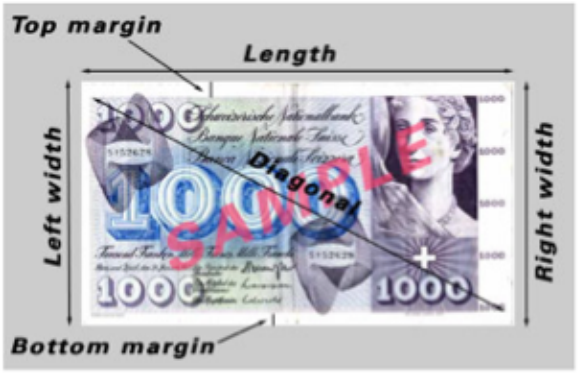
Statistical Data Mining II
Homework 2

Due: Thursday March 10th (11:59 pm)
45 points (15 points each)

Directions: See "homework guidelines" on UB learns for detailed information.

1. 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. (Ch 10, Q10 ISLR, adapted) In this problem, you will generate simulated data, and then perform. PCA and **K** means clustering on the data.

(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables. Hint: There are a number of functions in R that you can use to generate data. One example is the rnorm() function; runif() is another option. Be sure to add a mean shift to the observations in each class so that there are three distinct classes.

(b) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.

(c) Perform **K** -means clustering of the observations with **K** = 3. How well do the clusters that you obtained in **K** -means clustering compare to the true class labels? Hint: You can use the table() function in R to compare the true class labels to the class labels obtained by clustering. Be careful how you interpret the results: **K**-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same.

(d) Perform **K** -means clustering with **K** = 2 and **K** = 4. Describe your results.

(e) Now perform **K** -means clustering with **K** = 3 on the first two principal component score vectors, rather than on the raw data. Comment on the results.

(f) Using the scale() function, perform **K** -means clustering with **K** = 3 on the data after scaling each variable to have standard deviation one . How do these results compare to those obtained in (b)? Explain.

3. (Ch 10, Q11 ISLR, adapted) On the book website, www.StatLearning.com, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group. Load in the data using read.csv() . You will need to select header=F.

(b) Apply hierarchical clustering to the samples using correlation-based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

(c) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here.