Machine Learning Assignment - LLLL76

1 Discussion/details of approaches chosen and experimental procedure

1.1 kNN and Variations

K nearest neighbours was implemented and tested first. The benefits of kNN is that is is very fast to train only needing the time taken to plot all train data. The disadvantages of kNN is the slow query time, this is due to slow look up in high dimensional space. There is only a single value that can be changed in basic kNN to alter its behaviour, the number of neighbours that are used, k. There are a few more advancements on kNN based around the weighting of the neighbours. The weighting cane be done based on the inverse square of the distance between the query and its neighbour or the similarity between the query and its neighbour, both of which have been implemented.

1.2 SVM

Support Vector Machines are based around the idea of a linear separator, that any two classes can be split by a line, and to find the best line there should be a maximum separating margin. The implementation of this approach has a significant number of variables, kernel, degree, C, gamma and the class weighting.

1.3 Experimental Procedure

Different data splits were implemented, manual splits of 60-40, 70-30, 80-20 and 90-10 train to test were tested and shown to be: RESULTS
- k-fold
- k-fold stratified
- Different ratio of test:train
- used sklearn to do grid search for SVM
- did all of knn manually (show all grid search code in sklearn file)

2 Evidence of the performance of your chosen approaches on the data

2.1 Results for each one

results of k 1 - 100 and all weightings of them
results of the best SVM and a couple around it

2.2 Accuracy, Precision, Recall and F-Measure

another big table here
show these for all the best ones of each approach

2.3 Confusion Matrix

lots of confusion matrices
only the best ones

3 Conclusions from the experimentation

3.1 Results

talk about what was best and any limitations of any approach

3.2 Ethics

why is having this data an issue?

4 References

his code
all the sklearn stuff