Written Assignment 1

Marit Hoefsloot

16-09-2016

1

Suppose that we have historical data of result of soccer matches of teams playing against Ajax. We want to use this information to learn to predict at a certain moment whether a team will win, lose or draw against Ajax. Our approach will be based on Machine Learning.

1. Define the given and the goal of the prediction task and of the learning task that best matches our goal. Classify the learning task as supervised, unsupervised, reinforcement learning, and if supervised as classification or regression.

Prediction task: the given inputs are the teams that will play against Ajax, and where the game will be played (at the home stadium or at the opponents stadium). The goal is to predict whether or not Ajax will win, draw or loose against the given opponent. Learning task: the given inputs are the opponents, the place where the match took place and the result of the match. The goal of the learning task is to find a structure or a general rule between the input and the results, in order to be able to predict future outcomes. The learning task is a supervised and classification problem, as the program is given some example inputs and results.

2. What would be the form of training data for the learning task? Give a small training set.

For an example training set, see table 1.
Ajax loses: 0, Ajax wins: 1, draw: 2

Opponents	At home	At the opponent's stadium
ADO Den Haag	1	0
Vitesse	1	2
FC Twente	2	0
Feyenoord	0	1

Table 1: Example training set

2

Given the following data (table 2):

x	3	5	6
x	6	7	10

Table 2: The given data

1. 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.

This information gives us: $m = 3$, $\theta_0 = 0$, $\theta_0 = 1$ and $\alpha = 0.1$. The given data is plotted in figure 1. In the graph we can see that the value of θ_0 should converge to 2 and the value of θ_1 should converge to ~ 1.21.

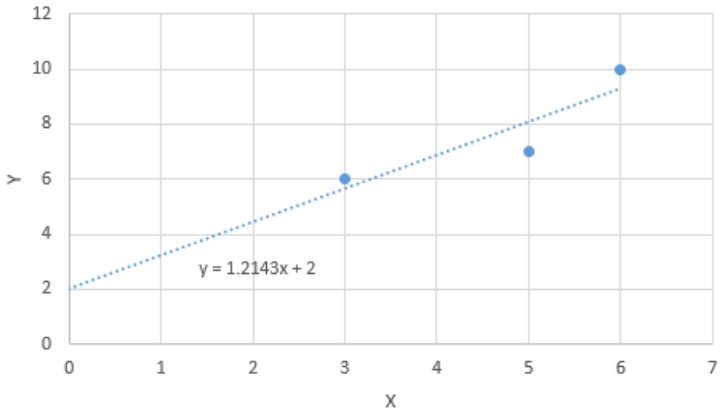

Figure 1: Graph of given data

Hypothesis:

$$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)} \tag{1}$$

Mean-squared error:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{2}$$

Gradient descent (with $j = 0, 1$):

$$\theta_j := \theta_j - \alpha \frac{\delta}{\delta \theta_j} J(\theta_0, \theta_1) \tag{3}$$

The mean-squared error value of the hypothesis with the original values, calculated with formula 2:

$$\begin{aligned}
J(\theta_0, \theta_1) &= \frac{1}{6} \sum_{i=1}^{3} (h_\theta(x^{(i)}) - y^{(i)})^2 \\
J(\theta_0, \theta_1) &= \frac{1}{6} (3(3-6)^2 + 5(5-7)^2 + 6(6-10)^2) \\
J(\theta_0, \theta_1) &= 4.83
\end{aligned}$$

With the functions given, I can manually find a formula that would give the straight line though the data in the plot. The first iteration, according to formula 3:

$$\begin{aligned}
\theta_0 &:= \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \\
&:= 0 - 0.1 \cdot \frac{1}{3} \cdot ((3-6) + (5-7) + (6-10)) \\
&:= \frac{-1}{30} \cdot -9 \\
&:= 0.3 \\
\theta_1 &:= \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)} \\
&:= 1 - 0.1 \cdot \frac{1}{3} \cdot (3(3-6) + 5(5-7) + 6(6-10)) \\
&:= 1 - \frac{1}{30} \cdot -43 \\
&:= 2.4333 \\
h_\theta(x^{(i)}) &= 0.3 + 2.4333 \cdot x^{(i)}
\end{aligned}$$

The mean-squared value, after the first iteration:

$$
\begin{aligned}
J(\theta_0, \theta_1) &= \frac{1}{6}\sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)})^2 \\
J(\theta_0, \theta_1) &= \frac{1}{6}((0.3 + 2.4333 \cdot 3 - 6)^2 \\
&\quad + (0.3 + 2.4333 \cdot 5 - 7)^2 \\
&\quad + (0.3 + 2.4333 \cdot 6 - 10)^2) \\
J(\theta_0, \theta_1) &= 9.41
\end{aligned}
$$

The second iteration:

$$
\begin{aligned}
\theta_0 &:= \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)}) \\
&:= 0.3 - \frac{1}{30} \cdot ((0.3 + 2.4333 \cdot 3 - 6) \\
&\quad + (0.3 + 2.4333 \cdot 5 - 7) + (0.3 + 2.4333 \cdot 6 - 10)) \\
&:= -0.0973 \\
\theta_1 &:= \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)} \\
&:= 2.4333 - \frac{1}{30} \cdot (3(0.3 + 2.4333 \cdot 3 - 6) \\
&\quad + 5(0.3 + 2.4333 \cdot 5 - 7) + 6(0.3 + 2.4333 \cdot 6 - 10)) \\
&:= 0.3822 \\
h_\theta(x^{(i)}) &= -0.0973 + 0.3822 \cdot x^{(i)}
\end{aligned}
$$

The mean-squared value after the second iteration:

$$
\begin{aligned}
J(\theta_0, \theta_1) &= \frac{1}{6}\sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)})^2 \\
J(\theta_0, \theta_1) &= \frac{1}{6}((-0.0973 + 0.3822 \cdot 3 - 6)^2 \\
&\quad + (-0.0973 + 0.3822 \cdot 5 - 7)^2 \\
&\quad + (-0.0973 + 0.3822 \cdot 6 - 10)^2) \\
J(\theta_0, \theta_1) &= 18.72
\end{aligned}
$$

2. Convert the data to z-scores (with $\mu = 0$, $\sigma = 1$) and repeat the calculations above. Compare the results with those for the original data.

The mean for x is $\mu = \frac{3+5+6}{3} = 4.6667$.

And the standard deviation for x is $\sigma = \sqrt{\frac{\sum_{i=1}^{m}(x^{(i)}-\mu)^2}{\#ofvalues}} = \sqrt{\frac{(3-\mu)^2+(5-\mu)^2+(6-\mu)^2}{3}} = 1.2472$.

The mean for y is $\mu = \frac{6+7+10}{3} = 7.6667$.

And the standard deviation for y is $\sigma = \sqrt{\frac{\sum_{i=1}^{m}(y^{(i)}-\mu)^2}{\#ofvalues}} = \sqrt{\frac{(6-\mu)^2+(7-\mu)^2+(10-\mu)^2}{3}} = 1.6997$.

With the formula for feature scaling (formula 4) we calculate the new values for x and y, see table 3 for the new values.

$$x' = \frac{x - \mu}{\sigma} \tag{4}$$

x	x'	y	y'
3	-1.3363	6	-0.9806
5	0.2673	7	-0.3923
6	1.0690	10	1.3728

Table 3: The new values

With these values, I again preform the gradient descent by hand. The first iteration:

$$\begin{aligned}
\theta_0 &:= \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)}) \\
&:= 0 - 0.1 \cdot \frac{1}{3} \cdot ((-1.3363 + 0.9806) + (0.2673 + 0.3923) \\
&\quad + (1.0690 - 1.3728)) \\
&:= -0.0000033.. \\
\theta_1 &:= \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)} \\
&:= 1 - 0.1 \cdot \frac{1}{3} \cdot (-1.3363(-1.3363 + 0.9806) \\
&\quad + 0.2673(0.2673 + 0.3923) + 1.0690(1.0690 - 1.3728)) \\
&:= 1 - \frac{1}{30} \cdot 0.3269 \\
&:= 0.9891 \\
h_\theta(x^{(i)}) &= -3.3333 \cdot 10^{-6} + 0.9891 \cdot x^{(i)}
\end{aligned}$$

The mean-squared value, after the first iteration:

$$
\begin{aligned}
J(\theta_0, \theta_1) &= \frac{1}{6}\sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)})^2 \\
J(\theta_0, \theta_1) &= \frac{1}{6}(-3.3333 \cdot 10^{-6} + 0.9891 \cdot -1.3363 + 0.9806)^2 \\
&\quad + (-3.3333 \cdot 10^{-6} + 0.9891 \cdot 0.2673 + 0.3923)^2 \\
&\quad + (-3.3333 \cdot 10^{-6} + 0.9891 \cdot 1.0690 - 1.3728)^2) \\
J(\theta_0, \theta_1) &= 0.1079
\end{aligned}
$$

The second iteration:

$$
\begin{aligned}
\theta_0 &:= \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)}) \\
&:= -3.3333 \cdot 10^{-6} - \frac{1}{30} \cdot ((-3.3333 \cdot 10^{-6} + 0.9891 \cdot -1.3363 + 0.9806) \\
&\quad + (-3.3333 \cdot 10^{-6} + 0.9891 \cdot 0.2673 + 0.3923) \\
&\quad + (-3.3333 \cdot 10^{-6} + 0.9891 \cdot 1.0690 - 1.3728)) \\
&:= -6.3333 \cdot 10^{-6} \\
\theta_1 &:= \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)} \\
&:= 0.9891 - \frac{1}{30} \cdot (-1.3363(-3.3333 \cdot 10^{-6} + 0.9891 \cdot -1.3363 + 0.9806) \\
&\quad + 0.2673(-3.3333 \cdot 10^{-6} + 0.9891 \cdot 0.2673 + 0.3923) \\
&\quad + 1.0690(-3.3333 \cdot 10^{-6} + 0.9891 \cdot 1.0690 - 1.3728)) \\
&:= 0.2942 \\
h_\theta(x^{(i)}) &= -6.3333 \cdot 10^{-6} + 0.2942 \cdot x^{(i)}
\end{aligned}
$$

The mean-squared value after the second iteration:

$$
\begin{aligned}
J(\theta_0, \theta_1) &= \frac{1}{6}\sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)})^2 \\
J(\theta_0, \theta_1) &= \frac{1}{6}((-6.3333 \cdot 10^{-6} + 0.2942 \cdot -1.3363 + 0.9806)^2 \\
&\quad + (-6.3333 \cdot 10^{-6} + 0.2942 \cdot 0.2673 + 0.3923)^2 \\
&\quad + (-6.3333 \cdot 10^{-6} + 0.2942 \cdot 1.0690 - 1.3728)^2) \\
J(\theta_0, \theta_1) &= 0.2811
\end{aligned}
$$

We can see that the mean-squared error value is a lot smaller in the second calculations (with the scaled features) than in the first calculations. This means that the hypothesis we get with the scaled features is more accurate than the first hypothesis. However, in both cases we can see that the error value becomes bigger after the second iteration and therefore that the values diverge which means that the gradient descent will not work. This is very likely due to the fact that α is too big; with a smaller α the values would converge and the gradient descent would give a formula for a fitting graph.

3

Suppose that X1 predicts Y, with some (mean squared) error MSE. We now extend the data with an additional variable X2 and use a learning algorithm that uses both X1 and X2 to predict Y. What will be the effect on the mean squared error of Y compared to just using X1 if X2 is equal to:

1. $a + b \cdot X1$

   I think the MSE would become linearly smaller.

2. $a + b \cdot X1^2$

   The MSE would become smaller, with the square root of what it was before.

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.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2$$

$$\frac{\delta}{\delta \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} ((\theta_0 + \theta_1 x^{(i)} - y^{(i)}) \cdot x^{(i)}) = 0$$

$$\sum_{i=1}^{m} (\theta_0 x^{(i)} + \theta_1 x^{2(i)} - x^{(i)} y^{(i)}) = 0$$

$$\sum_{i=1}^{m} (\theta_0 x^{(i)}) + \sum_{i=1}^{m} (\theta_1 x^{2(i)}) + \sum_{i=1}^{m} (x^{(i)} y^{(i)}) = 0$$

$$\sum_{i=1}^{m} (\theta_1 x^{2(i)}) = \sum_{i=1}^{m} (x^{(i)} y^{(i)}) - \sum_{i=1}^{m} (\theta_0 x^{(i)})$$

$$\sum_{i=1}^{m} \theta_1 = \sum_{i=1}^{m} \frac{x^{(i)} y^{(i)} - \theta_0 x^{(i)}}{x^{2(i)}}$$

$$m\theta_1 = \sum_{i=1}^{m} \frac{y^{(i)} - \theta_0}{x^{(i)}}$$

$$\theta_1 = \frac{1}{m} \sum_{i=1}^{m} \frac{y^{(i)} - \theta_0}{x^{(i)}}$$