CS 536 : Pruning Decision Trees

16:198:536

The purpose of this problem set is to look at the effect of pruning on decision trees. As before, we need a generative model for data so that we can run repeatable experiments. Let $\{(\underline{X}_1, Y_1), (\underline{X}_2, Y_2), \ldots, (\underline{X}_m, Y_m)\}$ denote a data set, where $\underline{X}_i$ represents a vector of k (binary) feature values, and Y_i is a corresponding binary class or label that we will need to learn to be able to predict from the $\underline{X}$-values.

We generate data via the following scheme, defining a distribution for our data set: Let $\underline{X} = (X_0, X_1, X_2, X_3, \ldots, X_{20})$ be a vector of binary values, satisfying the following

- $X_0 = 1$ with probability 1/2, $X_0 = 0$ with probability 1/2
- For $i = 1, \ldots, 14$, $X_i = X_{i-1}$ with probability 3/4, and $X_i = 1 - X_{i-1}$ with probability 1/4
- For $i = 15, \ldots, 20$, $X_i = 1$ with probability 1/2, $X_i = 0$ with probability 1/2.

The first feature is uniformly random, and the next 14 features are strongly correlated, but the last 5 features are independent of everything else. There are 21 X-variables, so there are $2^{21} \approx 2$ mil possible input $\underline{X}$. Some of these are more likely than others. In general, we expect the training data to cover only a fraction of the total possible inputs, so consider data sets of size m where m ranges from 10 to 10,000. We then define Y to be

$$Y = \begin{cases} \text{majority}(X_1, \ldots, X_7) \text{ if } X_0 = 0 \\ \text{majority}(X_8, \ldots, X_{14}) \text{ if } X_0 = 1. \end{cases} \tag{1}$$

That is, if $X_0 = 0$, we take the majority value of X_1 through X_7 - otherwise we take the majority value of X_8 through X_{14}. The values X_{15} through X_{20} are nothing but noise.

1) Write a function to generate m samples of $(\underline{X}, Y)$, and another to fit a tree to that data using **ID3**. Write a third function to, given a decision tree f, estimate the error rate of that decision tree on the underlying data, $\text{err}(f)$. Do this repeatedly for a range of m values, and plot the 'typical' error of a tree trained on m data points as a function of m. Does this agree with your intuition?

2) Note that X_{15} through X_{20} are completely irrelevant to predicting the value of Y. For a range of m values, repeatedly generate data sets of that size and fit trees to that data, and estimate the average number of irrelevant variables that are included in the fit tree. How much data would you need, typically, to avoid fitting on this noise?

3) Generate a data set of size $m = 10000$, and set aside 8000 points for training, and 2000 points for testing. The remaining questions should all be applied to this data set.

   a) **Pruning by Depth:** Consider growing a tree as a process - running ID3 for instance until all splits up to depth d have been performed. Depth $d = 0$ should correspond to no decisions - a prediction for Y is made just on the raw frequencies of Y in the data. Plot, as a function of d, the error on the training set and the error on the test set for a tree grown to depth d. What does your data suggest as a good threshold depth?

   b) **Pruning by Sample Size:** The less data a split is performed on, the less 'accurate' we expect the result of that split to be. Let s be a threshold such that if the data available at a node in your decision tree is less than or equal to s, you do not split and instead decide Y by simple majority vote (ties broken by coin flip). Plot, as a function of s, the error on the training set and the error on the testing set for a tree split down to sample size s. What does your data suggest as a good sample size threshold?

c) **Pruning by Significance:** If a variable X is independent of Y, then X has no value as a splitting variable. We can use something like the χ^2-test to estimate how likely a potential splitting variable is to be independent, based on the test statistic T compared to some threshold T_0 (in the usual 2-outcome case, $T_0 = 3.841$ is used to test at a significance level of $p = 5\%$ - see notes for more explanation). Given T_0, if given the data for X the value of T is less than T_0, it is deemed not significant and is not used for splitting. If given the data for X the value of T is greater than T_0, it is deemed significant, and used for splitting. Plot, as a function of T_0, the error on the training set and the error on the testing set for a tree split at significance threshold T_0. What does your data suggest as a good threshold for significance?

5) Repeat the computation of Problem 2, growing your trees only to depth d as chosen in 3.a. How does this change the likelihood or frequency of including spurious variables in your trees?

6) Repeat the computation of Problem 2, splitting your trees only to sample size s as chosen in 3.b. How does this change the likelihood or frequency of including spurious variables in your trees?

7) Repeat the computation of Problem 2, splitting your trees only at or above threshold level T_0 as chosen in 3.c. How does this change the likelihood or frequency of including spurious variables in your trees?