Written Assignment 2

Marit Hoefsloot

October 7th, 2016

1

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.

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

We multiply the two vectors pointwise, so the elements of $\boldsymbol{\theta}$ and $\boldsymbol{x^{(i)}}$ with the same indexes are multiplied to give the element of $h_{\boldsymbol{\theta}}(\boldsymbol{x^{(i)}})$ with that index:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x^{(i)}}) = \boldsymbol{\theta^T x^{(i)}}$$
$$h_{\boldsymbol{\theta}}(\boldsymbol{x^{(i)}}) = x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_n$$

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

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_{\boldsymbol{\theta}}(\boldsymbol{x^{(i)}}) - \boldsymbol{y^{(i)}})^2$$
$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}((x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_n) - y^{(i)})^2$$

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

$$\frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta})} = \begin{pmatrix} \frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta_0})} \\ \frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta_1})} \\ \vdots \\ \frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta_n})} \end{pmatrix} \tag{1}$$

$$\frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta})} = \frac{1}{m}\sum_{i=1}^{m}\frac{\delta}{\delta(\boldsymbol{\theta})}(h_{\boldsymbol{\theta}}(\boldsymbol{x^{(i)}}) - y^{(i)})$$
$$\frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta})} = \frac{1}{m}\sum_{i=1}^{m}(h_{\boldsymbol{\theta}}(\boldsymbol{x^{(i)}}) - y^{(i)})x^{(i)}$$
$$\frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta})} = \frac{1}{m}\sum_{i=1}^{m}((x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_n) - y^{(i)})x^{(i)}$$

4. What is the vectorized expression for the θ update rule in the gradient descent procedure.

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha\frac{\delta J(\boldsymbol{\theta})}{\delta(\boldsymbol{\theta})}$$

With one element of the vector with index n:

$$\theta_n := \theta_n - \alpha\frac{\delta J(\boldsymbol{\theta})}{\delta(\theta_n)}$$
$$\theta_n := \theta_n - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_{\boldsymbol{\theta}}(\boldsymbol{x^{(i)}}) - y^{(i)})x_n$$
$$\boldsymbol{\theta} := \begin{pmatrix} \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(((x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_2) - y^{(i)})x_0) \\ \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(((x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_2) - y^{(i)})x_1) \\ \vdots \\ \theta_n - \alpha\frac{1}{m}\sum_{i=1}^{m}(((x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_n) - y^{(i)})x_n) \end{pmatrix}$$

2

Consider two discrete random variables X and Y both with possible outcomes 0 or 1 (such discrete random variables are often called binary or boolean random variables).
We have done the experiment of drawing samples of X and Y simultaneously and found the following frequency table (2):

x	y	freq	P(X=x, Y=y)
0	0	a	
0	1	c	
1	0	b	
1	1	d	

Table 1: The table with P(X = x, Y = y) open

Here the frequency is the absolute number of times we found a particular combination of $X = x$ and $Y = y$ values.

1. Complete the table by estimating $P(X = x, Y = y)$ for all possible outcomes.

See table 2

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

Table 2: The table with P(X = x, Y = y) filled in

2. Calculate $P(X = 0)$

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

3. Calculate $P(X = 1|Y = 0)$

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

4. Calculate $P(X = 1 \cup Y = 0)$

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

5. Calculate the covariance $cov(X, Y)$

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

3

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

1. Estimate the mean μ and variance σ^2 of this normal distribution.

   The mean is $\mu = (2+5+7+7+9+25)/6 = 9\frac{1}{6}$ and the variance is

$$
\begin{aligned}
\sigma^2 = ((9\frac{1}{6} - 2)^2 &+ (9\frac{1}{6} - 5)^2 + (9\frac{1}{6} - 7)^2 + (9\frac{1}{6} - 7)^2 \\
&+ (9\frac{1}{6} - 9)^2 + (9\frac{1}{6} - 25)^2)/6 \simeq 54.81.
\end{aligned}
$$

2. Let $X \approx N(\mu, \sigma^2)$ be a random variable. Calculate the probability density $f_X(20)$.

$$
\begin{aligned}
\sigma = \sqrt{\sigma^2} &= \sqrt{54.81} \simeq 7.403 \\
f_X(x) &= \frac{1}{\sqrt{2\sigma^2\pi}}^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
f_X(20) &= \frac{1}{\sqrt{2 \cdot 54.81\pi}} e^{-\frac{(20-9\frac{1}{6})^2}{2 \cdot 54.81}} \\
f_X(20) &\simeq 0.0185
\end{aligned}
$$

3. Now consider six random variables $X_1, \cdots, X_n$. All independent of each other and all identically and normally distributed with mean μ and variable σ^2 as calculated above. Let $f_{X_1,\cdots,X_6}(x_1, \cdots, x_6)$ be the joint probability density function. Calculate $f_{X_1,\cdots,X_6}(2, 5, 7, 7, 9, 25)$.

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

$$f_{(X_1,\cdots,X_n)}(2, 5, 7, 7, 9, 25) = (\frac{1}{\sqrt{2 \cdot 54.81\pi}})^6 \cdot e^{-\frac{(2-9\frac{1}{6})^2}{2\cdot 54.81}} \cdot e^{-\frac{(5-9\frac{1}{6})^2}{2\cdot 54.81}}$$
$$\cdot 2e^{-\frac{(7-9\frac{1}{6})^2}{2\cdot 54.81}} \cdot e^{-\frac{(9-9\frac{1}{6})^2}{2\cdot 54.81}} \cdot e^{-\frac{(25-9\frac{1}{6})^2}{2\cdot 54.81}}$$

$$f_{(X_1,\cdots,X_n)}(2, 5, 7, 7, 9, 25) \simeq 4.919 \cdot 10^{-7}$$

4. Is $f_{X_1,\cdots,X_6}(2, 5, 7, 7, 8, 9)$ larger or smaller then the probability density calculated above?

The probability that one obtains this set of random variables is larger, as this set does not contain the value 25, which is located very far away from the mean ($\mu = 9\frac{1}{6}$). Therefore the variance is smaller and this gives a larger probability.

5. 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 $cov(X, Y)$.

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

$$\bar{x} = (2+5+7+7+9+25)/6 = 9\frac{1}{6}$$

$$\bar{y} = (4+4+5+6+8+10)/6 = 6\frac{1}{6}$$

$$\begin{aligned} cov(X,Y) = \frac{1}{5}((2-9\frac{1}{6})(4-6\frac{1}{6}) + (5-9\frac{1}{6})(4-6\frac{1}{6}) \\ + (7-9\frac{1}{6})(5-6\frac{1}{6}) + (7-9\frac{1}{6})(6-6\frac{1}{6}) \\ + (9-9\frac{1}{6})(8-6\frac{1}{6}) + (25-9\frac{1}{6})(10-6\frac{1}{6})) \end{aligned}$$

$$cov(X,Y) \simeq 17.57$$

6. 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.????????????????

4

Consider the binary random variables $X_1, \cdots, X_n$. Our goal is to build an anomaly detection system that can be used for an arbitrary number of binary variables.
The probability mass function is $p_{X_1 \cdots X_n}(x_1, \cdots, x_n)$. As in the case of the normally distributed features in the anomaly detection system discussed in the lectures we assume the features $X_1, \cdots, X_n$ are independant.

1. Given a lot of normal data points $x_1^{(i)}, \cdots, x_n^{(i)}$ for $i = 1, \cdots, m$, how would you estimate $p_{X_1 \cdots X_n}(x_1, \cdots, x_n)$?

2. Suppose we find the anomalies in case $p_{X_1 \cdots X_n}(x_1, \cdots, x_n) < 0.0001$. What will happen to the threshold value in case the number of features (n) increases? Should we decrease or increase the threshold?

3. Why?