Shrinkage Methods (Ridge and Lasso)

Petchara Pattarakijwanich

Introduction to Data Science, 28 October 2022

Goal of this week

- Subset Selection
  - Best Subset Selection
  - Forward/Backward Stepwise Subset Selection
- Shrinkage Methods
  - Ridge Regression
  - The Lasso

About the Final Project

Approaches

- Answer specific questions from data, using data science techniques
- Implement, invent, or test some data science techniques or algorithms

Data Source

- Your own research
- Publicly available datasets
  https://geekflare.com/open-datasets-for-data-science/
  https://www.dataquest.io/blog/free-datasets-for-projects/
- Simulated data (?)

Presentation (some time around final week or a bit later)

Model Selection

เพื่อใช้หาโมเดลที่ดีที่สุด

$$AIC = \frac{1}{\sigma^2}\left[MSE + \frac{2d\sigma^2}{n}\right]$$

$$BIC = \frac{1}{\sigma^2}\left[MSE + \frac{d\sigma^2 \ln n}{n}\right]$$

น้อย = โมเดลดี

ฟังก์ชันแต่ใหญ่

ราคาของจำนวนตัวแปร

Subset Selection

ตัวอย่าง Input $\{x_1, x_2, x_3\}$ output y

[สมมติว่า y ขึ้นกับ x_1 อย่างเดียว ไม่ขึ้นกับ x_2, x_3]

Linear Regression โดยเลือก subset ของ Input

จะได้โมเดลที่เป็นไปได้

การลดตัวแปร 1 ตัวขึ้นไป

- 0 ตัวแปร $y = \beta_0$
- 1 ตัวแปร $y = \beta_0 + \beta_1 x_1$, $y = \beta_0 + \beta_2 x_2$, $y = \beta_0 + \beta_3 x_3$
- 2 ตัวแปร $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, ... , ...
- 3 ตัวแปร $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Best Subset Selection

รู้ได้ไงว่า $y = \beta_0 + \beta_1 x_1$ ดีสุด

- ฟิตทั้ง 8 โมเดล ⇒ คำนวณ AIC, BIC ⇒ น้อยสุด = ดีสุด
  (หรือ test data)

[ถ้ามี p ตัวแปร จำนวนโมเดลที่ต้องฟิต $= 2^p$]

⇒ ฟิตหมดไม่ได้ ถ้า p มากพอ

[Brute force]

Forward Step-wise Subset Selection

0 parameter $\boxed{y = \beta_0}$

1 parameter $y \propto x_1$, $y \propto x_2$, $y \propto x_2$

- มีทั้งหมด 3 ตัว แต่ละตัวมีค่าความผิดพลาด
- เลือก $\boxed{y \propto x_1}$ ดีสุด

2 parameter $y \propto (x_1, x_2)$, $y \propto (x_1, x_3)$

- มีทั้งหมด 2 ตัว (ไม่มี $y \propto (x_2, x_3)$)
- เลือก $\boxed{y \propto (x_1, x_2)}$ ดีสุด

3 parameter $\boxed{y \propto (x_1, x_2, x_3)}$

$\Rightarrow$ เทียบ 4 โมเดลด้วย AIC, BIC $\Rightarrow$ โมเดลดีสุด

Forward Step-wise Subset Selection

กรณี p ตัวแปร

0 parameter : 1 โมเดล

1 parameter : p โมเดล ⇒ เก็บอันดีสุดไว้เป็นตัวแทน

2 parameters : $p-1$ โมเดล ⇒ เก็บอันดีสุดไว้เป็นตัวแทน

3 parameters : $p-2$ โมเดล

$\vdots$

$p-1$ parameters : 2 โมเดล

p parameters : 1 โมเดล

↑ Linear Regression ตรงๆ

จำนวนโมเดลทั้งหมด

$= 1 + p + (p-1) + (p-2) + \dots + 2 + 1$

$= 1 + \frac{1}{2} p(p+1)$

$\propto p^2$

Backward Step-wise Subset Selection

คล้าย Forward เริ่มจาก p ตัวแปร ลดทีละ 1

Ridge Regression

↑ Regularization technique

LR : Minimize $RSS = \sum (y - y_{model})^2$

$= \sum (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots)^2$

$$RSS = \sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2$$

Ridge Regression :

Minimize $\sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2 + \lambda \sum_i \beta_i^2$

RSS

penalty

ค่าของ hyperparameter

$= \beta_1^2 + \beta_2^2 + \cdots$

$=$ two-norm

Ridge Regression

Ridge Regression :

Minimize $\sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2 + \lambda \sum_i \beta_i^2$

ตัวแปรเลือกได้ $= \beta_1^2 + \beta_2^2 + \dots$ $= \text{two-norm}$

RSS

penalty

$\Rightarrow$ λ เป็นบวก เป็น tuning parameter

$\Rightarrow$ Minimize เพื่อ
- ให้แบบจำลองดีขึ้น (RSS น้อย)
- β_i มีค่าไม่มาก ($\sum_i \beta_i^2$ น้อย)

$\Rightarrow$ Shrinkage method

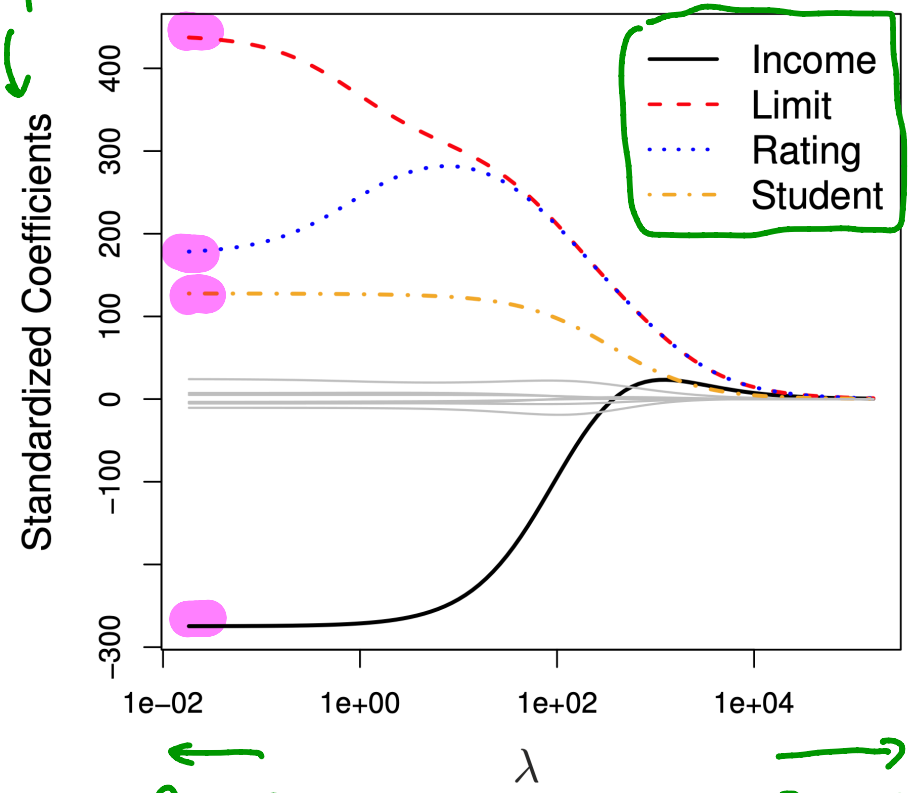
(Regularization)

Ridge Regression

$\|\beta\|_2$ = two-norm
$= \sum_i \beta_i^2$

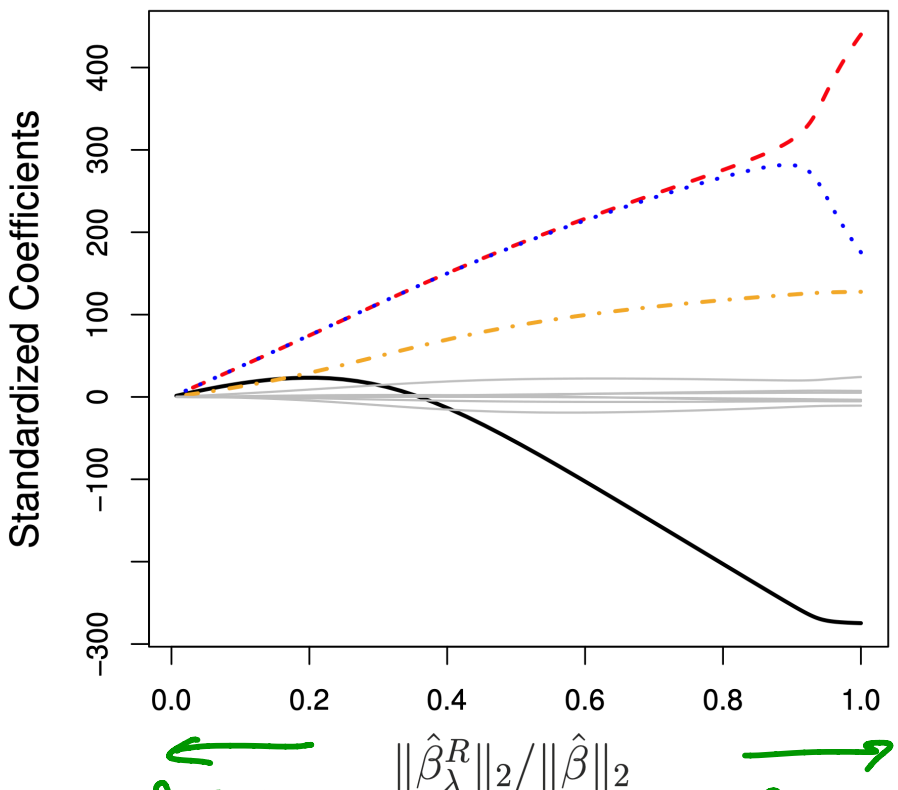
β_i

input x_i



$\lambda \to 0$
เหมือนกับ LR

$\lambda \to \infty$
$\beta_i = 0$
$y = \beta_0 = \bar{y}$



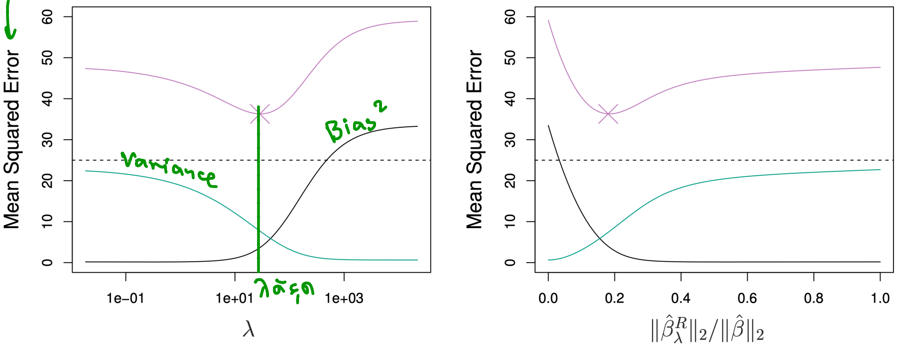
$\lambda \to \infty$
$\beta_i = 0$

$= \dfrac{(\sum \beta_i^2)_{Ridge}}{(\sum \beta_i^2)_{LR}}$

$\lambda \to 0$
LR

Ridge Regression

ของ test data



หา λ ดีสุด ⇒ ทดสอบกับ test data

λ เพิ่มขึ้น ⇒ โมเดล stiff → Bias มากขึ้น, → Variance น้อยลง ⇒ Bias Variance trade off

The Lasso

LR : Minimize $RSS = \sum (y - y_{model})^2$

$$= \sum (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots)^2$$

$$\boxed{RSS = \sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2}$$

Lasso :

Minimize

$$\boxed{\sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2 + \lambda \sum_i |\beta_i|}$$

ตัวแปรอิสระ

$= |\beta_1| + |\beta_2| + \cdots$

$=$ one-norm

RSS

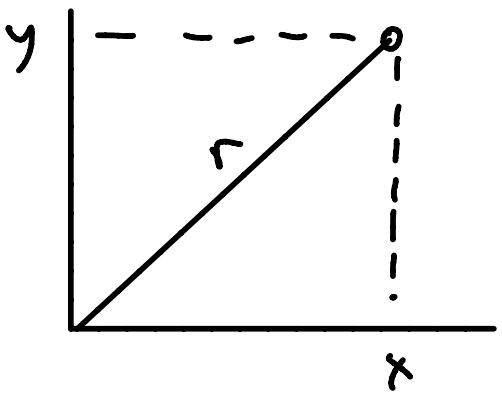
penalty

The Lasso

Two-norm $\|x\|_2$ vs One-norm $\|x\|_1$

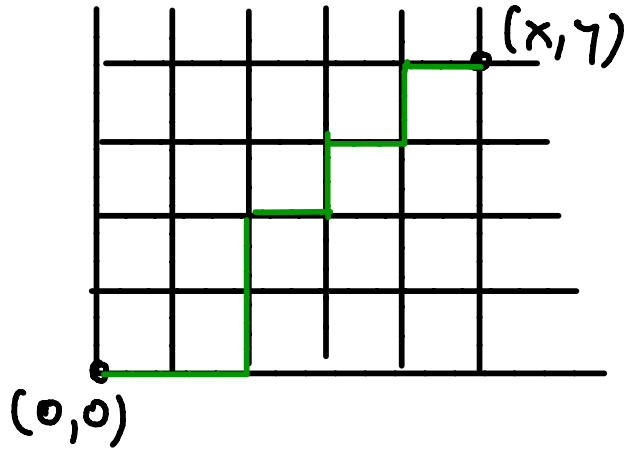
ระยะทางปกติ 2 มิติ
(Euclidean distance)

$$r^2 = x^2 + y^2 = \|x\|_2$$

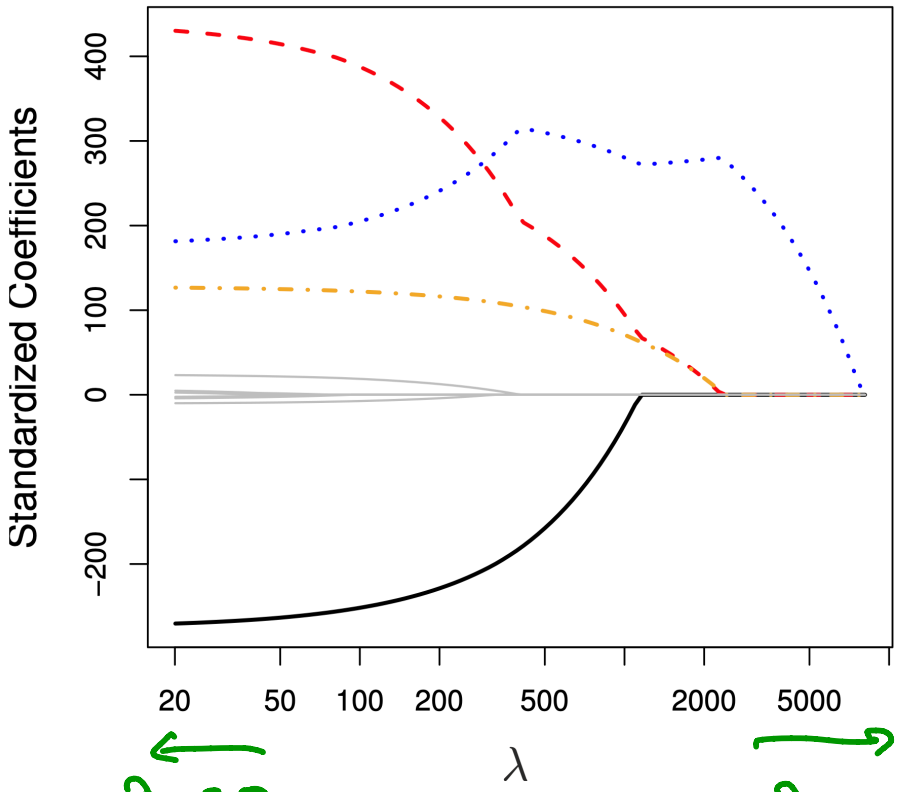


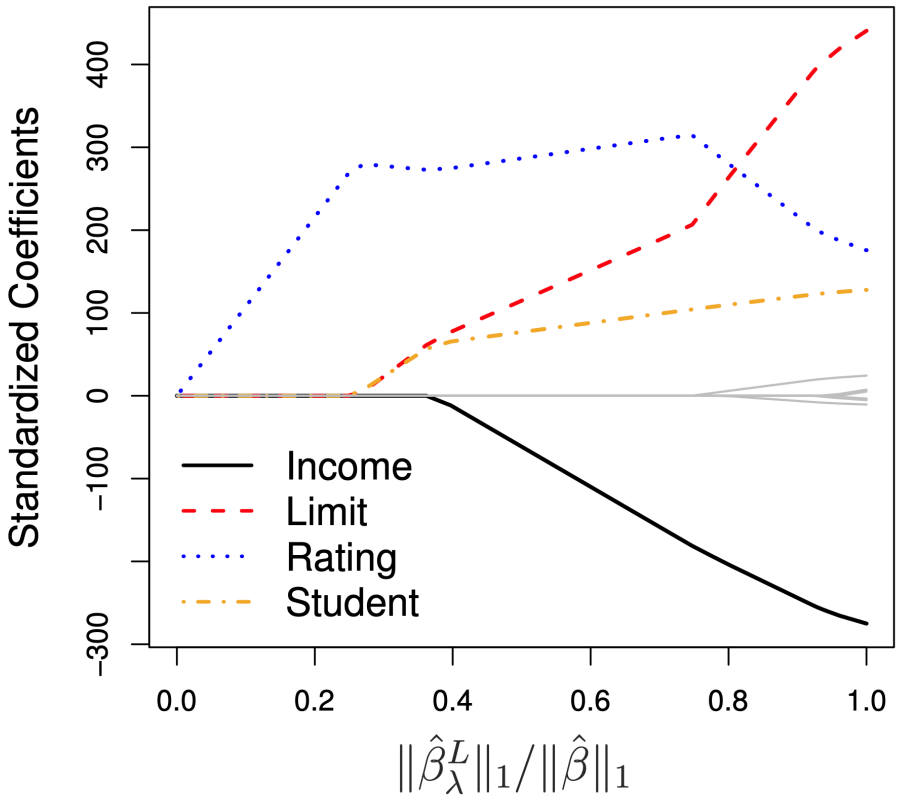
ระยะทาง 2 มิติ
(Taxi-driver distance)

$$r = |x| + |y| = \|x\|_1$$



The Lasso





$\lambda \to 0$ เป็น LR

$\lambda \to \infty$

$\beta_i = 0$

$y = \beta_0 = \bar{y}$

Parameter Selection : เมื่อ λ มากขึ้นจะทำให้ β_i บางตัว $= 0$ อัตโนมัติ

The Lasso

LR : Minimize $\text{RSS} = \sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2$

Ridge : Minimize $\sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2 + \lambda \sum_i \beta_i^2$

Lasso : Minimize $\sum \left(y - \beta_0 - \sum_i \beta_i x_i \right)^2 + \lambda \sum_i |\beta_i|$

LR ถ้า(หน่วย)ของ x_i ไม่เท่ากัน ⇒ ถ้า β_i ไม่เท่ากันถ้า x_i

⇒ Ridge & Lasso ถ้า x_i ไม่เท่ากัน ⇒ β_i ไม่เท่ากัน penalty

⇒ ต้อง normalize ถ้า x_i → ให้อยู่ในช่วง $[0, 1]$ หรือ → ให้ mean = 1, SD = 0

Ridge vs Lasso

วิธีไหนดีกว่า ? เลือก λ ที่เหมาะสมยังไง ?

- ทดสอบกับ test data (ทำ cross validation)

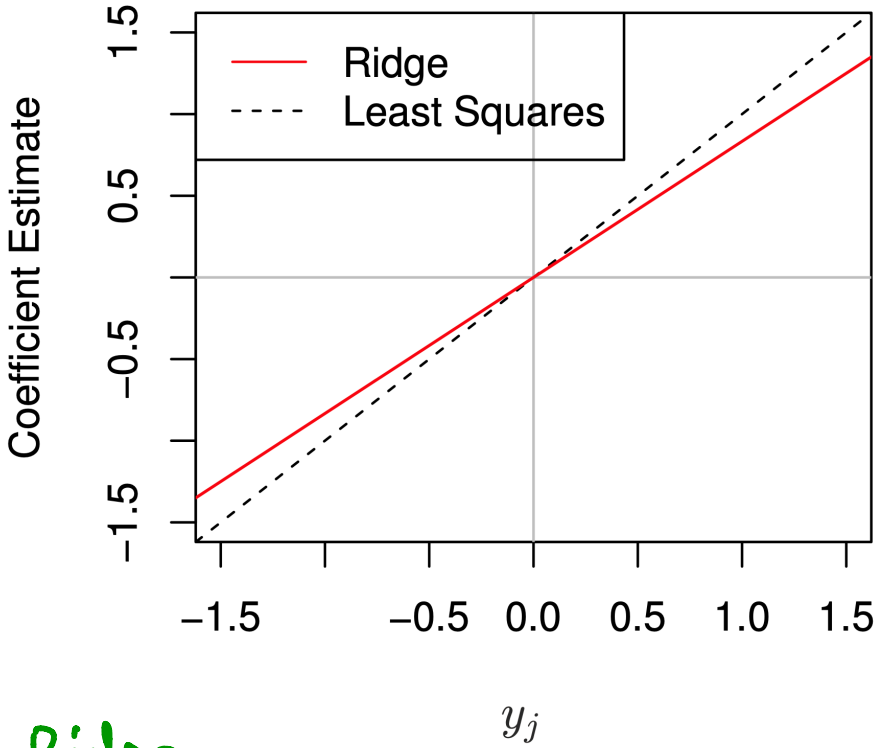
วิธีไหนดีกว่าเมื่อไหร่

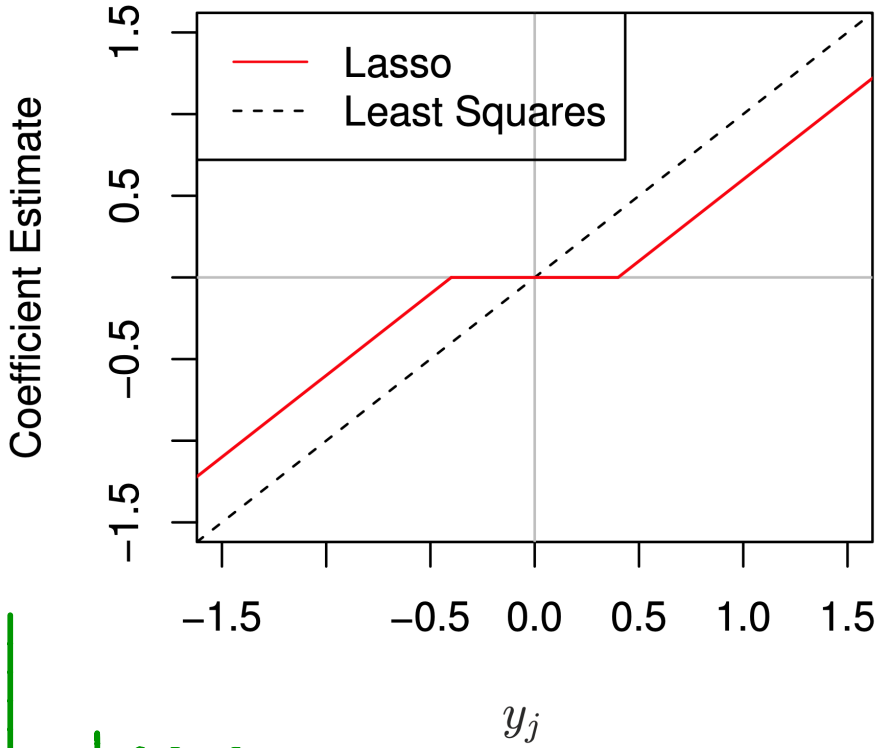
- Ridge ดีกว่าเมื่อ input ทุกตัวมีผลกับ output ($\beta_i \neq 0$)
- Lasso ดีกว่าเมื่อ input ไม่กี่ตัวมีผลต่อ output

(β_i ควรจะ $= 0$ สำหรับหลายๆตัว)

ทำ parameter selection โดยอัตโนมัติ

Ridge vs Lasso





Ridge

$$\beta_{i,ridge} \approx \frac{\beta_{i,LR}}{C(\lambda)}$$

Lasso

$$|\beta_{i,lasso}| \approx \begin{cases} |\beta_{i,LR}| - C(\lambda) \\ 0 \end{cases}$$

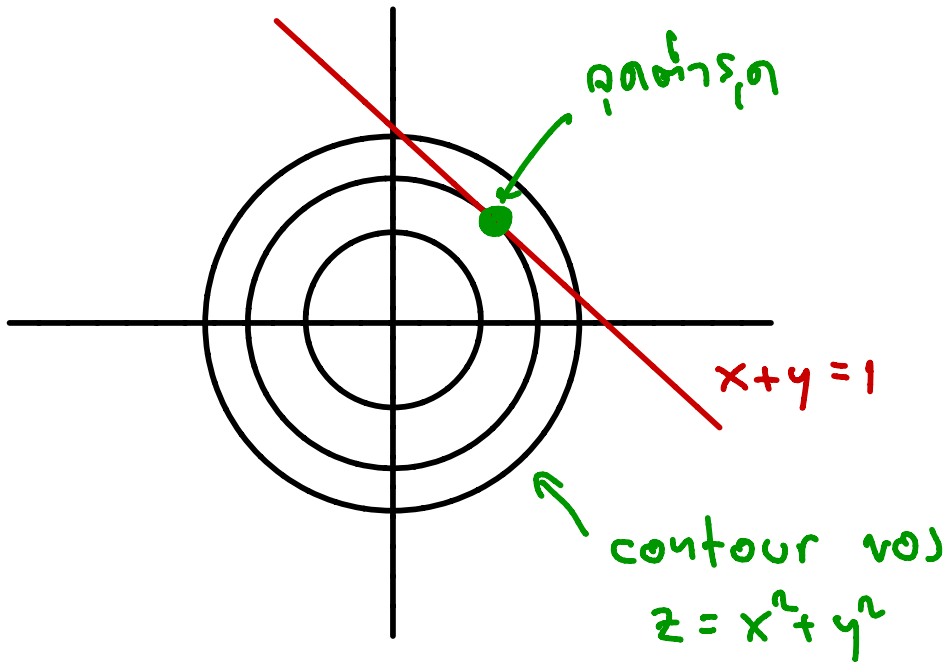
Ridge vs Lasso

Lagrange Multiplier

Ex. หา (x,y) ที่ทำให้ $z = x^2+y^2$ น้อยสุด โดยที่ $x+y=1$

$z = x^2+y^2$ objective function

$x+y=1$ constraint



Ridge vs Lasso

Ex. หา (x,y) ที่ทำให้ $z = x^2+y^2$ มีค่าสูงสุด โดยที่ $\underbrace{x+y=1}$

$\phi(x,y) = 0$

$x+y-1 = 0$

หา (x,y) ที่ทำให้ $z = x^2+y^2$ มีค่าสูงสุด โดยที่ $\phi = 0$

(Constrained optimization problem)

$\Downarrow$ Lagrange Multiplier

หา (x,y,λ) ที่ทำให้ $z+\lambda\phi$ มีค่าสูงสุด

$\Downarrow$

(Unconstrained optimization problem)

Ridge vs Lasso

Ridge

$$\left(y-\beta_0-\sum_i \beta_i x_i\right)^2+\lambda\sum_i \beta_i^2 \iff$$

minimize $\left(y-\beta_0-\sum_i \beta_i x_i\right)^2$ subject to $\sum_i \beta_i^2 = \text{constant}$

Lasso

$$\left(y-\beta_0-\sum_i \beta_i x_i\right)^2+\lambda\sum_i |\beta_i| \iff$$

minimize $\left(y-\beta_0-\sum_i \beta_i x_i\right)^2$ subject to $\sum_i |\beta_i| = \text{constant}$

Ridge vs Lasso