CS381: Numerical Computation & Softwares

Linear Least Square Problems

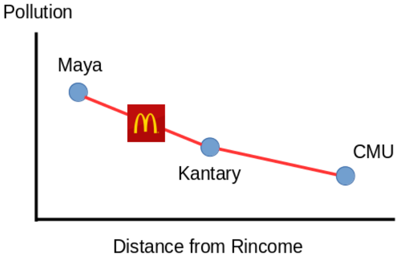
Jakramate Bootkrajang

Department of Computer Science

Chiang Mai University

Motivation

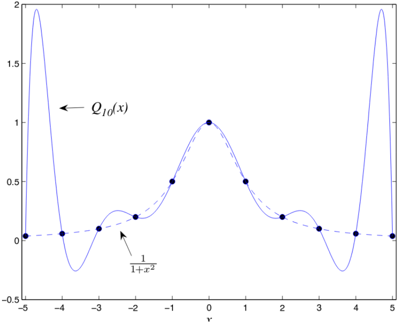
- ย้อนกลับไปยังปัญหาการประมาณค่าฟังก์ชัน ณ จุดที่ไม่มีข้อมูล
- ข้อมูลที่มีอยู่ในรูป $(x_i, y_i = f(x_i))$, $i = 0, \ldots, N$



- วิธีที่สามารถนำมาใช้แก้ปัญหานี้ได้คือการทำ interpolation แล้วประมาณ y_i ณ จุด x ที่สนใจ

Motivation

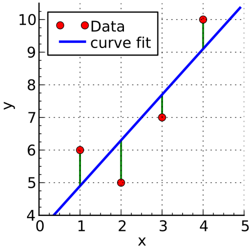
- Interpolation มีสมมติฐานว่าฟังก์ชัน $f()$ แสดงได้ด้วยพหุนามดีกรี $N-1$ ซึ่งอาจถูกหรือผิดก็ได้
- บางครั้งพหุนามที่หาได้อาจมีความซับซ้อนมากกว่า ความสัมพันธ์ของ x_i และ y_i จริง



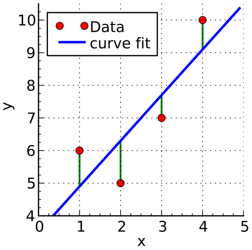
Motivation

- นอกจากการสมมติว่า $f()$ มีลักษณะเป็นพหุนามแล้ว เราสามารถตั้งสมมติฐานว่าฟังก์ชัน $f()$ มีลักษณะเป็นเชิงเส้นได้

- แน่นอนว่า การอธิบายข้อมูล $(x_i, y_i = f(x_i))$, $i = 0, \ldots, N$ ด้วยฟังก์ชันเชิงเส้นมีข้อจำกัด คือ เป็นการยากที่เส้นตรงดังกล่าวจะลากผ่านทุกจุดข้อมูล



Motivation



- ระยะห่างระหว่างค่าประมาณกับค่าจริง (เส้นสีเขียว) ของค่า x_i ใด ถือว่าเป็น error ที่เกิดขึ้นจากข้อจำกัดของฟังก์ชันเชิงเส้น (mathematical error)
- เป้าหมายของการอธิบายชุดข้อมูลด้วยฟังก์ชันเชิงเส้นคือการหาฟังก์ชัน $f(x) = ax$ ที่ทำให้เกิดค่า error รวมบนชุดข้อมูลน้อยที่สุด
- ส่วนมากแล้ว เรามักให้ค่า error รวมแทนด้วยผลรวมของค่าความคาดเคลื่อนยกกำลังสอง

Sum of square error

- ค่าความคาดเคลื่อนของการประมาณ y_i ของ x_i แต่ละตัวสามารถคิดจาก

$$\epsilon_i = y_i - f(x_i) \tag{1}$$
$$= y_i - ax_i \tag{2}$$

- เนื่องจากค่าประมาณ (ax_i) อาจมากกว่าหรือน้อยกว่า y_i ทำให้ ϵ_i เป็นได้ทั้งบวกและลบ เมื่อนำ ϵ_i มาบวกกันอาจเกิดการหักล้างกัน

- จึงทำการแปลงค่าความคาดเคลื่อนให้เป็นบวกเหมือนกันทั้งหมดโดยการยกกำลังสอง

$$\epsilon_i = (y_i - ax_i)^2 \tag{3}$$

- หากรวมค่าความคาดเคลื่อนของทุกๆ x_i เข้าด้วยกันจะได้ว่า

$$\epsilon = \sum_{i=0}^{N} (y_i - ax_i)^2 \tag{4}$$

In matrix form

- เราสามารถเขียนสมการความคาดเคลื่อนกำลังสองรวมในรูปของเมทริกซ์ได้

$$\epsilon = ||X\vec{a} - Y||_2^2 \tag{5}$$

- X คือเมทริกซ์ของข้อมูล มีขนาด N คูณ M (N จำนวนข้อมูล M มิติของข้อมูล)
- Y คือเมทริกซ์ของ response
- $\vec{a}$ คือเวคเตอร์สัมประสิทธิ์ของฟังก์ชันเชิงเส้นที่ต้องตามหา

In matrix form: Example

- กำหนดชุดข้อมูล $\{(3,14),(2,13),(5,19)\}$

$$X = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}, \vec{a} = \begin{bmatrix} a_1 \end{bmatrix}, Y = \begin{bmatrix} 14 \\ 13 \\ 19 \end{bmatrix} \quad (6)$$

- หรือ

$$X = \begin{bmatrix} 3 & 2 \\ 2 & 1 \\ 5 & 4 \end{bmatrix}, \vec{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, Y = \begin{bmatrix} 3.3 \\ 1.5 \\ 7.9 \end{bmatrix} \quad (7)$$

- สังเกตว่าปัญหาระบบสมการเชิงเส้นที่ไม่มีคำตอบ สามารถมองเป็นปัญหา least square ได้

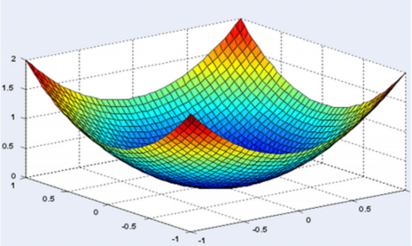
In matrix form: Example

- เพิ่มตัวอย่าง regression จริง

- เป้าหมายคือการหา $\vec{a}$ ที่ทำให้ค่า error มีค่าน้อยที่สุด โดยการ minimize ฟังก์ชันจุดประสงค์ (objective function)

$$L = \min_{\vec{a}} ||X\vec{a} - Y||_2^2 \tag{8}$$

- ฟังก์ชันจุดประสงค์ดังกล่าวเป็นฟังก์ชันแบบรูปเว้า (convex function) ที่มีจุดต่ำสุด



Where the best $\vec{a}$ is ?

- จากวิชา calculus เราทราบว่าจุดต่ำสุดของฟังก์ชันจุดประสงค์ดังกล่าวคือจุดที่อนุพันธ์ของฟังก์ชันมีค่าเท่ากับ 0 หรือ

$$
\begin{aligned}
\frac{\partial L}{\partial \vec{a}} &= \frac{\partial ||X\vec{a} - Y||_2^2}{\partial \vec{a}} && (9) \\
&= 2(X\vec{a} - Y)X && (10)
\end{aligned}
$$

- ต่อไปหาค่า $\vec{a}$ เมื่อสมการที่ (10) มีค่าเท่ากับ 0

$$
\begin{aligned}
(X\vec{a} - Y)X = 0 && (11) \\
X^T X\vec{a} - X^T Y = 0 && (12) \\
\vec{a} = (X^T X)^{-1} X^T Y && (13)
\end{aligned}
$$

Computing $\vec{a}$ is expensive

$$\vec{a} = (X^T X)^{-1} X^T Y$$

- Closed-form solution ของ $\vec{a}$ มีการหา inverse ของเมทริกซ์ $(X^T X)$ ซึ่งช้า หาก X มีจำนวน column มาก
- ในทางปฏิบัติคือกรณีที่ข้อมูลมี feature มาก
- การหาคำตอบของ least square problem อาจพิจารณาใช้วิธีอื่นที่เลี่ยงการคำนวณ inverse โดยตรง

Matrix decomposition

$$\vec{a} = (X^T X)^{-1} X^T Y$$

- แนวทางหนึ่งคือการพยายามแยกส่วน (decompose) เมทริกซ์ $(X^T X)$ ให้อยู่ในรูปของการคูณกันของเมทริกซ์อื่นที่เลี่ยงการหา inverse
- Cholesky decomposition ทำการ decompose $(X^T X) = R^T R$
- Singular Value Decomposition ทำการ decompose $(X^T X) = U \Sigma V^T$

Cholesky decomposition

$$X^T X = R^T R$$

- Cholesky decomposition ทำการ decompose $(X^T X) = R^T R$
- Example

$$\begin{bmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{bmatrix} \times \begin{bmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{bmatrix} \tag{14}$$

- The Cholesky–Banachiewicz / Cholesky–Crout algorithms คำนวณย้อนกลับ

$$A = R^T R = \begin{bmatrix} R_{11} & 0 & 0 \\ R_{21} & R_{22} & 0 \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \times \begin{bmatrix} R_{11} & R_{21} & R_{31} \\ 0 & R_{22} & R_{32} \\ 0 & 0 & R_{33} \end{bmatrix} \tag{15}$$

$$= \begin{bmatrix} R_{11}^2 & \cdots & \text{symmetric} \\ R_{21}R_{11} & R_{21}^2 + R_{22}^2 & \cdots \\ R_{31}R_{11} & R_{31}R_{21} + R_{32}R_{22} & R_{11}^2 + R_{22}^2 + R_{33}^2 \end{bmatrix} \tag{16}$$

- ดังนั้นแล้ว

$$R = \begin{bmatrix} \sqrt{A_{11}} & 0 & 0 \\ A_{21}/R_{11} & \sqrt{A_{22} - R_{21}^2} & 0 \\ A_{31}/R_{11} & (A_{32} - R_{31}R_{21})/R_{22} & \sqrt{A_{33} - R_{31}^2 - R_{32}^2} \end{bmatrix} \tag{17}$$

The patterns emerged

- สังเกตรูปแบบของพจน์ต่างๆ

$$R = \begin{bmatrix} \sqrt{A_{11}} & 0 & 0 \\ A_{21}/R_{11} & \sqrt{A_{22} - R_{21}^2} & 0 \\ A_{31}/R_{11} & (A_{32} - R_{31}R_{21})/R_{22} & \sqrt{A_{33} - R_{31}^2 - R_{32}^2} \end{bmatrix} \tag{18}$$

- Pattern คือ

$$L_{jj} = \sqrt{A_{jj} - \sum_{k=1}^{j-1} L_{jk}^2} \tag{19}$$

$$L_{ij} = \frac{1}{L_{jj}}(A_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk}) \text{ for } \; i > j \tag{20}$$

Cholesky least square algorithm

$$X^T X\vec{\alpha} = X^T Y \tag{21}$$

1. Compute $X^T X$ and $X^T Y$
2. Factorise $X^T X \rightarrow R^T R$
3. Solve lower triangular system $R^T w = X^T Y$ for w
4. Solve upper triangular system $R\vec{\alpha} = w$ for $\vec{\alpha}$

Once we found $\vec{\alpha}$

- เราสามารถใช้ในการทำนายค่า y ของ x ที่ไม่เคยเจอมาก่อน
- โดย $\hat{y} = x\vec{\alpha}$

References

- Numerical Linear Algebra by Greg Fasshauer

  http://www.math.iit.edu/~fass/477577_Chapter_5.pdf