Written Assignment 4

Isobel Smith

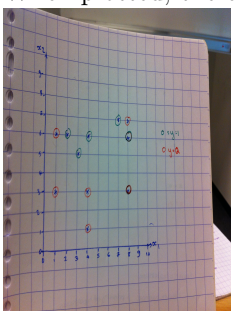
18/11/2016

1 Question 1: Drawing class boundaries

The dataset we have is:

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

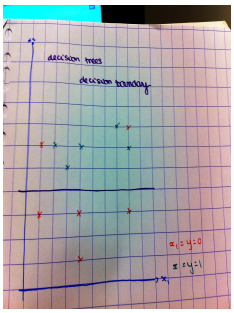
When plotted, the data looks like:



We have the class boundearies for the following:

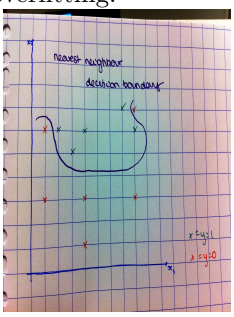
1.1 Decision trees

Decision trees find the best attribute (the one with the highest entropy) to split on. The data is subdivided into classes, and then tested to find the best fit. The decision boundaries are horizontal lines, dividing and seperating the classes, so can be ineffective.



1.2 nearest-neighbour

Nearest neighbour classifies points based on the points around it. It uses the voronoi tesselation to set up a voronoi cell around each cell, so is effective at finding the decision boundary. The danger of this effective decision boundary is overfitting.

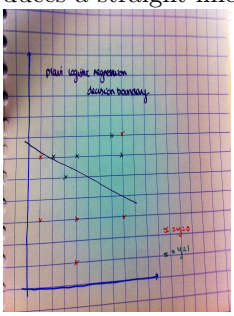


1.3 logistic regression

Logistic regression finds the decision boundary that best fits the data. This is calculated with the hypothesis:

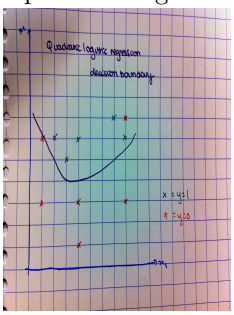
$$h_\theta(x) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x2$$

The hypothesis predicts y = 1 when $\theta^T x$ is greater than or equal to 0. This produces a straight line



1.4 logistic regression with quadratic terms

This would be expected to find a much better fit than "Plain" logistic regression. The quadratic logistic regression will produce a quatratic graph. (a curved line)



1.5 all classifiers

The graph below shows a summary of the different decision boundaries drawn by the classifiers