Written Assignment 4

Wendy Nieuwkamer

20 November 2016

1 Question 1

Consider the dataset:

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

1.1 Draw the data points and class boundaries in 2D for four methods

The methods are: (1) Decision Trees, (2) 1-nearest neighbor, (3) plain logistic regression and (4) logistic regression with quadratic terms. The drawing should illustrate the differences but does not need to be correct by the millimeter.

I chose to make approximations for the decision boundaries, the results can be found in figure 1. For the first algorithm I used the decision tree in figure 2, the purple boundary is the first step, the red one the second and the orange boundary is the last step in the tree. For the nearest neighbour I attempted to draw all the boundaries and then traced those which split the zeros from the ones in red.Thus, the dark line is the decision boundary. For the logistic regression boundaries I made an approximation. The plain logistic regression is a linear boundary, the quadratic is a hyperbole. Everything under the boundary is expected to be zero, everything above it is one.

1.2 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.

I would like to start with excluding Plain Logistic Regression as there is no decision boundary found by this algorithm that will fit the data well, it

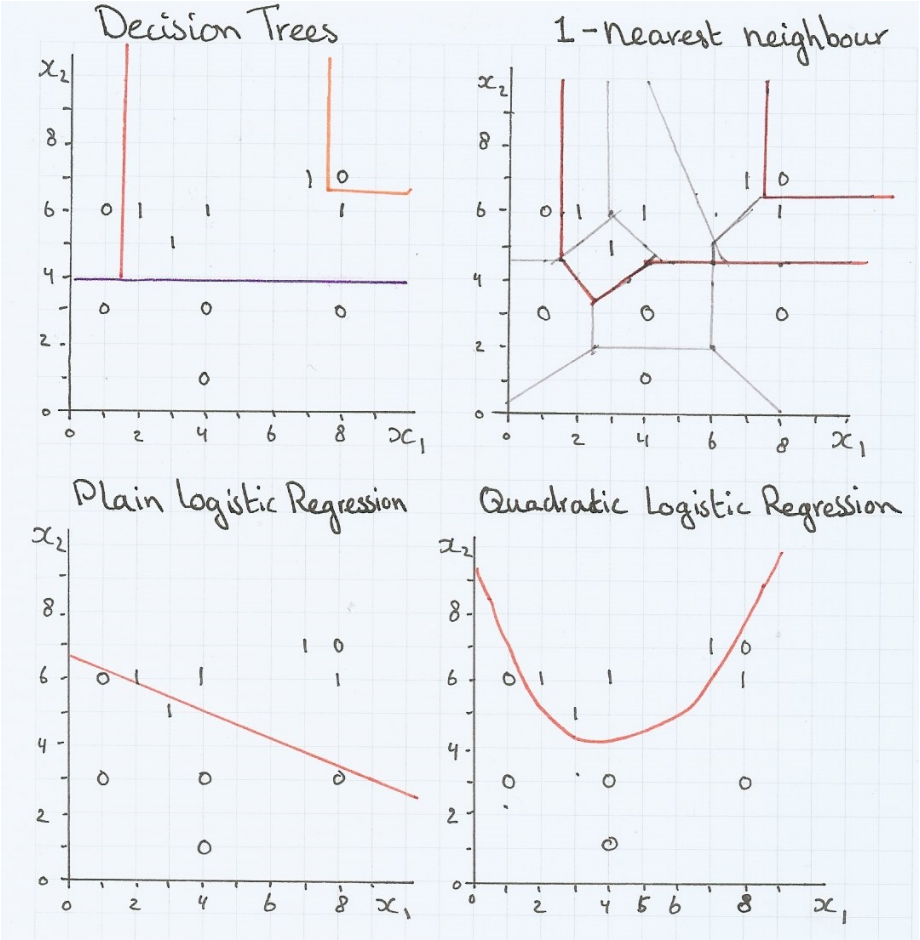

Figure 1: Four kinds of classification

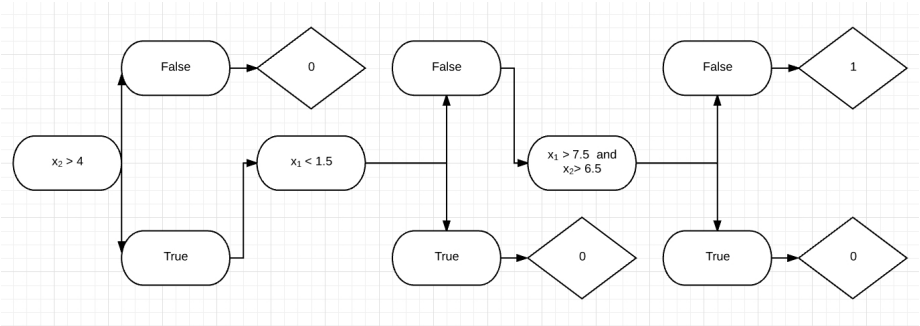

Figure 2: The decision tree used for classification

is biased. The data asks for a non-linear decision boundary. Then, we have Decision Trees, 1-Nearest Neighbour and Quadratic Logistic Regression left. The decision boundaries found by the first two are very similar. They divide the plane in three areas, two of which are zeros and one for ones. It seems strange to have three areas for two values, so these methods are probably over-fitting. The last method is Quadratic Logistic Regression. This method doesn't fit the data perfectly, but only has one value that will be predicted wrong with my decision boundary, so it is neither over-fitting nor biased. Also, it divides the plane in two areas which fit the two outcomes.

One thing that these three decision boundaries have in common is the fact that all inputs with $x_2 < 4$ have the value 0. We could combine the decision tree algorithm with quadratic logistic regression. we would first determine if $x_2 < 4$ is true, f it is we predict the example to belong to class 0; if it isn't we continue prediction using the boundary found by Quadratic Logistic Regression. The purpose of this is to prevent doing unnecessary computations, which could speed up predictions for big datasets considerably.