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 8

Problems: 4,7,11,17,21,22,23,24

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 *auto-mpg* sample dataset into the **Orange** application - ensure that *origin* is set as a *target* attribute type, as it will be used as a class label. Perform a **Hierarchical Clustering** using **Linkage** set to *Average*, after calculating **Distances**, with **Pruning** set to a *Max Depth* of 5. Also, set **Selection** to *Top N* with a value of 3. This will result in a shallow tree of depth 5, and a final cut resulting in 3 clusters. Examine the resulting clusters (C1,C2,C3) via **Distributions** analysis - is there a clear relationship between the cluster assignment and class label (1,2,3)? What are the probabilities calculated for each value of *origin* for each cluster? Does changing the **Max Depth** affect the results in any way?

2.2 Problem 2

Load the *breast-cancer-wisconsin-cont* dataset into the **Orange** application, and run a k-means analysis with the number of clusters **Optimized From** values for k from 2 to 5. Use **Silhouette** scoring - what is the score for each value of k? For the best score, what are the coordinates of the centroids? What are the distances between the centroids for the best score?

2.3 Problem 3

Load the *Boston* dataset (**sklearn.datasets.load_boston()**) into **Python** using a Pandas dataframe. Perform a K-Means analysis on *unscaled* data, with the number of clusters ranging from 2 to 6. Provide the *Silhouette* score to justify which value of k is optimal. What information do the values of *Homogeneity/Completeness* provide as well? Calculate the mean values for all features in each cluster for the optimal clustering - how do these values differ from the centroid coordinates?