**ASSIGNMENT 3**

**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.)

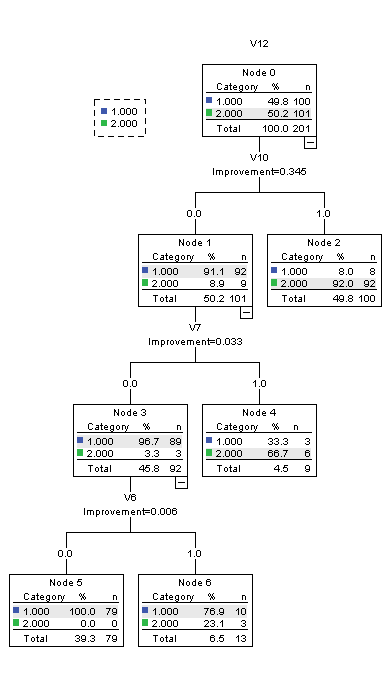
* Growing Method: CRT
* Validation: Split sample
* Impurity measure: Gini
* Maximum tree depth: 5
* Minimum cases in parent node: 18
* Minimum cases in child node: 9

1. How many nodes the final tree has and how many of them are terminal nodes;

|  |  |
| --- | --- |
| Min cases in Parent node | 18 |
| Min cases in child node | 9 |
| Number of nodes | 7 |
| Number of terminal nodes | 4 |
| Depth | 3 |
| Accuracy | 93% |

1. What are the most important three Lupus data features in building the tree? Explain your answer.

The most important three feature can be derived from viewing the decision tree. It can clearly observed from the below decision tree model, V10, V7 and V6 are the most important 3 features.



1. 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | Accuracy Training set | Accuracy Test set | # of nodes | # of terminal nodes | Depth |
| Np = 10 | Nc = 5 | 96% | 87% | 11 | 6 | 3 |
| Np = 12 | Nc = 6 | 95% | 88% | 7 | 4 | 3 |
| Np = 18 | Nc = 9 | 92% | 93% | 7 | 4 | 3 |
| Np = 22 | Nc = 11 | 91% | 92% | 5 | 3 | 2 |

As it could be seen in the table, with the increase in the number of cases in parent and child there is decrease in the tree complexity.

**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 datafrom 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |  |

There are six classes based on quality levels of the red wine. There are large number of cases in class 5 and class 6. Average in case of class 7 and very little data is falls under class 3, class 4 and class 8.

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

* Growing Method: CRT
* Validation: Split sample
* Impurity measure: Gini
* Maximum tree depth: 5
* Accuracy: 60.3%
* Minimum cases in parent node: 40
* Minimum cases in child node: 20
* Number of nodes: 21
* Number of terminal nodes: 11
* Depth: 5

The three important features are alcohol, density and chlorides.

By increasing the number of cases for every parent and child, the accuracy of the tree will be decreased.

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).

Class 3 and Class4 are binned into new value 🡺 1

Class 5 🡺 2

Class 6 🡺 3

Class 7 and Class 8 🡺 4.

The frequency distribution for the binned classes are shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Quality\_binnedLabel** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 1.00 | 63 | 3.9 | 3.9 | 3.9 |
| 2.00 | 681 | 42.6 | 42.6 | 46.5 |
| 3.00 | 638 | 39.9 | 39.9 | 86.4 |
| 4.00 | 217 | 13.6 | 13.6 | 100.0 |
| Total | 1599 | 100.0 | 100.0 |  |

* Growing Method: CRT
* Validation: Split sample
* Impurity measure: Gini
* Maximum tree depth: 5
* Minimum cases in parent node: 40
* Minimum cases in child node: 20
* Number of nodes: 31
* Number of terminal nodes: 16
* Depth: 5
* Accuracy: 64.7%
* The three important features are alcohol, density and chlorides.
* By increasing the number of cases for every parent and child, the accuracy of the tree will be decreased.

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?

With changes in the number of cases in parent/ child node, several models are derived on the original class variable. The best model has an accuracy of 65.8%

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min cases in Parent node | Min cases in child node | # of nodes | # of terminal nodes | Depth | Accuracy |
| 20 | 10 | 35 | 18 | 5 | 65.8% |
| 30 | 15 | 29 | 15 | 5 | 63.8% |
| 40 | 20 | 21 | 11 | 5 | 60.3% |
| 50 | 25 | 19 | 10 | 4 | 62% |

With changes in the number of cases in parent/ child node, several models are derived on the binned class variable. The best model has an accuracy of 67.7%

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min cases in Parent node | Min cases in child node | # of nodes | # of terminal nodes | Depth | Accuracy |
| 20 | 10 | 33 | 17 | 5 | 67.7% |
| 30 | 15 | 29 | 15 | 5 | 65.5% |
| 40 | 20 | 31 | 16 | 5 | 64.7% |
| 50 | 25 | 25 | 13 | 5 | 63.4% |

Binning makes the classes more balanced. It can observed from the tables the accuracy of the model derived on binned class variable is higher than the original class variable, indicates it has less complexity.

1. 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.

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

1. **Feature Selection and Feature Extraction**

**Feature Selection:** Selecting a subset of the existing features/ original feature set without a transformation.

**Feature Extraction:** Transforming the existing features to build a new set of features of lower dimensional space. It is not reversible because some information is lost in process of dimensionality reduction.

1. **Training and Testing**

In the Model Construction process, we use the input data set and expected output for every row to define the classification algorithms. This data set is called training data. The data set we are going to apply the model and test the outcome is called testing data.

1. **Parametric reduction techniques and non-parametric reduction techniques**

**A parametric data reduction technique** is a data reduction technique that assumes a certain model for the data. The model contains some parameters and the technique fits the data into the model to determine the parameters. Then data reduction can be performed.

**A nonparametric data reduction technique** is a data reduction technique that does not assume any model for the data and is applied to the data directly. It yields more uniform effectiveness irrespective of the data, but may not achieve as high data reduction as a well-suited parametric data reduction technique.

1. **Uniform binning and non-uniform binning**

Uniform binning means that all bins have the same property such as equal number of elements or same width for the intervals.

Non-uniform binning is when the bins are not the same. An example is produced through clustering which will create clusters/bins of not necessary the same size.

1. **Covariance matrix and correlation matrix**

A correlation matrix is used to investigate the dependence between multiple variables at the same time. The result is a table containing the correlation coefficients between each variable and the others. The diagonal elements of the matrix are 1.

A covariance matrix is a square matrix that contains the variances and covariance’s associated with several variables. The diagonal elements of the matrix contain the variances of the variables and the off-diagonal elements contain the covariance’s between all possible pairs of variables.

Covariance is the expected value of variations of two random variates from their expected values whereas correlation doesn’t include variations

Covariance involves the relationship of two variables or data set while correlation involve the relationship of multiple variables at the same time.

Correlation values range from [1, -1] whereas covariance value can exceed this range.