COMP47750 Tutorial
Ensembles

1.

(a) Load the *Wine* dataset using the CSV file provided, and assess the accuracy of a decision tree classifier using 10-fold cross-validation. What percentage of instances are correctly classified?

(b) Now, apply ensemble classification using *bagging* to achieve diversity and with a decision tree classifier. What percentage of instances are now correctly classified with an ensemble of size 10?

(c) Repeat (b), for ensembles of size 10, 50, 100, 200 and 300 classifiers. What level of improvement does this provide, in terms of percentage of instances correctly classified?

(d) Why does the level of improvement in accuracy often "level off" after an ensemble has been increased to a certain size?

(Note: An example of Bagging in scikit-learn is available in notebook 15 Ensembles.)

2.

(a) Load the Blood Alcohol Content (*BAC)* dataset using the CSV file provided. This dataset contains a mix of numeric and categorical data, use one-hot encoding to convert to a numeric format. When this dataset was collected the BAC limit for driving was 0.8mg/ml. Convert this to a classification task by adding a binary Over/Under feature where Over is a BAC level > 0.8mg/ml.

(b) Using 10-fold cross validation, compare the performance of:

  (a) a single decision tree,

  (b) a bagging ensemble (100 members) and

  (c) a boosting ensemble (also 100 members).

(c) Are the results from a single cross validation run stable?

(d) Repeat the 10-fold cross validation comparison 50 times to get a more robust comparison.

3.

(a) Load the *glass* dataset from *glass.cs*v. Evaluate a 1-NN classifier using 10-fold cross-validation. What is the overall accuracy achieved?

(b) Apply *bagging* with a 1-NN classifier for an ensemble size of 100. What is the improvement in terms of overall accuracy?

(c) Now apply *random subspacing* with a 1-NN classifier for an ensemble size of 100. How does it compare to the results from (b)? How do you explain this difference?

(d) What happens to the overall ensemble accuracy when we increase the *subspace size* to a value closer to 1 (e.g. max_features=0.8)? What is the explanation for the change in accuracy?

(Note: An example of Random Subspacing in scikit-learn is available in notebook 15 Ensembles.)

4.

(a) What does it mean for ensemble members to be diverse?

(b) Why is this diversity important?

(c) Briefly describe two methods to achieve diversity in ensembles.