# Machine Learning - Assignment 2

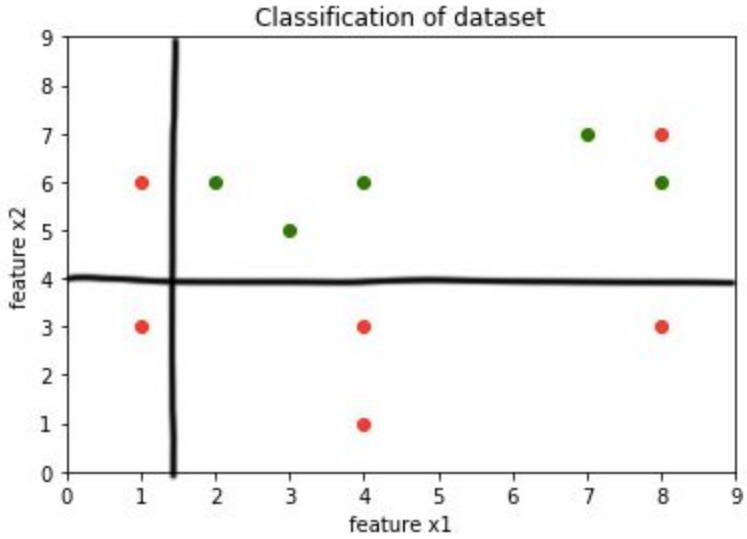
by
Maxime Gevers

Question 1a

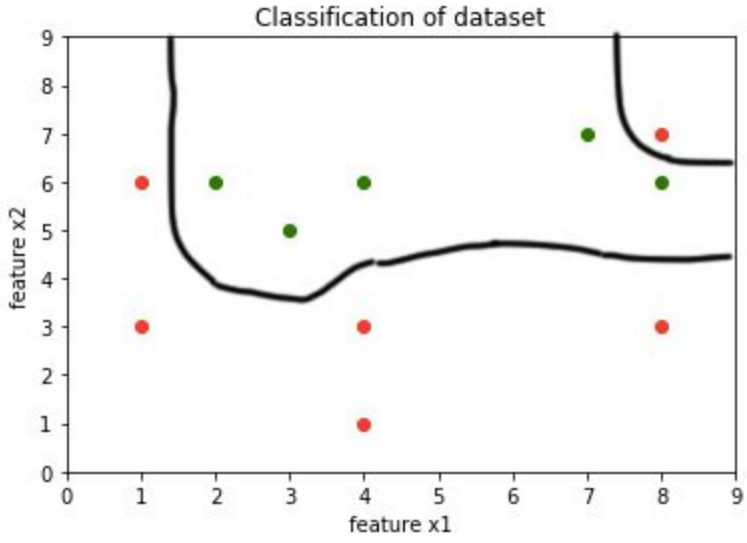
x1 = {1, 1, 2, 3, 4, 4, 4, 7, 8, 8, 8}
x2 = {3, 6, 6, 5, 1, 3, 6, 7, 6, 7, 3}
y = {0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0}

The following plots show the data points of the above dataset. The data points are classified as either 1 or 0. I have chosen the colour green to represent the data points that are labeled as 1 and the colour red to represent the data points that are labeled as 0.

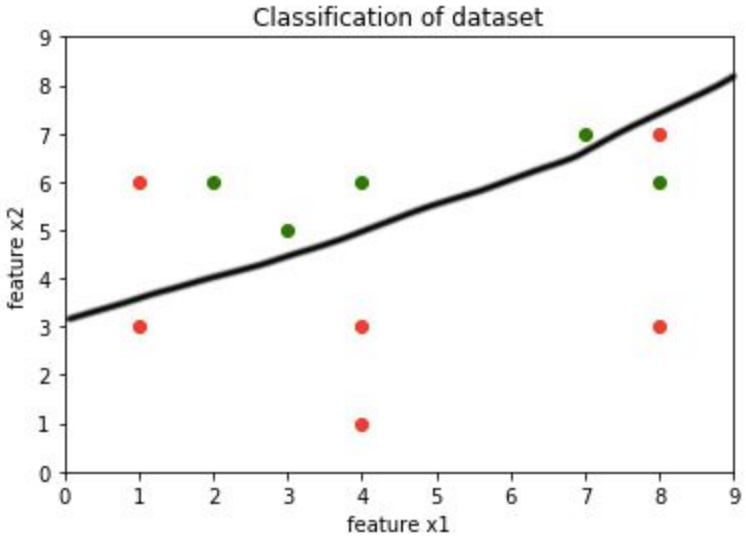
By Decision Tree:



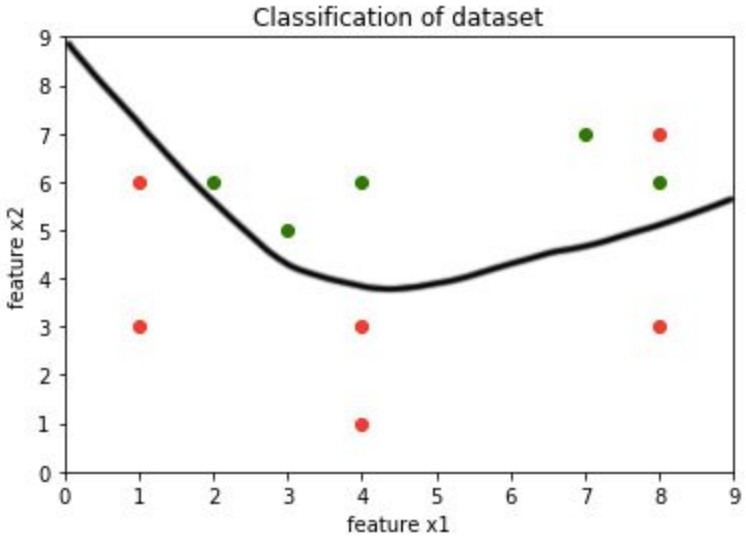

By 1 Nearest Neighbor:



By Logistic Regression:




By Logistic Regression with Quadratic Terms:



Question 1b

Yes, I think that some boundaries are better than others. As you see in the images above, it is clear that some data points are misclassified. The problem is that not all data points can be well-classified, since the data points don't form a clear specific cluster. For example, in the top right corner you can see that the red dot is very close to the green dots, but very far from all the other red dots. I would suggest that this is an outlier, meaning that this data point is a special case in comparison with the other data points. Nevertheless, this special case exists and has an influence on the decision boundary. For example, logistic regression doesn't work too well on this data set. It misclassified three data points, whereas knn does not misclassify any data points. In the end, it depends a lot on your data set (are there a lot of outliers?), if for example there only exists one data point 'far' from your decision boundary, then you can choose not to look at this data point. The accuracy of your decision boundary will still be high. When it happens that too many data points are misclassified, then you need to reevaluate your decision boundary or use another method for your decision boundary. In other words, some methods work better for certain datasets than others, therefore you need to investigate in advance with what kind of dataset you are working with and which method would work best.