Linear Regression

Data Mining

Prof. Sujee Lee

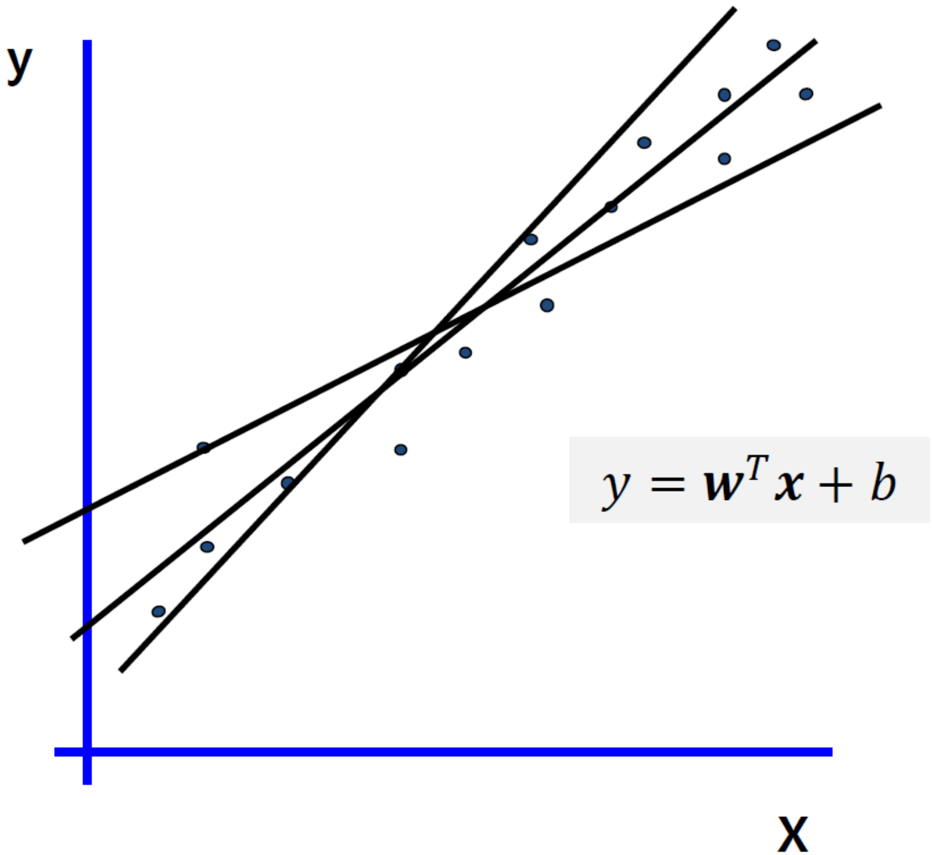
Department of Systems Management Engineering

Sungkyunkwan University

Linear Regression

Linear Regression

- **Linear Regression**
  - Linear regression models $\hat{y} = f(\boldsymbol{x})$ as a linear function of input features.



Learning a Linear Regression Model

- **Linear regression on single variate data** Assume $y_i \in \mathbb{R}, x_i \in \mathbb{R}$
  - Goal : Given data pairs $\{(x_i, y_i)\}_{i=1}^{n}$, find a "best fit" line through data
  - Model : $\hat{y} = f(x) = ax + b$
  - Least Squares :

$$\min_{a,b} \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2$$

Learning a Linear Regression Model

- To solve the following minimization problem (optimization problem)

$$\min_{a,b} \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2$$

  - take derivatives :

- System of linear equations :

- Solution :

$$a^* = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \quad b^* = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Learning a Linear Regression Model

- Q: How do we generalize to more dimensions?

- **Linear regression on multi variate data**
  - $y_i \in \mathbb{R}$, $\boldsymbol{x}_i = \left(x_{i,1}, x_{i,1}, \cdots, x_{i,d}\right) \in \mathbb{R}^d$
  - Goal : find best linear predictor of y_i using all components of vector $\boldsymbol{x}_i$
    - use $w_1\, x_{i,1} + \cdots + w_d x_{i,d} + b$ to predict y_i
  - we need to solve

$$\min_{w_0, w_1, w_2, \cdots, w_d} \sum_{i=1}^{n} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

Let $\quad f(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$

Take a gradient of $f(\boldsymbol{w})$ and set it equals to 0 : $\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = 0$

$$
\begin{aligned}
\nabla_{\boldsymbol{w}} \textstyle\sum_{i=1}^{n} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 \quad &= \textstyle\sum_{i=1}^{n} \nabla_{\boldsymbol{w}} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 \\
&= \textstyle\sum_{i=1}^{n} 2(y_i - \boldsymbol{w}^T \boldsymbol{x}_{\boldsymbol{i}}) \nabla_{\boldsymbol{w}} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i) \\
&= \textstyle\sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^T \boldsymbol{x}_i) \nabla_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{x}_i
\end{aligned}
$$

gradient of $\boldsymbol{w}^T \boldsymbol{x}_i$, compute partial derivative w.r.t each w_j

$$\frac{\partial}{\partial w_j} (\boldsymbol{w}^T \boldsymbol{x}_i) = \frac{\partial}{\partial w_j} \left(\textstyle\sum_{k=0}^{d} w_k x_{i,k} \right) = x_{i,j}, \forall j$$

$$\nabla_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{x}_i = \boldsymbol{x}_i$$

Thus, $\quad \nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = 0 \iff \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^T \boldsymbol{x}_i) \boldsymbol{x}_i = 0$

Solve for $\boldsymbol{w}$?

$$\sum_i -2(y_i - \boldsymbol{w}^T \boldsymbol{x}_i)\boldsymbol{x}_i = 0$$

$$\sum_i (\boldsymbol{w}^T \boldsymbol{x}_i)\boldsymbol{x}_i = \sum_i y_i \boldsymbol{x}_i$$

$$\left(\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)\boldsymbol{w} = \sum_i y_i \boldsymbol{x}_i$$

Ordinary Least Square estimate:

$$\therefore \boldsymbol{w}_{OLS} = \left(\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \cdot \sum_i y_i \boldsymbol{x}_i$$

Does it agree with 'a' and 'b' in 1-dim case?

$$\hat{y} = \boldsymbol{w}_{OLS}^T \boldsymbol{x}$$

Learning a Linear Regression Model

- **Matrix Representation**

Consider the matrix & vector $\boldsymbol{X} \in \mathbb{R}^{n\times(d+1)}, \boldsymbol{y} \in \mathbb{R}^n$

Least squares objective function can be written as

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2$$

Recall: if $\boldsymbol{v} \in \mathbb{R}^n$, then $||\boldsymbol{v}||_2^2 = \sum_{j=1}^{n} v_j^2$

Now, we want to $\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$

want to take a gradient of f :

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 = \nabla_{\boldsymbol{w}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$$= \nabla_{\boldsymbol{w}} [\boldsymbol{y}^T \boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w})^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} + (\boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{X}\boldsymbol{w})]$$

$$= \nabla_{\boldsymbol{w}} [\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w}]$$

$$= -2\boldsymbol{X}^T \boldsymbol{y} + \nabla_{\boldsymbol{w}} (\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w})$$

Claim: For any symmetric matrix $\boldsymbol{A}$, $\nabla_{\boldsymbol{\beta}} (\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}) = 2\boldsymbol{A}\boldsymbol{\beta}$

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} = 0$$

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = 0$$

Solve for $\boldsymbol{w}$:

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\therefore \boldsymbol{w}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad \boldsymbol{w}_{OLS} = \left(\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} \cdot \sum_i y_i\boldsymbol{x}_i$$

Linear Regression

- **Linear Regression (ordinary least squares (OLS))**
  - Linear regression finds the parameters $\mathbf{w}$ and b that minimize the mean squared error between predictions and the true regression targets on the training set.

$$\hat{y} = \mathbf{w}^T \boldsymbol{x} + b = w_1\, x_1 + \cdots + w_d x_d + b$$

$$\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d, \qquad y, \hat{y} \in \mathbb{R}$$

  - It solves

$$\min_{w} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

  - The solution is

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

  - The trained model is

$$f(\boldsymbol{x}) = \boldsymbol{w}^{*T}\boldsymbol{x}$$

  - Linear regression has no hyperparameters, thus has no way to control model complexity.

Other Views

Linear Regression

- **Linear Regression (ordinary least squares (OLS))**
  - Linear regression solves

$$\min_{w} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

  - The solution is

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

  - The trained model is

$$f(\boldsymbol{x}) = \boldsymbol{w}^{*T}\boldsymbol{x}$$

- **Three views of ordinary least squares**
  - Algebraic (matrices, gradients)
  - Geometric
  - Probabilistic

Probabilistic Approach for OLS

- **Probabilistic Approach for OLS**
  - Idea: Instead of thinking just in terms of data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ or data matrices $(\boldsymbol{X}, \boldsymbol{y})$, considers a generative model (probabilistic)
  - Consider a setting where $(\boldsymbol{x}_i, y_i)$ are generated in a random way:

$$y_i = \mathbf{w}^T \boldsymbol{x}_i + \epsilon_i, \quad \text{where } \boldsymbol{x}_i\text{'s are fixed vectors and } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

ϵ_i is normally distributed in $\mathbb{R}$, with variance σ^2

    - Start with $\boldsymbol{x}_i$, generate random ϵ_i, which gives y_i
  - Goal : Estimate $\mathbf{w}$ using the observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$

Probabilistic Approach for OLS

- **Maximum Likelihood Estimation (MLE)**
  - We will use maximum likelihood estimation (MLE) to find our estimate of $\mathbf{w}$
  - We write down the probability of seeing $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, assuming $\mathbf{w}$ was true regression vector, and maximize it over $\mathbf{w}$

Likelihood (probability of seeing the data):

$$L_w(\boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \boldsymbol{w}^T \boldsymbol{x_i})^2}{2\sigma^2}\right)$$

Recall: p.d.f. of a random variable x following $\mathcal{N}(\mu, \sigma^2)$ is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$L_{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{y}) = \mathrm{P}(\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n})$$

$$= \prod_i^n \mathrm{P}(\boldsymbol{x}_i, y_i)$$

$$= \prod_i^n \mathrm{P}(\epsilon_i = y_i - w^T x_i)$$

$$= \prod_i^n p(y_i - w^T x_i)$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2\right)$$

Now, we want to maximize likelihood $L_{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{y})$ with respect to $\boldsymbol{w}$

$$L_{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{y}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x_i})^2\right)$$

Maximizing likelihood $L_{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{y})$ with respect to $\boldsymbol{w}$ is the same as minimizing $\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x_i})^2$ with respect to $\boldsymbol{w}$ **(this is the same as OLS)**

Pros:

- We get a whole family of estimators (for different distributions of)
- We can do inference (i.e, confidence intervals, hypothesis, tests, etc.)

Cons:

- Assumptions on how data are generated (linear relationship, Gaussian errors)