**NIT6160 Assignment 2: Data Warehousing and Mining**

## ***Assignment 2***

This project is worth 20% of the total assessment of this unit and is due on Friday *week 12* 5 PM.

The goal of this project is to applying association rule mining, classification and clustering methods on the Mushroom and groceries data sets. For detailed information about the mush room data set, refer to the [Machine Learning Repository](http://archive.ics.uci.edu/ml/datasets/) provided by the University of California, Irvine. You can download and read more about the data there.

**The groceries Dataset**

Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer’s basket. That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a transaction and each column in a row represents an item.

#### Task 1: Data Pre-processing

Read the data in R. There are many ways to read in csv tables in R. For more details, please refer to data import/export in R

<https://cran.r-project.org/doc/manuals/r-release/R-data.pdf>

For the clustering experiments, the column for class labels need to be removed. Refer to lecture Module 10 to see how to do so.

Verify if any other pre-processing is beneficial for the analysis. For example, replacing missing values, attribute range normalization, converting numerical or string to nominal values etc.

#### Task 2: Data Mining

* **Association Rule Mining** experiments: Using R to explorer "association rules" on the groceries dataset. Try out different algorithms. Visualize the result you found. Report any interesting association rules discovered in the experiments and explain why they are interesting.
* **Classification** experiments: Using to construct classifiers on the mushroom dataset. Randomly split the data set in the training and test data set (80% v.s. 20%). Select at least one classifier from each of the following two categories of classifiers: Tree-based models, Bayes classifiers, and Rule-based classifiers. Compare the result of the chosen classifiers.
* **Clustering** experiments: Using R explorer clusters on the mushroom dataset. Select and compare two clustering algorithms from R (e.g. k-means v.s. density-based). Use R to visually explore the resulting clusters.
* For all the above experimentations, try different parameter settings to fine tune the outcome. In principle select methods that work well on the given data set.

#### Task 3: Prepare a report

Your report should contain the following:

* **Theoretical Discussion:** Limited to two pages discussing about data preprocessing steps, the motivation for selecting a particular method, and how the parameters are chosen.
* **Results:** Include results and screenshots of the above experimentations.
* **Discussion and error analysis:** Try to interpret the results of your model. Discuss intuitions or hypothesis that can be obtained by visual inspections of the resulting classes or clusters. Mention about assumptions if any, discuss issues that might have affected the model's performance.
* **References:** If you are using information from other sources apart from R manual and official website, you should cite them.

### Submission Instructions

This section is intended for submission instructions in learning systems.

### Grading

|  |  |
| --- | --- |
| **Report Section** | **Max. points** |
| Theoretical discussion and data-preprocessing | 5% |
| Results | 10% |
| Error analysis & references | 5% |
| **Total** | **20%** |