44251017 Huang Jiahui

Exercise 4: Derivation of linear regression

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$$

Find a linear regression model $\boldsymbol{t} = \boldsymbol{w}^T\boldsymbol{x}$ using the training samples

1. Express the sum of squared errors E as a function of $\boldsymbol{w}$
2. Derive the following (a)(b) to find the gradient $\nabla_w E$ for $\boldsymbol{w}$ of the sum of squared errors E

   (a) $\sum_{i=1}^N t_i \boldsymbol{x}_i = \boldsymbol{X}^T\boldsymbol{t}$

   (b) $\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{w}^T \boldsymbol{x}_i = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$
3. Show the gradient $\frac{\partial E}{\partial \boldsymbol{w}}$ in terms of $\boldsymbol{x}_i$ (or $\boldsymbol{X}$), $\boldsymbol{t}$
4. (Approximate solution) Show the parameter update equation for the linear regression model using the gradient descent method in terms of $\boldsymbol{x}_i$ (or $\boldsymbol{X}$), $\boldsymbol{t}$
   Initial solution $\boldsymbol{w}^0$, t–th update $\boldsymbol{w}^t$, step size parameter η
5. (Analytic solution) Show that $\boldsymbol{w}$ where $\frac{\partial E}{\partial \boldsymbol{w}}$ = 0 is

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

1. $E(w) = \frac{1}{2}\sum_{i=1}^{N}(t_i - w^T x_i)^2 = \frac{1}{2}\|t - Xw\|^2 = \frac{1}{2}(t - Xw)^T(t - Xw)$

2. (a) the right side of the equation:

$$X^T t = (x_1^T, x_2^T, \cdots, x_N^T)\begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix} = t_1 x_1^T + t_2 x_2^T + \cdots + t_N x_N^T = \sum_{i=1}^{N} t_i x_i^T = \text{left side}$$

(b) $\sum_{i=1}^{N} x_i (w^T x_i) = \sum_{i=1}^{N} x_i x_i^T w = \left(\sum_{i=1}^{N} x_i x_i^T\right) w$

Since $X^T X = \sum_{i=1}^{N} x_i x_i^T$

thus: $\sum_{i=1}^{N} x_i (w^T x_i) = X^T X w$

3. $E(w) = \frac{1}{2}(t^T t - 2w^T X^T t + w^T X^T X w)$

$$\frac{\partial E}{\partial w} = \nabla_w E = -X^T t + X^T X w = X^T(Xw - t)$$

which is (b) − (a).

4. $w_{t+1} = w_t - \eta \nabla_w E(w_t)$

$$= w_t - \eta X^T(Xw_t - t)$$

5. the squared error loss function is $E(w) = \frac{1}{2}\|t - Xw\|^2$

$\Rightarrow \nabla_w E(w) = X^T(Xw - t)$

To minimize the function, we set $\nabla_w E(w) = 0$

$\Rightarrow X^T(Xw - t) = 0$

$\Rightarrow X^T X w = X^T t$

$\Rightarrow w = (X^T X)^{-1} X^T t$

44251017 Huang Jiahui