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

1.1 Chapter 1

Exercises: 1

1.2 Chapter 2

Exercises: 2,7,15,16,17,18,19

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 *titanic* sample dataset from the **Seaborn** library into **Python** using a Pandas dataframe, and visualize the dataset. Create a distribution plot (histogram) of *survival* conditional on *age* and *gender* - 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 from the UCI Machine Learning Repository (**auto-mpg.data**) 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 (**Hint**: Use an Imputer from Scikit). Replace the missing values with the overall mean, median, and mode (**Hint**: Pandas makes this easy) - and calculate the variance of the feature. What imputation results in the lowest variance? Why? Is there a different method of imputing values that would match the distribution more accurately? Describe your method.

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 each of the Principal Components. Compare this to the percentage of variance explained by each of the original features. What do you observe?

2.4 Problem 4

Use *Matplotlib* to plot a projection of each feature onto the 1st Principal Component from the above problem against vs. the original feature itself. Which pair of features show a closer relationship to PC1 vs. the others? Why? (**Hint**: Think in terms of cosine distance/the angle θ). Calculate the correlation coefficient between the pair of features you have selected and their projections onto PC1 - do the result agree with the visual inspection?

2.5 Problem 5

Calculate the total variance of the original features and the total variance of the four eigenvectors from the above problem. What can you say about the corresponding values? If we wished to capture $> 95\%$ of the variance of the original data, how many principal components would we be selecting? How does this number correspond to the number of dimensions we are reducing our features to?