Machine Learning - Written Assignment 1

khaled tamimy

October 8, 2016

1 Question 1:

1.1 a)

We have,

$$\bar{\theta} = \begin{bmatrix}\theta_0 & \theta_1 & ... & \theta_n\end{bmatrix}^T$$

And,

$$\mathbf{x}^{(i)} = \begin{bmatrix}x_0^{(i)} & x_1^{(i)} & ... & x_n^{(i)}\end{bmatrix}^T$$

The hypothesis function then becomes:

$$h_\theta(x^i) = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + ... + \theta_n x_n^{(i)} \tag{1}$$

Hence, in matrix-vector form this is:

$$h_\theta(x^i) = \begin{bmatrix}\theta_0 & \theta_1 & ... & \theta_n\end{bmatrix}\begin{bmatrix}x_0^{(i)} \\ x_1^{(i)} \\ . \\ . \\ x_n^{(i)}\end{bmatrix} = \bar{\theta}^T \mathbf{x}^{(i)}$$

1.2 b)

We have:

$$J(\bar{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \tag{2}$$

Hence, in matrix-vector form this is:

$$J(\bar{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(\bar{\theta}^T \mathbf{x}^{(i)} - y^{(i)})^2 \tag{3}$$

1.3 c)

We have:

$$\frac{\partial J(\bar{\theta})}{\partial \theta} = \begin{bmatrix} \frac{\partial J(\bar{\theta})}{\partial \theta_0} \\ \frac{\partial J(\bar{\theta})}{\partial \theta_1} \\ . \\ . \\ \frac{\partial J(\bar{\theta})}{\partial \theta_n} \end{bmatrix}$$

And, we also have the following formula for the j^{th} row of the gradient vector:

$$\frac{\partial J(\bar{\theta})}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \tag{4}$$

Thus, in matrix-vector form this is:

$$\frac{\partial J(\bar{\theta})}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (\bar{\theta}^T \mathbf{x}^{(i)} - y^{(i)}) x_j^{(i)} \tag{5}$$

1.4 d)

For the update rule, we have:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) x_j^{(i)} \tag{6}$$

for the j^{th} row of $\bar{\theta}$.

Hence, in matrix-vector form this is:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (\bar{\theta}^T \mathbf{x}^{(i)} - y^{(i)}) x_j^{(i)} \tag{7}$$

for the j^{th} row of $\bar{\theta}$, where:

$$\bar{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ . \\ . \\ \theta_n \end{bmatrix}$$

1.5 e)

We have:

$$X = \begin{bmatrix} 1 & x_1^{(1)} & ... & x_n^{(1)} \\ 1 & x_1^{(2)} & ... & x_n^{(2)} \\ . & . & . & . \\ . & . & . & . \\ 1 & x_1^{(m)} & ... & x_n^{(m)} \end{bmatrix}$$

Then:

$$h_\theta(X) = X\bar{\theta} \tag{8}$$

And, since $(X\bar{\theta} - \mathbf{y})^2 = (X\bar{\theta} - \mathbf{y})^T(X\bar{\theta} - \mathbf{y})$, we have the following for the cost function:

$$J(\bar{\theta}) = \frac{1}{2m}(X\bar{\theta} - \mathbf{y})^T(X\bar{\theta} - \mathbf{y}) \tag{9}$$

And the gradient becomes:

$$\nabla J(\theta) = \frac{1}{m}X^T(X\bar{\theta} - \mathbf{y}) \tag{10}$$

With the update rule:

$$\bar{\theta} = \bar{\theta} - \frac{\alpha}{m}X^T(X\bar{\theta} - \mathbf{y}) \tag{11}$$

2 Question 3:

2.1 a)

We have the values: 2, 5, 7, 7, 9, 25. And, we also have the maximum likelihood estimates for mean, μ, and the variance, Var(X):

$$\mu = \frac{1}{n}\sum_{i=1}^{m} x_i \tag{12}$$

And,

$$\sigma^2 = Var(X) = \frac{1}{n}\sum_{i=1}^{m}(x_i - \mu)^2 \tag{13}$$

Hence, using the given sample set, we have:

$$\mu = 9.167 \tag{14}$$

And:

$$Var(X) = 54.81 \tag{15}$$

2.2 b)

Now, since X is normally distributed with the given μ and Var(X), we know that the pdf is given by:

$$f_{X_1} = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{16}$$

Now, if we substitute the μ and Var(X) calculated above, together with the given x=20, we have:

$$f_{X_1} = 0.0185 \tag{17}$$

2.3 c)

Now, we have $X_1, .., X_n$, all independent of each other and all identically and normally distributed with mean μ and variance Var(X) as calculated above.

Then, according to the normal model, we have:

$$f_{X_1X_2X_3X_4X_5X_6}(x_1, x_2, x_3, x_4, x_5, x_6) = \prod_{i=1}^{6} f_{X_i}(x_i) \tag{18}$$

Then, given the pdf in question b and given the respective values for $x_1..x_6$, we can easily calculate $f_{X_1X_2X_3X_4X_5X_6}(2, 5, 7, 7, 9, 25)$:

$$f_{X_1X_2X_3X_4X_5X_6}(2, 5, 7, 7, 9, 25) = 1.323 * 10^{-9} \tag{19}$$

2.4 d)

We do just as we did in question c, but replace the 25 with an 8. We have:

$$f_{X_1X_2X_3X_4X_5X_6}(2, 5, 7, 7, 8, 9) = 1.381 * 10^{-8} \tag{20}$$

Hence, it is larger than the probability density calculated above, which is logical, since 8 is closer to the mean than 25 is.

2.5 e)

We have:

$$Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y) \tag{21}$$

And, we estimate $E(XY) = \bar{XY}$ and $E(X) = \bar{X}$ and $E(Y) = \bar{Y}$. Where the bar represents the average of the corresponding variable.

Now, if we calculate these averages, we can easily plug this into equation 21 and we obtain:

$$Cov(X, Y) = 71.17 - 9.167 * 6.167 = 14.639 \tag{22}$$

2.6 f)

If we look at the definition for the MSE, we can easily see the following:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X-\mu)^2 = \frac{1}{m}\sum_{i=1}^{m}(X-\mu)(X-\mu) = E((X-\mu)(X-\mu)) \quad (23)$$

Where μ represents E(X), and thus:

$$MSE = E((X-E(X))(X-E(X))) = Cov(X,X) \quad (24)$$

So, yes they are related according to the above. The covariance X with itself equals the MSE of X. However, they are not the same; there is a difference. The covariance is a measure of how two random variables change together, while the MSE is a measure of offset of a certain variable. So in the sense of Cov(X,X) = MSE(X); the two values are the same numerically, but represent something conceptually different; the one represents the offset of a variable from what is estimated, while the Cov(X,X) represents the growth of the variable X with respect to itself.