Written Assignment 2

Leah Dickhoff
due October 7th, 2016
Machine Learning

1.

a. The vectorial expression for the hypothesis function $h_\theta(x)$ is given by the dot product between vectors θ and $x^{(i)}$:

$$\begin{aligned} h_\theta(x) &= \theta_0 + \theta_1 x + \theta_2 x_2 + \ldots + \theta_n x_n \\ &= (\theta_0\, \theta_1 \cdots \theta_n)^T \cdot (x_0\, x_1 \cdots x_n)^T \\ &= \theta^T \cdot x^{(i)} \end{aligned}$$

b.

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

c.

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta} &= \begin{pmatrix} \frac{\partial J(\theta)}{\partial \theta 0} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta n} \end{pmatrix} \\ &= \frac{1}{m} \sum_{i=1}^{m} [\frac{\partial}{\partial \theta}(h_\theta(x^{(i)}) - y^{(i)})] \\ &= \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x^{(i)} \\ &= \frac{1}{m} \sum_{i=1}^{m} (\theta^T \cdot x^{(i)} - y^{(i)})\, x^{(i)} \end{aligned}$$

d.

$$\begin{aligned} \theta_j &:= \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)} \\ &:= \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (\theta^T \cdot x^{(i)} - y^{(i)})\, x_j^{(i)} \\ &:= \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta j} \end{aligned}$$

so for θ_n:

$$\begin{aligned} \theta_n &:= \theta_n - \alpha \frac{1}{m} \sum_{i=1}^{m} (\theta^T \cdot x^{(i)} - y^{(i)})\, x_n^{(i)} \\ &:= \begin{pmatrix} \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} [\theta^T \cdot x^{(i)} - y^{(i)}]\, x_0^{(i)} \\ \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} [\theta^T \cdot x^{(i)} - y^{(i)}]\, x_1^{(i)} \\ \vdots \\ \theta_n - \alpha \frac{1}{m} \sum_{i=1}^{m} [\theta^T \cdot x^{(i)} - y^{(i)}]\, x_n^{(i)} \end{pmatrix} \end{aligned}$$

2.

a.

x	y	freq	P(X=x, Y=y)
0	0	a	$\frac{a}{a+b+c+d}$
0	1	c	$\frac{c}{a+b+c+d}$
1	0	b	$\frac{b}{a+b+c+d}$

1	1	d	$\frac{d}{a+b+c+d}$

b. P(X=0) = $\frac{a+b}{a+b+c+d}$

c. P(X=1 I Y=0) = $\frac{\mathrm{P(X=1 \cap Y=0)}}{\mathrm{P(Y=0)}}$

$$= \frac{\frac{b}{a+b+c+d}}{\frac{a+b}{a+b+c+d}}$$

$$= \frac{b}{a+b}$$

d. P(X=1 ∪ Y=0) = $\frac{a+b+d}{a+b+c+d}$

e. cov(X,Y) = $\frac{1}{m}\sum_{i=1}^{m}(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})$

where $\bar{\mathrm{x}} = \frac{0+0+1+1}{4}$ = ½

and $\bar{\mathrm{y}} = \frac{0+1+0+1}{4}$ = ½

and (the number of values) m = 4

so cov(X,Y) = $\frac{1}{4}\sum_{i=1}^{4}(\mathrm{x_i}$ - ½)($\mathrm{y_i}$ – ½)

$$= \frac{1}{4}\,[(0 - ½)(0 - ½) + (0 - ½)(1 - ½) + (1 - ½)(0 - ½) + (1 - ½)(1 - ½)]$$

$$= 0$$

3. random values from a normal distribution: 2, 5, 7, 7, 9, 25

a. Mean estimation: $\mu = \frac{2+5+7+7+9+25}{number\ of\ values} = \frac{2+5+7+7+9+25}{6} = \frac{55}{6}$

Variance estimation:

$$\sigma^2 = [(\tfrac{55}{6} - 2)^2 + (\tfrac{55}{6} - 5)^2 + (\tfrac{55}{6} - 7)^2 + (\tfrac{55}{6} - 7)^2 + (\tfrac{55}{6} - 9)^2 + (\tfrac{55}{6} - 25)^2]/6$$

$$\approx 54{,}806$$

b. $f_X(x) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$

where $\sigma = \sqrt{\sigma^2} = \sqrt{54{,}81}$ (≈ 7,40)

so:

$$f_X(20) = \frac{1}{\sqrt{2\cdot 54{,}81\pi}}\, e^{-\left(20-\frac{55}{6}\right)^2/(2\cdot 54{,}81)}$$

$$\approx 0{,}0185$$

c. $f_{X1\,\dots\,Xn}(x_1, \dots\, x_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\left(x^{(i)}-\mu\right)^2/2\sigma^2}$

so:

$f_{X1\,\dots\,X6}(2, 5, 7, 7, 9, 25)$

$$= \left(\frac{1}{\sqrt{2\cdot 54{,}81\pi}}\right)^6 \cdot e^{-\left(2-\frac{55}{6}\right)^2/(2\cdot 54{,}81)} \cdot e^{-\left(5-\frac{55}{6}\right)^2/(2\cdot 54{,}81)}$$

$$\cdot\, e^{-2\cdot\left(7-\frac{55}{6}\right)^2/(2\cdot 54{,}81)} \cdot e^{-\left(9-\frac{55}{6}\right)^2/(2\cdot 54{,}81)} \cdot e^{-\left(25-\frac{55}{6}\right)^2/(2\cdot 54{,}81)}$$

$$\approx 1{,}2193 \cdot 10^{-9}$$

d. The probability density $f_{X1 \dots X6}(2, 5, 7, 7, 8, 9)$ would be considerably larger than the probability density $f_{X1 \dots X6}(2, 5, 7, 7, 9, 25)$. The reason for this is the fact that in the second one the maximum value of 25 (which is a lot larger than the mean μ) is included, meaning that the variance is larger, and thus the probability density smaller. In the first probability density, however, all values are closer to the mean μ, so the variance is smaller and the probability density larger.

e. Covariance estimation:

$$\operatorname{cov}(X,Y) = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y}) \qquad \| \text{ with } \bar{x} = \frac{2+5+7+7+9+25}{6} = \frac{55}{6}$$

$$\| \text{ and } \bar{y} = \frac{4+4+5+6+8+10}{6} = \frac{37}{6}$$

$$= \frac{1}{5}\Big[(2 - \frac{55}{6})(4 - \frac{37}{6}) + (5 - \frac{55}{6})(4 - \frac{37}{6}) + (7 - \frac{55}{6})(5 - \frac{37}{6}) + (7 - \frac{55}{6})(6 - \frac{37}{6})$$
$$+ (9 - \frac{55}{6})(8 - \frac{37}{6}) + (25 - \frac{55}{6})(10 - \frac{37}{6})\Big]$$
$$= 17{,}58$$

f. $\operatorname{cov}(X,Y) = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})$

$\text{MSE} = \frac{1}{m-1} \sum_{i=1}^{m} (y_i - \bar{y})^2$

Hence, by comparing the two formulas, we can conclude that cov(X,Y) = MSE $\forall(x=y)$:

$$\operatorname{cov}(X,Y) = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{1}{m-1} \sum_{i=1}^{m} (y_i - \bar{y})(y_i - \bar{y})$$
$$= \frac{1}{m-1} \sum_{i=1}^{m} (y_i - \bar{y})^2$$
$$= \text{MSE}$$

This result seems accurate, as the MSE measures the distance of the values to the hypothesis and the covariance measures how much the variables show a similar behavior. If the covariance is positive and large, then if x is small (or large), so is y. The MSE is positive anyway (mathematically because of the square, geometrically because it is a distance). Thus, intuitionally, if y=x (i.e. they are identical: so both small (or large)) then cov(X,Y) is positive and large just like the MSE, so cov(X,Y) = MSE.

4.

a.

b. If the number of features (*n*) increases, then the threshold value increases, too. We should decrease the threshold.

c. We should decrease the threshold, because with more features, we (statistically) have a more accurate and precise dataset, allowing us to decrease the threshold value to make it more accurate, too. The goal is to find the minimum value that this threshold can be, and increasing the number of features is a way to come closer to this goal.