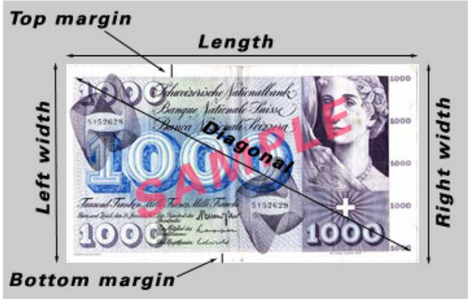
Statistical Data Mining II
Homework ~~2~~

Directions: Submit all source codes with write up. You must provide thorough explanations with output. See "homework guidelines" on UB learns for detailed information.

(1) (10 points) Access the SwissBankNotes data (UB learns). The data consists of six variables measured on 200 old Swiss 1,000-franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height of the bank note, measured on the right, distance of the inner frame to the lower border, distance of inner frame to upper border, and length of the diagonal. Carry out a PCA of the 100 genuine bank notes, of the 100 counterfeit bank notes, and all of the 200 bank notes combined. Do you notice any differences in the results? Show all work in the selection of Principal Components, including diagnostic plots.



(2) (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 in words, for all three methods. Calculate the misclassification rate for all three methods. Which method performed the best and which method performed the worst? Was the result in line wit your expectations?

b) Cluster the data based on K-means **or** K-medoids. Explain your choice for the number grouping K and calculate the misclassification rate. How did the performance compare to the hierarchical clustering of part a? Which did you feel was a better method for this data?

(3) (10 points) Exercise 14.6 in your book.