# IoT·인공지능·빅데이터 개론 및 실습

## Regression (1)

서울대학교 전기정보공학부
오성회

# Contents

# 1 Univariate Linear Regression

- Univariate linear function: $y = w_1 x + w_0$, where $w_0, w_1 \in \mathbb{R}$.
- Let $\mathbf{w} = [w_0, w_1]^T$ and define

$$h_{\mathbf{w}}(x) = w_1 x + w_0.$$

- **Linear regression**: Given a training set $\{(x_i, y_i) : 1 \leq i \leq N\}$, find $h_{\mathbf{w}}$ that best fits the training set.

$$Loss(h_{\mathbf{w}}) = \sum_{j=1}^{N} L_2(y_j, h_{\mathbf{w}}(x_j)) = \sum_{j=1}^{N} (y_j - h_{\mathbf{w}}(x_j))^2 = \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2.$$

- Our goal is to find $\mathbf{w}^* = \arg\min_{\mathbf{w}} Loss(h_{\mathbf{w}})$.

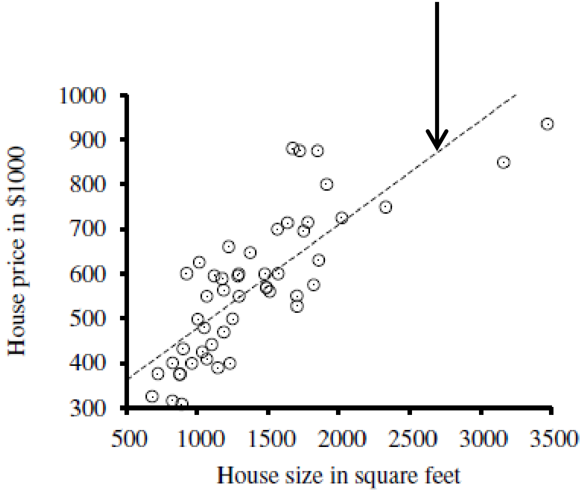
$$\frac{\partial}{\partial w_0} \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2 = 0 \text{ and } \frac{\partial}{\partial w_1} \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2 = 0$$

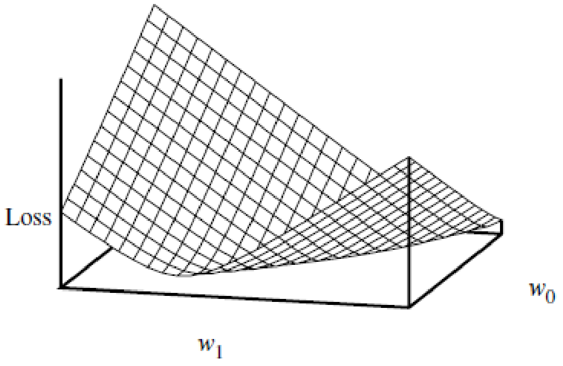
$$w_1 = \frac{N(\sum x_j y_j) - (\sum x_j)(\sum y_j)}{N(\sum x_j^2) - (\sum x_j)^2}; \quad w_0 = (\sum y_j - w_1(\sum x_j))/N\ .$$

# Univariate Linear Regression

## Example

$y = 0.232x + 246.$




$$\mathbf{w}^* = [w_0^*, w_1^*]^T = [246, 0.232]$$

$$Loss(h_{\mathbf{w}}) = \sum_{j=1}^{N} (y_j - (w_1 x_j + w_0))^2.$$

(Source: Artificial Intelligence: A Modern Approach, Stuart Russell and Peter Norvig, Prentice Hall, 2010.)

# ① Univariate Linear Regression

## Gradient Descent

$\mathbf{w} \leftarrow$ any point in the parameter space
**loop** until convergence **do**
  **for each** $w_i$ **in** $\mathbf{w}$ **do**
$$w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} Loss(\mathbf{w})$$

$\alpha$: step size or learning rate.

$$\begin{aligned}
\frac{\partial}{\partial w_i} Loss(\mathbf{w}) &= \frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(x))^2 \\
&= 2(y - h_{\mathbf{w}}(x)) \times \frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(x)) \\
&= 2(y - h_{\mathbf{w}}(x)) \times \frac{\partial}{\partial w_i}(y - (w_1 x + w_0))
\end{aligned}$$

$$\frac{\partial}{\partial w_0} Loss(\mathbf{w}) = -2(y - h_{\mathbf{w}}(x))$$

$$\frac{\partial}{\partial w_1} Loss(\mathbf{w}) = -2(y - h_{\mathbf{w}}(x)) \times x$$

$$\boxed{w_0 \leftarrow w_0 + \alpha \, (y - h_{\mathbf{w}}(x)) \, ; \quad w_1 \leftarrow w_1 + \alpha \, (y - h_{\mathbf{w}}(x)) \times x}$$

# Gradient Descent

$\mathbf{w} \leftarrow$ any point in the parameter space
**loop** until convergence **do**
  **for each** $w_i$ **in** $\mathbf{w}$ **do**
    $w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} Loss(\mathbf{w})$

$\alpha$: step size or learning rate.

**Batch gradient descent** (steepest descent):

$$w_0 \leftarrow w_0 + \alpha \sum_j (y_j - h_{\mathbf{w}}(x_j))\,; \quad w_1 \leftarrow w_1 + \alpha \sum_j (y_j - h_{\mathbf{w}}(x_j)) \times x_j$$

**Stochastic gradient descent**: processes one data point at a time

# 2 Multivariate Linear Regression

- When $\mathbf{x}_j \in \mathbb{R}^n$, the hypothesis space for linear regression is spanned by functions of the following form.

$$h_{\mathbf{w}}(\mathbf{x}_j) = w_0 + \sum_{i=1}^{n} w_i x_{j,i}.$$

- Now, redefine $\mathbf{x}_j = [1, x_{j,1}, x_{j,2}, \ldots, x_{j,n}]^T \in \mathbb{R}^{n+1}$. Then

$$h_{\mathbf{w}}(\mathbf{x}_j) = \sum_{i=0}^{n} w_i x_{j,i} = \mathbf{w}^T \mathbf{x}_j.$$

- Solution to linear regression:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j L_2(y_j, \mathbf{w}^T \mathbf{x}_j).$$

- Gradient descent update equation:
  $w_i = w_i + \alpha \sum_j x_{j,i}(y_j - h_{\mathbf{w}}(\mathbf{x}_j))$.
- Closed form solution: $\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.

## Closed-Form Solution

- Let $\mathbf{y} = [y_1, \ldots, y_N]^T$ and $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}$ be the data matrix (or design matrix).

  Then

$$Loss(h_{\mathbf{w}}) = \sum_{j=1}^{N} (y_j - \mathbf{w}^T \mathbf{x}_j)^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}).$$

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}} Loss(h_{\mathbf{w}}) &= \frac{\partial}{\partial \mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \\ &= \frac{\partial}{\partial \mathbf{w}} (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) \\ &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} \end{aligned}$$

- Setting it to zero, we have $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$, which is known as a **normal equation**. If $\mathbf{X}^T\mathbf{X}$ is invertible, we have

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

# 3 General Linear Regression

- Linear model: $h_{\mathbf{w}}(\mathbf{x}_j) = \sum_{i=0}^{n} w_i x_{j,i} = \mathbf{w}^T \mathbf{x}_j$.
- General linear model:

$$h_{\mathbf{w}}(\mathbf{x}_j) = \sum_{i=0}^{n} w_i \phi_i(\mathbf{x}_j) = \mathbf{w}^T \phi(\mathbf{x}_j),$$

  where $\phi(\mathbf{x}_j) = [1 \ \phi_1(\mathbf{x}_j) \cdots \phi_n(\mathbf{x}_j)]^T$ and $\phi_i : \mathbb{R}^n \to \mathbb{R}$ are nonlinear functions.

- Examples

$$\begin{aligned} \phi_i(x) &= x^i \qquad \text{(polynomial regression)} \\ \phi_i(\mathbf{x}) &= \exp\left(-\frac{1}{2s^2}(\mathbf{x} - \mu_i)^2\right), \end{aligned}$$

  where $s$ and $\mu_i$ are fixed parameters.

# 3 General Linear Regression

- General linear model: $h_{\mathbf{w}}(\mathbf{x}_j) = \mathbf{w}^T \phi(\mathbf{x}_j)$
- Let

$$\mathbf{\Phi} = \begin{bmatrix} 1 & \phi_1(\mathbf{x}_1) & \cdots & \phi_n(\mathbf{x}_1) \\ 1 & \phi_1(\mathbf{x}_2) & \cdots & \phi_n(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(\mathbf{x}_N) & \cdots & \phi_n(\mathbf{x}_N) \end{bmatrix}.$$

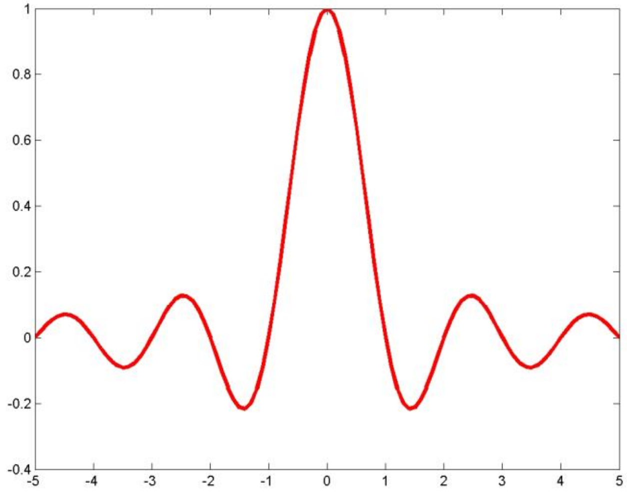
- Then

$$Loss(h_{\mathbf{w}}) = \sum_{j=1}^{N} (y_j - \mathbf{w}^T \phi(\mathbf{x}_j))^2 = (\mathbf{y} - \mathbf{\Phi}\mathbf{w})^T (\mathbf{y} - \mathbf{\Phi}\mathbf{w}).$$

- Hence

$$\mathbf{x}^* = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y}.$$

# Example: Linear modeling of the SINC function




Model 1: $y = w_0 + w_1 x$

Model 2: $y = \sum_{i=1}^{n} w_i \phi_i(x), \quad \phi_i(x) = \exp\left(-\frac{1}{2s^2}(x-\mu_i)^2\right), \quad \mu = -5 : 0.5 : 5, s = 1$

# Example: Linear modeling of the SINC function