迴歸分析

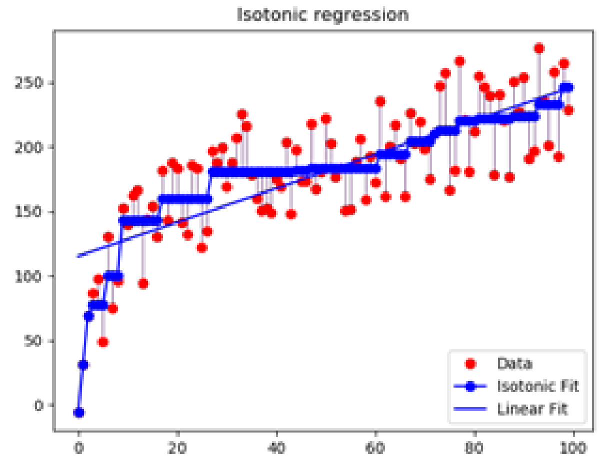
黃志勝 (Tommy Huang)
義隆電子 人工智慧研發部
國立陽明交通大學 AI學院 合聘助理教授
國立台北科技大學 電資學院合聘助理教授

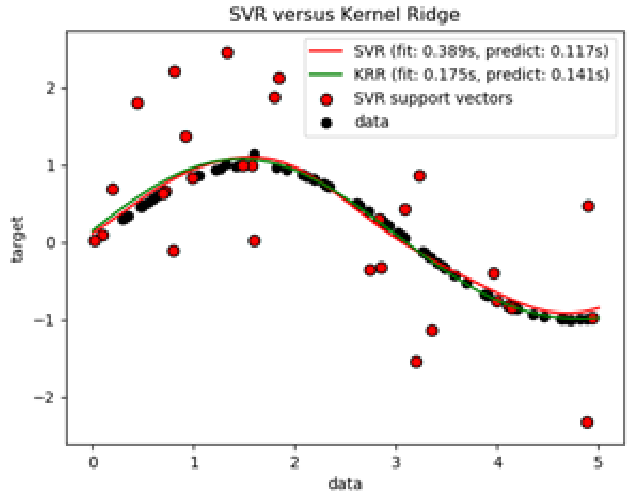
Regression

- Introduction for regression
- Linear Regression
- Regularized Regression (L1 & L2)

Regression

- In the last presentation, we brief introduce ML topic.
- Regression: predicting a continuous-valued attribute associated with an object.
- What to do?
- How to do?





Regression

●What to do?

independent variables

predict

dependent variables

Which are dependent variables?

Depend on your problem : specific definition (salary prediction or bodyfat prediction)

Which are independent variables?

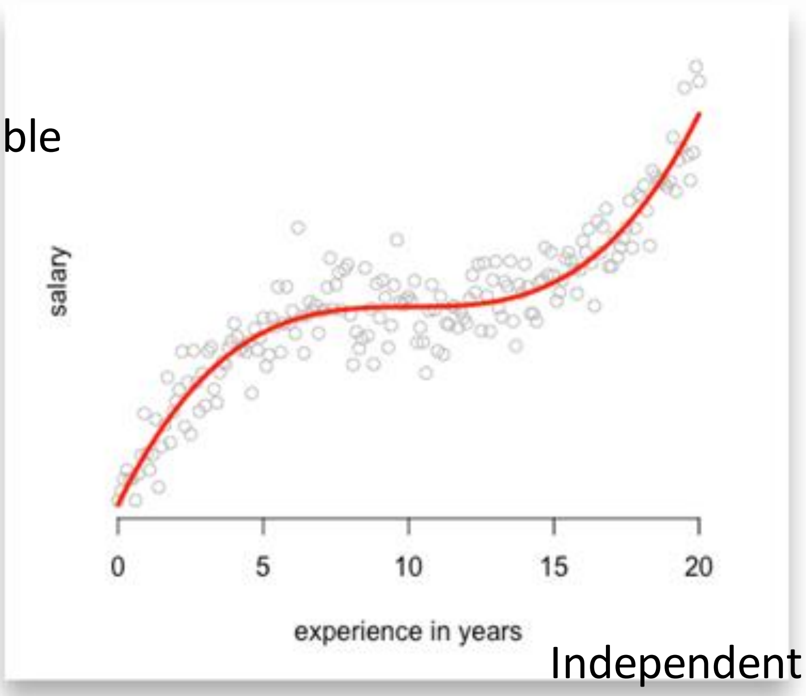
Depend on your collecting data.

Regression

●How to do?

Finding the curve that best fits your data is called regression.

Dependent variable
(y)



Independent variable
(x)

$$y = f(x)$$

f is a linear function：
linear regression

f is non-linear function：
nonlinear regression

Regression

y: salary, x: experience in years

$$y = f(x) = \beta_0 + \beta_1 x$$

→ Simple linear regression

β_0: intercept
β_1: Slope

Regression

If there are more than one independent variables.

y: salary

x_1: experience in years

x_2: career

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \longrightarrow \text{Multiple linear regression}$$

Regression

- How to do nonlinear?
- Let your independent variables as a other independent variable by

1. polynomial.

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

2. Interact.

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

3. Nonlinear function (ϕ): sigmoid function,…

$$y = f(x) = \phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

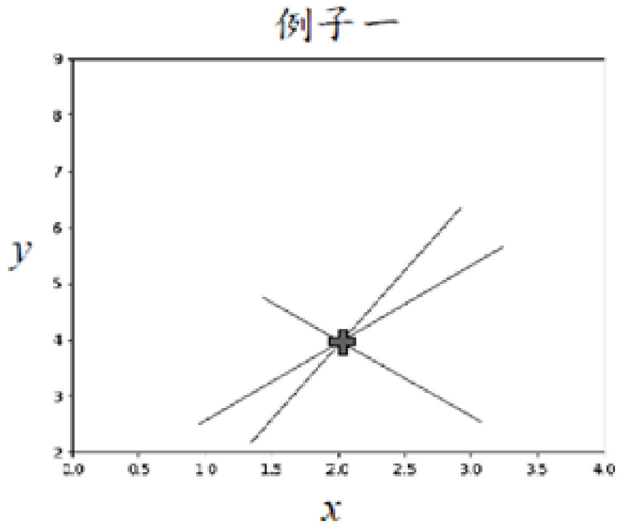
Regression(Example)

$$y = f(x) = \beta_0 + \beta_1 x$$

- 訓練資料只有一筆資料(*x*, *y*) = {(2, 4)}，我們將此資料代入方程式
- 內：

$$4 = \beta_0 + 2\beta_1$$

- β_0和β_1的解有無限多組。

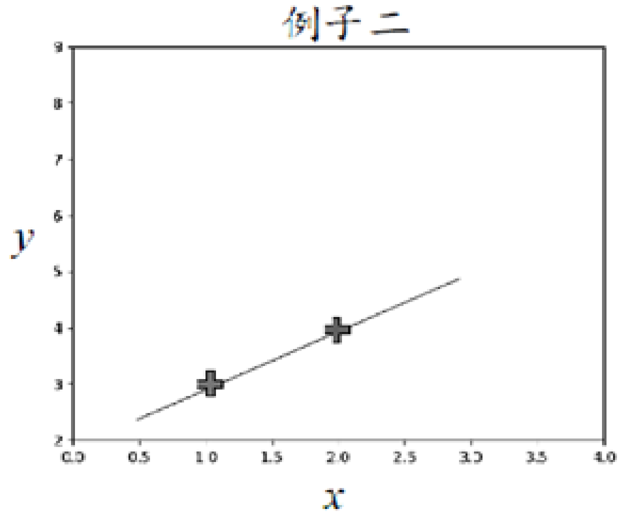


Regression(Example)

$$y = f(x) = \beta_0 + \beta_1 x$$

訓練資料只有一筆資料(x, y) = {(2, 4),(1,3)}，我們將此資料代入方程式內：

$$\begin{cases} 4 = \beta_0 + 2\beta_1 \\ 3 = \beta_0 + 1\beta_1 \end{cases} \Rightarrow \begin{cases} \beta_0 = 2 \\ \beta_1 = 1 \end{cases}$$

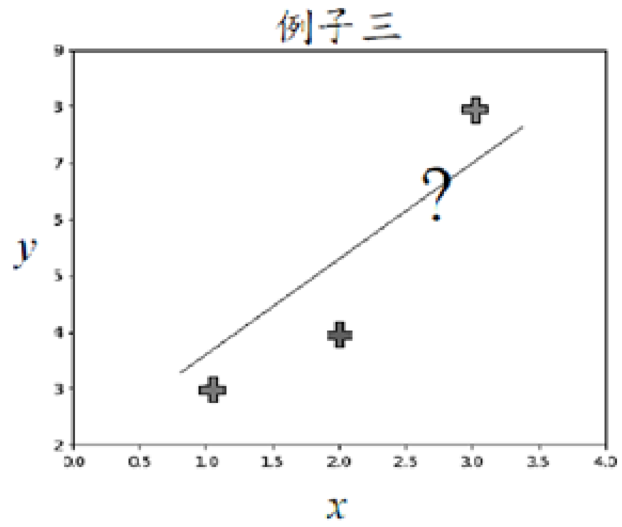


Regression(Example)

$$y = f(x) = \beta_0 + \beta_1 x$$

訓練資料只有一筆資料(*x*, *y*) = {(2, 4),(1,3),(3,8)}，我們將此資料代入方程式內：

$$\begin{cases} 4 = \beta_0 + 2\beta_1 \cdots (1) \\ 3 = \beta_0 + 1\beta_1 \cdots (2) \\ 8 = \beta_0 + 3\beta_1 \cdots (3) \end{cases}$$



Regression

- For now, we clearly understand what is regression.

Recall: How to do?

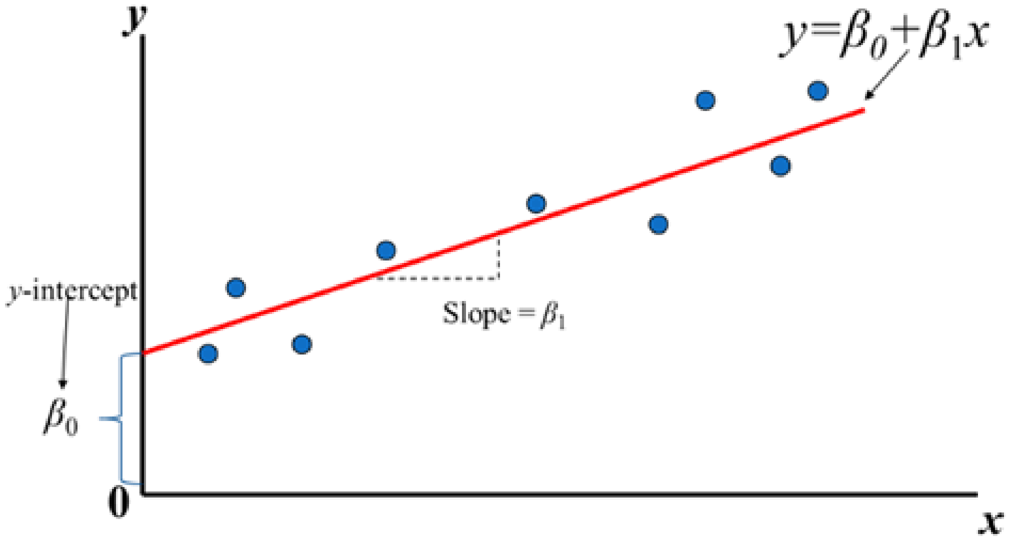
Finding the curve that best fits your data is called regression.

Two key points: **1. data, 2. curve.**

Data is the blue point

Curve is the red line

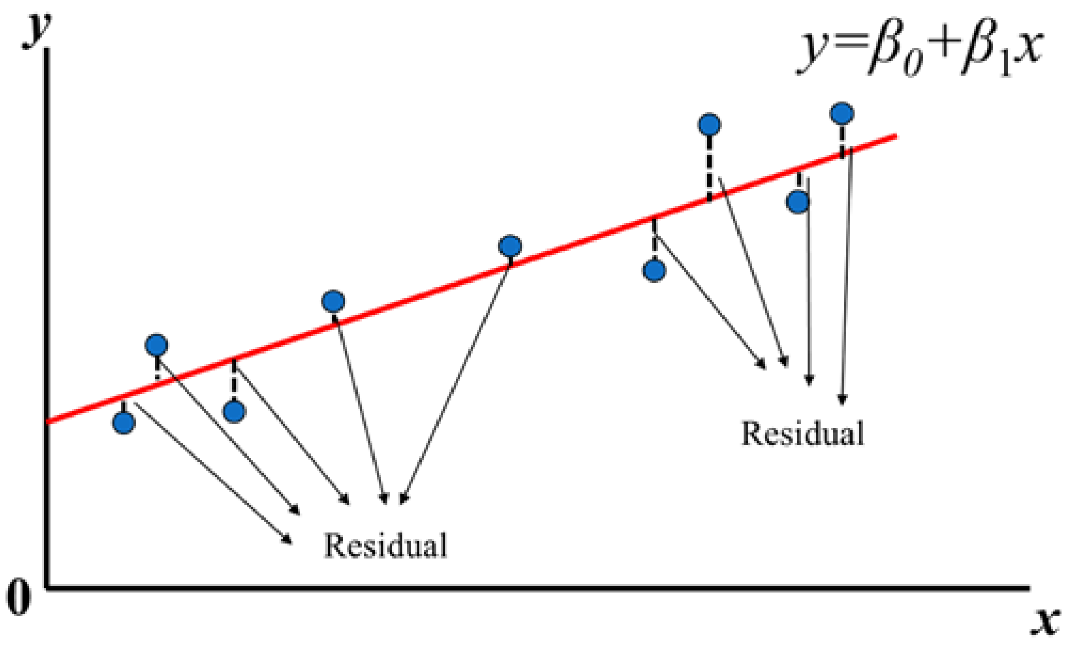
Using the data to find the β_0 and β_1



Regression

- Using the data to find the β_0 and β_1 .

How to achieve this goal?



Ideal:

All the data can fix on this line.

Real:

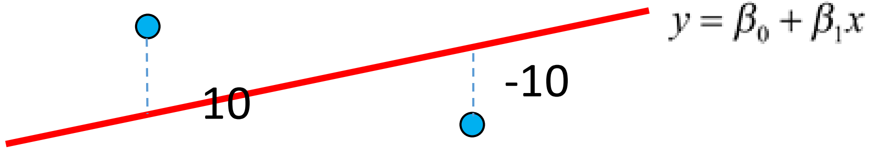
Fix on the line as best as possible.
Residuals are as small as possible.

Regression

- Residuals are as small as possible.

$$residual = \hat{y} - y$$

- Residuals can be positive and negative.



$$sum\ error = \sum_i (\hat{y}_i - y_i) = 10 - 10 = 0$$

$$sum\ square\ error = \sum_i (\hat{y}_i - y_i)^2 = 100 + 100 = 200$$

Regression

- We usually hope the can let the sum square error as small as possible.

$$sum\ square\ error(SSE) = \sum_i (\hat{y}_i - y_i)^2$$

$$mean\ square\ error(MSE) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

- SO in regression, the objective/loss function is MSE.

$$\min_{\beta_0,\beta_1}\left\{ loss(\beta_0,\beta_1) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}((\beta_0+\beta_1 x) - y_i)^2 \right\}$$

Regression

- In calculation, using derivative to find the minima.

$$\min_{\beta_0,\beta_1}\left\{ loss(\beta_0,\beta_1) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}((\beta_0+\beta_1 x)-y_i)^2 \right\}$$

$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_0} = 0$$

$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_1} = 0$$

Regression

Find β_0 (intercept)

$$\frac{\partial loss(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2}{\partial \beta_0} = 0$$

$$\Rightarrow \frac{2}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(\beta_0) + \sum_{i=1}^{n}(\beta_1 x_i - y_i) = 0$$

$$\Rightarrow n\beta_0 = \sum_{i=1}^{n}(y_i - \beta_1 x_i)$$

$$\Rightarrow \beta_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i) - \beta_1 \frac{1}{n}\sum_{i=1}^{n}(x_i) = \bar{y} - \beta_1 \bar{x}$$

Regression

Find β_1 (Slope)

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\frac{\partial loss(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2}{\partial \beta_1} = 0$$

$$\Rightarrow \frac{2}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i) x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (\bar{y} - y_i) x_i + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i = 0$$

$$\Rightarrow \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \sum_{i=1}^{n} (y_i - \bar{y}) x_i$$

$$\Rightarrow \beta_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Details

$$\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$$

分母:

$$\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i x_i - \bar{x}x_i) = \sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}\bar{x}x_i = \sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \ldots(1)$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \ldots(2)$$

$$\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

分子:

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n}(x_i y_i - \bar{y}x_i) = \sum_{i=1}^{n}x_i y_i - \bar{y}\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y} \ldots(3)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n}x_i y_i - \bar{x}\sum_{i=1}^{n}y_i - \bar{y}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}\bar{y} = \sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y} \ldots(4)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

Ordinary Least Square Estimation (OLSE)

We hope the loss as small as possible, so this approach is called ordinary least square estimation.

Recall:

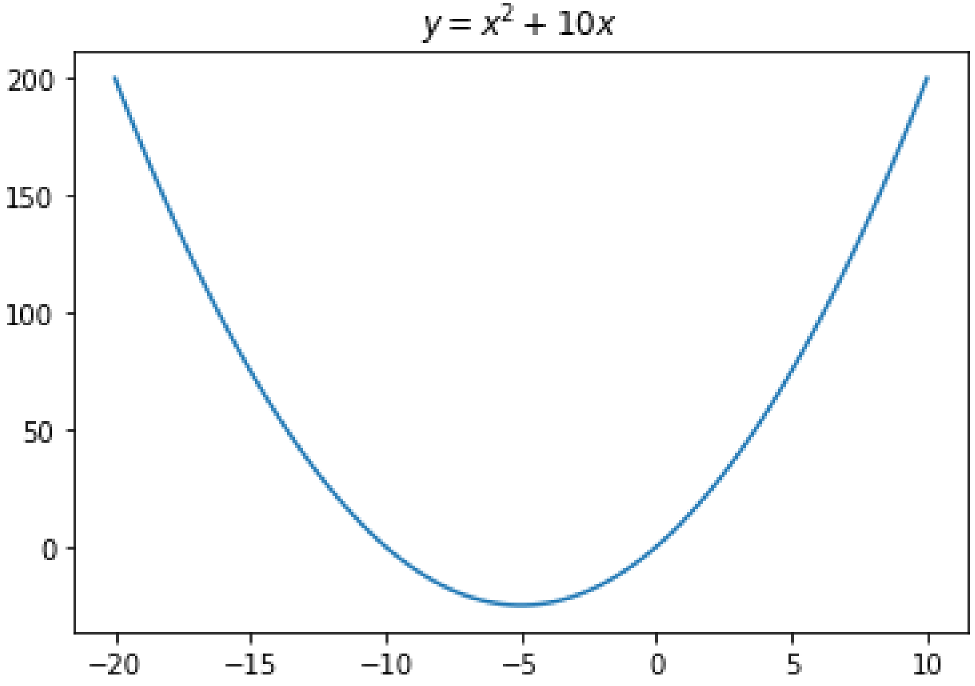
$$\min_{\beta_0,\beta_1}\left\{loss(\beta_0,\beta_1)=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2\right\}$$

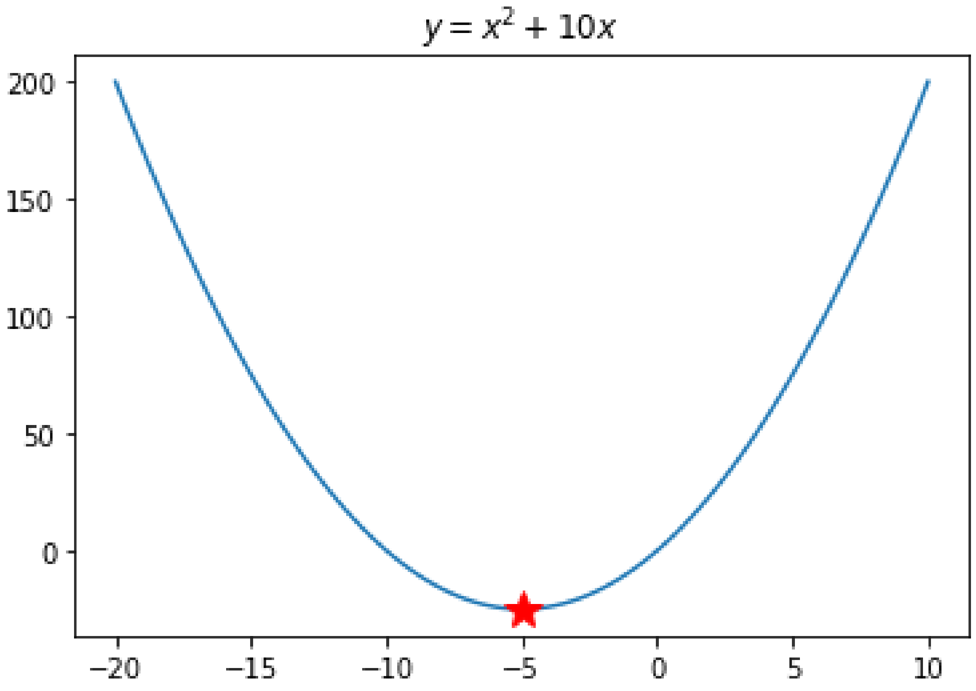
$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_0}=0 \Rightarrow \beta_0=\bar{y}-\beta_1\bar{x}$$

$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_1}=0 \Rightarrow \beta_1=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

簡易數值分析

$$\min_x x^2 + 10\,x$$



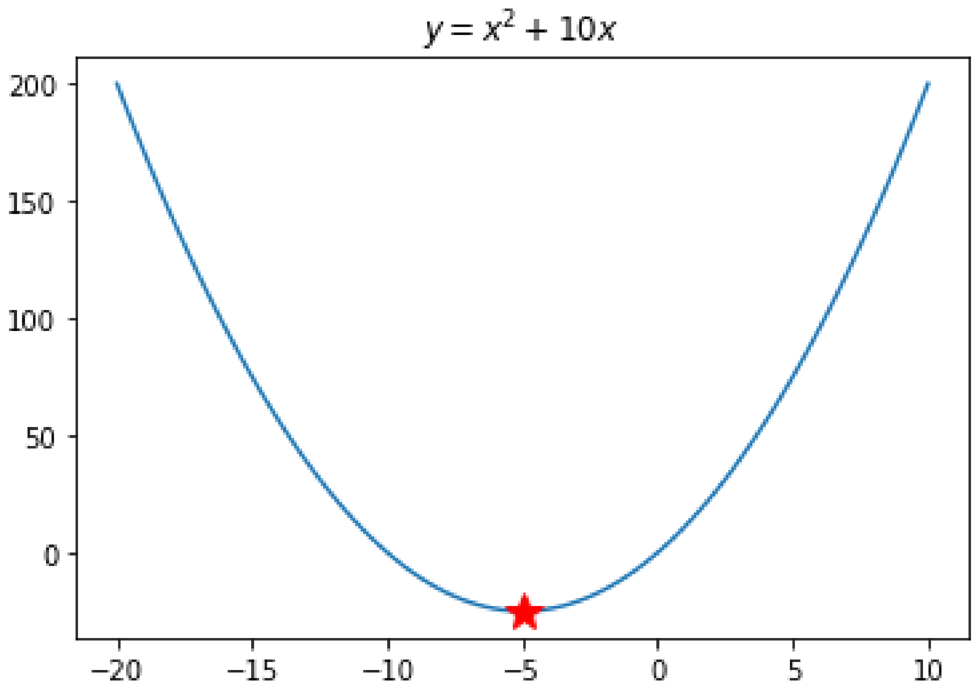


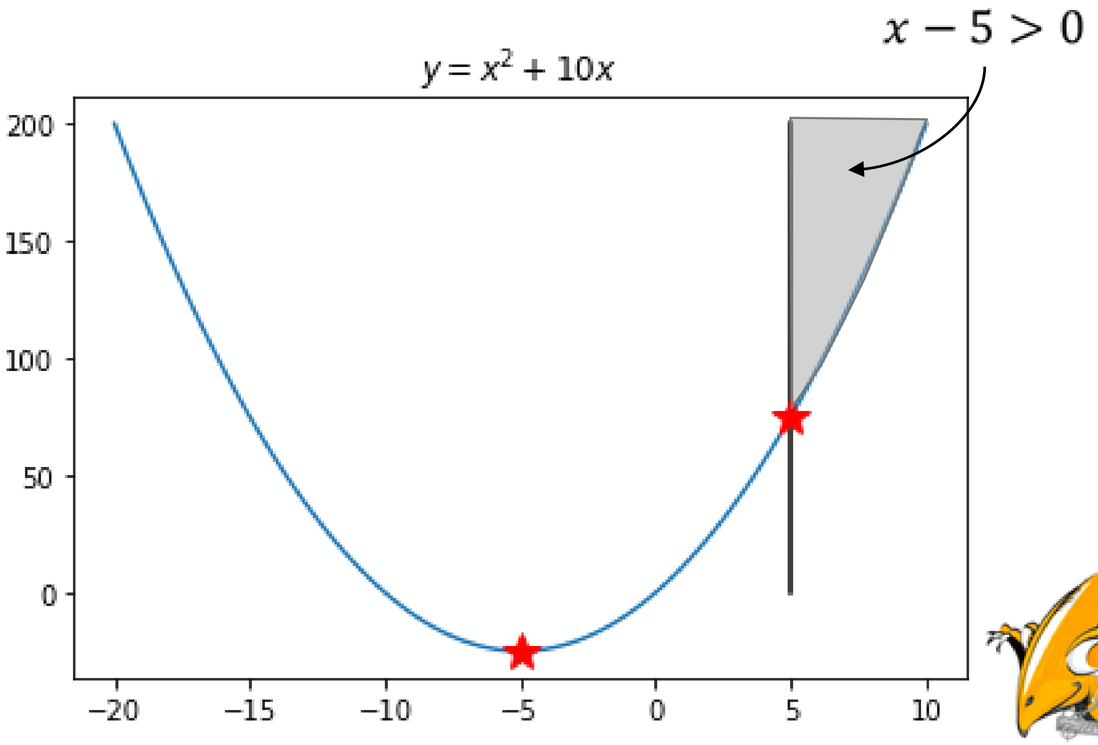
簡易數值分析

- $$\min_{x} x^2 + 10\,x$$

subject to

$$x - 5 > 0$$



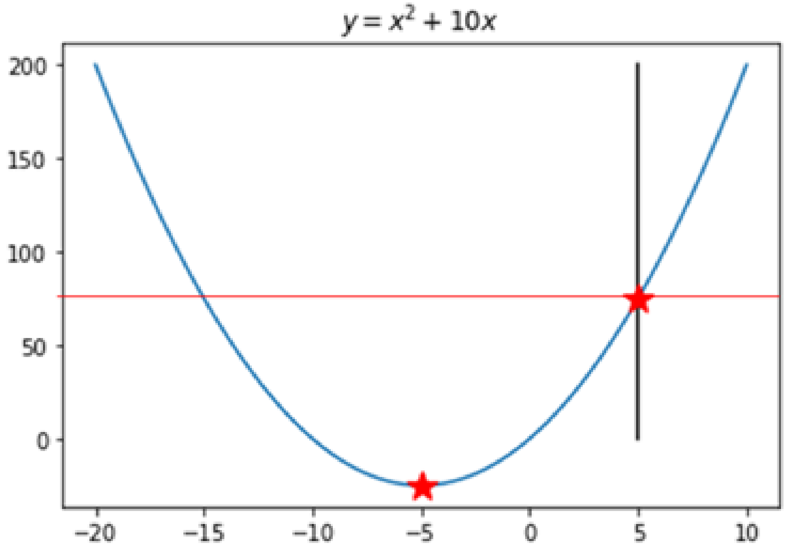


簡易數值分析

$$\min_x x^2 + 10x$$

subject to

$$x - 5 > 0$$



等價於

$$\min_x\{x^2 + 10x - \lambda(x-5)\}$$

$$\frac{\partial x^2 + 10x - \lambda(x-5)}{\partial\lambda} = x - 5 = 0 \Rightarrow x = 5$$

$$\frac{\partial x^2 + 10x - \lambda(x-5)}{\partial x} = 2x + 10 - \lambda = 0 \Rightarrow \lambda = 2x + 10 = 20$$

所以在x=5有最小值，$x^2 + 10x - \lambda(x-5)$= $5^2 + 50 - 20(5-5) = 75$。

Regularized Regression

Regularized Regression

- Regularized term, also call penalized term, is using to control the coefficients in regression model. (This trick is also using in deep learning).
- In regularized regression,

$$\min_{\beta}\left\{\mathrm{MSE}(\hat{y}, y) + p_{\beta}\right\}$$

- Regularized term is a way to overcome the overfitting problem in learning algorithm.

Regularized Regression

Ridge regression

$$\min_{\beta}\left\{\mathrm{MSE}(\hat{y}, y) + \lambda L_2\mathrm{norm}(\beta)\right\}$$

$$L_2\mathrm{norm}(\beta) = \sum_i \beta_i^2$$

Least absolute shrinkage and selection operator (LASSO)

$$\min_{\beta}\left\{\mathrm{MSE}(\hat{y}, y) + \lambda L_1\mathrm{norm}(\beta)\right\}$$

$$L_1\mathrm{norm}(\beta) = \sum_i \left|\beta_i\right|$$

Regularized Regression

Absolutely

Elastic Net

$$\min_{\beta}\left\{\mathrm{MSE}(\hat{y}, y) + \lambda_1 L_1\mathrm{norm}(\beta) + \lambda_2 L_2\mathrm{norm}(\beta)\right\}$$

Regularized Regression

$$\min_{\beta}\{\mathrm{MSE}(\hat{y}, y) + \lambda L_2\mathrm{norm}(\beta)\}$$

$\lambda = 0$

regularized regression = linear regression

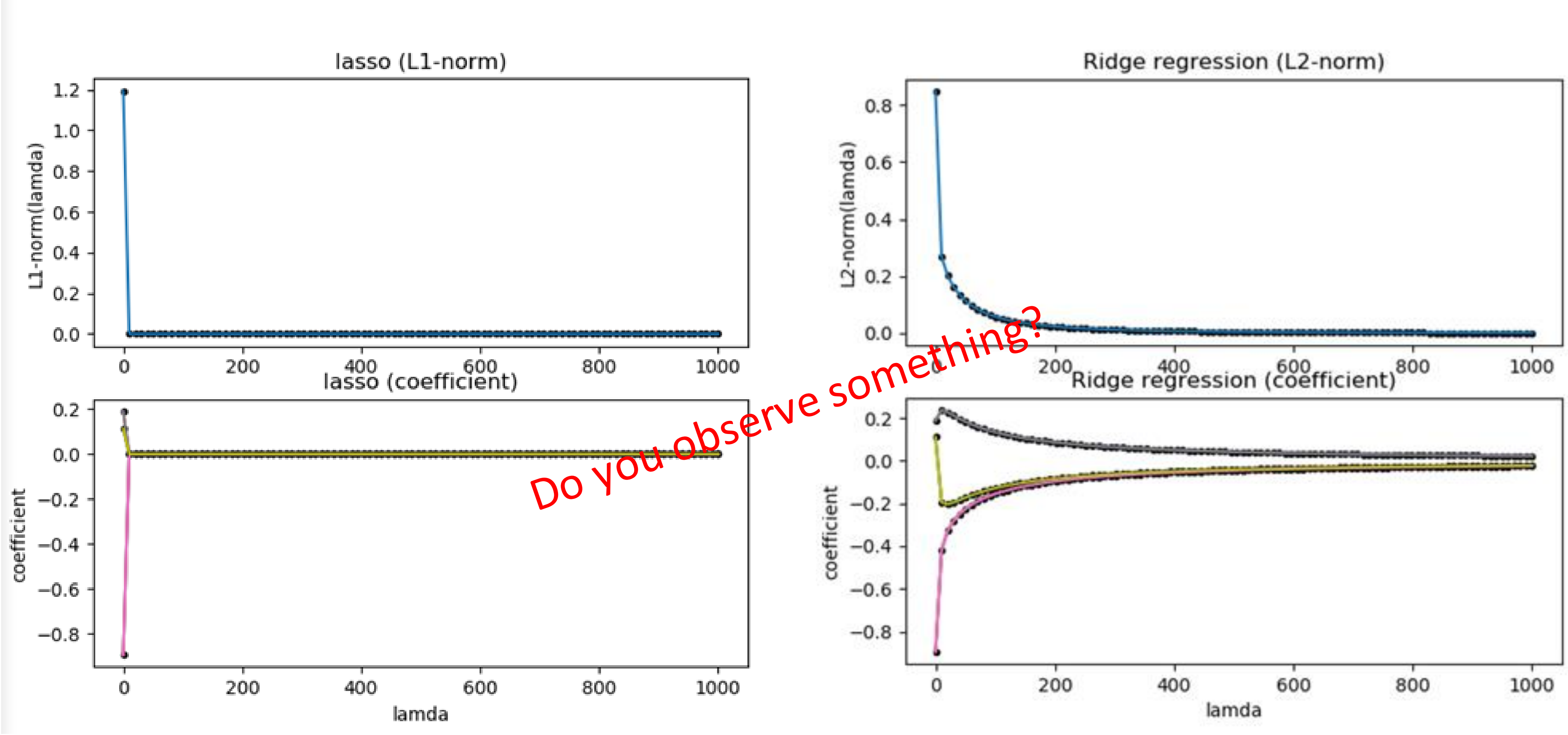
$\lambda \longrightarrow \infty$

$$\lambda L_2\mathrm{norm}(\beta) > \mathrm{MSE}(\hat{y}, y)$$

$$\beta \longrightarrow 0$$

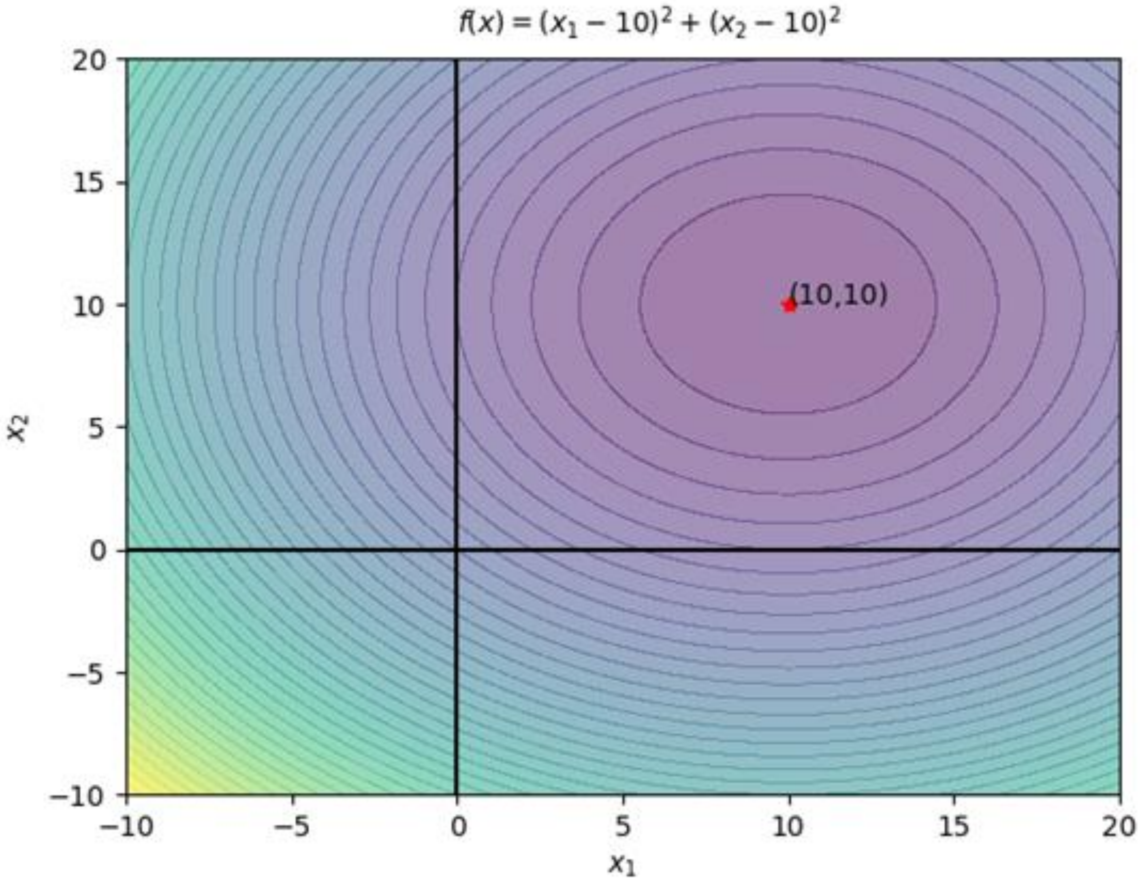
Regularized Regression

Example for a three independent variables with one dependent variable.



Do you observe something?

Regularized Regression



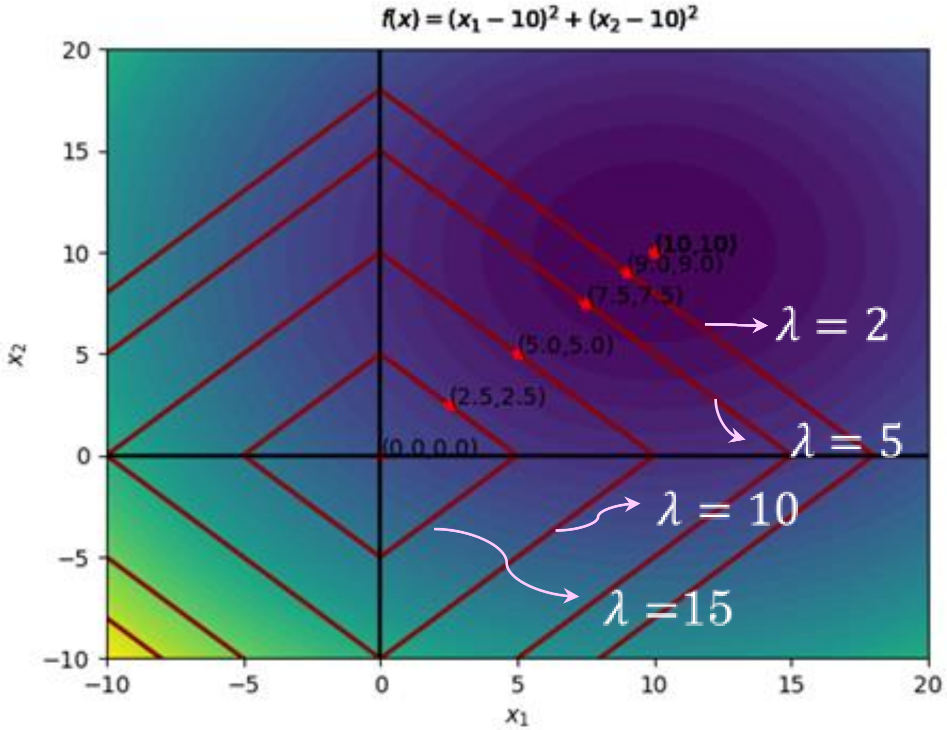
$$\min_{x1,x2}\{f(x) = (x_1 - 10)^{2} + (x_2 - 10)^{2}\}$$

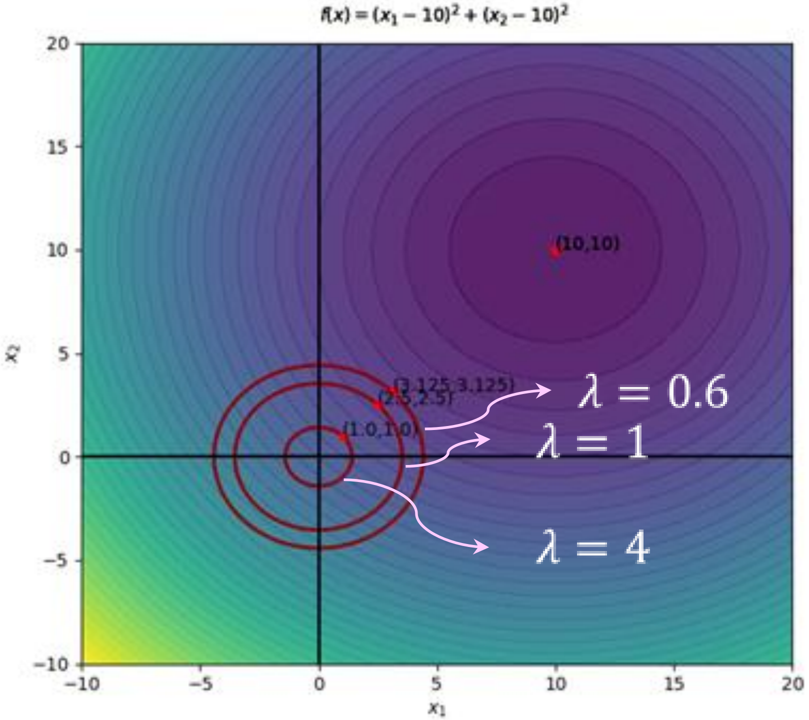
ANS: $x_1 = 10, x_2 = 10$

Regularized Regression (L1&L2)

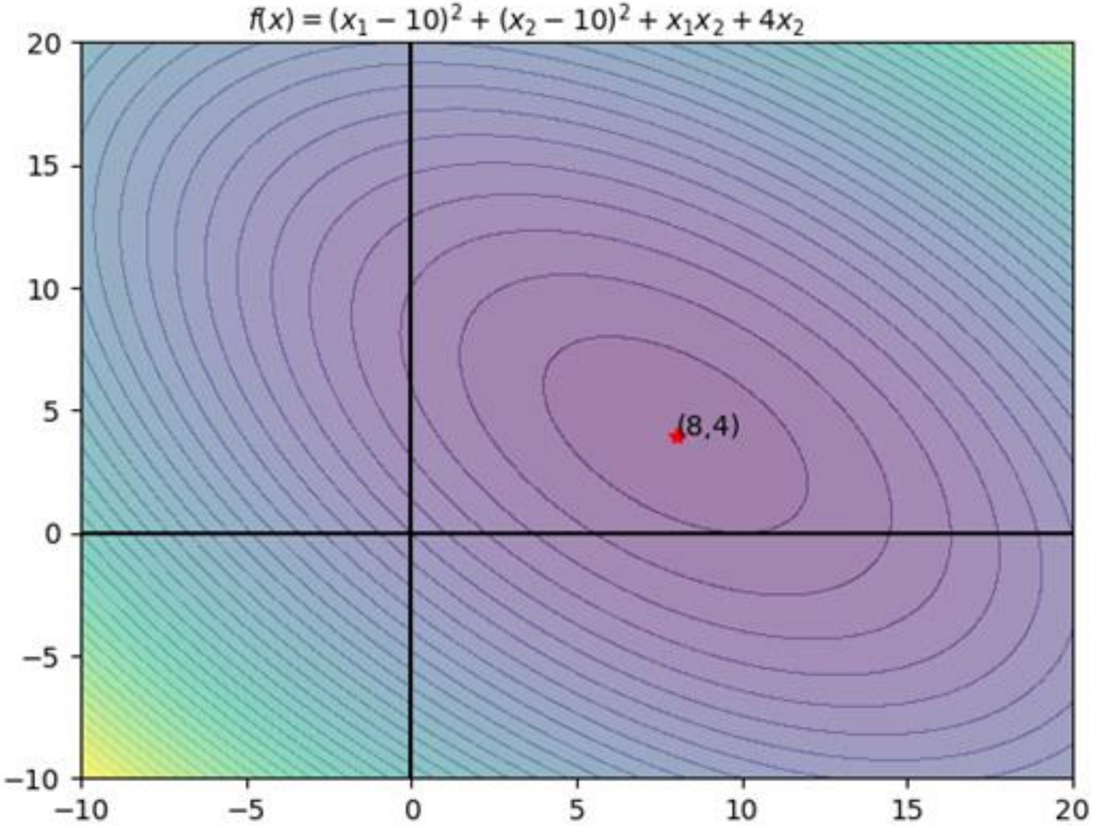
$$\min_{x1,x2}\{f(x)+\lambda\sum_{i=1}^{2}|x_i|\}$$

$$\min_{x1,x2}\{f(x)+\lambda\sum_{i=1}^{2}{x_i}^2\}$$





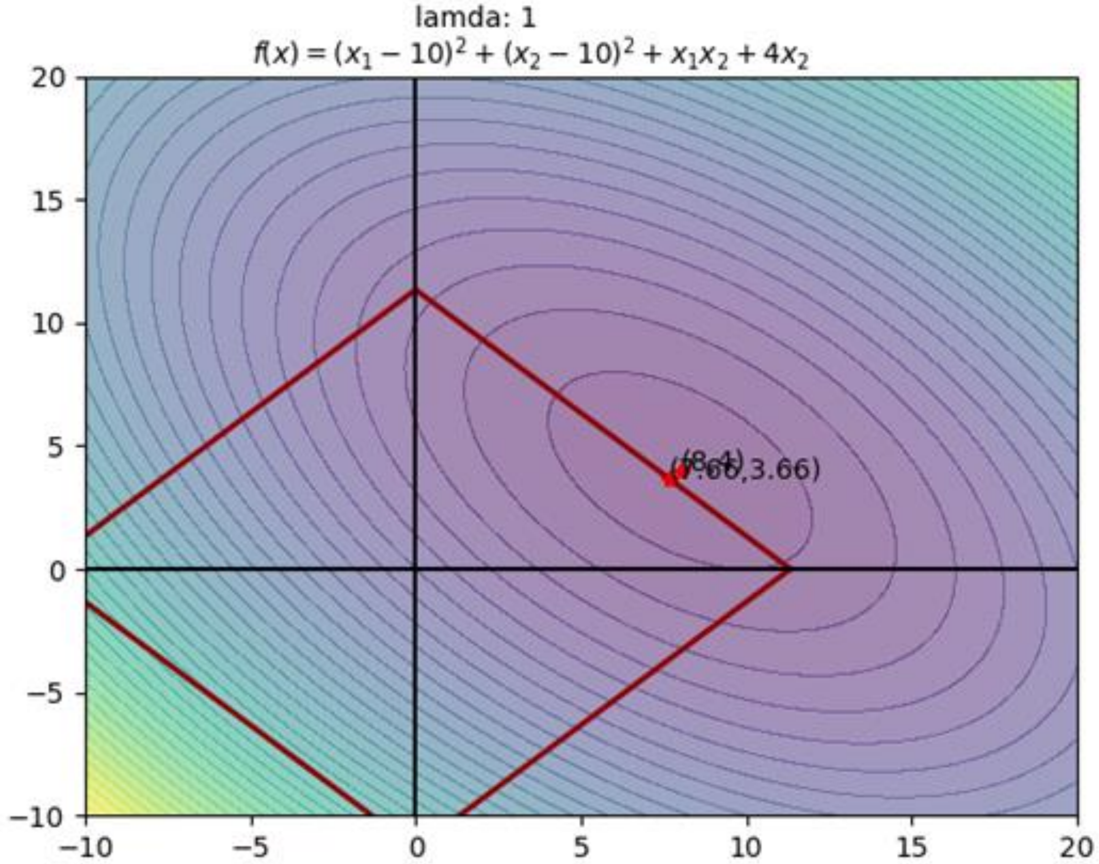
Regularized Regression



$$\min_{x1,x2}\{f(x) = (x_1 - 10)^{\ 2} + (x_2 - 10)^{\ 2} - x_1x_2 + 4x_2\}$$

ANS: $x_1 = 8, x_2 = 4$

Regularized Regression (L1)



$$\min_{x1,x2}\{f(x) + \lambda \sum_{i=1}^{2} |x_i|\}$$

$$\lambda = 12, x_1 = 4, x_2 = 0$$

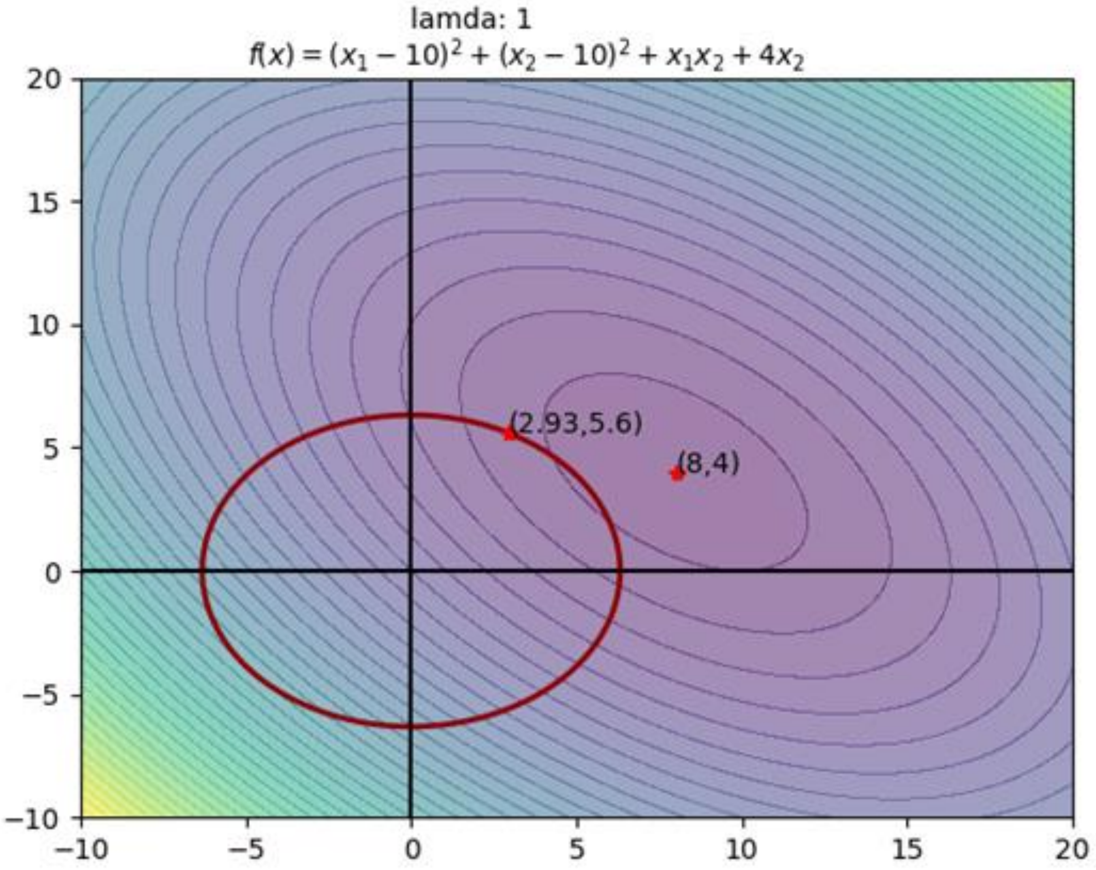
Advantage:

L1 norm has corners, it' very likely that the joint minima is at one of the corners. →Sparsity

Disadvantage:

Not differentiable everwhere.

Regularized Regression (L2)



$$\min_{x1,x2}\{f(x) + \lambda \sum_{i=1}^{2} x_i^{\,2}\}$$

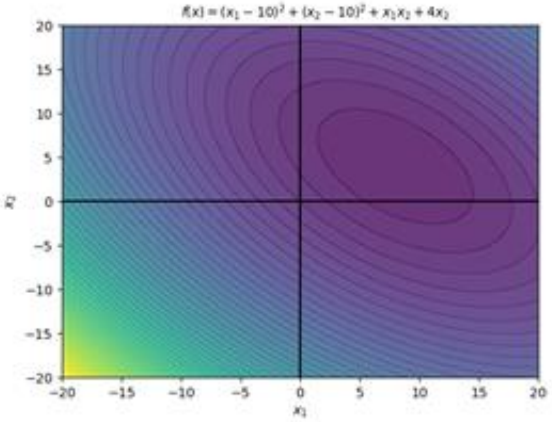
Advantage:

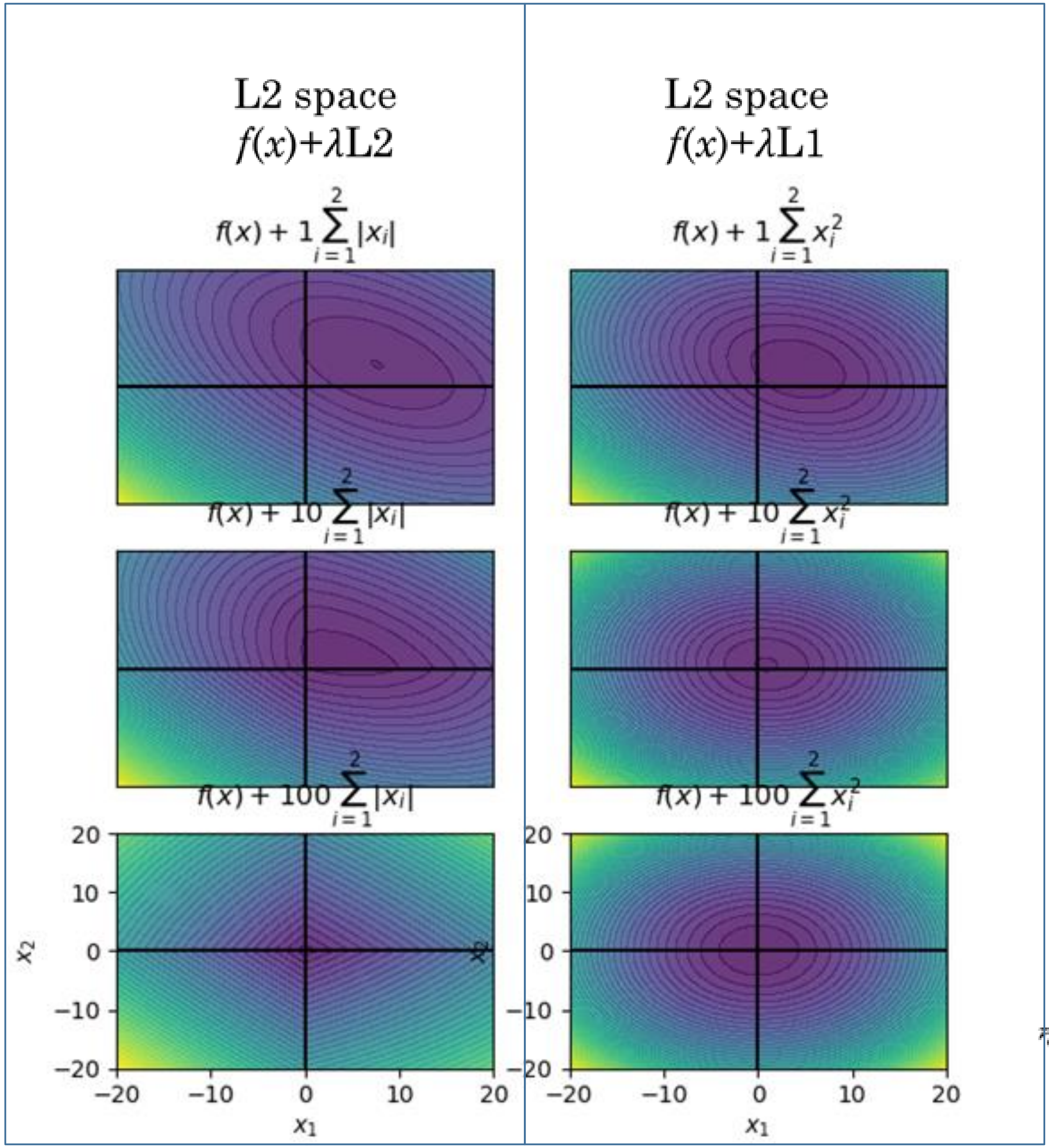
L2 norm has no corners, it' very likely that the joint minima is on any of axes.

Differentiable and easy to optimize.

Regularized Regression

Original space
f(*x*)





Regularized Regression

Can we give different penalized terms for each variable?

$$\min_{x1,x2}\{f(x) + \lambda_1 x_1{}^2 + \lambda_2 x_2{}^2\}$$

$\lambda_i \to \infty$, $x_i \to 0$, so we can use the regularized term to control the model.

$$\{\mathbf{x}_i, y_i\}, \mathbf{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{id} \end{bmatrix} \in R^{d+1}, \forall i = 1,2,\ldots,n$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

Regression: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$

$$loss(\boldsymbol{\beta}) = (\mathbf{Y} - \hat{\mathbf{Y}})^{\mathrm{T}} (\mathbf{Y} - \hat{\mathbf{Y}})$$

找 $\boldsymbol{\beta}$ 的 closed-form solution

- Code Example