Machine Learning - Written Assignment 1

khaled tamimy

September 16, 2016

1 Question 1:

1.1 a)

Our goal is: use information of former matches to learn to predict at a certain moment whether a team will win, lose or draw against Ajax.

The given for the prediction task is the good-fitting hypotheses function found in the learning task.

The goal for the prediction task is to be able to predict the outcome for a given feature x that is not in the training set.

The given for the learning task is the training set with training examples, which in this case is the historical data of results of teams playing against Ajax.

The goal for the learning task is to have a good fitting hypothesis function for the given training set. The learning task is a supervised classification learning problem.

1.2 b)

The form would be a table of a certain team playing against Ajax and all the outcomes: win, lose, or draw.

Small trainingset: Ajax - Feyenoord
Ajax won 28 times
Ajax lost 22 times
A draw 26 times

2 Question 2:

2.1 a)

Since this problem is a linear regression problem, we can directly use the following algorithm for gradient descent:

Repeat until convergence:

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) \tag{1}$$

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

Now, first, we initialize the parameters to the given conditions:
- the regression function must pass through the origin; and,
- has an angle of 45 degrees.

This then corresponds to $\theta_0 = 0$ and $\theta_1 = 1$, so the hypothesis function becomes:

$$h_\theta(x) = \theta_0 + \theta_1 x = x \tag{3}$$

Also, we use a learning rate α of $0.1 --> \alpha = 0.1$

x	3	5	6
y	6	7	10
h(x)	3	5	6

Thus, now we start substituting the initial values we have in formula (1) and (2) and iterate twice:

$$\theta_0 := 0 - 0.1 \frac{1}{3} \sum_{i=1}^{3} (h_\theta(x^i) - y^i) \tag{4}$$

$$\theta_0 := -0.1 \frac{1}{3} [(3-6) + (5-7) + (6-10)] \tag{5}$$

$$\theta_0 := -0.1 \frac{1}{3} (-3-2-4) = -\frac{1}{30}(-9) \tag{6}$$

$$\theta_0 := 0.3 \tag{7}$$

And:

$$\theta_1 := 1 - 0.1 \frac{1}{3} \sum_{i=1}^{3} (h_\theta(x^i) - y^i) x^i \tag{8}$$

$$\theta_1 := 1 - 0.1 \frac{1}{3} [3(3-6) + 5(5-7) + 6(6-10)] \tag{9}$$

$$\theta_1 := 1 - 0.1 \frac{1}{3} [-9-10-24] = 1 - \frac{1}{30}(-43) \tag{10}$$

$$\theta_1 := 2.433 = 2.4 \tag{11}$$

Now, we continue with the second iteration:
We have $\theta_0 = 0.3$ and $\theta_1 = 2.43$, thus:

x	3	5	6
y	6	7	10
h(x)	7,59	12,45	14,88

Thus:

$$\theta_0 := 0.3 - 0.1\frac{1}{3}\sum_{i=1}^{3}(h_\theta(x^i) - y^i) \tag{12}$$

$$\theta_0 := 0.3 - 0.1\frac{1}{3}(1.6 + 5.46 + 4.9) \tag{13}$$

$$\theta_0 := -0.0988 = -0.099 \tag{14}$$

And:

$$\theta_1 := 2.433 - 0.1\frac{1}{3}\sum_{i=1}^{3}(h_\theta(x^i) - y^i)x^i \tag{15}$$

$$\theta_1 := 2.433 - 0.1\frac{1}{3}(4.8 + 27.33 + 29.4) \tag{16}$$

$$\theta_1 := 0.3822 = 0.38 \tag{17}$$

Now, to compute the MSE:

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \tag{18}$$

$$J(-0.098, 0.38) = \frac{1}{6}\sum_{i=1}^{3}(h_\theta(x^i) - y^i)^2 \tag{19}$$

$$J(-0.098, 0.38) = \frac{1}{6}(112.36) = 18.73 = 19 = MSE \tag{20}$$

2.2 b)

The formula for obtaining z-scores is as follows:

$$z = \frac{x - \mu}{\sigma} \tag{21}$$

with $\mu = \frac{3+5+6}{3} = \frac{14}{3} = 4\frac{2}{3}$
and:
$(3 - 4\frac{2}{3})^2 = \frac{25}{9}$
$(5 - 4\frac{2}{3})^2 = \frac{1}{9}$
$(6 - 4\frac{2}{3})^2 = \frac{16}{9}$

Thus: $\sigma^2 = \frac{\frac{25}{9} + \frac{1}{9} + \frac{16}{9}}{3} = \frac{14}{9}$
And: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{14}{9}} = 1.247 = 1.2$

Thus, now we can compute the z-scores:

$$x^{(1)} standardized : z^{(1)} = \frac{3 - 4\frac{2}{3}}{\sqrt{\frac{14}{9}}} = -1.336 = -1.3 \tag{22}$$

$$x^{(2)} standardized : z^{(2)} = \frac{5 - 4\frac{2}{3}}{\sqrt{\frac{14}{9}}} = .2673 = .27 \tag{23}$$

$$x^{(3)} standardized : z^{(3)} = \frac{6 - 4\frac{2}{3}}{\sqrt{\frac{14}{9}}} = 1.069 = 1.0 \tag{24}$$

z	-1,3	0,27	1,0
y	6	7	10
h(x)	-1,3	0,27	1,0

Now, just as done above, we are computing two iterations of the gradient descent, but with the standardized scores:

$$\theta_0 := 0 - 0.1\frac{1}{3}\sum_{i=1}^{3}(h_\theta(z^i) - y^i) \tag{25}$$

$$\theta_0 := -0.1\frac{1}{3}[(-1.3 - 6) + (0.27 - 7) + (1.0 - 10)] \tag{26}$$

$$\theta_0 := -0.1\frac{1}{3}(-23) = -\frac{1}{30}(-23) \tag{27}$$

$$\theta_0 := 0.766 = 0.77 \tag{28}$$

And:

$$\theta_1 := 1 - 0.1\frac{1}{3}\sum_{i=1}^{3}(h_\theta(z^i) - y^i)x^i \tag{29}$$

$$\theta_1 := 1 - 0.1\frac{1}{3}[-1.3(-1.3 - 6) + 0.27(0.27 - 7) + 1.0(1.0 - 10)] \tag{30}$$

$$\theta_1 := 1 - 0.1\frac{1}{3}[-1.54] = 1 - \frac{1}{30}(-1.54) \tag{31}$$

$$\theta_1 := 1.05 = 1.1 \tag{32}$$

Now, we continue with the second iteration:
We have $\theta_0 = 0.766$ and $\theta_1 = 1.05$, thus:

z	-1,3	0,27	1,0
y	6	7	10
h(x)	0,07	1,8	2,6

Thus:

$$\theta_0 := 0.77 - 0.1\frac{1}{3}\sum_{i=1}^{3}(h_\theta(z^i) - y^i) \tag{33}$$

$$\theta_0 := 0.77 - 0.1\frac{1}{3}(-20.7) \tag{34}$$

$$\theta_0 := 1.457 = 1.5 \tag{35}$$

And:

$$\theta_1 := 1.05 - 0.1\frac{1}{3}\sum_{i=1}^{3}(h_\theta(z^i) - y^i)x^i \tag{36}$$

$$\theta_1 := 1.05 - 0.1\frac{1}{3}(-1.37) \tag{37}$$

$$\theta_1 := 1.096 = 1.1 \tag{38}$$

Now, to compute the MSE:

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \tag{39}$$

$$J(1.457, 1.096) = \frac{1}{6}\sum_{i=1}^{3}(h_\theta(z^i) - y^i)^2 \tag{40}$$

$$J(1.457, 1.096) = \frac{1}{6}(118.04) = 19.673 = 20 = MSE \tag{41}$$

If we compare the results of before and after standardization, we see that there is a difference of approximately 1 in the MSE. Generally feature scaling, or standardization, is used to make the gradient descent algorithm converge faster. This scaling is performed on all the features, such that all the features lie within the same range of values. However, here, we only have one feature, namely x, and thus the standardization of this feature isn't very succesfull.

3 Question 3:

3.1 a)

The MSE value becomes smaller, or stays equally in the worst case

3.2 b)

The MSE value becomes smaller, as now a polynomial approximation can be made instead of a linear one.

4 Question 4:

To find the optimal value of the parameter θ_1, we can look for the optimal value of θ, since θ_0 is fixed.
We start by rewriting hypothesis function in matrix-vectorial form:

$$h_\theta(x) = \begin{bmatrix}1 & x\end{bmatrix}\begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix} = \bar{X}\bar{\theta} \tag{42}$$

Then,

$$Error = \bar{X}\bar{\theta} - \bar{y} = E \tag{43}$$

And,

$$\bar{E^T}\bar{E} = (\bar{X}\bar{\theta} - \bar{y})^T(\bar{X}\bar{\theta} - \bar{y}) \tag{44}$$

$$\bar{E^T}\bar{E} = (\bar{X^T}\bar{\theta^T} - \bar{y^T})(\bar{X}\bar{\theta} - \bar{y}) \tag{45}$$

$$\bar{E^T}\bar{E} = \bar{\theta^T}\bar{X^T}\bar{X}\bar{\theta} - \bar{\theta^T}\bar{X^T}\bar{y} - \bar{y^T}\bar{X}\bar{\theta} + \bar{y^T}\bar{y} \tag{46}$$

$$\bar{E^T}\bar{E} = \bar{\theta^T}\bar{X^T}\bar{X}\bar{\theta} - 2\bar{y^T}\bar{X}\bar{\theta} + \bar{y^T}\bar{y} \tag{47}$$

-

Now, since $z^T z = \sum_{i=1}^{m} z^2$, and thus $(\bar{X}\bar{\theta} - \bar{y})^T(\bar{X}\bar{\theta} - \bar{y} = \sum_{i=1}^{m}(\bar{X}\bar{\theta} - \bar{y})^2$,

$$\bar{E^T}\bar{E} = 2J(\theta) \tag{48}$$

-

And, to find the optimal θ, we minimize $J(\theta)$, hence:

$$\nabla J(\theta) = \nabla\frac{1}{2}(\bar{E^T}\bar{E}) = \frac{1}{2}\nabla(\bar{\theta^T}\bar{X^T}\bar{X}\bar{\theta} - 2\bar{y^T}\bar{X}\bar{\theta} + \bar{y^T}\bar{y}) \tag{49}$$

$$\nabla J(\theta) = \frac{1}{2}\nabla tr(\bar{\theta^T}\bar{X^T}\bar{X}\bar{\theta} - 2\bar{y^T}\bar{X}\bar{\theta} + \bar{y^T}\bar{y}) \tag{50}$$

$$\nabla J(\theta) = \frac{1}{2}\nabla(tr\bar{\theta^T}\bar{X^T}\bar{X}\bar{\theta} - tr2\bar{y^T}\bar{X}\bar{\theta}) \tag{51}$$

And, since $trA = trA^T$:

$$\nabla J(\theta) = \frac{1}{2}(2\bar{X^T}\bar{X}\bar{\theta} - 2\bar{X^T}\bar{y}) = \bar{X^T}\bar{X}\bar{\theta} - \bar{X^T}\bar{y} \tag{52}$$

-

We set $J(\theta) = 0$:

$$\bar{X^T}\bar{X}\bar{\theta} - \bar{X^T}\bar{y} = 0 \tag{53}$$

$$\bar{X^T}\bar{X}\bar{\theta} = \bar{X^T}\bar{y} \tag{54}$$

$$\bar{\theta} = (\bar{X^T}\bar{X})^{-1}\bar{X^T}\bar{y} \tag{55}$$

QED