Homework 4: Decision Trees + Analysis

CS412

Due: April 23rd, 11:30pm on Gradescope

1 Trees/Ensemble Methods

For this portion of the assignment, use the digits data and ther two features from the digits data. You will be testing decision trees, random forests and AdaBoost. For all trees, you should use `criterion="entropy"`.

a) Experiment with the value `max leaf nodes` which is the maximum number of nodes that the tree will build, greedy first. Plot a graph of cross validation errors where the x-axis is values of `max leaf nodes` in $\{5, 10, 15, 20, 30, 40, 50, 75, 100, 200, 500, 1000, 10000\}$. Use a logarithmic scale. **Label this Figure 4.1**

b) Was there any evidence of over or underfitting? Which values of `max leaf nodes` exhibited this and explain why you think that is the case.

c) Select the tree from part a) that has the lowest cross validation error. Plot the 2D decision region for this optimal model. **Label this Figure 4.2**

d) **Graduate Student Question:** `min impurity decrease` and `min impurity split` are pre-pruning techniques. Let `max leaf nodes` go to default and find the optimal values of `min impurity decrease` and `min impurity split` in terms of cross-validation error. You may give your results in a graph or a table.

e) Now, experiment with the Random Forest Classifier. For this, you will produce 3 graphs each corresponding to a `max leaf nodes` value of 10,100,1000. For each of these three, plot the cross-validation error with respect to `n estimators` in $\{5, 10, 15, 20, 30, 40, 50, 75, 100, 200, 500, 1000, 10000\}$. **Label these 4.3,4.4 and 4.5**

f) Which values of `max leaf nodes` was most impacted by the bagging approach? Explain why you think this is the case. If there is no significant difference, explain that.

g) To what extent does n estimators impact the result? Was there a point at which more estimators did not make an impact? Identify this point.

h) Plot the decision region for the Random Forest classifier from above which has lowest cross-validation error. **Label this Figure 4.6**

i) **Graduate Student Question:** Experiment with different float values for `max features`. Create a table of cross validation errors for `max leaf nodes` in 10,100,1000, `n estimators` in 1,10,100,100 and `max features` in 0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0.

j) Now you will experiment with the AdaBoost Classifier. For `base estimator = DecisionTreeClassi depth=1)` plot a graph of 10-fold cross-validation errors where the x-axis is `n estimators` in 1,5,10,100,1000,10000. **Label this figure 4.7**.

k) Repeat this for `base estimator = DecisionTreeClassifier(max depth=10)`, **Label this figure 4.8**.

l) Repeat this for `base estimator = DecisionTreeClassifier(max depth=1000)`, **Label this figure 4.9**.

m) Plot the decision boundary of the AdaBoost classifier which has the lowest cross-validation error. **Label this figure 4.10**.

2 Reporting Final Error

Now, we will finally use the testing data we have set aside. From HW2,3 and 4. Select the model with the lowest cross-validation for each of the following.

- Polynomial SVM
- Neural Network
- Random Forest
- AdaBoost

For each of these, train your model on the entire training set and test your data with the entire testing set. Use the 2D data and do not use cross-validation.

Clearly state which model you have selected for each category and give the following upper concentration bounds at the 75%, 95% and 99% confidence intervals. For the chebysev bound, treat the error as the average of n independent Bernoulli trials. Show your work, there are no packages in `sklearn` for this.

- Markov
- Chebyshev
- Hoeffding

a) Which of the following models changed most dramatically with respect to confidence interval?

b) Given this data, which model would you select if presented with this problem by a client? Defend your answer.

c) Are there any considerations other than final reported error that went into your decision? Why or why not?

Extra Credit

Repeat the testing process from section 2 with some other dataset where the model selected is different than the one you would have selected from section 2. *This should encourage you to automate this entire process, which will help you with your final project*