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 Textbook 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 1

Problems: 1

1.2 Chapter 2

Problems: 2,7,18,19

1.3 Chapter 3

Problems: 8

1.4 Chapter 4

Problems: 2,3

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, and visualize the dataset. Create a scatterplot between *mpg* and *weight* - what is the basic relationship between these variables using just visual inspection? Do the results make sense? Why?

2.2 Problem 2

Load the *auto-mpg* sample dataset into **Python** using a Pandas dataframe. The *horsepower* feature has a few missing values with a *?* - replace these with a NaN from NumPy, and calculate summary statistics for each numerical column. How do the summary statistics vary when excluding the NaNs, vs. imputing

them with the mean (**Hint**: Use an Imputer from Scikit) - can we do better than just using the overall sample mean?

2.3 Problem 3

Load the *iris* sample dataset into **Python** using a Pandas dataframe. Perform a PCA using the Scikit *Decomposition* component, and provide the percentage of variance explained by the 1st Principal Component. Use *Matplotlib* to plot the 1st/2nd Principal Components to recreate the scatterplot shown in class, with colored classes for each flower type.

2.4 Problem 4

Build two classification trees using the *iris* sample dataset within the **Orange** application. Keep all parameters for both classifiers the same (Feature Selection, Pruning), and modify the Limit Depth parameter to a smaller value than the default (e.g., from 10 to 2). How does this affect the Precision and Recall of the classifier? What types of flowers are misclassified? Why? What does Tan refer to as the border where these misclassifications occur?