CS 534: Machine Learning

Homework #3

Due: March 21st at 11:59 PM on Canvas

Submission Instructions: The homework should be submitted electronically on Canvas. The code can be done in Python, R, or Matlab with the hints mostly focusing on those who will be programming in Python. Also be sure to attach the code to ensure you'll get partial / full credit if the answer is not what was expected.

1. **(10 points) AdaBoost Update**

   Derive the expression for the update parameter in AdaBoost (Exercise 10.1 in HTF).

   $$\beta_m = \frac{1}{2} \log \frac{1 - \text{err}_m}{\text{err}_m}$$

2. **(5 + 5 + 10 + 5 = 25 pts) Detecting Thyroid Disease with Decision Trees**

   We will be using the hyperthyroid dataset `allhyper.data`, which is a subset of the Thyroid Disease Data Set found at `http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease`. The problem is to determine whether a patient referred to the clinic is hypothyroid. The original problem is not a binary classification and contains negative (not hypothyroid), hyperthyroid, and some subnormal functioning. For the purpose of the homework, you will want use hyperthyroid as the positive class and the others as the negative class. Partition the data into 70%–30% train–test split that you will use for the rest of the homework.

   (a) Build a decision tree using the gini splitting criterion using the default values. Plot the tree.

   (b) Build a decision tree using the entropy criterion using the default values. Plot the tree.

   (c) Explore building different decision trees by altering the maximum depth and minimum number of samples in a leaf for both splitting criteria. What do you notice about the generalization error as you adjust the two parameters?

   (d) Report the misclassification rate, F1 score, and AUC for training and test set for all of the above. How do they compare?

3. **(10 pts) Detecting Thyroid Disease with AdaBoost**

   Build AdaBoost models using decision stumps for the following number of estimators: 5, 10, 50, 100, and 200. Report the misclassification rate, F1 score, and AUC for training and test set for all of the above. How do they compare?

4. **(10 + 10 = 20 pts) Detecting Thyroid Disease with Gradient Boosting**

   (a) Build gradient boosting models with the deviance loss function using the existing gradient boosting classifier in sklearn and varying the number of estimators to be 10, 50, 100, and 200. Report the misclassification rate, F1 score, and AUC for training and test set for all of the above. How do they compare?

   (b) Build gradient boosting models with xgboost and varying the number of estimators to be 5, 10, 50, 100, and 200. Report the misclassification rate, F1 score, and AUC for training and test set for all of the above. How does it compare with the sklearn default gradient boosting algorithm?

5. **(25 points) Detecting Thyroid Disease with Random Forest** Train random forest models varying the splitting criteria and the number of estimators to be 5, 10, 50, 100, and 200. Report the out of bag misclassification rate, misclassification rate, F1 score, and AUC for training and test set for all of the above. How does the out of bag misclassification rate generalize on the test error?

6. **(10 points) Comparison of AdaBoost, Gradient Boosting, and Random Forest**

   (a) How do the relative importance of the features compare across the three different models? Report the table with the rankings and comment on the differences.

   (b) Comment on the performance of the three different models compare in terms of misclassification rate, F1 score, and AUC. Also comment on the computational complexity of the algorithms.