเลือกคุณลักษณะที่จำเป็นในการจำแนกข้อมูล คือการดูว่าคุณลักษณะนั้น สามารถแบ่งข้อมูลออกเป็นสองกลุ่ม ตามที่ต้องการได้ดีเพียงใด ลักษณะที่พึง

[1] ผู้สนใจเรื่อง Information Theory สามารถอ่านเพิ่มเติมได้จาก [MacKay, 2003])

ประสงค์ประการหนึ่งสำหรับการสร้างตัวตันแบบเพื่อจำแนกกลุ่มข้อมูล คือการที่คุณลักษณะสามารถนำไปใช้ในการจำแนกกลุ่ม โดยที่ค่ากลางของกลุ่มข้อมูลสองกลุ่มมีค่าต่างกันมากๆ และข้อมูลที่อยู่ในกลุ่มเดียวกันจะต้องมีการกระจายตัวน้อย เพื่อเราจะใช้เกณฑ์ที่เรียกว่า อัตราส่วนของสัญญาณต่อสัญญาณรบกวน (Signal–2–Noise Ratio) ซึ่งนิยามโดย

$$S2N = \frac{\text{signal}}{\text{noise}} \tag{3.22}$$

มาประยุกต์ เพื่อชี้ว่าคุณลักษณะให้ลักษณะที่พึงประสงค์ดังกล่าวมากน้อยแค่ไหน ในบริบทของการจำแนกข้อมูล จะถือว่าสัญญาณคือระยะห่างของกลุ่มสองกลุ่ม เป็นส่วนที่เราต้องการให้มีค่าสูง ส่วนสัญญาณรบกวนเป็นส่วนที่เราไม่ต้องการและควรมีค่าต่ำ เปรียบเทียบได้กับการกระจายตัวภายในกลุ่มของข้อมูล ดังนั้นแล้วเราสามารถปรับสมการข้างต้นให้เข้ากับบริบทของการจำแนกข้อมูล เพื่อให้ได้สมการทางคณิตศาตร์ ที่ใช้วัดประโยชน์ของคุณลักษณะดังนี้

$$S2N = \frac{\mu_0^k - \mu_1^k}{\sigma_0^k + \sigma_1^k} \tag{3.23}$$

ในที่นี้ กำหนดให้กลุ่มข้อมูลสองกลุ่มมีชื่อว่า กลุ่ม 0 และกลุ่ม 1 ตามลำดับ โดยข้อมูลที่อยู่ในกลุ่มที่ 0 คือข้อมูลที่มีฉลากเป็นเลข 0 กำกับ เช่นเดียวกับกรณีของกลุ่ม 1 ที่เป็นกลุ่มของข้อมูลที่มีฉลากเป็น 1 μ_0^k แสดงค่ากลางของข้อมูลที่มีฉลากเป็น 0 โดยใช้เฉพาะคุณลักษณะที่ k มาเป็น ตัวแบ่งข้อมูลสองกลุ่มออกจากกัน ส่วน σ_0^k ก็คือ การกระจายตัวของข้อมูลที่มีฉลากเป็น 0 โดยพิจารณาเฉพาะคุณลักษณะที่ k นั่นคือ

$$\mu_0^k = \frac{\sum_{n=1}^{N_0} x_n^k}{N_0} \tag{3.24}$$

$$\sigma_0^k = \frac{\sum_{n=1}^{N_0} (x_n^k - \mu_0^k)^2}{N_0} \tag{3.25}$$

ในที่นี้ N_0 คือจำนวนข้อมูลที่มีป้ายของกลุ่มเป็น 0 สำหรับ μ_1^k และ σ_1^k ก็สามารถหาได้ด้วย วิธีการคล้ายๆกัน เมื่อคำนวณ S2N ของคุณลักษณะทุกตัวได้ครบแล้ว เราจะทำการเรียงค่า S2N จากมากไปหาน้อย คุณลักษณะที่มีค่า S2N มากที่สุดในบรรดาคุณลักษณะทั้งหมด M ตัวจะถือว่าเป็นคุณลักษณะที่มีประโยชน์ต่อการจำแนกข้อมูลมากที่สุด

อัตราการขยายของสารสนเทศ

อัตราการขยายของสารสนเทศ (Information Gain) ระหว่างตัวแปรสุ่มสองตัว X, Y เป็นตัวชี้วัดว่า สารสนเทศเกี่ยวกับตัวแปรสุ่ม Y จะเพิ่มมากขึ้นแค่ไหน หากทราบค่าของตัวแปรสุ่ม X ในกรณีที่การทำเหมืองข้อมูลมีจุดมุ่งหมายเพื่อหาค่าของ Y อย่างเช่นการจำแนกข้อมูล ซึ่งโดยทั่วไปแล้วกำหนดให้ Y แทนป้ายของกลุ่ม และ X แทนคุณลักษณะ เราอาจต้องการเลือกคุณลักษณะ X ที่เมื่อทราบค่าแล้ว ทำให้ได้สารสนเทศเกี่ยวกับป้ายของกลุ่ม Y เพิ่มขึ้นมากที่สุด เพราะนั่นหมายถึงการทำนายป้ายของกลุ่ม ก็จะผิดพลาดน้อยลงด้วย

โดยทั่วไปแล้ว ปริมาณสารสนเทศในตัวแปรสุ่มจะเกี่ยวข้องกับหน่วยวัดที่เรียกว่า เอนโทรปี (Entropy) ซึ่งเป็นหน่วยวัด ที่ใช้วัด *ความไม่แน่นอน* ของผลลัพธ์ของเหตุการณ์ใดๆ เอนโทรปีจะมีค่าสูงสุดเมื่อความน่าจะเป็นของการเกิดเหตุการณ์ใดเหตุการณ์หนึ่ง ในจำนวนเหตุการณ์ทั้งหมดมีค่าเท่ากัน ทั้งนี้เนื่องจากเราจะไม่สามารถคาดการณ์อะไรได้เลย ค่าเอนโทรปีซึ่งเป็นหน่วยวัดความไม่แน่นอน จึงมีค่าสูงสุด ตรงกันข้าม หากความน่าจะเป็นของการเกิดเหตุการณ์หนึ่งใดมีค่าเป็น 1 ส่วนความน่าจะเป็นที่จะเกิดเหตุการณ์ที่เหลือเป็น 0 แล้ว เอนโทรปีจะมีค่าต่ำสุด เพราะว่าเราสามารถทราบได้แน่นอนว่า เหตุการณ์ไหนจะเกิดขึ้น หรืออีกนัยหนึ่งคือไม่มีความไม่แน่นอนอยู่นั่นเอง การนิยามเอนโทรปีจะอยู่ในรูปของค่าคาดหวัง (Expected Value) ของปริมาณสารสนเทศ (Information Content) ซึ่งอยู่ในรูปของ ค่าลบของลอการิทึมของความน่าจะเป็นของเหตุการณ์ทั้งหมด

$$I(X) = \log_2(\frac{1}{P(X)}) \tag{3.26}$$

$$\begin{aligned} H(X) &= E[I(X)] && (3.27)\\ &= E[-\log_2(I(X))] && (3.28)\\ &= -\sum_{n=1}^{N} P(x_n)\log_2 P(x_n) && (3.29) \end{aligned}$$

ยกตัวอย่างให้เห็นภาพ ลองพิจารณาการออกหัวก้อยของเหรียญบาทหนึ่งเหรียญ หากเหรียญนั้นเป็นเหรียญที่เที่ยง (Fair) เราจะพบว่าความไม่แน่นอนมีค่าสูงสุด นั่นคือ

$$\begin{aligned} H(X) &= -\sum_{n=1}^{N} P(x_n)\log_2 P(x_n) && (3.30)\\ &= -[0.5\log_2(2^{-1}) + 0.5\log_2(2^{-1})] && (3.31)\\ &= 1 && (3.32) \end{aligned}$$

หากเหรียญนั้นเป็นเหรียญที่ไม่เที่ยง (Biased) เช่นออกหัวตลอด เราจะพบความไม่แน่นอนมีค่าต่ำสุด นั่นคือ

$$H(X) = -\sum_{n=1}^{N} P(x_n) \log_2 P(x_n) \tag{3.33}$$

$$= -[1 \log_2(1) + 0 \log_2(0)] \tag{3.34}$$

$$= 0 \tag{3.35}$$

เมื่อทราบการหาเอนโทรปีแล้ว เราสามารถคำนวณหาอัตราการขยายของสารสนเทศ ซึ่งแสดงในรูปส่วนต่างของเอนโทรปีก่อนทราบค่าตัวแปรสุ่ม X กับเอนโทรปีหลังทราบค่าตัวแปรสุ่ม X ได้โดย

$$IG(X) = H(Y) - H(Y|X) \tag{3.36}$$

สำหรับเหตุการณ์ที่นับได้ (Discrete Event) ของ Y เราจะได้ว่า

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2(P(Y = y_i)) \tag{3.37}$$

และ

$$H(Y|X) = \sum_{j=1}^{r} P(X = x_j) H(Y|X = x_j) \tag{3.38}$$

โดยที่ $H(Y|X = x_j)$ หมายถึงเอนโทรปี ของ Y เมื่อ X มีค่าเท่ากับ x_j นั่นคือ

$$H(Y|X = x_j) = -\sum_{i=1}^{k} P(Y = y_i|X = x_j) \log_2(P(Y = y_i|X = x_j)) \tag{3.39}$$

คุณลักษณะที่มีค่าอัตราการขยายของสารสนเทศสำหรับป้ายของกลุ่มสูง แสดงว่าเป็นคุณลักษณะที่มีประโยชน์สำหรับงานจำแนกข้อมูล

สารสนเทศร่วม

จริงอยู่ว่าเราสามารถเลือกคุณลักษณะโดยอาศัยเกณฑ์สองตัวที่ศึกษาไปตอนต้น แต่ปัญหาที่อาจตามมา คือ คุณลักษณะที่ดีสองตัวใดๆ อาจจะสื่อถึงสารสนเทศเดียวกัน ทำให้เกิดการซ้ำซ้อนขึ้น เพราะเราสามารถ เลือกคุณลักษณะสองตัวดังกล่าว ไว้เพียงตัวเดียวได้

เกณฑ์อันหนึ่ง ที่สามารถนำมาใช้ประกอบในการคัดกรองคุณลักษณะเพื่อแก้ปัญหาความซ้ำซ้อน ในการเลือกคุณลักษณะคือ สารสนเทศร่วม (Mutual Information) สารสนเทศร่วมเป็นเกณฑ์ที่ใช้บอกว่า ตัวแปรสุ่ม X กับตัวแปรสุ่ม Y ให้ สารสนเทศแบบเดียวกันมากน้อยแค่ไหน ตัวอย่างที่อาจทำให้เราเห็นภาพสารสนเทศร่วม ของตัวแปรสุ่มสองตัวในชีวิตจริง ได้แก่ ชื่อบุคคลกับเพศ ถือว่ามีสารสนเทศร่วมสูง เพราะหากทราบชื่อบุคคลก็สามารถเดาได้ว่าคนคนนั้นเป็นเพศหญิงหรือชาย ส่วน อายุและเพศ เป็นตัวแปรสุ่มที่มีสารสนเทศร่วมต่ำ เพราะเมื่อทราบเพศก็อาจจะไม่ง่ายที่จะบอกอายุของบุคคลคนนั้นได้

การคำนวณสารสนเทศร่วมสามารถทำได้โดย

$$I(x,y) = \sum_{x}\sum_{y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3.40}$$

การนำสารสนเทศร่วมไปประยุกต์ใช้ในการเลือกคุณลักษณะ มีสองทางคือ

1. ใช้เพื่อลดความซ้ำซ้อนของคุณลักษณะดังที่กล่าวไปตอนต้น สำหรับกรณีนี้ หากคุณลักษณะสองตัว คือ x_1 และ x_2 มีสารสนเทศร่วมสูง แสดงว่าคุณลักษณะสองตัวสื่อถึงสารสนเทศเดียวกัน ทำให้เราสามารถเลือกตัดคุณลักษณะตัวใดตัวหนึ่งออกไปได้

2. ใช้เพื่อเลือกคุณลักษณะที่มีประโยชน์ ในกรณีนี้จะกำหนดให้ X แทนคุณลักษณะ และ Y แทนป้ายของกลุ่ม ในการจำแนกข้อมูลเราจะถือว่าคุณลักษณะที่เป็นประโยชน์ จะคือคุณลักษณะที่มีสารสนเทศร่วมกับป้ายของกลุ่มมากที่สุด

เมื่อทราบประโยชน์ของคุณลักษณะจากเกณฑ์วัดแต่ละตัวแล้ว วิธีคัดกรองจะทำการสร้างเซตย่อยของคุณลักษณะ โดยเลือกคุณลักษณะที่มีประโยชน์มากสุด $M' < M$ ตัวแรก โดยอาจมีการพิจารณาความซ้ำซ้อนไปด้วย ทว่าการคัดเลือกตามลำดับแบบนี้ มีปัญหาตรงที่ เซตย่อยที่ได้อาจยังไม่ใช่เซตย่อยที่ดีที่สุด จากเหตุผลสองประการคือ หนึ่ง เซตดังกล่าวอาจจะมีคุณลักษณะที่มีสารสนเทศร่วมสูงหลงเหลืออยู่ภายในเซต ทำให้เซตย่อยที่ได้มีขนาดใหญ่เกินควร และ สองคือมีงานวิจัยที่แสดงให้เห็นว่าคุณลักษณะอยู่โดดๆแล้วเหมือนจะไม่มีประโยชน์ (ค่าจากเกณฑ์วัดแต่ละตัวมีค่าต่ำมากๆ) แต่เมื่อนำคุณลักษณะดังกล่าวมารวมกับคุณลักษณะตัวอื่นๆ อาจพบว่ากลุ่มของคุณลักษณะเหล่านั้นมีคุณภาพสูงมากก็เป็นได้ [Guyon and Elisseeff, 2003] ด้วยเหตุนี้จึงมีแนวคิดที่จะทำการเลือกเซตย่อยของคุณลักษณะที่ซับซ้อนขึ้น แทนที่จะอาศัยการจัดลำดับคุณลักษณะอย่างเดียว

3.2.2 วิธีแบบตัวคลุม

สำหรับวิธีการแบบตัวคลุมจะต่างจากวิธีการแบบคัดกรอง โดยจะมีการใช้ประสิทธิภาพของขั้นตอนวิธี เข้ามาเป็นเกณฑ์ในวัดประสิทธิภาพของเซตย่อยของคุณลักษณะอีกชั้นหนึ่ง นั่นคือหลังจากได้เซตย่อยของคุณลักษณะที่น่าจะดีแล้ว เซตย่อยดังกล่าวจะถูกนำไปทดสอบกับขั้นตอนวิธี ที่จะนำมาใช้งานจริง เพื่อประเมินว่าเซตย่อยที่เลือกมาได้ สามารถใช้งานได้ดีแค่ไหน ท้ายสุดแล้วเซตย่อยที่จะถูกเลือกคือ เซตย่อยที่เหมาะสมกับขั้นตอนวิธีนั้นได้ดีที่สุด ข้อดีที่เห็นได้ชัดคือว่า เซตย่อยที่ได้จะมีความจำเพาะกับขั้นตอนวิธีมากที่สุด และอาจทำให้ประสิทธิภาพโดยรวมสูงกว่าการใช้เซตย่อยที่ได้จากวิธีคัดกรอง

เนื่องจากวิธีการนี้อาจไม่ได้เลือกเซตย่อยของคุณลักษณะ บนพื้นฐานของประโยชน์ของคุณลักษณะนั้นเสมอไป วิธีการแบบตัวคลุมจำเป็นจะต้องมียุทธวิธีในการค้นหาเซตย่อย (Search Strategy) ที่ดีและมีประสิทธิภาพ และต้องกำหนดว่าจะวัดประสิทธิภาพของเซตย่อยนั้นได้อย่างไร อย่างเช่น สำหรับการจำแนกข้อมูล อาจจะวัดประสิทธิภาพของเซตย่อย โดยการวัดความแม่นยำในการทำนายฉลากของข้อมูลทดสอบ หรือหากเป็นการวิเคราะห์ถดถอย ประสิทธิภาพอาจจะวัดด้วย ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error)

ในทางปฏิบัติแล้วยุทธวิธีสำหรับการค้นหาเซตย่อย มีได้หลายวิธี ตั้งแต่ การลองทุกกรณี (Brute Force) วิธีแบบละโมบ (Greedy Approach) ไปจนถึง การใช้ขั้นตอนวิธีเชิงวิวัฒนาการ (Evolutionary Algorithm) แน่นอนว่าการลองทุกกรณีที่เป็นไปได้ จะเจอเซตย่อยของคุณลักษณะที่ดีที่สุดแน่นอน แต่ว่าก็ใช้เวลาใน การคำนวณหามากที่สุดเช่นกัน โดยเฉพาะอย่างยิ่งเมื่อจำนวนคุณลักษณะมีเยอะมากๆ ซึ่งจำนวนรูปแบบที่จะต้องค้นหาเพิ่มขึ้นแบบก้าวกระโดด วิธีนี้จึงไม่เป็นที่นิยม เนื่องจากขั้นตอนวิธีเชิงวิวัฒนาการมีรายละเอียดพอสมควรและอยู่นอกเหนือจากบทเรียน [2] ในที่นี้เราจะศึกษาการเลือกเซตย่อยโดยวิธีละโมบ ซึ่งแบ่งเป็น 2 แนวทางหลักคือ การเลือกแบบเดินหน้า (Forward Selection) และการกำจัดแบบย้อนหลัง (Backward Elimination) วิธีค้นหาเซตย่อยแบบละโมภ จะอาศัยเกณฑ์วัดประโยชน์ของคุณลักษณะที่แนะนำไปข้างต้นมาประยุกต์ใช้ด้วย

การเลือกแบบเดินหน้า

การเลือกคุณลักษณะแบบเดินหน้า จะเริ่มจากเซตว่าง ที่ไม่มีคุณลักษณะใดเลย ต่อจากนั้นจะทยอย เพิ่มคุณลักษณะที่เป็นประโยชน์ (โดยพิจารณาเทียบกับคุณลักษณะที่มีอยู่ในเซตอยู่แล้วด้วย) เข้ามาทีละตัว และจะหยุดเพิ่มคุณลักษณะเมื่อเซตย่อยที่ได้มีคุณสมบัติดีพอ ในทางปฏิบัติ เชื่อกันว่าการเลือกแบบเดินหน้าประหยัดทรัพยากรทางด้านการคำนวณมากกว่า การกำจัดแบบย้อนหลัง [Guyon and Elisseeff, 2003]

[2] ผู้สนใจสามารถอ่านเพิ่มเติมได้จาก [Eiben and Smith, 2003]

การกำจัดแบบย้อนหลัง

การกำจัดแบบย้อนหลังคือยุทธวิธีการเลือกเซตย่อย ที่ตรงกันข้ามกับการเลือกแบบเดินหน้า นั่นคือ ในกรณีนี้จะเริ่มจากเซตของคุณลักษณะทั้งหมด แล้วพยายามตัดคุณลักษณะที่ไร้ประโยชน์ ออกทีละตัวจนกระทั่งได้เซตย่อยของคุณลักษณะ ที่มีคุณสมบัติดีพอ ถึงแม้ว่าการกำจัดแบบย้อนหลังอาจจะต้องใช้ทรัพยากรทางการคำนวณสูงกว่าการเลือกแบบเดินหน้า แต่การกำจัดแบบย้อยหลัง อาจเจอเซตย่อยของคุณลักษณะที่มีประสิทธิภาพสูงกว่า ทั้งนี้เนื่องจากคุณลักษณะที่มีประโยชน์น้อยบางตัว อาจมีประโยชน์มาก เมื่ออยู่ร่วมกัน การเลือกแบบเดินหน้าอาจมองข้ามกลุ่มของคุณลักษณะจำพวกนี้ไปก็ได้ [Guyon and Elisseeff, 2003]

3.2.3 วิธีแบบฝังตัว

วิธีแบบฝังตัวเป็นการเลือกคุณลักษณะอีกประเภทหนึ่งที่เฉพาะเจาะจงกับขั้นตอนวิธีมาก ในกรณีนี้กลวิธีและขั้นตอนการคัดเลือกคุณลักษณะจะถูกฝังอยู่ในตัวขั้นตอนวิธี สำหรับวิธีนี้ การประมาณค่าพารามิเตอร์ของแบบจำลองและการคัดเลือกคุณลักษณะจะถูกทำไปพร้อมๆกัน ยกตัวอย่างวิธีการแบบฝังตัว ได้แก่การใช้ เรกูลาไรเซชัน (Regularisation) ให้กับแบบจำลองสำหรับทำนาย

$$f_{obj} = \text{objective} - \text{regularisation} \tag{3.41}$$

ในรูปของฟังก์ชันจุดประสงค์ลบด้วยเรกูลาไรเซชัน ยกตัวอย่าง การถดถอยแบบโลจิสติกที่ใช้ L1 เรกูลาไรเซชัน มีฟังก์ชันจุดประสงค์เป็นดังต่อไปนี้

$$f_{obj} = \underbrace{\arg\min_{w} \sum_{i=1}^{N} y_i(w^T x_i + b)}_{\text{objective function}} + \underbrace{\lambda \sum_{j=1}^{M} |w_j|}_{\text{regularisation}} \tag{3.42}$$

ในที่นี้ λ คือตัวแปรที่เรียกว่า เรกูลาไรเซชันพารามิเตอร์ (Regularisation Parameter) มีไว้กำหนดระดับของการบังคับใช้เรกูลาไรเซชัน ในสมการนี้เรกูลาไรเซชันเป็นแบบ L1 เรกูลาไรเซชัน ซึ่งมีคุณสมบัติทำหน้าที่คัดกรองคุณลักษณะ ให้เหลือแต่คุณลักษณะที่จำเป็น นั่นคือ การประมาณค่าพารามิเตอร์ของแบบจำลอง (w) จะพยายามทำให้ w_j ส่วนใหญ่มีค่าน้อยที่สุด (ในที่นี้คือ 0 เนื่องมาจากค่าสัมบูรณ์) ซึ่งในกรณีที่ w_j เท่ากับ 0 จะเป็นการบอกว่าคุณลักษณะที่ j ไม่มีความสำคัญต่อการทำนาย และเสมือนว่าขั้นตอนวิธีได้ตัดคุณลักษณะที่ j ออกจากการเรียนรู้นั่นเอง รายละเอียดเกี่ยวกับการใช้เรกูลาไรเซชัน สามารถอ่านเพิ่มเติมได้จาก [Ng, 2004]

3.3 โปรเจคชันแบบสุ่ม

การทำงานของการวิเคราะห์องค์ประกอบหลัก (PCA) ที่เคยศึกษาไปเป็นวิธีที่เสมือนกับการตามหาฐานหลักใหม่ในมิติที่ต่ำลง ที่เมื่อฉายข้อมูลลงไปแล้วจะเหมือนกับการลดมิติของข้อมูลลง และง่ายต่อการนำไปคำนวณต่อ แต่เราพบว่า ปัญหาของ PCA คือการคำนวณหาเมทริกซ์ความแปรปรวนร่วม และการหาเวคเตอร์ลักษณะเฉพาะและค่าลักษณะเฉพาะ ซึ่งทำงานได้ช้าในกรณีที่มิติตั้งต้นของข้อมูลมีค่ามากๆ

เพื่อแก้ข้อจำกัดทางด้านเวลา จึงมีแนวคิดที่ว่า หากสามารถหาฐานหลักใหม่โดยคร่าวๆแต่รวดเร็ว และเป็นฐานหลักที่ดีเทียบเท่าหรือใกล้เคียงกับ ฐานหลักที่ได้จาก PCA คงจะดีไม่น้อย เราพบว่า มีทฤษฎีบทตัวหนึ่งที่พิสูจน์ให้เห็นว่า หากเราสร้างฐานหลักโดยสุ่มเลือก สมาชิกที่ประกอบขึ้นมาเป็นเมทริกซ์แสดงฐานหลักเหล่านั้นจากการแจกแจงแบบปกติ และฉายข้อมูลลงไปบนฐานหลักดังกล่าว จะพบว่าค่าเฉลี่ยของข้อมูลหลังจากการฉายแบบสุ่มหลายๆครั้ง โครงสร้างของข้อมูลยังคงถูกรักษาไว้ ไม่สูญหายไปไหน โครงสร้างของข้อมูลในที่นี้วัดจาก การคงสภาพของระยะห่างระหว่างข้อมูลสองตัวใดๆ ก่อนและหลังจากการฉาย วิธีการนี้มีชื่อเรียกว่า โปรเจคชันแบบสุ่ม (Random Projection) ซึ่งเป็นวิธีการที่ถูกคิดค้นบนพื้นฐานของทฤษฎีที่ชื่อว่า Johnson Lindenstrauss Lemma ซึ่งกล่าวไว้ว่า

Lemma 1 ([Johnson et al., 1986]). *Let* $\epsilon \in (0,1)$. *Let* $k, M, N \in \mathcal{N}$ *such that* $k \geq C\epsilon^{-2}\log N$, *for a large enough absolute constant* C. *Let* $V \subseteq \mathcal{R}^M$ *be a set of* N *points. Then there exists a linear mapping* $R : \mathcal{R}^M \rightarrow \mathcal{R}^k$, *such that for all* $u, v \in V$:

$$(1-\epsilon)||u-v||_{l_2^d}^2 \leq ||Ru-Rv||_{l_2^d}^2 \leq (1+\epsilon)||u-v||_{l_2^d}^2 \tag{3.43}$$

จากทฤษฎี V คือเมทริกซ์ของชุดข้อมูล M คือมิติของข้อมูล และ N คือจำนวนข้อมูล เราพบว่า หากคูณข้อมูลสองตัวใดๆ ด้วยเมทริกซ์สุ่ม (Random Matrix) R ทฤษฎีจะรับประกันว่า ระยะห่างสัมพัทธ์ (Relative distance) ของข้อมูลทั้งสองตัวในมิติเดิมและมิติใหม่จะเปลี่ยนแปลงไปไม่มากกว่า $1 \pm \epsilon$ เท่า ซึ่งแปลความได้ว่า โครงสร้างของข้อมูลทั้งหมดไม่ถูกทำลายด้วยการคูณด้วยเมทริกซ์สุ่ม R

เราสามารถนำผลลัพธ์ดังกล่าว ไปประยุกต์ใช้ในการลดมิติของข้อมูลที่มีมิติสูงมากๆ โดยสร้างเมทริกซ์สุ่ม ขนาด $k \times M$ โดยที่ k น้อยกว่า M มากๆ เมทริกซ์ดังกล่าวมีสมาชิกที่สุ่ม มาจากการแจกแจงแบบปกติและฉายชุดข้อมูลลงบนฐานหลัก ซึ่งกำหนดโดยเมทริกซ์สุ่ม ในการฉายแต่ละครั้งผลลัพธ์ที่ได้อย่างมีทั้งดีและไม่ดี เพราะทฤษฎีรับประกันความแม่นยำถึงแค่ $(1 \pm \epsilon)$ แต่เราสามารถทำซ้ำหลายๆครั้ง โดยหวังว่าค่าเฉลี่ยข้อมูลของการสุ่มฉายจะมีคุณภาพมากขึ้น

แบบฝึกหัด

1. จงอธิบายความหมายของเวคเตอร์ลักษณะเฉพาะที่หาได้ในกระบวนการวิเคราะห์องค์ประกอบหลัก
2. จงอธิบายความหมายของค่าลักษณะเฉพาะที่หาได้ในกระบวนการวิเคราะห์องค์ประกอบหลัก
3. เราสามารถนำวิธีการวิเคราะห์องค์ประกอบหลักไปประยุกต์ใช้ในงานสร้างมโนภาพได้อย่างไรบ้าง
4. ลองเสนอแนวคิดในการเลือกจำนวนองค์ประกอบหลักว่าควรจะเลือกองค์ประกอบหลักกี่ตัว
5. การคัดเลือกคุณลักษณะด้วยวิธีแบบตัวคลุมชนิดเลือกเดินหน้า และกำจัดย้อนหลัง มีข้อดีข้อเสียแตกต่างกันอย่างไร

บทที่ 4

การทำเหมืองเพื่อหาแบบรูปและความสัมพันธ์

เนื้อหาในส่วนนี้จะกล่าวถึงการทำเหมืองข้อมูลที่มุ่งหาความสัมพันธ์ระหว่างเซตของสิ่งของ ที่มักพบเจอพร้อมกันบ่อยๆ ความสัมพันธ์ที่ได้สามารถนำไปประยุกต์ใช้ในด้านธุรกิจ เช่นการพบว่าสินค้าสองสิ่งมักขายได้พร้อมกันอาจนำไปสู่การจัดการส่งเสริมการขาย นอกจากการหาความสัมพันธ์ของเซตของสิ่งของแล้ว เราจะทำความรู้จักและศึกษาระบบแนะนำ (Recommendation Systems) ซึ่งมีเป้าหมายอยู่ที่การทำนายความชอบของบุคคลต่อสินค้าหรือสิ่งของอย่างใดอย่างหนึ่ง ผลลัพธ์ที่ได้อาจนำไปสู่การแนะนำสินค้าที่ผู้ใช้อาจจะชอบ แต่ยังไม่เคยซื้อหรือใช้ ให้แก่ผู้ใช้ได้ ระบบดังกล่าวก็เริ่มมีการนำมาใช้งานกันมากขึ้นในวงการธุรกิจในปัจจุบันด้วย

4.1 การหากฎความสัมพันธ์

กฎความสัมพันธ์ (Association Rule) คือกฎทางตรรกศาสตร์ ที่ใช้ระบุความสัมพันธ์ของเซตสองเซต ในรูปแบบ ถ้า. . . แล้ว ซึ่งสามารถแสดงในเชิงสัญลักษณ์ทางคณิตศาสตร์ได้ว่า $A \implies B$ ในที่นี้ A และ B แสดงเซตของสิ่งของ 2 เซตดังกล่าว ก่อนจะเข้าสู่ขั้นตอนการหากฎความสัมพันธ์ เราจะมาทำความรู้จักกับ สิ่งที่เรียกว่า แบบรูปที่พบบ่อย (Frequent Pattern) ซึ่งก็คือแบบรูปที่เกิดขึ้นบ่อย แบบรูปที่พบบ่อยอาจใช้เรียกรวมแบบรูปต่างๆต่อไปนี้

- ไอเทมเซตที่พบบ่อย (Frequent Itemset) คือ เซตของสิ่งของที่มักเกิดขึ้นพร้อมๆกัน อย่างเช่น เรามัก

จะพบว่า ลูกค้าจะซื้อครีมกันแดดกับน้ำแข็งพร้อมกันเสมอ

- แบบรูปเชิงลำดับที่พบบ่อย (Frequent Sequential Pattern) คือลักษณะการเกิดขึ้นพร้อมๆกัน แบบมีลำดับ เช่น เรามักจะพบว่าลูกค้าที่ซื้องคอมพิวเตอร์ไปแล้วช่วงหนึ่ง จะกลับมาซื้อปริ้นท์เตอร์เสมอ

- แบบรูปเชิงโครงสร้างที่พบบ่อย (Frequent Structured Pattern) คือลักษณะการเกิดขึ้นพร้อมๆกันของสิ่งของแบบมีโครงสร้าง เช่น หากร่างกายมีการสร้างกลุ่มโปรตีนที่มีโครงสร้างแบบ A แล้ว เราจะพบว่าร่างกายก็จะสร้างโปรตีนที่มีโครงสร้างแบบ B เสมอ

การหาแบบรูปที่พบบ่อย ถือเป็นก้าวสำคัญเพื่อนำไปสู่การสร้างกฎความสัมพันธ์ หรือความสัมพันธ์แบบอื่นๆ ในข้อมูล เนื้อหาต่อจากนี้ไป เราจะศึกษาขั้นตอนวิธีที่สามารถนำมาใช้ในการหาแบบรูปที่พบบ่อยได้อย่างมีประสิทธิภาพ

ตัวอย่างที่ทำให้เห็นภาพและจินตนาการตามได้ง่าย ในการหาแบบรูปที่พบบ่อย คือตัวอย่างการวิเคราะห์ตะกร้าสินค้า (Market Basket Analysis) กระบวนการดังกล่าว จะทำการวิเคราะห์อุปนิสัยการจับจ่ายของลูกค้า โดยการหาความสัมพันธ์ระหว่างสินค้าที่ลูกค้าซื้อไป ผลลัพธ์ของการวิเคราะห์ จะทำให้ได้ความสัมพันธ์ ซึ่งสามารถนำไปประยุกต์ใช้ในการวางแผนการตลาดต่อไป ยกตัวอย่างเช่น หากเราพบว่าลูกค้ามักจะซื้อนมพร้อมกับขนมปัง แบบรูปที่พบบ่อยซึ่งในที่นี้จะเรียกว่าไอเทมเซตที่พบบ่อย สามารถนำไปสู่การจัดวางร้านค้าใหม่ โดยจัดวางผลิตภัณฑ์นมไว้ใกล้กับขนมปังหรือเบเกอรี่ หรือในอีกมุมมองหนึ่ง เราอาจจะวางนมไว้ห่างจากขนมปังพอประมาณ แล้วจัดวางแยม น้ำผึ้งไว้ระหว่างทาง จากชั้นวางนมไปยังชั้นวางขนมปัง ซึ่งทั้งนี้ก็แล้วแต่ผู้ประกอบการจะนำไปวางแผนต่อไป

4.1.1 ไอเทมเซตที่พบบ่อยและกฎความสัมพันธ์

ในที่นี้เพื่อให้ง่ายต่อการทำความเข้าใจ เราจะแทนไอเทมด้วยสินค้า ทั้งนี้ไอเทมไม่จำเป็นจะต้องเป็นสินค้าเท่านั้น อาจจะเป็นสิ่งของอย่างอื่น แล้วแต่บริบทของการนำไปใช้งาน

กำหนดให้ $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ แทนเซตของสินค้าทั้งหมดในร้าน และ กำหนดให้ $D = \{t_1, t_2, \ldots, t_m\}$ แทนเซตของการซื้อขาย (Transaction Set) โดยที่การซื้อขายที่ t_i เป็นเซตย่อยของ เซตของสิ่งของข้างต้น $t_i \subseteq \mathcal{I}$ จากนั้นกำหนดให้ A แทนเซตของสินค้าที่ลูกค้าซื้อ เราจะถือว่าการซื้อขายที่ t_i รวม A อยู่ด้วยถ้า $A \subseteq t_i$ จากนิยามข้างต้น เราสามารถนิยามกฎความสัมพันธ์ ซึ่งคือ ประพจน์แบบมีเงื่อนไข (Implication) ได้ในรูปของ

$$A \implies B \tag{4.1}$$

โดยที่ $A \subset \mathcal{I}, B \subset \mathcal{I}$ และ $A \cap B = \emptyset$ สำหรับเซตของการซื้อขาย D ใดๆ เราสามารถคำนวณอัตราส่วน ระหว่างการซื้อขายใน D ที่ประกอบด้วยสินค้าในเซต $A \cup B$ (A หรือ B) กับการซื้อขายทั้งหมด ค่าดังกล่าวจะถูกเรียกว่า ค่าสนับสนุน (Support) ของกฎความสัมพันธ์หรือใช้สัญลักษณ์ว่า s จากนั้น เราจะนิยามสิ่งที่เรียกว่า ค่าความเชื่อมั่น (Confidence) ของกฎความสัมพันธ์ หรือใช้สัญลักษณ์ c ให้เป็นอัตราส่วนระหว่างการซื้อขายใน D ที่ประกอบด้วย สินค้าในเซต $A \cup B$ (A หรือ B) กับการซื้อขายใน D ที่ประกอบด้วย สินค้าในเซต A นั่นคือ

$$\begin{aligned} support(A \implies B) &= P(A \cup B) \\ &= \frac{\#\text{ of } t_i \text{ containing } A \cup B}{\#\text{ of } t_i} \end{aligned} \tag{4.2}$$

$$\begin{aligned} confidence(A \implies B) &= P(B|A) \\ &= \frac{support(A \cup B)}{support(A)} \\ &= \frac{\#\text{ of } t_i \text{ containing } A \cup B}{\#\text{ of } t_i \text{ containing } A} \end{aligned} \tag{4.3}$$

ในที่นี้, A ถูกเรียกว่าเหตุ (Antecedence) และ B จะถูกเรียกว่า ผล (Consequence) ของกฎดังกล่าว กฎที่มี s มากกว่าค่าสนับสนุนขั้นต่ำ (Minimum Support Threshold) และมี c มากกว่าค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence Threshold) จะถือว่าเป็นกฎที่หนักแน่น (ในภาษาอังกฤษใช้คำว่า Strong) ในขณะที่เซตของสิ่งของ $\mathcal{F} = A \cup B$ ซึ่งมีจำนวนสมาชิก k ตัวและมี s มากกว่าค่าสนับสนุนขั้นต่ำ จะถูกเรียกว่าไอเทมเซตที่พบบ่อย k ตัว (Frequent k-itemset) โดยทั่วไปแล้ว เซตของไอเทมเซตที่พบบ่อย k ตัว มักจะถูกแทนด้วยสัญลักษณ์ L_k ต่อไปเป็นตัวอย่างของการคำนวณค่าสนับสนุนและค่าความเชื่อมั่นของกฎความสัมพันธ์และของไอเทมเซต

ตัวอย่าง 4.1.1. *กำหนดให้ร้านค้าแห่งหนึ่ง ขายสินค้า 4 ชนิด และมีประวัติการซื้อขายเป็นดังตารางที่ 4.1*

จากตารางเราพบว่าเซตของสินค้าคือ $\mathcal{I} = \{$*นมเปรี้ยว, โยเกิร์ต, มันฝรั่งทอดกรอบ, น้ำอัดลม*$\}$ *สมมุติว่า เราต้องการจะคำนวณค่าสนับสนุน และค่าความเชื่อมั่นของกฎ* $A \implies B$ *โดย* $A = \{$*นมเปรี้ยว*$\}$ *และ* $B = \{$*โยเกิร์ต*$\}$ *เราจะได้ว่า*

$$support(A \implies B) = \frac{2}{5} = 0.4 = 40\%$$

$$confidence(A \implies B) = \frac{0.4}{0.4} = 1 = 100\%$$

ตาราง 4.1: ตัวอย่างของรายการขายสินค้าในร้านค้าแห่งหนึ่ง

เลขบิล	นมเปรี้ยว	โยเกิร์ต	มันฝรั่งทอดกรอบ	น้ำอัดลม
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

สมมุติว่าค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำ มีค่าเท่ากับ 40% and 60% ตามลำดับ เราจะได้ว่ากฎ นมเปรี้ยว $\Longrightarrow$ โยเกิร์ต นั้นหนักแน่น และถือว่า $\mathcal{F} = A \cup B = \{$นมเปรี้ยว, โยเกิร์ต$\}$ เป็นไอเทมเซตที่พบบ่อย

โดยทั่วไปแล้ว การทำเหมืองข้อมูลเพื่อหากฎความสัมพันธ์ สามารถมองเป็นกระบวนการซึ่งประกอบด้วย 2 ขั้นตอนย่อย ได้แก่

1. การหาไอเทมเซตที่พบบ่อยซึ่งมีค่าสนับสนุนเกินค่าสนับสนุนขั้นต่ำ
2. การหากฎความสัมพันธ์ที่หนักแน่น จากไอเทมเซตที่พบบ่อยที่พบข้างต้น ซึ่งมีค่าสนับสนุนเกินค่าสนับสนุนขั้นต่ำ และมีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นขั้นต่ำ

4.1.2 ขั้นตอนวิธีอะพริออรี

จากนี้ไป เราจะลองมาศึกษาวิธีสำหรับการค้นหาแบบรูปที่พบบ่อย ซึ่งมีชื่อเรียกว่า ขั้นตอนวิธีอะพริออรี (Apriori Algorithm) ขั้นตอนวิธีนี้อาศัยเทคนิคการค้นหาไอเทมเซตที่พบบ่อย แบบแนวกว้าง (Breadth–first–search) โดยที่ข้อมูลของไอเทมเซต k ตัว จะถูกนำมาใช้เพื่อสร้างไอเทมเซต (k+1) ตัวต่อไป อัลกอริทึม จะเริ่มจากการค้นหาไอเทมเซต 1 ตัว (1–itemset) ที่มีค่าสนับสนุนเกินกว่าค่าสนับสนุนขั้นต่ำที่ผู้ใช้ตั้งไว้ สมมุติให้เซตของไอเทมเซต 1 ตัวที่หาได้เรียกว่า L_1 จากนั้นเราจะสร้าง L_2 โดยอาศัยไอเทมเซต ที่อยู่ใน L_1 ทั้งนี้ L_2 ก็จะได้ชื่อว่าเป็นไอเทมเซต 2 ตัวที่พบบ่อย (Frequent 2–itemsets) ต่อจากนั้น L_2 ก็จะถูกใช้ ในการสร้าง L_3 และต่อไปเรื่อยๆ จนกว่าเราจะไม่เจอไอเทมเซต k ตัว ที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ

อย่างไรก็ตาม วิธีการสร้างไอเทมเซต k ตัว ที่กล่าวไปข้างต้นยังไม่มีประสิทธิภาพเท่าที่ควร เพราะในการหาไอเทมเซต (k+1) ตัวอาจมีไอเทมเซต (k+1) ตัวบางเซตที่ไม่มีทางจะมีค่าสนับสนุน มากกว่าค่าสนันสนุนขั้นต่ำได้ ทำให้เปลืองเนื้อที่ในการจัดเก็บและเวลาการคำนวณ ฉะนั้นแล้วเราจะนำคุณสมบัติที่เรียกว่า คุณสมบัติ

อะพริออรี (Apriori Property) ซึ่งเป็นคุณสมบัติเกี่ยวกับ ไอเทมเซตที่พบบ่อย มาใช้เพื่อตัดไอเทมเซตบางตัวที่ไม่สามารถเป็นไอเทมเซตที่พบบ่อยได้แน่นอน ออกไปจากการคำนวณ คุณสมบัติดังกล่าวมีความว่า

Apriori Property: *เซตย่อยที่ไม่ใช่เซตว่างทั้งหมดของไอเทมเซตที่พบบ่อย จะต้องเป็นไอเทมเซตที่พบบ่อยด้วย*

การที่ไอเทมเซตจะเป็นไอเทมเซตที่พบบ่อยได้ เซตย่อยทุกตัวของไอเทมเซตนั้นจะต้องเป็นไอเทมเซตที่พบบ่อยมาก่อน ที่เป็นเช่นนี้ เพราะว่าโดยนิยามถ้าไอเทมเซต I มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำ เราจะถือว่า I ไม่ใช่ไอเทมเซตที่พบบ่อย จากนั้นถ้าหากเราเพิ่มไอเทม A เข้าไปใน I เราก็จะพบว่าไอเทมเซต $I \cup A$ ไม่สามารถจะมีค่าสนับสนุนที่มากกว่าไอเทมเซต I ได้ เพราะเพียงลำพังแค่ I ยังไม่ใช่ ไอเทมเซตที่พบบ่อย $I \cup A$ ก็ไม่สามารถจะเป็นไอเทมเซตที่พบบ่อยได้ เมื่อนำคุณสมบัติอะพริออริมาประยุกต์ใช้หาไอเทมเซต เราจะได้ขั้นตอนวิธีที่เรียกว่าขั้นตอนวิธีอะพริออริ ดังแสดงข้างล่าง

Algorithm 2 ขั้นตอนวิธีอะพริออรี

1: Find all strong 1–itemsets
2: **while** L_{k-1} is non–empty set **do**
3: $C_k =$ apriori–gen(L_{k-1})
4: For each c in C_k, initialise c.count to zero
5: **for** records r in the database **do**
6: C_r = subset(C_k, r) ; for each c in C_r , c.count++
7: Set L_k to all c in C_k whose support is greater than minimum support
8: **end for**
9: **end while**
10: Return all of the L_k sets

ในที่นี้จะมีโปรแกรมย่อย apriori–gen ซึ่งอาศัยคุณสมบัติอะพริออริมาช่วย ในการสร้างไอเทมเซตที่ใหญ่ขึ้น ซึ่งประกอบด้วยขั้นตอนสองขั้นตอน คือ การรวม (Joining) และการตัดแต่ง (Pruning) ดังรายละเอียดต่อไปนี้

ขั้นตอนการรวม

การหาไอเทมเซต k ตัวที่อาจจะเป็น ไอเทมเซตที่พบบ่อย (แคนดิเดตไอเทมเซต k ตัว) L_k สามารถทำได้โดยการนำ L_{k-1} มาต่อรวมกับ L_{k-1} ด้วยกันเอง เราจะให้ชื่อ แคนดิเดตไอเทมเซต นี้ว่า C_k เพราะว่าไอเท

มเซตเหล่านั้น อาจจะมีค่าสนับสนุนต่ำกว่าค่าสนับสนุนขั้นต่ำ ตามธรรมเนียมแล้วเราจะสมมุติให้ไอเทมแต่ละตัวในไอเทมเซต จัดเรียงตามลำดับอักษร การจะนำไอเทมเซตสองเซตมารวมกัน $L_{k-1} \bowtie L_{k-1}$ สามารถทำได้เมื่อ ไอเทมเซตทั้งสองมีไอเทม $(k-2)$ ตัวแรกเหมือนกัน นั่นคือสมาชิก l_1 และ l_2 ของ L_{k-1} จะถูกนำมาต่อกันถ้า

$$(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \cdots \wedge (l_1[k-2] = l_2[k-2]) \\ \wedge (l_1[k-1] < l_2[k-1])$$

ข้อแม้สุดท้ายมีไว้เพื่อป้องกันการสร้างไอเทมเซตที่ซ้ำกันขึ้น ผลลัพธ์ของการนำ l_1 และ l_2 มาต่อกันก็คือ

$$l_1[1], l_1[2], \ldots, l_1[k-1], l_1[k-1], l_2[k-1]$$

ขั้นตอนการตัดแต่ง

C_k ที่ได้จากการต่อในขั้นที่แล้ว ถือว่าเป็นซูเปอร์เซตของ L_k เพราะว่าไอเทมเซตที่พบบ่อย k ตัวนั้นจะอยู่ใน C_k แน่นอน เพียงแต่ว่าอาจจะมีสมาชิกบางตัวของ C_k ที่ไม่ใช่ไอเทมเซตที่พบบ่อย ดังนั้นเราจำเป็นที่จะต้องคัดกรองสมาชิกที่มีค่าสนับสนุนต่ำกว่าค่าสนับสนุนขั้นต่ำออก เพื่อให้ได้ L_k สำหรับฐานข้อมูลขนาดเล็กที่ C_k มีขนาดไม่ใหญ่มาก เราสามารถคิดค่าสนับสนุนของสมาชิกของ C_k ได้โดยตรง แต่สำหรับฐานข้อมูลขนาดใหญ่ C_k อาจจะมีขนาดใหญ่ตามไปด้วย ด้วยเหตุนี้เราจะนำคุณสมบัติอะพริออริมาใช้เพื่อลดขนาดของ C_k ในเบื้องต้นก่อน เราพบว่าไอเทมเซต (k–1) ตัวใดๆ ที่ตัวมันเองไม่ใช่ไอเทมเซตที่พบบ่อย จะไม่สามารถเป็นเซตย่อยของไอเทมเซต k ตัวได้ ดังนั้นถ้าไอเทมเซต (k–1) ตัวใดๆ ของแคนดิเดตไอเทมเซต k ตัว ไม่ได้อยู่ใน L_{k-1} แล้วแคนดิเดตไอเทมเซตนั้นจะไม่สามารถเป็นไอเทมเซตที่พบบ่อยได้

การสร้างกฎความสัมพันธ์

หลังจากที่เราได้ไอเทมเซตที่พบบ่อยมาแล้ว เราสามารถสร้างกฎความสัมพันธ์ที่หนักแน่นขึ้นมาได้ จากไอเทมเซตเหล่านั้น จากนิยามก่อนหน้านี้ กฎเหล่านั้นจะเป็นกฎที่หนักแน่นหากเป็นกฎที่มีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นขั้นต่ำ และมีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ การสร้างกฎความสัมพันธ์มีขั้นตอนดังต่อไปนี้

- สำหรับไอเทมเซตที่พบบ่อย l, หาเซตย่อยที่ไม่ใช่เซตว่างทุกตัวของ l.

- สำหรับทุกๆเซต $s : s \subset l$ เราจะสร้างกฎในรูปแบบของ

$$s \implies (l - s) \quad if \quad \frac{support\_count(l)}{support\_count(s)} \geq min\_confidence$$

ความท้าทายของการหาไอเทมเซตที่พบบ่อยจากฐานข้อมูลขนาดใหญ่ก็คือ เรามักจะพบว่าจำนวนไอเทมเซต ที่มีค่าสนับสนุนผ่านเกณฑ์ ค่าสนับสนุนขั้นต่ำมีมากเกินไป นั่นเป็นเพราะว่าหากเซต A เป็นไอเทมเซตที่พบบ่อยแล้ว เซตย่อยของเซต A ทั้งหมดก็ต้องถือว่าเป็นไอเทมเซตที่พบบ่อยด้วย เพราะว่าเซตย่อยเหล่านั้นต้องมีค่าสนับสนุนที่เท่ากับ A ดังนั้นหากเราพบไอเทมเซตที่พบบ่อยที่มีขนาด 100 เท่ากับว่าเราอาจจะต้องเก็บไอเทมเซตที่พบบ่อยทั้งหมด กว่า 2^{100} เซตไว้เพื่อใช้ในการคำนวณต่อ ซึ่งถือว่าเยอะมากและอาจจะเกินหน่วยความจำที่มีอยู่ได้ เพื่อที่จะลดปัญหานี้ เราจะมาศึกษาประเภทไอเทมเซตที่พบบ่อยเพิ่มเติมสองแบบ ได้แก่ ไอเทมเซตที่พบบ่อยแบบปิด (Closed Frequent Itemset) และไอเทมเซตที่พบบ่อยใหญ่สุด (Maximal Frequent Itemset)

เราจะถือว่าไอเทมเซต A เป็นไอเทมเซตแบบปิด ในชุดข้อมูล S ถ้าไม่มีซูเปอร์เซต (Superset) B ซึ่งมีค่าสนับสนุนเท่ากับ A หากกลับไปดูที่ตัวอย่างข้างต้น $A = \{\text{นมเปรี้ยว}\}$ ไม่ถือว่าเป็นไอเทมเซตแบบปิด เพราะว่ายังมีเซต $B = \{\text{นมเปรี้ยว,โยเกิร์ต}\}$ ซึ่งเป็นซูเปอร์เซตของ A และ ยังมีค่าสนับสนุนเท่ากับเซต A ด้วย อย่างไรก็ตาม B ถือว่าเป็น ไอเทมเซตแบบปิด เนื่องจากไม่มีซูเปอร์เซตที่มีค่าสนับสนุนเท่ากับ B อีกแล้วในชุดข้อมูล S หากไอเทมเซต A เป็นไอเทมเซตแบบปิด และมีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ (หรือเรียกว่ามีคุณสมบัติ พบบ่อย) จะถือว่า เซต A เป็นไอเทมเซตที่พบบ่อยแบบปิด

เราจะถือว่าไอเทมเซต A เป็น ไอเทมเซตที่พบบ่อยใหญ่สุด บนชุดข้อมูล S ถ้า A มีคุณสมบัติพบบ่อย (Frequent) และไม่มีซูเปอร์เซต B ที่ $A \subset B$ และ B ก็มีคุณสมบัติพบบ่อย ซึ่งคล้ายกับนิยามของไอเทมเซตที่พบบ่อยแบบปิด แต่ต่างกันตรงที่ว่า B ไม่จำเป็นจะต้องมีค่าสนับสนุนเท่ากับ A

ดังนั้นแล้วในกรณีที่ไอเทมเซตที่พบบ่อยในระบบมีเพิ่มมากขึ้นเรื่อยๆ เราอาจจะพิจารณาคัดกรอง ไอเทมเซตที่พบบ่อยบางตัวทิ้งไป และเลือกเก็บเฉพาะไอเทมเซตที่พบบ่อยแบบปิด หรือไอเทมเซตที่พบบ่อยใหญ่สุดเท่านั้นไว้เพื่อนำไปสร้างกฎความสัมพันธ์ต่อไป

4.2 ระบบแนะนำ

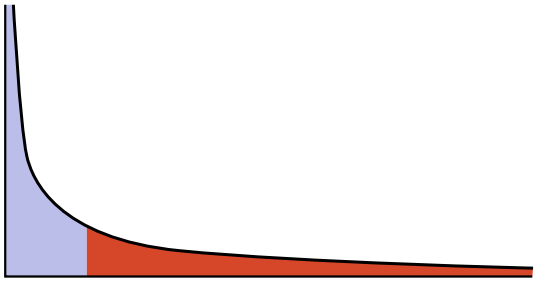
การทำเหมืองข้อมูลถูกนำมาประยุกต์ใช้กับข้อมูลทางธุรกิจได้อีกแบบหนึ่ง คือการสร้างระบบแนะนำ ระบบแนะนำมีเป้าหมายเพื่อคาดเดาผลตอบรับของผู้ใช้ต่อทางเลือกต่างๆที่มี เช่น ในระบบร้านค้าออนไลน์ เราต้องการจะทำนายว่าลูกค้าแต่ละคนจะตอบรับกับสินค้าต่างๆอย่างไรบ้าง จากนั้นเราสามารถ เลือกสินค้าที่

ลูกค้าสนใจมากที่สุดมาแนะนำให้กับลูกค้า แน่นอนว่าการทำนายต้องอยู่บนพื้นฐานของประวัติการซื้อของหรือ การเข้าชมของของลูกค้า อีกตัวอย่างหนึ่ง อาจเป็นการแนะนำบทความที่น่าสนใจ จากจำนวนบทความเยอะๆ ให้กับผู้อ่าน ร้านค้าออนไลน์ ที่นำระบบแนะนำมาใช้ที่เราอาจจะรู้จัก ได้แก่ Netflix ซึ่งเป็นบริษัทให้บริการวิดีโอออนไลน์ ได้ประยุกต์ใช้ระบบแนะนำ ในการแนะนำหนังหรือสารคดีที่ผู้ใช้อาจจะชอบให้กับผู้ใช้ ซึ่งถือว่าเป็นประโยชน์มาก แทนที่จะต้องให้ ผู้ใช้ต้องเลือกหนังเอง จากวิดีโอที่มีอยู่ในฐานข้อมูลเกือบ 1 ล้านเรื่อง Netflix ให้ความสำคัญกับระบบแนะนำของตนมาก ถึงขนาดประกาศการแข่งขันพัฒนาระบบแนะนำ ที่มีประสิทธิภาพสูง โดยผู้ชนะได้ประกาศผลออกมาแล้ว ว่าคือ ทีม BellKor's Pragmatic Chaos จากบริษัท AT&T โดยทีมดังกล่าวคว้าเงินรางวัล 1 ล้านเหรียญสหรัฐไปครอง อีกบริษัทหนึ่งได้แก่ Amazon เวบไซต์ขายของออนไลน์ขนาดใหญ่แห่งหนึ่ง ก็ได้นำระบบแนะนำเข้ามาประยุกต์ใช้ เพื่อแนะนำสินค้าที่มีจำนวนมากให้กับผู้ซื้อ โดยวิเคราะห์จากสถิติการซื้อและการเข้าชมสินค้าของผู้ใช้

จะเห็นได้ว่าระบบแนะนำถือว่ามีประโยชน์มากสำหรับบริษัทที่ทำธุรกิจในทำนองนี้ แน่นอนว่าเรายังสามารถนำระบบ ไปปรับใช้เพื่อแนะนำ อาหาร สถานที่ท่องเที่ยว วิชาที่เลือกเรียน ได้อีก

โดยทั่วไประบบแนะนำแบ่งออกเป็นสองแนวทางคือ หนึ่ง การแนะนำบนพื้นฐานของเนื้อหา (Content-based System) ที่มุ่งแนะนำสินค้าหรือผลิตภัณฑ์ที่ มีคุณสมบัติตรงกับสิ่งที่ผู้ใช้ชอบ เช่นการแนะนำหนังตลกให้กับผู้ใช้ที่ดูหนังที่มีเนื้อหาสนุกสนานบ่อยๆ อีกแนวทางหนึ่งก็คือ การคัดกรองร่วมกัน (Collaborative Filtering) แนวทางนี้ไม่ได้แนะนำสิ่งที่ผู้ใช้อาจจะชอบโดยดูจากคุณสมบัติของสินค้า แต่จะแนะนำสินค้าโดยมีพื้นฐานจากข้อมูลของผู้ใช้คนอื่นที่มีลักษณะเหมือนกับผู้ใช้คนนี้

ระบบแนะนำถือว่าเป็นส่วนสำคัญที่ทำให้ร้านค้าออนไลน์ได้เปรียบร้านค้าแบบมีหน้าร้านจริงๆ ทั้งนี้เนื่องมาจากปรากฏการณ์ที่มีชื่อว่า ปรากฏการณ์หางยาว (Long-tail Phenomenon) จากภาพ 4.1 จะเห็นได้ว่า ร้าน



รูปภาพ 4.1: ปรากฏการหางยาว

ค้าแบบมีหน้าร้านมีข้อจำกัดด้านกายภาพ ทำให้ไม่สามารถนำสินค้าทุกอย่าง มาจัดแสดงได้ โดยส่วนใหญ่ก็ทางร้านก็จะเลือกสินค้าที่ได้รับความนิยมอันดับต้นๆเท่านั้น (จากเส้นแบ่งไปทางซ้าย) โดยสินค้าที่อยู่ด้านขวาของเส้นซึ่งมีอีกหลายชนิด อาจจะกลายเป็นสินค้าขายดีได้ หากได้รับการโปรโมท แต่ร้านค้าแบบมีหน้าร้าน ไม่

ผู้ใช้	ปอบ 1	ปอบ 2	บุญชู 1	บุญชู 2	บุญชู 3	ฟรีแลนซ์	อัพ
เจ๋	3			5	1		
สุชาติ	1		5			2	
ยุ่น	2			1	1		
อิม	4	1			3		

ตาราง 4.2: ตารางแสดงยูทิลิตีเมทริกซ์ของผู้ใช้ 4 คนต่อไอเทม 7 ชิ้น

สามารถทำเช่นนั้นได้ ผิดกับร้านค้าออนไลน์ซึ่งสามารถแสดงสินค้าได้ทุกชนิดที่มีอยู่ในคลังสินค้า ซึ่งแน่นอนจะต้องมีวิธีในการเลือกสินค้ามาแสดงให้เหมาะสม ในการนี้ ระบบแนะนำจึงเข้ามามีบทบาทในการเลือกสินค้าที่อาจจะยังไม่เป็นที่นิยมสำหรับคนส่วนใหญ่ แต่อาจเป็นที่นิยมของคนบางกลุ่ม มาแสดงบนหน้าเวปได้ และภายหลังอาจพบว่าสินค้าดังกล่าว อาจกลายมาเป็นสินค้ายอดนิยมได้

4.2.1 โครงสร้างข้อมูล

ในการทำระบบแนะนำมีหัวใจหลักอยู่ที่การสร้างเมทริกซ์ที่เก็บความสัมพันธ์ระหว่างผู้ใช้และสินค้า ซึ่งเรียกกันว่า ยูทิลิตีเมทริกซ์ (Utility Matrix) ค่าในเมทริกซ์ดังกล่าวบ่งบอกถึงระดับความชอบของผู้ใช้ต่อสินค้านั้นๆ โดยทั่วไปแล้ว ยูทิลิตีเมทริกซ์เป็นเมทริกซ์ ที่ค่าส่วนใหญ่จะยังว่างอยู่ ซึ่งนั่นแสดงว่า ผู้ใช้เคยแสดงความชอบต่อสินค้าเพียงแค่ส่วนหนึ่งจากสินค้าทั้งหมด ในที่นี้ระบบแนะนำที่มีหน้าที่ที่จะทำนายความชอบของผู้ใช่ต่อส่วนที่ยังว่างอยู่ ให้ใกล้เคียงกับความชอบของผู้ใช้มากที่สุด ตัวอย่างยูทิลิตีเมทริกซ์ที่ใช้เก็บความชอบของผู้ใช้ 4 คนโดยมีสเกลจาก 1–5 ต่อภาพยนตร์เจ็ดเรื่องแสดงไว้ในตารางที่ 4.2

ถึงแม้ว่าเป้าหมายของระบบแนะนำคือการทำนายค่าที่ควรจะเป็นของช่องที่ว่างอยู่ ซึ่งก็คือช่องที่ผู้ใช้ยังไม่ได้ให้คะแนน แต่หากมีช่องว่างมากเกินไป การทำนายก็อาจจะไม่แม่นยำเท่าที่ควร เพื่อลดปัญหาในจุดนี้ ยูทิลิตีเมทริกซ์จึงไม่ควรจะว่างมากเกินไป ซึ่งนั่นหมายความว่า เราต้องหาทางเก็บรวบรวมคะแนนจากผู้ใช้ให้ได้มากที่สุด วิธีส่วนใหญ่ที่นิยมใช้กันคือการชักชวนให้ผู้ใช้ให้คะแนนสินค้า หลังจากที่ผู้ใช้ได้ซื้อหรือเลือกสินค้าไปใช้แล้ว วิธีนี้ถือว่าเป็นวิธีที่จะทำให้ได้คะแนนแบบแม่นยำที่สุด แต่ปัญหาคือผู้ใช้ส่วนใหญ่มักจะไม่ให้ความร่วมมือ และไม่ทำแบบสอบถาม วิธีที่สอง ผู้พัฒนาอาจจะเลือกการอนุมานโดยใช้สถานการณ์รอบข้าง เช่นถึงแม้ว่าผู้ใช้จะไม่ได้ตัดสินใจซื้อสินค้า แต่หากเข้ามายังหน้าสินค้าบ่อยๆ และใช้เวลาในหน้านั้นนานๆ เราอาจจะอนุมานได้ว่าผู้ใช้ชอบของชิ้นนั้นก็ได้ เป็นที่น่าสังเกตว่า หากเราเลือกวิธีในการเติมยูทิลิตีเมทริกซ์แบบแรก เราสามารถได้คะแนนความชอบเป็นลำดับจาก 1 ไปถึง K แต่หาก เราใช้วิธีหลัง เราจะไม่สามารถให้คะแนนละเอียดแบบนั้นได้ เราจะให้อย่างมากได้แค่ 0 หรือ 1 ซึ่งแทนความหมายว่า พอใจกับสินค้าหรือไม่พอใจกับ

สินค้าเท่านั้น

4.2.2 การแนะนำบนพื้นฐานของเนื้อหา

ในส่วนนี้เราจะมาศึกษาการทำงานของ ระบบแนะนำที่อาศัยคุณลักษณะของสิ่งของมาช่วยในการแนะนำ หรือที่เรียกว่าการแนะนำบนพื้นฐานของเนื้อหา โครงสร้างข้อมูลที่จำเป็นสำหรับแนวทางนี้นอกเหนือจากยูทิลิตีเมทริกซ์ ก็คือ ตารางที่ใช้เก็บว่าสินค้าหรือสิ่งของที่อยู่ในร้านมีคุณลักษณะอย่างไรบ้าง ซึ่งตารางนี้เราจะเรียกว่า ไอเทมโปรไฟล์ (Item Profile) และตารางที่เอาไว้เก็บความชอบของผู้ใช้ (User) ต่อคุณลักษณะต่างๆข้างต้น ซึ่งเราจะเรียกว่า ยูเซอร์โปรไฟล์ (User Profile) ขั้นตอนการแนะนำจะเริ่มโดยการหาสินค้าที่มีคุณลักษณะคล้ายกับความชอบของผู้ใช้ ในที่นี้คือสินค้าที่มีคุณลักษณะตรงกับคุณลักษณะที่ผู้ใช้ชอบ และต้องเป็นสินค้าที่ผู้ใช้ยังไม่เคยซื้อหรือใช้บริการ มาทำการแนะนำให้กับลูกค้า การวัดความคล้ายในที่นี้ สามารถใช้ความคล้ายแบบโคไซน์ หรือระยะห่างมินโควสกี้ ที่เคยกล่าวไว้ในบทก่อนหน้านี้มาใช้ได้

ไอเทมโปรไฟล์

ในการสร้างไอเทมโปรไฟล์ ผู้พัฒนาระบบแนะนำจะต้องเป็นคนกำหนดเซตของคุณลักษณะไว้ล่วงหน้า และเมื่อทำการ เพิ่มสินค้าเข้าไปในฐานข้อมูล จะต้องมีการระบุคุณลักษณะของสินค้าใหม่นั้นด้วย ส่วนของไอเทมโปรไฟล์ มักจะมีการเปลี่ยนแปลงไม่บ่อย ตัวอย่างเช่น สินค้าประเภทซีดีภาพยนตร์ สามารถแสดงได้ด้วยคุณลักษณะต่างๆดังต่อไปนี้

- นักแสดงนำ
- ผู้กำกับ
- ปีที่ฉาย
- ประเภทของหนัง

คุณลักษณะเหล่านี้เป็นคุณลักษณะที่มาพร้อมกับภาพยนตร์อยู่แล้ว เราสามารถหาข้อมูลมาใช้ได้เลย แต่ถ้าหากเป็นสินค้าที่ไม่มีคุณลักษณะสากลที่ ใช้อธิบายสินค้าเหล่านั้นตั้งแต่ต้น เช่นสินค้าประเภท รูปภาพ บทความ หรือสถานที่ท่องเที่ยว เราก็สามารถกำหนดคุณลักษณะขึ้นมาเอง ให้เหมาะสมได้ ตัวอย่างของ ไอเทมโปรไฟล์ ของสินค้าที่เป็นซีดีภาพยนตร์สามารถแสดงได้ดังตาราง 4.3

หนัง / คุณลักษณะ (นักแสดง)	S.Johansson	C.Evans	R.Downey
Ant man		1	
Avenger	1	1	1
Iron man	1		1

ตาราง 4.3: ไอเทมโปรไฟล์สำหรับซีดีภาพยนตร์

ยูเซอร์โปรไฟล์

ส่วนสำคัญอีกอย่างสำหรับการแนะนำบนพื้นฐานของเนื้อหาก็คือ ยูเซอร์โปรไฟล์ ซึ่งเป็นตารางที่ประกอบด้วยคุณลักษณะชุดเดียวกับที่อยู่ในไอเทมโปรไฟล์ แต่ในที่นี้จะเอาไว้เก็บความชอบที่ผู้ใช้มีต่อคุณลักษณะนั้นๆ โดยปกติแล้วข้อมูลในส่วนของยูเซอร์โปรไฟล์ สามารถประมาณค่าโดยอาศัยไอเทมโปรไฟล์ และยูทิลิตีเมทริกซ์ หากมีการเปลี่ยนแปลงในยูทิลิตีเมทริกซ์ นั่นคือเมื่อผู้ใช้เข้ามาให้คะแนนไอเทมตัวใหม่ ซึ่งแปลว่าเราจะต้องทำการคำนวณค่าของยูเซอร์โปรไฟล์ใหม่ด้วย เพราะว่าความชอบต่อคุณลักษณะบางตัวอาจจะเปลี่ยนไป ตามคะแนนที่ผู้ใช้ได้ให้มาใหม่ ตัวอย่างของยูเซอร์โปรไฟล์แสดงไว้ในตาราง 4.4

ผู้ใช้ / คุณลักษณะ (นักแสดง)	S.Johansson	C.Evans	R.Downey
เจ๋		1	
สุชาติ	1	1	1
ยุ่น	1		1
อิม	1		1

ตาราง 4.4: ยูเซอร์โปรไฟล์สำหรับผู้ใช้ 4 คน

โดยภาพรวมแล้วการแนะนำบนพื้นฐานของเนื้อหามีขั้นตอนดังนี้

1. สร้างไอเทมโปรไฟล์ และยูทิลิตีเมทริกซ์
2. เชิญชวนให้ผู้ใช้ ใส่ค่าความชอบหรือความพึงพอใจในยูทิลิตีเมทริกซ์ หลังจากใช้สินค้าไปแล้ว
3. เมื่อมีการเปลี่ยนแปลงของยูทิลิตีเมทริกซ์ของผู้ใช้คนใด ให้ทำการปรับปรุงค่ายูเซอร์โปรไฟล์ของผู้ใช้นั้นด้วย

โดยหากเรทติ้ง (Rating) ถูกเก็บไว้เป็นแบบทวิภาค ความชอบต่อคุณลักษณะ j ของ ผู้ใช้ i คำนวณได้โดย

$$p_{i,j} = \sum_{S_i} (\delta(j \in k))/|S_i| \tag{4.4}$$

ในที่นี้ S_i คือเซตของไอเทมที่ผู้ใช้ i เคยโหวตให้คะแนน และ $\delta(j \in k)$ มีค่าเท่ากับ 1 เมื่อ ไอเทมชิ้นที่ k มีคุณลักษณะ j และเท่ากับ 0 หากไอเทม k ไม่มีคุณลักษณะนั้น

หากเรทติ้งถูกเก็บไว้เป็นแบบจำนวนเต็ม ความชอบของผู้ใช้ที่ i ต่อคุณลักษณะ j จะคำนวณได้โดย

$$p_{i,j} = \sum_{c \in C_j} (v_c - \bar{v}_i)/|C_j| \tag{4.5}$$

โดย C_j คือเซตของสินค้าที่ผู้ใช้ i เคยโหวตและมีคุณลักษณะ j เป็นหนึ่งในคุณลักษณะของสินค้า ส่วน $\bar{v}_i$ แสดงค่าเฉลี่ยของเรทติ้งที่ผู้ใช้ i เคยให้สินค้าทั้งหมด ในที่นี้ $|C_j|$ แสดงจำนวนสมาชิกของเซต C_j

4. เมื่อยูเซอร์โปรไฟล์เปลี่ยน ระบบจะแนะนำไอเทมที่มีไอเทมโปรไฟล์ใกล้เคียงกับยูเซอร์โปรไฟล์มากที่สุดให้กับผู้ใช้

ตัวอย่าง 4.2.1. *บริษัทเช่าวิดีโอต้องการจะแนะนำหนังให้กับผู้ใช้ โดยหนังถูกนำเสนอโดยคุณลักษณะซึ่งคือนักแสดงที่แสดงหนังเรื่องนั้น เราพบว่า ไอเทมโปรไฟล์ ของหนังต่างๆสามารถแสดงได้ดังตารางต่อไปนี้ และ*

หนัง / คุณลักษณะ	S.Johansson	C.Evans	R.Downey
Ant man		1	
Avenger	1	1	1
Iron man	1		1

ตาราง 4.5: ไอเทมโปรไฟล์

เช่นกันยูทิลิตีเมทริกซ์ สามารถแสดงได้เป็น สมมุติว่าเจ๋ได้เข้ามาให้เรทติ้งกับ Iron man ว่าตนชอบหนังเรื่องดังกล่าว (ในกรณีนี้ ยูทิลิตีเมทริกซ์เก็บค่าแบบทวิภาคเท่านั้นคือ ชอบหรือไม่ชอบ) คำถามคือ ระบบควรจะแนะนำหนังใหม่เรื่องไหนให้กับเจ๋ ขั้นตอนจะเริ่มจากการอัพเดทค่ายูเซอร์โปรไฟล์ของเจ๋ ซึ่งเดิมมีค่าว่างเปล่า

ผู้ใช้ / หนัง	Ant man	Avenger	Iron man
สุชาติ	1		
เจ๋			1
ยุ่น	1		1

ตาราง 4.6: ยูทิลิตีเมทริกซ์แบบทวิภาค

ผู้ใช้ / คุณลักษณะ	S.Johansson	C.Evans	R.Downey
เจ๋	0	0	0

ตาราง 4.7: ยูเซอร์โปรไฟล์ของเจ๋ ก่อนการประมวลผล

ผู้ใช้ / คุณลักษณะ	S.Johansson	C.Evans	R.Downey
เจ๋	1	0	1

ตาราง 4.8: ยูเซอร์โปรไฟล์ของเจ๋ หลังการประมวลผล

โดยใช้หลักการคำนวณว่าความชอบต่อคุณลักษณะ j *ของผู้ใช้* i *คำนวณได้โดย* $p_{i,j} = \sum_{S_i}(\delta(j \in k))/|S_i|$ *เมื่อคำนวณโดยสูตรข้างต้นเราจะพบว่า ยูเซอร์โปรไฟล์ของเจ๋มีค่าเท่ากับ ซึ่งหมายความว่าเจ๋อาจจะชอบ S.Johansson และ R.Downey ซึ่งเป็นนักแสดงในหนังเรื่อง Iron man ซึ่งเจ๋ชอบดู ต่อไปเราจะใช้ยูเซอร์โปรไฟล์ของเจ๋ เพื่อหาหนัง (ที่เจ๋ยังไม่ได้ดู) ที่มี โปรไฟล์ใกล้เคียงกับ โปรไฟล์ของเจ๋มากที่สุด จากการคำนวณความเหมือนโดยใช้ความคล้ายแบบโคไซน์ เราพบว่า* $S_{\text{เจ๋, Avenger}} = (1\times1)+(1\times0)+(1\times1) = 2$ $S_{\text{เจ๋, Ant man}} = (1 \times 0) + (0 \times 1) + (1 \times 0) = 0$ *ฉะนั้นแล้ว จากหนังสองเรื่องที่เจ๋ยังไม่ได้ดู เราควรจะแนะนำให้เจ๋ดู Avenger เนื่องจากโปรไฟล์ของหนังใกล้เคียงกับ ความชอบของเจ๋ นั่นคือใกล้เคียงกับโปรไฟล์ของเจ๋นั่นเอง หากเรามีข้อมูลใหม่ ในที่นี้หมายถึงหากเจ๋เข้ามาให้เรทติ้งของหนังเรื่องใหม่ เราก็จะทำการปรับปรุงค่ายูเซอร์โปรไฟล์ของเจ๋อีกรอบหนึ่ง และทำการแนะนำหนังเรื่องใหม่ให้เจ๋ต่อไป*

จากตัวอย่างจะสังเกตได้ว่าการให้เรทติ้งยังเป็นแบบทวิภาค นั่นคือค่าในยูทิลิตีเมทริกซ์จะมีค่าเท่ากับ 0 เมื่อผู้ใช้ไม่ชอบสินค้านั้น และมีค่าเท่ากับ 1 เมื่อผู้ใช้ชอบสินค้านั้น แต่ในความเป็นจริงแล้วเราอาจจะอยากปรับเรทติ้งให้ละเอียดมากขึ้น นั่นคือผู้ใช้สามารถระบุความชอบ ได้ละเอียดเป็นตัวเลขตั้งแต่ (ยกตัวอย่าง) 1 ถึง 5 เราจะได้ ยูทิลิตีเมทริกซ์ ที่หน้าตาเปลี่ยนไปเป็น ในกรณีนี้การคำนวณความชอบของผู้ใช้ที่ i ต่อคุณลักษณะ j ต้องปรับเล็กน้อยเป็น $p_{i,j} = \sum_{c\in C_j}(v_c - \bar{v}_i)/|C_j|$

ผู้ใช้ / หนัง	Ant man	Avenger	Iron man
สุชาติ	3		
เจ๋			2
ยุ่น	5		5

ตาราง 4.9: ยูเซอร์โปรไฟล์แบบเก็บเป็นคะแนน (Score-based)

4.2.3 การคัดกรองร่วมกัน

อีกวิธีการหนึ่งสำหรับการทำระบบแนะนำ คือการใช้ข้อมูลของผู้ใช้ที่คล้ายๆกันมาเป็น ปัจจัยในการแนะนำสินค้าวิธีนี้เราจะไม่ใช้คุณลักษณะของสินค้ามาเป็นตัวตัดสิน แต่จะใช้ความเหมือนของผู้ใช้แทนวิธีนี้รู้จักกันในชื่อ การคัดกรองร่วมกัน (Collaborative Filtering) ขั้นตอนของการคัดกรองร่วมกัน มีสองขั้นตอนคือ

- หากลุ่มของผู้ใช้ที่ใกล้เคียงกับผู้ใช้ที่ต้องการแนะนำสินค้าให้
- จากนั้นแนะนำสินค้าที่ผู้ใช้คนนี้ยังไม่เคยซื้อแต่กลุ่มของผู้ใช้ที่คล้ายกันได้เคยซื้อไปใช้แล้ว

ในที่นี้เราจะมาศึกษาขั้นตอนวิธีพื้นฐานสำหรับการคัดกรองร่วมกัน กำหนดให้ $v_{i,j}$ แทนเรทติ้งของผู้ใช้ i ต่อไอเทม j และเซต I_i แทนเซตของไอเทมที่ผู้ใช้ i เคยให้คะแนน เราจะสามารถหาคะแนนเฉลี่ยของการให้คะแนนของผู้ใช้ i ได้โดย

$$\bar{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j} \tag{4.6}$$

จากนั้นเราจะสามารถทำนายว่า ผู้ใช้ที่เราต้องการจะแนะนำสินค้าให้ ในที่นี้จะเรียกว่าเป็น 'active user" a จะให้คะแนนสินค้าชิ้นที่ j เป็น

$$p_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^{n} w(a,i)(v_{i,j} - \bar{v}_i) \tag{4.7}$$

ในที่นี้ κ คือตัวปรับช่วงให้ $p_{a,j}$ มีค่าในช่วงของคะแนนโหวตที่เป็นไปได้ และ $w(a,i)$ คือน้ำหนักที่ได้จากการวัดความคล้ายระหว่าง ผู้ใช้ a กับผู้ใช้คนที่ i

สิ่งที่เราต้องหาต่อไปก็คือ เราจะวัดความคล้ายระหว่างผู้ใช้สองคนได้อย่างไรบ้าง จากบทก่อนหน้า เราพบว่าการวัดระยะทางหรือความคล้ายสำหรับข้อมูลเชิงตัวเลขมีได้หลายวิธี ในบทนี้ เราจะลองศึกษาการวัดความคล้ายบางตัวที่มีความเหมาะสมกับข้อมูลที่มีลักษณะเป็นเซตเพิ่มเติม ได้แก่

- วิธีเพื่อนบ้านใกล้เคียงบนพื้นฐานของระยะห่างแบบมินคอว์สกี้ (Minkowski distance–based Nearest Neighbour)
- สัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's Correlation Coefficient)
- ระยะห่างแบบแจคการ์ด (Jaccard Distance)
- ระยะห่างแบบโคไซน์ (Cosine Distance)

เราจะลองมาดูการใช้มาตรวัดเหล่านั้นในการวัดความคล้ายระหว่างผู้ใช้สองคนกัน

วิธีเพื่อนบ้านใกล้เคียง

การวัดความคล้ายโดยใช้วิธีเพื่อนบ้านใกล้เคียง คือการกำหนดให้ค่าน้ำหนักเป็นดิสครีตซึ่งนิยามได้โดย

$$w(a,i) = \begin{cases} 1, & \text{if } i \in \text{neighbour(a)} \\ 0, & \text{otherwise} \end{cases}$$

ในที่นี้ neighbour(a) คือเซตของผู้ใช้ k คนที่มีระยะทางห่างจากผู้ใช้ a น้อยที่สุด ซึ่งวัดโดยระยะห่างแบบมินโควสกี้ $d(u_i, u_j) = (|u_{i1} - u_{j1}|^h + \cdots + |u_{iM} - u_{jM}|^h)^{\frac{1}{h}}$

หากไม่ต้องการให้น้ำหนักมีค่าเป็นดิสครีต ก็สามารถใช้ผลลัพธ์ของระยะห่างแบบมินโควสกี้มาใช้โดยตรง ซึ่งจะได้ว่า

$$w(a,i) = \begin{cases} d(u_a, u_i), & \text{if } i \in \text{neighbour(a)} \\ 0, & \text{otherwise} \end{cases}$$

สัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน

วิธีต่อไปสำหรับใช้หาน้ำหนัก ก็คือการใช้สัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน ซึ่งนิยามไว้โดย

$$w(i,j) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}}$$

การตีความของการวัดแบบนี้คือ หากค่าสหสัมพันธ์เป็นไปในทางบวก จะเป็นการบอกว่าผู้ใช้สองคนมักจะให้เรทติ้งคล้ายๆกัน ซึ่งก็แปลได้ว่า ผู้ใช้สองคนนั้นอาจจะมีความชอบที่เหมือนกัน แต่ถ้าหากค่า สหสัมพันธ์เป็น

ไปในทิศทางลบ เราอาจจะบอกได้ว่า ผู้ใช้สองคนมักให้ค่าที่มักจะขัดแย้งกันเสมอ ค่าสหสัมพันธ์ในทิศทางลบอาจจะนำมาใช้ได้ หากยูทิลิตีเมทริกซ์เก็บเรทติ้งแบบทวิภาค คือ ชอบหรือไม่ชอบ ในกรณีดังกล่าวหากพบสหสัมพันธ์แบบลบ ระบบก็จะแนะนำสิ่งที่ตรงกันข้ามให้กับผู้ใช้แทน ทว่าหากยูทิลิตีเมทริกซ์เก็บเรทติ้งแบบเป็นค่าที่กว้างกว่านั้น การนำสหสัมพันธ์ทางลบมาใช้อาจจะไม่ง่ายนัก โดยสรุปคือ สหสัมพันธ์ที่มีความเป็นบวกหรือลบมากๆ สามารถนำมาเป็นข้อมูลในการแนะนำได้ ผิดจากสหสัมพันธ์ที่มีค่าต่ำๆ ใกล้ 0 จะไม่ค่อยบอกข้อมูลอะไร ทำให้ไม่มีประโยชน์ต่อการนำมาใช้งาน

คำถามต่อมาคือว่า ทำไมเราถึงจำเป็นจะต้องใช้สหสัมพันธ์ทางลบ ในเมื่อการใช้สหสัมพันธ์ทางบวกง่ายและสื่อความหมายมากกว่า คำตอบสำหรับคำถามนี้เป็นไปได้กรณีหนึ่งก็คือ ในกรณีที่เรามีข้อมูลไม่เพียงพอ นั่นคือไม่สามารถหาผู้ใช้ที่มีสหสัมพันธ์ทางบวกกับผู้ใช้คนนี้ได้ แทนที่จะใช้ค่าสหสัมพันธ์ที่เป็นบวกแต่มีค่าน้อยๆใกล้ศูนย์ เราอาจจะพิจารณาใช้ค่าสหสัมพันธ์ทางลบที่มีค่าสูงมาใช้แทน และอาจจะให้ผลการทำนายเรทติ้งที่ดีกว่า

ระยะห่างแบบแจคการ์ด

การวัดระยะห่างแบบแจคการ์ด ถือว่าเป็นการวัดความคล้ายของเซตจำกัด (Finite Set) สองเซต ระยะห่างแบบแจคการ์ดระหว่างเซต S และเซต T นิยามไว้ว่า

$$1 - |S \cap T|/|S \cup T| \tag{4.8}$$

นั่นคือ หนึ่งลบด้วยอัตราส่วนของจำนวนสมาชิกที่เหมือนกันของ S และ T ต่อจำนวนสมาชิกรวม (ไม่นับตัวซ้ำ) ของเซต S และ T ในทางคณิตศาสตร์ตัวเศษคือ อินเทอร์เซกชัน (Intersection) ของเซตทั้งสอง ส่วนตัวส่วนก็คือยูเนี่ยน (Union) ของเซตทั้งสองนั่นเอง

ตัวอย่าง 4.2.2. *กำหนดเซต S = {dog, cat, parrot, monkey} และเซต T = {dog, monkey, snake} เราพบว่าความคล้ายแบบแจคการ์ดระหว่าง S และ T มีค่าเท่ากับ* $SIM_{Jaccard}(S,T) = 2/5 = 0.4$ *ในขณะเดียวกันระยะห่างแบบแจคการ์ดก็จะมีค่าเท่ากับ 1 -* $SIM_{Jaccard}$ *ตามลำดับ*

การนำความคล้ายแบบแจคการ์ดไปปรับใช้กับระบบแนะนำก็คือ ให้น้ำหนักระหว่างผู้ใช้สองคนมีค่าเท่ากับระยะห่างแบบแจคการ์ดของทั้งคู่ แต่ก่อนจะวัดได้ เราจำเป็นจะต้องแปลงยูทิลิตีเมทริกซ์ที่ไม่ได้อยู่ในรูปทวิภาคมาเป็นทวิภาคก่อน หลังจากนั้นเราถึงจะสามารถวัดระยะห่างแบบแจคการ์ดได้

ระยะห่างแบบโคไซน์

วิธีสุดท้าย คือการวัดระยะทางหรือความเหมือนในเชิงมุมที่กระทำกันของเรทติ้งเวคเตอร์ (Rating Vector) ของผู้ใช้สองคน เรทติ้งเวคเตอร์ในที่นี้หมายถึงแถวหนึ่งแถวของยูทิลิตีเมทริกซ์นั่นเอง มุมดังกล่าวสามารถวัดได้

โดย

$$SIM_{cos}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{4.9}$$

โดยที่ A คือ เรทติ้งเวคเตอร์ของผู้ใช้ A และ B คือ เรทติ้งเวคเตอร์ของผู้ใช้ B สังเกตว่าการจะหามุมระหว่างเวคเตอร์สองตัว เวคเตอร์ทั้งสองต้องมีจำนวนมิติเท่ากัน หากเรานำหลักการวัดนี้มาใช้กับเรทติ้งเวคเตอร์ เราก็ต้องมั่นใจก่อนว่า เรทติ้งเวคเตอร์มีจำนวนมิติเท่ากันด้วย ในกรณีนำระยะห่างแบบโคไซน์ไปใช้คำนวณความเหมือนของเวคเตอร์จากยูทิลิตีเมทริกซ์ อาจจะมีปัญหาคือ ช่องที่ผู้ใช้ยังไม่ได้ให้เรทติ้งจะมีค่าว่างไว้ ทำให้แถวสองแถวใดๆของยูทิลิตีเมทริกซ์ อาจมีจำนวนมิติไม่เท่ากัน การแก้ไขคืออาจจะกำหนดค่าว่างให้เป็น 0 เพื่อบอกว่า ผู้ใช้ยังไม่เคยโหวต ด้วยวิธีนี้จะทำให้จำนวนมิติของเรทติ้งเวคเตอร์ ใดๆ มีค่าเท่ากัน และทำให้ สามารถคำนวณระยะห่างด้วยระยะห่างแบบโคไซน์ได้

การปัดข้อมูล

นอกเหนือจากการเรียกใช้ค่าที่อยู่ในยูทิลิตีเมทริกซ์ตามที่เก็บไว้แล้ว เราอาจต้องการดิสครีไทซ์ (Discretise) ข้อมูล เพื่อให้ความหลากหลายของข้อมูลลดลง และเพื่อเพิ่มความเร็วในการคำนวณความคล้าย หรือแปลงเพื่อให้สามารถคำนวณความคล้ายบางอย่างได้ (เช่น ระยะห่างแบบแจคการ์ด) เพื่อการนี้ ยกตัวอย่างยูทิลิตีเมทริกซ์ที่เก็บค่า 1–5 สามารถแปลงโดยให้เรทติ้ง 3,4,5 แทนด้วย 1 และแปลงค่าที่เก็บไว้เป็นคะแนน 1 และ 2 ด้วย 0 เพื่อแสดงว่าสินค้าชิ้นนั้นยังไม่ได้ถูกให้คะแนน

การปรับบรรทัดฐานเรทติ้ง

ในบางโอกาสเราอาจต้องการปรับบรรทัดฐานของเรทติ้ง ให้อยู่ในช่วงที่เราต้องการ เพื่อสะดวกในการคำนวณ วิธีหนึ่งที่สามารถทำได้คือการนำค่าเฉลี่ยของเรทติ้งที่ผู้ใช้เคยให้ ลบออกจากเรทติ้งทุกๆตัวที่ผู้ใช้คนนั้นเคยให้ ผลลัพธ์ที่ได้คือ เรทติ้งที่มีทั้งค่าบวกและค่าลบ โดยที่เรทติ้งค่าบวกจะสื่อถือว่าผู้ใช้มีความคิดบวกกับสินค้าเหล่านั้น ส่วนค่าลบคือผู้ใช้อาจจะไม่ค่อยชอบสินค้าชนิดนั้น สังเกตได้ว่าวิธีนี้ทำให้เราสามารถตีความเรทติ้งที่ผู้ใช้ให้ได้ง่าย กว่าเรทติ้งที่เป็นเฉพาะค่าบวกอย่างเดียว การปรับบรรทัดฐานนี้นิยมใช้เมื่อเราใช้ ระยะห่างแบบโคไซน์ในการวัดความคล้าย

การประเมินประสิทธิภาพของระบบ

ถึงจุดนี้ เราได้ศึกษาการทำนายเรทติ้งและนำไปสู่การแนะนำสินค้าให้ผู้ใช้ ต่อไปเราจะมาดูกันว่าเราจะมีวิธีการวัดประสิทธิภาพ ของระบบแนะนำอย่างไรบ้าง วิธีที่นิยมกันมากที่สุดก็คือ การกันเรทติ้งบางส่วนออกจาก

ยูทิลิตีเมทริกซ์ จากนั้นเราจะทำการทำนายเรทติ้งที่ได้กันออกไปตั้งแต่ต้น หากยูทิลิตีเมทริกซ์เก็บค่าแบบทวิภาค การวัดประสิทธิภาพจะได้ออกมาใช้รูปแบบของความแม่นยำเฉลี่ย นั่นคือ ในบรรดาไอเทมที่เราทายว่าผู้ใช้จะชอบ มีกี่ตัวที่ผู้ใช้ชอบจริงๆ ซึ่งวัดได้โดยใช้สมการ

$$acc = \frac{\sum_{i=1}^{n} |\hat{r} - r|}{n} \tag{4.10}$$

หากยูทิลิตีเมทริกซ์เก็บเรทติ้งที่มีค่าตั้งแต่ 1 ถึง K การวัดประสิทธิภาพอาจทำได้ด้วยค่าคลาดเคลื่อนกำลังสองเฉลี่ย

$$M.S.E = \frac{\sum_{i=1}^{N} (p_{i,j} - vi, j)^2}{N} \tag{4.11}$$

โดยที่ N คือจำนวนเรทติ้งที่ทำการทำนายทั้งหมด

แบบฝึกหัด

1. สมมุติฐานอะพริออรีมีใจความสำคัญอย่างไร
2. ข้อแตกต่างสำคัญของวิธีการแนะนำบนพื้นฐานของเนื้อหากับการคัดกรองร่วมกันคืออะไร
3. ในกรณีไหนที่การแนะนำบนพื้นฐานของเนื้อหาอาจให้ผลลัพธ์ที่ดีกว่าการคัดกรองร่วมกัน
4. การเลือกเก็บคะแนนความชอบเป็นแบบทวิภาคหรือแบบคะแนนให้ผลต่างกันอย่างไร
5. กำหนดข้อมูลการซื้อขายสินค้าของร้านค้าแห่งหนึ่งเป็นดังต่อไปนี้

	อัลมอนด์	โค้ก	ครีม	กล้วย	ไข่	ดอกไม้	ขิง
ลูกค้า 1	1	1		1	1		
ลูกค้า 2			1	1	1		
ลูกค้า 3	1	1			1		1
ลูกค้า 4		1	1			1	
ลูกค้า 5	1		1	1		1	1
ลูกค้า 6		1	1			1	
ลูกค้า 7	1	1		1	1		1

- จงหาค่าสนับสนุนของไอเทมเซต

  [โค้ก, ดอกไม้] และ [อัลมอนด์, กล้วย, ขิง]

- จงหาค่าความเชื่อมั่นของกฎ

  IF อัลมอนด์ & โค้ก THEN ไข่

บทที่ 5

การจำแนกข้อมูลและการทำนาย

มีหลายครั้งที่เป้าหมายของการวิเคราะห์ข้อมูล คือการจำแนกข้อมูลที่มีอยู่ออกเป็นกลุ่มๆ ตามคุณลักษณะบางอย่างที่คล้ายกันภายในกลุ่ม ยกตัวอย่างเช่น ผู้ดูแลระบบอาจต้องการจำแนกอีเมล (E-mail) ที่เข้ามายังองค์กร ออกเป็นกลุ่มของอีเมลปกติ กับกลุ่มที่เป็นอีเมลขยะ (Spam E-mail) ผู้ดูแลระบบอาจมีสมมุติฐานที่ว่าอีเมลขยะมีคุณลักษณะบางอย่างที่เฉพาะกลุ่ม ซึ่งแตกต่างจากอีเมลปกติ แต่ผู้ดูแลระบบก็ไม่สามารถบอกได้ชัดเจนว่าคุณลักษณะนี้คืออะไร

ในกรณีนี้ ผู้ดูแลระบบอาจต้องการสร้างขั้นตอนวิธีที่สามารถเรียนรู้คุณลักษณะเฉพาะตัวดังกล่าว จากชุดข้อมูลฝึกหัดโดยอัตโนมัติ เพื่อนำมาช่วยในการคัดแยกกลุ่มอีเมลที่จะเข้ามาในอนาคตได้อย่างถูกต้องแม่นยำ ชุดข้อมูลฝึกหัดในที่นี้ประกอบด้วย กลุ่มของอีเมลที่ติดป้ายบอกกลุ่มว่าเป็นอีเมลขยะ และกลุ่มของอีเมลที่ติดป้ายบอกกลุ่มว่าเป็น อีเมลปกติ

โดยทั่วไปแล้ว การจำแนกกลุ่มข้อมูล (Classification) คือการหาฟังก์ชัน $h\colon X \to Y$ โดยอาศัย ชุดข้อมูลฝึกหัด $S = \{x_i, y_i\}_{i=1}^{N} \sim D$ ที่ประกอบไปด้วยคู่อันดับ (x_i, y_i) โดย $x_i \in X$ แทนข้อมูลขาเข้า (Input Vector) และ $y_i \in Y$ แทนป้ายบอกกลุ่ม (Class Label) ของ x_i คู่อันดับเหล่านั้นถูกสุ่มมาจากการแจกแจงร่วม $\mathcal{D}$ ป้ายบอกกลุ่มดังกล่าวส่วนใหญ่จะได้จากการนำชุดข้อมูลไปให้ผู้เชี่ยวชาญวิเคราะห์และติดป้ายเพื่อบอกว่า ข้อมูลแต่ละตัวอยู่กลุ่มไหน ในที่นี้ผู้เชี่ยวชาญมีหน้าที่สอนขั้นตอนวิธีในการจำแนกกลุ่มข้อมูลให้จำแนกข้อมูลตามป้ายที่ตนติดกำกับไว้ เราจึงเรียกวิธีการเรียนรู้แบบนี้ว่า การเรียนรู้แบบมีผู้สอน (Supervised Learning)

เป้าหมายของการหาฟังก์ชันนี้ก็เพื่อที่จะใช้ h ในการทำนาย y ของ x ใดๆที่ไม่เคยเจอมาก่อน ได้อย่างถูกต้อง นั่นคือ คำทำนายซึ่งให้สัญลักษณ์เป็น $\hat{y}$ ควรมีค่าตรงกับ y หรือป้ายบอกกลุ่มของ x ซึ่งมาจากการแจกแจงร่วมแบบ เดียวกันกับชุดข้อมูลฝึกหัด ในที่นี้ h จะถูกเรียกว่า ตัวจำแนก (Classifier)

5.1 การเรียนรู้แบบเบส์

การเรียนรู้แบบเบส์ (Bayesian Learning) เป็นแนวทางการสร้างตัวจำแนกแบบหนึ่ง ที่มีพื้นฐานมาจากการใช้กฎของเบส์ (Bayes' Rule) การสร้างตัวจำแนกตามแนวทางนี้มีจุดเริ่มต้นจากการประยุกต์ใช้กฎของเบส์ ซึ่งมีอยู่ว่า

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{5.1}$$

กฎข้างต้นอาจจะจำได้ยาก เพื่อการจำกฎให้ได้ง่ายขึ้น เราสามารถสร้างกฎดังกล่าวจากนิยามพื้นฐานทางสถิติ นั่นคือ เราสามารถพิจารณาว่าความน่าจะเป็นร่วม (Joint Probability) $p(x, y)$ สามารถกระจายพจน์ได้สองแบบคือ

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x) \tag{5.2}$$

ถ้าหากนำการกระจายทั้งสองแบบมาเทียบกัน แล้วย้ายข้าง $p(x)$ ไปหารอีกฝั่งหนึ่ง เราก็จะได้กฎของเบส์ตามสมการที่ 5.1

ก่อนจะนำกฎของเบส์ไปใช้ เราจะมาทำความรู้จักพจน์แต่ละพจน์ก่อน ว่ามีความหมายอย่างไรบ้าง ในบริบทของการจำแนกข้อมูล หากให้ y แทนป้ายบอกกลุ่มและ x แทนเวคเตอร์ของคุณลักษณะ เราจะเรียกพจน์ต่างๆของกฎของเบส์ดังต่อไปนี้

1. $p(y|x)$ คือ ความน่าจะเป็นที่ป้ายบอกกลุ่มของข้อมูลขาเข้า x จะมีค่าเท่ากับ y พจน์นี้มีชื่อเรียกว่า ความน่าจะเป็นภายหลัง (Posterior Probability)

2. $p(x|y)$ คือ ความน่าจะเป็นที่ข้อมูลขาเข้า x จะมาจากกลุ่ม y พจน์นี้มีชื่อเรียกว่า ความควรจะเป็น (Likelihood)

3. $p(y)$ คือ ความน่าจะเป็นที่จะพบป้ายบอกกลุ่มที่มีค่าเท่ากับ y พจน์นี้มีชื่อเรียกว่า ความน่าจะเป็นก่อน (Prior Probability)

4. $p(x)$ คือ ความน่าจะเป็นที่จะพบข้อมูลขาเข้า x พจน์นี้มีชื่อเรียกว่า หลักฐาน (Evidence)

การประยุกต์ใช้กฎของเบส์เพื่อสร้างตัวจำแนก ประกอบด้วยขั้นตอน 2 ขั้นตอน ขั้นตอนที่หนึ่งคือการหาค่าความน่าจะเป็นภายหลังของกลุ่มที่เป็นไปได้ทั้งหมด และขั้นตอนที่สองคือการเลือกทำนายว่าป้ายบอก

กลุ่มของข้อมูลขาเข้าใหม่ ควรจะเป็นไปตามกลุ่มที่ให้ค่าความน่าจะเป็นภายหลังสูงสุด เพื่อลดความซับซ้อนในการทำความเข้าใจการใช้งานกฎของเบส์ สมมุติว่าปัญหาการจำแนกเป็นปัญหาแบบสองกลุ่ม (Binary Classification Problem) โดยกำหนดให้ป้ายบอกกลุ่มทั้งสองแทนด้วยเลข 0 และ 1

ในกรณีข้างต้น ค่าของ $p(y=1|x)$ และ $p(y=0|x)$ สามารถหาได้โดย

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)} \tag{5.3}$$

$$p(y=0|x) = \frac{p(x|y=0)p(y=0)}{p(x)} \tag{5.4}$$

จะเห็นได้ว่าความน่าจะเป็นภายหลังทั้งสอง ขึ้นอยู่กับความควรจะเป็น ความน่าจะเป็นก่อน และหลักฐาน

ดังนั้นเราก็ต้องตามไปหาค่าดังกล่าวให้ครบ จึงจะสามารถคำนวณความน่าจะเป็นภายหลังได้ ทั้งนี้ ขึ้นอยู่กับประเภทของข้อมูล ค่าความควรจะเป็นอาจมีได้หลายรูปแบบ ตั้งแต่ $p(x|y)$ ที่ได้จากคำบอกเล่า ได้จากการนับ หรือ ได้จากฟังก์ชันความหนาแน่นของความน่าจะเป็น (Probability Density Function (PDF)) ของการแจกแจงทางสถิติ เช่น การแจกแจงแบบปกติ ในขั้นนี้ เราจะลองฝึกหาค่าความควรจะเป็น แบบที่ง่ายที่สุดก่อน นั่นคือค่าความควรจะเป็นที่ได้จากคำบอกเล่า โดยการพิจารณาตัวอย่างต่อไปนี้

ตัวอย่าง 5.1.1. *ผู้ป่วยคนหนึ่งไปทำการตรวจสุขภาพที่ห้องทดลองแห่งหนึ่ง พบว่าผลตรวจเป็นบวก (Positive) โดยที่ทราบเพิ่มเติมว่า เครื่องมือที่ใช้ในการตรวจมีความคลาดเคลื่อนเล็กน้อย นั่นแปลว่าในกรณีที่ผู้ป่วยเป็นโรคจริงจะมีโอกาส 98% ที่เครื่องจะตรวจพบความผิดปกติดังกล่าว (True Positive) เช่นเดียวกัน ในกรณีที่ผู้ป่วยไม่ได้เป็นโรค ก็มีโอกาส 97% ที่เครื่องจะรายงานว่าผลตรวจเป็นลบ ในฐานะหมอซึ่งต้องนำผลตรวจมาวิเคราะห์อีกทีหนึ่ง จะวินิจฉัยว่าผู้ป่วยคนดังกล่าวเป็นโรคหรือไม่*

กำหนดให้ $x = \text{positive}$ แทนข้อมูลที่ว่าผลตรวจเป็นบวก และกำหนดให้ $y = 0$ แทนกลุ่มผู้ป่วยที่ไม่เป็นโรค และ $y = 1$ แทนกลุ่มผู้ป่วยที่เป็นโรค เพื่อประกอบการวินิจฉัย แพทย์จะต้องคำนวณ $p(y=0|\text{positive})$ และ $p(y=1|\text{positive})$

5.1.1 การประมาณค่าความควรจะเป็นสูงสุด

สืบเนื่องจากตัวอย่างข้างต้น ความน่าจะเป็นภายหลังทั้งสอง สามารถคำนวณได้จากกฎของเบส์

$$p(y=0|\text{positive}) = \frac{p(\text{positive}|y=0)p(y=0)}{p(\text{positive})} \tag{5.5}$$

$$= \frac{p(\text{positive}|y=0)p(y=0)}{p(\text{positive}|y=0)p(y=0) + p(\text{positive}|y=1)p(y=1)} \tag{5.6}$$

$$p(y=1|\text{positive}) = \frac{p(\text{positive}|y=1)p(y=1)}{p(\text{positive})} \tag{5.7}$$

$$= \frac{p(\text{positive}|y=1)p(y=1)}{p(\text{positive}|y=0)p(y=0) + p(\text{positive}|y=1)p(y=1)} \tag{5.8}$$

จากข้อมูลที่ให้มา พบว่าค่าความควรจะเป็น $p(\text{positive}|y=1)$ ซึ่งแทนความน่าจะเป็นที่ เครื่องตรวจให้ค่าเป็นบวกในกรณีที่ผู้นั้นป่วยจริง มีค่าเท่ากับ 0.98 และพบอีกว่า $p(\text{negative}|y=0)$ ซึ่งแทนความน่าจะเป็นที่เครื่องตรวจให้ค่าเป็นลบในกรณีที่ผู้นั้นแข็งแรง มีค่าเท่ากับ 0.97 จากค่าทั้งสอง เราสามารถคำนวณหาค่าความควรจะเป็นสองค่าที่เหลือได้โดย

$$p(\text{positive}|y=1) + p(\text{negative}|y=1) = 1 \tag{5.9}$$

$$p(\text{negative}|y=1) = 1 - p(\text{positive}|y=1) \tag{5.10}$$

$$= 1 - 0.98 = 0.02 \tag{5.11}$$

$$p(\text{positive}|y=0) + p(\text{negative}|y=0) = 1 \tag{5.12}$$

$$p(\text{positive}|y=0) = 1 - 0.97 = 0.03 \tag{5.13}$$

ถึงจุดนี้ เราทราบพจน์ที่จำเป็นสำหรับการคำนวณความน่าจะเป็นภายหลังเกือบทั้งหมดแล้ว ยกเว้นแต่ ค่าความน่าจะเป็นก่อน ซึ่งข้อมูลไม่ได้กำหนดมาให้

โดยปกติแล้ว $p(y)$ ซึ่งเป็นค่าความน่าจะเป็นที่บอกอัตราการเกิดขึ้นของกลุ่มใดๆ สามารถประมาณได้โดย คำนวณอัตราส่วนระหว่างข้อมูลที่มีป้ายบอกกลุ่มเท่ากับ y ส่วนด้วยจำนวนข้อมูลทั้งหมด แต่ในบางกรณี (อย่างเช่นกรณีนี้) การคำนวณ $p(y)$ ไม่สามารถทำได้ เพราะไม่มีข้อมูลดิบมาให้ด้วย สำหรับกรณีที่ไม่

สามารถคำนวณ $p(y)$ ได้ เราอาจจะอยากสมมุติว่ากลุ่มทั้งสองมีความน่าจะเป็นในการเกิดขึ้นเท่าๆกัน

$$p(y=1) = p(y=0) = 0.5$$

เมื่อความน่าจะเป็นก่อนของทั้งสองกลุ่มมีค่าเท่ากัน เราพบว่า พจน์ $p(y)$ ในกฎของเบส์จะหักล้างกันไป ดังนี้

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)} \tag{5.14}$$

$$= \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0) + p(x|y=1)p(y=1)} \tag{5.15}$$

$$= \frac{p(x|y=1)0.5}{p(x|y=0)0.5 + p(x|y=1)0.5} \tag{5.16}$$

$$= \frac{p(x|y=1)}{p(x|y=0) + p(x|y=1)} \tag{5.17}$$

โดยอาศัยสมการ $p(x) = p(x|y=0)p(y=0) + p(x|y=1)p(y=1)$ มาช่วย สังเกตได้ว่า ค่า $p(x)$ ดังกล่าวมีหน้าที่ปรับบรรทัดฐานให้ ความน่าจะเป็นภายหลังที่คำนวณได้ มีค่าตั้งแต่ 0 ถึง 1 ตามนิยามของความน่าจะเป็นนั่นเอง

กล่าวโดยสรุปคือ การคำนวณความน่าจะเป็นภายหลัง โดยสมมุติว่าความน่าจะเป็นก่อน มีค่าเท่ากันสำหรับทุกกลุ่ม สิ่งที่ต้องคำนวณจริงๆ ก็มีเพียงค่าความควรจะเป็นเท่านั้น เพราะความน่าจะเป็นก่อนจะตัดกันไปเอง สำหรับค่าหลักฐานที่เป็นตัวหาร นั้นไม่ได้มีผลกับการทำนาย เพียงแต่ทำหน้าที่เป็นตัวปรับบรรทัดฐานเท่านั้น ฉะนั้นแล้วสิ่งที่มีผลต่อค่าความน่าจะเป็นภายหลัง และโยงไปถึงผลการทำนาย ก็มีแต่ค่าความควรจะเป็นเท่านั้น หากการทำนายอยู่บนพื้นฐานของการเลือก ความน่าจะเป็นภายหลังที่มีค่าสูงที่สุด ก็เปรียบเสมือนว่า การทำนายอยู่บนพื้นฐานของการเลือกกลุ่มที่ให้ค่าควรจะเป็นสูงที่สุดนั่นเอง

$$y = \arg\max_{y \in (0,1)} p(y|x) = \arg\max_{y \in (0,1)} p(x|y) \tag{5.18}$$

ด้วยเหตุนี้ วิธีการนี้จึงได้ชื่อว่า วิธีการประมาณค่าความควรจะเป็นสูงสุด (Maximum Likelihood Estimate (ML))

กลับมาดูตัวอย่างที่ค้างไว้ พบว่าเราขาดข้อมูล $p(y)$ ไป นั่นคือเราไม่ทราบ ว่าสัดส่วนของคนที่เป็นโรคกับคนสุขภาพดีมีมากน้อยแค่ไหน สิ่งที่สามารถทำได้ก็คือ การสมมุติว่าสัดส่วนของคนที่เป็นโรคกับคนที่ไม่

เป็นโรคนั้นเท่ากัน $p(y=0) = p(y=1) = 0.5$ การตั้งสมมุติฐานแบบนี้อาจจะดูไม่เป็นธรรมชาติ แต่หากเราไม่มีข้อมูลส่วนนี้ การเลือกใช้ ML ก็ดูเหมือนจะเป็นทางเลือกที่ดี เมื่อดำเนินการต่อเราพบว่า

$$p(y=0|\text{positive}) = \frac{p(\text{positive}|y=0) \times 0.5}{p(\text{positive}|y=0) \times 0.5 + p(\text{positive}|y=1) \times 0.5} \tag{5.19}$$

$$= \frac{0.03 \times 0.5}{0.03 \times 0.5 + 0.98 \times 0.5} \tag{5.20}$$

$$= 0.03 \tag{5.21}$$

$$p(y=1|\text{positive}) = \frac{p(\text{positive}|y=1) \times 0.5}{p(\text{positive}|y=0) \times 0.5 + p(\text{positive}|y=1) \times 0.5} \tag{5.22}$$

$$= \frac{0.98 \times 0.5}{0.03 \times 0.5 + 0.98 \times 0.5} \tag{5.23}$$

$$= 0.98 \tag{5.24}$$

$$> p(y=0|\text{positive}) \tag{5.25}$$

ดังนั้นเมื่อใช้ ML แพทย์ควรจะวินิจฉัยว่าบุคคลคนนี้เป็นโรค

5.1.2 ความน่าจะเป็นภายหลังสูงสุด

สมมุติว่าเราได้ข้อมูลเพิ่มเติมมาว่า ประชากรทั้งหมดแค่ 0.008 % เท่านั้นที่จะเป็นโรคร้ายนี้ ผลการวินิจฉัยของแพทย์จะแตกต่างไปจากเดิมหรือไม่ จากข้อมูลเพิ่มเติมดังกล่าวทำให้เราทราบว่า $p(y=1) = 0.008$ และ $p(y=1) = 0.992$ ซึ่งสามารถนำไปใช้ในการคำนวณความน่าจะเป็นภายหลังได้

เราเรียก การนำความน่าจะเป็นก่อนเข้ามาใช้เพื่อคำนวณความน่าจะเป็นภายหลังว่า ความน่าจะเป็นภายหลังสูงสุด (Maximum A Posterior (MAP))

$$y = \arg\max_{y \in (0,1)} p(y|x) = \arg\max_{y \in (0,1)} p(x|y)p(y) \tag{5.26}$$

หากลองใช้วิธี MAP เราจะพบว่า

$$p(y=0|\text{positive}) = \frac{p(\text{positive}|y=0)p(y=0}{p(\text{positive}|y=0)p(y=0)+p(\text{positive}|y=1)p(y=1)} \tag{5.27}$$

$$= \frac{0.03 \times 0.992}{0.03 \times 0.992 + 0.98 \times 0.008} \tag{5.28}$$

$$\approx 0.79 \tag{5.29}$$

$$p(y=1|\text{positive}) = \frac{p(\text{positive}|y=1)p(y=1)}{p(\text{positive}|y=0)p(y=0)+p(\text{positive}|y=1)p(y=1)} \tag{5.30}$$

$$= \frac{0.98 \times 0.008}{0.03 \times 0.992 + 0.98 \times 0.008} \tag{5.31}$$

$$\approx 0.21 \tag{5.32}$$

$$< p(y=0|\text{positive}) \tag{5.33}$$

ดังนั้นเมื่อใช้ MAP แพทย์ควรจะวินิจฉัยว่าบุคคลคนนี้ยังแข็งแรงอยู่

จากตัวอย่างเราพอจะทราบวิธีในการสร้างตัวจำแนก โดยใช้กฎของเบส์มาช่วยในการหาความน่าจะเป็นภายหลัง ของป้ายทั้งสองกลุ่ม ซึ่งนำไปสู่การทำนายป้ายบอกกลุ่มของข้อมูลตัวใหม่บนพื้นฐานของความน่าจะเป็นภายหลังสูงสุด แน่นอนว่า ความน่าจะเป็นภายหลังดังกล่าวอาจจะได้จากการใช้ค่าความควรจะเป็นอย่างเดียว ตามวิธี ML หรืออาจจะใช้ความน่าจะเป็นก่อนเข้ามาช่วยคำนวณด้วย ตามวิธี MAP ซึ่งก็แล้วแต่ความเหมาะสมของสถานการณ์ และข้อมูลที่เรามีอยู่ เราเรียกตัวจำแนกแบบข้างต้นว่า ตัวจำแนกแบบเบส์ (Bayes Classifier)

5.1.3 ตัวจำแนกนาอีฟเบส์

ข้อมูลขาเข้าของตัวอย่างข้างต้นมีแค่หนึ่งมิติ แต่ในชีวิตจริง ข้อมูลขาเข้าอาจจะมีมากกว่าหนึ่งมิติ เราจะมาศึกษาว่า ในกรณีที่ข้อมูลมีมิติที่สูงมากขึ้น การประยุกต์ใช้ตัวจำแนกแบบเบส์จะสามารถทำได้อย่างไรบ้าง

เราจะมาดูการประยุกต์ใช้ตัวจำแนกดังกล่าวผ่านตัวอย่างต่อไปนี้ สมมุติว่าเรามีชุดข้อมูลฝึกหัด แสดงลักษณะของแมว 14 ตัวดังตารางที่ 5.1 ในที่นี้ ข้อมูล $x_i = \{x_i^1, x_i^2, \ldots, x_i^M\}$ คือเวคเตอร์ที่ใช้อธิบายลักษณะทางกายภาพของแมวแต่ละตัว และป้ายบอกกลุ่มเพื่อแสดงว่าแมวตัวดังกล่าวเป็นมิตรหรือดุร้าย เป้า

หมายของเราในที่นี้คือ ต้องการทำนายว่าแมวตัวที่ 15 ซึ่งเราบังเอิญไปเจอจะมีนิสัยดุร้ายหรือเป็นมิตร

$$x_{15} = \{\text{ความยาวขน=ปานกลาง,ลาย=ไม่มีลาย, สีขน=สีเข้ม, รูปร่าง=อ้วน}\}$$

แมว	ความยาวขน	ลาย	สีขน	รูปร่าง	สภาพอารมณ์
1	ปานกลาง	ลายทาง	สีเข้ม	ผอม	ดุร้าย
2	ปานกลาง	ลายทาง	สีเข้ม	อ้วน	ดุร้าย
3	เกรียน	ลายทาง	สีเข้ม	ผอม	เป็นมิตร
4	ยาว	ลายจุด	สีเข้ม	ผอม	เป็นมิตร
5	ยาว	ไม่มีลาย	สีอ่อน	ผอม	เป็นมิตร
6	ยาว	ไม่มีลาย	สีอ่อน	อ้วน	ดุร้าย
7	เกรียน	ไม่มีลาย	สีอ่อน	อ้วน	เป็นมิตร
8	ปานกลาง	ลายจุด	สีเข้ม	ผอม	ดุร้าย
9	ปานกลาง	ไม่มีลาย	สีอ่อน	ผอม	เป็นมิตร
10	ยาว	ลายจุด	สีอ่อน	ผอม	เป็นมิตร
11	ปานกลาง	ลายจุด	สีอ่อน	อ้วน	เป็นมิตร
12	เกรียน	ลายจุด	สีเข้ม	อ้วน	เป็นมิตร
13	เกรียน	ลายทาง	สีอ่อน	ผอม	เป็นมิตร
14	ยาว	ลายจุด	สีเข้ม	อ้วน	ดุร้าย

ตาราง 5.1: ชุดข้อมูลแสดงความยาวขน ลาย สีขนและรูปร่างของแมว 14 ตัว พร้อมป้ายบอกกลุ่มที่ว่าแมวตัวนั้นมีสภาพอารมณ์ดุร้ายหรือเป็นมิตร

หากจะใช้ตัวจำแนกแบบเบส์เพื่อการทำนายว่าแมวตัวนั้นมีสภาพอารมณ์เป็นอย่างไร เราจำเป็นต้องหา

$$p(y|x_{15}) = p(\{x_{15}^1, x_{15}^2, \ldots, x_{15}^M\}|y)p(y) \tag{5.34}$$

โดยที่ $y = \{\text{เป็นมิตร, ดุร้าย}\}$ ปัญหาอยู่ตรงที่ว่า การหา $p(\{x^1, x^2, \ldots, x^M\}|y)$ เป็นไปได้ลำบาก หากมิติของข้อมูลขาเข้า M สูงมากๆ โดยทั่วไปแล้วหากคุณลักษณะของ x มีแค่ 1 มิติ $p(x|y)$ สามารถประมาณได้จากการนับการเกิดขึ้นของคุณลักษณะดังกล่าวในชุดข้อมูลฝึกหัด ยกตัวอย่างเช่น หากต้องการหา $p(x^1 = \text{ปานกลาง}|y = \text{เป็นมิตร})$ จะสามารถทำได้โดยคำนวณอัตราส่วนระหว่างแถวที่มีป้ายบอกกลุ่ม $y = \text{เป็นมิตร}$ และมีคุณลักษณะ $x^1 = \text{ปานกลาง}$ ต่อจำนวนแถวที่มี $y = \text{เป็นมิตร}$ ซึ่งจาก

ตัวอย่างข้างต้นเราจะได้ว่า $p(x^1 = \text{ปานกลาง} | y = \text{เป็นมิตร}) = \frac{2}{9}$ เห็นได้ว่าหากคุณลักษณะของ x มีหลายมิติมากขึ้น การคำนวณ $p(\{x^1, x^2, \ldots, x^M\}|y)$ ก็ทำได้ยากขึ้นด้วย เนื่องจากเราต้องหา x ที่คุณลักษณะที่ $x^1, x^2, \ldots, x^M$ เป็นจริงพร้อมกันทุกตัว

ถึงแม้ว่าเราจะสามารถกระจายความน่าจะเป็นร่วม (Joint Probability) $p(\{x^1, x^2, \ldots, x^M\}|y)$ ออกเป็นพจน์ย่อยได้ดังนี้

$$p(\{x^1, x^2, \ldots, x^M\}|y) = p(x^1|y)p(x^2, \ldots, x^M|y) \tag{5.35}$$
$$= p(x^1|y)p(x^2|y, x^1)p(x^3, \ldots, x^M|y) \tag{5.36}$$
$$= \ldots \tag{5.37}$$

น่าเสียดายว่าความน่าจะเป็นแต่ละพจน์ ก็ยังอยู่ในรูปของความน่าจะเป็นมีเงื่อนไข (Conditional Probability) ซึ่งขึ้นอยู่กับคุณลักษณะตัวก่อนหน้า การคำนวณความน่าจะเป็นมีเงื่อนไขดังกล่าวก็ไม่ง่ายเช่นกัน เพื่อความสะดวกในการคำนวณความน่าจะเป็นต่างๆดังกล่าว จึงมีการเสนอให้ตั้งสมมุติฐานเพิ่มว่า คุณลักษณะแต่ละตัวเป็นอิสระต่อกันหากทราบป้ายบอกกลุ่ม สมมุติฐานนี้รู้จักกันในชื่อที่ว่า สมมุติฐานแบบนาอีฟ (Naive Assumption)

เมื่อนำสมมุติฐานแบบนาอีฟมาปรับใช้ในการกระจายพจน์ในสมการที่ 5.35 เราจะได้ว่า

$$p(\{x^1, x^2, \ldots, x^M\}|y) = p(x^1|y)p(x^2, \ldots, x^M|y, x^1) \tag{5.38}$$
$$= p(x^1|y)p(x^2|y)p(x^3, \ldots, x^M|y, x^1, x^2) \tag{5.39}$$
$$= \prod_{i=1}^{M} p(x^i|y) \tag{5.40}$$

เห็นได้ว่าการคำนวณค่าความควรจะเป็นจะง่ายขึ้น เราเรียกตัวจำแนกแบบเบส์ที่ปรับใช้สมมุติฐานแบบนาอีฟเข้ามาใช้ในการคำนวณความน่าจะเป็นภายหลังว่า ตัวจำแนกนาอีฟเบส์ (Naive Bayes Classifier)

ย้อนกลับมายังโจทย์ปัญหาของเรา หากเราต้องการจะทำนายว่าแมวตัวที่ 15 จะดุร้ายหรือเป็นมิตรเราจะต้องเริ่มจาก

$$\hat{y} = \arg \max_{y \in \{\text{เป็นมิตร, ดุร้าย}\}} p(x_{15}|y)p(y) \tag{5.41}$$
$$= \arg \max_{y \in \{\text{เป็นมิตร, ดุร้าย}\}} p(y) \prod_{i=1}^{4} p(x^i|y) \tag{5.42}$$

สำหรับ $y =$ เป็นมิตร เราต้องคำนวณ

$$p(y = \text{เป็นมิตร}|x_{15}) = p(y = \text{เป็นมิตร})p(\text{ปานกลาง}|y = \text{เป็นมิตร})p(\text{ไม่มีลาย}|y = \text{เป็นมิตร})$$
$$p(\text{สีเข้ม}|y = \text{เป็นมิตร})p(\text{อ้วน}|y = \text{เป็นมิตร}) \quad (5.43)$$
$$= \frac{9}{14}\frac{2}{9}\frac{2}{9}\frac{3}{9}\frac{3}{9}/p(x_{15}) \quad (5.44)$$
$$p(y = \text{ดุร้าย}|x_{15}) = p(y = \text{ดุร้าย})p(\text{ปานกลาง}|y = \text{ดุร้าย})p(\text{ไม่มีลาย}|y = \text{ดุร้าย})$$
$$p(\text{สีเข้ม}|y = \text{ดุร้าย})p(\text{อ้วน}|y = \text{ดุร้าย}) \quad (5.45)$$
$$= \frac{5}{14}\frac{2}{9}\frac{2}{9}\frac{3}{9}\frac{3}{9}/p(x_{15}) \quad (5.46)$$

เราจะพบว่า $p(y = \text{ดุร้าย}|x_{15}) > p(y = \text{เป็นมิตร}|x_{15})$ ฉะนั้นตามกฎของเบส์ เราจะทำนายว่าแมวตัวนี้น่าจะดุร้าย

ตัวจำแนกนาอีฟเบส์ที่ดำเนินการบนข้อมูลที่เก็บคุณลักษณะดีสครีต ยังนิยมใช้ในการจัดการข้อมูลประเภทข้อความ โดยนำไปใช้ในการจำแนกเนื้อความของข้อความออกเป็นกลุ่มหัวข้อต่างๆเช่น การจำแนกข่าวออกเป็นหมวดข่าวต่างๆ ชุดข้อมูลที่รวบรวมข้อมูลข่าวที่นิยมใช้ในการทดสอบประสิทธิภาพตัวจำแนกนาอีฟเบส์ ได้แก่ชุดข้อมูล 20 Newsgroups ซึ่งเป็นการรวบรวมเนื้อข่าวจาก 20 หมวดข่าวเข้าไว้ด้วยกัน

การเตรียมข้อมูลซึ่งอยู่ในรูปแบบของข้อความ มีรายละเอียดปลีกย่อยพอสมควร และเนื้อหาส่วนดังกล่าวก็อยู่นอกเหนือจากเนื้อหารายวิชานี้ ในที่นี้จะแนะนำโดยย่อว่า การสร้างเวคเตอร์ตัวแทนข้อมูลแบบข้อความวิธีหนึ่ง สามารถทำได้โดยกำหนดคำสำคัญ (Keyword) ที่สนใจไว้ล่วงหน้า จากนั้นเราจะนับจำนวนคำสำคัญที่พบในข้อความดังกล่าว ยกตัวอย่างเซตของคำสำคัญ

$$v \in \{\text{มหึมา, คน, หุบเขา, เมือง, แสดแดด, ท้องฟ้า}\}$$

และข้อความต่อไปนี้

ในความรู้สึกของข้าพเจ้า ล้อส แซนเจลิสเป็นเมืองขนาดมหึมาจนยากจะทำความรู้จักและคุ้นเคย มันนั่งอยู่ริมฝั่งมหาสมุทรแล้วเหยียดกายคลุมไปเหนือหุบเขาและที่รายกว้างไกลเกือบห้าร้อยตารางไมล์ (ว่ากันอย่างนั้น) มันเติบโตขึ้นมาด้วยแสงแดดและความแปรปรวนของลมฟ้าอากาศ มันเป็นเมืองที่ใครก็ตามจะรู้สึกเป็นคนแปลกถิ่นตลอดเวลา ความรู้สึกบอกข้าพเจ้าอีกว่า มันเป็นเมืองที่ไม่มีบุคลิก เป็นของตัวเองอย่างแท้จริง มันเป็นเมืองที่รีบร้อนเหมือนนิวยอร์ค แต่บรรยากาศของมันชวนให้เหงายิ่งกว่า ในขณะที่แซน แฟรนซิสโกมีความเป็นกันเอง และเราอาจจำใบหน้าของคนที่เดินสวนกันบนถนนได้ แต่ล้อส แซนเจลิส ข้าพเจ้าไม่คิดว่าผู้คนจะมีความสนใจในกันและกันบ้างเลย

(ใต้ถุนป่าคอนกรีต: 'รงค์ วงษ์สวรรค์)

สามารถแปลงเป็นเวคเตอร์ขาเข้าได้เท่ากับ

$$x = \{\text{มหึมา=1, คน=3, หุบเขา=1, ทะเล=0, ท้องฟ้า=0, เมือง=4}\}$$

หรือแบบย่อในรูปของค่าของความถี่ได้โดย

$$x = \{1,3,1,0,0,4\}$$

เมื่อได้ชุดข้อมูลฝึกหัดในรูปเวคเตอร์แล้ว ก็สามารถนำชุดข้อมูลที่ได้ไปสร้างตัวจำแนกนาอีฟเบส์ต่อได้

5.2 การวิเคราะห์ตัวแบ่งแยกแบบปกติ

ข้อมูลที่เรานำมาวิเคราะห์ก่อนหน้านี้มีลักษณะเป็นดิสครีต นั่นคือคุณลักษณะถูกบันทึกในเชิงทวิภาคหรือเชิงนาม ฉะนั้นแล้วการประมาณหาค่าความควรจะเป็น และความน่าจะเป็นก่อน สามารถทำได้โดยการนับการเกิดขึ้นของเหตุการณ์ที่อยู่ในความสนใจ แต่บางครั้งคุณลักษณะอาจไม่ได้อยู่ในรูปแบบดิสครีตเสมอไป ทำให้การคำนวณค่าความควรจะเป็น และความน่าจะเป็นก่อนมีปัญหา

เนื้อหาในส่วนนี้จะเบนความสนใจ จากสร้างตัวจำแนกเพื่อจำแนกข้อมูลที่คุณลักษณะถูกแสดงด้วยค่าแบบดิสครีต มาศึกษาการสร้างตัวจำแนก เพื่อการจำแนกคุณลักษณะที่มีค่าแบบต่อเนื่อง วิธีนี้ยังคงตั้งอยู่บนพื้นฐานของกฎของเบส์ ตัวจำแนกดังกล่าวมีชื่อเรียกว่า การวิเคราะห์ตัวแบ่งแยกแบบปกติ (Normal Discriminant Analysis)

เราจะเริ่มจากการพิจารณาชุดข้อมูลตัวอย่าง ที่เก็บความสูงของประชากรชายและหญิงอย่างละ 50 คน ไว้ในรูปของจำนวณจริง เป้าหมายสำหรับงานนี้ ก็คือต้องการสร้างตัวจำแนกบนพื้นฐานของชุดข้อมูลฝึกหัด

100 ตัวที่มี เพื่อทำนายว่าบุคคลคนหนึ่งที่เรามีข้อมูลเฉพาะความสูงของเขา สมมุติว่าเป็น $x_q = 173.2$ ควรจะเป็นประชากรเพศใด

หากลองสร้างตัวจำแนกตามวิธีตัวจำแนกแบบเบส์ เราจะพบว่าค่า $p(x_q = 173.2|y = \text{ชาย})$ อาจจะมีค่าเป็น 0 เนื่องจากอาจจะพบว่าไม่มีคนที่สูง 173.2 พอดีอยู่ในชุดข้อมูลฝึกหัดเลย เมื่อค่าความควรจะเป็นเป็นศูนย์ ก็ส่งผลให้ความน่าจะเป็นภายหลัง $p(y = \text{ชาย}|x_q = 173.2)$ มีค่าเท่ากับศูนย์ด้วย ซึ่งอาจจะเป็นเช่นเดียวกัน สำหรับกรณีของ $p(y = \text{หญิง}|x_q = 173.2)$ เราอาจจะอยากแก้ปัญหานี้โดยการเพิ่มจำนวนข้อมูลในชุดข้อมูลฝึกหัดให้มากขึ้น แต่มันจะเป็นไปไม่ได้เลยที่เราจะเก็บข้อมูลได้ครบและครอบคลุมทุกๆค่าที่เป็นไปได้ของค่าส่วนสูงได้ เนื่องจากค่าของข้อมูลเป็นแบบต่อเนื่อง (มีจำนวนค่าที่ต่างกันไม่สิ้นสุด)

จากปัญหาดังกล่าว เราจำเป็นจะต้องหาวิธีในการประมาณค่าความควรจะเป็นแบบใหม่ วิธีหนึ่งที่เป็นได้คือ การสร้างตัวต้นแบบเพื่ออธิบายการแจกแจงทางสถิติของข้อมูลแบบต่อเนื่องดังกล่าว (Data Distribution Model) เมื่อได้รูปแบบของการแจกแจงที่เหมาะสมแล้ว ค่าความควรจะเป็นจะหาได้จาก การคำนวณฟังก์ชันความหนาแน่นของความน่าจะเป็นของการแจกแจงรูปแบบที่เลือก

เนื่องจากการแจกแจงทางสถิติมีอยู่หลายรูปแบบ คำถามต่อมาคือ การแจกแจงแบบไหนถึงจะเหมาะสมกับข้อมูลมากที่สุด ในอุดมคติแล้ว ไม่มีการแจกแจงที่เหมาะสมกับข้อมูลทุกประเภท ผู้ใช้งานจำเป็นจะต้องศึกษาลักษณะของข้อมูลก่อนในเบื้องต้น เพื่อใช้ในการประเมินว่าการแจกแจงของข้อมูลที่อยู่ในมือเป็นแบบใด แต่ในทางปฏิบัติ หากเราไม่ทราบการแจกแจงของข้อมูลอย่างชัดเจนแล้ว เราอาจเริ่มจากการทดลองตั้งสมมุติฐานว่า ข้อมูลมีการแจกแจงแบบปกติ (Normally Distributed) สมมุติฐานนี้ตั้งอยู่บนทฤษฎีที่เรียกว่า Central Limit Theorem [Wasserman, 2013] ทฤษฎีดังกล่าวอ้างว่า หากจำนวนของคุณลักษณะที่เป็นอิสระต่อกัน (Independent Feature) เพิ่มมากขึ้นเรื่อยๆ เราพบว่าการแจกแจงทางสถิติของข้อมูลจะเข้าใกล้การแจกแจงแบบปกติมากขึ้น ด้วยเหตุนี้ หากไม่ทราบการแจกแจงของข้อมูลโดยชัดเจนแล้ว ก็มักนิยมกำหนดให้ตัวต้นแบบของการแจกแจงเป็นการแจกแจงแบบปกติไปด้วยเลย

เมื่อกำหนดตัวต้นแบบของการแจกแจงของข้อมูลได้แล้ว การคำนวณค่าความควรจะเป็น ก็จะเปลี่ยนจากการนับ มาเป็นการคำนวณโดยใช้ PDF ของการแจกแจงแบบปกติแทน สำหรับข้อมูลขาเข้าใน 1 มิติ เราจะใช้ PDF ของการแจกแจงแบบปกติแบบตัวแปรเดี่ยว (Univariate Normal Distribution) ซึ่งมีรูปแบบดังนี้

$$p(x|y = k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{(x - \mu_k)^2}{2\sigma_k^2}\} \tag{5.47}$$

จากนั้น การคำนวณความน่าจะเป็นภายหลังก็สามารถทำได้โดย

$$p(y=0|x) = p(x|y=0)p(y=0) \tag{5.48}$$
$$= p(x|\mu_0, \sigma_0)p(y=0) \tag{5.49}$$
$$= \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\{-\frac{(x-\mu_0)^2}{2\sigma_0^2}\}p(y=0) \tag{5.50}$$
$$p(y=1|x) = p(x|y=1)p(y=1) \tag{5.51}$$
$$= p(x|\mu_1, \sigma_1)p(y=1) \tag{5.52}$$
$$= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\}p(y=1) \tag{5.53}$$

ในที่นี้ เรากำหนดตัวต้นแบบของข้อมูลสองกลุ่มแยกกัน โดย $\mu_0, \mu_1, \sigma_0, \sigma_1$ คือค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของการแจกแจงแบบปกติทั้งสองตัว

เพื่อการคำนวณค่าควรจะเป็น ผู้ใช้งานต้องทราบค่าของ $\mu_0, \mu_1, \sigma_0, \sigma_1$ ให้ครบทุกตัวก่อน ค่าพารามิเตอร์ดังกล่าว สามารถประมาณได้โดยอาศัยชุดข้อมูลฝึกหัด นั่นคือข้อมูลที่มี $y = 0$ จะถูกนำมาใช้หา μ_0, σ_0 และข้อมูลที่มี $y = 1$ จะถูกนำมาใช้หา μ_1, σ_1 ดังนี้

$$\mu_k = \frac{\sum_{n=1}^{N_k} x_n}{N_k} \tag{5.54}$$
$$\sigma_k = \frac{\sum_{n=1}^{N_k}(x_n - \mu_k)^T(x_n - \mu_k)}{N_k} \tag{5.55}$$
$$\pi_k = p(y=k) = \frac{N_k}{N} \tag{5.56}$$

ตัวจำแนกทางสถิติ (Probabilistic Classifier) ที่คำนวณความน่าจะเป็นภายหลัง (ผ่านทางการคำนวณค่าควรจะเป็น) โดยอาศัยการสร้างตัวต้นแบบเพื่อจำลองการเกิดขึ้นของข้อมูล ถือเป็นตัวจำแนกแบบเจเนอเรทีฟ (Generative Classifier)

เมื่อทราบความน่าจะเป็นภายหลังของทั้งสองกลุ่มแล้ว การจะทำนายกลุ่มก็ทำเหมือนตัวจำแนกแบบเบส์ คือเลือกทำนายฉลากตามความน่าจะเป็นภายหลังที่มีค่าสูงกว่า

5.2.1 ฟังก์ชันแบ่งแยก

นอกจากอาศัยการเทียบความน่าจะเป็นภายหลังเพื่อทำนายกลุ่มแล้ว เราสามารถสร้างฟังก์ชันซึ่งเรียกว่าฟังก์ชันแบ่งแยก (Discriminant Function) ฟังก์ชันดังกล่าวเป็นฟังก์ชันที่รับข้อมูลขาเข้าเป็น x แล้วให้ผลลัพธ์คือ y สำหรับการจำแนกข้อมูลแบบทวิภาค เราสามารถสร้างฟังก์ชันแบ่งแยกให้อยู่ในรูปของ

$$f_1(x) = \mathbb{1}\Big(\frac{p(y=1|x)}{p(y=0|x)} > 1\Big) \tag{5.57}$$

ฟังก์ชันดังกล่าวจะให้ผลลัพธ์เป็น 1 เมื่อ $p(y=1|x) > p(y=0|x)$ และให้ผลลัพธ์เป็น 0 หาก $p(y=0|x)$ มีค่าสูงกว่า นอกจากนี้ เราสามารถใช้ฟังก์ชันลอการิทึมมาเปลี่ยนรูปแบบของ $f_1(x)$ ให้เป็น

$$f_2(x) = \mathbb{1}\Big(\log\frac{p(y=1|x)}{p(y=0|x)} > 0\Big) \tag{5.58}$$

เพื่อให้ฟังก์ชันแบ่งแยกมีจุดเปลี่ยนของการทำนาย (Prediction Threshold) ที่ค่า 0 การทำเช่นนี้จะทำให้ การทำนายสะดวกขึ้น เพราะการทำนายจะขึ้นอยู่กับว่าฟังก์ชัน $f_2(x)$ ให้ค่าบวกหรือค่าลบนั่นเอง

5.2.2 การวิเคราะห์ตัวแบ่งแยกแบบปกติหลายตัวแปร

ในกรณีที่ข้อมูลมีมากกว่า 1 มิติ เรายังคงสามารถใช้การแจกแจงแบบปกติ ในการสร้างตัวต้นแบบของข้อมูลได้เพียงแต่ต้องปรับการคำนวณ PDF จากการแจกแจงปกติตัวแปรเดี่ยวเป็นการแจกแจงปกติแบบหลายตัวแปร ซึ่งมีรูปแบบดังนี้

$$p(x|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}\exp\{-\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)\} \tag{5.59}$$

ในที่นี้ Σ แทนเมทริกซ์ความแปรปรวนร่วม ส่วน μ แทนเวคเตอร์ของค่าเฉลี่ยและ $|\Sigma|$ คือดีเทอมิแนนต์ของเมทริกซ์ความแปรปรวนร่วม ซึ่งแต่ละตัวสามารถประมาณค่าโดยใช้ชุดข้อมูลฝึกหัดดังนี้

$$\mu_k = \frac{\sum_{n=1}^{N_k} x_n}{N_k} \tag{5.60}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N_k}(x_n-\mu_k)^T(x_n-\mu_k)}{N_k} \tag{5.61}$$

ความแม่นยำของการประมาณค่าพารามิเตอร์ของตัวต้นแบบทางสถิติ ขึ้นอยู่กับจำนวนของข้อมูลฝึกหัด ในกรณีที่ ชุดข้อมูลฝึกหัดมีขนาดเล็ก การประมาณค่าพารามิเตอร์อาจจะผิดเพี้ยนไปจากความจริงได้ ปัญหานี้จะเกิดขึ้นเมื่อข้อมูล อยู่ในมิติที่สูงมากๆและข้อมูลฝึกหัดมีจำกัด วิธีการแก้ไขปัญหาทางหนึ่งก็คือการลดจำนวนพารามิเตอร์ของตัวต้นแบบให้น้อยลง หรือในอีกนัยหนึ่งก็คือ พยายามเลือกใช้ตัวต้นแบบที่ซับซ้อนน้อยนั่นเอง

ยกตัวอย่าง การใช้การแจกแจงแบบปกติเพื่อสร้างแบบจำลองของข้อมูลใน M มิติ มีจำนวนพารามิเตอร์ที่ต้องประมาณค่าทั้งหมดเท่ากับ $O(M^2)$ ซึ่งถือว่าเยอะมาก จำนวนพารามิเตอร์ส่วนใหญ่ในที่นี้ ก็คือค่าของสมาชิกแต่ละตัวของเมทริกซ์ความแปรปรวนร่วม การลดจำนวนพารามิเตอร์ดังกล่าว สามารถทำได้โดยการจำกัดรูปแบบของเมทริกซ์ความแปรปรวนร่วม ให้เป็นแบบพิเศษ 2 อย่างดังต่อไปนี้

เมทริกซ์ความแปรปรวมร่วมแบบเหมือนกัน

เมทริกซ์ความแปรปรวมร่วมแบบเหมือนกัน (Common Covariance Matrix) อยู่บนสมมุติฐานที่ว่า ประชากรของทั้งสองกลุ่มมีการแจกแจงที่ใกล้เคียงกัน ดังนั้น เมทริกซ์ความแปรปรวมร่วมของทั้งสองกลุ่มควรมีค่าใกล้เคียงกันได้ด้วย การประมาณค่าเมทริกซ์ความแปรปรวมร่วม สามารถทำได้โดยการใช้ค่าเฉลี่ยของเมทริกซ์ความแปรปรวมร่วมของทั้งสองกลุ่ม

$$\Sigma = \frac{\Sigma_0 + \Sigma_1}{2} \tag{5.62}$$

เมทริกซ์ความแปรปรวมร่วมแบบทะแยงมุม

ส่วนอีกวิธีหนึ่งคือการเลือกใช้เมทริกซ์ความแปรปรวมร่วมแบบทะแยงมุม (Diagonal Covariance Matrix) ซึ่งเป็นเทคนิคอาศัยสมมุติฐานแบบนาอีฟ ซึ่งหากยังจำได้สมมุติฐานดังกล่าวถือว่าคุณลักษณะสองตัวใดๆ เป็นอิสระต่อกันเมื่อทราบป้ายของกลุ่ม ดังนั้นแล้วความแปรปรวนร่วมของคุณลักษณะสองตัวใดๆ ก็จะถือว่ามีค่าเป็น 0 ตามไปด้วย เมื่อเป็นเช่นนี้ จะได้ว่าสมาชิกที่อยู่นอกเส้นทะแยงมุมของ เมทริกซ์ความแปรปรวมร่วมจะมีค่าเป็น 0 ไปด้วย ด้วยเหตุนี้จำนวนพารามิเตอร์ที่ต้องประมาณค่า จะลดลงไปด้วย

นอกจากการสร้างตัวจำแนก เพื่อจำแนกข้อมูลออกเป็นสองกลุ่มแล้ว การวิเคราะห์ตัวแบ่งแยกแบบปกติ ยังรองรับการใช้งานในกรณีที่กลุ่มข้อมูลมีมากกว่าสองกลุ่มด้วย สำหรับกรณีดังกล่าว เราเพียงแต่ต้องคำนวณความน่าจะเป็นภายหลังของกลุ่มที่เพิ่มขึ้นมาให้ครบ การทำนายก็สามารถทำได้โดย ตรวจดูว่าความน่าจะเป็นภายหลังของกลุ่มไหนมีค่าสูงสุด

5.3 การถดถอยแบบโลจิสติก

เนื้อหาก่อนหน้านี้ได้กล่าวถึงการสร้างตัวจำแนกในแบบเจเนอเรทีฟ ซึ่งเป็นแนวทางหนึ่งของการสร้างตัวจำแนก หากมองย้อนกลับไปยังปรัชญาตามแนวเจเนอเรทีฟ พบว่าในกรณีดังกล่าวเราพยายามที่จะสร้างตัวต้นแบบเพื่ออธิบาย การเกิดขึ้นของข้อมูล โดยอาศัยรูปแบบการแจกแจงทางสถิติ ซึ่งจากจุดนั้นทำให้เราสามารถคำนวณหาค่าความควรจะเป็นของกลุ่มได้ นำมาซึ่งค่าของความน่าจะเป็นภายหลัง และการทำนายตามลำดับ ทว่ามีผู้ตั้งข้อสงสัยเกี่ยวกับความเหมาะสมของปรัชญานี้ [1] [Vapnik and Vapnik, 1998] ว่าจุดประสงค์ของการจำแนกข้อมูล ก็เพียงแค่ต้องการหาเส้นที่แบ่งข้อมูลออกเป็นกลุ่ม ได้อย่างถูกต้องเท่านั้น ไม่ได้ต้องการรู้ว่าข้อมูลเกิดขึ้นได้อย่างไร และไม่ได้อยากรู้ว่าการแจกแจงของข้อมูลเป็นแบบไหน การเดินตามแนวทางปรัชญาแบบเจเนอเรทีฟ ถือว่าเป็นการแก้ปัญหาที่ไม่ตรงจุด จึงได้เสนอแนวคิดแบบดิสคริมิเนทีฟ (Discriminative) ที่มุ่งหาเส้นที่แบ่งข้อมูลออกเป็นกลุ่ม พร้อมให้ความเห็นว่าแนวทางนี้น่าจะให้ผลการจำแนกที่แม่นยำกว่า

จากนั้นไม่นานก็งานวิจัยเชิงทฤษฎี ที่มุ่งตรวจสอบข้อสังเกตดังกล่าวหลายงานวิจัย แต่งานที่เด่นมากได้แก่ การเปรียบเทียบประสิทธิภาพของการวิเคราะห์ตัวแบ่งแยกแบบปกติกับ การถดถอยแบบโลจิสติก (Logistic Regression) [Ng and Jordan, 2002] ซึ่งตัวจำแนกทั้งสอง ถือว่าเป็นตัวแทนของฝ่ายเจเนอเรทีฟ และฝ่ายดิสคริมิเนทีฟ ที่ได้รับความนิยมสูง

งานวิจัยนั้นรายงานว่า การวิเคราะห์ตัวแบ่งแยกแบบปกติ มักจะทำได้ดีเมื่อชุดข้อมูลฝึกหัดมีขนาดเล็ก แต่หากมีข้อมูลฝึกหัดเยอะขึ้น การถดถอยแบบโลจิสติกมักจะมีประสิทธิภาพดีกว่า นอกจากการถดถอยแบบโลจิสติกแล้ว ตัวจำแนกที่มีแนวทาง แบบดิสคริมิเนทีฟ ที่สำคัญก็ได้แก่ เครื่องกลเวคเตอร์สนับสนุน (Support Vector Machine) และโครงข่ายประสาทเทียม (Artificial Neural Network) ในส่วนนี้ เราจะลองมาศึกษาการสร้างตัวจำแนกตามแนวทางดิสคริมิเนทีฟที่เรียกว่าการถดถอยแบบโลจิสติก (Logistic Regression)

การจะเข้าใจการถดถอยแบบโลจิสติก อาจจะต้องเริ่มจากการย้อนไปทำความเข้าใจวิธีถดถอยเชิงเส้น (Linear Regression) ก่อน เป็นที่ทราบว่า การวิเคราะห์การถดถอยเชิงเส้นคือการหาความสัมพันธ์เชิงเส้นของข้อมูลขาเข้า x กับข้อมูลขาออก y ซึ่งอยู่ในรูปของจำนวนจริง เราสามารถกำหนดรูปแบบของฟังก์ชันนี้ได้โดย

$$y = w^T x + \alpha \tag{5.63}$$

สิ่งที่ต้องประมาณค่าก็คือ เวคเตอร์ w ที่นิยมเรียกว่า เวคเตอร์น้ำหนัก (Weight Vector) โดยที่คุณภาพของ w จะวัดโดย ค่าความคลาดเคลื่อนซึ่งนิยามไว้ว่า $|y - \hat{y}|$ ในที่นี้ y คือคำตอบที่ต้องการ ส่วน $\hat{y}$ คือคำตอบที่วิธีถดถอยทำนาย เป้าหมายของการวิเคราะห์ถดถอยแบบเชิงเส้นก็คือพยายามหา w ที่ทำให้ความคลาดเคลื่อนมีค่าน้อยที่สุด

[1] ซึ่งก็คือ Vladimir Vapnik ผู้คิดค้น Support Vector Machine (SVM)

ภายใต้พื้นฐานเดียวกัน หากเราต้องการจะหาเส้นแบ่ง (Decision Hyperplane) ที่แบ่งข้อมูลออกเป็นกลุ่มๆ เราสามารถดิสครีไทซ์ผลลัพธ์ของวิธีถดถอยซึ่งเป็นจำนวนจริง ให้เป็นผลลัพธ์แบบดิสครีต เราเรียกการกระทำนี้ว่าการแปลงวิธีถดถอยให้เป็นตัวจำแนกนั่นเอง การแปลงดังกล่าวสามารถทำได้โดยการกำหนดขีดแบ่ง (Threshold) ไว้ที่ 0 หากวิธีถดถอยทำนาย $\hat{y} \geq 0$ เราก็จะเทียบโอนว่าตัวจำแนกจะทำนายป้ายของกลุ่มเป็น 1 หรือหาก $\hat{y} < 0$ ให้ทำนายว่าป้ายของกลุ่มเป็น 0 หากเราเทียบตัวจำแนกที่เกิดจากการกำหนดขีดแบ่งของผลลัพธ์จากวิธีถดถอยแบบเชิงเส้น เราจะเห็นว่า ตัวจำแนกนี้มีลักษณะเหมือนฟังก์ชันแบ่งแยกในสมการ 5.58

$$f_2(x) = \mathbb{1}\Big(\log \frac{p(y=1|x)}{p(y=0|x)} > 0 \Big) = \mathbb{1}\Big(\hat{y} > 0 \Big) \tag{5.64}$$

เราสามารถเชื่อมโยงฟังก์ชันแบ่งแยกไปยังฟังก์ชันเชิงเส้น $w^T x$ ผ่าน $\hat{y}$

$$\log \frac{p(y=1|x)}{p(y=0|x)} = \hat{y} = w^T x \tag{5.65}$$

ซึ่งโดยสรุปแล้วสมการข้างต้นบอกว่า เราสามารถสร้างตัวต้นแบบของฟังก์ชันแบ่งแยกได้ด้วยสมการเชิงเส้น $w^T x$

นี่คือจุดเริ่มต้นของการถดถอยแบบโลจิสติก จากนั้นเมื่อทราบว่าเราสามารถสร้างตัวต้นแบบของฟังก์ชันแบ่งแยก ได้ด้วยสมการเชิงเส้นแล้ว เราอยากที่จะหาความน่าจะเป็นที่มาสนับสนุนการการทำนายของตัวจำแนกดังกล่าว นั่นคือ เราต้องการหา $p(y=1|x)$ และ $p(y=0|x)$ จากความสัมพันธ์ในสมการ 5.65 ซึ่งพบว่า

$$\log \frac{p(y=1|x)}{p(y=0|x)} = w^T x \tag{5.66}$$

$$\frac{p(y=1|x)}{p(y=0|x)} = \exp(w^T x) \tag{5.67}$$

$$\frac{p(y=1|x)}{1-p(y=1|x)} = \exp(w^T x) \tag{5.68}$$

$$p(y=1|x) = \exp(w^T x) - p(y=1|x)\exp(w^T x) \tag{5.69}$$

$$p(y=1|x) = \frac{\exp(w^T x)}{1-\exp(w^T x)} \tag{5.70}$$

$$= \frac{1}{1-\exp(-w^T x)} \tag{5.71}$$

ด้วยวิธีเดียวกัน เราจะได้ว่า

$$p(y=0|x) = \frac{\exp(-w^T x)}{1-\exp(-w^T x)} \tag{5.72}$$

เราเรียกฟังก์ชันในรูปแบบ

$$\frac{1}{1-\exp(-w^T x)} \tag{5.73}$$

ว่าฟังก์ชันซิกมอยด์ (Sigmoid Function)

เมื่อได้ความน่าจะเป็นภายหลังของทั้งสองกลุ่มแล้ว เราจะใช้ความน่าจะเป็นทั้งคู่ในการวัดคุณภาพของ w แน่นอนว่า w ที่ดีนั้น จะต้องสามารถอธิบายชุดข้อมูลฝึกหัดได้ถูกต้อง ถ้าข้อมูลขาเข้า x มีป้ายบอกกลุ่มเป็น 0 โมเดลของเราจะต้องให้ค่า $p(y=0|x) > p(y=1|x)$ หรือในทางกลับกัน ถ้าข้อมูลขาเข้ามีป้ายบอกกลุ่มเป็น 1 เราก็ต้องการให้โมเดลของเราให้ค่า $p(y=1|x) > p(y=0|x)$ หรือพูดอีกแบบหนึ่งว่า หากข้อมูลขาเข้ามีป้ายบอกกลุ่มเป็น 0 เราต้องการจะหาเวคเตอร์น้ำหนักที่ทำให้ $p(y=0|x)$ มีค่าสูงสุดและ หากข้อมูลขาเข้ามีป้ายบอกกลุ่มเป็น 1 เราต้องการให้ $p(y=1|x)$ มีค่าสูงสุด ข้อกำหนดทั้งสองประการสามารถสรุปออกเป็นสมการทางคณิตศาสตร์ได้ว่า

$$\mathcal{L}_1 = \prod_{n=1}^{N} p(y_n=1|x_n,w)^{y_n}(1-p(y_n=1|x_n,w))^{1-y_n} \tag{5.74}$$

สมการข้างต้นเรียกว่า ฟังก์ชันค่าควรจะเป็น (Likelihood Function) ในทางสถิติฟังก์ชันค่าควรจะเป็นถือว่าเป็นฟังก์ชันของพารามิเตอร์ที่เราสนใจ นั่นคือเราต้องการจะหาพารามิเตอร์ w ที่ทำให้ฟังก์ชันค่าควรจะเป็นมีค่าสูงสุด โดยตรึงค่าของข้อมูลไว้ระหว่างการหาพารามิเตอร์เหล่านั้น ในบางโอกาสการหาค่าที่ดีที่สุดของฟังก์ชันค่าควรจะเป็น จะสะดวกกว่าหากเราแปลงฟังก์ชันดังกล่าวให้อยู่ในรูปของลอการิทึม

$$\mathcal{L}_2 = \sum_{n=1}^{N} y_n \log p(y_n=1|x_n,w) + (1-y_n)\log(1-p(y_n=1|x_n,w)) \tag{5.75}$$

ในภาษาอังกฤษเราเรียก $\mathcal{L}_2$ ว่า Log-likelihood Function หรือบางทีเราอาจจะใส่เครื่องหมายลบให้กับ Log-likelihood Function แล้วเปลี่ยนเป้าหมายของการหาค่าที่มากที่สุดของ Log-likelihood เป็นการหาค่าที่น้อยที่สุดของค่าลบของ Log-likelihood หรือที่เรียกว่า Negative Log-likelihood Function แทน

คุณสมบัติอย่างหนึ่งของฟังก์ชันจุดประสงค์ $\mathcal{L}_2$ คือเป็นคอนเวกซ์ฟังก์ชัน (Convex Function) จากวิชาแคลคูลัส เราทราบว่าที่คอนเวกซ์ฟังก์ชันมีค่าน้อยที่สุดคือจุดที่อนุพันธ์ของฟังก์ชันเป็นศูนย์ $\mathcal{L}_2' = 0$ ทฤษฎีการหาค่าที่ดีที่สุด (Optimisation Theory) เป็นอีกสาขาวิชาที่มีซับซ้อน ผู้สนใจสามารถหาข้อมูลเพิ่มเติมได้จาก [Boyd and Vandenberghe, 2004] ในที่นี้เราจะมาทำความรู้จักกับวิธีสำหรับการหาค่าที่ดีที่สุด วิธีหนึ่งที่ใช้งานกันอย่างแพร่หลาย ที่รู้จักกันในชื่อของ วิธีของนิวต้น (Newton's Method) วิธีนี้เดิมถูกใช้ในการหารากของฟังก์ชัน กำหนดให้ฟังก์ชันบวก $f(w)$ มีตัวแปรคือ w เราต้องการหาค่าพารามิเตอร์ w ที่ทำให้ฟังก์ชันมีค่าน้อยที่สุด เนื่องจากฟังก์ชันดังกล่าวเป็นฟังก์ชันบวก เราพบว่าค่าที่น้อยที่สุดก็คือ $f(\tilde{w}) = 0$ วิธีของนิวต้นจะทำการประมาณค่า w ที่ดีขึ้นเรื่อยๆ โดยใช้กฎการคำนวณที่ว่า

$$w_{i+1} = w_i - \lambda \frac{f(w_i)}{f'(w_i)} \tag{5.76}$$

ย้อนกลับมายังสิ่งที่เราต้องการนั่นคือ $\mathcal{L}_2' = 0$ เราก็สามารถนำวิธีของนิวต้นมาประยุกต์ใช้ได้โดย

$$w_{i+1} = w_i - \lambda \frac{\mathcal{L}_2'(w_i)}{\mathcal{L}_2''(w_i)} \tag{5.77}$$

ในที่นี้ λ คืออัตราการเรียนรู้ (Learning Rate) ที่บอกถึงอัตราค่าพารามิเตอร์เดิมจะถูกเปลี่ยนแปลง ซึ่งปกติแล้วจะกำหนดให้มีค่าสูงในขั้นต้นๆ ของการหาค่าดีสุดและจะลดลงจนเหลือค่าน้อยๆ ในตอนท้ายๆ ของกระบวนการ ส่วน $\mathcal{L}_2'$ และ $\mathcal{L}_2''$ คือ อนุพันธ์อันดับหนึ่ง (First Derivative) และอนุพันธ์อันดับสอง (Second Derivative) ของ $\mathcal{L}_2$ ซึ่งมีค่าเท่ากับ

$$\mathcal{L}_2' = \sum_{n=1}^{N} [y_n p(y_n = 1|x_n, w) - (1 - y_n) p(y_n = 0|x_n, w)] x_n \tag{5.78}$$

$$\mathcal{L}_2'' = -\sum_{n=1}^{N} x_n p(y_n = 1|x_n, w) p(y_n = 0|x_n, w) x_n^T \tag{5.79}$$

เมื่อทราบค่าของอนุพันธ์ทั้งสองแล้ว เราจะใช้วิธีของนิวตันในการอัพเดทค่าของ w จนถึงจุดที่ดีที่สุด

เห็นได้ว่าวิธีถดถอยแบบโลจิสติกนั้น ไม่ได้สร้างตัวต้นแบบของการเกิดขึ้นของข้อมูลเลย แต่ ขั้นตอนวิธีมุ่งที่จะสร้างตัวต้นแบบของฟังก์ชันแบ่งแยก ด้วยฟังก์ชันเชิงเส้นในรูปของ $w^T x$ เท่านั้นเอง แนววิธีการสร้างตัวจำแนกที่มุ่งหาเส้นแบ่งข้อมูลถูกเรียกว่าตัวจำแนกแบบดิสคริมิเนทีฟ (Discriminative Classifier)

5.4 วิธีเพื่อนบ้านใกล้เคียง

ตัวจำแนกที่มีพื้นฐานมาจากทฤษฎีทางสถิติสองตัวข้างต้น ถือว่าเป็นตัวจำแนกในกลุ่มของตัวต้นแบบแบบพาราเมทริก (Parametric Model) ซึ่งหมายความว่า ตัวจำแนกเหล่านั้นสร้างขึ้นโดยอาศัยตัวต้นแบบอย่างใดอย่างหนึ่ง ซึ่งตัวต้นแบบดังกล่าว ก็ถูกกำหนดไว้ด้วยค่าพารามิเตอร์ซึ่งจำเป็นจะต้องมีการประมาณค่า เพื่อให้ตัวต้นแบบเข้ากันกับชุดข้อมูล

ในส่วนนี้ เราจะลองมาศึกษาตัวจำแนกที่เป็นแบบไม่ใช่พาราเมทริก (Non-parametric) ดูบ้าง ตัวอย่างตัวจำแนกในประเภทนี้ ที่ได้รับความนิยมสูงสุดน่าจะเป็นตัวจำแนกที่มีชื่อว่า วิธีเพื่อนบ้านใกล้เคียง k ตัว (k-Nearest Neighbour (kNN) ทั้งนี้เนื่องมาก จากเป็นตัวจำแนกที่สามารถทำความเข้าใจได้ง่าย และนำไปใช้งานได้สะดวก kNN เป็นตัวจำแนกที่ตั้งอยู่บนสมมุติฐานว่าข้อมูลที่อยู่ใกล้กันควรจะมีป้ายบอกกลุ่มที่เหมือนกัน

กำหนดข้อมูลขาเข้าเป็น x เพื่อนบ้านของ x คือข้อมูลที่อยู่ใกล้กับ x เมื่อวัดด้วยมาตรวัดอย่างใดอย่างหนึ่งซึ่งจะกล่าวต่อไป หากมีข้อมูลขาเข้า x_q ที่ต้องการทำนายป้ายบอกกลุ่ม kNN จะทำการหาจากบรรดาข้อมูลฝึกหัดทั้งหมดที่มีว่า ข้อมูลที่อยู่ใกล้กับ x_q มีใครบ้าง จากนั้น kNN จะตรวจดูว่าเพื่อนบ้านส่วนใหญ่ของ x_q มีป้ายบอกกลุ่มเป็นอะไร และทำนายว่าป้ายบอกกลุ่มของ x_q ควรจะมีค่าเหมือนกับป้ายบอกกลุ่มของเพื่อนบ้านส่วนใหญ่ โดยปกติแล้วเราจะเลือกพิจารณาป้ายบอกกลุ่มข้อมูลที่ใกล้ที่สุดจำนวน k ตัว และนั่นจึงเป็นที่มาของชื่อขั้นตอนวิธีแบบ kNN

ในทางคณิตศาสตร์ หากกำหนดให้ข้อมูลฝึกหัดอยู่ในรูป (x, y) นั่นคือ คู่ลำดับของตัวข้อมูลและป้ายบอกกลุ่มซึ่งได้มาจากผู้เชี่ยวชาญ $y = h(x)$ หลักการทำนายของตัวจำแนกแบบ kNN สามารถเขียนได้ดังนี้

$$h_{knn}(x_q) = \text{majority}(h_1(x_q), h_2(x_q), \ldots, h_k(x_q)) \tag{5.80}$$

โดยที่ k คือจำนวนเพื่อนบ้านของ x_q

ตัวจำแนกแบบ kNN ถือเป็นตัวจำแนกที่อยู่ในกลุ่มของ ผู้เรียนที่ขี้เกียจ (Lazy Learner) ด้วยเหตุผลว่าการใช้งาน kNN นั้น เราไม่จำเป็นจะต้องประมาณค่าพารามิเตอร์ของตัวต้นแบบใดๆเลย จึงถือว่าวิธีนี้ไม่มีช่วงที่ต้องฝึกหัด (หรือเรียนรู้) เมื่อได้ชุดข้อมูลฝึกหัดมา สิ่งที่ต้องทำก็คือการเก็บข้อมูลฝึกหัดไว้ในฐานข้อมูลเท่านั้น ซึ่งถือว่าเป็นข้อดีของตัวจำแนกประเภทนี้ คือ เวลาในการเรียนรู้ (Training Time) สั้นมาก แต่จะมีข้อเสียคือ เมื่อถึงเวลาที่ต้องการจะทำนายว่า x_q ที่เข้ามาใหม่มีป้ายบอกกลุ่มเป็นค่าไหน ตัวจำแนก kNN จำเป็นจำต้องทำงานหนักกว่าตัวจำแนกที่อยู่ในกลุ่มของ ผู้เรียนที่กระตือรือร้น (Eager Learner) โดยสิ่งที่ kNN จะต้องทำในช่วงการทำนายคือการคำนวณระยะทางจากจุด x_q ไปยังจุดอื่นๆทุกจุดที่อยู่ในข้อมูลฝึกหัด ซึ่งถึงแม้ว่าจะเป็นแค่การคำนวณระยะทาง ในการใช้งานจริงอาจจะต้องเสียเวลาคำนวณระยะทางมากพอสมควร โดยเฉพาะอย่างยิ่งกรณีที่ชุดข้อมูลฝึกหัดมีขนาดใหญ่ ด้วยเหตุนี้ทำให้ kNN ไม่เหมาะสมกับการทำนายแบบทันที

ทันใด (Real Time) เมื่อเทียบกับ ผู้เรียนที่กระตือรือร้น อย่างวิธีถดถอยโลจิสติกหรือการวิเคราะห์ตัวแบ่งแยกแบบปกติ

ข้อสังเกตอีกประการของผู้เรียนที่ขี้เกียจคือ จะเก็บชุดข้อมูลฝึกหัดไว้ทั้งหมด ทำให้ข้อมูลต่างๆยังอยู่ครบ ต่างจากผู้เรียนที่กระตือรือร้นซึ่งส่วนใหญ่จะสรุปข้อมูลให้อยู่ในรูปของเซตของพารามิเตอร์ของตัวต้นแบบ ทำให้ข้อมูลบางอย่างที่มีความสำคัญ แต่อาจจะไม่สามารถสรุปให้อยู่ในรูปของตัวต้นแบบหรือฟังก์ชันที่เลือกได้จะต้องหายไป

ทางเทคนิคแล้ว ผู้เรียนที่กระตือรือร้นจะสร้างเส้นแบ่งกลุ่ม โดยอาศัยการประมาณแบบครอบคลุม (Global Approximation) ซึ่งการประมาณนี้อาจจะจำกัดโดยรูปแบบของฟังก์ชันที่นำมาใช้ แต่ผู้เรียนที่ขี้เกียจสามารถสร้างเส้นแบ่งกลุ่ม โดยอาศัยการประมาณเฉพาะที่ (Local Approximation) (ในบริเวณเพื่อนบ้าน) ทำให้เส้นแบ่งกลุ่มที่ได้มีความยืดยุ่นกว่าในบางกรณี ความยืดหยุ่นนี้ทำให้ kNN สามารถสร้างเส้นแบ่งกลุ่มที่ซับซ้อนได้ โดยไม่จำกัดว่าจะต้องเป็นฟังก์ชันเส้นตรงเท่านั้น

ข้อควรระวังของการใช้ kNN คือ kNN ไม่เหมาะสมกับข้อมูลที่อยู่ในมิติสูงมากๆ เนื่องจากคำสาปของมิติข้อมูลบางประการ ยกตัวอย่างเช่นระยะห่างแบบยูคลิดของจุดสองจุดใดๆในมิติสูงๆ จะมีค่าลู่เข้าหากัน เหตุนี้ส่งผลให้เพื่อนบ้านที่หาเจออาจจะมาจากคนละกลุ่ม วิธีการแก้ไขคือ เราอาจจะต้องหาการวัดระยะทางแบบอื่นที่ดีกว่าระยะทางแบบยูคลิด

ปัญหาอีกประการหนึ่งของ kNN คือการเลือกค่า k ที่เหมาะสม ซึ่งยังไม่มีวิธีที่บอกได้ชัดเจนว่า k ควรเป็นเท่าไร แต่ต้องปรับให้เหมาะสมกับข้อมูลที่นำมาวิเคราะห์ โดยส่วนใหญ่แล้วจะใช้วิธีการตรวจสอบแบบไขว้ (Cross Validation) ที่จะกล่าวในส่วนถัดไป หรือหากไม่ต้องการเลือกค่า k เราอาจจะเลือกใช้ kNN แบบถ่วงน้ำหนักตามระยะทาง (Distance-weighted kNN) แทนได้ โดยวิธี kNN แบบถ่วงน้ำหนักตามระยะทางนี้ จะถือข้อมูลฝึกหัดทุกตัวเป็นเพื่อนบ้าน แต่จะลดทอนความเห็นของเพื่อนบ้านเหล่านั้นตามระยะห่างของเพื่อนบ้านนั่นคือ

$$h_{knn}(x_q) = \frac{\sum_{n \in N} w_n h(x_n)}{\sum_{n=1}^{N} w_n} \tag{5.81}$$

$$w_n = \frac{1}{d(x_q, x_n)} \tag{5.82}$$

ในที่นี้ $d(x_q, x_n)$ แทนมาตรวัดระยะห่าง ระหว่าง x_q และ x_n

5.5 การประเมินตัวจำแนก

เนื้อหาช่วงต้นของบทนี้ได้แนะนำตัวจำแนกไปหลายรูปแบบ จากนี้สิ่งที่จำเป็นจะต้องสนใจก็คือ การประเมินประสิทธิภาพของตัวจำแนกที่สร้างขึ้นมา โดยปกติแล้ว เราจะสนใจว่าตัวจำแนกที่เราสร้างขึ้นมา สามารถทำนายป้ายบอกกลุ่มของข้อมูลได้แม่นยำมากน้อยแค่ไหน ซึ่งมักจะวัดกันด้วยค่าความผิดพลาด ซึ่งนิยามไว้ว่า

$$\text{ค่าความผิดพลาด} = \frac{\text{จำนวนข้อมูลที่จำแนกถูก}}{\text{จำนวนข้อมูลทั้งหมด}} \tag{5.83}$$

ค่าความผิดพลาดมีอยู่ด้วยกันสองประเภทคือ ค่าความผิดพลาดบนข้อมูลฝึกหัด (Training Error) และค่าความผิดพลาดบนข้อมูล ที่ไม่เคยเห็นมาก่อน (Generalisation Error) โดยค่าความผิดพลาดบนข้อมูลฝึกหัดสามารถคำนวณได้ เนื่องจากสำหรับชุดข้อมูล ฝึกหัดเรามีคำตอบที่ถูกต้องอยู่แล้ว แต่ในทางปฏิบัติ เราไม่สามารถคำนวณค่าความผิดพลาดบนข้อมูลที่ไม่เคยเห็นมาก่อนได้ เพราะข้อมูลที่ไม่เคยเห็นมาก่อนมีมากมายเกินกว่าเราจะเก็บรวบรวมมาได้หมด (หากเก็บมาได้หมด เราก็ไม่จำเป็นจะต้องสร้าง ตัวจำแนก เพราะเราจะสามารถบอกกลุ่มข้อมูลได้ทันทีโดยไม่ต้องเรียนรู้อะไร) ถึงกระนั้นก็ตามเราก็ยังให้ความสนใจ ค่าความผิดพลาดบนข้อมูลที่ไม่เห็นมาก่อนอยู่ดี ถึงแม้ว่าจะคำนวณโดยตรงไม่ได้ แต่เราจะประมาณค่านั้นโดยอาศัยสิ่งที่เรียกว่าค่าความผิดพลาดบนข้อมูลทดสอบ (Test Error)

ในการสร้างตัวจำแนก เรามักแบ่งชุดข้อมูลฝึกหัดออกเป็นสองส่วน คือ

1. ชุดข้อมูลฝึกหัด (Training Dataset)

   คือส่วนของข้อมูลที่ใช้ในการสร้างตัวจำแนก หรืออีกนัยหนึ่งคือใช้ในปรับค่าพารามิเตอร์ของตัวต้นแบบ โดยตัวจำแนกจะสามารถเข้าถึง x และ y ได้

2. ชุดข้อมูลทดสอบ (Testing Dataset)

   คือส่วนของข้อมูลที่ถูกกันไว้ เพื่อใช้ทดสอบความแม่นยำในการทำนาย ในส่วนนี้ตัวจำแนกจะสามารถเข้าถึง x ได้เท่านั้น และจะต้องพยายามทำนาย y ให้ได้ถูกต้อง

จากประเภทของข้อมูลฝึกหัดข้างต้น ค่าความผิดพลาดที่วัดได้ก็จะถูกเรียกให้สอดคล้องกันดังนี้

1. ค่าความผิดพลาดที่เกิดจากการทำนายป้ายบอกกลุ่มของข้อมูลที่อยู่ในชุดข้อมูลฝึกหัด เรียกว่า ค่าความผิดพลาดบนข้อมูลฝึกหัด ปกติแล้วค่าในส่วนนี้จะมีค่าต่ำเนื่องจากเป็นข้อมูลที่ตัวจำแนกเคยเห็นมาก่อน

2. ค่าความผิดพลาดที่เกิดจากการทำนายป้ายบอกกลุ่มของข้อมูลที่อยู่ในชุดข้อมูลทดสอบ เรียกว่า ค่าความผิดพลาดบนข้อมูลทดสอบ ค่าในส่วนนี้จะวัดความสามารถในการทำนายข้อมูลที่ไม่เคยเจอมาก่อน เป็นค่าที่ใช้ประมาณค่าความผิดพลาดบนข้อมูลที่ไม่เห็นมาก่อน ซึ่งไม่สามารถคำนวณได้โดยตรง

5.5.1 โอเวอร์ฟิตติงและอันเดอร์ฟิตติง

คำว่าโอเวอร์ฟิตติง (Overfitting) ในสื่อถึงเหตุการณ์ที่ตัวจำแนกสามารถทำนายป้ายบอกกลุ่มของชุดข้อมูลฝึกหัดได้ดีมาก แต่ไม่สามารถทำนายป้ายบอกกลุ่มของข้อมูลที่ไม่เคยเจอมาก่อนได้อย่างเป็นที่น่าพอใจ ปกติแล้วโอเวอร์ฟิตติง เป็นสิ่งที่ไม่พึงประสงค์ เพราะถึงแม้ว่าตัวจำแนกจะทำงานได้ดีบนชุดข้อมูลฝึกหัด แต่ต้องไม่ลืมว่า ชุดข้อมูลฝึกหัดเป็นชุดที่เรามีคำตอบอยู่แล้ว ไม่จำเป็นจะต้องให้ตัวจำแนกมาช่วยจำแนก แต่การที่ตัวจำแนกตัวเดียวกันไม่สามารถ ทำนายข้อมูลที่เราไม่รู้คำตอบได้อย่างแม่นยำนั้น กลายเป็นข้อด้อยของตัวจำแนกนั้นไป เพราะทำให้ไม่มีประโยชน์ในการใช้งานเท่าไร

ตรงข้ามกันก็คือ เหตุการณ์ที่ตัวจำแนกจะไม่ปรับตัวให้เข้ากับข้อมูลฝึกหัด ทำให้ค่าความผิดพลาดบนข้อมูลฝึกหัดมีค่าสูง เราเรียกเหตุการณ์นั้นว่า อันเดอร์ฟิตติง (Underfitting) ในกรณีที่ชุดข้อมูลฝึกหัดมีสัญญาณรบกวนมาก การเลือกใช้ตัวจำแนกที่มีลักษณะแบบนี้อาจเป็นข้อดี เพราะตัวจำแนกจะพยายามหลีกเลี่ยงสัญญาณรบกวน แต่หากชุดข้อมูลฝึกหัดมีคุณภาพดี ตัวจำแนกที่อันเดอร์ฟิต ถือว่าเป็นตัวจำแนกที่ไร้ประโยชน์เช่นกัน

5.5.2 มาตรวัดประสิทธิภาพ

ในข้างต้นเรากล่าวถึงประสิทธิภาพ ในรูปแบบของค่าความผิดพลาดในการทำนาย แต่ความจริงแล้ว มีการวัดประสิทธิภาพที่หลากหลายกว่านั้น ปัญหาของการวัดค่าความผิดพลาดในการทำนาย คือยังไม่ละเอียดเพียงพอ เราเพียงแต่รู้ว่าตัวจำแนกทายผิด บางกรณีเราอาจจะต้องการแจกแจงว่าตัวจำแนกทายผิดแบบไหน เช่น การทำนายว่าผู้ป่วยเป็นโรค ก็จะมีการทายผิดอยู่สองแบบคือ แบบที่ 1 ผู้ป่วยเป็นโรคแต่ ตัวจำแนกทำนายว่าไม่เป็น กับ แบบที่ 2 ผู้ป่วยไม่ได้เป็นโรคแต่ทำนายว่าเป็น ซึ่งก็ถือว่าทำนายผิดทั้งคู่ และความผิดพลาดทั้งสองแบบก็ส่งผลต่างกันอีกด้วย ด้วยเหตุนี้เราจึงอยากแจกแจงความผิดพลาด หรือในทางกลับกัน ความถูกต้องให้ออกเป็นกรณีที่ละเอียดมากขึ้น ในที่นี้เราจะสมมุติว่าเราทำงานกับตัวจำแนกแบบทวิภาคซึ่ง ประสิทธิภาพของตัวจำแนกแบบทวิภาค สามารถสรุปออกมาเป็น ตารางที่เรียกว่า เมทริกซ์ค่าความสับสน (Confusion Matrix) ได้ จากตารางข้างต้น เราจะได้ว่า

- Accuracy หาได้โดย $\frac{a+d}{a+b+c+d} = 1 - \text{error}$

		ป้ายที่ทำนาย	
		Negative	Positive
ป้ายจริง	Negative	a — true negative	b — false positive
	Positive	c — false negative	d — true positive

ตาราง 5.2: เมทริกซ์ค่าความสับสนของตัวจำแนกแบบทวิภาค

เป็นการวัดความแม่นยำในการทำนาย ของทั้งสองกลุ่ม

- True Positive Rate (Recall) หาได้โดย $\frac{d}{c+d}$

  แสดงความสามารถในการจำข้อมูลที่เป็นกลุ่มบวกได้มากน้อยแค่ไหน

- True Negative Rate (Specificity) หาได้โดย $\frac{a}{a+b}$

  แสดงความสามารถในการจดจำข้อมูลที่เป็นกลุ่มลบ

- False Positive Rate (False Alarm) หาได้โดย $\frac{b}{a+b}$

  แสดงอัตราส่วนการแจ้งเตือนที่ผิด ยกตัวอย่างเช่น การเตือนว่าผู้ป่วยเป็นโรคทั้งที่จริงแล้วไม่ได้เป็น

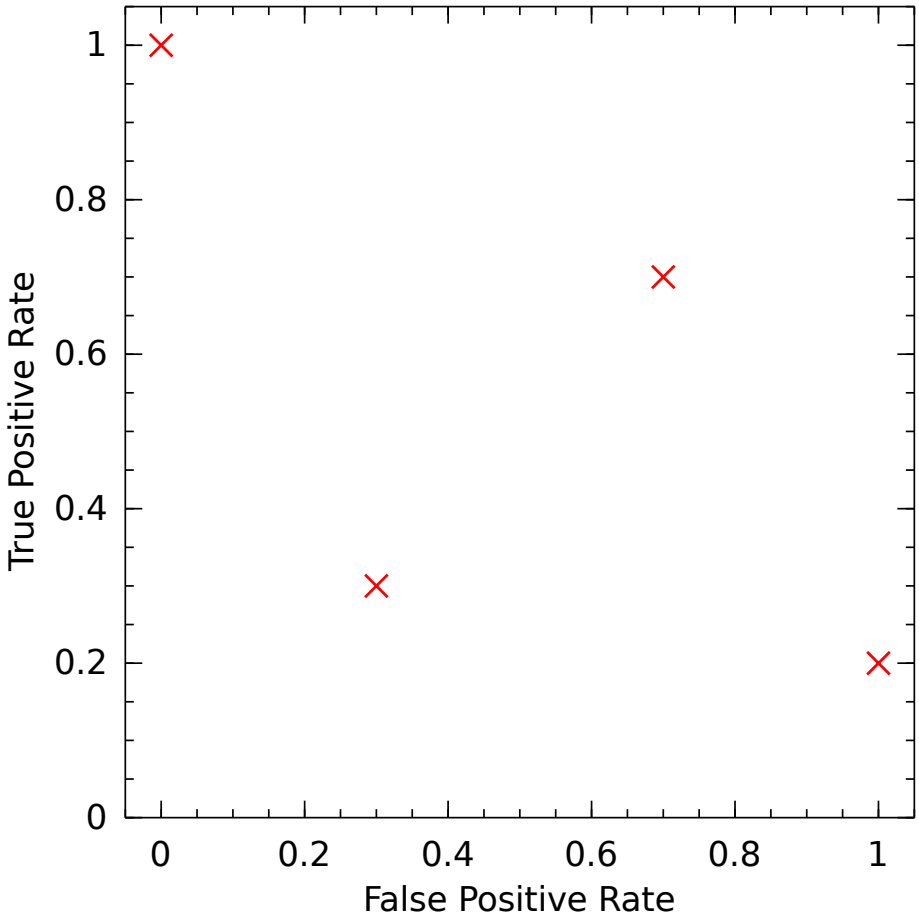
- False Negative Rate หาได้โดย $\frac{c}{c+d}$

  แสดงความผิดพลาดในการตรวจจับ เช่น การไม่แจ้งเตือนทั้งที่ผู้ป่วยเป็นโรค

โดยส่วนใหญ่แล้ว การจำแนกข้อมูล มักจะวัดประสิทธิภาพกันด้วยความแม่นยำแต่ในบางกรณี ที่เราสนใจการแจ้งเตือนที่ผิด (False Alarm) หรือการจำได้ (Recall) เราอาจจะต้องเลือกใช้การวัดประสิทธิภาพให้เหมาะสม

5.5.3 การวิเคราะห์คุณสมบัติการทำงานของเครื่องรับ

วิธีการวิเคราะห์คุณสมบัติการทำงานของเครื่องรับ (Receiver Operation Characteristic Analysis) คือวิธีแสดงประสิทธิภาพของตัวจำแนกโดยอาศัยกราฟ 2 มิติ ที่แนวแกน x แสดงค่า False Positive Rate และแนวแกน y แสดง True Positive Rate ซึ่งกราฟดังกล่าวมีชื่อเรียกว่า รอคกราฟ (ROC Graph) ตัวอย่างของรอคกราฟแสดงไว้ในรูปภาพที่ 5.1



รูปภาพ 5.1: รอคกราฟแสดงประสิทธิภาพของตัวจำแนกสี่ตัว

จากรอคกราฟ จุดแต่ละจุดจะแสดงถึงประสิทธิภาพของตัวจำแนกตัวหนึ่ง โดยตัวจำแนกที่มีประสิทธิภาพดีที่สุด ก็คือตัวจำแนกที่แสดงด้วยจุด (0,1) ซึ่งอยู่ซ้ายบนของกราฟ จุดนี้เป็นจุดที่ False Positive Rate มีค่าเท่ากับ 0 และ True Positive Rate มีค่าเท่ากับ 1 จุดที่แสดงว่าประสิทธิภาพในการทำนายของตัวจำแนกไม่ต่างจากการเดาสุ่มเลย ก็คือบรรดาจุดบนเส้นทะแยงมุมที่ลากตั้งแต่ (0.0) ถึง (1,1) ตัวจำแนกเหล่านี้เป็นตัวจำแนกที่เราไม่ต้องการ เพราะมีค่าความถูกต้องไม่ต่างจากการเดาสุ่ม ในทางกลับกันจุดซึ่งอยู่ต่ำกว่าแนวเส้นทแยงมุม เช่น จุด (1,0) (หรือจุดใกล้เคียง) เป็นพื้นที่ที่บอกว่าตัวจำแนกดังกล่าวส่วนใหญ่แล้วทำนายผิด อย่างไรก็ตาม หากนำตัวจำแนกที่อยู่ต่ำกว่าแนวเส้นทแยงมุมไป เทียบกับตัวจำแนกที่มีตำแหน่งอยู่บนเส้นทแยงมุมแล้ว ถือว่าเป็นตัวจำแนกที่อาจยังนำมาใช้ประโยชน์ได้ เพราะหากกลับคำทำนายที่ได้จากตัวจำแนกดังกล่าว

ทุกครั้ง เราก็จะได้ตัวจำแนกที่ทายถูกมากขึ้น

เมื่อย้อนกลับไปพิจารณาหลักการตัดสินใจที่เราใช้ในกรณีของตัวจำแนกแบบเบส์ หรือวิธีถดถอยแบบโลจิสติกในปัญหาการจำแนกแบบทวิภาค เราจะทำนายว่าป้ายบอกกลุ่มเป็น 1 หาก $p(y = 1|x) > 0.5$ ไม่เช่นนั้นจะทำนายว่า ป้ายบอกกลุ่มเป็น 0 ค่าที่เท่ากับ 0.5 ในที่นี้เรียกกันว่าค่าขั้นต่ำของการตัดสินใจ (Decision Threshold) ที่สามารถปรับให้สอดคล้องกับค่าเสียหาย (Cost) จากการทำนายผิดได้ เราได้ศึกษาไปแล้วว่าการทายผิดมีอยู่สองแบบคือ False Positive (ความผิดพลาดแบบที่ 1 (Type I Error)) ซึ่งแสดงเหตุการณ์ที่ตัวจำแนกทายว่าป้ายบอกกลุ่มเป็นลบ แต่ป้ายบอกกลุ่มที่แท้จริงเป็นบวก อาจเทียบได้กับการทำนายว่าผู้ป่วยเป็นโรค แต่แท้จริงคนคนนั้นมีสุขภาพดี การทายผิดแบบที่สองคือ False Negative (ความผิดพลาดแบบที่ 2 (Type II Error)) ซึ่งแสดงเหตุการณ์ที่ตัวจำแนกทำนายว่าป้ายบอกกลุ่มเป็นบวก แต่ป้ายบอกกลุ่มที่แท้จริงเป็นลบ อาจเทียบได้กับการทำนายว่าคนคนนั้นสุขภาพดี แต่แท้จริงแล้วเค้าเป็นโรค

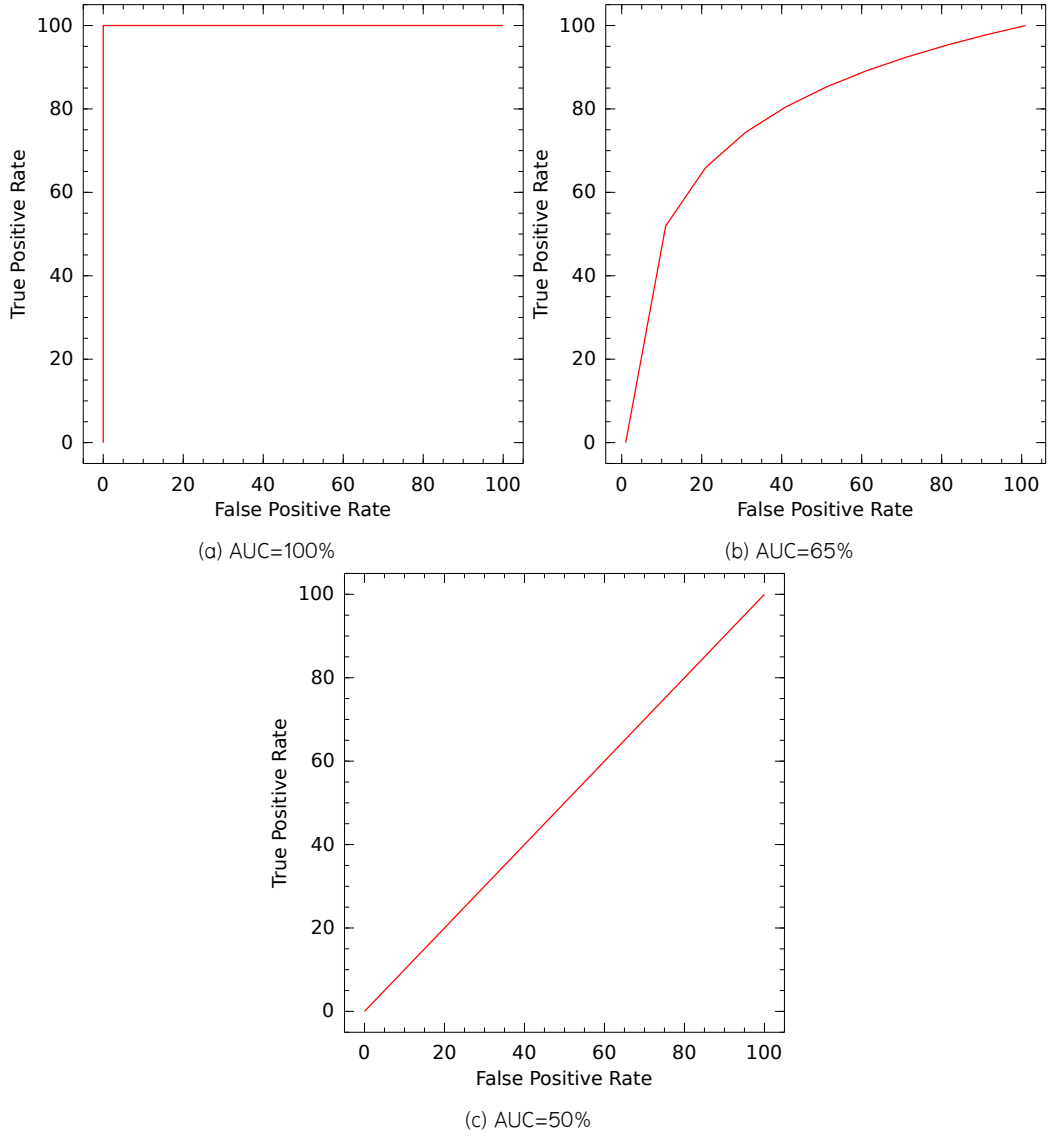
ในการนำตัวจำแนกดังกล่าวไปใช้งานจริง หากค่าเสียหายของการทำทายผิดทั้งสองแบบข้างต้นมีค่าเท่าๆกัน ก็เป็นการเหมาะสมที่จะกำหนดให้ค่าขั้นต่ำของการตัดสินใจอยู่ที่ 0.5 ซึ่งผิดทั้งสองแบบมีผลเสียเท่าๆกัน

แต่หากเราให้ความสำคัญของความผิดพลาดแบบแรก (False Positive) มากกว่า ซึ่งแปลว่าเราไม่อยากให้ตัวจำแนกทำนายผิดว่าผู้ป่วยเป็นโรค หรือในความหมายเดียวกันคืออยากให้ค่าความผิดพลาดแบบที่ 1 มีค่าน้อยๆ เราอาจอยากปรับค่าขั้นต่ำของการตัดสินใจให้สูงกว่า 0.5 นั่นแปลว่าตัวจำแนกจะต้องให้ค่าความน่าจะเป็น $p(y = 1|x)$ ที่สูงจริงๆ จึงจะทำนายว่าผู้ป่วยเป็นโรค ในทำนองเดียวกันหากเราให้ความสำคัญกับความผิดพลาดแบบที่ 2 มากกว่า เราอาจจะพิจารณาลดค่าขั้นต่ำของการตัดสินใจ ให้ต่ำกว่า 0.5 ได้เช่นกัน จะเห็นได้ว่า ถึงแม้ตัวต้นแบบของตัวจำแนกจะเป็นตัวเดียวกัน แต่เมื่อปรับค่าขั้นต่ำของการตัดสินใจให้ต่างกัน ค่าความผิดพลาดที่ได้ ก็อาจจะต่างกันไปด้วย ด้วยเหตุนี้ การที่จะวัดประสิทธิภาพของตัวต้นแบบหนึ่งๆ ให้ครบถ้วนที่สุด จำเป็นจะต้องพิจารณา ค่าความผิดพลาดที่เป็นผลมาจากค่าขั้นต่ำของการตัดสินใจทุกๆค่าที่เป็นไปได้

เมื่อเราแปรผันค่าขั้นต่ำของการตัดสินใจให้กับตัวจำแนกตัวหนึ่ง สิ่งที่ได้ก็จะคือเซตของจุดที่ แสดง False Positive Rate และ True Positive Rate ณ แต่ละค่าของค่าขั้นต่ำของการตัดสินใจ หากเรานำเซตของจุดต่างๆ เหล่านั้นมาวาดลงบนรอคกราฟ เราก็จะได้สิ่งที่เรียกว่า รอคเคิร์ฟ (ROC Curve) ซึ่งเป็นเส้นโค้งที่สรุปภาพรวมของประสิทธิภาพของตัวจำแนก ณ ทุกค่าที่เป็นไปได้ของค่าขั้นต่ำของการตัดสินใจ

ดังที่ได้อธิบายไปข้างต้น ตัวจำแนกมีประสิทธิภาพดีควรจะอยู่เหนือเส้นทแยงมุมขึ้นไปทางซ้ายบนมากๆ ด้วยเหตุนี้ รอคเคิร์ฟของตัวต้นแบบการจำแนก (Classification Model) ที่มีประสิทธิภาพดีควรจะให้เส้นโค้งที่มีลักษณะเหมือน หรือใกล้เคียงกับภาพในรูปที่ 5.2a ในหลายกรณีเส้นโค้งที่ได้จากตัวจำแนกสองเส้น อาจมีลักษณะที่ใกล้เคียงกันมากจนแยกไม่ออกด้วยตาเปล่า จึงมีการพยายามจะสรุปประสิทธิภาพที่ผูกกับเส้นโค้งออกมาในเชิงตัวเลข ค่าดังกล่าวเรียกว่า พื้นที่ใต้เส้นโค้ง (Area Under Curve (AUC)) โดยขนาดพื้นที่ที่มากที่สุด

ที่เป็นไปได้มีค่าเท่ากับ 1 และพื้นที่น้อยที่สุดก็มีค่าเท่ากับ 0 เราจะพบว่าค่า AUC ของตัวต้นแบบการจำแนกที่มีประสิทธิภาพสูง จะมีค่าใกล้ 1 ส่วนตัวต้นแบบที่มีประสิทธิภาพต่ำ จะมี AUC น้อยลงไปตามลำดับ ดังแสดงในรูปที่ 5.2 ทั้งนี้ค่า AUC สามารถนำไปใช้เพื่อเปรียบเทียบประสิทธิภาพในการทำนายตัวจำแนกแบบต่างๆ ได้เช่นเดียวกันกับการใช้ค่าความผิดพลาดอย่างทั่วๆไป นอกเหนือจากการใช้ค่าความผิดพลาด



รูปภาพ 5.2: ค่าพื้นที่ใต้เส้นโค้งแสดงประสิทธิภาพของตัวจำแนก

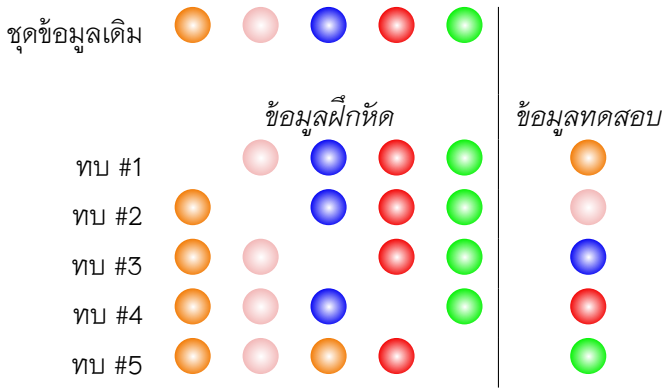
ผู้อ่านสามารถหาข้อมูลเพิ่มเติมเกี่ยวกับการใช้งานการวิเคราะห์คุณสมบัติการทำงานของเครื่องรับ ได้จาก [Fawcett, 2006]

5.5.4 การขยายขนาดของชุดข้อมูลในทางทฤษฎี

เป็นที่ทราบว่าค่าความผิดพลาดบนข้อมูลที่ไม่เคยเห็นมาก่อน ไม่สามารถคำนวณได้โดยตรง เนื่องจากเราไม่ทราบการแจกแจงร่วม ของ (x, y) ดังนั้นเราจำเป็นจะต้องประมาณค่าค่าความผิดพลาดบนข้อมูลที่ไม่เคยเห็นมาก่อน โดยใช้ ค่าความผิดพลาดบนข้อมูลทดสอบแทน ตามหลักสถิติแล้ว เพื่อให้ได้มาซึ่งค่าความผิดพลาดบนข้อมูลทดสอบที่ใกล้เคียงกับค่าความผิดพลาดบนข้อมูลที่ไม่เคยเห็นมาก่อน เราจำเป็นจะต้องหาค่าความผิดพลาดบนข้อมูลทดสอบหลายๆครั้งแล้วนำมาหาค่าเฉลี่ย

ในกรณีนี้ เราจำเป็นต้องมีวิธีจัดการกับชุดข้อมูลฝึกหัด 1 ชุดที่มีอยู่ในมือ ให้เสมือนกับว่าเรามีชุดข้อมูลฝึกหัดหลายๆชุด วิธีการสร้างชุดข้อมูลฝึกหัดหลายๆชุด จากชุดข้อมูลชุดเดียวมีหลายวิธี โดยในที่นี้จะกล่าวถึงวิธีที่ใช้กันมากที่สุดคือ วิธีโฮลเอาท์ (Hold-out Method)

การทำโฮลเอาท์คือการแบ่งข้อมูลออกเป็นสองส่วน คือชุดข้อมูลฝึกหัดและชุดข้อมูลทดสอบ โดยอาศัยการสุ่มเลือก จากนั้นจะสร้างตัวจำแนกโดยใช้ชุดข้อมูลฝึกหัดที่แบ่งไว้ และทดสอบตัวจำแนกที่สร้างเสร็จบนชุดข้อมูลทดสอบอีกต่อหนึ่ง โดยปกติจะทำกระบวนการนี้ซ้ำ k ครั้ง แล้วนำค่าความผิดพลาดบนชุดข้อมูลทดสอบ k ตัวมาเฉลี่ย



รูปภาพ 5.3: วิธีการตรวจสอบไขว้แบบ 5 ทบ

ปัญหาของวิธีโฮลเอาท์ก็คือ ข้อมูลฝึกหัดตัวหนึ่งๆอาจไม่มีโอกาสเข้าไปอยู่ในชุดข้อมูลทดสอบเลย (เนื่องมาจากการสุ่มเลือก) ทำให้การประมาณค่าความผิดพลาดบนข้อมูลที่ไม่เคยเห็นมาก่อนยังไม่ดีเท่าที่ควร เนื่องจากเราไม่ได้ลองทดสอบบนข้อมูลทุกตัวที่มี ด้วยเหตุนี้จึงมีการปรับปรุงวิธีโฮลเอาท์ โดยการแบ่งจะไม่

ใช้วิธีสุ่มอีกต่อไป แต่จะแบ่งข้อมูลออกเป็น k ช่วงเท่าๆกัน แล้วใช้ $k-1$ ช่วงเป็นชุดข้อมูลฝึกหัด ส่วน 1 ช่วงที่เหลือเป็นชุดข้อมูลทดสอบหมุนเวียนกันไปจนครบ k ครั้ง วิธีนี้สามารถรับรองได้ว่า ข้อมูลฝึกหัดทุกตัวจะได้ทำหน้าที่เป็นทั้งข้อมูลฝึกหัดและข้อมูลทดสอบ วิธีนี้มีชื่อเรียกว่า วิธีการตรวจสอบไขว้แบบ k ทบ (k-fold Cross Validation) ตัวอย่างการทำการตรวจสอบไขว้แบบ 5 ทบ แสดงไว้ดังรูปที่ 5.3 ในกรณีพิเศษที่ $k = N$ หรือแบ่งส่วนเท่ากับจำนวนข้อมูลในชุดข้อมูลฝึกหัด เราจะเรียกวิธีการตรวจสอบไขว้แบบนั้นว่า Leave-one-out Cross Validation

แบบฝึกหัด

1. เหตุใดจึงเรียกการจำแนกข้อมูลว่าเป็นการเรียนรู้แบบมีผู้สอน
2. อะไรคือข้อแตกต่างระหว่างตัวจำแนกแบบเจเนอเรทีฟกับตัวจำแนกแบบดิสคริมิเนทีฟ
3. ใจความสำคัญของสมมุติฐานแบบนาอีฟคืออะไร และสามารถนำไปประยุกต์ใช้กับการจำแนกได้อย่างไรบ้าง
4. ในกรณีไหนที่ตัวจำแนกแบบเจเนอเรทีฟอาจมีประสิทธิภาพดีกว่าตัวจำแนกแบบดิสคริมิเนทีฟ
5. เราจะมีวิธีการเลือกค่า k ในการใช้งานวิธีเพื่อนบ้านใกล้เคียงได้อย่างไรบ้าง
6. วิธีเพื่อนบ้านใกล้เคียง k ตัวอาจให้ผลการทำนายดีกว่าวิธีถดถอยโลจิสติกในกรณีไหนบ้าง
7. อธิบายว่าทำไมการวัดความแม่นยำในการทำนายอย่างเดียว อาจจะไม่พอสำหรับการวัดประสิทธิภาพของตัวจำแนก

บทที่ 6

การจัดกลุ่มข้อมูล

ในบทที่แล้วเราได้ศึกษาการจำแนกกลุ่มข้อมูลซึ่งถือเป็น การเรียนรู้แบบมีผู้สอน (Supervised Learning) ในเชิงที่ว่า ข้อมูลขาเข้าถูกแบ่งออกเป็นกลุ่มล่วงหน้าไว้แล้วโดยผู้เชี่ยวชาญ โดยอาศัยป้ายบอกกลุ่มเป็นตัวกำหนด ภารกิจของการจำแนกก็คือ สร้างตัวจำแนกโดยใช้ชุดข้อมูลฝึกหัดดังกล่าว ให้สามารถจำแนกข้อมูลขาเข้าใหม่ให้ตรงกับการจำแนกของผู้เชี่ยวชาญให้มากที่สุด

ในบทนี้เราจะมาศึกษาวิธีการแบ่งกลุ่มอีกวิธีหนึ่ง ซึ่งมีปรัชญาต่างจากการเรียนรู้แบบมีผู้สอน โดยที่ชุดข้อมูลขาเข้าไม่ได้ถูกกำหนดป้ายบอกกลุ่มมาไว้ก่อน อาจจะเนื่องมาจากขาดผู้เชี่ยวชาญที่จะมากำหนดป้ายบอกกลุ่ม หรือยังไม่ได้ดำเนินการถึงขั้นนั้น ในกรณีดังกล่าวข้อมูลขาเข้าจะไม่มีป้ายบอกกลุ่มมาด้วย หรือที่เรียกกันว่า ข้อมูลไม่ติดป้าย (Unlabelled Data) ในภาษาอังกฤษนั้น เราจะเรียกกลุ่มของข้อมูลที่ติดป้ายว่า คลาส (Class) แต่จะเรียกกลุ่มของ ข้อมูลที่ไม่ติดป้ายว่า คลัสเตอร์ (Cluster) ในภาษาไทยเราจะใช้คำว่ากลุ่มสำหรับข้อมูลติดป้ายและข้อมูลไม่ติดป้าย ดังนั้นแล้วการดำเนินการเพื่อหาคลัสเตอร์เราจะเรียกว่า การจัดกลุ่มข้อมูล (Clustering)

เป้าหมายของการจัดกลุ่มข้อมูล คือเพื่อทำความเข้าใจกลุ่มก้อนของข้อมูลในระยะเริ่มต้น หรือเพื่อค้นพบโครงสร้างบางอย่างที่ซ่อนอยู่ในกลุ่มข้อมูลนั้น ยกตัวอย่างเช่น การจัดกลุ่มเอกสารที่มีความคล้ายคลึงกันของเนื้อหา หรือการจัดกลุ่มผู้ป่วยที่มีอาการใกล้เคียงกัน

ในที่นี้เราจะศึกษาแนวทางการจัดกลุ่มข้อมูลสองแนวทาง แนวทางแรกเรียกว่า วิธีการจัดกลุ่มแบบแบ่ง (Partitioning Clustering) ซึ่งเป็นการจัดกลุ่มที่แบ่งแยกข้อมูลออกเป็นกลุ่มก้อนอย่างชัดเจน ข้อมูลแต่ละตัวจะเป็นสมาชิกของกลุ่มเพียงกลุ่มเดียวเท่านั้น แนวทางที่สองเรียกว่า วิธีการจัดกลุ่มแบบลำดับขั้น (Hierachical Clustering) ซึ่งเป็นการแบ่งแยกข้อมูล ออกเป็นกลุ่มย่อยที่ซึ่งจะเป็นส่วนหนึ่งของกลุ่มที่ใหญ่ขึ้นไปเรื่อยๆ มีลักษณะความสัมพันธ์ระหว่างกลุ่มเป็นลำดับขั้นแบบโครงสร้างข้อมูลแบบต้นไม้ (Tree)

6.1 วิธีการจัดกลุ่มแบบแบ่ง

เนื่องจากชุดข้อมูลสำหรับทำการจัดกลุ่มข้อมูลไม่มีป้ายบอกกลุ่มมาด้วย การจัดกลุ่มจำเป็นจะต้องอาศัยวิธีการอื่นเพื่อบอกว่าข้อมูลสองตัวควรอยู่ในกลุ่มเดียวกัน โดยทั่วไปแล้ว เราจะใช้ความคล้ายกันของข้อมูลเป็นตัววัดว่าข้อมูลสองตัวควรจะอยู่ในกลุ่มเดียวกันหรือไม่

กำหนดให้ชุดข้อมูลแทนด้วย $\mathcal{D} = (x_1, \ldots, x_N)$ โดยข้อมูลขาเข้า $x_i = [x_i^1, \ldots, x_i^M]$ คือเวคเตอร์ที่อยู่ใน M มิติ จากนั้นกำหนดให้ฟังก์ชันระยะห่าง (Distance Function) ซึ่งใช้วัดระยะห่างระหว่างข้อมูล x_i และ x_j แทนด้วย $d(x_i, x_j)$ ในที่นี้ เราจะเลือกใช้ระยะห่างแบบยูคลิดซึ่งนิยามไว้โดย

$$d(x_i, x_j) = \sum_{m=1}^{M} (x_i^m - x_j^m)^2 = ||x_i - x_j||$$

ภารกิจของการจัดกลุ่มก็คือ การหาค่าของเวคเตอร์ z ที่ใช้ระบุว่าข้อมูลแต่ละตัวเป็นสมาชิกของกลุ่มไหน ซึ่ง z ถูกนิยามไว้ว่า

$$z = [z_1, \ldots, z_N] \in \{1, \ldots, k\}$$

โดยที่ค่าของ z ที่หาได้ต้องทำให้ลดฟังก์ชันจุดประสงค์ต่อไปนี้ให้มีค่าต่ำสุด

$$F_{obj} = \frac{1}{2} \sum_{i=1}^{n} ||x_n - m_{z_n}||^2 \tag{6.1}$$

หนึ่งในวิธีจัดกลุ่มข้อมูลแบบแบ่งที่ใช้กันอย่างแพร่หลายก็คือ ขั้นตอนวิธีเคมีนส์ (k-means Algorithm)

6.1.1 ขั้นตอนวิธีเคมีนส์

ขั้นตอนวิธีเคมีนส์เป็นวิธีการจัดกลุ่มข้อมูลที่แทนกลุ่มก้อนของข้อมูลด้วยค่าเฉลี่ย (Mean) ของสมาชิกในกลุ่มนั้นๆ เนื่องจากวิธีเคมีนส์ใช้ค่าเฉลี่ยในการระบุกลุ่ม วิธีการนี้จะเหมาะสมกับข้อมูลที่สามารถหาค่าเฉลี่ยได้โดยไม่ขัดต่อธรรมชาติของข้อมูล หมายความว่า ข้อมูลควรจะถูกแสดงโดยคุณลักษณะเชิงตัวเลขที่เป็นค่าต่อเนื่อง หากคุณลักษณะเป็นแบบเชิงนาม หรือแบบดิสครีตแล้ว ค่าเฉลี่ยที่หาได้อาจจะอยู่นอกโดเมนของค่าที่เป็นไปได้ของคุณลักษณะดังกล่าว จึงอาจจะต้องพิจารณาวิธีอื่นๆที่ไม่ขึ้นกับการหาค่าเฉลี่ย ในที่นี้เราจะสมมุติว่าข้อมูลขาเข้ามีลักษณะเชิงตัวเลขแบบต่อเนื่อง

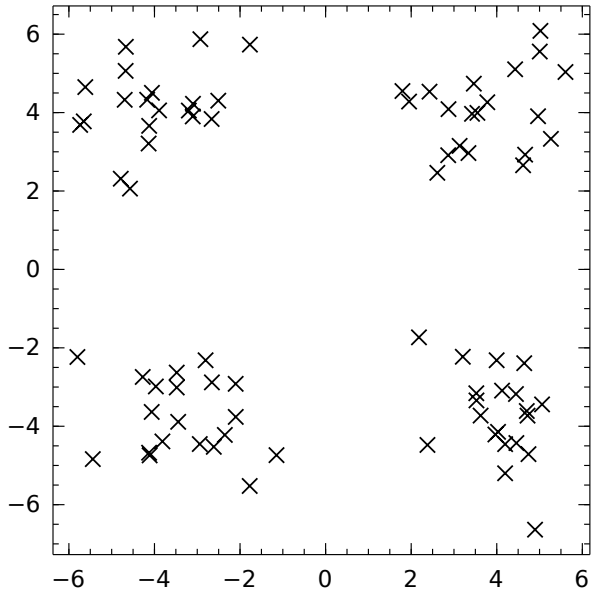
Algorithm 3 วิธีการจัดกลุ่มแบบเคมีนส์

1: Randomly choose $\mu_{1:k}$ ▷ กำหนดค่าตั้งต้นให้เวคเตอร์ค่าเฉลี่ย
2: Initilise z ▷ กำหนดค่าตั้งต้นให้เวคเตอร์บอกกลุ่ม
3: **do**
4: $z_n = \arg\min_{i \in \{1,\ldots,k\}} d(x_n, \mu_i)$ ▷ จัดข้อมูลแต่ละตัวเข้ากลุ่มที่มีค่าเฉลี่ยใกล้สุด
5: $\mu_k = \frac{1}{N_k} \sum_{n:z_n=k} x_n$ ▷ ปรับค่าเฉลี่ยของกลุ่ม
6: **while** z does not change
7: **return** z ▷ เวคเตอร์ z ที่ใช้บอกกลุ่มของข้อมูลแต่ละตัว

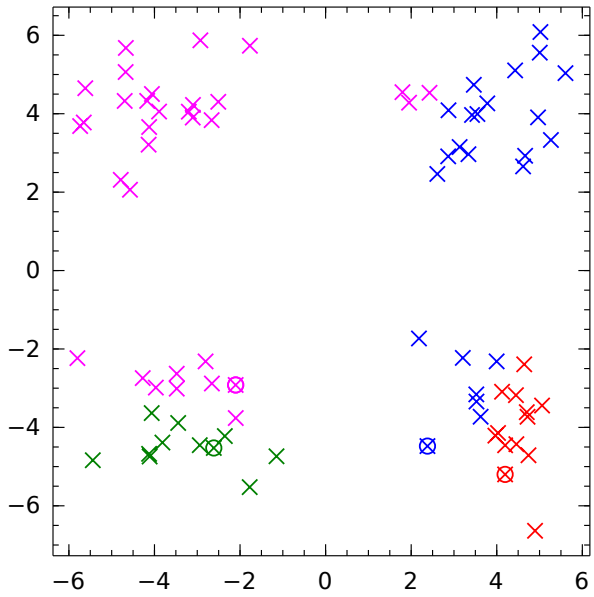
ภาพรวมของวิธีเคมีนส์ ซึ่งแสดงใน Algorithm 3 เริ่มจากการกำหนด ค่าเฉลี่ยของกลุ่ม k ตัวโดยอาศัยการสุ่ม เมื่อได้ค่าเฉลี่ยเริ่มต้นแล้ว ขั้นตอนวิธีก็จะระบุกลุ่มให้กับข้อมูลแต่ละตัว โดยจะพิจารณาเลือกกลุ่มที่ค่าเฉลี่ยของกลุ่มอยู่ใกล้ที่สุด พร้อมกับปรับค่า z ให้สอดคล้องกัน ทั้งนี้อาจมีข้อมูลบางตัว ย้ายจากกลุ่มหนึ่งไปอีกกลุ่มหนึ่งได้ เมื่อปรับค่า z เรียบร้อยแล้ว ขั้นตอนวิธีก็จะคำนวณหาค่าเฉลี่ยใหม่ของข้อมูลที่อยู่ในกลุ่มเดียวกัน ทั้งนี้ค่าเฉลี่ยอาจเปลี่ยนแปลง เพราะอาจมีข้อมูลตัวใหม่เข้ามาในกลุ่ม ขั้นตอนวิธีจะดำเนินต่อไปเรื่อยๆ จนกว่าค่า z จะไม่เปลี่ยนแปลง นั่นคือ ถึงจุดที่ข้อมูลทุกตัว ถูกจัดเข้าไว้ในกลุ่มที่ควรจะเป็นแล้วนั่นเอง

ตัวอย่างการทำงานของวิธีเคมีนส์แสดงไว้ใน รูปภาพที่ 6.1–6.3 โดย รูปภาพที่ 6.1 แสดงการกระจายของข้อมูล 80 ตัวที่สุ่มมาจากการแจกแจงแบบปกติ 4 กลุ่มกลุ่มละ 20 ตัว การแจกแจงสี่กลุ่มดังกล่าวมีจุดกลางอยู่ที่ (4,4), (−4,4), (4,−4), (−4,−4) ตามลำดับ โดยการแจกแจงแบบปกติทั้งหมดมีค่าความแปรปรวนเท่ากับ 1 จะเห็นว่าในมุมมองของขั้นตอนวิธีเคมีนส์ ข้อมูลทุกตัวถือว่ายังไม่มีกลุ่ม เพราะไม่มีป้ายบอกกลุ่มกำกับ ในที่นี้แสดงด้วยการที่ข้อมูลทุกตัวอยู่กลุ่มสีดำเหมือนกัน ทั้งหมด แต่ในมุมมองของมนุษย์เราเห็นชัดเจนแล้วว่าข้อมูลมีการรวมกลุ่มก้อนเป็น 4 กลุ่มอย่างชัดเจน ทั้งนี้เนื่องจากข้อมูลอยู่ในมิติที่มนุษย์เราสามารถรับรู้ได้ด้วยการรับรู้ทางสายตา แต่หากข้อมูลอยู่ในมิติที่สูงเกินกว่าประสาทรับรู้ทางสายตาของมนุษย์แล้ว เราไม่อาจจะทราบได้เลยว่า ข้อมูลมีลักษณะเป็นกลุ่มก้อนอย่างไรบ้าง

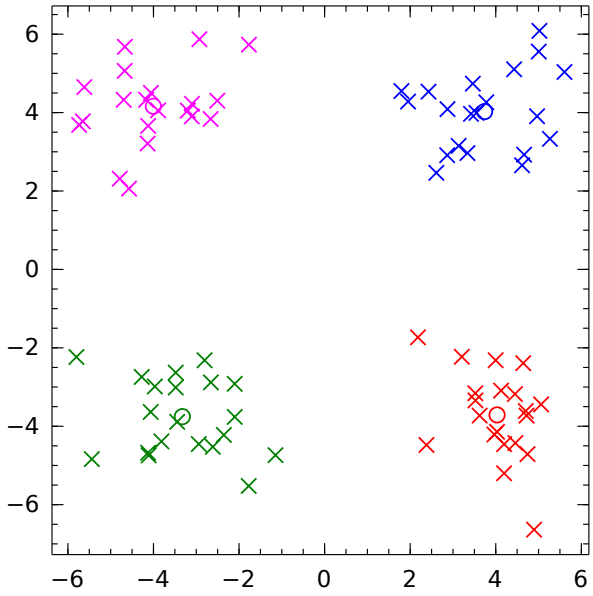
กลับมาที่การทำงานของขั้นตอนวิธีเคมีนส์ ซึ่งในช่วงแรกจะทำการปรับและจัดกลุ่มของข้อมูลใหม่เพื่อ ลดค่าของฟังก์ชันจุดประสงค์ ดังแสดงในภาพที่ 6.2 สังเกตว่าจุดศูนย์กลางของกลุ่มจะถูกแสดง ด้วยสัญลักษณ์วงกลม และข้อมูลต่างๆจะถูกจัดเข้ากลุ่มที่อยู่ใกล้ที่สุด เมื่อวัดจากจุดศูนย์กลางของกลุ่มดังกล่าว สังเกตอีกว่า ในช่วงแรกๆการจัดกลุ่มอาจยังไม่เป็นไปตามที่เราคาดหวังเท่าไหร่ เพราะจุดศูนย์กลางเริ่มต้นเป็นจุดที่ได้มาจากการสุ่มค่าขึ้นมา แต่เมื่อขั้นตอนวิธีดำเนินต่อไป จุดศูนย์กลางของกลุ่มก็จะเคลื่อนไปยังจุดที่ควรจะเป็น ดังที่เห็นในรูปที่ 6.3 ซึ่ง ณ ตำแหน่งดังกล่าว จะเป็นจุดที่ทำให้ค่าของฟังก์ชันจุดประสงค์มีค่าต่ำสุด



รูปภาพ 6.1: แผนภาพการกระจายของชุดข้อมูลที่ยังไม่ได้จัดกลุ่ม



รูปภาพ 6.2: แผนภาพการกระจายแสดงกลุ่มข้อมูลในช่วงแรกของการจัดกลุ่มโดยวิธีเคมีนส์



รูปภาพ 6.3: แผนภาพการกระจายแสดงกลุ่มข้อมูลในช่วงสุดท้ายของการจัดกลุ่มโดยวิธีเคมีนส์

เราสามารถแสดงได้ว่าวิธีการเคมีนส์มุ่งที่จะหาค่าต่ำสุดให้กับฟังก์ชันจุดประสงค์ในสมการที่ 6.1 หรือสามารถเขียนในรูปที่ชัดเจนมากขึ้นโดยระบุพารามิเตอร์ของฟังก์ชันลงไปด้วย ได้ดังนี้

$$F(z_{1:n}, \mu_{1:k}) = \frac{1}{2}\sum_{i=1}^{n} ||x_n - \mu_{z_n}||^2 \tag{6.2}$$

สังเกตว่า

1. การที่วิธีการเคมีนส์ตรึงค่าของค่าเฉลี่ยไว้ แล้วย้ายข้อมูลให้เข้าไปอยู่ในกลุ่มที่มีค่าเฉลี่ยอยู่ใกล้ที่สุด เป็นการลดค่าของ F โดยการปรับ z

2. การที่วิธีการเคมีนส์ตรึงค่าของ z ไว้ แล้วคำนวณค่าเฉลี่ยของกลุ่มใหม่ ก็เป็นการลดค่าของ F โดยการปรับ μ

สองขั้นตอนดังกล่าวเป็นการแสดงว่าวิธีเคมีนส์จะทำการหาค่า z และ μ ที่ลดค่าของ F ลงเรื่อยๆ จนถึงค่าต่ำสุดเฉพาะที่ (Local Minimum) กลยุทธ์ดังกล่าวมีลักษณะของขั้นตอนวิธีแบบลดลงตามพิกัด (Coordinate Descend Algorithm) ซึ่งจะผลัดกันหาค่าที่ดีที่สุดของพารามิเตอร์ แต่ละตัว จนกว่าจะเจอจุดต่ำสุดของฟังก์ชัน

6.1.2 ขั้นตอนวิธีเคเมดอยส์

ในบางกรณีที่ค่าเฉลี่ยใช้ไม่ได้ เช่น กรณีที่คุณลักษณะเป็นแบบดิสครีต ค่าเฉลี่ยของค่าดิสครีตที่ได้อาจอยู่นอกโดเมน เช่น หากข้อมูลของเรามีคุณลักษณะหนึ่งที่ใช้บ่งบอกอาชีพ 3 อย่างคือ หมอ ทหาร นักธุรกิจ ซึ่งคุณลักษณะเชิงนามสามตัวดังกล่าว ถูกแปลงให้อยู่ในรูปของตัวเลขคือ หมอ=1 ทหาร=2 และ นักธุรกิจ=3 จากนั้นในการคำนวณค่าเฉลี่ยของคุณลักษณะที่เก็บอาชีพที่ถึงแม้จะถูกแปลงเป็นคุณลักษณะเชิงตัวเลขแล้ว เราอาจจะพบว่าค่าเฉลี่ยอาชีพของคนส่วนใหญ่คือ 2.2 ซึ่งไม่ตรงกับอาชีพใดเลย การแก้ไขก็คือแทนที่จะคำนวณค่าเฉลี่ย เราก็จะใช้ค่ามัธยฐานแทน ซึ่งค่ากลางแบบมัธยฐาน ก็คือข้อมูลตัวหนึ่งในกลุ่มนั่นเอง โดยวิธีการเคเมดอยส์ (k-medoids) ก็จะคล้ายกับวิธีการเคมีนส์ทุกอย่างยกเว้นการหาค่ากลาง เท่านั้น ขั้นตอนวิธีเคเมดอยส์ แสดงไว้ใน Algorithm 4

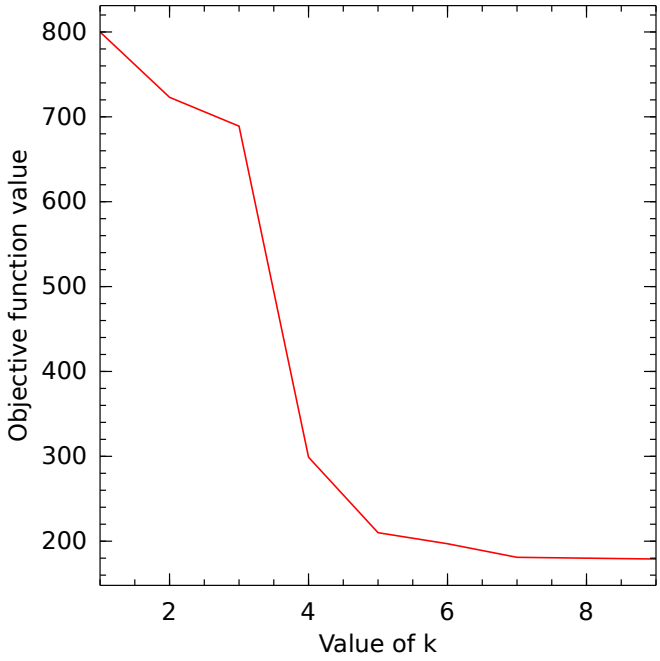
Algorithm 4 วิธีการจัดกลุ่มแบบเคเมดอยส์

1: Randomly choose $m_{1:k}$ ▷ กำหนดค่าตั้งต้นให้เวคเตอร์ค่ามัธยฐาน
2: Initilise z ▷ กำหนดค่าตั้งต้นให้เวคเตอร์บอกกลุ่ม
3: **do**
4: $z_n = \arg\min_{i \in \{1,\dots,k\}} d(x_n, m_i)$ ▷ จัดข้อมูลแต่ละตัวเข้ากลุ่มที่มีค่ามัธยฐานใกล้สุด
5: $m_k = \text{median}(x_n)$ where $z_n = k$ ▷ ปรับค่ามัธยฐานของกลุ่ม
6: **while** z does not change
7: **return** z ▷ เวคเตอร์ z ที่ใช้บอกกลุ่มของข้อมูลแต่ละตัว

6.1.3 การเลือกจำนวนกลุ่มที่เหมาะสม

การเลือกค่า k ที่เหมาะสมอาจเป็นเรื่องที่ไม่ง่ายนัก จนถึงปัจจุบันยังไม่มีวิธีที่สามารถใช้ในการบอกค่า k ที่เหมาะสมได้ แต่ยังมีวิธีที่เป็นวิธีการในเชิงศึกษาสำนึก (Heuristic) อยู่ โดยวิธีดังกล่าวคือการทดลองค่า k ต่างๆ ตั้งแต่ k=1 ไปจนถึงค่าสูงสุดประมาณหนึ่ง แล้วลองสังเกตการเปลี่ยนแปลงของค่าฟังก์ชันจุดประสงค์ (สมการ 6.1) ในที่นี้ ค่า k ที่เหมาะสมจะคือ ค่า k ที่ทำให้ค่าฟังก์ชันจุดประสงค์เปลี่ยนแปลงจาก ค่า k ก่อนหน้ามากที่สุด ตัวอย่างของการพลอตค่าฟังก์ชันจุดประสงค์ต่ำสุดเมื่อกำหนดจำนวนกลุ่ม (ค่า k) ตั้งแต่ 1 ถึง 9 สำหรับชุดข้อมูลกลุ่มหนึ่ง แสดงไว้ในรูปภาพที่ 6.4 จากภาพจะสังเกตได้ว่าเมื่อค่า k=1 ค่าฟังก์ชันจุดประสงค์มีค่าต่ำสุดได้ประมาณ 800 และลดต่ำลงเรื่อยๆ เมื่อค่า k เพิ่มมากขึ้น และค่าฟังก์ชันจุดประสงค์ลดลงมากที่สุดเมื่อค่า k เปลี่ยนจาก 3 เป็น k=4 โดยค่าฟังก์ชันจุดประสงค์ ลดจากประมาณ 690 เป็นประมาณ 300 แต่พอเพิ่ม

ค่า k สูงขึ้นพบว่าความเปลี่ยนแปลงของฟังก์ชันจุดประสงค์มีน้อยมาก จากกรณีดังกล่าว เราจะถือว่าจำนวนกลุ่มที่เป็นธรรมชาติมากที่สุดก็คือ 4 กลุ่ม



รูปภาพ 6.4: การเลือกจำนวนกลุ่มที่เหมาะสมคือจำนวนที่ทำให้ฟังก์ชันจุดประสงค์มีค่าลดลงจากจำนวนกลุ่มก่อนหน้ามากที่สุด ในที่นี้เราพบว่าค่า k=4 เหมาะสมที่สุด

6.2 วิธีการจัดกลุ่มแบบลำดับขั้น

การจัดกลุ่มแบบลำดับขั้น (Hierachical Clustering) เป็นอีกแนวทางหนึ่งในการวิเคราะห์กลุ่มก้อนของข้อมูล การจัดกลุ่มแบบลำดับขั้นต่างจากการจัดกลุ่มแบบแบ่งตรงที่ แนวทางนี้จะมุ่งหาต้นไม้แบบทวิภาคของกลุ่มข้อมูล โดยที่ในแต่ละชั้นของต้นไม้แบบทวิภาค จะเกิดการรวมกันของกลุ่มข้อมูล ที่ใกล้เคียงกันมากที่สุดสองกลุ่มเข้าด้วยกัน กล่าวอีกแบบก็คือ การจัดกลุ่มแบบลำดับขั้นคือการสร้างต้นไม้แบบทวิภาคของกลุ่มข้อมูลนั่นเอง

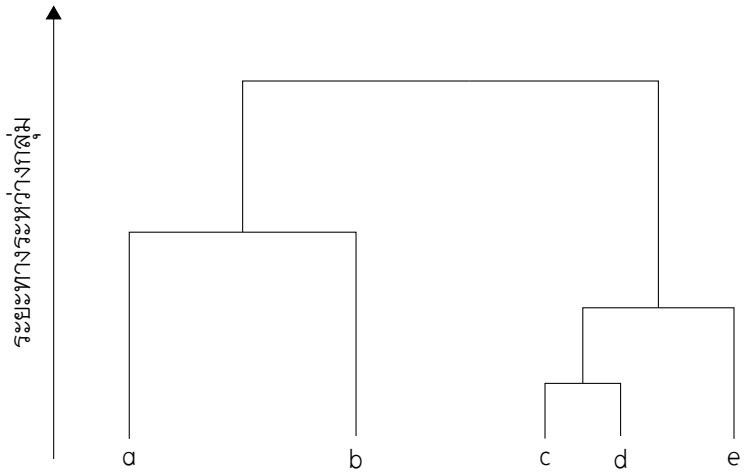
การจัดกลุ่มแบบลำดับขั้นไม่จำเป็นจะต้องเลือกค่า k ที่เหมาะสม ซึ่งต่างจากการจัดกลุ่มโดยวิธีเคมีนส์หรือวิธีเคเมดอยส์ นอกจากนั้นการจัดกลุ่มแบบลำดับขั้น ยังไม่จำเป็นจะต้องคำนวณหาค่ากลางของกลุ่ม สิ่ง

ที่การจัดกลุ่มแบบลำดับขั้นจำเป็นต้องใช้ ก็คือจะต้องมีการนิยามและเลือกวิธีการวัดระยะห่างระหว่างกลุ่มของข้อมูล ให้เหมาะสมเท่านั้น

ในที่นี้เราจะกล่าวถึงการจัดกลุ่มแบบลำดับขั้นที่เรียกว่า แอกโกลเมอเรทีฟคลัสเตอริง (Agglomerative Clustering) ซึ่งมีลักษณะเป็นการสร้างและรวมกลุ่มจากล่างขึ้นบน (Bottom–up) นั่นคือในขั้นเริ่มต้น ข้อมูลทุกตัวจะถือว่าอยู่ในกลุ่มที่มีสมาชิก เพียงหนึ่งตัว คือตัวมันเอง จากนั้นขั้นตอนวิธีจะหากลุ่มสองกลุ่มที่อยู่ใกล้กันมากที่สุด เมื่อวัดโดยมาตรวัดความคล้ายที่กำหนด และรวมกลุ่มสองกลุ่มเข้าด้วยกันเป็นกลุ่มใหญ่ ขั้นตอนวิธีจะทำซ้ำแบบนี้ไปเรื่อยๆ จนถึงขั้นสุดท้ายซึ่งก็คือชั้นบนสุดของต้นไม้ที่กลุ่มต่างๆ ถูกยุบรวมกัน จนเหลือกลุ่มใหญ่อันเดียว

6.2.1 เดนโดรแกรม

ต้นไม้แบบทวิภาคที่ได้จากแอกโกลเมอเรทีฟคลัสเตอริง เรียกว่าเดนโดรแกรม (Dendrogram) ตัวอย่างของเดนโดรแกรมแสดงในภาพที่ 6.5 สังเกตว่าแต่ละครั้งที่เกิดการรวมกันของกลุ่ม ความเหมือนของกลุ่มสองกลุ่มที่ถูกยุบรวมกันจะน้อยลงเรื่อยๆ ไปในทิศทางเดียวกัน (Monotonic) นั่นเอง การสรุปข้อมูลด้วยเดนโดรแกรมทำให้เรา ทราบความคล้ายสัมพัทธ์ของข้อมูลสองกลุ่มใดๆบนเดนโดรแกรม



รูปภาพ 6.5: ตัวอย่างเดนโดรแกรมสำหรับข้อมูล 5 ตัว

จากเดนโดรแกรมข้างต้น เราพบว่าไอเทม c กับไอเทม d มีความคล้ายกันมากกว่าไอเทมอื่นๆในชุดข้อมูล จึงถูกจัดรวมกันก่อนเข้าเป็นกลุ่มที่มีสมาชิก 2 ตัว จากนั้นไอเทม e ก็ถูกรวมเข้ากับกลุ่มของ c และ d กลาย

เป็นกลุ่มที่ใหญ่ขึ้น ในขณะเดียวกันไอเทม a และไอเทม b ก็ถูกรวมเข้าด้วยกันเป็นกลุ่ม ก่อนที่กลุ่มใหญ่ ทั้งสองจะถูกยุบรวมในขั้นสุดท้าย กลายเป็นคลัสเตอร์ขนาดใหญ่คลัสเตอร์เดียว

6.2.2 มาตรวัดความคล้ายระหว่างกลุ่ม

การจัดกลุ่มข้อมูลแบบลำดับขั้น จำเป็นจะต้องมีการนิยามการวัดความคล้ายของกลุ่มสองกลุ่ม ซึ่งจะถูกยุบรวมกัน ในส่วนนี้เราจะมาศึกษาการวัดความคล้ายระหว่างกลุ่มสองกลุ่ม 3 วิธีที่ใช้กันมาก

การเชื่อมโยงแบบเดี่ยว

ความเหมือนหรือความคล้ายระหว่างกลุ่มสองกลุ่มเมื่อวัดด้วยการเชื่อมโยงแบบเดี่ยว (Single Linkage) คือ ระยะทางที่น้อยที่สุดของข้อมูลสองตัวจากทั้งสองกลุ่ม ซึ่งนิยามไว้ว่า

$$d_{SL}(G, H) = \min_{i \in G, j \in H} d(x_i, x_j) \tag{6.3}$$

ในที่นี้ G และ H แทนเซตของข้อมูลในแต่ละกลุ่มตามลำดับ การเชื่อมโยงแบบเดี่ยวสนใจเฉพาะระยะห่างของข้อมูลคู่ที่ใกล้กันที่สุดเท่านั้น ผลข้างเคียงของการใช้วิธีการวัดระยะห่างแบบนี้ คืออาจนำไปสู่สิ่งที่เรียกว่า เชนนิง (Chaining) ที่หมายถึง เหตุการณ์ที่กลุ่มถูกยุบรวมกันเร็วเกินไป เพราะบังเอิญมีข้อมูลสองตัว (ที่อาจเป็นค่าสุดโต่ง) มาอยู่ใกล้กัน ทั้งๆที่ข้อมูลส่วนใหญ่นั้นรวมตัวกันอยู่ห่างออกไปมาก

การเชื่อมโยงแบบสมบูรณ์

ความเหมือนหรือความคล้ายระหว่างกลุ่มสองกลุ่มเมื่อวัดด้วย การเชื่อมโยงแบบสมบูรณ์ (Complate Linkage) คือ ระยะทางที่มากที่สุดระหว่างข้อมูลสองตัวจากทั้งสองกลุ่ม ซึ่งนิยามไว้ว่า

$$d_{CL}(G, H) = \max_{i \in G, j \in H} d(x_i, x_j) \tag{6.4}$$

ผลข้างเคียงของการใช้การเชื่อมโยงแบบสมบูรณ์ ต่างจาก การเชื่อมโยงแบบเดี่ยว ตรงที่ว่ากลุ่มที่ควรจะถูกยุบรวมกัน กลับกลายเป็นไม่ถูกเลือกมารวมกัน เพราะอาจมีค่าสุดโต่งอยู่ในกลุ่มทั้งสอง ซึ่งอยู่ไกลกันมากๆ ทั้งๆที่ข้อมูลส่วนใหญ่เกาะกลุ่มรวมตัวอยู่ใกล้กัน แต่เมื่อวัดด้วยการเชื่อมโยงแบบสมบูรณ์แล้วจะถือว่ากลุ่มทั้งสองอยู่ห่างกันเกินไป

ระยะห่างโดยอาศัยค่าเฉลี่ยของกลุ่ม

ความเหมือนหรือคล้ายระหว่างกลุ่มอาจวัดโดยอาศัยค่าเฉลี่ยของระยะห่าง ของข้อมูลทุกตัวในกลุ่มทั้งสอง แบบพบกันหมด (Group Average) ซึ่งนิยามไว้ว่า

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d(x_i, x_j) \tag{6.5}$$

โดยที่ N_G และ N_H แทนจำนวนสมาชิกในกลุ่ม G และกลุ่ม H ตามลำดับ สังเกตว่าการใช้ค่าเฉลี่ยระยะทางของกลุ่มอาจเป็นวิธีที่ลดปัญหาเชนนิงของการเชื่อมโยงแบบเดี่ยว และปัญหาการรวมกันช้าไปของกลุ่มในการเชื่อมโยงแบบสมบูรณ์ได้ เนื่องจากเป็นการพิจารณาข้อมูลทุกๆตัวในกลุ่ม

การเลือกมาตรวัดความคล้ายระหว่างกลุ่ม มีผลอย่างมากต่อรูปร่างของเดนโดรแกรมผลลัพธ์ ทำให้การแปลผลที่ได้จากการจัดกลุ่มแบบลำดับขั้นยากขึ้นด้วย ปัญหาในการแปลผลที่สำคัญอีกประการของการจัดกลุ่มแบบลำดับขั้น ก็คือ ความจริงที่ว่าในท้ายที่สุดแล้วกลุ่มทุกกลุ่มก็จะถูกรวมกันเป็นกลุ่มใหญ่กลุ่มเดียว ถึงแม้กลุ่มต่างๆเหล่านั้น จะไม่มีความสัมพันธ์กันเลยก็ตาม การแปลความสัมพันธ์ของกลุ่มจึงต้องระวังมากขึ้น ซึ่งปัญหานี้ จะไม่เจอในการจัดกลุ่มแบบแบ่ง เพราะเราสามารถระบุจำนวนกลุ่มที่เราคิดว่าเหมาะสมให้กับขั้นตอนวิธีได้

แบบฝึกหัด

1. เหตุใดจึงเรียกวิธีการจัดกลุ่มข้อมูลว่าเป็นการเรียนรู้แบบไม่มีผู้สอน
2. แนววิธีการจัดกลุ่มข้อมูลแบ่งออกเป็นกี่แนวทาง อะไรบ้าง
3. ปัญหาของการจัดกลุ่มข้อมูลที่วัดระยะห่างระหว่างกลุ่มโดยใช้การเชื่อมโยงแบบสมบูรณ์คืออะไร
4. ในกรณีใดที่การใช้วิธีเคมีนส์อาจไม่เหมาะสม ยกตัวอย่างชุดข้อมูลในกรณีดังกล่าวมาหนึ่งตัวอย่าง
5. จงหาเดนโดรแกรมของข้อมูลต่อไปนี้โดยใช้การเชื่อมโยงแบบเดี่ยวและใช้มาตรวัดระยะห่างแบบแมนฮัตตัน

ชื่อปลา	ความยาว	น้ำหนัก
ปลาคาร์พ	5	10
ปลาแซลมอน	6	19
ปลาทูน่า	8	15
ปลาไหล	12	5
ปลานิล	7	13

เอกสารอ้างอิง

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Cleveland, 1993] Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[Eiben and Smith, 2003] Eiben, A. E. and Smith, J. E. (2003). *Introduction to evolutionary computing*, volume 53. Springer.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874.

[Fawcett and Provost, 1997] Fawcett, T. and Provost, F. (1997). Adaptive fraud detection. *Data mining and knowledge discovery*, 1(3):291–316.

[Fayyad et al., 2002] Fayyad, U. M., Wierse, A., and Grinstein, G. G. (2002). *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann.

[Friedman, 1997] Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

[Han et al., 2011] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[Hoffman and Grinstein, 2002] Hoffman, P. E. and Grinstein, G. G. (2002). A survey of visualizations for high-dimensional data mining. *Information visualization in data mining and knowledge discovery*, pages 47–82.

[Indyk and Motwani, 1998] Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM.

[Jindal and Liu, 2007] Jindal, N. and Liu, B. (2007). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM.

[Johnson et al., 1986] Johnson, W. B., Lindenstrauss, J., and Schechtman, G. (1986). Extensions of lipschitz maps into banach spaces. *Israel Journal of Mathematics*, 54(2):129–138.

[Ke and Sukthankar, 2004] Ke, Y. and Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE.

[Lillesand et al., 2014] Lillesand, T., Kiefer, R. W., and Chipman, J. (2014). *Remote sensing and image interpretation*. John Wiley & Sons.

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Manjunath et al., 2001] Manjunath, B. S., Ohm, J.-R., Vasudevan, V. V., and Yamada, A. (2001). Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):703–715.

[Ng, 2004] Ng, A. Y. (2004). Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM.

[Ng and Jordan, 2002] Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, pages 841–848.

[Ojala et al., 2002] Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray–scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1–2):1–135.

[Richards, 1999] Richards, J. A. (1999). *Remote sensing digital image analysis*, volume 3. Springer.

[Sonka et al., 2014] Sonka, M., Hlavac, V., and Boyle, R. (2014). *Image processing, analysis, and machine vision*. Cengage Learning.

[Vapnik and Vapnik, 1998] Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.

[Wasserman, 2013] Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.