PSS718 - Data Mining
Assignment 2

due 23 Nov 2016

Follow the steps below. When you see R provide R code. When you see V provide visualizations. When you see E provide explanation in text. When you see a figure along the question, that is provided as a guide and an example. You may try to replicate the figure as much as possible but you should be answering the question using the figure you have generated on your own.

To answer some of the following, you may need to create your own supplementary data sets. When that is so, also provide the R code for creating those sets (or a CSV file output).

You may prepare your report in Latex or Word, I will accept only hard copies, and I will accept hard copies only if everything is clearly readable.

Try to do/answer the following:

1. The *iris* dataset in the *datasets* package provides 150 observations of 5 variables on the iris flower genus. The dataset contains 50 observations of each of the *setosa*, *versicolor* and *virginica* species. Load the dataset and remove the *species* variable. R

2. Does the dataset require any transformation prior to further analysis? If so, what are these transformations and why are they necessary. If not, why not? E (R if necessary)

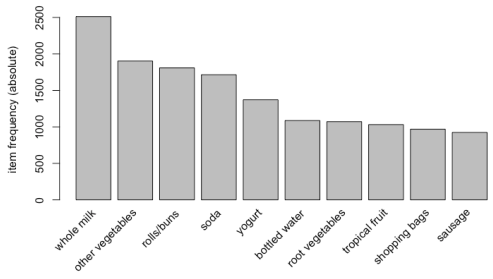
3. Consider the kmeans clustering of the dataset for values of k between 1 and 10. Which k value provides the best clustering? Consider the following measures of clustering quality and try to justify your answer E V :

   - betweenss
   - withinss
   - betweenss / tot.withinss

4. Construct a comparison table between your clustering and the actual species information in the iris table. [Hint: Check use of table() with two vectors]

5. Load the Groceries dataset from the arules package. This is a transaction dataset (already in the transactions format). Also install and load the arulesViz package. This package helps in visualizing transaction data and association rules. R

6. Draw a plot showing the top 10 highest support items. [Hint: itemFrequencyPlot] V



7. List the top 5 highest confidence association rules with at least 0.0015 support and a minimum rule length of 2. Also list the top 5 highest lift arules. R

8. You are designing the shelves of a market. Based on this groceries data, which item would you place next to the "cereals"? R E

9. Let assume your answer to the above question is X. Now, what would you place between "cereals" and "X" ? R E

10. It looks like "whole milk" goes with almost anything, ...almost. Which item would you never place close to the "whole milk" section? R E

11. Consider a model with 0.005 support and 0.65 confidence. Visualize the rules using the plot function. V