Please complete the assigned problems to the best of your abilities. Ensure that the work you do is entirely your own, external resources are only used as permitted by the instructor, and all allowed sources are given proper credit for non-original content.

1 Recitation Exercises

These exercises are to be found in: **Introduction to Data Mining, 2nd Edition** by *Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar.*

1.1 Chapter 5

Exercises: 2,6,8,9,12,13,20

2 Practicum Problems

These problems will primarily reference the *lecture materials and the examples given in class* using **Python**. It is suggested that a *Jupyter/IPython* notebook be used for the programmatic components.

2.1 Problem 1

Load the *Online Retail* dataset (**Online Retail.xlsx**) from the UCI Machine Learning Repository into **Python** using a Pandas dataframe. Using the *apriori* module from the **MLxtend** library, generate Frequent Itemsets for all transactions for the country of France. What itemset has the largest support? Set the minimum support threshold to 5% and extract frequent itemsets, and use them as input to the *association_rules* module. Use each of the confidence and lift metrics to extract the association rules with the highest values, respectively. What are the antecedents and consequents of each rule? Is the rule with the highest confidence the same as the rule with the highest lift? Why or why not?

2.2 Problem 2

Load the *Extended Bakery* dataset (**75000-out2-binary.csv**) into **Python** using a Pandas dataframe. Calculate the binary correlation coefficient Φ for the *Chocolate Coffee* and *Chocolate Cake* items. Are these two items symmetric binary variables? Provide supporting calculations. Would the association rules $\{Chocolate\ Coffee\} \implies \{Chocolate\ Cake\}$ have the same value for Φ as $\{Chocolate\ Cake\} \implies \{Chocolate\ Coffee\}$?