

ปฏิบัติการที่ 7

การใช้เครื่องมือวิเคราะห์การถดถอย

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์การถดถอยของข้อมูลได้
2. เพื่อให้สามารถใช้เครื่องมือวัดประสิทธิภาพของวิธีการวิเคราะห์การถดถอยได้

ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Student Marks (สำหรับการสาธิต)
- ชุดข้อมูล Real Estate (สำหรับการฝึกปฏิบัติการ)

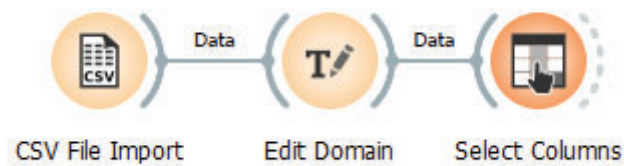
ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

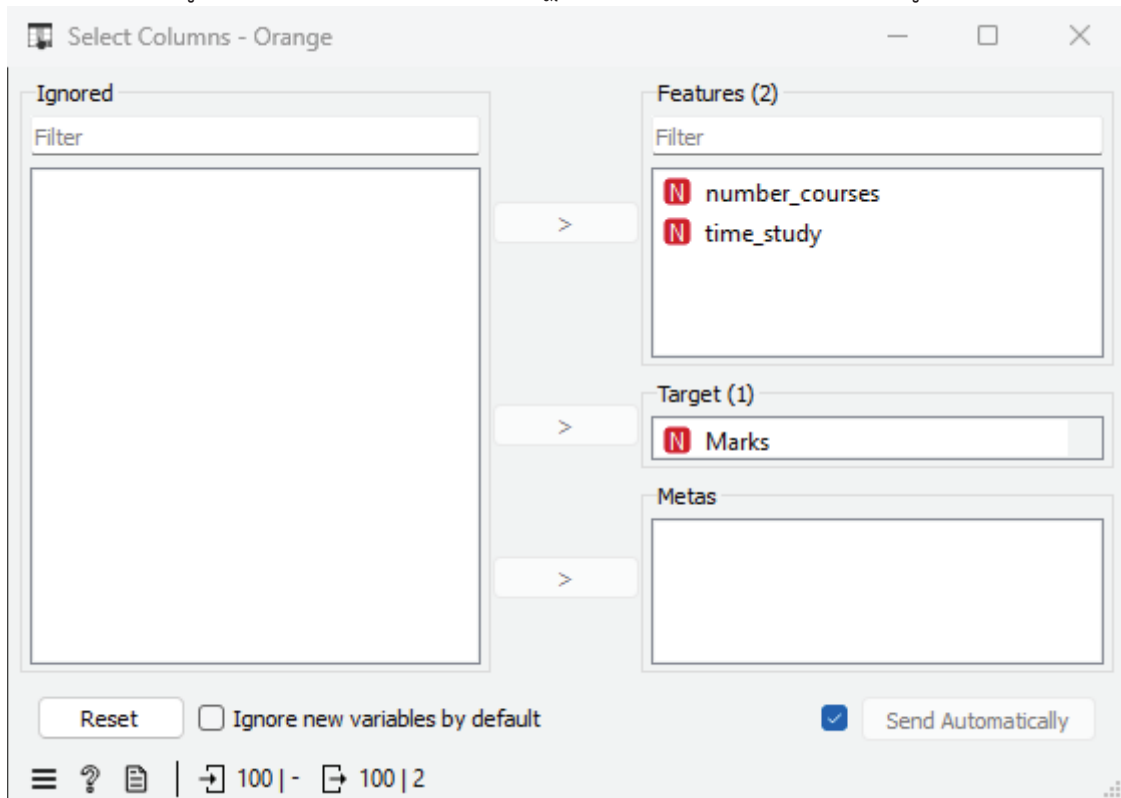
1. เปิดโปรแกรม Orange
2. ทำการบันทึก workspace โดยไปที่เมนู File เลือก Save จากนั้นทำการตั้งชื่อไฟล์ในรูปแบบ Practice_07_id.ows โดยแทน id ด้วยรหัสนักศึกษา แล้วกดปุ่ม Save
3. นำชุดข้อมูลจากแฟ้มข้อมูล Student_Marks.csv เข้าสู่โปรแกรม Orange โดยใช้โมดูล CSV File Import
4. ตรวจสอบและเปลี่ยนชนิดข้อมูลของตัวแปร โดยใช้โมดูล Edit Domain (ดูปฏิบัติการที่ 2) กำหนดให้แต่ละตัวแปรต้องมีชนิดข้อมูล ดังนี้

ตัวแปร	ชนิดข้อมูล
number_courses	Numeric Feature
time_study	Numeric Feature
Marks	Numeric Feature

5. ใช้โมดูล Select Columns ในการเลือกตัวแปรต้นและตัวแปรตาม โดยคลิกเลือกโมดูล Select Columns จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Edit Domain (นำผลลัพธ์จากโมดูล Edit Domain ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล Select Columns ด้าน input ดังรูป



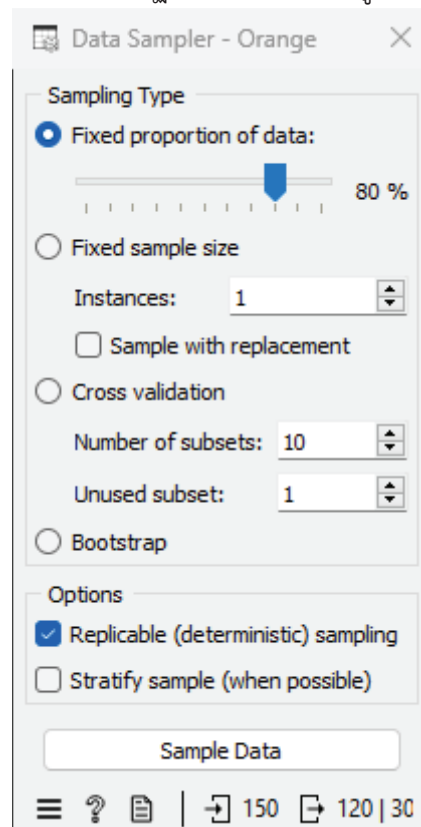
6. ดับเบิลคลิกที่โมดูล Select Columns จากปรากฏหน้าตาต่างสำหรับเลือกตัวแปร ดังรูป



7. เลือกตัวแปร number_courses และ petal_width ที่จะใช้เป็นตัวแปรต้นในการวิเคราะห์ไว้ในส่วน Features และเลือกตัวแปรเป้าหมายหรือตัวแปรตามไว้ในส่วน Target ในที่นี้ คือ ตัวแปร Marks
8. ต่อมาทำการแบ่งชุดข้อมูล iris ออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้ (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) โดยใช้โมดูล Data Sampler คลิกเลือกโมดูล Data Sampler จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Select Columns (นำผลลัพธ์จากโมดูล Select Columns ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล Data Sampler ด้าน input ดังรูป



9. ดับเบิลคลิกที่โมดูล Data Sampler จากปรากฏหน้าต่างตั้งค่า ดังรูป



10. โมดูล Data Sampler สามารถสุ่มแบ่งข้อมูลออกเป็น 2 ชุดข้อมูลได้ 3 วิธี ในที่นี้จะใช้วิธีการกำหนดสัดส่วนระหว่างข้อมูลทั้ง 2 ชุด โดยเลือกตัวเลือก Fixed proportion of data และกำหนดสัดส่วนเท่ากับ 80% นั่นคือ ทำการสุ่มชุดข้อมูลแรกจำนวนร้อยละ 80 ของข้อมูลทั้งหมด และข้อมูลที่เหลืออีกร้อยละ 20 ของข้อมูลทั้งหมด จะเป็นข้อมูลชุดที่สอง

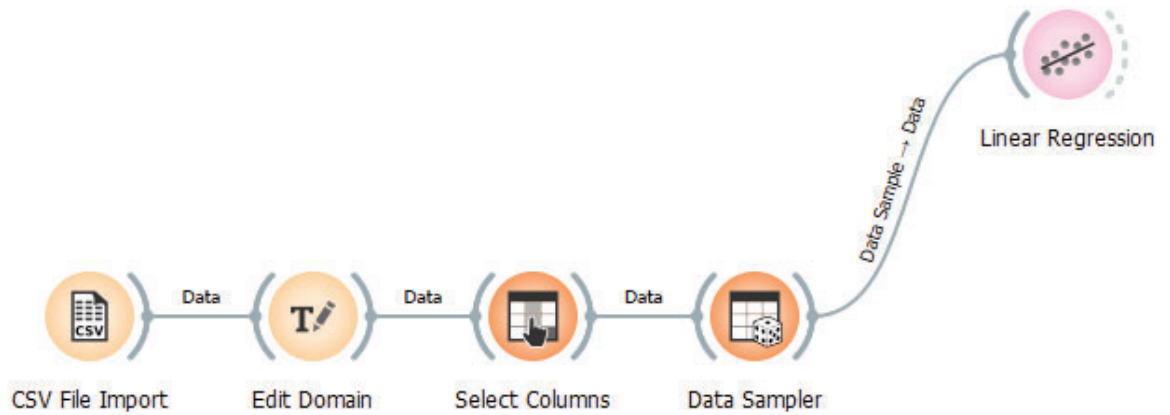
11. คลิกปุ่ม Sample Data

12. ในปฏิบัติการนี้จะสร้างแบบจำลองการวิเคราะห์การถดถอยเชิงเส้นตรง (Linear Regression Model) เพื่อทำนายคะแนนสอบของนักเรียน (ตัวแปร Masks) จากจำนวนกระบวนวิชาที่เรียน (ตัวแปร number_courses) และเวลาที่ใช้ในการเรียน (ตัวแปร time_study)

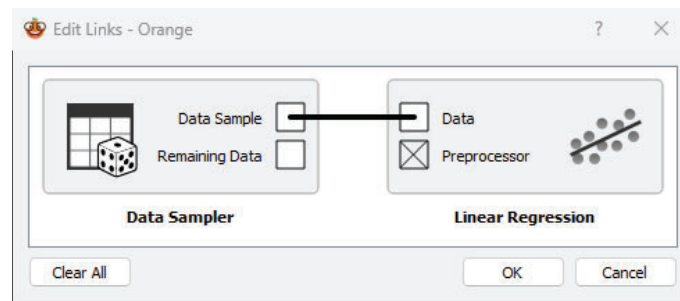
$$Masks = \beta_0 + (\beta_1 \times number_courses) + (\beta_2 \times time_study)$$

13. คลิกเลือกโมดูล Linear Regression ตามลำดับ จะปรากฏโมดูลใน workspace

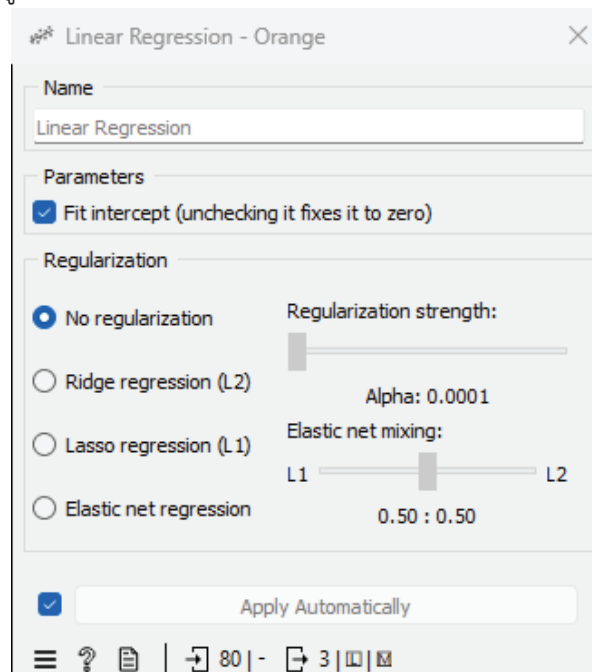
14. ทำการสร้างแบบจำลอง โดยคลิกเชื่อมโมดูลโมดูล Data Sampler จากด้าน output เข้าสู่โมดูล Linear Regression ด้าน input ดังรูป




15. จากนั้น ดับเบิลคลิกที่เส้นเชื่อมระหว่างโมดูล Data Sampler และ Linear Regression จากปรากฏหน้าต่างดังรูปด้านล่าง ให้ลากเส้นเชื่อมระหว่าง Data Sample และ Data เป็นการกำหนดให้ข้อมูลที่ถูกลสุ่มชุดแรกเป็นข้อมูลสำหรับนำไปใช้สอนแบบจำลอง



16. คลิกปุ่ม OK
17. กำหนดค่าไฮเปอร์พารามิเตอร์ของ Linear Regression ดับเบิลคลิกที่โมดูล Linear Regression จากปรากฏหน้าต่างตั้งค่า ดังรูป



18. ในที่นี่จะใช้แบบจำลอง Linear Regression โดยกำหนดให้ใช้พจน์ของ intercept นั่นคือ พารามิเตอร์ β_0 ในแบบจำลอง และไม่ทำการ Regularization ใดๆ
19. จากนั้น คลิกปุ่ม  เพื่อดูแบบจำลองที่ได้จากการสอน จะปรากฏหน้าต่างแสดงค่าพารามิเตอร์ต่างๆ ของแบบจำลอง ดังรูป

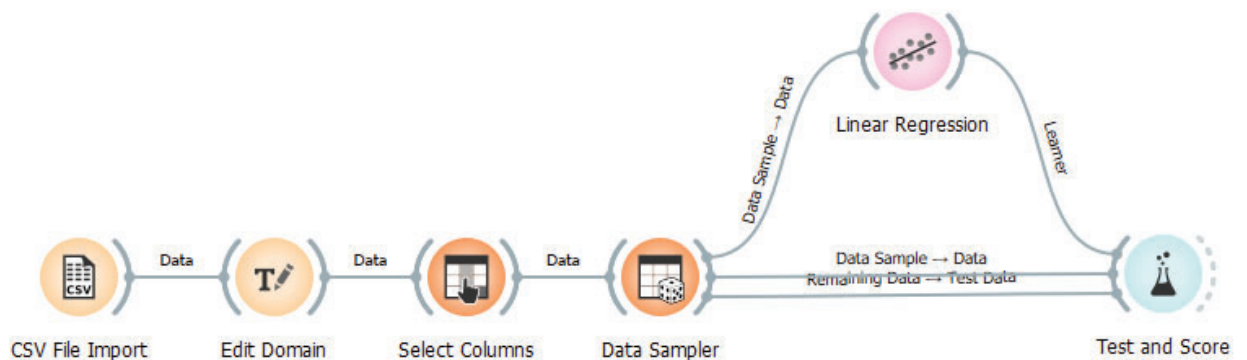
Coefficients: coefficients: 3 instances, 2 variables
Features: numeric (no missing values)
Metas: string

	name	coef
1	intercept	-8.10945
2	number_courses	1.94721
3	time_study	5.46312

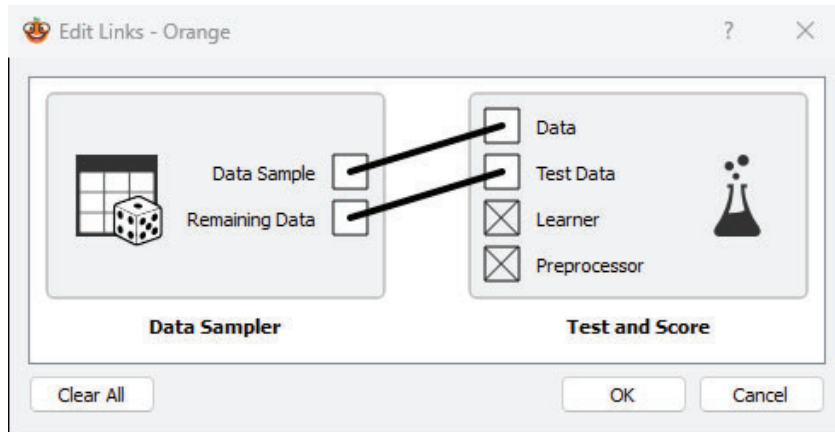
จากภาพ จะสามารถเขียนแบบจำลองได้ ดังนี้

$$Masks = -8.10945 + (1.94721 \times number_courses) + (5.46312 \times time_study)$$

20. คลิกเลือกโมดูล Test and Score จะปรากฏโมดูลใน workspace เพื่อทำการสร้างและทดสอบประสิทธิภาพของแต่ละแบบจำลอง จากนั้นคลิกเชื่อมโมดูล Linear Regression และ Data Sampler จากด้าน output เข้าสู่โมดูล Test and Score ด้าน input ดังรูป



21. จากนั้น ดับเบิลคลิกที่เส้นเชื่อมระหว่างโมดูล Data Sampler และ Test and Score จากปรากฏหน้าต่าง ดังรูปด้านล่าง ให้ลากเส้นเชื่อมระหว่าง Data Sample และ Data และเส้นเชื่อมระหว่าง Remaining Data และ Test Data เป็นการกำหนดให้ข้อมูลที่ถูกสุ่มชุดแรกเป็นข้อมูลสำหรับนำไปใช้สอนแบบจำลอง และข้อมูลที่เหลือจากการสุ่มนำไปใช้ทดสอบแบบจำลอง



22. คลิกปุ่ม OK

23. ดับเบิลคลิกที่โมดูล Test and Score จะปรากฏหน้าต่าง ผลการวัดประสิทธิภาพของแบบจำลอง ดังรูป

Model	MSE	RMSE	MAE	MAPE	R2
Linear Regression	12.665	3.559	3.285	0.232	0.912

Compare models by: Mean sq Negligible diff.: 0.1

Linear Regression	Linear Re...
-------------------	--------------

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

24. ในที่นี้ให้เลือกตัวเลือก Test on test data เพื่อนำเฉพาะข้อมูลในชุดข้อมูลทดสอบมาใช้ในการประเมินประสิทธิภาพเท่านั้น ทางด้านขวาของหน้าต่าง คือ ผลการประเมินประสิทธิภาพที่ได้ของแต่ละแบบจำลอง

แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำชุดข้อมูล Real Estate จากแฟ้มข้อมูล Real_Estate.csv เข้าสู่โปรแกรม Orange
2. ทำการเปลี่ยนชนิดข้อมูลของตัวแปรทั้งหมด ให้เป็นชนิดข้อมูล Numeric Feature
3. กำหนดตัวแปรทั้งหมด ยกเว้นตัวแปร House_price เป็นตัวแปรต้น และ กำหนดตัวแปร House_price เป็นตัวแปรตาม
4. แบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ในอัตราส่วน 70 ต่อ 30
5. สร้างแบบจำลองการวิเคราะห์การถดถอยเชิงเส้นตรง (Linear Regression Model) เพื่อทำนายราคาบ้าน (ตัวแปร House_price) จากอายุของบ้าน (ตัวแปร House_age) ระยะห่างจากสถานีรถไฟฟ้าที่ใกล้ที่สุด (ตัวแปร Distance_to_the_nearest_MRT_station) จำนวนร้านค้าสะดวกซื้อ (ตัวแปร Number_of_convenience_stores) พิกัดที่ตั้งละติจูด (ตัวแปร Latitude) และพิกัดที่ตั้งลองจิจูด (ตัวแปร Longitude)

$$\begin{aligned} \text{House_price} = & \beta_0 + (\beta_1 \times \text{House_age}) \\ & + (\beta_2 \times \text{Distance_to_the_nearest_MRT_station}) \\ & + (\beta_3 \times \text{Number_of_convenience_stores}) + (\beta_4 \times \text{Latitude}) \\ & + (\beta_5 \times \text{Longitude}) \end{aligned}$$

6. สร้างและทดสอบประสิทธิภาพของแบบจำลองทั้งหมด
7. ศึกษาแบบจำลองที่ได้และอธิบายผลการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูล
8. **สิ่งที่ต้องส่งเป็นการบ้าน** ทำการบันทึกไฟล์ workspace ของนักศึกษา โดยตั้งชื่อไฟล์ในรูปแบบ Lab_07_id.ows โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>