Machine Learning 2016/2017: Assignment 1
Deadline: September 16
Isabel Walter

1. A) Given is the historical data. Goal is finding out whether a team will lose, win or draw against Ajax. The learning task is supervised because you have data of known results. Furthermore, it is a classification, because you only have three different outcomes (win/lose/draw).
   B) The data will consist of the score after a certain time has passed in the match and the end result of the match for the competing team, that is whether they won/lost/played a draw.
   E.g.:

Score after 45 min of match (competing team-Ajax)	End result competing team
2-1	win
1-1	draw
2-2	draw
3-1	win
0-0	draw
0-1	lose

2. A)
   Gradient descent formula:
$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$
   Starting situation: $\alpha = 0.1$ $\theta_0 = 0$ $\theta_1 = 1$
   θ_0 = 0 – 0.1 * (1/3) ((1 * 3 – 6) * 1 + (1 * 5 – 7) * 1 + (1 * 6 – 10) * 1) = 0.3
   θ_1 = 1 – 0.1 * (1/3) ((1 * 3 – 6) * 3 + (1 * 5 – 7) * 5 + (1 * 6 – 10) * 6) = 2.4333333 ..
   θ_0 = 0.3 – 0.1 * (1/3) ((0.3 + 2.433 * 3 – 6) * 1 + (0.3 + 2.433 * 5 – 7) * 1 + (0.3 + 2.433 * 6 – 10) * 1) = -0.09888 ..
   θ_1 = 2.433 – 0.1 * (1/3) ((0.3 + 2.433 * 3 – 6) * 3 + (0.3 + 2.433 * 5 – 7) * 5 + (0.3 + 2.433 * 6 – 10) * 6) = 0.3822 ..

   So:
   $h_{(\theta)}$ = -0.098 + 0.3822 x

   mean-squared error:
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$
   $J(\theta_0, \theta_1)$ = (1/6) ((-0.098 + 0.3822*3 – 6)^2 + (-0.098 + 0.3822*5 – 7)^2 + (-0.098 + 0.3822*6 – 10)^2) = 18.72205

   B)
   z-score: (x – mean)/standard deviation

   z-scores X:
   mean = (3 + 5 + 6)/3 = 4.666666667
   sd = sqrt(1/2((3 – mean)^2 + (5 – mean)^2 + (6 – mean)^2)) = 1.5275 ..
   x1 = (3 – mean)/ sd = -1.0910 ..
   x2 = (5 – mean)/ sd = 0.2182 ..
   x3 = (6 – mean)/ sd = 0.87287 ..

Starting situation: $\alpha = 0.1$ $\theta_0 = 0$ $\theta_1 = 1$
(Using the gradient descent formula from 2A)

$\theta_0 = 0 - 0.1 * (1/3)((1 * -1.0910 - 6) * 1 + (1 * 0.2182 - 7) * 1 + (1 * 0.87287 - 10) * 1) = 0.7666643333$
$\theta_1 = 1 - 0.1 * (1/3)((1 * -1.0910 - 6) * -1.0910 + (1 * 0.2182 - 7) * 0.2182 + (1 * 0.87287 - 10) * 0.87287) = 1.057$ ..
$\theta_0 = 0.7667 - 0.1 * (1/3)((0.7667 + 1.057 * -1.0910 - 6) * 1 + (0.7667 + 1.057 * 0.2182 - 7) * 1 + (0.7667 + 1.057 * 0.87287 - 10) * 1) = 1.4566621$
$\theta_1 = 1.057 - 0.1 * (1/3)((0.7667 + 1.057 * -1.0910 - 6) * -1.0910 + (0.7667 + 1.057 * 0.2182 - 7) * 0.2182 + (0.7667 + 1.057 * 0.87287 - 10) * 0.87287) = 1.110218304$

So:
$h_{(\theta)} = 1.4566621 + 1.110218304\, x$

mean-squared error:

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

$J(\theta_0, \theta_1) = (1/6)((1.4566621 + 1.110218304 * -1.0910 - 6)^2 + (1.4566621 + 1.110218304 * 0.2182 - 7)^2 + (1.4566621 + 1.110218304 * 0.87287 - 10)^2) = 19.76437286$

Thus the mean squard error has become bigger using scaled values of x. You scale the values because you want to make gradient descent quicker and therefore make the convergence much faster (so in less iterations). However, it seems that the scaling did not really help for the convergence. Perhaps the the learning rate was too big for the scaled values: Since scaled values are much smaller than the normal values a step of 0.1 could already have a way bigger impact.

3.

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

A) $h_{(\theta)} = \theta_0 + \theta_1 x1 + a\theta_2 + b\theta_1 x1$ instead of $h_{(\theta)} = \theta_0 + \theta_1 x1$, regarding the mean squared error formula …..
B) $h_{(\theta)} = \theta_0 + \theta_1 x1 + a\theta_2 + b\theta_1 x1^2$ instead of $h_{(\theta)} = \theta_0 + \theta_1 x1$

4. $$\frac{\partial}{\partial\theta_1} J(\theta 1) = \frac{1}{m} \sum_{i=1}^{m} ((\theta_0 + \theta_1 x^{(i)}) - y^{(i)})\, x^{(i)} = 0$$
$$\frac{\partial}{\partial\theta_1} J(\theta 1) = \sum_{i=1}^{m} ((\theta_0 - y^{(i)}) x^{(i)} + \theta_1 x^{(i)2}) = 0$$
$$\frac{\partial}{\partial\theta_1} J(\theta 1) = \sum_{i=1}^{m} \theta_0 x^{(i)} - \sum_{i=1}^{m} y^{(i)} x^{(i)} + \sum_{i=1}^{m} \theta_1 x^{(i)2} = 0$$
$$\frac{\partial}{\partial\theta_1} J(\theta 1) = \theta_0 \sum_{i=1}^{m} x^{(i)} - \sum_{i=1}^{m} y^{(i)} x^{(i)} + \theta_1 \sum_{i=1}^{m} x^{(i)2} = 0$$
$$\theta_1 \sum_{i=1}^{m} x^{(i)2} = -\theta_0 \sum_{i=1}^{m} x^{(i)} + \sum_{i=1}^{m} y^{(i)} x^{(i)}$$

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