CSE 6363 - *Machine Learning*

Project 2- Spring 2021

Due Date: Apr. 18, 2021

Decision Trees

1. Consider the problem where we want to predict whether we are going to win a game of Tic-Tac-Toe from the current board configuration. To make this decision we have access to the state of the board in the form of 9 attributes reflecting the locations on the board, each one with 3 possible values (x, o, b) representing the two players or blank, respectively. There is a training and a test data set for this problem on Canvas (the datasets are derived from the more expansive UCI machine learning tic-tac-toe data set).

    a) Show the construction of a 2 level decision tree using minimum Entropy as the construction criterion on the training data set. You should include the entropy calculations and the construction decisions for each node you include in the 2-level tree.

    b) Implement a decision tree learner for this particular problem that can derive decision trees with an arbitrary, pre-determined depth (up to the maximum depth where all data sets at the leaves are pure) using the information gain criterion.

    c) Apply the tree from part b) to the test data set for all possible tree depths (i.e. 1 - 9) and compare the classification accuracy on the test set with the one on the training set for each. For which depths does the result indicate overfitting ?

Ensemble Classifiers

2. Using the data and decision tree algorithm from Problem 1, chose a decision tree depth that does not overfit but achieves some baseline classification performance (but at least depth 4) and apply bagging to the problem.

    a) Implement a bagging routine for the decision tree classifier.

    b) Apply bagging 10, 50, and 100 times to the training data. For each of the three cases, evaluate the resulting ensemble classifier on the test data set and compare the error rates for a single classifier of the chosen depth and the three ensemble classifiers. Briefly discuss the results you obtained.

3. Using the data and decision tree algorithm from Problem 1 and the depth chosen for Question 2, apply boosting to the problem.

    a) Implement AdaBoost on top of your decision tree classifier.

    b) Apply boosting 10, 25, and 50 times to the training data. For each of the three cases, evaluate the resulting ensemble classifier on the test data set and compare the error rates for a single classifier with the chosen depth and the three ensemble classifiers. Briefly discuss the results you obtained.