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 Problems

These problems are to be found in: **Introduction to Data Mining, 1st Edition** by *Pang-Ning Tan, Michael Steinbach, Vipin Kumar.*

1.1 Chapter 6

Problems: 2,6,7,9,11,12,18,19,20

2 Practicum Problems

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

2.1 Problem 1

Load the *market-basket* sample dataset into the **Orange** application, and run both frequent itemset as well as association rule modules. Set the *support* threshold to 10% and observe the *antecedent* in the rules with the highest lift. What item is observed to be there, and what is its support? Is this a valuable association rule? Why or why not?

2.2 Problem 2

Load the *Extended Bakery* dataset (**75000-out2-final.csv**) into the **Orange** application, and run both frequent itemset as well as association rule modules. Set the *support* threshold to 1% and the *confidence* threshold to 90%. Observe the association rules containing the *Cherry Tart* item within the *antecedent*. What other item appears with it? When the *confidence* threshold is lowered to 45%, does the *Cherry Tart* item now appear without another item in the *antecedent*? Is the same *consequent* observed in both cases? How did lowering the confidence threshold lead to this change? Hint: Reference the Simpson's Paradox section of the text.

2.3 Problem 3

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. Show whether the two items are symmetric binary variables via their *co-presence* and *co-absence*. Would an association rule between these items as *antecedent* and *consequent* have a high confidence level? Why or why not?