CSE 151: Introduction to Machine Learning Summer 2019, Session 2

Programming Assignment 2

Instructor: Joseph Geumlek **Due on:** Aug 22, 2018

Instructions

- This is a 20 point homework. The assignment should be done individually.
- You are free to use any programming language that you wish.
- The programming assignment should be submitted as a single pdf file to Gradescope. Enter your answers to the questions first, followed by a copy of your code.

Problem 1: Programming Assignment: 20 points

In this problem, we will look at the task of classifying whether a client is likely to default on their credit card payment based on their past behaviour and other characteristics. We will use a decision tree for this purpose.

Download the files `pa2train.txt`, `pa2validation.txt` and `pa2test.txt` from the class website. These are your training, validation and test sets respectively. The files are in ASCII text format, and each line of the file contains a feature vector followed by its label. Each feature vector has 22 coordinates; they are named Feature 1, Feature 2, ..., Feature 22, respectively. The coordinates are separated by spaces. The last (23rd) coordinate represents the label of an example, that is, whether the card-holder defaults on their credit card bill in October, where 1 means yes, and 0 means no.

1. First, build a full ID3 Decision Tree classifier based on the data in `pa2train.txt`. **Do not use pruning.** Draw the first three levels decision tree that you obtain (i.e. root, root's children, root's grandchildren). For each node that you draw, if it is a leaf node, write down the label that will be predicted for this node, as well as how many of the training data points lie in this node. If it is an internal node, write down the splitting rule for the node, as well as how many of the training data points would reach this node. (Hint: If your code is correct, the root node will involve the rule Feature $5 \leq 0.5$.)

   Your code does not need to directly draw the tree (i.e. you may draw it by hand), but we do expect to see code for getting the values you use while drawing the tree.

2. What is the training error and test error of your classifier trained in part (1)? You can find your test data in `pa2test.txt`.

3. Now, prune the decision tree developed in part (1) using the data in `pa2validation.txt`. While selecting nodes to consider for pruning, select them in Breadth-First order (aka, layer-by-layer, starting at the root), going from left to right (aka, from the Yes branches to the No branches). Write down the validation and test error after 1 and 2 rounds of pruning (that is, after you have pruned 1 and 2 nodes from the tree). You may stop after selecting your second node to prune.

4. **Interpreting a model:** Unlike k-NN, decision trees are a little more transparent in how the values of the feature vector affect the prediction. In this problem, you will try to examine what our ID3 tree has learned. Download the file `pa2features.txt` from the class website. This file provides a descriptive name for the 22 coordinates of the feature vectors, presented in the corresponding order. Based on the feature descriptions, what do you think is the most salient or prominent feature the ID3 tree found for predicting credit card defaults? (Hint: More salient features should be used in decisions higher up in the ID3 Decision tree.)