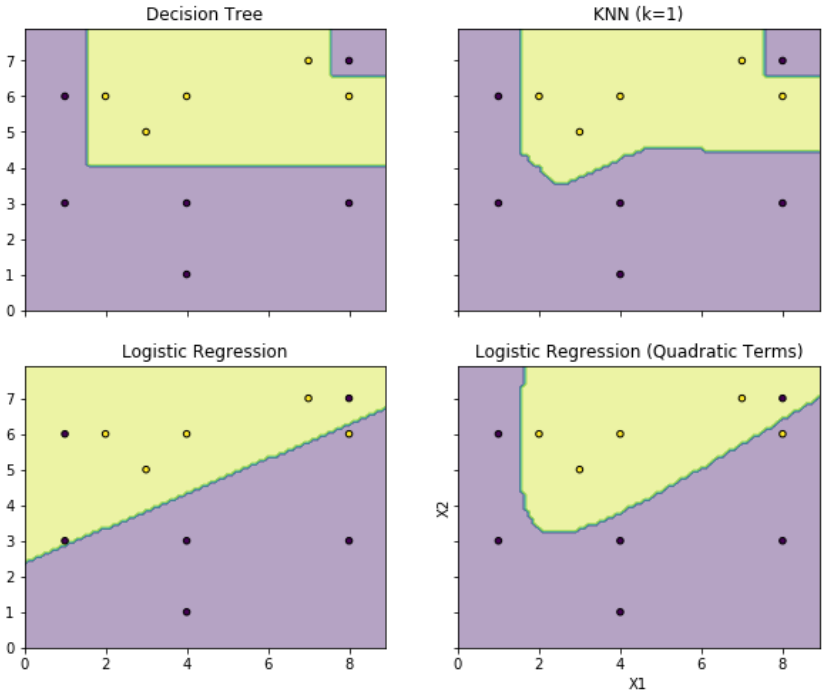
Machine Learning 2017/2018

Graded Assignment 2 Part 2

1.



After taking a look at the "Plot the decision boundaries of a VotingClassifier" tutorial on the scikit learn website and modifying it to work with the Logistic Regression and Single Nearest Neighbour classifiers, I got this as an output. The purple indicates elements with y=0, and the greenish indicates elements with y=1. The blue line between the two coloured areas shows the decision boundary.
As you can see, both Decision Tree and KNN fit snugly around each element of the set, while both Logistic Regression classifiers seem to fit more loosely around the data. I will discuss the pro's and cons of this in the next exercise.

2.

It's hard to tell from just the training input if one is better than the other, since we don't know what the actual data it's going to be applied to in real life is going to look like. However, if I would have to pick one that I think would generalise best, it's the quadratic term logistic regression classifier. This is mainly because it seems to strike the best balance between fitting the data right, while not overfitting it. Whereas the decision tree and KNN classifiers have the lowest error rate, they seem to ignore the fact that the purple dot on the top-right might be just an anomaly. On the other hand, while plain logistic regression doesn't get affected by this anomaly, it seems to underfit the data by ignoring the two purple dots on the top-left. The logistic regression classifier with quadratic terms seems to take the best of these three classifiers, but like I said, this is not sure since we don't really have enough data for that.