Assignment 2

Wendy Nieuwkamer

07-10-2016

1 Question 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.

Let θ be the parameter vector $\theta = \begin{pmatrix}\theta_0 & \theta_1 & \cdots & \theta_n\end{pmatrix}^T$ and let the i-th data vector be: $x^{(i)} = \begin{pmatrix}x_0 & x_1 & \cdots & x_n\end{pmatrix}^T$ where $x_0 = 1$. m is the amount of learning examples, n is the amount of features.

1.1 Write the hypothesis function $h_\theta(x)$ as a vectorial expression.

The summation notation for the hypothesis function is:

$$h_\theta(x^{(i)}) = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + ... + \theta_n x_n^{(i)}$$

This is the same as the result of the following matrix multiplication:

$$h_\theta(x^{(i)}) = \theta^T x^{(i)}, \tag{1}$$

which is a vectorial expression.

1.2 What is the vectorized expression for the cost function: $J(\theta)$?

The cost function in the notation used up to now:

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

If we simply insert the vectorial notation of the hypothesis function from last question (1) we get:

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

1.3 What is the vectorized expression for the gradient of the cost function?

i.e. what is:

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

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

The notation we used up until now is:

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

If we integrate the vectorized notation of the hypothesis (1) again we get:

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

Substituting this summation for the ∂ notation in the given vector will give us the following:

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{m}\begin{pmatrix} \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)}) \\ \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_1^{(i)} \\ \vdots \\ \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_n^{(i)} \end{pmatrix} \tag{2}$$

1.4 What is the vectorized expression for the θ update rule in the gradient descent procedure?

The original notation for the update rule for one theta was:

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

By writing it in vector notation we update the entire theta instead of each element separately. Using the formulas from (1) and (2) we get:

$$\theta := \theta - \alpha \frac{1}{m} \begin{pmatrix} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) \\ \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_1^{(i)} \\ \vdots \\ \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_n^{(i)} \end{pmatrix}$$

1.5 (BONUS) Remove the explicit summation by using a matrix vector multiplication

We start by defining the data matrix X; every row of X is a training example, the first column containing $x_0^{(1)}, x_0^{(2)}, ..., x_0^{(n)}$.

$$X = \begin{pmatrix} x_0^{(1)} & x_1^{(1)} & \cdots & x_n^{(1)} \\ x_0^{(2)} & x_1^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & & \vdots \\ x_0^{(m)} & x_1^{(m)} & \cdots & x_n^{(m)} \end{pmatrix}$$

The hypothesis function will then be $h_\theta(X) = X\theta$, which results in the matrix:

$$h_\theta(X) = X\theta = \begin{pmatrix} x_0^{(1)} & x_1^{(1)} & \cdots & x_n^{(1)} \\ x_0^{(2)} & x_1^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & & \vdots \\ x_0^{(m)} & x_1^{(m)} & \cdots & x_n^{(m)} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} = \begin{pmatrix} \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + ... + \theta_n x_n^{(1)} \\ \theta_0 x_0^{(2)} + \theta_1 x_1^{(2)} + ... + \theta_n x_n^{(2)} \\ \vdots \\ \theta_0 x_0^{(m)} + \theta_1 x_1^{(m)} + ... + \theta_n x_n^{(m)} \end{pmatrix}$$

Let Y be the matrix $Y = \begin{pmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(m)} \end{pmatrix}^T$. Then we can write the derivative of $J(\theta)$ as follows:

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= \frac{1}{m} X^T (X\theta - Y) \\
&= \frac{1}{m} \begin{pmatrix} x_0^{(1)} & x_0^{(2)} & \cdots & x_0^{(m)} \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ \vdots & \vdots & & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \end{pmatrix} \left(\begin{pmatrix} \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + ... + \theta_n x_n^{(1)} \\ \theta_0 x_0^{(2)} + \theta_1 x_1^{(2)} + ... + \theta_n x_n^{(2)} \\ \vdots \\ \theta_0 x_0^{(m)} + \theta_1 x_1^{(m)} + ... + \theta_n x_n^{(m)} \end{pmatrix} - \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} \right)
\end{aligned}
$$

The first X has to be transposed in order to make the dimensions of the matrices match. Thus, the final update rule would be:

$$\theta := \theta - \alpha \frac{1}{m} X^T (X\theta - Y)$$

2 Question 2

Consider events with two binary outcomes, X and Y. We encode the two values as 0 and 1. We can represent the outcomes of an experiment in a frequency table:

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

2.1 Complete the table by estimating $P(X = x, Y = y)$ for every combination

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

2.2 Calculate $P(X = 0)$

$X = 0$ for both the combinations $X = 0, Y = 0$, and $X = 0, Y = 1$. Thus, we are looking at a and c. Then, according to the standard formula:

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

2.3 Calculate $P(X = 1|Y = 0)$

This is the probability of $X = 1$ given $Y = 0$..

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

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

The probability of $X = 1$ and/or $Y = 0$.

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

3 Question 3

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

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

$$\begin{aligned}\mu &= \frac{1}{m}\sum_{i=1}^{m} x^i \\ &= \frac{2+5+7+7+9+25}{6} = 9.17\end{aligned}$$

$$\begin{aligned}\sigma^2 &= \frac{1}{m}\sum_{i=1}^{m}(x^i-\mu)^2 \\ &= \frac{(2-9.17)^2+(5-9.17)^2+(7-9.17)^2+(7-9.17)^2+(9-9.17)^2+(25-9.17)^2}{6} \\ &= \frac{328.83}{6} = 54.81\end{aligned}$$

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

The expression $X \sim N(\mu, \sigma^2)$ means X is distributed as $N(\mu, \sigma^2)$, which is the normal distribution of the given values. The probability function for this distribution is:

$$\begin{aligned}f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi 54.81}} e^{-\frac{(x-9.17)^2}{2*54.81}}\end{aligned}$$

Then, the probability densitty of $f_X(20)$ is:

$$\begin{aligned}f_X(20) &= \frac{1}{\sqrt{2\pi 54.81}} e^{-\frac{(20-9.17)^2}{2*54.81}} \\ &=\end{aligned}$$