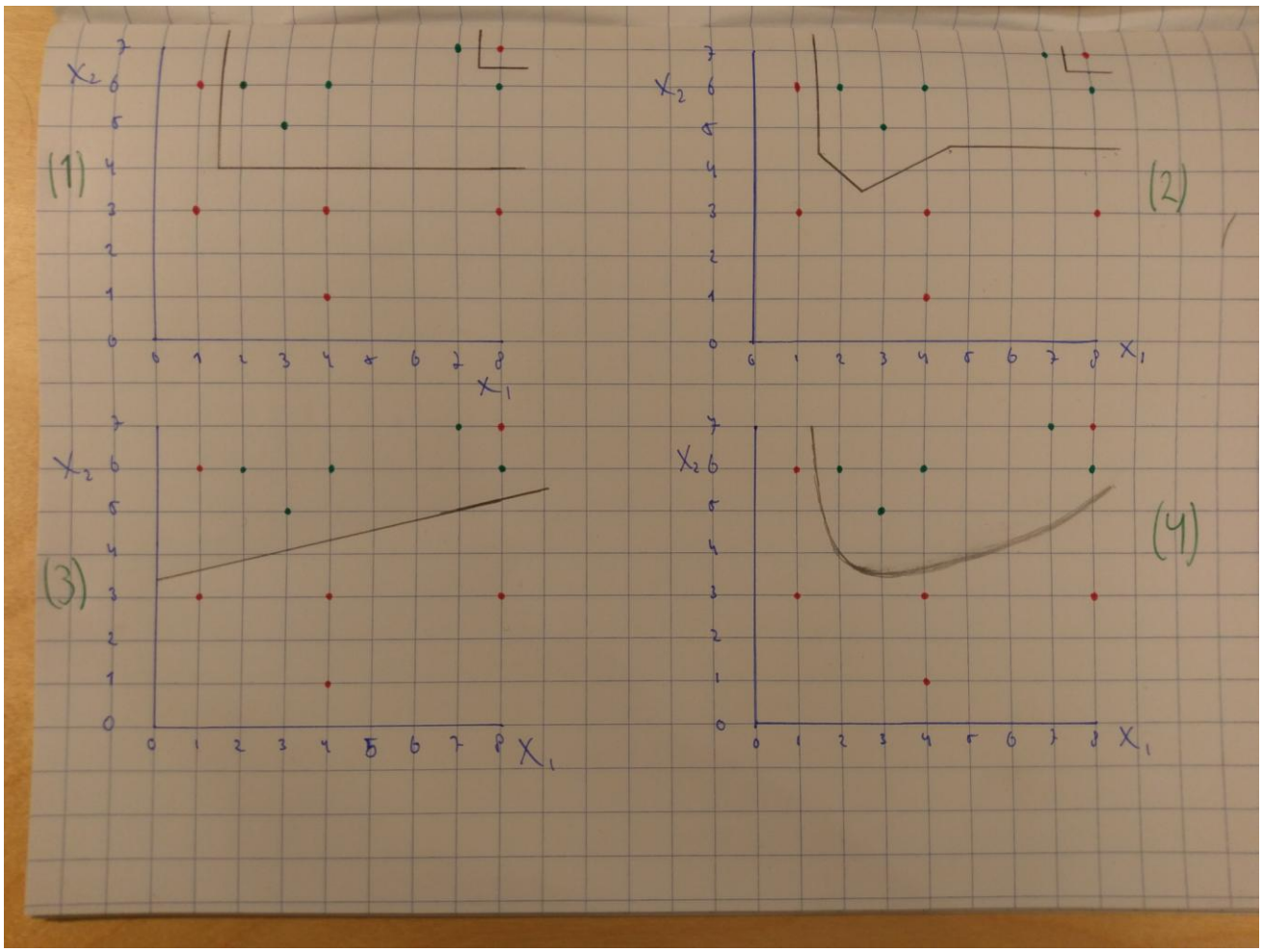
Machine Learning – Assignment 2 Part 2 - Bobbie van Gorp

(a)



(1) = Decision Trees
(2) = 1 Nearest Neighbor
(3) = Plain Logistic Regression
(4) = Logistic Regression with quadratic terms

(b) In this case there are not a lot of known data points and by looking at the points it is not immediately clear what the relation between x1, x2 and y is, which makes it difficult to say something about the best boundary. Of course the decision trees and 1 nearest neighbor boundaries fully separate the points with different y values from another by definition (while in both logistic regression classifications there are misclassified points even in this 'training set'), but this does not mean that they generalize well to new unseen data points.

If the training set is not going to be larger than these 11 points, a combination of these classifiers is certainly the best option in order to predict unseen data point as well as possible. An example of this could be 'majority voting': the class of a new unseen data point is predicted by all four (or even more) of these classifiers and the class that is predicted by most classifiers will be the prediction. If

there is a draw, it might be an option to do the majority voting again, but this time omitting the one classifier that seems most likely to predict the wrong class. In this case, if I follow my intuition, I would omit the plain logistic regression classifier from the majority voting. This classifier seems most likely to be making wrong predictions, because of underfitting: the dataset seems too complicated to be separated by a straight line.