Machine Learning – Written Assignment 1

Philip Hartout

September 16, 2016

1. (a) This is a supervised, classification problem. The given is the historical data of soccer matches of team playing against Ajax.
(b) A small training set would look like so:

x	**y**
Ajax - Feyenoord	Win
Ajax - AZ Alkmaar	Loose
Ajax - FC Utrecht	Draw
Ajax - FC Utrecht	Win

Table 1: Training sample

2. (a) Given the following data, manually (using only a calculator) calculate two iterations of the gradient descent algorithm for univariate linear regression function. Initialize the parameters such that the regression function passes through the origin (0, 0) and has an angle of 45 degrees. Use a learning rate of 0.1. Give the intermediate results of your calculations and also compute the mean-squared error of the function after 2 iterations.

x	**y**
3	6
5	7
6	10

Table 2: Training sample

We set our learning rate $\alpha = 0.1$. Recall that $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \cdot x^{(i)}$
Initialisation
$\theta_0 = 0$, $\theta_1 = 1$. So, $h_\theta(x) = 0 + 1 \cdot x$
Variable update #1
Now we update θ_0 and θ_1

$$
\begin{aligned}
\theta_0 &:= \theta_0 - \alpha \cdot \frac{\partial}{\partial \theta_0} J(\theta_0) \\
\Leftrightarrow \theta_0 &:= \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) - y^{(i)} \\
\Leftrightarrow \theta_0 &:= 0 - 0.1 \frac{1}{3} \left[(3-6) + (5-7) + (6-10) \right] \\
\Leftrightarrow \theta_0 &:= 0.3
\end{aligned}
$$

Now, we update θ_1 simultaneously, i.e. we still use our hypothesis function we had in the beginning: $h(x) =$

$0 + 1 \cdot x$. This yields:

$$\begin{aligned}
\theta_1 &:= \theta_1 - \alpha \cdot \frac{\partial}{\partial \theta_1} J(\theta_1) \cdot x^{(i)} \\
\Leftrightarrow \theta_1 &:= \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)} \\
\Leftrightarrow \theta_1 &:= 1 - 0.1 \frac{1}{3} \left[(3-6) \cdot 3 + (5-7) \cdot 5 + (6-10) \cdot 6 \right] \\
\Leftrightarrow \theta_1 &:= \frac{73}{30}
\end{aligned}$$

Variable update #2
Now we update θ_0 and θ_1 with $h_\theta(x) = 0.3 + \frac{73}{30} \cdot x$

$$\begin{aligned}
\theta_0 &:= \theta_0 - \alpha \cdot \frac{\partial}{\partial \theta_0} J(\theta_0) \\
\Leftrightarrow \theta_0 &:= \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) - y^{(i)} \\
\Leftrightarrow \theta_0 &:= 0.3 - 0.1 \frac{1}{3} \left[(7.6 - 6) + (\frac{187}{15} - 7) + (\frac{149}{10} - 10) \right] \\
\Leftrightarrow \theta_0 &:= -\frac{89}{900}
\end{aligned}$$

Now, we update θ_1 simultaneously, i.e. we still use our hypothesis function we had in the beginning: $h(x) = 0 + 1 \cdot x$. This yields:

$$\begin{aligned}
\theta_1 &:= \theta_1 - \alpha \cdot \frac{\partial}{\partial \theta_1} J(\theta_1) \cdot x^{(i)} \\
\Leftrightarrow \theta_1 &:= \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)} \\
\Leftrightarrow \theta_1 &:= \frac{73}{30} - 0.1 \frac{1}{3} \left[(7.6 - 6) \cdot 3 + (\frac{187}{15} - 7) \cdot 5 + (\frac{149}{10} - 10) \cdot 6 \right] \\
\Leftrightarrow \theta_1 &:= \frac{86}{225}
\end{aligned}$$

Conclusion
We end up with the following hypothesis fitting our data: $h_\theta(x) = -\frac{89}{900} + \frac{86}{225} \cdot x$.
The mean square error (MSE) for this hypothesis is calculated as follows:

$$\begin{aligned}
\text{MSE} &= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x) - y^{(i)} \right)^2 \\
\Leftrightarrow \text{MSE} &= \frac{1}{6} \left[\left(\frac{943}{900} - 6 \right)^2 + \left(\frac{1631}{900} - 7 \right)^2 + \left(\frac{79}{36} - 10 \right)^2 \right] \\
\Leftrightarrow \text{MSE} &= \frac{6067669}{324000} \approx 18.7
\end{aligned}$$

(b) Convert the data to z-scores (with mean = 0, sd = 1) and repeat the calculations above. Compare the results with those for the original data.
The mean of the data set is $\mu = \frac{14}{3}$ and its standard deviation is approximately $\sigma = 1.25$.

x	**y**
-1.33	6
0.27	7
1.07	10

Table 3: z-scores of the training sample

We set our learning rate $\alpha = 0.1$. Recall that $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(z^{(i)}) - y^{(i)})^2 \cdot z^{(i)}$
Initialisation

$\theta_0 = 0$, $\theta_1 = 1$. So, $h_\theta(x) = 0 + 1 \cdot x$

Variable update #1

Now we update θ_0 and θ_1

$$\begin{aligned}
&\theta_0 := \theta_0 - \alpha \cdot \frac{\partial}{\partial \theta_0} J(\theta_0) \\
\Leftrightarrow &\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} h_\theta(z^{(i)}) - y^{(i)} \\
\Leftrightarrow &\theta_0 := 0 - 0.1 \frac{1}{3} \left[(-1.33 - 6) + (0.27 - 7) + (1.07 - 10)\right] \\
\Leftrightarrow &\theta_0 := 0.77
\end{aligned}$$

Now, we update θ_1 simultaneously, i.e. we still use our hypothesis function we had in the beginning: $h(x) = 0 + 1 \cdot x$. This yields:

$$\begin{aligned}
&\theta_1 := \theta_1 - \alpha \cdot \frac{\partial}{\partial \theta_1} J(\theta_1) \cdot z^{(i)} \\
\Leftrightarrow &\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(z^{(i)}) - y^{(i)}) \cdot z^{(i)} \\
\Leftrightarrow &\theta_1 := 1 - 0.1 \frac{1}{3} \left[(-1.33 - 6) \cdot (-1.33) + (0.27 - 7) \cdot 0.27 + (1.07 - 10) \cdot 1.07\right] \\
\Leftrightarrow &\theta_1 := 1.05
\end{aligned}$$

Variable update #2

Now we update θ_0 and θ_1 with $h_\theta(x) = 0.77 + 1.05 \cdot x$

$$\begin{aligned}
&\theta_0 := \theta_0 - \alpha \cdot \frac{\partial}{\partial \theta_0} J(\theta_0) \\
\Leftrightarrow &\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} h_\theta(z^{(i)}) - y^{(i)} \\
\Leftrightarrow &\theta_0 := 0.77 - 0.1 \frac{1}{3} \left[(-0.63 - 6) + (1.05 - 7) + (1.89 - 10)\right] \\
\Leftrightarrow &\theta_0 := 1.46
\end{aligned}$$

Now, we update θ_1 simultaneously, i.e. we still use our hypothesis function we had in the beginning: $h(x) = 0.77 + 1.05 \cdot x$. This yields:

$$\begin{aligned}
&\theta_1 := \theta_1 - \alpha \cdot \frac{\partial}{\partial \theta_1} J(\theta_1) \cdot z^{(i)} \\
\Leftrightarrow &\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(z^{(i)}) - y^{(i)}) \cdot z^{(i)} \\
\Leftrightarrow &\theta_1 := 1.05 - 0.1 \frac{1}{3} \left[(-0.63 - 6) \cdot (-1.33) + (1.05 - 7) \cdot 0.27 + (1.89 - 10) \cdot 1.07\right] \\
\Leftrightarrow &\theta_1 := 1.1
\end{aligned}$$

Conclusion

We end up with the following hypothesis fitting our data: $h_\theta(x) = 1.46 + 1.1 \cdot x$.

The mean square error (MSE) for this hypothesis is calculated as follows:

$$\begin{aligned}
&\text{MSE} = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(z^{(i)}) - y^{(i)}\right)^2 \\
\Leftrightarrow &\text{MSE} = \frac{1}{6} \left[\left(3 \cdot 10^{-3} - 6\right)^2 + (1.16 - 7)^2 + (0.28 - 10)^2\right] \\
\Leftrightarrow &\text{MSE} \approx 27.2
\end{aligned}$$

We notice that the mean square error is smaller in the previous application of gradient descent. When doing multivariate linear regression, we would expect the MSE to be smaller after having applied feature scaling but

this is not the case here, since we only have one feature.
3.
4. Derive an equation that can be used to find the optimal value of the parameter θ_1 for univariate linear regression without doing gradient descent. This can be done by setting the value of the derivative equal to 0. You may assume that the value of θ_0 is fixed.

Let us define a vector $\vec{x}$ containing all the values of x:

$$\vec{x} = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix}$$

Similarly, we define a vector y containing all the values of y:

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Now, since $h_{\theta_1}(x^i) = \vec{x} \cdot \theta_1$, we can verify that:

$$\vec{x}\theta_1 - \vec{y} = \begin{bmatrix} x^{(1)}\theta_1 \\ \vdots \\ x^{(m)}\theta_1 \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} h(x^{(1)}) - y^{(1)} \\ \vdots \\ h(x^{(m)}) - y^{(m)} \end{bmatrix}$$

We also know that, for any vector $\vec{z}$, we have the following property: $zz^T = \sum_i z_i^2$ which, when applied to the formula above, yields:

$$\frac{1}{2}(\vec{x}\theta_1 - \vec{y})^T(\vec{x}\theta_1 - \vec{y}) = \frac{1}{2}\sum_{i=1}^{m}(h_{\theta_1}(x^{(i)}) - y^{(i)})^2 = J(\theta_1)$$

Now, using properties of the tr operator, we obtain the following:

$$\begin{aligned}
\nabla_{\theta_1} J(\theta_1) =_{\theta_1} & \frac{1}{2}(\vec{x}\theta_1 - \vec{y})^T(\vec{x}\theta_1 - \vec{y}) \\
&= \frac{1}{2}\nabla_{\theta_1}(\theta_1^T\vec{x}^T\vec{x}\theta_1 - \theta_1^T\vec{y} - \vec{y}^T\vec{x}\theta_1 + \vec{y}^T\vec{y}) \\
&= \frac{1}{2}\nabla_{\theta_1}\mathrm{tr}(\theta_1^T\vec{x}^T\vec{x}\theta_1 - \theta_1^T\vec{y} - \vec{y}^T\vec{x}\theta_1 + \vec{y}^T\vec{y}) \\
&= \frac{1}{2}\nabla_{\theta_1}(\mathrm{tr}\theta_1^T\vec{x}^T\vec{x}\theta_1 - 2\mathrm{tr}\vec{y}^T\vec{x}\theta_1) \\
&= \frac{1}{2}(\vec{x}^T\vec{x}\theta_1 + \vec{x}^T\vec{x}\theta_1 - 2\vec{x}^T\vec{y}) \\
&= \vec{x}^T\vec{x}\theta_1 - \vec{x}^T\vec{y}
\end{aligned}$$

We minimise J by setting its derivatives to zero, which yields the following:

$$\vec{x}^T\vec{x}\theta_1 = \vec{x}^T\vec{y}$$

Which allows us to isolate θ_1 and obtain it's optimal value:

$$\theta_1 = (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}$$

.