University of Wollongong

School of Computing and Information Technology

CSCI946 **Big Data Analytics** **Autumn 2020**

Assignment 1 **(Due: 23:59, 26th October 2020, Beijing Time)** **10 marks**

Aim

This assignment is intended to provide basic experience in conducting data analytics experiments with R language. After having completed this assignment you should know how to load and save data and workspace; conduct hypothesis testing; perform clustering; and generate association rules.

Preliminaries

Read through the lecture notes and recommended readings on hypothesis testing, clustering and association rules. Study all example programs therein so that you fully understand these techniques and know how to perform them with R.

Task 1 – Hypothesis Testing

To improve student learning performance, a teacher developed two new learning approaches, called "approach1" and "approach2" in short. To analyze the effectiveness of these approaches, the teacher randomly selected N students. For N_1 of them, he applied "approach1" and for N_2 of them, he applied "approach2". For the rest ($N- N_1- N_2$) students, he applied nothing. After a period of time, the teacher conducted a test on all the N students and evaluated the performance of each student with a performance score. The evaluation result is stored in "A1_performance_test.csv", which is provided with this assignment. In this task, you will use hypothesis testing to help this teacher to answer the following questions:

1. Whether the two new learning approaches can effectively improve student learning performance?
2. In terms of improving student learning performance, whether the two approaches are significantly different from each other?

In your report, you need to

1. State the null hypothesis and alternative hypothesis, and report the key parameters you set for the testing.
2. Answer the above two questions and carefully justify your conclusions by using your analysis results
3. Show the output of your code on this data and attach your code at the end of the report

Task 2 – Clustering

Iris dataset was collected by Sir Ronald Aylmer Fisher, a great mathematician and statistician, in 1936. This dataset has been provided with standard R distribution. Load this dataset into your R workspace and study it. In this task, you will perform clustering on this dataset based on its four attributes of "Sepal.Length", "Sepal.Width", "Petal.Length", and "Petal.Width".

In your report, you need to

1. Describe your observation on this dataset such as the number of examples, the number of features, and the meaning of these features. You shall also use `summary()` function to help you gain more understanding.
2. Plot the scatterplot matrix of Iris dataset to visualize the pairwise relationships between the four attributes.
3. Perform K-means clustering analysis on the Iris dataset and report your result. Explain how you choose the number of clusters and justify your choice.
4. Find ways to visualize your clustering result and perform diagnostics to answer the following questions:
    a. Are the clusters well separated from each other?
    b. Do any of the clusters have only a few points?
    c. Do any of the centroids appear to be too close to each other?
5. Learn to perform hierarchical agglomerative clustering via `hclust()` function and compare the clustering result with that obtained with K-means clustering.
6. Show the output of your code on this data and attach your code at the end of the report.

Task 3 – Association Rule

Students of different grade, gender, and enrolment took part in a test. The test result "Success" or "Not Success" is recorded for each student and saved in "A1_success_data.csv" provided in this assignment. In this task, you will use association rule to mine interesting relationships between these four attributes.

In your report, you need to

1. Generate frequent itemsets by applying various "support" thresholds and inspect these itemsets by displaying their support, confidence, and lift values
2. Set the right hand side (rhs) as the attribute "Success" to generate the frequent itemsets that can help to predict if a student can pass this test or not based on his/her grade, gender and/or enrolment.
3. Visualize the rules generated in the last step by 1) showing the relationship among support, confidence and lift and 2) using the graph visualization based on the sorted lift value.
4. Show the output of your code on this data and attach your code at the end of the report.

Submit:

Important:

1. **The report must be in PDF format.**
2. **The report shall contain sufficient and detailed description, explanation, justification and discussion. Marks will be deducted for a BRIEF report.**
3. **Sufficient annotation shall be provided in your code to make it easy to understand.**

Also submit your source code in the file A1.zip via Moodle

Note: Failure of your code to run may attract zero marks. Code or reports considered to be unreasonably same due to copying will attract zero marks. You may be requested to demonstrate and explain your program when necessary. Marks will be awarded for correct design, implementation and style. Any request for an extension of the submission deadline or demonstration time limit must be made to the Subject Coordinator before the submission deadline. Supporting documentation must accompany the request for any extension. Late assignment submissions without granted extension will be marked but the mark awarded will be reduced by 25% of the assignment mark for each (or part of) day (including weekends) late.

--- END ---