IS467: Assignment 3

Jasmine Dumas

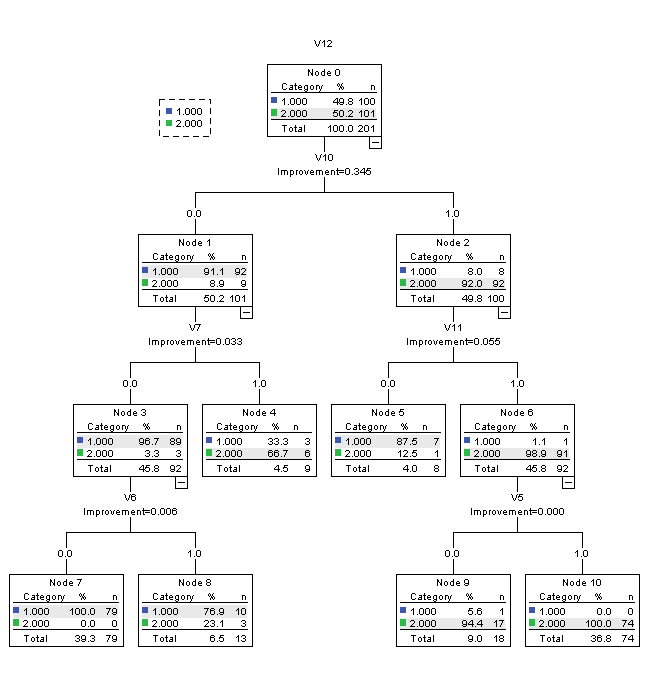
February 12, 2016

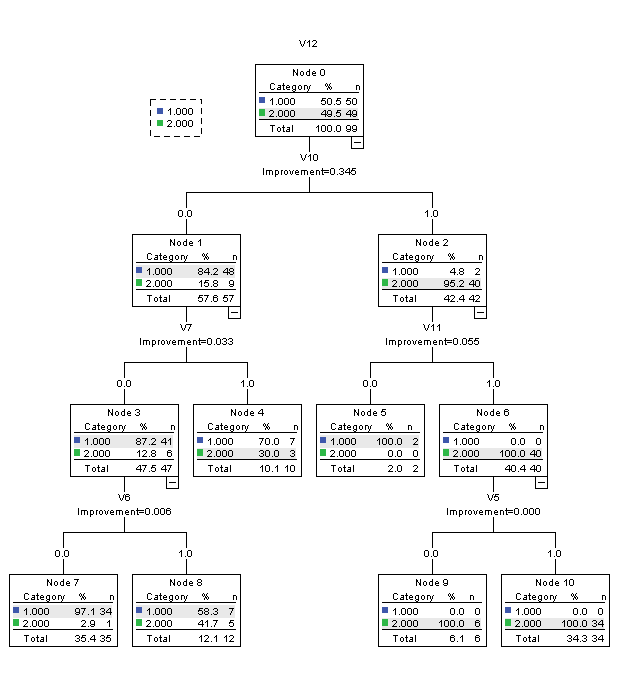
### Problem 1 (20 points): This problem illustrates the classification approach by using decision trees and the Lupus data (you can download the data file “sledata” from D2L site, course documents for week 5). The data consists of 300 patient records. Each record contains 12 elements. The first 11 elements stand for different symptoms and the final element of each record indicates the diagnosis. Build a decision tree and report:

1. The decision tree and the criteria used for building the tree for deciding the best split and the stopping condition (such as which impurity measure, how many cases for parents and children per node, etc): **The decision tree was build with a maximum depth of 20; Minimum cases in parent node of 10; Minimum cases in Child node of 5 with a growing method of CRT (CRT growing method attempts to maximize within-node homogeneity) and the default impurity index of GINI.**

|  |  |  |
| --- | --- | --- |
| **Model Summary** | | |
| Specifications | Growing Method | CRT |
| Dependent Variable | V12 |
| Independent Variables | V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V11 |
| Validation | Split Sample |
| Maximum Tree Depth | 20 |
| Minimum Cases in Parent Node | 10 |
| Minimum Cases in Child Node | 5 |
| Results | Independent Variables Included | V10, V7, V6, V1, V9, V11, V3, V5, V4, V8, V2 |
| Number of Nodes | 11 |
| Number of Terminal Nodes | 6 |
| Depth | 3 |

**Training Sample**



**Test Sample**

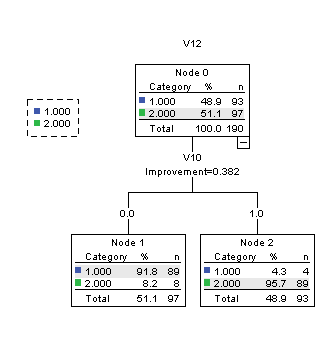
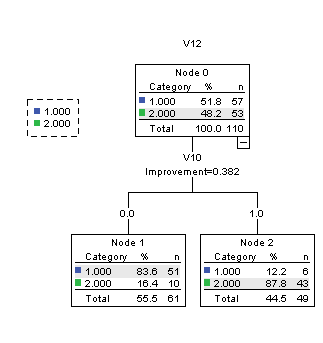
|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .040 | .014 |
| Test | .131 | .034 |
| Growing Method: CRT  Dependent Variable: V12 | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | | | | |
| Sample | Observed | Predicted | | |
| 1 | 2 | Percent Correct |
| Training | 1 | 96 | 4 | 96.0% |
| 2 | 4 | 97 | 96.0% |
| Overall Percentage | 49.8% | 50.2% | 96.0% |
| Test | 1 | 43 | 7 | 86.0% |
| 2 | 6 | 43 | 87.8% |
| Overall Percentage | 49.5% | 50.5% | 86.9% |
| Growing Method: CRT  Dependent Variable: V12 | | | | |

1. How many nodes the final tree has and how many of them are terminal nodes: **The final tree has 11 (including the Root node) nodes with 6 of those being terminal nodes.**
2. What are the most important three Lupus data features in building the tree? Explain your answer: **The most important features in building the lupus tree are V10, V7, & V11 (symptoms 7, 10, 11) as these are at the top nodes of the decision tree since a decision tree implicitly performs variable selection.**
3. Increase the number of cases for each parent and child. What do you notice with the complexity (number of nodes) of the tree? Does it increase? Explain your answer. **The complexity of the decision tree decreased (less nodes) because increasing the cases creates wider bins to fit more data in & essentially partitioning the data less specifically.**

|  |  |  |
| --- | --- | --- |
| **Model Summary** | | |
| Specifications | Growing Method | CRT |
| Dependent Variable | V12 |
| Independent Variables | V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V11 |
| Validation | Split Sample |
| Maximum Tree Depth | 30 |
| Minimum Cases in Parent Node | 20 |
| Minimum Cases in Child Node | 10 |
| Results | Independent Variables Included | V10, V7, V9, V11, V1, V6, V3, V8, V5, V2, V4 |
| Number of Nodes | 3 |
| Number of Terminal Nodes | 2 |
| Depth | 1 |

**Training Sample Test Sample**



|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .063 | .018 |
| Test | .145 | .034 |
| Growing Method: CRT  Dependent Variable: V12 | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | | | | |
| Sample | Observed | Predicted | | |
| 1 | 2 | Percent Correct |
| Training | 1 | 89 | 4 | 95.7% |
| 2 | 8 | 89 | 91.8% |
| Overall Percentage | 51.1% | 48.9% | 93.7% |
| Test | 1 | 51 | 6 | 89.5% |
| 2 | 10 | 43 | 81.1% |
| Overall Percentage | 55.5% | 44.5% | 85.5% |
| Growing Method: CRT  Dependent Variable: V12 | | | | |

Problem 2 (30 points): This problem illustrates the effect of the class imbalance of the accuracy of the decision trees. Download the red wine quality data from the UCI machine learning repository at: <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

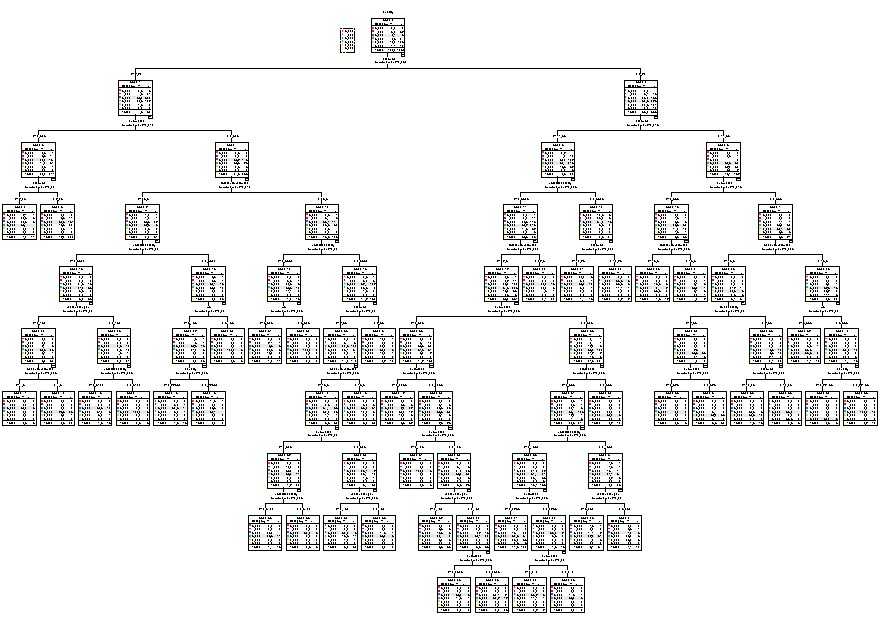
1. Report how many classes (treat each quality level as a different class) are and what is the distribution of these classes for the red wine data is: **There are 6 classes within the quality variable and the distribution is below:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **quality** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 3 | 10 | .6 | .6 | .6 |
| 4 | 53 | 3.3 | 3.3 | 3.9 |
| 5 | 681 | 42.6 | 42.6 | 46.5 |
| 6 | 638 | 39.9 | 39.9 | 86.4 |
| 7 | 199 | 12.4 | 12.4 | 98.9 |
| 8 | 18 | 1.1 | 1.1 | 100.0 |
| Total | 1599 | 100.0 | 100.0 |  |

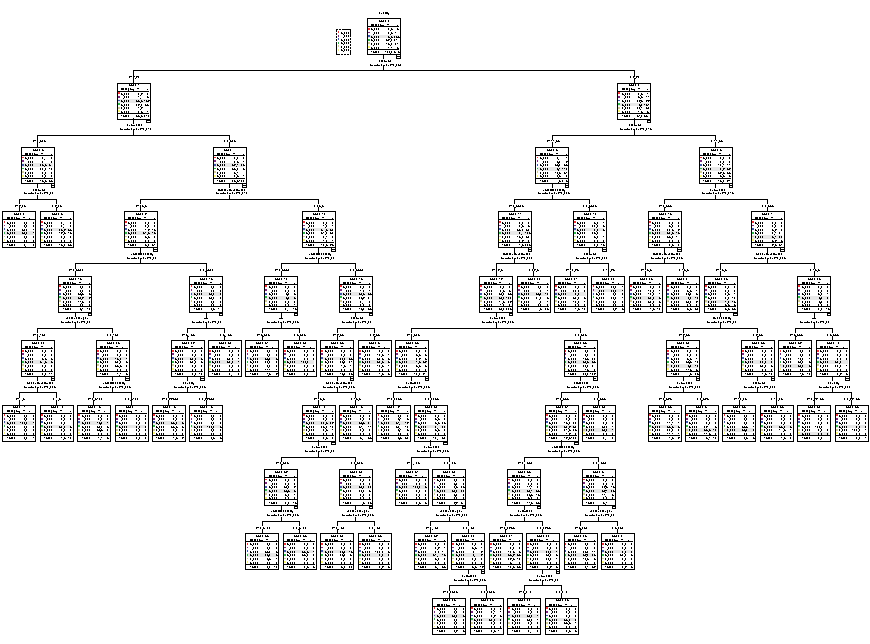
1. Repeat **Problem 1** on the red wine data.

|  |  |  |
| --- | --- | --- |
| **Model Summary** | | |
| Specifications | Growing Method | CRT |
| Dependent Variable | quality |
| Independent Variables | fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol |
| Validation | Split Sample |
| Maximum Tree Depth | 20 |
| Minimum Cases in Parent Node | 10 |
| Minimum Cases in Child Node | 5 |
| Results | Independent Variables Included | alcohol, density, totalsulfurdioxide, sulphates, volatileacidity, chlorides, citricacid, pH, residualsugar, freesulfurdioxide, fixedacidity |
| Number of Nodes | 79 |
| Number of Terminal Nodes | 40 |
| Depth | 9 |

**Training Sample**



**Test Sample**



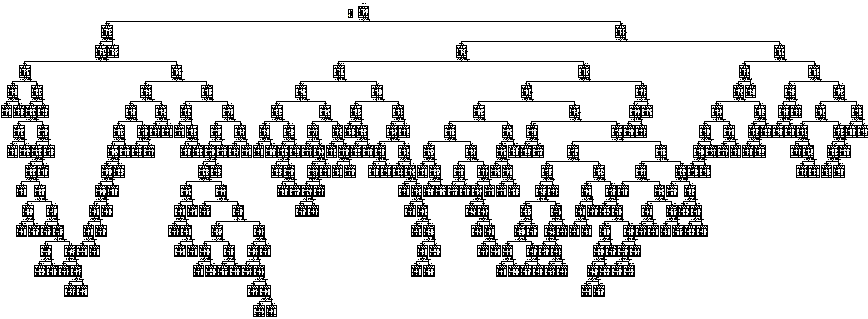
|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .276 | .014 |
| Test | .421 | .021 |
| Growing Method: CRT  Dependent Variable: quality | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classification** | | | | | | | | |
| Sample | Observed | Predicted | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | Percent Correct |
| Training | 3 | 0 | 0 | 3 | 4 | 0 | 0 | 0.0% |
| 4 | 0 | 5 | 25 | 9 | 0 | 0 | 12.8% |
| 5 | 0 | 3 | 374 | 65 | 1 | 0 | 84.4% |
| 6 | 0 | 1 | 85 | 322 | 14 | 2 | 75.9% |
| 7 | 0 | 0 | 15 | 55 | 58 | 0 | 45.3% |
| 8 | 0 | 0 | 0 | 7 | 2 | 3 | 25.0% |
| Overall Percentage | 0.0% | 0.9% | 47.7% | 43.9% | 7.1% | 0.5% | 72.4% |
| Test | 3 | 0 | 1 | 2 | 0 | 0 | 0 | 0.0% |
| 4 | 0 | 1 | 8 | 5 | 0 | 0 | 7.1% |
| 5 | 0 | 2 | 166 | 69 | 1 | 0 | 69.7% |
| 6 | 0 | 1 | 69 | 130 | 10 | 4 | 60.7% |
| 7 | 0 | 0 | 7 | 45 | 19 | 0 | 26.8% |
| 8 | 0 | 0 | 1 | 3 | 2 | 0 | 0.0% |
| Overall Percentage | 0.0% | 0.9% | 46.3% | 46.2% | 5.9% | 0.7% | 57.9% |
| Growing Method: CRT  Dependent Variable: quality | | | | | | | | |

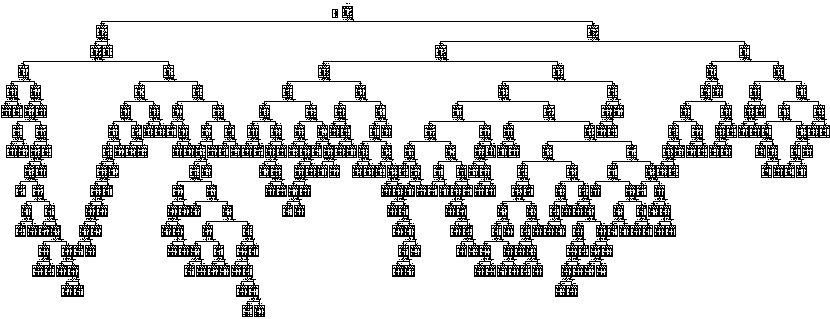
1. Now bin the class variable in such a way that data is not so imbalanced with respect to the class variable. Repeat **Problem 1** but on the wine data with less number of classes (the binned class variable).

|  |  |  |
| --- | --- | --- |
| **Model Summary** | | |
| Specifications | Growing Method | CRT |
| Dependent Variable | quality |
| Independent Variables | fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol |
| Validation | Split Sample |
| Maximum Tree Depth | 20 |
| Minimum Cases in Parent Node | 5 |
| Minimum Cases in Child Node | 2 |
| Results | Independent Variables Included | alcohol, totalsulfurdioxide, density, sulphates, chlorides, volatileacidity, pH, freesulfurdioxide, residualsugar, citricacid, fixedacidity |
| Number of Nodes | 289 |
| Number of Terminal Nodes | 145 |
| Depth | 15 |

**Training Sample**



**Test Sample**



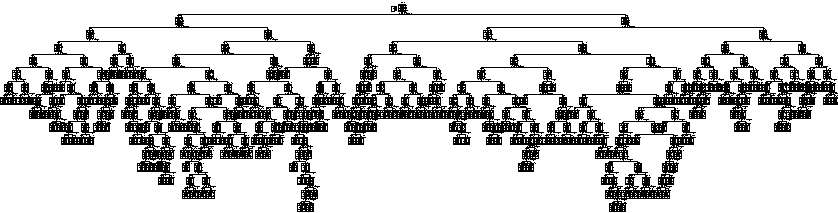
|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .118 | .010 |
| Test | .389 | .022 |
| Growing Method: CRT  Dependent Variable: quality | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classification** | | | | | | | | |
| Sample | Observed | Predicted | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | Percent Correct |
| Training | 3 | 2 | 2 | 1 | 3 | 0 | 0 | 25.0% |
| 4 | 1 | 7 | 11 | 9 | 3 | 0 | 22.6% |
| 5 | 0 | 0 | 449 | 23 | 6 | 0 | 93.9% |
| 6 | 1 | 1 | 26 | 388 | 7 | 0 | 91.7% |
| 7 | 0 | 0 | 4 | 23 | 116 | 0 | 81.1% |
| 8 | 0 | 0 | 2 | 3 | 3 | 4 | 33.3% |
| Overall Percentage | 0.4% | 0.9% | 45.0% | 41.0% | 12.3% | 0.4% | 88.2% |
| Test | 3 | 0 | 0 | 1 | 0 | 1 | 0 | 0.0% |
| 4 | 2 | 2 | 11 | 4 | 3 | 0 | 9.1% |
| 5 | 2 | 2 | 145 | 48 | 6 | 0 | 71.4% |
| 6 | 0 | 0 | 59 | 131 | 24 | 1 | 60.9% |
| 7 | 0 | 0 | 7 | 17 | 30 | 2 | 53.6% |
| 8 | 0 | 0 | 0 | 3 | 3 | 0 | 0.0% |
| Overall Percentage | 0.8% | 0.8% | 44.2% | 40.3% | 13.3% | 0.6% | 61.1% |
| Growing Method: CRT  Dependent Variable: quality | | | | | | | | |

1. How the performance of the best classification model on the original class variable compares with the accuracy of the best classification model on the binned classification variable? **The binned (reduced parent/child nodes) resulted in an increase in accuracy for prediction/classification of the quality class as seen in the classification table above.**
2. Do you have any other ideas on how you can improve the results further?

Showing that your idea will actually work will be graded with five extra credit points.: **I could improve the results by creating a balanced dataset to pick the train/test values from as to remove any bias from not having a equal representation of each class level of the target variable. To create the “balanced” dataset I re-ran the improved binned classification tree on the entire dataset without validation (which random selects a 66/34%) which includes a better distribution of the wine quality rating. The overall percentage is 88.4%, an 0.2% increase from above on training and 27.3% increase from above for the test tree. Another option would be to create special clusters that represented an equal distribution of the quality ratings before running the decision tree.**

|  |  |  |
| --- | --- | --- |
| **Model Summary** | | |
| Specifications | Growing Method | CRT |
| Dependent Variable | quality |
| Independent Variables | fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol |
| Validation | None |
| Maximum Tree Depth | 20 |
| Minimum Cases in Parent Node | 5 |
| Minimum Cases in Child Node | 2 |
| Results | Independent Variables Included | alcohol, density, chlorides, volatileacidity, citricacid, sulphates, totalsulfurdioxide, fixedacidity, pH, residualsugar, freesulfurdioxide |
| Number of Nodes | 391 |
| Number of Terminal Nodes | 196 |
| Depth | 15 |



|  |  |
| --- | --- |
| **Risk** | |
| Estimate | Std. Error |
| .116 | .008 |
| Growing Method: CRT  Dependent Variable: quality | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Classification** | | | | | | | |
| Observed | Predicted | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | Percent Correct |
| 3 | 2 | 0 | 5 | 2 | 1 | 0 | 20.0% |
| 4 | 0 | 22 | 17 | 14 | 0 | 0 | 41.5% |
| 5 | 0 | 2 | 640 | 34 | 5 | 0 | 94.0% |
| 6 | 0 | 1 | 51 | 570 | 14 | 2 | 89.3% |
| 7 | 0 | 0 | 7 | 23 | 168 | 1 | 84.4% |
| 8 | 0 | 0 | 0 | 3 | 3 | 12 | 66.7% |
| Overall Percentage | 0.1% | 1.6% | 45.0% | 40.4% | 11.9% | 0.9% | 88.4% |
| Growing Method: CRT  Dependent Variable: quality | | | | | | | |

### Problem 3 (5 points): Differentiate between the following terms:

1. feature selection and feature extraction: **Feature selection is selecting a subset of the the variables from the original while feature extraction is the transformation of the original variables into a new set of variables.**
2. training and testing: **Training (data) is a partition of the data to which a model is formulated on and refined and the testing (data) is the the data you run your completed model on for verification/validation of a valid predictive model.**
3. parametric reduction techniques and non-parametric reduction techniques: **Parametric reduction techniques include: N(mean, standard deviation) and multiple linear regression which quantify boundaries of data exclusion (reduction) and assume a certain model form (and depends on it) where as non-parametric reduction techniques include: histograms, clustering, sampling which does not assume any model initially and is more effective uniformly.**
4. uniform binning and non-uniform binning: **Uniform binning is when there is equal amount of observations in each bin (therefore information can be lost) as opposed to non-uniform amounts of observations in each bin (or creating thresholds) where the distribution can be seen clearer.**
5. covariance matrix and correlation matrix: **Correlation is just the normalized covariance. Sample variance measures the average of the square differences from the sample mean. ( E[ x – xbar)2]), covariance does that for two different variables rather than x with itself ( E[ x – xbar)( y - ybar]).**