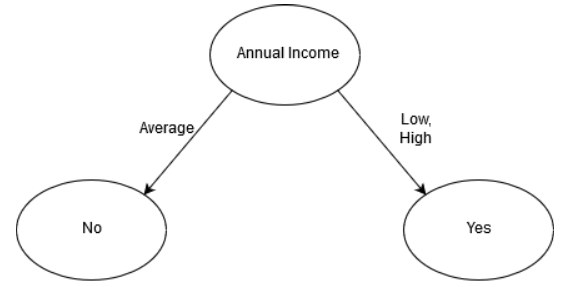
Richard Vargas

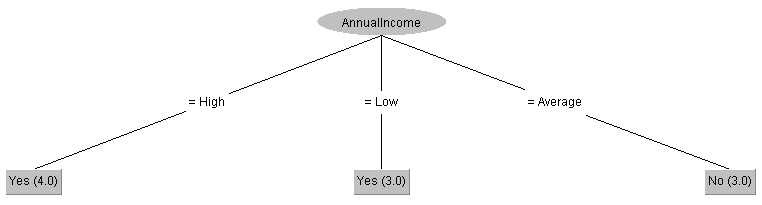
CSE 516 Machine Learning

Dr Zhang

1. a) The first step is to look at our features from our label data. I have chosen Annual Income as the root of the tree. From the children of this node will be the next features. This process continues until all features have been exhausted, and correct labels have been posted. The Gini of Annual Income is represented by the following equation (as parent node),  
   ,

The children of this node will be leaves of that are either yes or no decisions, both with

 b)

 c)

2.

b)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | M=2 | | | M=10 | | | M=50 | | | M=100 | | |
|  | CR | TS | NL | CR | TS | NL | CR | TS | NL | CR | TS | NL |
| C=0.40 | 97.8048 | 269 | 135 | 95.2619 | 117 | 59 | 91.0237 | 23 | 12 | 89.0459 | 11 | 6 |
| C=0.25 | 97.1745 | 207 | 104 | 95.0228 | 99 | 50 | 91.0237 | 23 | 12 | 89.0459 | 11 | 6 |
| C=0.05 | 95.3054 | 113 | 57 | 94.3273 | 75 | 38 | 90.8715 | 19 | 10 | 89.0459 | 11 | 6 |

As we increase M, the number of minimum objects per leaf, we see a decrease across the board for number of leaves and tree size. C, the confidence factor, has an effect on the pruning process. Where the lower the confidence factor the smaller the tree and smaller amount of leaves. This process of reducing the amount of nodes and leaves is called pruning.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | M=2 | | | M=10 | | | M=50 | | | M=100 | | |
|  | CR | TS | NL | CR | TS | NL | CR | TS | NL | CR | TS | NL |
| C=0.40 | 92.6972 | 269 | 135 | 92.1321 | 117 | 59 | 90.263 | 23 | 12 | 88.5242 | 11 | 6 |
| C=0.25 | 92.9798 | 207 | 104 | 92.1104 | 99 | 50 | 90.1543 | 23 | 12 | 88.546 | 11 | 6 |
| C=0.05 | 92.5016 | 113 | 57 | 91.9583 | 75 | 38 | 90.0891 | 19 | 10 | 88.5894 | 11 | 6 |

It appears the the correctly classified instance rate is higher in using the training set rather than using the cross-validation. So it would appear using the training set is a better method of determining the correct tree.