Written Assignment 4

Machine Learning

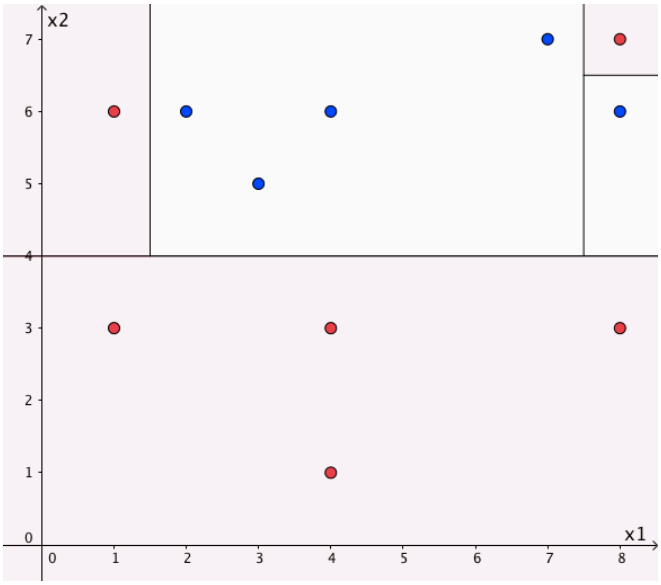
Leah Dickhoff
S11155515

due November 18th, 2016

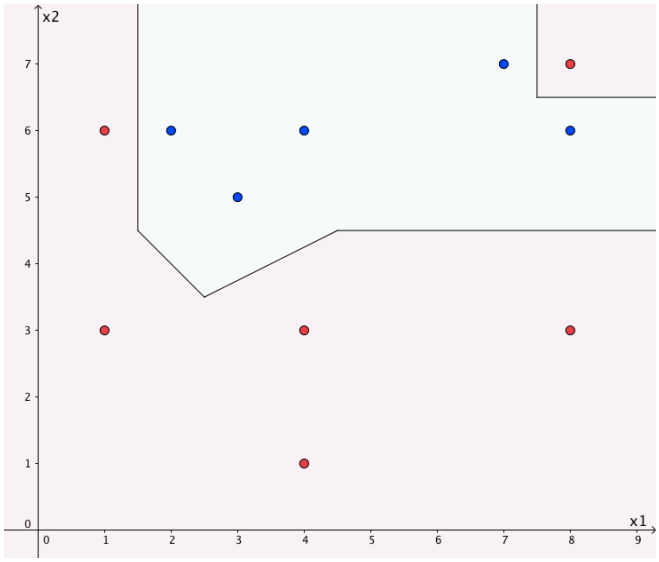
1.

a) In the following graphs, the colour red indicates a y-output of 0, and the colour blue indicates a y-output of 1.

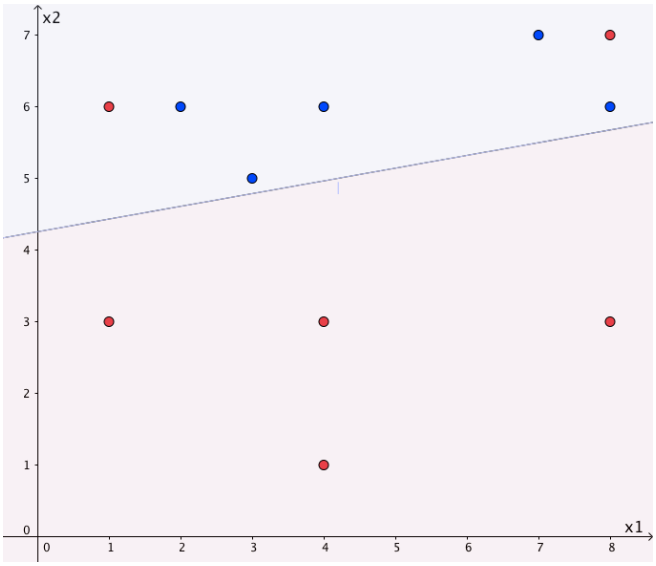
Boundary found by Decision Trees:



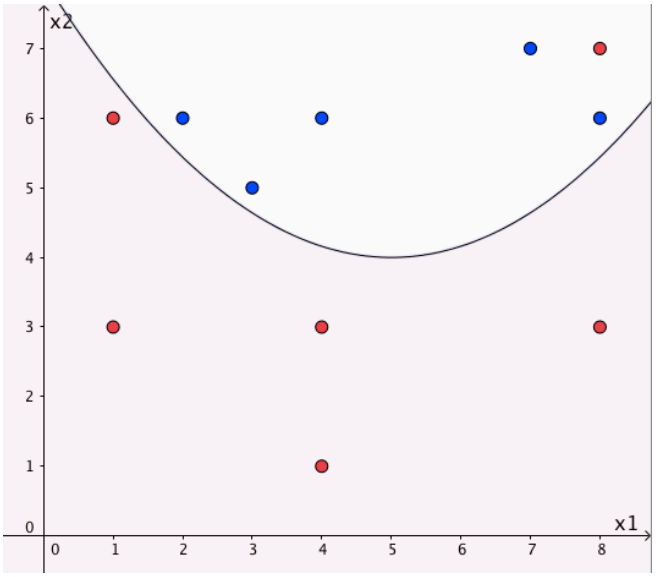
Boundary found by 1-nearest neighbour:



Boundary found by plain logistic regression:



Boundary found by logistic regression with quadratic terms:



b) As to the advantages of the different algorithms (that one can clearly see in a)), the one of 1-nearest neighbour seems to be the best one to fit every single point of the dataset, as it deals best with outliers; whereas the logistic regression with quadratic terms has a better overall fit on the data. Hence, if one could find a way to combine both, then the results (good overall and outlier fit) are likely to be even better than the results found by the two algorithms separately. The 1-nearest algorithm would succeed in perfectly separating the red and blue regions for every single data point, and the logistic regression would 'even out' the boundary so it has a better general fit, and does not get affected too much by extreme outliers.

2. Dataset: 1, 2, 3, 3,4, 5, 5, 7,10, 11, 13, 14, 15,17, 20, 21

There are 3 clusters with 3 different means, so assigning each data point to the nearest mean, we get:

data point	closest mean	assigned cluster
1	1	1
2	1	1
3	3	2
3	3	2
4	3	2
5	3	2
5	3	2
7	8	3
10	8	3
11	8	3
13	8	3
14	8	3
15	8	3
17	8	3
20	8	3
21	8	3

note: data point '2' is as close to mean 1 as to mean 3, so it is randomly assigned

Before the 1st iteration (as there are 16 data points, m=16):

$J(c^{(1)},\ldots,c^{(16)},\mu_{(1)},\mu_{(16)}) = \frac{1}{16}\sum_{i=1}^{16}\left\|x^{(i)}-\mu_{(i)}\right\|^2 = 33$

where: $x^{(i)}$ is the i^{th} data point and

$\mu_{(i)}$ the mean associated with that data point

Now, the means are updated to:

data point	mean after update	assigned cluster	final mean for each cluster
1	1,00	1	1,50
2	1,50	1	1,50
3	3,00	2	4,00
3	3,00	2	4,00
4	3,33	2	4,00
5	3,75	2	4,00
5	4,00	2	4,00
7	7,00	3	14,23
10	8,50	3	14,23
11	9,33	3	14,23
13	10,25	3	14,23
14	11,00	3	14,23
15	11,66	3	14,23
17	12,43	3	14,23

20	13,37	3	14,23
21	14,23	3	14,23

The cost function, after this step, is:

$J(c^{(1)},\ldots,c^{(16)},\mu_{(1)},\mu_{(16)}) = \frac{1}{16}\sum_{i=1}^{16}\left\|x^{(i)}-\mu_{(i)}\right\|^2 = 10{,}86$

which is already a lot smaller than the cost before the first iteration of the k-means algorithm.