Homework 6

AM 218 Machine Learning

Due on May 30 at 11:59 PM on Canvas

Programming assignment

In this assignment, you will continue to work on the spambase dataset (please review Homework 3 for details about this dataset) predict whether the email is spam or not. Since you just learned tree models such as decision tree and ensemble models such as random forests. You would like to evaluate and compare the performance of those algorithms. Specifically, we will implement the decision tree model, the bagging decision trees (bagging + decision tree), the random forest, and the AdaBoost decision tree (AdaBoost + decision tree) on the spambase dataset with different training set sizes, feature space, tree depths, and the number of learners in the ensemble models.

You can either use the functions from libraries such as scikit-learn or write the code from scratch to implement the algorithm. After you run the sample code (see, the demo code "tree_demo.py") successfully (this step is not required, but recommended), you should implement the tree models and the ensemble methods following the instructions as below.

In this homework, the default number of features is 57, the default maximum depth of a decision tree for both tree models and ensemble models is 5, the default number of decision trees (base learners) in those ensemble models is set to 50, and the default minimum number of samples for the leaf node is set to 5, if they are not specified.

1. Explore the effect of training set size (TSS) (15pts)

    a. For each TSS in [0.01, 0.05, 0.1, 0.15, 0.20, 0.25, 0.5]

        i. Split the dataset into 10 folds.

        ii. For each fold, reserve it for testing and sampling TSS*100% samples from the rest of nine folds for training.

iii. Train the decision tree, the bagging decision trees, the random forest, and the AdaBoost decision trees on the selected training samples. Evaluate their accuracy on the validation set.

b. Plot the average accuracy with error bars (one standard error) for each model with different TSS. Discuss the results.

c. Compare the performance of the random forecast and bagging decision trees with different TSS using the paired t-test. Please state the hypotheses and conduct the hypothesis testing to justify your conclusion.

2. Explore the effect of feature space (15pts)

a. For each W in [10, 20, 30, 40, 50, 57]

i. Select the first W features for each sample in the dataset.

ii. Train and evaluate the decision tree, the bagging decision trees, the random forest, and the AdaBoost decision trees on the training samples using 10-fold cross validation on the whole dataset.

b. Plot the average accuracy with error bars (one standard error) for each model under different number of features W. Discuss the results.

c. Compare the performance of the random forecast and bagging decision trees with different W using the paired t-test. Please state the hypotheses and conduct the hypothesis testing to justify your conclusion.

3. Explore the effect of the maximum depth of the decision tree (15pts)

a. For each D in [1, 5, 10, 15, 20]

i. Set the maximum depth of decision tree for both tree models and ensembles models to D.

ii. Train and evaluate the decision tree, the bagging decision trees, the random forest, and the AdaBoost decision trees using the 10-fold cross validation on the whole dataset.

b. Plot the average accuracy with error bars (one standard error) for each model under different maximum depth of tree D. Discuss the results.

c. Compare the performance of the random forecast and bagging decision trees with different D using the paired t-test. Please state the hypotheses and conduct the hypothesis testing to justify your conclusion.

4. Explore the effect of the number of base learners (decision trees) in the ensemble models (15pts)

   a. For each N in [1, 2, 5, 10, 15 ,20, 25, 50]

      i. Set the number of decision trees (base learners) for the ensembles models to N.

      ii. Train and evaluate the decision tree, the bagging decision trees, the random forest, and the AdaBoost decision trees on the training samples using the 10-fold cross validation on the whole dataset.

   b. Plot the average accuracy with error bars(one standard error) for each model under different number of base learners N. Discuss the results.

   c. Compare the performance of the random forecast and bagging decision trees with different N using the paired t-test. Please state the hypotheses and conduct the hypothesis testing to justify your conclusion.

Hint

1. You may use `sklearn.tree.DecisionTreeClassifier()` to train the decision tree model.

2. You may use `sklearn.tree.BaggingClassifier()` to train the bagging decision tree. Note that you need to specify the argument `base_estimator` to make sure the base learner is a decision tree.

3. You may use `sklearn.tree.RandomForestClassifier()` to train the random forest model.

4. You may use `sklearn.tree.AdaBoostClassifier()` to train the AdaBoost decision tree model. Note that you need to specify the argument `base_estimator` to make sure the base learner is a decision tree.

5. You may use `matplotlib.pyplot.errorbar()` to plot the average accuracy of those models with error bars. Note that the argument `yerr` is the standard error of the mean accuracy, which is obtained by dividing the sample standard deviation of accuracy by $\sqrt{k}$ where k is the number of folds.

Submission

1. The source code in python.

   Name your file as "tree.py". You should include the code that you write for the above assignments.

2. Your evaluation & analysis in .pdf format.

   Note that your analysis should include the graphs as well as a discussion of results.