Machine Learning
Written Assignment 2

F. Raaijmakers 10886869

7 October 2016

Question 1

A $\boldsymbol{\theta} = (\theta_0 \; \theta_1 \ldots \theta_n)^T$
$\boldsymbol{x}^{(i)} = (x_0 \; x_1 \ldots x_n)^T where \; x_0 = 1$

$$\begin{aligned}
h_\theta(x^{(i)}) &= \sum_{i=0}^{n} \theta_i x_i \\
&= (\theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n) \\
&= \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} * \begin{pmatrix} x_0 & x_1 & \ldots x_n \end{pmatrix} \\
&= \boldsymbol{\theta}^T \boldsymbol{x}^{\boldsymbol{(i)}}
\end{aligned}$$

B I replaced the predicted x value in the cost function by the vectorized hypothesis expression.

$$\begin{aligned}
J(\boldsymbol{\theta}) &= \frac{1}{2m} \sum_{i=0}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \\
&= \frac{1}{2m} \sum_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{\boldsymbol{(i)}} - y^{(i)})^2
\end{aligned}$$

C The gradient of the cost function is a vector with the gradient of the cost function with respect to all n θ values. Hence we differentiate the cost function with respect to every θ value.

$$\frac{\partial}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)})^2 \tag{1}$$

$$= \frac{1}{2m} \frac{\partial}{\partial \theta_j} \sum_{i=0}^{m} ((\theta_0 x_0 + \ldots + \theta_n x_n) - y^{(i)})^2 \tag{2}$$

Applying the chain rule, function (2) is differentiated with respect to an arbritary θ_j value.

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} &= \frac{1}{2m}\sum_{i=0}^{m} 2 * ((\theta_0 x_0 + \ldots + \theta_n x_n) - y^{(i)}) * x_j \\
&= \frac{1}{m}\sum_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * x_j \\
\frac{\partial J \boldsymbol{\theta}}{\partial \theta} &= \begin{pmatrix} \frac{\partial J\boldsymbol{\theta}}{\partial \theta_0} \\ \vdots \\ \frac{\partial J\boldsymbol{\theta}}{\partial \theta_n} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{m}\sum\limits_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * x_0 \\ \vdots \\ \frac{1}{m}\sum\limits_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * x_n \end{pmatrix} \\
&= \frac{1}{m}\sum_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * \boldsymbol{x}^{(i)}
\end{aligned}
$$

D

$$
\begin{aligned}
\boldsymbol{\theta} &:= \boldsymbol{\theta} - \alpha \frac{\partial J(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} \\
\theta_j &:= \theta_j - \frac{\alpha}{m}\sum_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * x_j \\
\boldsymbol{\theta} &:= \begin{pmatrix} \theta_0 - \frac{\alpha}{m}\sum\limits_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * x_0 \\ \theta_1 - \frac{\alpha}{m}\sum\limits_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * x_1 \\ \vdots \\ \theta_n - \frac{\alpha}{m}\sum\limits_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * x_n \end{pmatrix} \\
&:= \boldsymbol{\theta} - \frac{\alpha}{m}\sum_{i=0}^{m} (\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} - y^{(i)}) * \boldsymbol{x}^{(i)}
\end{aligned}
$$

E $\boldsymbol{X} = \begin{pmatrix} (\boldsymbol{x}^{(0)})^T \\ (\boldsymbol{x}^{(1)})^T \\ \vdots \\ (\boldsymbol{x}^{(m)})^T \end{pmatrix}$

$$\boldsymbol{y} = \begin{pmatrix} y^{(0)} \\ y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix}$$

The cost function is defined as a vectorized expression:

$$J(\boldsymbol{\theta}) = \frac{1}{2m}(\boldsymbol{X\theta} - \boldsymbol{y})^T(\boldsymbol{X\theta} - \boldsymbol{y})$$

Since the gradient of the cost function is its derivative, the gradient function is defined accordingly:

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta})}{\partial \theta} &= \frac{\partial}{\partial \theta}\frac{1}{2m}(\boldsymbol{X\theta} - \boldsymbol{y})^T(\boldsymbol{X\theta} - \boldsymbol{y}) \\ &= \frac{1}{m}(\boldsymbol{X\theta} - \boldsymbol{y})^T\boldsymbol{X} \end{aligned}$$

Question2

A e = a + c + b + d

x	**y**	**freq**	**P(X=x,Y=y)**
0	0	a	$^a/_e$
0	1	c	$^c/_e$
1	0	b	$^b/_e$
1	1	d	$^d/_e$

B

$$\begin{aligned} P(X=0) &= P(X=0, Y=0) + P(X=0, Y=1) \\ &= \frac{a}{e} + \frac{c}{e} \\ &= \frac{a+c}{e} \end{aligned}$$

C

$$\begin{aligned}P(X=1|Y=0) &= \frac{P(X=1\cap Y=0)}{P(Y=0)}\\&= \frac{\frac{b}{e}}{\frac{a+b}{e}}\\&= \frac{b*e}{(a+b)*e}\\&= \frac{b}{a+b}\end{aligned}$$

D

$$\begin{aligned}P(X=1\cup Y=0) &= P(X=1,Y=0)+P(X=1,Y=1)+P(X=0,Y=0)\\&= \frac{a+b+d}{e}\end{aligned}$$

E

$\bar{x} = \frac{0+0+1+1}{4} = 0.5$
$\bar{y} = \frac{0+1+0+1}{4} = 0.5$
m=4

$$\begin{aligned}cov(X,Y) &= \frac{1}{m}\sum_{i=0}(x_i-\bar{x})(y_i-\bar{y})\\&= \frac{1}{4}\sum_{i=0}(x_i-0.5)(y_i-0.5)\\&= \frac{1}{4}[(0-0.5)(0-0.5)+(0-0.5)(1-0.5)+(1-0.5)(0-0.5)+(1-0.5)(1-0.5)]\\&= 0\end{aligned}$$

Question 3

A

$\mu = \frac{2+5+7+7+9+25}{6} = 9\frac{1}{6}$

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=0}^{n}(x_i-\mu)^2\\&= \frac{1}{6}[(2-9\frac{1}{6})^2+(5-9\frac{1}{6})^2+(7-9\frac{1}{6})^2+(7-9\frac{1}{6})^2+(9-9\frac{1}{6})^2+(25-9\frac{1}{6})^2]\\&= \frac{1}{6}*328.83\bar{3}\\&\simeq 54.81\end{aligned}$$

B

For a normal distribution, the PDF is the following:

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

$$f_X(20) = \frac{1}{\sqrt{2*54.81\pi}}^{-\frac{(20-9\frac{1}{6})^2}{2*54.81}}$$
$$\simeq 0.019$$

C

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_6) = \prod_{i=1}^{6} \frac{1}{\sqrt{2\sigma^2\pi}}^{-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}}$$

$$f_{X_1,\ldots,X_n}(2,5,7,7,9,25) = \frac{1}{\sqrt{2*54.81\pi}}^{-\frac{(2-9\frac{1}{6})^2}{2*54.81}} * \frac{1}{\sqrt{2*54.81\pi}}^{-\frac{(5-9\frac{1}{6})^2}{2*54.81}} * 2\frac{1}{\sqrt{2*54.81\pi}}^{-\frac{(7-9\frac{1}{6})^2}{2*54.81}}$$
$$* \frac{1}{\sqrt{2*54.81\pi}}^{-\frac{(9-9\frac{1}{6})^2}{2*54.81}} * \frac{1}{\sqrt{2*54.81\pi}}^{-\frac{(25-9\frac{1}{6})^2}{2*54.81}}$$
$$= 0.054^{-0.49} * 0.054^{-0.16} * 2 * 0.054^{-0.043} * 0.054^{-0.00025} * 0.054^{-2.29}$$
$$\simeq 11270.86$$

D The probability density function will be larger, because the variability between the data is smaller (smaller variance).

E $m = 6$
$\bar{x} = 9\frac{1}{6}$
$\bar{y} = \frac{4+4+5+6+8+10}{6} = 6\frac{1}{6}$

$$cov(X,Y) = \frac{1}{m-1}\sum_{i=0}^{m}(x_i-\bar{x})(y_i-\bar{y})$$
$$= \frac{1}{5}[(2-9\frac{1}{6})(4-6\frac{1}{6} + (5-9\frac{1}{6})(4-6\frac{1}{6} + (7-9\frac{1}{6})(5-6\frac{1}{6} + (7-9\frac{1}{6})(6-6\frac{1}{6}+$$
$$(9-9\frac{1}{6})(8-6\frac{1}{6} + (25-9\frac{1}{6})(10-6\frac{1}{6}]$$
$$\simeq 17.57$$

F Placing these definitions next to each other, we can see that they are highly similar, and conclude that the MSE is equal to the covariance when x=y.
$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})$
$MSE(Y) = \frac{1}{n}\sum_{i=1}^{m}(y_i-\bar{y})^2$

Question 4

A $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}} * exp(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu))$

B smaller