Data Mining II
Homework 2
Due: March 29th (11:59 pm)
40 points

**ISLR – Introduction to Statistical Learning

1) (10 points) ISLR text: Chapter 10 Question 9

2) (10 points) ISLR text: Chapter 10 Question 11

3) (10 points) Access the data "primate.scapulae" (on UB learns).
   a) Cluster the data based on single-linkage, average linkage, and complete-linkage agglomerative hierarchical clustering. Decide on the groupings, and justify it, for all three methods. Calculate the misclassification rate. Which method performed the best and which method performed the worst? Was the result in line with your expectations?

   b) Cluster the data based on K-means or K-medoids. Use an analytical technique to justify your choice in "k". How did the performance compare to the hierarchical clustering of part a? Which did you feel was a better method for this data?

4) (10 points) Run a batch-SOM analysis on the Wisconsin Breast-Cancer data (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)). Describe how well the SOM methods cluster the tumor cases into benign and malignant. Compute the U-matrix and discuss its representation for these data.