AUC Course: Machine Learning

Solutions to Assignment 2

Vincent Roest, 10904816
vincentr@live.nl

I Worked together with Bas Straathof in my quest to correct solutions

Problem 1:
a: We know that the hypothesis function looks like:

$$h(x) = \theta_0 + \theta_1 x_1 + ...\theta_n x_n$$

We can easily rewrite this function as a vectorial expression in this way, with $x_0 = 1$:

$$h_\theta(x) = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} \theta_0 & \theta_1 & \cdots & \theta_n \end{bmatrix}$$

This is, in the notation of the question, then:

$$h_\theta(x) = \theta^T x$$

$$h_\theta(x) = \begin{bmatrix} \theta^T x^{(1)} \\ \theta^T x^{(2)} \\ \vdots \\ \theta^T x^{(m)} \end{bmatrix} = \begin{bmatrix} 1 + \theta_1 x_1^{(1)} + \ldots + \theta_n x_n^{(1)} \\ 1 + \theta_1 x_1^{(2)} + \ldots + \theta_n x_n^{(2)} \\ \vdots \\ 1 + \theta_1 x_1^{(m)} + \ldots + \theta_n x_n^{(m)} \end{bmatrix}$$

b: We know that the cost function is defined as follows:

$$J(\theta_0, \theta_1, \cdots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Now using the vector-notation we found in **a**, we can write this as:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)} \right)^2$$

c: We know that the derivative of the cost function with respect to for instance θ_0 is given by:

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta_0} &= \frac{\partial}{\partial \theta_0}\left(\frac{1}{2m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2\right) \\ &= \frac{1}{2m}\sum_{i=1}^{m}\frac{\partial}{\partial \theta_0}\left(\left(\theta^T x^{(i)} - y^{(i)}\right)^2\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)x_0^{(i)}\end{aligned}$$

Now, the gradient of the cost function is given by:

$$\frac{\partial J(\theta)}{\partial \theta} = \left(\frac{\partial J(\theta)}{\partial \theta_0}\cdots\frac{\partial J(\theta}{\partial \theta_n}\right)^T$$

Combining the two expressions we found, the gradient of the cost function is given by:

$$\frac{\partial J(\theta)}{\partial \theta} = \begin{bmatrix}\frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)x_0^{(i)} \\ \vdots \\ \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)x_n^{(i)}\end{bmatrix}$$

d: We know that the update rule for θ_j is defined as:

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

Which was derived from the fact that:

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial \theta_j}J(\theta)$$

So combining the results of subquestions **a**, **b** and **c** and the expression above, the update rule for a θ_j can be written in the following vectorial way:

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)x_j^{(i)}$$

e: For the previous questions we were ignoring the jth feature of the ith data set, since I assumed that they are present in the x^ith data vector. In this question, we are asked not to ignore this. Then every vector of x would replaced by a matrix in the form of the following matrix, where n is the total number of the features and m the number of of data points. In other words, the

features are on the rows and the examples on the columns. Let's say we have training set of houses, then the 1 feature of house 4 would be on the 4th row in the 1st column:

$$X = \begin{bmatrix} x_0^{(1)} & \cdots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}$$

However, if we use the matrix notation for X, as described in the above, we have to write for the definition of the hypothesis function:

$$h_\theta(x) = X\theta$$

We also have to think as to how we write the vector **y**:

$$y = \left(y^{(1)} y^{(2)} \dots y^{(m)}\right)^T$$

. Then, we also get an updated form of the cost function. Matrix squaring is different from scalar squaring in the sense that the square of a matrix is given by the matrix times its transpose. Let us write this for the cost function and simplify:

$$\begin{array}{rcl} J(\theta) & = & \frac{1}{2m}\left(X \cdot \theta - y\right)^T \left(X \cdot \theta - y\right) \\ & = & \frac{1}{2m}\left(\theta^T X^T - y^T\right)\left(X\theta - y\right) \\ & = & \frac{1}{2m}\left(\theta^T X^T X\theta - 2\theta^T X^T y + yy^T\right) \end{array}$$

Now that we have a different expression for the cost function, vectorial gradient expression with respect to θ from part **c** changes as well:

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{m}\left(X^T X\theta - X^T y\right)$$

And finally, the update rule for θ_j in vector notation:

$\theta_j := \theta_j - \alpha \frac{1}{m}\left(X^T X\theta - X^T y\right)$ where θ is a zero vector except for the j-th element in the vector

Problem 3:
a: The mean μ is defined as:

$$\mu = \frac{\sum_{i=1}^{n}(X_i)}{n}$$

where X are the values for the random variables and n is the sample size. So:

$$\mu = \frac{55}{6}$$

The variance is given by:

$$\sigma^2 = \frac{\sum_{i=1}^{n}|X_i - \mu|^2}{n}$$

So, it is the squared distance from each point with respect to the mean divided by the sample size:

$$\sigma^2 = \frac{(2-\frac{55}{6})^2 + (5-\frac{55}{6})^2 + (7-\frac{55}{6})^2 + (7-\frac{55}{6})^2 + (9-\frac{55}{6})^2 + (25-\frac{55}{6})^2}{6} = \frac{1973}{36}$$

From the dusty probability book I had lying around somewhere I found that a random variable is Normal distributed if:

$$X \ N(\mu, \sigma^2)$$

So, the normal distribution is simply:

$$X \ N(\frac{55}{6}, \frac{1973}{36})$$

We could also numerically write the normal distribution now as:

$$f(x = \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

We can fill this in and find the distribution:

$$f(x) = \frac{1}{\sqrt{\frac{1973}{36}}\sqrt{2\pi}} e^{\frac{-(x-\frac{55}{6})^2}{2\frac{1973}{36}}}$$

b: The last expression comes in very useful for this question, since we have to plug in x = 20 to find our answer:

$$f(20) = \frac{1}{\sqrt{\frac{1973}{36}}\sqrt{2\pi}} e^{\frac{-(20-\frac{55}{6})^2}{2\frac{1973}{36}}} \approx 0.018$$

c: The fact that all variables are identically distributed means that their joint probability can be written as:

$$f_{X_1 \ldots X_6}(x_1, \ldots x_6) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_6}(x_6)$$

Now we can easily calculate the values for the densities, just as we did in part **b**. For instance:

$$f_{X_1}(2) = \frac{1}{\sqrt{\frac{1973}{36}}\sqrt{2\pi}} e^{\frac{-(2-\frac{55}{6})^2}{2\frac{1973}{36}}} \approx 0.034$$

We do the same for $i = 2..n$ and find the following values: $f_{X_2}(5) = 0.046, f_{X_3}(7) \approx 0.052, f_{X_4}(7) \approx 0.052, f_{X_5}(9) \approx 0.054$ and $f_{X_6}(25) \approx 0.0055$. We find the following probability:

$$f_{X_1 \ldots X_6}(x_1, \ldots x_6) = 0.034 \cdot 0.046 \cdot 0.052 \cdot 0.052 \cdot 0.054 \cdot 0.0055 \approx 1.26 \cdot 10^{-9}$$

d: From the last equation, we can already see that X = 25 has a smaller (order 10 smaller) probability than the rest of the points in the data set. Because X = 8, which replaces the value 25 in the second set, is closer to the sample mean, it will have a higher probability than X = 25, so the total joint probability will be larger. In short: $f_X(8) > f_X(25)$, so the joint probability of the second data set will be higher than that of the first data set.

e: The covariance between X and Y is given by:

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

We know the mean for the X was $\mu_X = \frac{55}{6}$. And using the formula below we can also determine the mean for Y:

$$\mu = \frac{\sum_{i=1}^{n}(Y_i)}{n} = \frac{65}{12}$$

Now we have everything to fill in the formula for the covariance:

$$cov(X, Y) \approx \frac{(-\frac{43}{6})(-\frac{17}{12}) + (-\frac{25}{6})(-\frac{17}{12}) + (-\frac{13}{6})(-\frac{5}{12}) + (-\frac{13}{6})(\frac{7}{12}) + (-\frac{1}{6})(\frac{31}{12}) + (\frac{95}{6})(\frac{55}{12})}{6} \approx 14.64$$

f: The difference between the MSE and the covariance is that the MSE measures the mean of the squared errors of 1 data set i.e. the vertical spread around the linear regression line, whereas the covariance measures the dependency of two different on each other. A similarity might be that they are both dependent on the data size and on the mean of their data. They measure a different thing, but are related through their components so to say.