[機器與深度學習基礎知識初探] Regression and Classification

黃志勝

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

基礎機器學習

針對前述的介紹，每個topic都介紹一個演算算法

1. Regression: Linear regression & Regularization

2. Classification: Linear and Quadratic Discriminant Analysis

3. Clustering: K-means

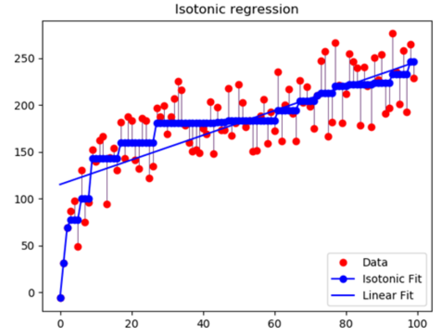
4. Dimension Reduction: PCA

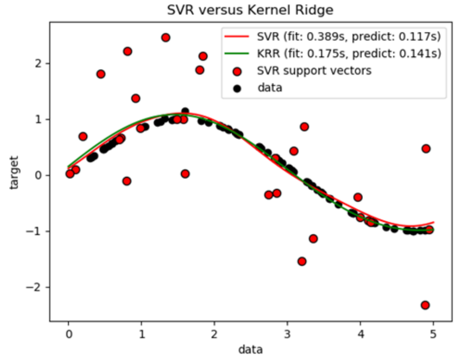
5. Ensemble learning: 不介紹。

Regression

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

Regression



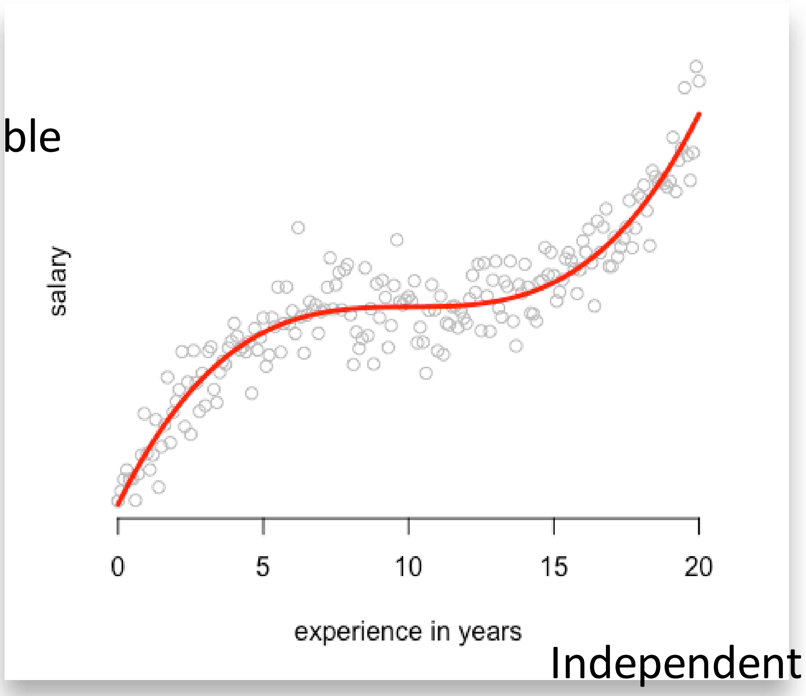


- **How to do?**

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

$$y = f(x)$$

Dependent variable (y)



Independent variable (x)

f is a linear function : linear regression

f is non-linear function : nonlinear regression

Regression

y: salary, x: experience in years

$$y = f(x) = \beta_0 + \beta_1 x \longrightarrow \text{Simple linear regression}$$

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

- For now, we clearly understand what is regression.

Recall: How to do?

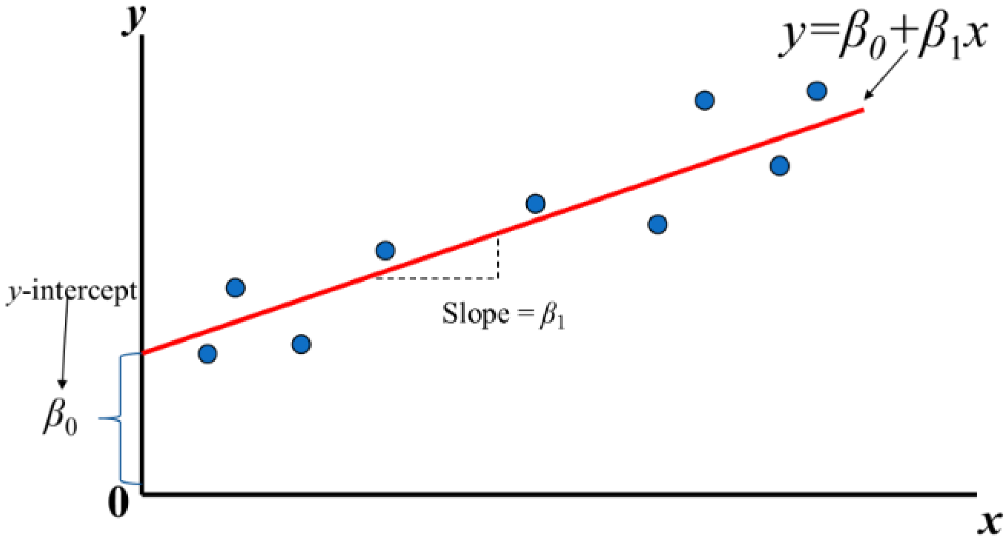
Finding the curve that best fits your data is called regression.

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

Data is the blue point

Curve is the red line

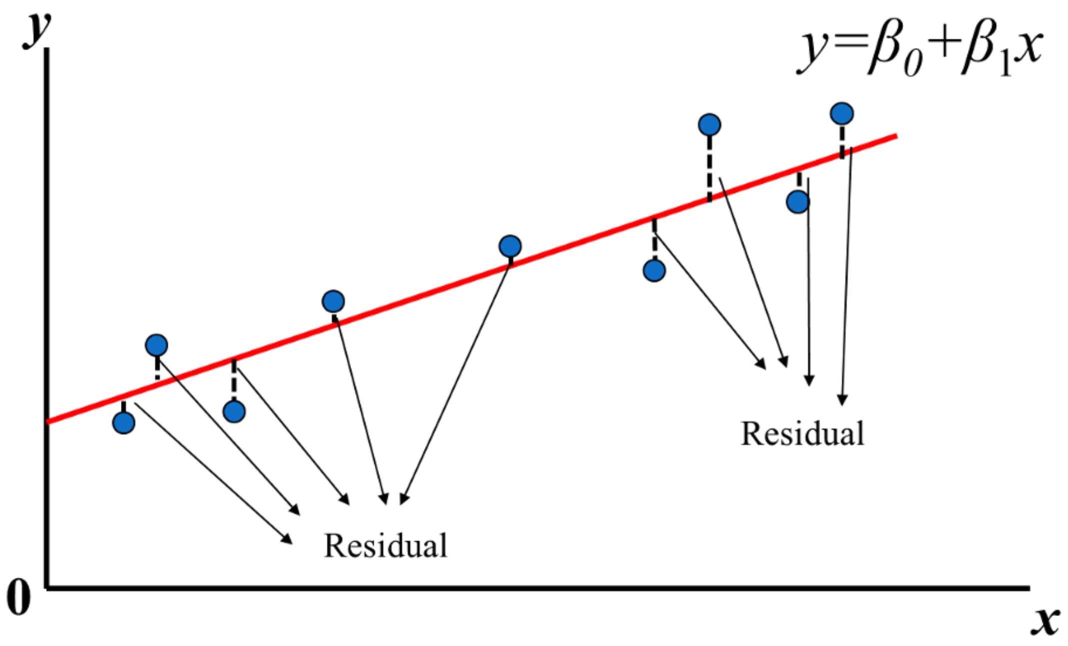
Using the data to find the β_0 and β_1



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.

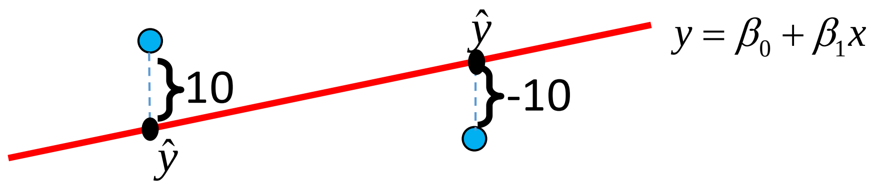
Residuals are as small as possible.

Regression

- Residual: 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.

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

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

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

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

Optimization for MSE loss

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

- In calculation, using derivative to find the minima.

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

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

Find β_0 (intercept)

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

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

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

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

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

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

Find β_1 (Slope)

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

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

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

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

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

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

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

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:

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

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

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

Ordinary Least Square Estimation (OLSE)

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \mathbf{X}^T \boldsymbol{\beta}$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}_{(d+1)\times 1}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(d)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & \cdots & x_n^{(d)} \end{bmatrix}_{n\times(d+1)}$$

$$\begin{aligned} Loss(\boldsymbol{\beta}) &= (\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) \\ &= (\mathbf{Y} - \hat{\boldsymbol{\beta}}^T\mathbf{X})^T(\mathbf{Y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}) \\ &= \mathbf{Y}^T\mathbf{Y} + \mathbf{X}^T\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T\mathbf{X} - 2\mathbf{X}^T\hat{\boldsymbol{\beta}}\mathbf{Y} \end{aligned}$$

$$\frac{\partial Loss(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \partial \frac{\mathbf{Y}^T\mathbf{Y} + \mathbf{X}^T\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T\mathbf{X} - 2\mathbf{X}^T\hat{\boldsymbol{\beta}}\mathbf{Y}}{\partial \boldsymbol{\beta}} = 0$$

$$\Rightarrow 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}^T\mathbf{Y} = 0$$

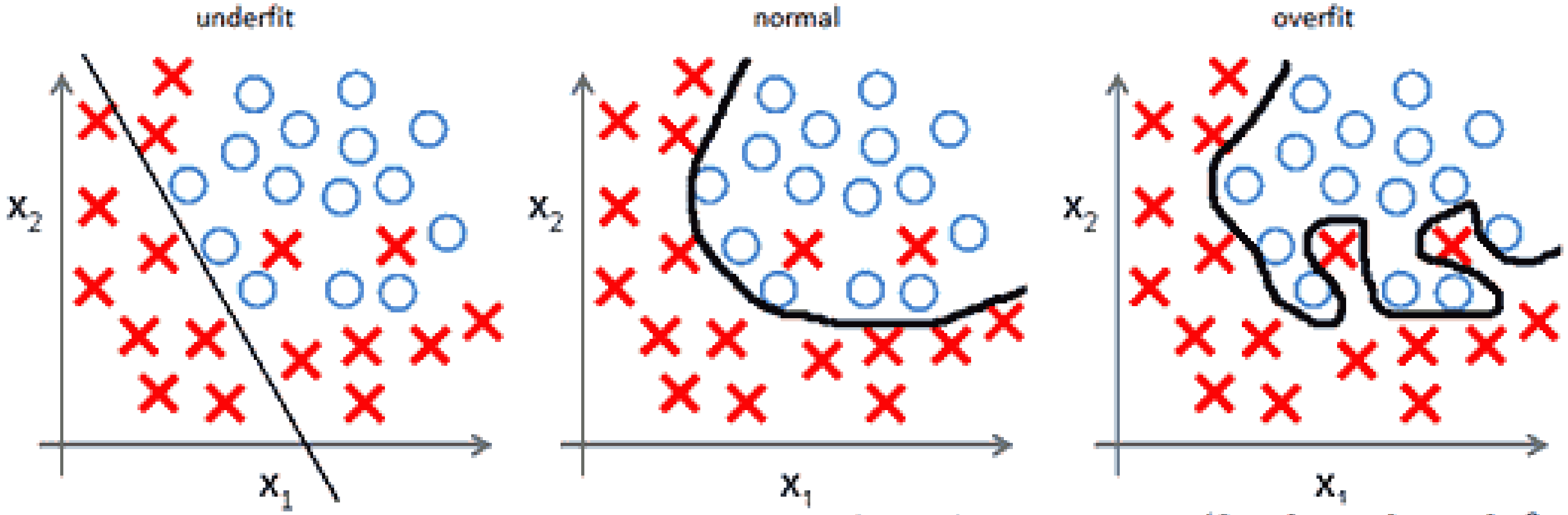
$$\Rightarrow \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

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).



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \theta_6 x_1^2 x_2 + \theta_7 x_1 x_2^2 + \theta_8 x_1^2 x_2^2 + \theta_9 x_1^3 + \ldots)$

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.
- In deep learning, called **weight decay**.

Regression

$$y = \beta_0 + \beta_1 x$$

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)

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

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

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

Regression

$$y = \beta_0 + \beta_1 x$$

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

$\lambda = 0$

regularized regression = linear regression

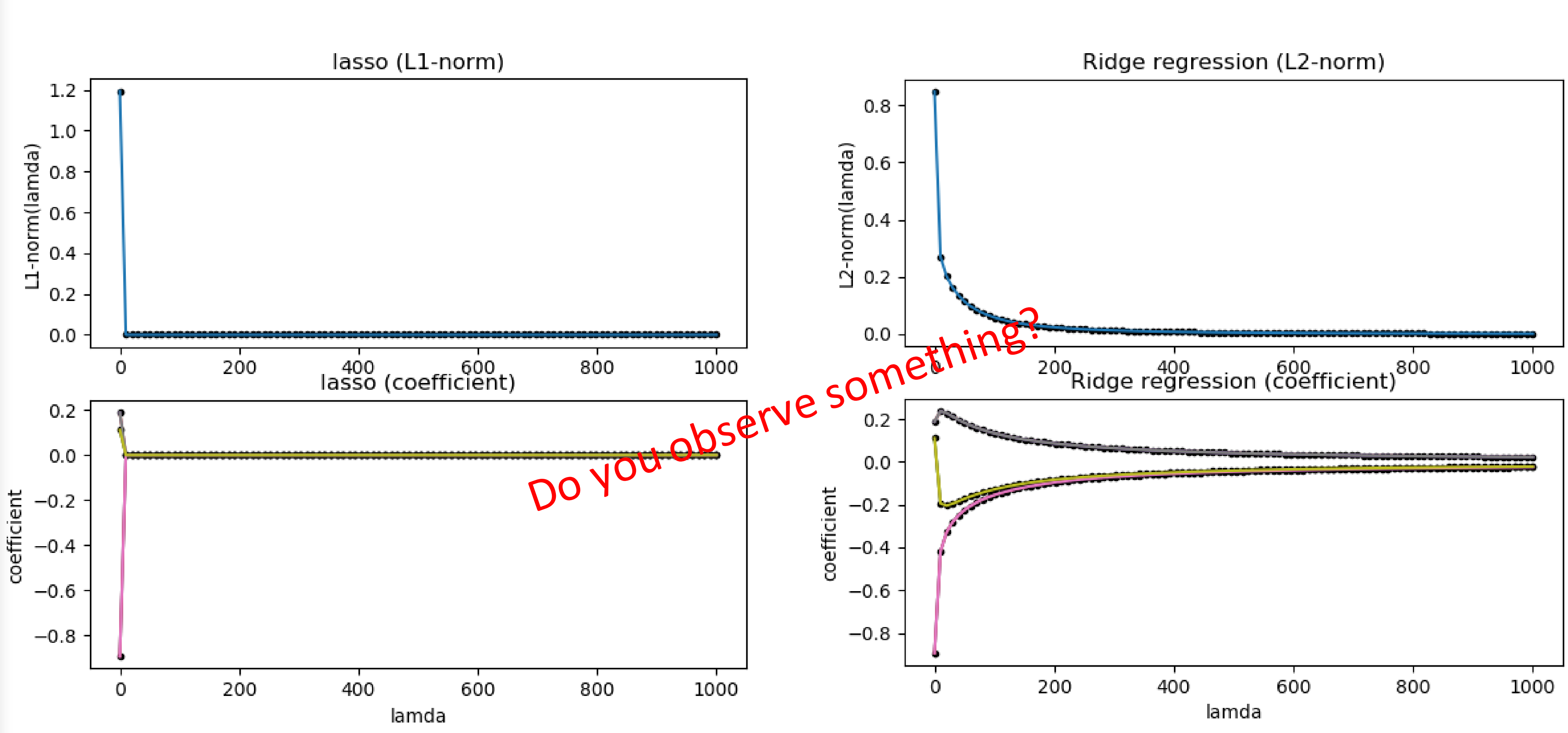
$\lambda \to \infty$

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

$$\beta \to 0$$

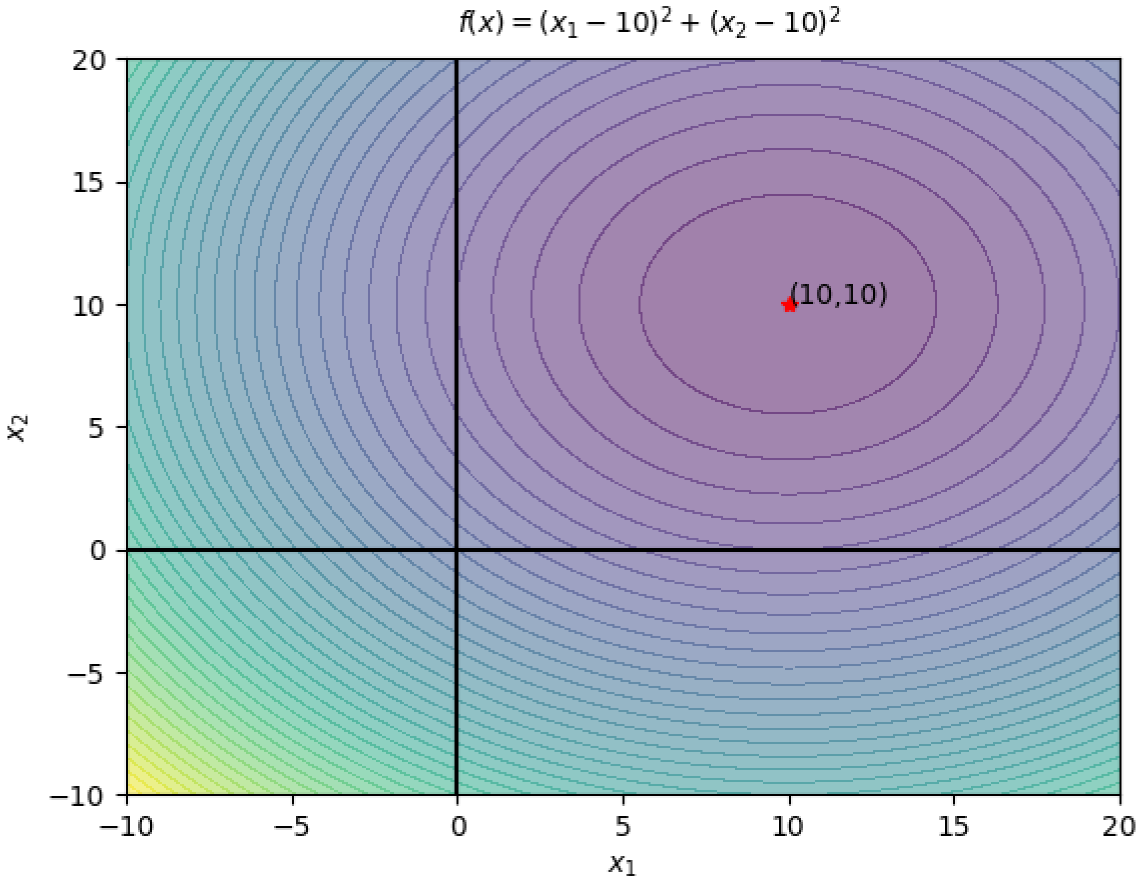
Regularized Regression

Example for a three independent variables with one dependent variable.



Do you observe something?

Regularized Regression



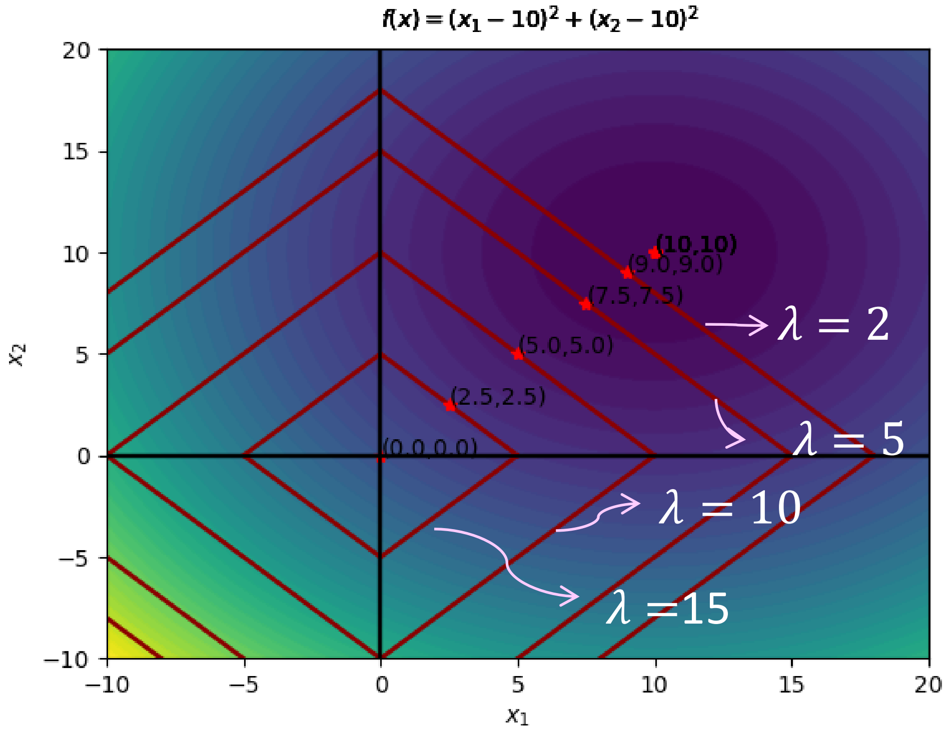
$$\min_{x1,x2}\{f(x) = (x_1 - 10)^2 + (x_2 - 10)^2\}$$

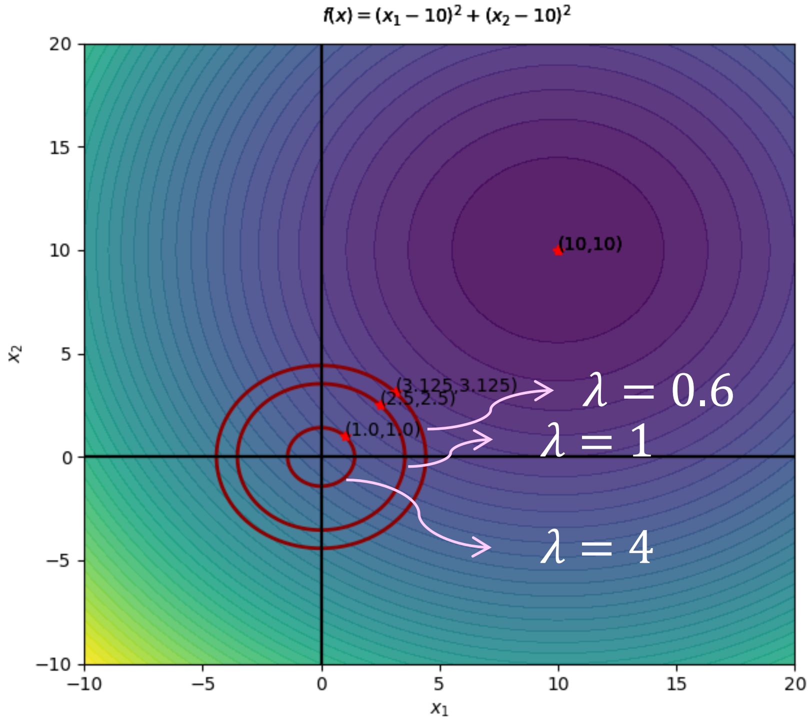
ANS: $x_1 = 10, x_2 = 10$

Regularized Regression (L1&L2)

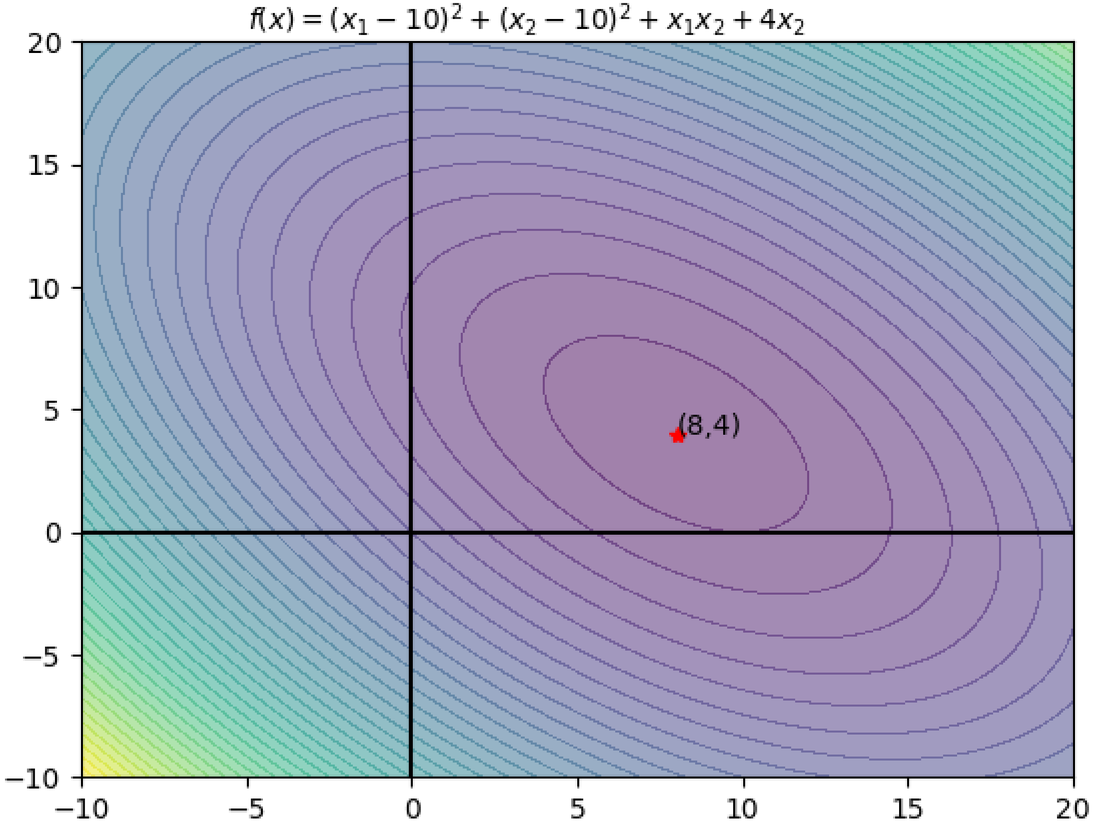
$$\min_{x1,x2}\{f(x)+\lambda\sum_{i=1}^{2}|x_i|\}$$

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



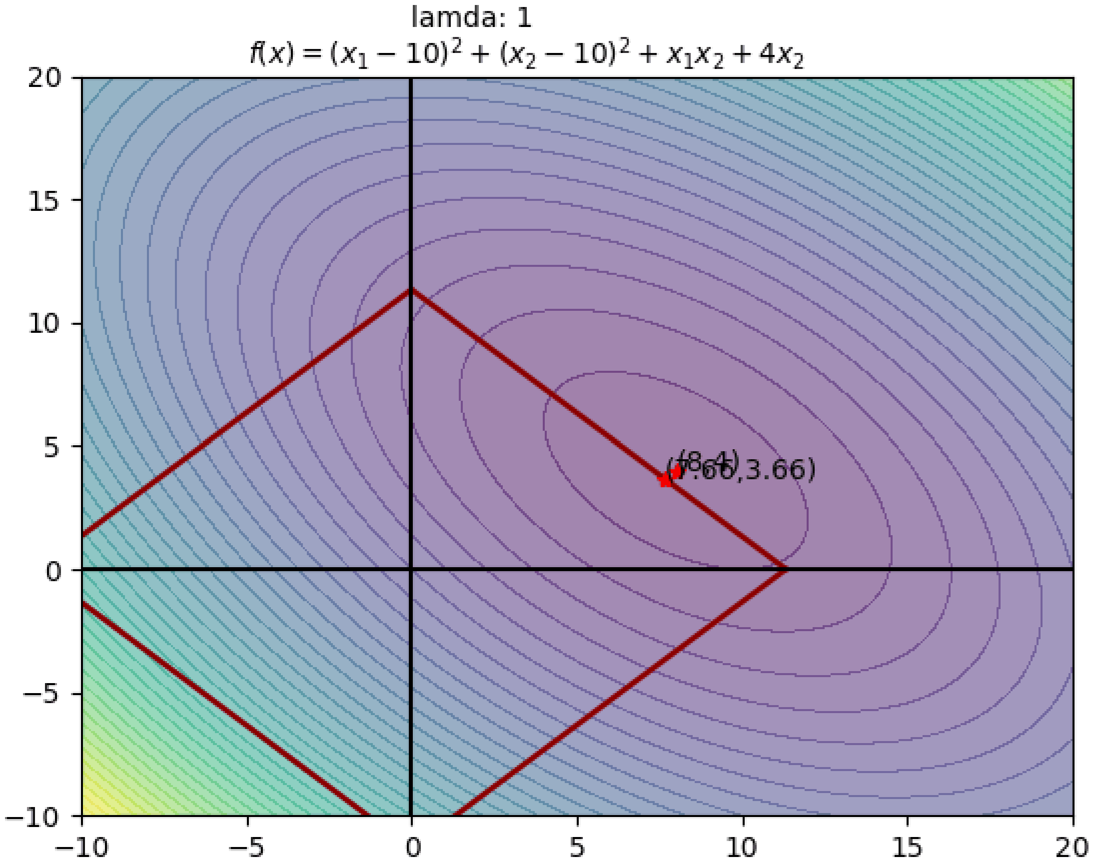


Regularized Regression



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

Regularized Regression (L1)



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

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

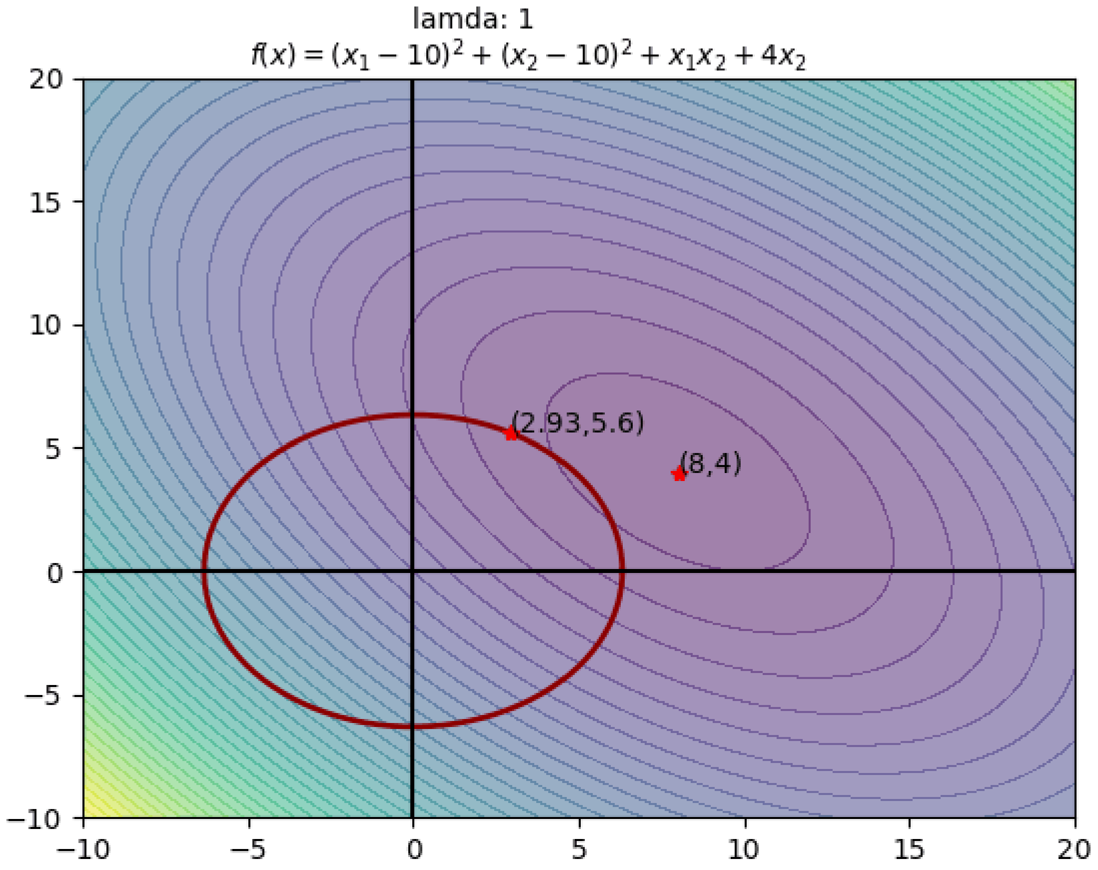
Advantage:

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

Disadvantage:

Not differentiable everywhere.

Regularized Regression (L2)



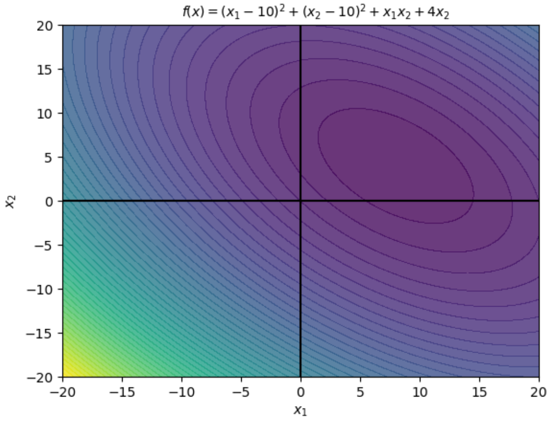
$$\min_{x1,x2} \{ f(x) + \lambda \sum_{i=1}^{2} x_i^2 \}$$

<u>Advantage:</u>

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

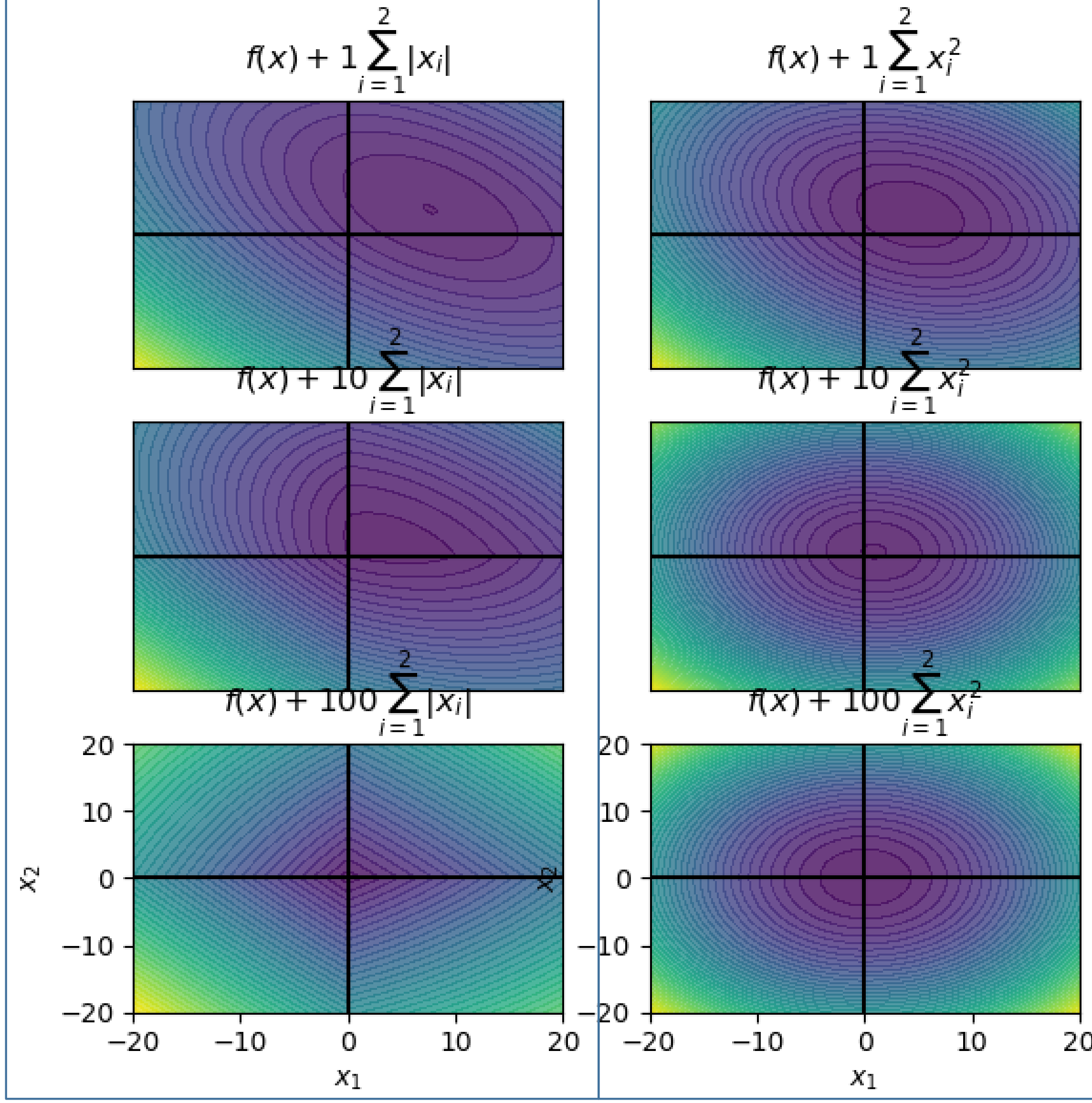
Regularized Regression

Original space
$f(x)$



L2 space
$f(x)+\lambda$L2

L2 space
$f(x)+\lambda$L1



Regularized Regression

Can we give different penalized terms for each variable?

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

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

基礎機器學習

針對前述的介紹，每個topic都介紹一個演算算法

1. Regression: Linear regression

2. Classification: Linear and Quadratic Discriminant Analysis

3. Clustering: K-means

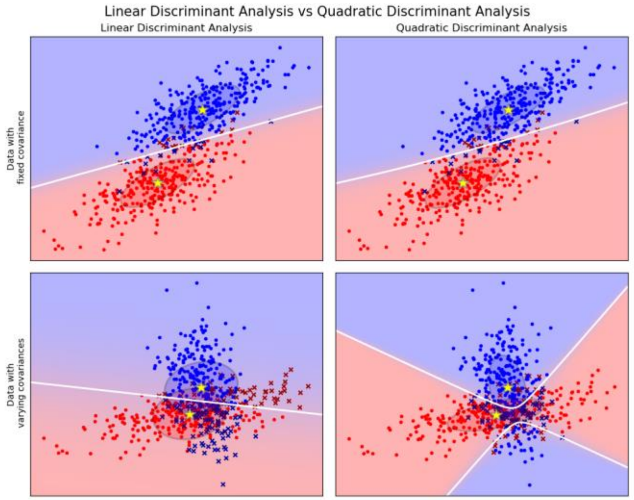
4. Dimension Reduction: PCA

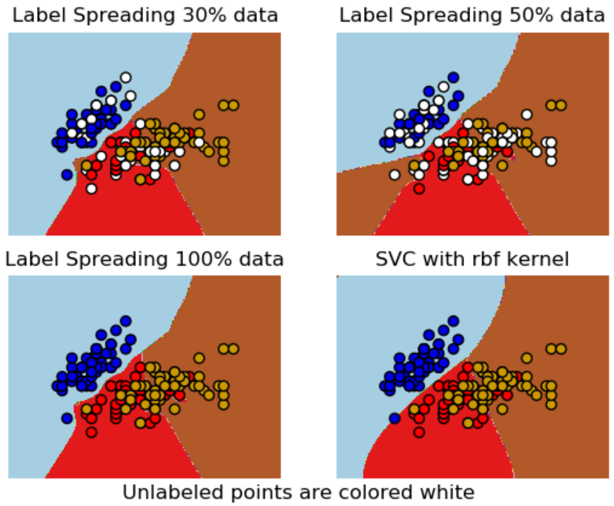
5. Ensemble learning: 不介紹。

Classification

Identifying to which category an object belongs to.

- Logistic Regression
- **Linear and Quadratic Discriminant Analysis**
- Support Vector Machine
- Nearest neighbors
- Random forest
- Neural Network





Classification

A Very simple classification problem

"How to classify {male or female} by a measured feature (body fat)?"

<u>**Collected data (body fat(%))**</u>

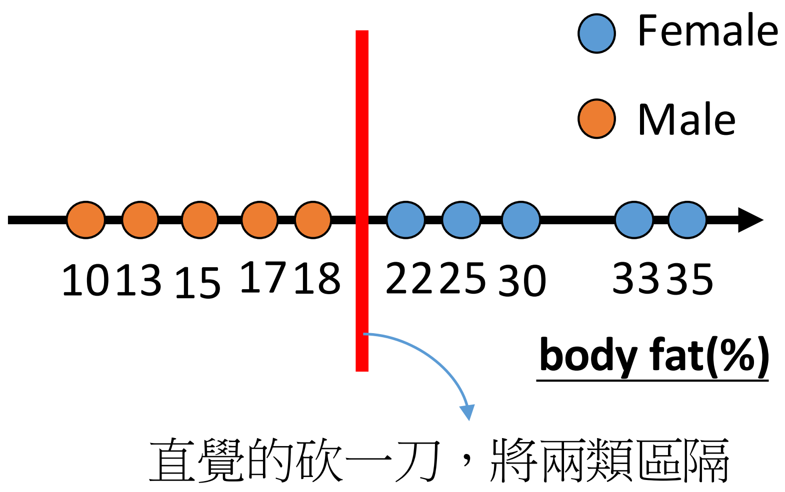
Female:{22, 25, 30, 33, 35}

Male:{ 10, 15, 20, 25, 30}

Classification

Female: {22, 25, 30, 33, 35}

Male: { 10, 13, 15, 17, 18}



1. 專家經驗決定閾(ㄩˋ)值(threshold)在體脂肪為20%。(專家系統)

2. 資料的觀察法，將資料分布畫出來，然後人工決定閾值。

3. 用資料去推算閾值在哪裡(機器學習)。

Classification

Female: {22, 25, 30, 33, 35}

Male: { 10, 13, 15, 17, 18}

The simplest way:

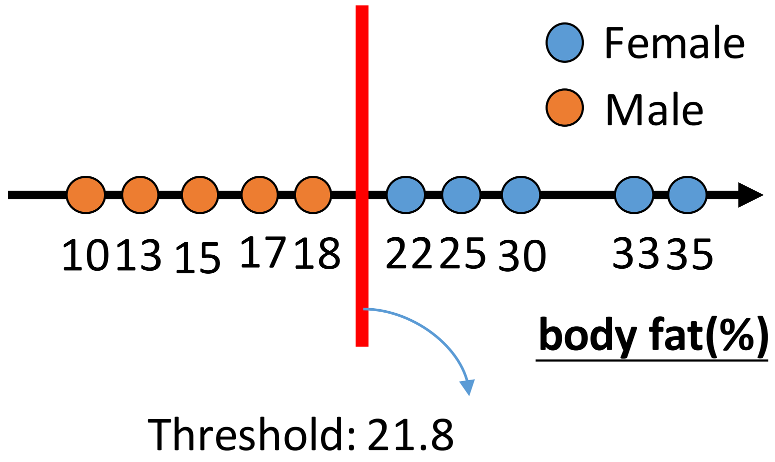
Using mean value as decision rule.

$$\frac{\text{Mean value (Female)} + \text{Mean value (Male)}}{2}$$

$$= \frac{29 + 14.6}{2} = 21.8$$

Body fat>21.8 → Female

Body fat<21.8 → Male



Classification

Female: {22, 25, 30, 33, 35}

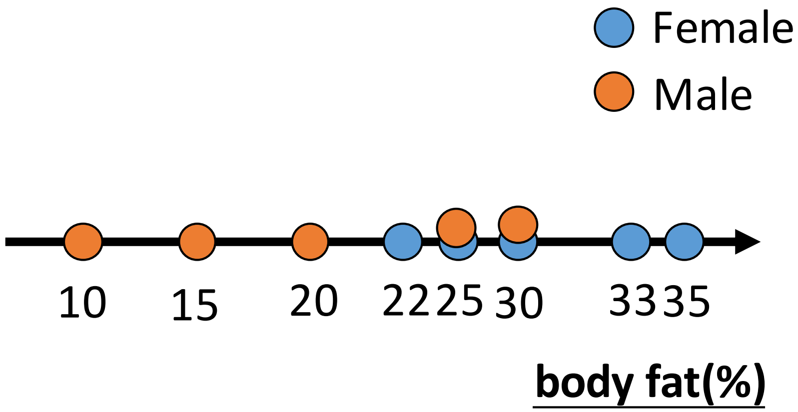
Male: { 10, 15, 20, 25, 30}

當資料有overlap的時候要怎麼辦?

1. 統計方法: type I and type II error

2. Minima risk learning

3. …



Classification

Classification by mean value

Female with 100 data, Male with 100 data (Body fat).

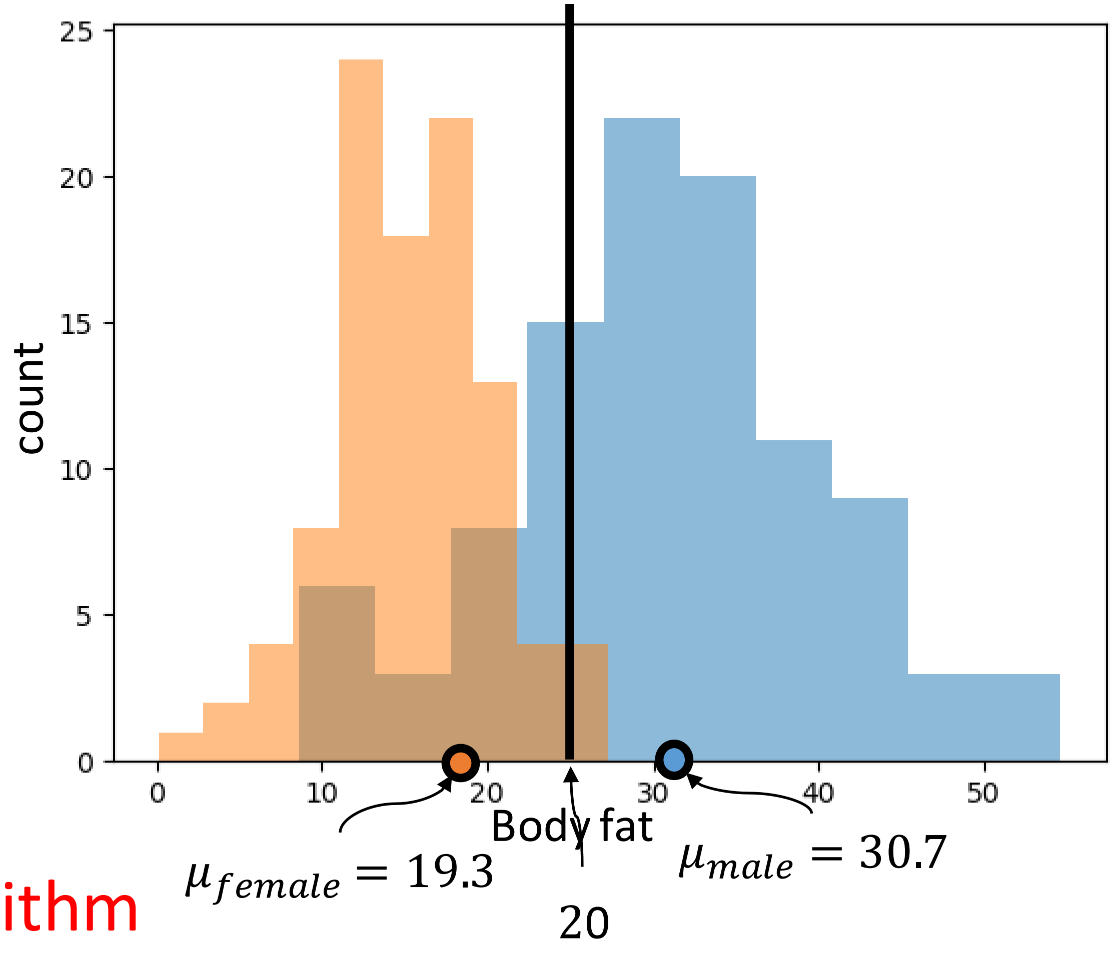
Visualization by histogram.

Blue: Male

Red: Female

$$\frac{\text{Mean value (Female)} + \text{Mean value (Male)}}{2}$$

$$= \frac{30.7 + 19.3}{2} = 25.0$$



You Just learn a classification algorithm

Classification

Classification by mean value

$\{x_i\}, \forall i, x: baby\ fat$

For a unknown label data x^*, which class it is ?

$$\mu_{male} = \frac{1}{n_{male}} \sum_{i=1}^{n_{male}} x_i\,, \qquad \mu_{female} = \frac{1}{n_{female}} \sum_{i=1}^{n_{female}} x_i\,,$$

$$f_{male}(x^*) = x^* - \mu_{male} = 40 - 30.7 = 9.3$$

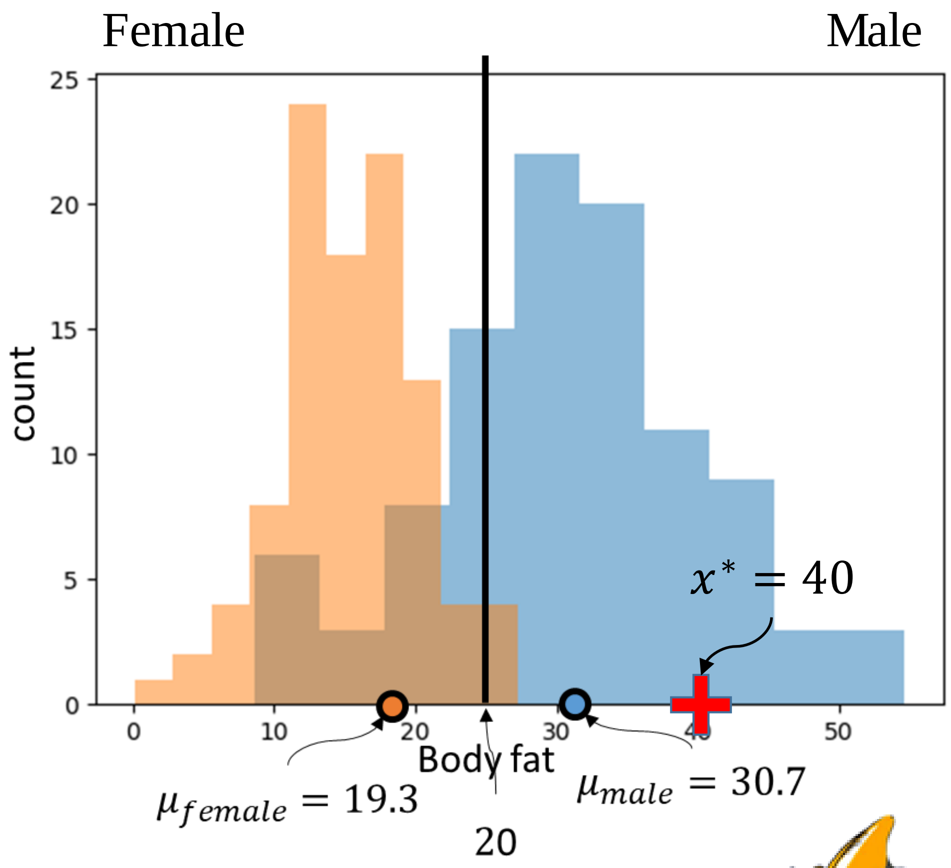
$$f_{female}(x^*) = x^* - \mu_{female} = 40 - 19.3 = 21.7$$

Decision rule: feature value(x) is closed to which class, and classify this x to which class.

Decision value $= f_{male}(x) - f_{female}(x) = 9.3 - 21.7 = -12.4$

Decision rule:

$$Decision(x) = \begin{cases} female & \text{Decision value} \geq 0 \\ male & \text{Decision value} < 0 \end{cases}$$



Classification by Density

Likelihood function

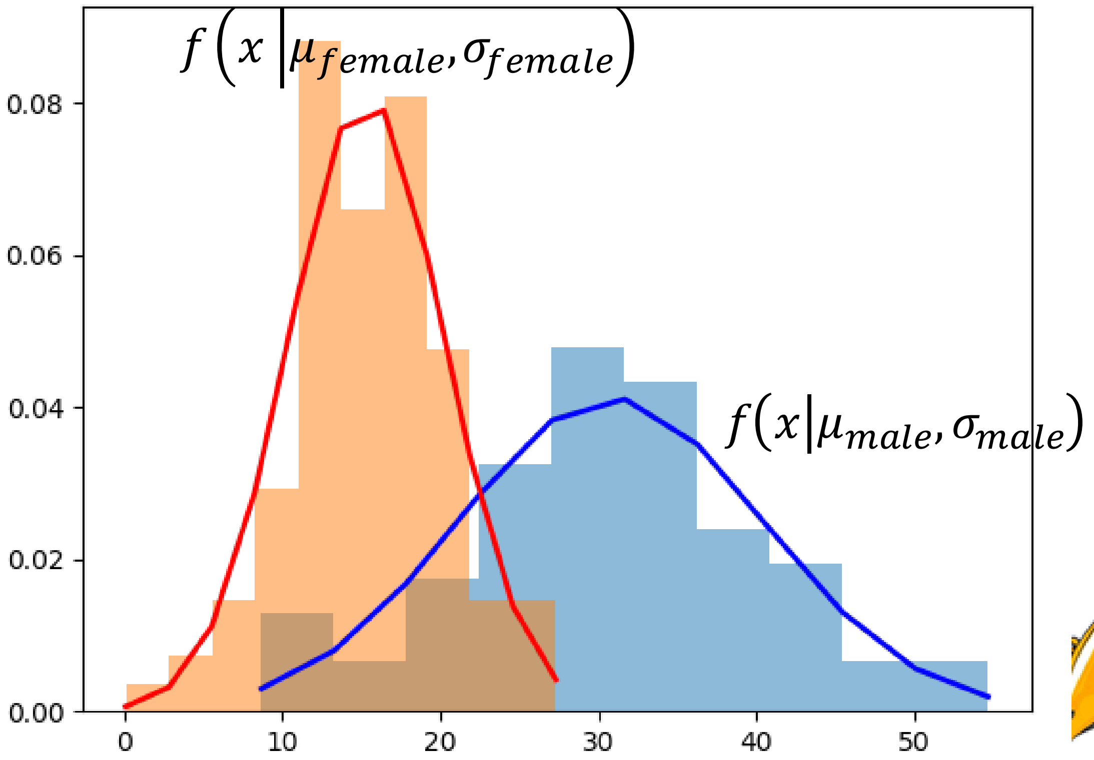
We can assume the histogram (density) is a Gaussian (normal)-like distribution.

That means

$$x_{male} \sim N(\mu_{male}, \sigma_{male})$$

$$x_{female} \sim N(\mu_{female}, \sigma_{female})$$

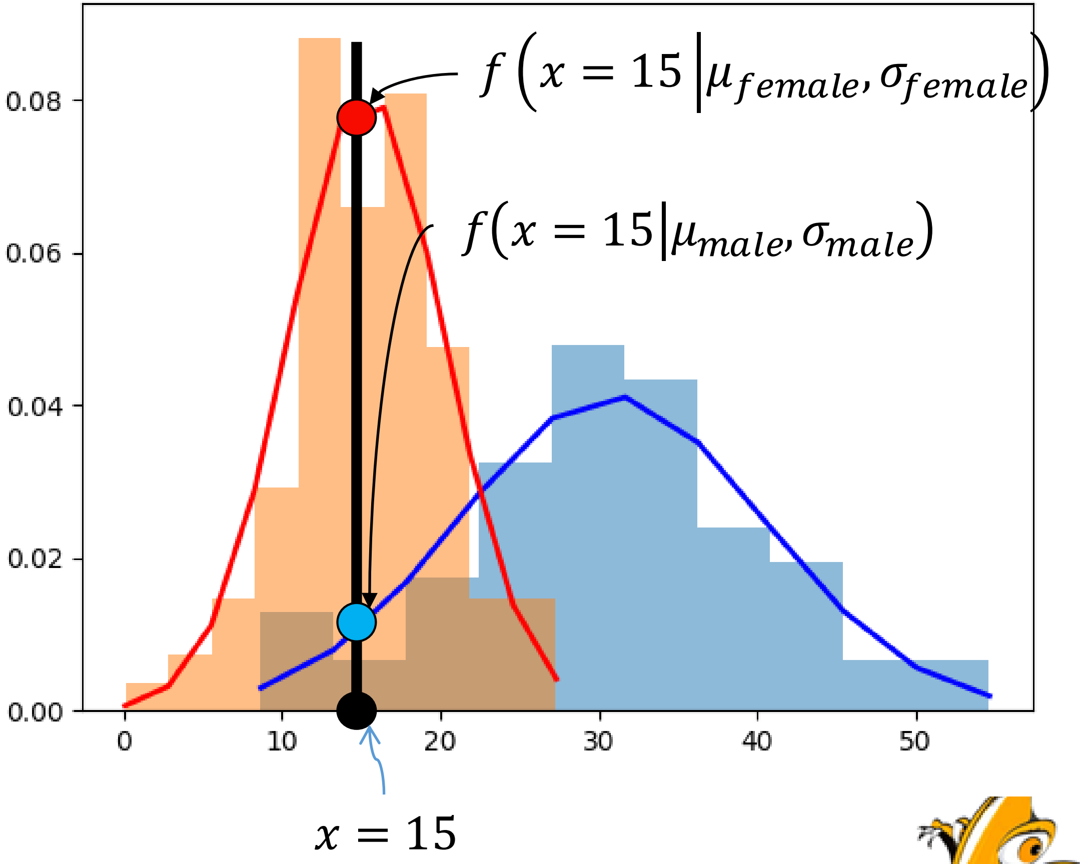
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Classification by Density

Likelihood function

$$x_{male} \sim N(\mu_{male}, \sigma_{male})$$
$$x_{female} \sim N(\mu_{female}, \sigma_{female})$$
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



A unlabeled x: 15% body fat

$$f\left(x = 15 \middle| \mu_{female}, \sigma_{female}\right)$$
$$> f(x = 15|\mu_{male}, \sigma_{male})$$

So this unlabeled x would be classify to Female.

Likelihood function

If we get multi-features (i.e. body fat and height), how to do?

$$\boldsymbol{x}_i = \begin{bmatrix} x_{bodyfat} \\ x_{height} \end{bmatrix} \qquad f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-0.5} exp\{-0.5(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

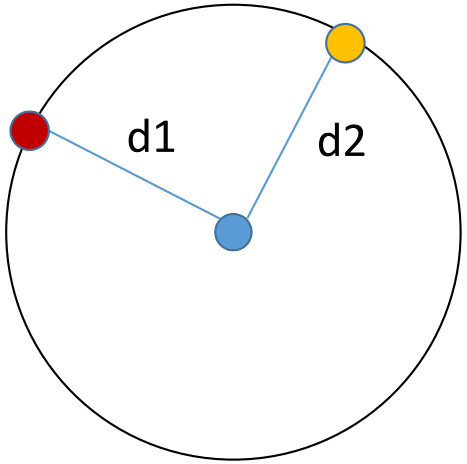
Euclidean Distance ($\boldsymbol{\Sigma} = \mathrm{I}$, Classification by mean vector) =$(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})$

Mahalanobis Distance (Classification by density) = $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$

Distance

Euclidean Distance

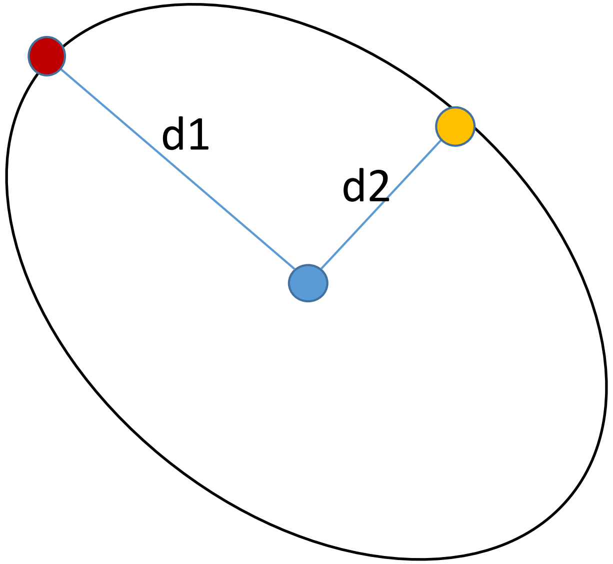
$$(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})$$



d1 = d2

Mahalanobis Distance

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$



d1 = d2

Prior probability

我們將問題改成，我們到一個百貨公司，隨機抽 一個人出來問他的體脂肪，來猜測他是男生還是女生?

$p(c)$, in Chinese = 先驗機率 (在還沒有建模前，得到的先天訊息)

Male : Female = 25: 75

Sample a people.

Female
75%

Gender?

Male
25%

Classification

- 所以除了前面提的likelihood function，我們還需要考慮先驗機率。

統計學習/機器學習上我們會採用**後驗機率**來做分類稱為Maximum a posterior (MAP)。

$p(c|\boldsymbol{x})$: posterior probability of data $\boldsymbol{x}$ for class c.

Maximum a posterior (MAP)

- $p(c|x)$: posterior probability of data $\boldsymbol{x}$ for class c.

$$p(c|\boldsymbol{x}) = \frac{p(c)p(\boldsymbol{x}|c)}{p(\boldsymbol{x})}$$

$p(c)$: prior probability for class c (前面學的先驗機率)

$p(\boldsymbol{x}|c)$: likelihood function for class c (前面學的概似函數)

$p(\boldsymbol{x}) = \sum_{c=1}^{L} p(c)\, p(\boldsymbol{x}|c)$: normalizing constant

Maximum a posterior (MAP)

$$p(c|\boldsymbol{x}) = \frac{p(c)p(\boldsymbol{x}|c)}{p(\boldsymbol{x})}$$

Female 75%: $p(c = female) = 0.75$

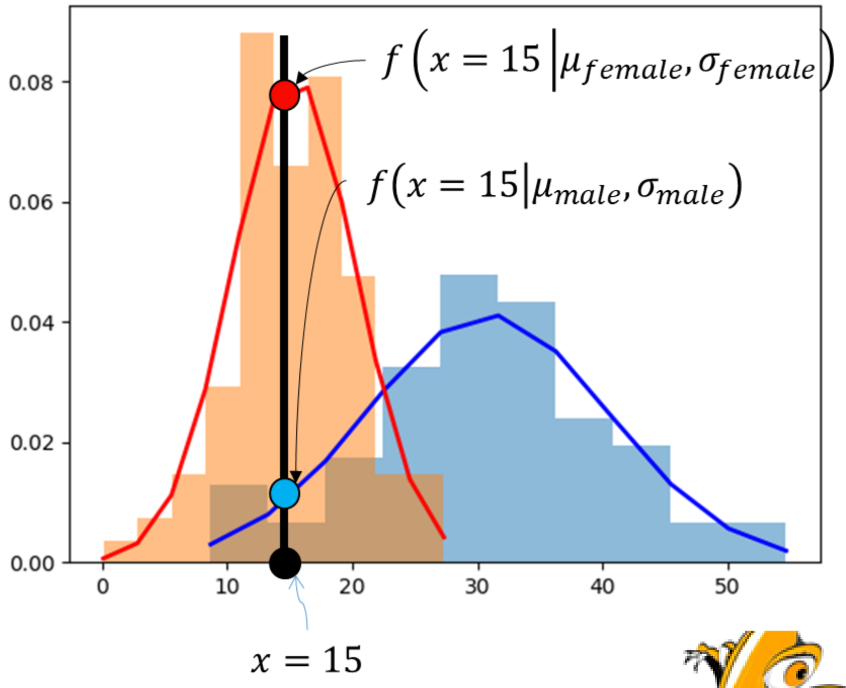
Male 25%: $p(c = male) = 0.25$

$$f(x = 15|female) = 0.075$$
$$f(x = 15|male) = 0.01$$
$$p(c = female)f(x = 15|female) = 0.75 * 0.075 = 0.05625$$
$$p(c = male)f(x = 15|male) = 0.25 * 0.01 = 0.025$$

這兩個值相乘總和不是1，機率論有說全機率要為1。

所以我們只需要除上兩個的總和全機率就為1

$$p(c = female|\boldsymbol{x} = 15) = \frac{0.05625}{0.05625 \ + 0.025} = 0.692$$
$$p(c = male|\boldsymbol{x} = 15) = \frac{0.025}{0.05625 \ + 0.025} = 0.308$$



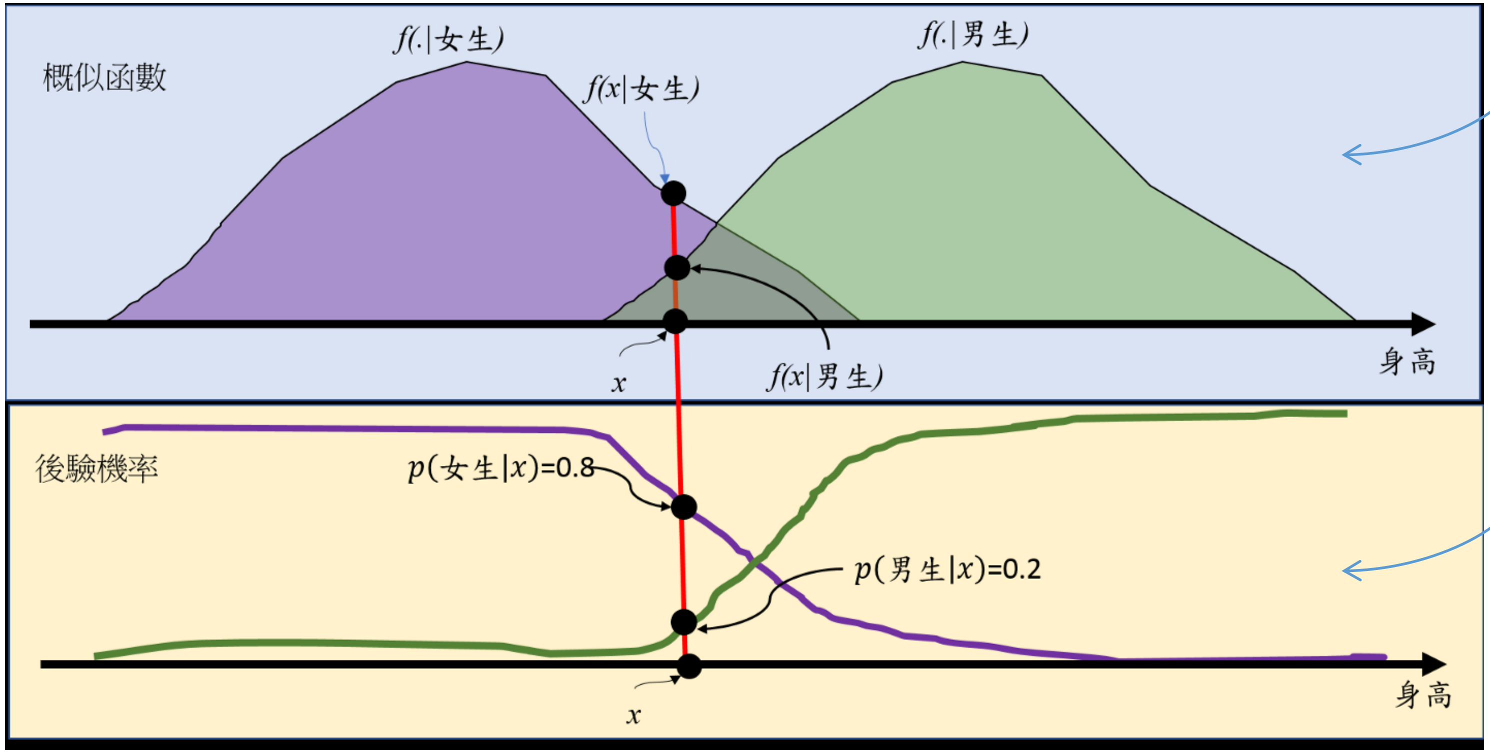
Maximum a posterior (MAP)

Posterior probability: $p(c|\boldsymbol{x}) = \dfrac{p(c)p(\boldsymbol{x}|c)}{p(\boldsymbol{x})}$

Likelihood : $f(\boldsymbol{x}|\mu,\sigma)$

Likelihood function

Posterior probability



Maximum a posterior (MAP)

How to make decision in an L-classes classification problem?

By checking the posterior probability for all class.

$$\text{Decision}(\boldsymbol{x}) = \underset{c=\{1,2,\dots,L\}}{arg\max}\{p(c|\boldsymbol{x})\}$$

$$\underset{c=\{1,2,\dots,L\}}{arg\max}\{p(c)p(\boldsymbol{x}|c)\}$$

MAP with Gaussian function

Gaussian function:

$$f(x|\mu_c,\Sigma_c)=(2\pi)^{-d/2}\left|\Sigma_c\right|^{-0.5}\exp\left\{-0.5(x-\mu_c)^T\Sigma^{-1}(x-\mu_c)\right\}$$

MAP:

$$c_{MAP}=\mathop{\arg\max}_{c=\{1,2,\ldots,L\}}\{p(c|x)\}=\mathop{\arg\max}_{c=\{1,2,\ldots,L\}}\left\{p(c)f(x|c)\right\}$$

MAP with Gaussian function

Gaussian function + MAP:

$$
\begin{aligned}
c_{MAP}^{GC} &= \underset{c=\{1,2,\ldots,L\}}{\arg\max}\{p(c|x)\} = \underset{c=\{1,2,\ldots,L\}}{\arg\max}\left\{p(c)f(x|c)\right\} \\
&= \underset{c=\{1,2,\ldots,L\}}{\arg\min}\{-2\ln(p(c)) + \ln(|\Sigma_c|) - 0.5(x-\mu_c)^T\Sigma^{-1}(x-\mu_c)\}
\end{aligned}
$$

MAP with Gaussian function

$$c_{MAP}^{GC} = \underset{c=\{1,2,\ldots,L\}}{\arg\min}\{-2\ln(p(c)) + \ln(|\Sigma_c|) - 0.5(x-\mu_c)^T\Sigma^{-1}(x-\mu_c)\}$$

The most important term of this formula is measuring the distance between x and center of distribution

$$\text{QDC:}\ \ w_{MAP} = \underset{i=\{1,2,\ldots,L\}}{\arg\max}\{lnp\left(w_i\right) - 0.5ln\left|\boldsymbol{\Sigma}_i\right| - 0.5\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)^T\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)\}$$

$$\text{LDC:}\ \ w_{MAP} = \underset{i=\{1,2,\ldots,L\}}{\arg\max}\{lnp\left(w_i\right) - 0.5\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)\}$$

$$\text{MEC:}\ \ w_{MAP} = \underset{i=\{1,2,\ldots,L\}}{\arg\max}\{lnp\left(w_i\right) - 0.5\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)^T I_d\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)\}$$

I_d: Identity matrix $(d\times d)$

Classification

We just learned **model-based** algorithm.

Model-based: data is assumed following the normal distribution. (parameters: mean vector and covariance matrix)

Can we learn without model (model-free)?

ANS: Yes. Nearest neighbors, SVM, neural network.