Machine Learning Assignment 2 - Julian Tsang

1. (GRADED) This question is about *vectorization*, i.e. writing expressions in matrix-vector form. The goal is to vectorize the update rule for multivariate linear regression.

   (a) Let $\boldsymbol{\theta}$ be the parameter vector $\boldsymbol{\theta} = (\theta_0\ \theta_1 \cdots \theta_n)^T$ and let the i-th data vector be: $\boldsymbol{x}^{(i)} = (x_0\ x_1\ \cdots x_n)^T$ where $x_0 = 1$. What is the vectorial expression for the hypothesis function $h_{\boldsymbol{\theta}}(\boldsymbol{x})$?

   (b) What is the vectorized expression for the cost function: $J(\boldsymbol{\theta})$ (still using the explicit summation over all training examples).

   (c) What is the vectorized expression for the gradient of the cost function, i.e. what is:

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_0} \\ \vdots \\ \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_n} \end{pmatrix} \tag{1}$$

   Again the explicit summation over the data vectors from the learning set is allowed here.

   (d) What is the vectorized expression for the $\boldsymbol{\theta}$ update rule in the gradient descent procedure.

   (e) (bonus points) Vectorization can be taken one step further. We can remove the explicit summation over the training samples by 'hiding´ it in a matrix vector multiplication. Start by collecting all training samples in a data matrix $\boldsymbol{X}$ such that every *row* of $\boldsymbol{X}$ is a vector from the training set (with the augmented $x_0 = 1$ elements, i.e. the first column of $\boldsymbol{X}$ has elements equal to 1).

a. $\mathrm{h}_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = \theta^T x$

b. $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\mathrm{h}_\theta(x^{(i)}) - y^{(i)})^2 \rightarrow \frac{1}{2m}\sum_{i=1}^{m}[\theta^T x - y^{(i)}]^2$

c. $\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{m}\sum_{i=1}^{m}[\theta^T x - y^{(i)}]$

d. $\theta_j := \theta_j - \frac{\alpha}{m}[\theta^T x - y^{(i)}]x^{(i)}$ $where\ j = 0,\ 1,\ \ldots n$

3. (GRADED) We assume the value 2, 5, 7, 7, 9, 25 are random values from a normal distribution.

   (a) Estimate the mean μ and variance σ^2 of this normal distribution.
   (b) Let $X \sim N(\mu, \sigma^2)$ be a random variable. Calculate the probability density $f_X(20)$.
   (c) Now consider six randowm variables $X_1, \ldots, X_n$. All *independent of eachother* and all identically and normally distributed with mean μ and variable σ^2 as calculated above. Let $f_{X_1 \cdots X_6}(x_1, \ldots, x_6)$ be the joint probability density function. Calculate $f_{X_1 \cdots X_6}(2, 5, 7, 7, 9, 25)$.
   (d) Is $f_{X_1 \cdots X_6}(2, 5, 7, 7, 8, 9)$ larger or smaller then the probability density calculated above?
   (e) Now consider two random variables X and Y and six random samples of this multivariate distribution:

x	y
2	4
5	4
7	5
7	6
9	8
25	10

   Estimate the covariance $\mathrm{cov}(X, Y)$.

   (f) Compare the definition of the covariance with the mean squared error that is used in the cost function in linear regression. Are they related? Is there a difference? If so, what? Explain your answer.

a. $Mean\ \mu = \frac{2+5+7+7+9+25}{6} = \frac{55}{6} \approx 9.1667$

$Variance\ \sigma^2 = \frac{1}{6}\left[\left(2-\frac{55}{6}\right)^2 + \left(5-\frac{55}{6}\right)^2 + \left(7-\frac{55}{6}\right)^2 + \left(7-\frac{55}{6}\right)^2 + \left(9-\frac{55}{6}\right)^2 + \left(25-\frac{55}{6}\right)^2\right] \approx 54.8$

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

$$P(20) = \frac{1}{\sqrt{54.8}\sqrt{2\pi}} e^{-\left(20-\frac{55}{6}\right)^2/(2(54.8))} \approx 0.01847$$

c. $P(2) = \frac{1}{\sqrt{54.8}\sqrt{2\pi}} e^{-\left(2-\frac{55}{6}\right)^2/(2(54.8))} \approx 0.0337287$

$$P(5) = \frac{1}{\sqrt{54.8}\sqrt{2\pi}} e^{-\left(5-\frac{55}{6}\right)^2/(2(54.8))} \approx 0.0459966$$
$$P(7) = \frac{1}{\sqrt{54.8}\sqrt{2\pi}} e^{-\left(7-\frac{55}{6}\right)^2/(2(54.8))} \approx 0.0516319$$
$$P(9) = \frac{1}{\sqrt{54.8}\sqrt{2\pi}} e^{-\left(9-\frac{55}{6}\right)^2/(2(54.8))} \approx 0.0538778$$
$$P(25) = \frac{1}{\sqrt{54.8}\sqrt{2\pi}} e^{-\left(25-\frac{55}{6}\right)^2/(2(54.8))} \approx 0.00547183$$

$$f_{x_1 \dots x_6}(2,\ 5,\ 7,\ 7,\ 9,\ 25) = P(2) * P(5) * P(7) * P(7) * P(9) * P(25) \approx 1.219 * 10^{-9}$$

d. $f_{x_1 \dots x_6}(2,\ 5,\ 7,\ 7,\ 8,\ 9) \approx 0.0532263$ would be larger than $f_{x_1 \dots x_6}(2,\ 5,\ 7,\ 7,\ 9,\ 25)$

e. $cov(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$

$$= \frac{\left(2-\frac{55}{6}\right)\left(4-\frac{37}{6}\right) + \left(5-\frac{55}{6}\right)\left(4-\frac{37}{6}\right) + \left(7-\frac{55}{6}\right)\left(5-\frac{37}{6}\right) + \left(7-\frac{55}{6}\right)\left(6-\frac{37}{6}\right) + \left(9-\frac{55}{6}\right)\left(8-\frac{37}{6}\right) + \left(25-\frac{55}{6}\right)\left(10-\frac{37}{6}\right)}{6}$$

$$= \frac{527}{36} \approx 14.63889$$

f. Covariance versus Mean Squared Error
   Covariance is a measure of correlation between X and Y values, showing how changes in one variable are related to changes in another variable. Mean squared error for linear regression shows the difference between an estimated fitted value and a given value, which is the vertical spread of the data around the regression line.