CSE 151: Introduction to Machine Learning Spring 2018

Programming Assignment 2

Instructor: Kamalika Chaudhuri **Due on:**

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 pair of files – a pdf file, containing the answers to the questions, and a zip file containing the code. Please email these two files to cse151homeworks@gmail.com.

Problem 4: 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 folks default on their credit card bill in October, 1 means yes, 0 means no.

1. First, build an 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. 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 lie in this node. (Hint: If your code is correct, the root node will involve the rule Feature $5 < 0.5$.)
2. What is the training and test error of your classifier in part (1), where test error is measured on the data in `pa2test.txt`?
3. Now, prune the decision tree developed in part (1) using the data in `pa2validation.txt`. While selecting nodes to prune, select them in Breadth-First order, 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.)
4. Download the file `pa2features.txt` from the class website. This file provides a description in order of each of the features – that is, it tells you what each coordinate means. Based on the feature descriptions, what do you think is the most salient or prominent feature that predicts credit card default? (Hint: More salient features should occur higher up in the ID3 Decision tree.)

Solutions

1. Figure 1 shows the first three levels of the decision tree. There are multiple possible decision trees even if we ignore the effect of spacing in the training examples. Nevertheless, all of them share the same top 3 layers. In this solution, if multiple splittings achieves maximum information gain, we pick one uniformly at random and recurse.

When splitting each node, we employ Information Gain as the criterion for selecting a (feature, threshold) pair. The set of thresholds for a particular feature to be considered at a node are chosen according to the following approach. First, we sort the training samples S on the feature f being considered. There are only a finite number of these values, so let us denote them in sorted order by $v_1 < v_2 < \cdots < v_n$. Any threshold value lying between v_i and v_{i+1} will have the same effect of dividing the data points associated with the node into those whose value of the feature f lies in $\{v_k : k \leq i\}$ and those whose value is in $\{v_k : k > i\}$. There are thus only $n-1$ possible splits on f. We choose the midpoint of each interval, i.e. $\frac{v_i + v_{i+1}}{2}$, as the representative threshold.

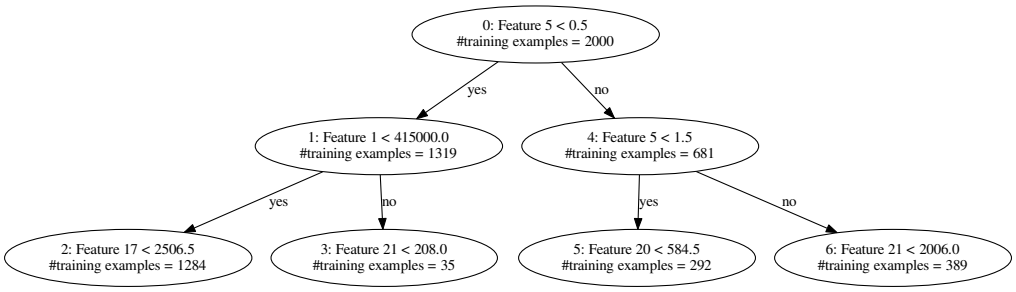

Figure 1: The top 3 layers of the trained ID3 Decision Tree

For the next two problems, depending on the randomness in tie breaking in the decision tree training procedure, we might end up with different trees. The results on the table below is from one of the trees generated by ID3. In general the results might slightly deviate from the result below.

For demonstration purposes, for each node, we also present (1) **"#validation error examples"**, the number of validation examples that passes through this node such that the subtree rooted here classifies them incorrectly, and (2) **"#validation error examples if this is leaf"**, the number of validation examples that passes through this node such that their labels does not agree with the label of this node. If the latter is less than the former, the subtree rooted at this node can be pruned to a leaf.

2. Table 1 shows the errors:

#Prunings	Training Error	Validation Error	Test Error
-	**0**	0.167	**0.175**

Table 1: Training, validation and test errors of decision tree without pruning.

The decision trees generated is shown (again) in Figure 2, with the feature names incorporated.

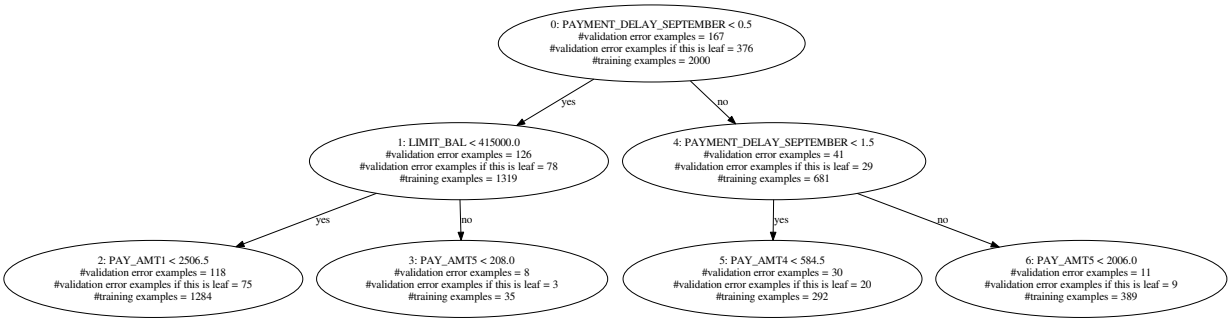

Figure 2: The top 3 layers of the trained decision tree without pruning.

3. The decision tree after pruning 1 node:

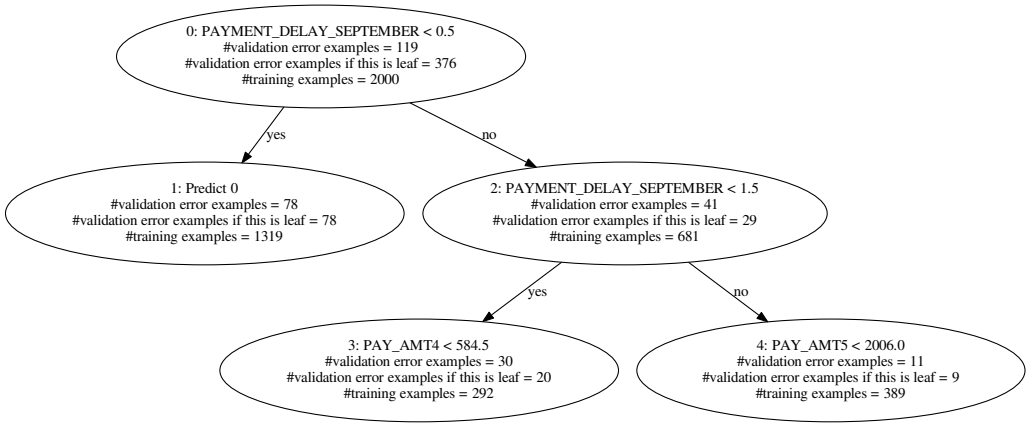

Figure 3: The top 3 layers of the trained decision tree, with 1 node pruned.

The decision tree after pruning 2 nodes:

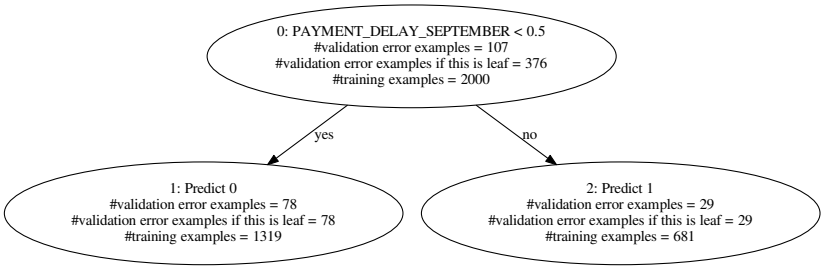

Figure 4: The top 3 layers of the trained decision tree, with 2 nodes pruned.

Table 2 shows the errors.

#Prunings	Training Error	Validation Error	Test Error
1	0.0845	**0.119**	**0.118**
2	0.105	**0.107**	**0.103**

Table 2: Training, validation and test errors of decision tree with pruning.

4. As shown in Figure 2, the most salient feature is PAYMENT_DELAY_SEPTEMBER. Features BILL_AMT1, LIMIT_BAL, PAY_AMT4, PAY_AMT5 are also relevant.