COMP4702/COMP7703 - Machine Learning

Prac W10 – Ensemble Methods

Aims:

- To explore the performance of bagging and boosting ensemble techniques.
- To produce some assessable work for this subject.

Procedure

In this prac, you will gain some experience at applying bagging and boosting ensemble methods for classification.

Start with the 2D classification dataset from Prac W3 (w3classif.csv). Shuffle and split this dataset into 70% (training) and 30% hold-out.

Question 1: Fit a single Decision Tree classifier to the training data with default parameters, and calculate E_{train} and $E_{hold\text{-}out}$.

Bagging ensemble method

Question 2: Fit a bagging ensemble method to the same training data using the Random Forest algorithm. Calculate E_{train} and $E_{hold\text{-}out}$ for the ensemble.

Question 3: Compare your results from Questions 1 and 2. Are they as expected? You can try varying the depth of the tree (Question 1) and the size of the ensemble in Question 2, but you should NOT look at $E_{hold\text{-}out}$ when you do this (recall from lectures that this would invalidate your results!).

Question 4: In lectures you have heard that bagging can improve performance by reducing the variance component of the error. Estimate the variance of your single decision tree and your bagged ensemble and see if this is the case for this problem.

> [Hint 1: you can generate a much larger version of the dataset for yourself: for each input variable, half of the input data points (for class 0) are drawn from a Normal distribution with mean = 0 and standard deviation = 1, the other half (for class 1) from a Normal distribution with mean = 3 and standard deviation = 1.5.]
> [Hint 2: refer to Section 4.4 of the book/notes, in particular Eqn.4.19. it is only an estimate of the variance, but you can split your large dataset into (e.g 10) equal parts, repeat the model fits. Three levels of averaging will be required (the first one is using the average of the model fits as an estimate of $f(x_*)$!]

Boosting ensemble method

Question 5: Fit a boosting ensemble method to the training data using the AdaBoost algorithm. Calculate E_{train} and $E_{hold\text{-}out}$ for the ensemble.

Bagging vs Boosting

Question 6: Compare the performance of the bagging model with that of the boosting model and interpret the results.

Question 7 (if you have time available): Explore ensemble methods further by:

- Performing hyperparameter tuning for the models
- Use a different base algorithm in your ensembles
- Evaluate the algorithms on a different dataset. The UCI machine learning repository is one source of commonly used benchmark datasets.