CS 536 : Decision Trees

16:198:536

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 $\underline{X} = (X_1, X_2, X_3, \ldots, X_k)$ be a vector of binary values, satisfying the following

- $X_1 = 1$ with probability $1/2$, $X_1 = 0$ with probability $1/2$
- For $i = 2, \ldots, k$, $X_i = X_{i-1}$ with probability $3/4$, and $X_i = 1 - X_{i-1}$ with probability $1/4$.

In this way, the first feature value is uniformly random, but every successive feature is strongly correlated with the value of the feature before it. We can then define Y to be a function of $\underline{X}$ as

$$Y = \begin{cases} X_1 \text{ if } w_2 X_2 + w_3 X_3 + \ldots + w_k X_k \geq 1/2 \\ 1 - X_1 \text{ else.} \end{cases} \tag{1}$$

In other words, if the 'weighted average' of $X_2, \ldots X_k$ tilts high, Y will agree with X_1; if the weighted average of $X_2, \ldots, X_k$ tilts low, Y will disagree with X_1. Take the weights to be defined by $w_i = 0.9^i/(0.9^2 + 0.9^3 + \ldots + 0.9^k)$.

1) For a given value of k, m, (number of features, number of data points), write a function to generate a training data set based on the above scheme.

2) Given a data set, write a function to fit a decision tree to that data based on splitting the variables by maximizing the information gain. Additionally, return the training error of this tree on the data set, $\text{err}_{\text{train}}(\hat{f})$. *It may be useful to have a function that takes a data set and a variable, and returns the data set partitioned based on the values of that variable.*

3) For $k = 4$ and $m = 30$, generate data and fit a decision tree to it. Does the ordering of the variables in the decision tree make sense, based on the function that defines Y? Why or why not? Draw the tree.

4) Write a function that takes a decision tree and estimates its typical error on this data $\text{err}(\hat{f})$; i.e., generate a lot of data according to the above scheme, and find the average error rate of this tree over that data.

5) For $k = 10$, estimate the value of $|\text{err}_{\text{train}}(\hat{f}) - \text{err}(\hat{f})|$ for a given m by repeatedly generating data sets, fitting trees to those data sets, and estimating the true and training error. Do this for multiple m, and graph this difference as a function of m. What can you say about the marginal value of additional training data?

6) Design an alternative metric for splitting the data, not based on information content / information gain. Repeat the computation from (5) above for your metric, and compare the performance of your trees vs the ID3 trees.