AUC Course: Machine Learning

Solutions to Assignment 4

Vincent Roest, 10904816
vincentr@live.nl

I Worked together with Bas Straathof in my quest to correct solutions

Question 1

1a) The data set is represented in Figure . The blue squares are the (x_1, x_2) data points that correspond to $y = 0$, and conversely the red circles represent (x_1, x_2) data points that have $y = 1$. Note that the figure only represents an approximation, no calculations were done to create it. In this and all the following graphs the vertical axis accounts for the x_2 values and the x_1 values are represented on the horizontal axis.

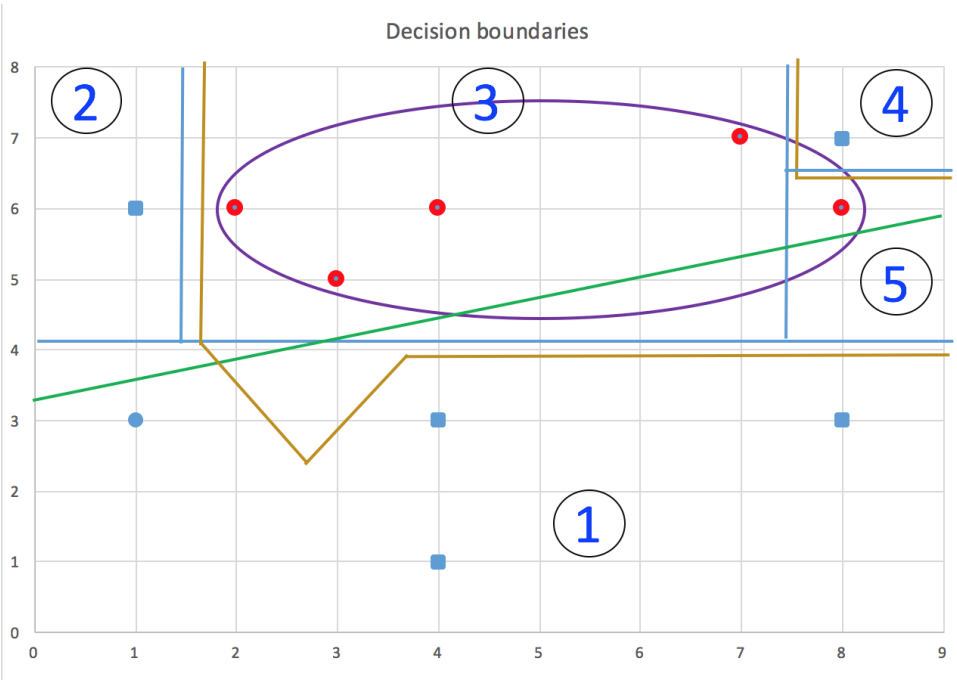

Figure 1: Decision Boundaries: Decision tree (blue), 1-nearest-neighbor (golden), plain logistic regression (green) and logistic regression with quadratic terms (purple)

Decision tree decision boundary:
By using decision trees you are restricted to an axis aligned split of the data. Given this data

set, we definitely run into difficulty in classifying this data set properly. If we use a sufficiently large decision tree, however, we might be able to classify the data set to perfection, because all the distinct data points will be singled out into separate points. If we choose this decision tree, however, our model will (likely) be unable to generalize for new examples. In order to combat this high variance, we could use pruning on the decision tree in order to obtain a more general classification. The approximate classification of the data set after this pruning might look as shown in Figure 2. In the regions *1, 2* and *4* we would classify a point as a $y = 0$ blue square, whereas in region *3* we would classify a data point as a red circle $y = 1$. If we use this boundary, we can see that we have to allow one misclassification.

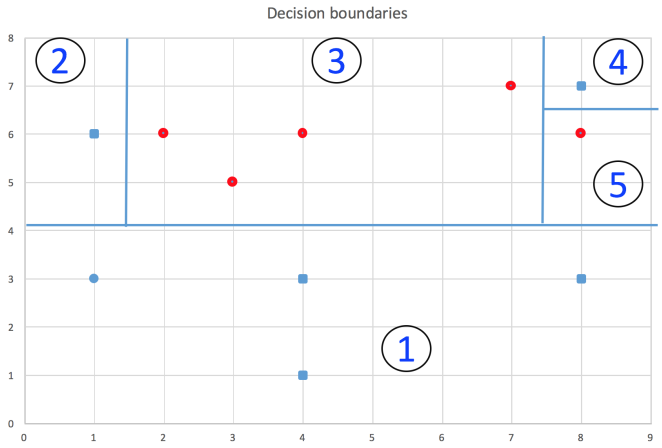

Figure 2: Decision tree boundary

1-nearest neighbor decision boundary:
In using the 1-nearest classification, which is the most simple adaptation of the k-nearest neighborhood, we construct the boundaries in such a way that should we introduce a new data point $(x_1, x2)$, it is classified $y = 1$ or $y = 0$ according to the value of the existing data point which is closest to the new data point. An approximate decision boundary for the data set presented might look like Figure 3. A new data point that falls in regions *1* & *3* will be classified as a $y = 0$ blue square, whereas new points in region *2* will be classified as $y = 1$ red circles.

Plain logistic regression decision boundary:
Plain logistic regression is here conceived of as a classification using a regression line $\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$ that defines the boundary. As we can immediately see, it is impossible given this data set to classify all points correctly using the plain logistic regression. However, if we approximately draw a line with choices $\theta_0 = 3$ and $x_2 \approx \frac{1}{3}x_1$ we get a boundary $x_2 = \frac{1}{3}x_1 + 3$ that fits the data set pretty well, see Figure 4. Naturally, we use the Sigmoid function in logistic regression. If we define the vectors $x = [x_0, x_1, x_2]$ and $\theta = [3, \frac{1}{3}, -1]$, we can use the Sigmoid function as $g = \frac{1}{1+e^{-\theta^T x}}$. Then, the hypothesis function for plain logistic regression

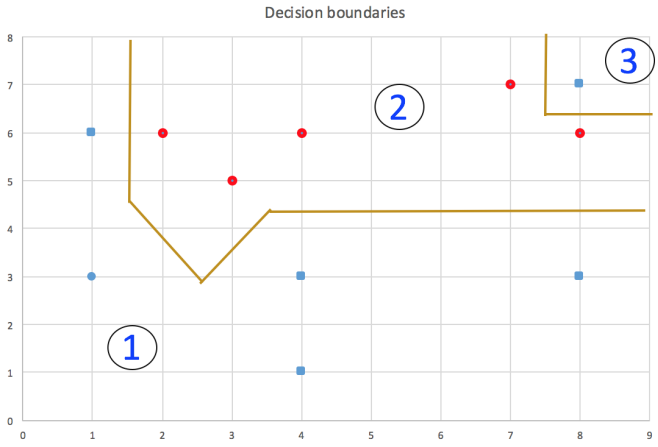

Figure 3: 1-nearest neighbor boundary

$h_\theta(x_1, x_2) = g(-3 - \frac{x_1}{3} + x_2)$ will classify a data point as a blue square ($y = 0$) if $h_\theta(x_1, x_2) < \frac{1}{2}$ and similarly classify a data point as a red circle ($y = 1$) if $h_\theta(x_1, x_2) > \frac{1}{2}$.

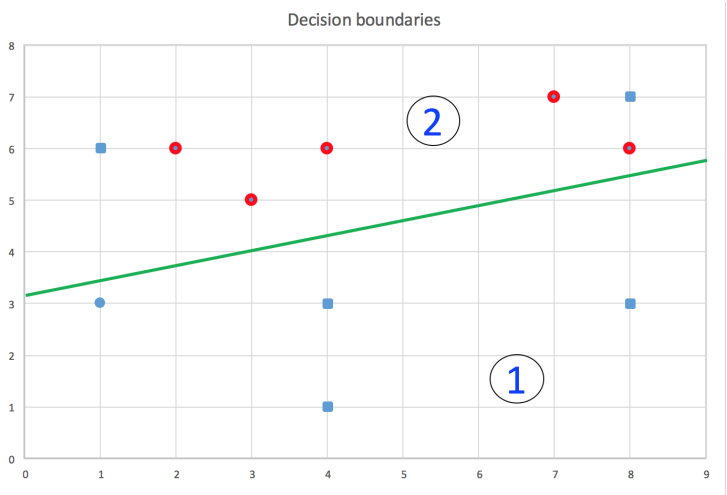

Figure 4: Plain logistic regression boundary

Logistic regression with quadratic terms:
Let us continue the line of reasoning from the previous method, and we simply add quadratic terms to the classifier. For this data set, we can correctly classify all data points by using an ellipse, as shown in Figure 5. The formula for this ellipse is approximately (my Excel drawing

skills fall miserably short sometimes):

$$\frac{(x_1-5)^2}{3.25^2}+\frac{(x_2-6)^2}{1.5^2}=1$$

Then again, similar to the previous classifier, we get the following two cases: classify as a blue square ($y=0$) if the hypothesis $h_\theta(x_1,x_2)=g(1-\frac{(x_1-5)^2}{3.25^2}-\frac{(x_2-6)^2}{1.5^2})<\frac{1}{2}$ and classify as a red circle ($y=1$) if the hypothesis $h_\theta(x_1,x_2)=g(1-\frac{(x_1-5)^2}{3.25^2}-\frac{(x_2-6)^2}{1.5^2})>\frac{1}{2}$.

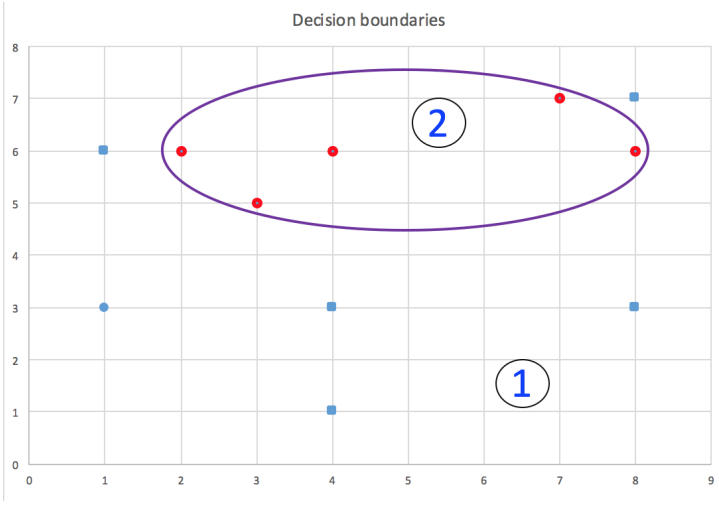

Figure 5: Quadratic logistic regression boundary.

1b) From the decision boundaries above, we can immediately see that the decision tree and plain logistic regression would not be the most suitable choice if we had to classify this data set. They even do not really fit the given (training) data well, and future examples will probably not be classified correctly. The logistic regression with quadratic terms, however, seems to perfectly fit the data set. The approximate boundary that was provided, however, only just captures every data point (that was the purpose of adding the quadratic terms) and is pretty strict for examples $y=1$, and perhaps future examples will fall just outside the region and be classified incorrectly, as some of the data points are already close to the boundary and are most probably not outliers. The main advantage of the 1-nearest neighborhood method over the logistic regression with quadratic terms is that it's regime where points are classified as $y=1$ is a great deal larger, paving way for a better classification of future data points (it has lower variance). In addition, it shares the key property with logistic regression, namely it also separates the data set given perfectly. For this data set, it would be sufficient to implement the 1-nearest neighbor method, which for large data sets is usually quite slow (it has to calculate a lot of distances if the data set is large), but it will be pretty efficient for the data set at hand. Intuitively, I would like combine the two methods of 1-nearest neighborhood and quadratic logistic regression. The space where

the logistic regression classifies a data point $x_1, x_2)$ as $y = 1$ is almost completely contained in the $y = 1$ decision space of the 1-nearest neighbor classifier. Suppose that we encounter a new data point that is contained in both $y = 1$ spaces, we could with a good certainty classify that as a $y = 1$ point. If the new point lies inside the $y = 1$ space of the nearest neighbor but not in the ellipse space, we could choose to classify the point as $y = 1$ with a certain probability, and otherwise classify it as 0 value. In this way, we rely on the strict classification boundary of the logistic regression, but we also accommodate for its outliers.

Question 2 (k-means only)

As the question dictates we can assume that there are 3 clusters, whose means are defined as follows: $\mu_1 = 1, \mu_2 = 3,$ and $\mu_3 = 8$, and we also have 16 data points Basically, every iteration consists of two separate steps, namely:
For every i (16 times), we set the calculate $c_(i)$:

$$c_(i) := argmin_j ||x(i) - \mu_j||^2$$

And then, for every j, we "shift" the means:

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

The first inner-loop does the following calculations:

$$c_{(1)} := argmin_j ||1 - \mu_j||^2 = 1$$

$$c_{(2)} := argmin_j ||2 - \mu_j||^2 = 1$$

$$c_{(3)} := argmin_j ||3 - \mu_j||^2 = 2$$

$$c_{(4)} := argmin_j ||3 - \mu_j||^2 = 2$$

$$c_{(5)} := argmin_j ||4 - \mu_j||^2 = 2$$

$$c_{(6)} := argmin_j ||5 - \mu_j||^2 = 2$$

$$c_{(7)} := argmin_j ||5 - \mu_j||^2 = 2$$

$$c_{(8)} := argmin_j ||7 - \mu_j||^2 = 3$$

$$c_{(9)} := argmin_j ||10 - \mu_j||^2 = 3$$

$$c_{(10)} := argmin_j ||11 - \mu_j||^2 = 3$$

$$c_{(11)} := argmin_j ||13 - \mu_j||^2 = 3$$

$$
\begin{aligned}
c_{(12)} &:= argmin_j ||14 - \mu_j||^2 = 3 \\
c_{(13)} &:= argmin_j ||15 - \mu_j||^2 = 3 \\
c_{(14)} &:= argmin_j ||17 - \mu_j||^2 = 3 \\
c_{(15)} &:= argmin_j ||20 - \mu_j||^2 = 3 \\
c_{(16)} &:= argmin_j ||21 - \mu_j||^2 = 3
\end{aligned}
$$

And afterwards, we perform the calculations of the second inner-loop:

$$
\begin{aligned}
\mu_1 &:= \frac{\sum_{i=1}^{m=16} 1\{c^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m=16} 1\{c^{(i)}=1\}} \\
&:= \frac{1+2}{2} \\
&:= \frac{3}{2}
\end{aligned}
$$

$$
\begin{aligned}
\mu_2 &:= \frac{\sum_{i=1}^{m=16} 1\{c^{(i)}=2\}x^{(i)}}{\sum_{i=1}^{m=16} 1\{c^{(i)}=2\}} \\
&:= \frac{3+3+4+5+5}{5} \\
&:= 4
\end{aligned}
$$

$$
\begin{aligned}
\mu_3 &:= \frac{\sum_{i=1}^{m=16} 1\{c^{(i)}=3\}x^{(i)}}{\sum_{i=1}^{m=16} 1\{c^{(i)}=3\}} \\
&:= \frac{7+10+11+13+14+15+17+20+21}{9} \\
&:= 14.22
\end{aligned}
$$