CSE 6363 - *Machine Learning*

Homework/Project 2- Spring 2022

Due Date: Apr. 13 2022, 3:30 pm

Data Sets

This assignment consists a number of implementation and result analysis questions. For these we will again consider the problem and data sets from Questions 2 and 3 in the first assignment where we had height, weight, age, and gender information. Assume the same datasets you generated for the first assignment. Make sure that you use the data you got using your own student ID. Results on data for other student ID numbers will not be considered correct solutions.

Decision Trees

1. Consider the problem from the previous assignments where we want to predict gender from information about height, weight, and age. Here we will use Decision Trees to make this prediction. Note that as the data attributes are continuous numbers yo have to use the $\leq$ attribute and determine a threshold for each node in the tree. As a result you need to solve the information gain for each threshold that is half way between two data points and thus the complexity of the computations increases with the number of data items.

    a) Show the construction steps in the construction of a 2 level decision tree using a single step lookahead search and maximum information gain as the construction criterion. You should include the entropy calculations and the construction decisions for each node you include in the 2-level tree. Since the size of the depth-limited search used in the construction of the tree depends on the training set size, you should limit the data to only the first 3 data items for each gender in the data set your generated for Questions 2 a) and 3a) (the smaller data set for manual work) in Homework 1.

    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) Divide the data set from Question 2c) in Homework 1 (the large training data set) into a training set comprising the first 90 data points and a test set consisting of the last 30 data elements. Use the resulting training set to derive trees of depths 1 - 8 and evaluate the accuracy of the resulting trees for the 90 training samples and for the test set containing the last 30 data items. Compare the classification accuracy on the test set with the one on the training set for each tree depth. 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.