Practical 3

Dr Simon Denman
CAB420: Machine Learning

This weeks practical will focus on classification, and three classification methods:

- Support Vector Machines;
- K-Nearest Neighbours Classification.
- Random Forests

Problem 1. Binary Classification. Consider the data set *redwine-binary.csv.* This data contains both objective measurements on chemical and physical properties of the red wines, and subjective measurements of quality based on expert judegments. In this data, wine quality is the response variable and is either "above average" (1) or "below average" (0).

Using this data set:

1. Fit a Support Vector Machine to the data, and select appropriate values of C and an appropriate kernel to maximise accuracy.
2. Fit a K-Nearest Neighbours Classifier to the data, and select appropriate values of K and the distance metric to maximise accuracy.

For both models, repeat your experiments with and without standardising the data, and note any differences in performance.

Problem 2. Multi-Class Classification. In this question we are using the multi-class version of the red wine data, *redwine-multiclass.csv*, which contains fine-grained quality ratings, but is otherwise the same as the binary data.

Using this data:

1. Train a Random Forest to classify a wine's quality. Select appropriate hyper-parameters for the forest, and evaluate the impact of including class weights.
2. Train an ensemble of SVMs to predict a wine's quality. In doing this you should:
   (a) Train both a one vs one and one vs all model. Comment on differences in performance and training speed.
   (b) Compare the performance of the two models, while also considering the class imbalance in the data. Explore how the error costs can be changed to improve performance for those classes with limited data.

3. Obtain precision, recall and F1 score measures for both classifiers, and compare the results.

In developing your models, consider the impact you observed when standardising the data in Question 1.