AUC Course: Machine Learning

Written Assignment 2

Simon Fang
simon.fang@student.auc.nl
10898492

Exercise 1:

(a) We assume the following:

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_n \end{pmatrix}; \qquad \boldsymbol{x}^{(i)} = \begin{pmatrix} x_o \\ \vdots \\ x_n \end{pmatrix}$$

where $\theta_0 = 1$. We have to rewrite the hypothesis function using $\boldsymbol{\theta}$ and $\boldsymbol{x}^{(i)}$ as defined above. The hypothesis function is given as follows:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n \tag{1}$$

By the definition of the dot product, we know that for $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{n+1}$ we have the following relation:

$$\boldsymbol{a}^T \cdot \boldsymbol{b} = \begin{pmatrix} a_0 & \cdots & a_n \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ \vdots \\ b_n \end{pmatrix} = a_0 b_0 + a_1 b_1 + \ldots + a_n b_n \tag{2}$$

Using the result from equation (2), we can rewrite equation (1) using the $\boldsymbol{\theta}$ and $\boldsymbol{x}^{(i)}$ as defined above:

$$h_\theta(\boldsymbol{x}^{(i)}) = \boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n \tag{3}$$

where we assume that the input is a column vector. If $\boldsymbol{x} \in \mathbb{R}^{(n+1)\times m}$, then our output will be a row vector of length m.

(b) From the lecture notes, we are given the scalarized version of the cost function for linear regression:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{4}$$

Note that $\boldsymbol{y} \in \mathbb{R}^{m1}$ is a column vector and $y^{(i)}$ denotes the i^{th} element of $\boldsymbol{y}$. Using our formula from part (a), we can vectorize the cost function in the following way:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{2m}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m}\sum_{i=1}^{n}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)})^2 \end{aligned}$$

(c) In order to answer this question, we have to introduce a new notation for extracting elements from $\boldsymbol{x}$: we denote $x_j^{(i)}$ as the element on the j^{th} row and in the i^{th} column. For the scalarized version of the gradient of the cost function we had the following equations:

$$\begin{aligned} \frac{\partial J}{\partial \theta_0} &= \frac{1}{m}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)}) \\ \frac{\partial J}{\partial \theta_1} &= \frac{1}{m}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})x_1 \\ &\vdots \\ \frac{\partial J}{\partial \theta_n} &= \frac{1}{m}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})x_n \end{aligned}$$

For the vectorized version, we will have to extract elements from the matrix $\boldsymbol{x}$. By doing this, we can vectorize the above into the following way:

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \begin{pmatrix} \frac{\partial J\boldsymbol{\theta}}{\partial \theta_0} \\ \frac{\partial J\boldsymbol{\theta}}{\partial \theta_1} \\ \vdots \\ \frac{\partial J\boldsymbol{\theta}}{\partial \theta_n} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{n}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)}) \\ \frac{1}{m}\sum_{i=1}^{n}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)})x_1^{(i)} \\ \vdots \\ \frac{1}{m}\sum_{i=1}^{n}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)})x_n^{(i)} \end{pmatrix} \end{aligned}$$

(d) Using the results from part (c), we can also vectorize the update rule in the gradient descent procedure:

Repeat until convergence{

$$\begin{aligned}\boldsymbol{\theta} &= \boldsymbol{\theta} - \alpha\frac{1}{m}\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \boldsymbol{\theta} - \alpha\frac{1}{m}\begin{pmatrix} \frac{1}{m}\sum_{i=1}^{n}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)}) \\ \frac{1}{m}\sum_{i=1}^{n}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)})x_1^{(i)} \\ \vdots \\ \frac{1}{m}\sum_{i=1}^{n}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)})x_n^{(i)} \end{pmatrix}\end{aligned}$$

}
where α is the learning rate and m is the size of vector $\boldsymbol{y}$

(e) If we use matrix multiplication instead of the explicit summation, then the gradient of the cost function will be re-written to:

$$\frac{\partial J\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = \frac{1}{m}(\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{\theta} - \boldsymbol{\theta}\boldsymbol{y}) \tag{5}$$

And from this we can re-write the result from part (d) into the following:

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha\frac{1}{m}(\boldsymbol{x}^T(\boldsymbol{x} \cdot \boldsymbol{\theta} - y)) \tag{6}$$

Exercise 3:

(a) The mean μ is defined as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

Using this equation, we can compute the mean of the given set as follows:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{6}(2+5+7+7+9+25) = \frac{55}{6} = 9\frac{1}{6} \approx 9.2 \tag{8}$$

In order to calculate the variance σ^2, we make use of the following identity:

$$\sigma^2 = E(X^2) - (E(X))^2 \tag{9}$$

The variance of the given set is

$$\sigma^2 = \frac{1}{6}(4 + 25 + 49 + 49 + 81 + 625) - (55/6)^2 = \frac{1973}{36} \approx 54.8 \quad (10)$$

(b) The probability density function (PDF) of a normal distribution with mean μ and variance σ is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2} \quad (11)$$

Hence, for $X = 20$, $\mu = 9.2$ and $\sigma^2 = 54.8$ the PDF is

$$\begin{aligned} f_X(20) &= \frac{1}{\sqrt{2 \cdot 54.8 \cdot \pi}} \exp \frac{-(20-9.2)^2}{2 \cdot 54.8} \\ &\approx 0.019 \end{aligned}$$

(c) If $X_1, ...X_n \sim N(\mu, \sigma^2)$ and they are IID, then we the joint the probability density function is

$$f_{X_1,...,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n) \quad (12)$$

Using equation (12), we can calculate the joint PDF

$$\begin{aligned} f_{X_1,...,X_n}(x_1, \ldots, x_6) &= f_{X_1}(x_1) f_{X_2}(x_2) f_{X_3}(x_3) f_{X_4}(x_4) f_{X_5}(x_5) f_{X_6}(x_6) \\ &= \left(\frac{1}{\sqrt{2 \cdot \pi \cdot 54.8}}\right)^6 \exp\left(-\frac{1}{2\sigma^2}((2 - \frac{55}{6})^2\right. \\ &\quad \left. + (5 - \frac{55}{6})^2 + (7 - \frac{55}{6})^2 + (7 - \frac{55}{6})^2 + (9 - \frac{55}{6})^2 + (25 - \frac{55}{6})^2\right) \\ &\approx 1.2 \cdot 10^{-9} \end{aligned}$$

(d) We note that

$$\begin{aligned} f_X(8) &> f_X(25) \\ 0.053 &> 0.0055 \end{aligned}$$

If we use the above information and we assume that all the variables are IID, then we can conclude that

$$f_X(8) > f_X(25)$$
$$f_{X_1}(8)f_{X_2}(2)f_{X_3}(5)f_{X_4}(7)f_{X_5}(7)f_{X_6}(9) > f_{X_1}(25)f_{X_2}(2)f_{X_3}(5)f_{X_4}(7)f_{X_5}(7)f_{X_6}(9)$$
$$f_{X_1,\dots,X_6}(8,2,5,7,7,9) > f_{X_1,\dots,X_6}(25,2,5,7,7,9)$$

(e) The formula for the covariance is defined as

$$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) \tag{13}$$

Using equation (7), we calculate the mean of Y

$$\mu_Y = \frac{1}{6}(4+4+5+6+8+10) = \frac{37}{6} \approx 6.2 \tag{14}$$

and the variance of Y is

$$\sigma_Y^2 = \frac{1}{6}(16+16+25+36+64+100) - (\frac{37}{6})^2 = \frac{246}{36} \approx 6.8 \tag{15}$$

If we combine all the information, then we can calculate the covariance of X and Y using equation (13)

$$\begin{aligned}\text{cov}(X,Y) = \frac{1}{6}[&\left(2-\frac{55}{6}\right)\left(4-\frac{37}{6}\right) + \left(5-\frac{55}{6}\right)\left(4-\frac{37}{6}\right) + \\ &\left(7-\frac{55}{6}\right)\left(5-\frac{37}{6}\right) + \left(7-\frac{55}{6}\right)\left(6-\frac{37}{6}\right) + \\ &\left(9-\frac{55}{6}\right)\left(8-\frac{37}{6}\right) + \left(25-\frac{55}{6}\right)\left(10-\frac{37}{6}\right)] \approx 14.6\end{aligned}$$

(f) By definition, the mean squared error (MSE) measures the spread around a line, while the covariance measures how much two random variables change together. The main difference is that the MSE looks at one dataset, while the covariance looks at two or more datasets and measures the spread between them. When we look at the two formulae, the cost function is essentially the covariance for X and X, which is the variance, with the mean in the variance replaced with the hypothesis function. So in both cases, they are essentially looking at the spread of the random variables.