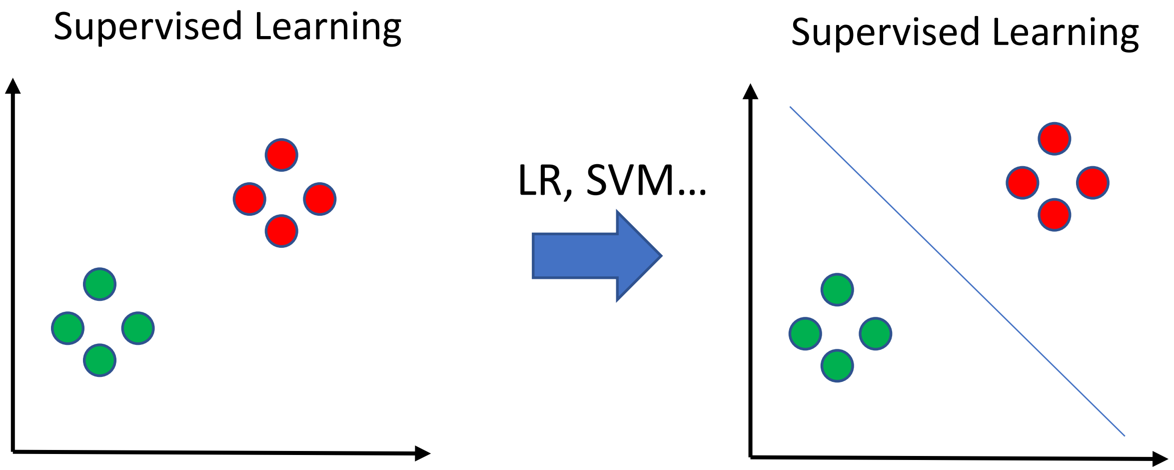
Linear Regression

Hongchang Gao

Spring 2024

Supervised Learning

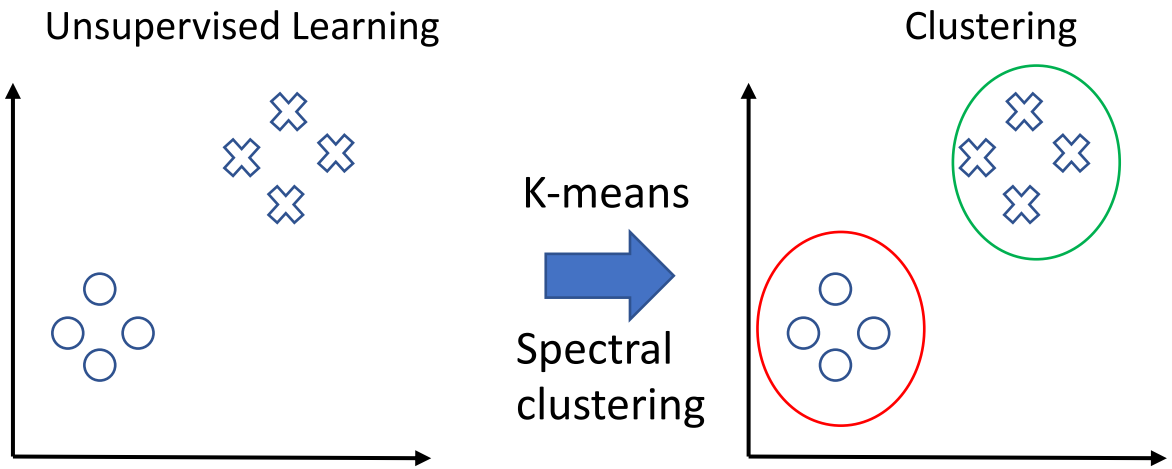
- Supervised Learning:
  - Given: feature and label/response/ground truth of the sample
  - Goal: predict samples' labels



Know ALL labels

Unsupervised Learning

- Unsupervised Learning
  - Given: only features, NO labels
  - Clustering: find meaningful groups of samples s.t.
    - Samples in the same group are “similar”
    - Samples in different groups are “dissimilar”



Know NOTHING about labels

Supervised Learning

- Supervised Learning Methods:

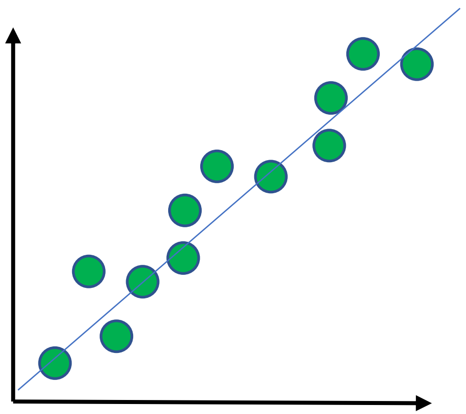
Given n samples: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

Learn a mapping function $x_i \xrightarrow{f(x)} y_i$

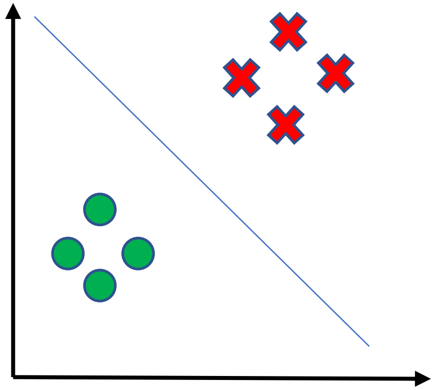
Y is continuous: regression

Y is discrete: classification

Regression



classification



Supervised Learning

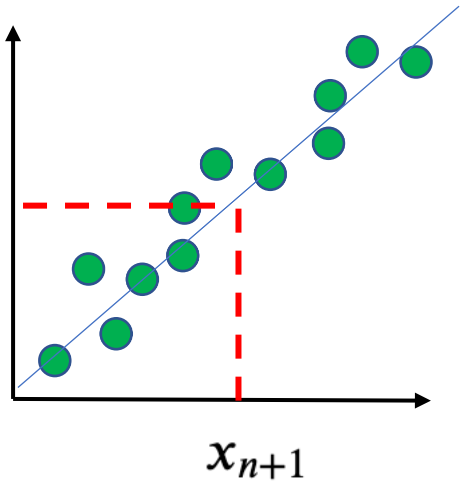
- Supervised Learning Methods:

Given n samples: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

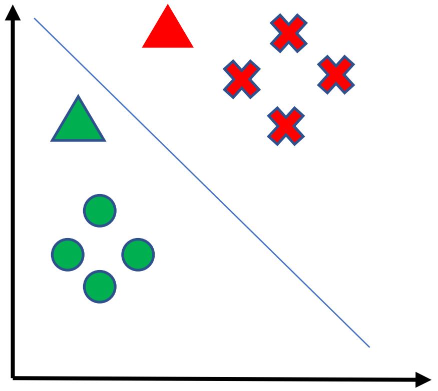
Learn a mapping function $x_i \xrightarrow{f(x)} y_i$

Given a new sample: $x_{n+1} \xrightarrow{f(x)} ?$

Regression

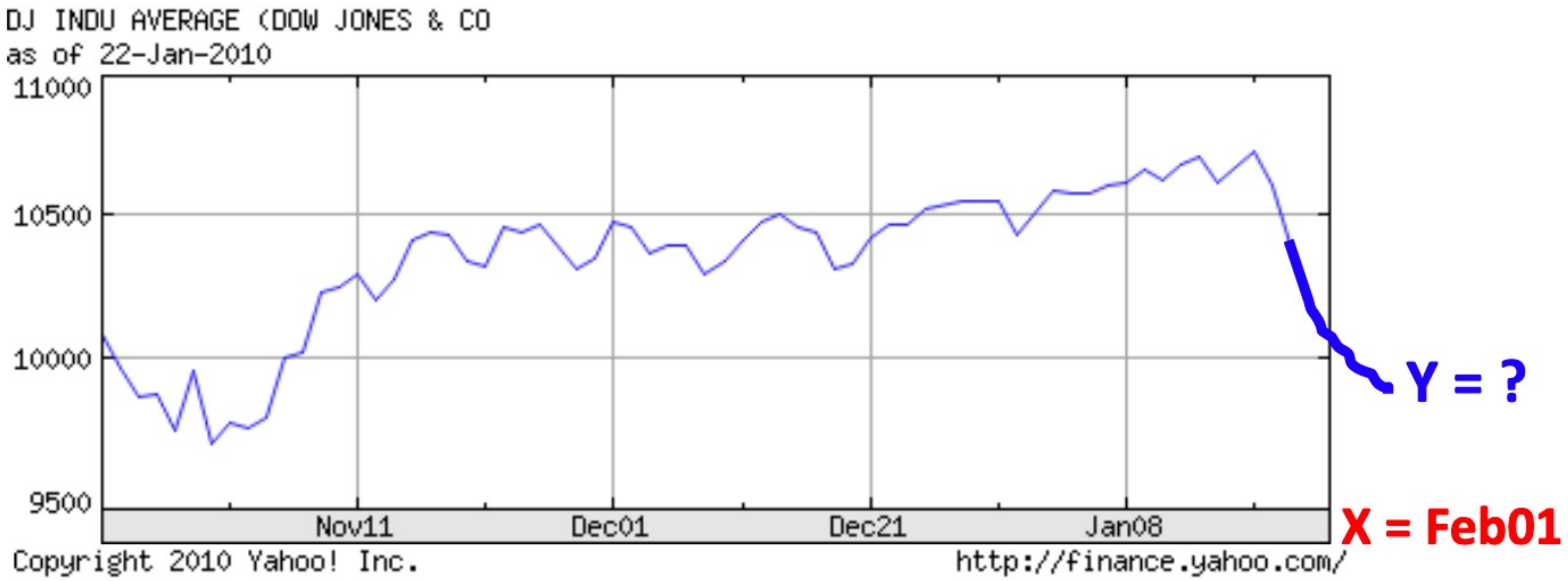


classification



Regression

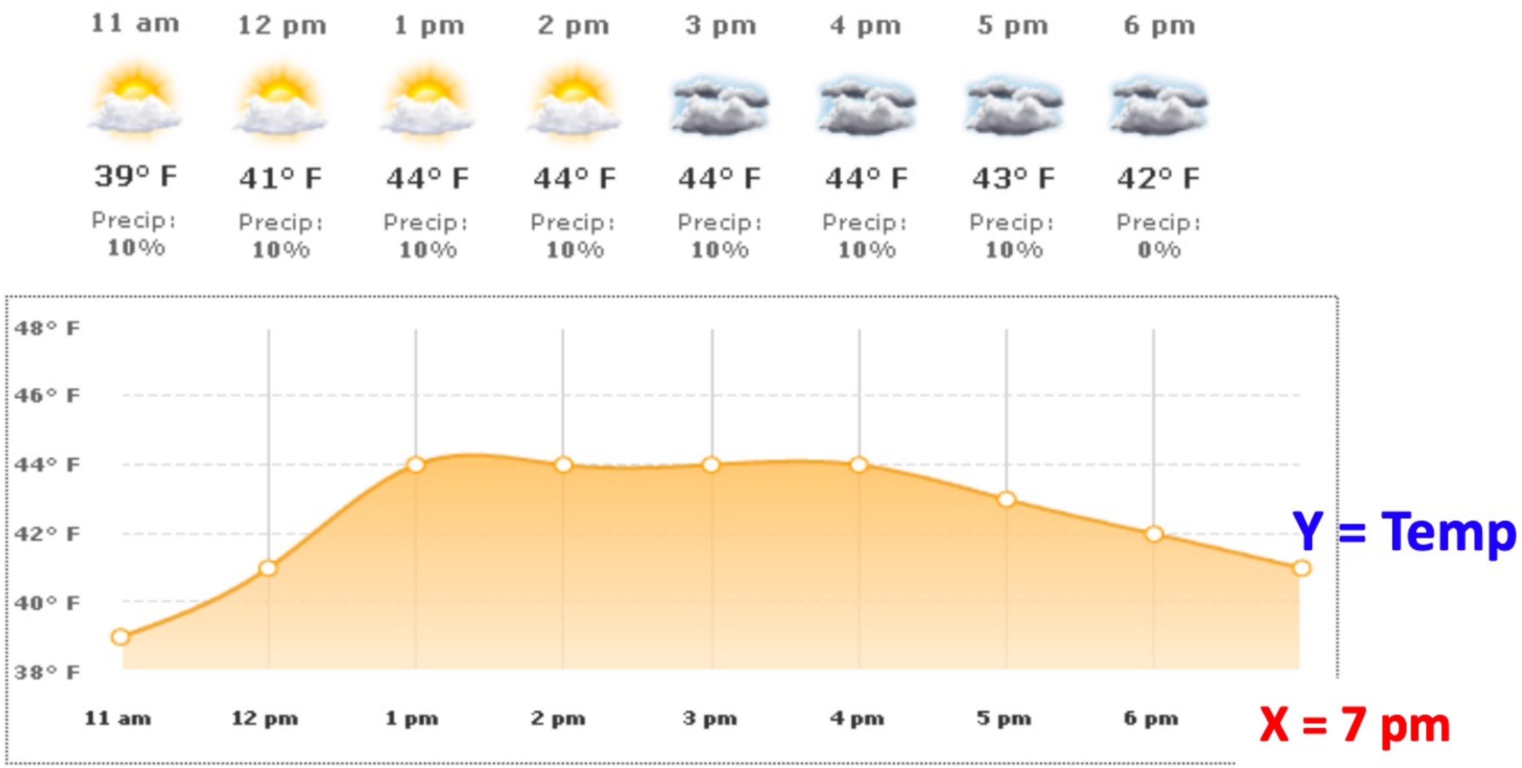
- Stock Market Prediction



Ref: https://www.cs.cmu.edu/~ninamf/courses/401sp18/lectures/linear_regression-2018.pdf

Regression

- Weather Prediction



Ref: https://www.cs.cmu.edu/~ninamf/courses/401sp18/lectures/linear_regression-2018.pdf

Linear Regression : build model

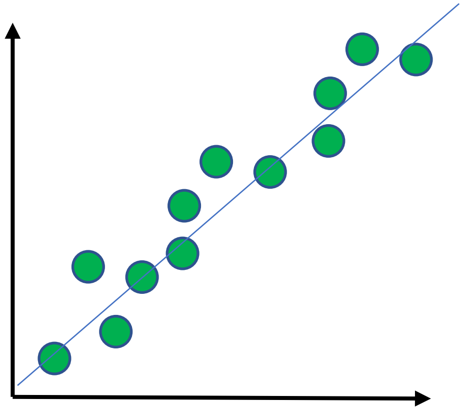
- Linear regression
  - Learn a mapping function $x_i \xrightarrow{f(x)} y_i$
  - $f(x)$ is a linear combination of input features

    $$f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$$

    - $x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,d})$ is the feature vector of the i-th sample
    - $w = (w_0, w_1, w_2, \cdots, w_d)$ is the model parameter
    - w_0 is bias

$$f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d} = w^T x_i$$

Linear Regression



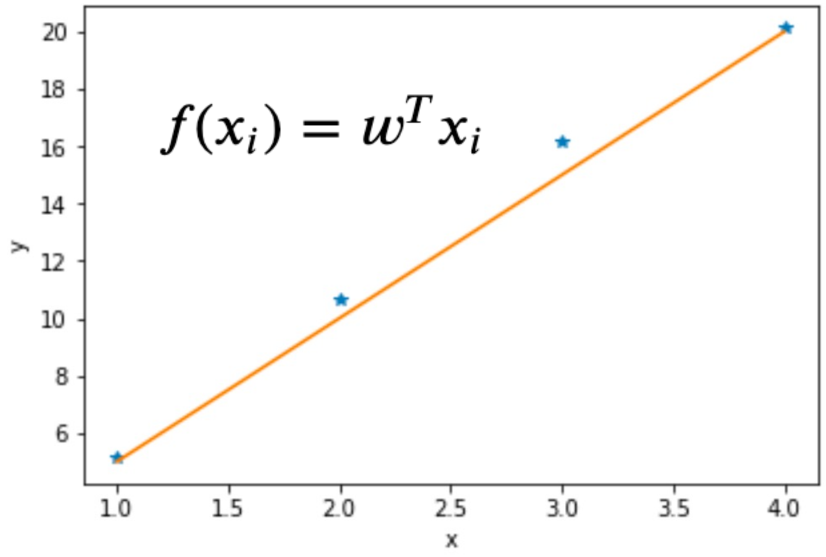
Linear Regression : build model

- Linear regression
  - $f(x)$ is a linear combination of input features

$$f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$$

x	y
1	5.14
2	10.67
3	16.17
4	20.12

$$f(x_1) = w_0 + w_1 * 1$$
$$f(x_2) = w_0 + w_1 * 2$$
$$f(x_3) = w_0 + w_1 * 3$$
$$f(x_4) = w_0 + w_1 * 4$$



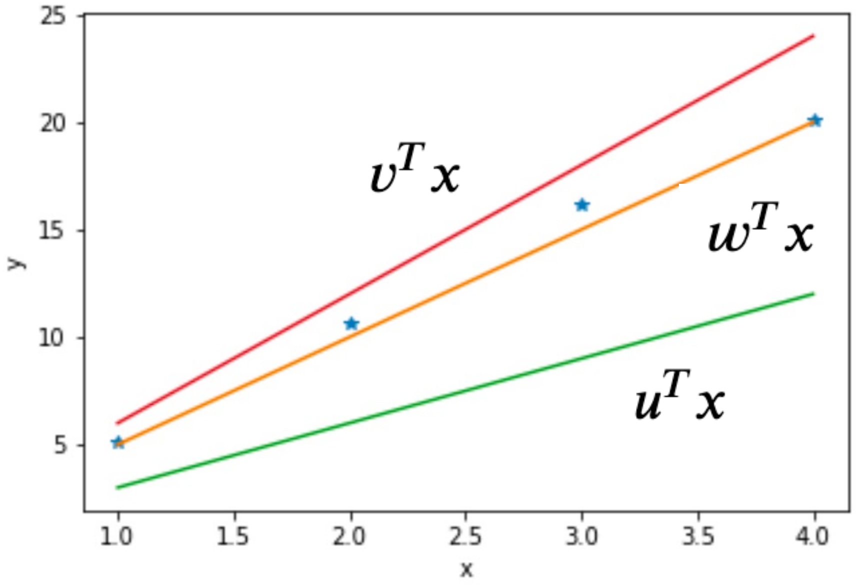
Linear Regression : optimize model

- How to determine the model parameter?
  - $f(x)$ is a linear combination of input features

$$f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$$

x	y
1	5.14
2	10.67
3	16.17
4	20.12

$$f(x_1) = w_0 + w_1 * 1$$
$$f(x_2) = w_0 + w_1 * 2$$
$$f(x_3) = w_0 + w_1 * 3$$
$$f(x_4) = w_0 + w_1 * 4$$



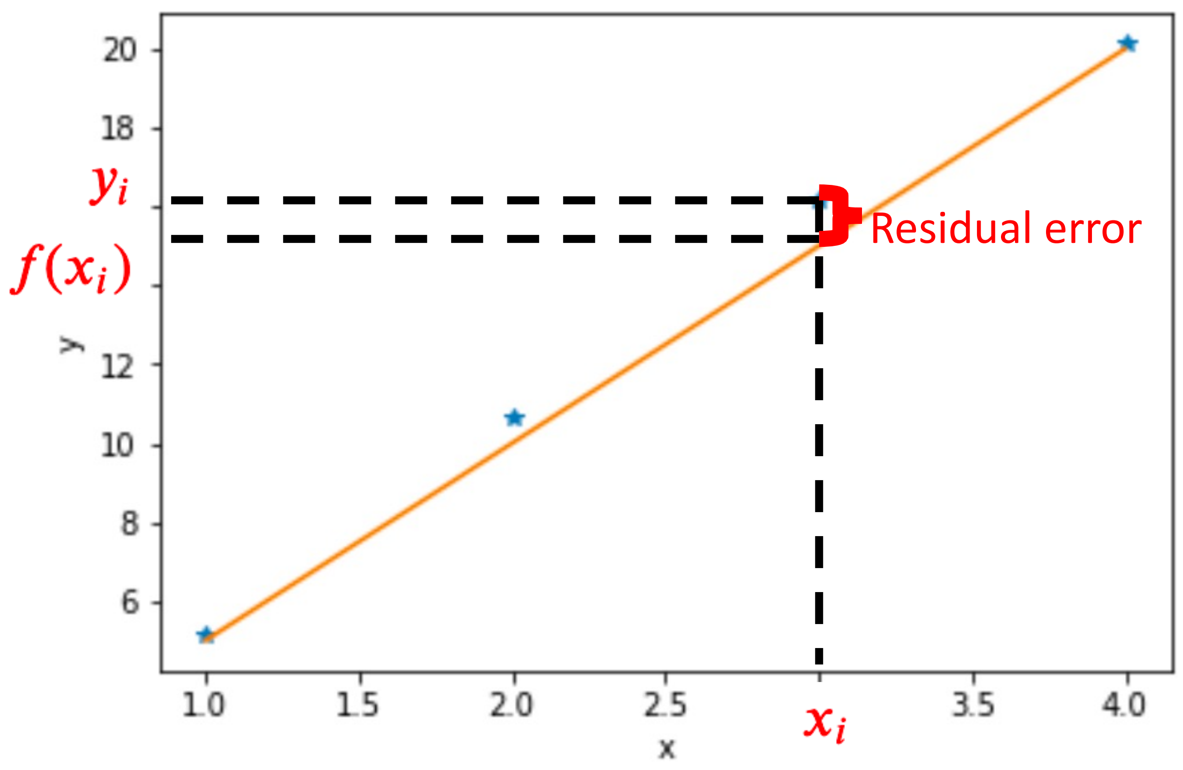
Linear Regression : optimize model

- How to determine the model parameter?
  - Residual error/prediction error

$$y_i - f(x_i)$$

  - Minimize residual error

Loss function $\min_w \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$



Linear Regression : optimize model

- How to determine the model parameter?
  - Residual error

$$y_i - f(x_i)$$

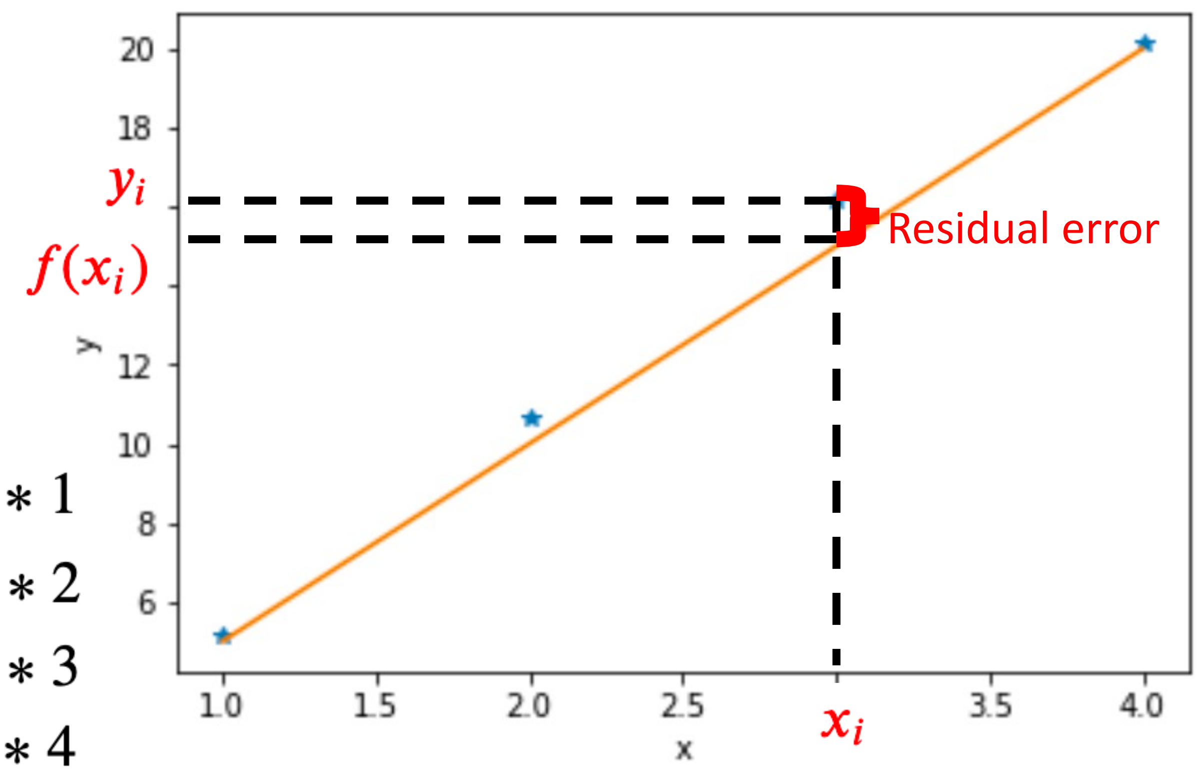
  - Minimize residual error

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Loss function

x	y
1	5.14
2	10.67
3	16.17
4	20.12

$$f(x_1) = w_0 + w_1 * 1$$
$$f(x_2) = w_0 + w_1 * 2$$
$$f(x_3) = w_0 + w_1 * 3$$
$$f(x_4) = w_0 + w_1 * 4$$



Linear Regression : optimize model

- Vector representation $y_{n\times 1}$ $X_{n\times(d+1)}$ $w_{(d+1)\times 1}$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad f(\mathbf{X}) = \mathbf{X}\mathbf{w} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & a_{2d} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} \sum_{j=0}^{d} x_{1j} w_j \\ \sum_{j=0}^{d} x_{2j} w_j \\ \vdots \\ \sum_{j=0}^{d} x_{mj} w_j \end{pmatrix}$$

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \Longrightarrow \min_w \frac{1}{n} \|y - Xw\|_2^2$$

Linear Regression : optimize model

- How to get the optimal parameter?
  - gradient of the loss function with respect to the model parameter should be 0
    - The gradient represents the slope of the loss function curve

$$\mathcal{L}(w) = \frac{1}{n}\|y - Xw\|_2^2$$

$$\Downarrow$$

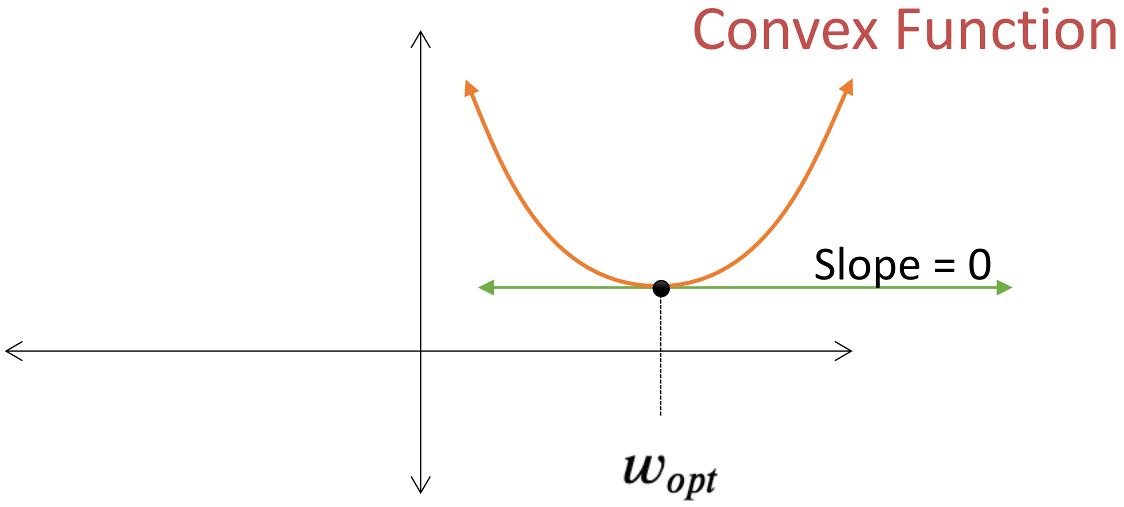
$$\frac{\partial \mathcal{L}(w)}{w} = -\frac{2}{n}X^T(y - Xw) = 0$$

$$\Downarrow$$

$$X^T X w = X^T y$$

$$\Downarrow$$

$$w = (X^T X)^{-1} X^T y$$



Linear Regression : optimize model

- Example

x	y
1	5.14
2	10.67
3	16.17
4	20.12

$$\mathcal{L}(w) = \frac{1}{n}\|y - Xw\|_2^2$$

$$\Downarrow$$

$$w = (X^T X)^{-1} X^T y$$

$$\Downarrow$$

$$w = \left[\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}\right]^{-1}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}\begin{pmatrix} 5.14 \\ 10.67 \\ 16.17 \\ 20.12 \end{pmatrix}$$

Linear Regression : evaluation

- 1. Mean Absolute Error (MAE)

$$\frac{1}{n}\sum_{i=1}^{n}|e_i|$$

- 2. Mean Squared Error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}e_i^2$$

- 3. Root Mean Squared Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}$$