Linear regression

Seung-Hoon Na[1]

[1]Department of Computer Science
Chonbuk National University

2018.11.12

Linear regression

- $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$: training data for regression problem where $y \in \mathbb{R}$
- Linear regression function:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$

  where $\mathbf{w}$ and b are parameters.
- J: the error for the least squares estimator on $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ (i.e., the sum of squares of the errors)

$$\begin{aligned} J &= \sum_i \left(f(\mathbf{x}_i) - y_i\right)^2 \\ &= \sum_i \left(\mathbf{w}^T \mathbf{x}_i + b - y_i\right)^2 \\ &= \sum_i \left(\mathbf{w}^T \mathbf{x}_i + b - y_i\right)^T \left(\mathbf{w}^T \mathbf{x}_i + b - y_i\right) \\ &= \sum_i \left(\mathbf{w}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{w} - 2(y_i - b)\mathbf{w}^T \mathbf{x}_i + (y_i - b)^2\right) \end{aligned}$$

Jacobian

- $\mathbf{y} = \mathbf{f}(\mathbf{x})$: a vector of m scalar-valued functions that each take a vector $\mathbf{x}$

$$\begin{aligned} y_1 &= f_1(\mathbf{x}) \\ &\vdots \\ y_m &= f_m(\mathbf{x}) \end{aligned}$$

- Jacobian matrix: has m rows for m equations (row-oriented)

$$\begin{aligned} \frac{\partial \mathbf{y}}{\partial \mathbf{x}} &= \begin{bmatrix} \nabla f_1(\mathbf{x}) \\ \cdots \\ \nabla f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{x}} f_1(\mathbf{x}) \\ \cdots \\ \frac{\partial}{\partial \mathbf{x}} f_m(\mathbf{x}) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \vdots & \cdots & \vdots \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix} \end{aligned}$$

- Jacobian matrix: has m columns for m equations (column-oriented)

Matrix calculus

- The case that Jacobian is column-oriented:

$\mathbf{y}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$
$\mathbf{A}\mathbf{x}$	$\mathbf{A}^T$
$\mathbf{x}^T\mathbf{A}$	$\mathbf{A}$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\mathbf{A}\mathbf{x}+\mathbf{A}^T\mathbf{x}$

Linear regression: Deriv J wrt $\mathbf{w}$

$$\frac{\partial \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{w}}{\partial \mathbf{w}} = 2\mathbf{x}_i \mathbf{x}_i^T \mathbf{w}$$

$$\frac{\partial \mathbf{w}^T \mathbf{x}_i}{\partial \mathbf{w}} = \mathbf{x}_i$$

All together lead to:

$$\frac{\partial J}{\partial \mathbf{w}} = \sum_i \left(2\mathbf{x}_i \mathbf{x}_i^T \mathbf{w} - 2(y_i - b)\mathbf{x}_i \right) = \mathbf{0}$$

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

Linear regression: Deriv J wrt $\mathbf{w}$

- Let $\mathbf{X}$ and $\mathbf{y}$ be:

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} \mathbf{x}_1 & \cdots \mathbf{x}_N \end{bmatrix} \\ \mathbf{y}' &= \begin{bmatrix} y_1 - b & \cdots y_N - b \end{bmatrix}^T \end{aligned}$$

- Then, $\frac{\partial J}{\partial \mathbf{w}} = \mathbf{0}$ is rewritten by:

$$\mathbf{X}\mathbf{X}^T\mathbf{w} = \mathbf{X}\mathbf{y}'$$

- Thus, $\mathbf{w}$ to minimize the sum of the squared errors is:

$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{y}'$$

Linear regression: Deriv J wrt b

Pseudo inverse

- For $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{A}^+$ is a pseudo-inverse of $\mathbf{A}$, satisfying all of the following four criteria, known as the Moore-Penrose conditions:

$$\begin{aligned} \mathbf{A}\mathbf{A}^+\mathbf{A} &= \mathbf{A} \\ \mathbf{A}^+\mathbf{A}\mathbf{A}^+ &= \mathbf{A}^+ \\ \left(\mathbf{A}\mathbf{A}^+\right)^* &= \mathbf{A}\mathbf{A}^+ \\ \left(\mathbf{A}^+\mathbf{A}\right)^* &= \mathbf{A}^+\mathbf{A} \end{aligned}$$

- When $\mathbf{A}$ has linearly independent columns (i.e., $\mathbf{A}^*\mathbf{A}$ is invertible), $\mathbf{A}^+$ is:

$$\mathbf{A}^+ = \left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^*$$

- When $\mathbf{A}$ has linearly independent rows (i.e., $\mathbf{A}\mathbf{A}^*$ is invertible), $\mathbf{A}^+$ is:

$$\mathbf{A}^+ = \mathbf{A}^*\left(\mathbf{A}\mathbf{A}^*\right)^{-1}$$

Pseudo inverse and linear least squares

- The pseudoinverse provides a least squares solution to a system of linear equations:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

  where $\mathbf{A} \in \mathbb{R}^{m \times n}$

- The least squares solution is:

$$\mathbf{z} = \mathbf{A}^{+}\mathbf{b}$$

  in the sense that for all $\mathbf{x}$, the following holds:

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \geq \|\mathbf{A}\mathbf{z} - \mathbf{b}\|^2$$

Pseudo inverse for linear regression

- Our linear regression problem is formulated as finding a least squares solution to the following linear equation system:

$$\mathbf{X}^T \mathbf{w} = \mathbf{y}'$$

- The pseudo-inverse of $\mathbf{X}^T$ is:

$$\left(\mathbf{X}^T\right)^+ = \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}$$

- Thus, the least squares solution $\mathbf{z}$ is

$$\begin{aligned} \mathbf{z} &= \left(\mathbf{X}^T\right)^+ \mathbf{y}' \\ &= \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{y}' \end{aligned}$$

Linear regression: Bias-inclusive setting

- $\mathcal{D}' = \{(\mathbf{x}_i, y_i)\}$: training data for regression problem where $y \in \mathbb{R}$ where $x_{i1} = 1$ for the bias part.
- Linear regression function:

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

  where $\mathbf{w}$ is a parameter vector and w_1 corresponds to a bias parameter (b term is absorbed into $\mathbf{w}$.
- J: the error for the least squares estimator on $\mathcal{D}'$:

$$\begin{aligned} J &= \sum_i \left(f(\mathbf{x}_i) - y_i\right)^2 \\ &= \sum_i \left(\mathbf{w}^T \mathbf{x}_i - y_i\right)^2 \\ &= \sum_i \left(\mathbf{w}^T \mathbf{x}_i - y_i\right)^T \left(\mathbf{w}^T \mathbf{x}_i - y_i\right) \\ &= \sum_i \left(\mathbf{w}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{w} - 2 y_i \mathbf{w}^T \mathbf{x}_i + y_i^2\right) \end{aligned}$$