Written Assignment 4

Marit Hoefsloot

November 18th, 2016

Exercise 1

1. The class boundaries

We consider the data set:

x_1	x_2	y
1	3	0
1	6	0
2	6	1
3	5	1
4	1	0
4	3	0
4	6	1
7	7	1
8	6	1
8	7	0
8	3	0

The red dots represent $y = 0$ and the blue dots represent $y = 1$.

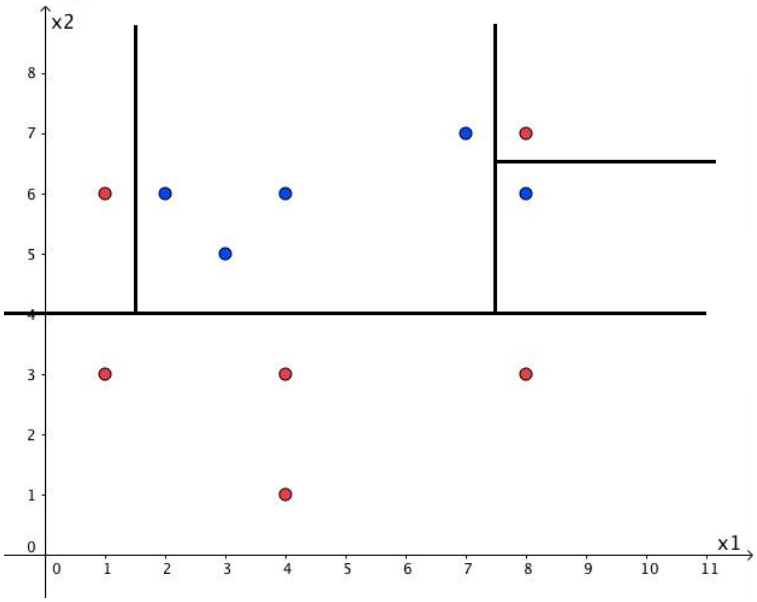

Figure 1: Class boundaries found by decision trees

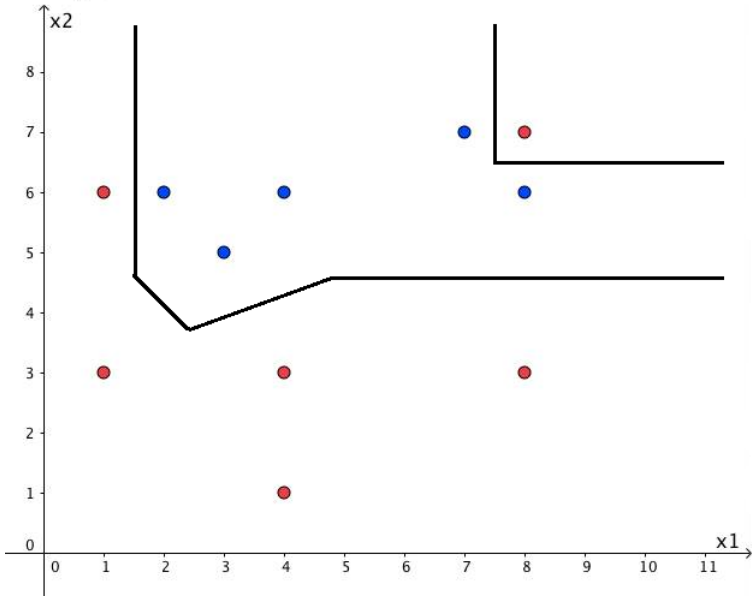

Figure 2: Class boundaries found by the 1-nearest neighbour

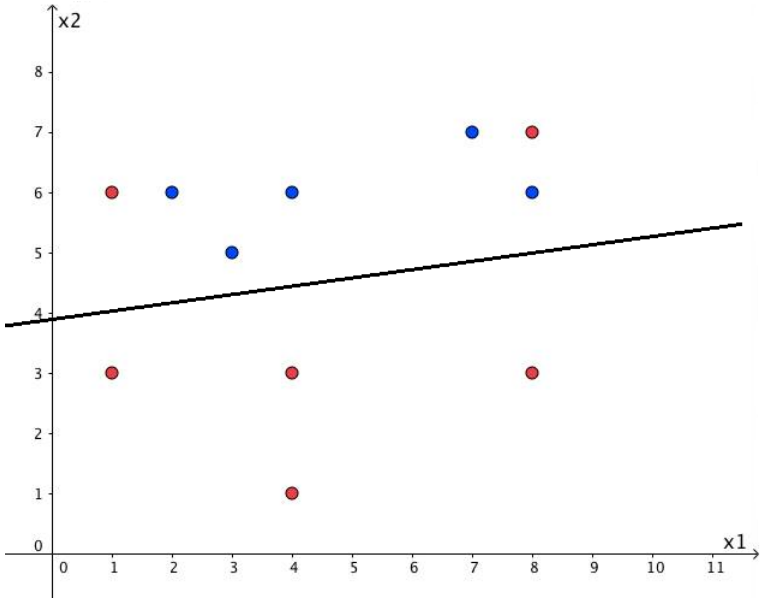

Figure 3: Class boundaries found by plain logistic regression

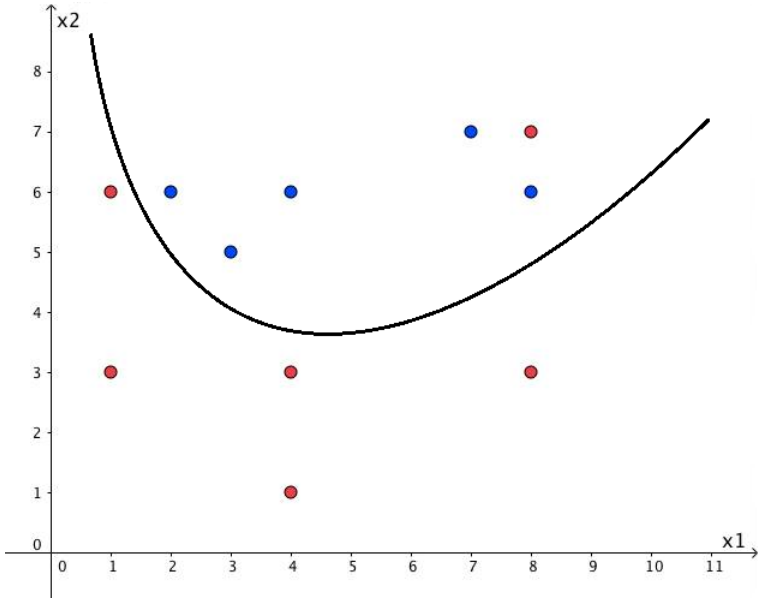

Figure 4: Class boundaries found by logistic regression with quadratic terms

2. Do you intuitively think that one boundary is better than another? It may be possible to use such an intuition to invent method that uses multiple learning algorithms and combine the results, using your intuition as a prior probability. Explore this line of thought.

As is clearly visible from looking at the plots of the different algorithms, there are some advantages and disadvantages to each algorithm. I think that the nearest neighbour algorithm adapts best to identifying different regions. For example, we could have a data set that includes red dots in the areas where both x_1 and x_2 are small or both are big and blue dots everywhere else (see Figure 5). The 1-nearest neighbour algorithm differentiates between the three regions in a very successful way, better than for example logistic regression could. However, logistic regression with quadratic terms fits better to the general data without getting too affected by outliers. If it would be possible to combine those two, you would get an algorithm that differentiates the different regions, does not get affected too much by outliers and that gives a good overall fit to the data.

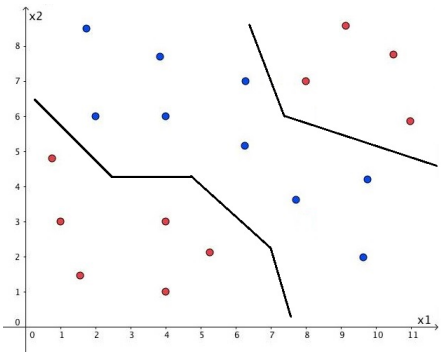

Figure 5: Example of regions found by 1-nearest neighbour

Exercise 2

Manually calculate 1 iteration of k-means clustering for the 1-dimensional data below. Assume that there are 3 clusters and initialize the means with 1, 3 and 8. Calculate the cost for k-means before and after this step. Data set: $(1, 2, 3, 3, 4, 5, 5, 7, 10, 11, 13, 14, 15, 17, 20, 21)$ and $m = 16$.

We have three clusters with three respective means (1, 3 and 8), so we can assign the data points to the nearest cluster:

Data point	Cluster	Mean
1	1	1
2	1	1
3	2	3
3	2	3
4	2	3
5	2	3
5	2	3
7	3	8
10	3	8
11	3	8
13	3	8
14	3	8
15	3	8
17	3	8
20	3	8
21	3	8

The second data point (with number 2) is exactly in the middle of two clusters, so I flipped a coin and assigned the data point to the first cluster.

Firstly, I calculate the cost before the first iteration:

$$\begin{aligned} J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K) &= \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2 \\ &= \frac{1}{16} \cdot (|1-1|^2 + |2-1|^2 + |3-3|^2 + |3-3|^2 + |4-3|^2 \\ &+ |5-3|^2 + |5-3|^2 + |7-8|^2 + |10-8|^2 + |11-8|^2 + |13-8|^2 \\ &+ |14-8|^2 + |15-8|^2 + |17-8|^2 + |20-8|^2 + |21-8|^2) \\ &= \frac{528}{16} = 33 \end{aligned}$$

Secondly, I update the means by taking the new means of the clusters. So for every data point that gets added, I update the mean for that cluster:

Data point	Cluster	Updated mean
1	1	1
2	1	1.50
3	2	3
3	2	3
4	2	3.33
5	2	3.75
5	2	4
7	3	7
10	3	8.50
11	3	9.33
13	3	10.25
14	3	11
15	3	11.67
17	3	12.43
20	3	13.38
21	3	14.22

So the final means for the different clusters are: $\mu_1 = 1.50$, $\mu_2 = 4$ and $\mu_3 = 14.22$.

Lastly, I calculate the cost function again:

$$\begin{aligned}
J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K) &= \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2 \\
&= \frac{1}{16} \cdot (|1 - 1.50|^2 + |2 - 1.50|^2 + |3 - 4|^2 + |3 - 4|^2 \\
&+ |4 - 4|^2 + |5 - 4|^2 + |5 - 4|^2 + |7 - 14.22|^2 + |10 - 14.22|^2 \\
&+ |11 - 14.22|^2 + |13 - 14.22|^2 + |14 - 14.22|^2 + |15 - 14.22|^2 \\
&+ |17 - 14.22|^2 + |20 - 14.22|^2 + |21 - 14.22|^2) \\
&= \frac{174.06}{16} \approx 10.88
\end{aligned}$$

So, only one iteration of the k-means algorithm can already make a significant difference in the cost.