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, 2nd Edition** by *Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar.*

1.1 Chapter 3

Problems: 2,3,5,6,7

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 *iris* sample dataset from sklearn (**load_iris()**) into **Python** using a Pandas dataframe. Induce a set of binary Decision Trees with a minimum of 2 instances in the leaves, no splits of subsets below 5, and an maximal tree depth from 1 to 5 (you can leave the majority parameter to 95%). Which depth values result in the highest Recall? Why? Which value resulted in the lowest Precision? Why? Which value results in the best F1 score? Explain the difference between the micro/macro/weighted methods of score calculation.

2.2 Problem 2

Load the *Breast Cancer Wisconsin (Diagnostic)* sample dataset from the UCI Machine Learning Repository (The *discrete* version at: **breast-cancer-wisconsin.data**) into **Python** using a Pandas dataframe. Induce a binary Decision Tree with a minimum of 2 instances in the leaves, no splits of subsets below 5, and a maximal tree depth of 2 (use the default Gini criterion). Calculate the Entropy, Gini, and Misclassification Error of the first split - what is the Information Gain? What is the feature selected for the first split, and what value determines the decision boundary?

2.3 Problem 3

Load the *Breast Cancer Wisconsin (Diagnostic)* sample dataset from the UCI Machine Learning Repository (The *continuous* version at: **wdbc.data**) into

Python using a Pandas dataframe. Induce the same binary Decision Tree as above (now using the continuous data) but perform a PCA dimensionality reduction beforehand. Using only the first principal component of the data for a model fit, what is the F1, Precision, and Recall of the PCA-based single factor model compared to the original (continuous) data? Repeat using the first and second principal components. Using the Confusion Matrix, what are the values for FP and TP as well as FPR/TPR? Is using continuous data in this case beneficial within the model? How?

2.4 Problem 4

Simulate a binary classification dataset with a single feature using a mixture of normal distributions with *NumPy* (**Hint**: Generate two data frames with the random number and a class label, and combine them together). The normal distribution parameters (**np.random.normal**) should be (5,2) and (-5,2) for the pair of samples. Induce a binary Decision Tree of maximum depth 2, and obtain the threshold value for the feature in the first split. How does this value compare to the empirical distribution of the feature?