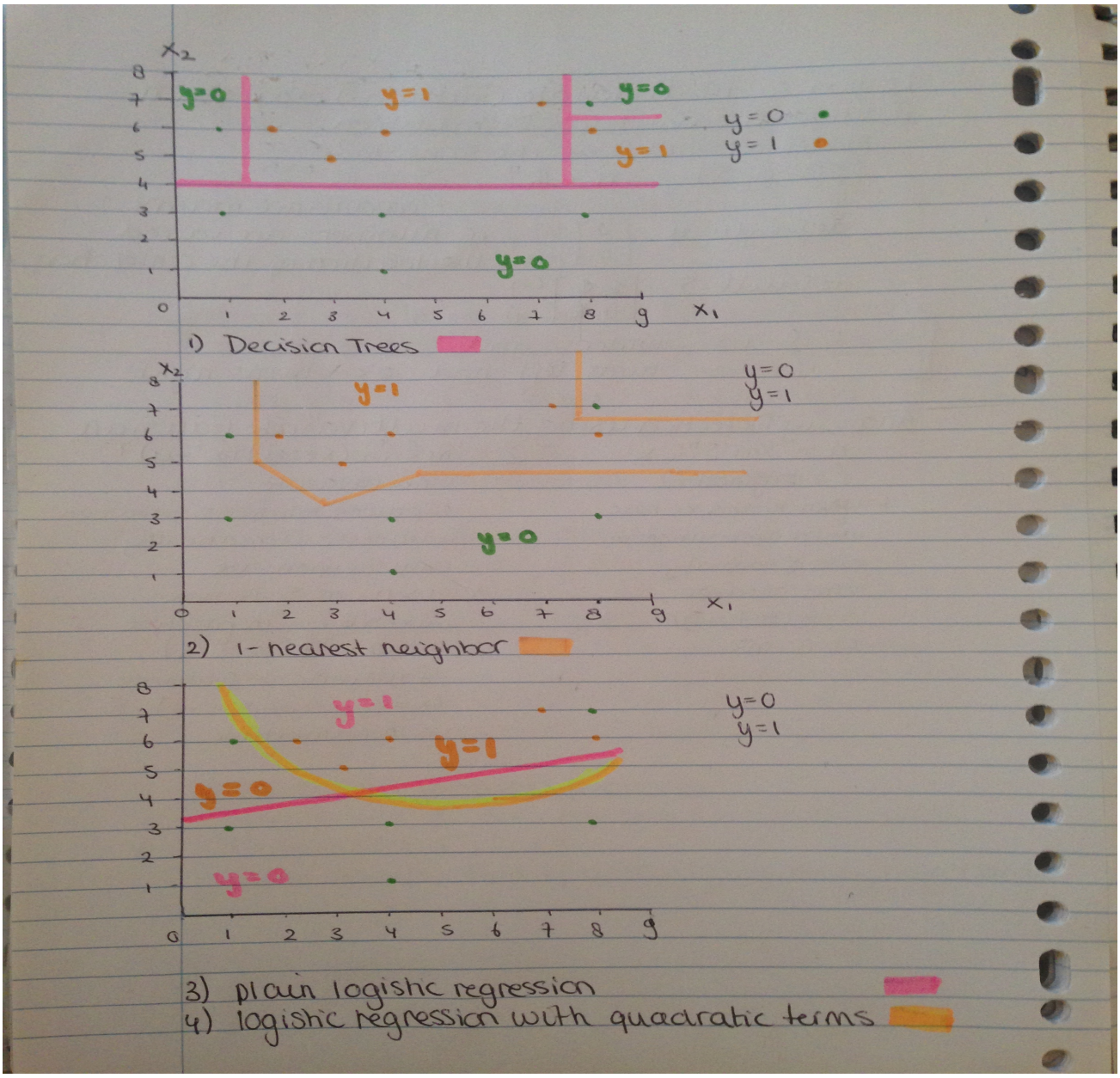
Machine Learning - Assignment 4

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}

a) Draw the data points in 2D and draw the class boundaries that would be found by **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. You are most welcome to use your own or other programs to calculate the class boundaries but is also OK if you make a reasonable approximation without calculating it precisely. Also, nice if plot everything in colors but is also OK if you make a clear drawing (or more) and include this in the submission.



(b) 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 believe in this case the 1-nearest neighbor and the decision trees are not optimal, and neither is plain logistic regression so I believe the best boundry is the one predicted by logistic regression with quadratic terms. I believe the method that is described above is possible, but would require a lot of knowledge about the data. In this case that is not a big problem because there are only two features x1 and x2, but with more features the visual representation is not as clear or not possible in which case deciding which boundries to combine will be hard.

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