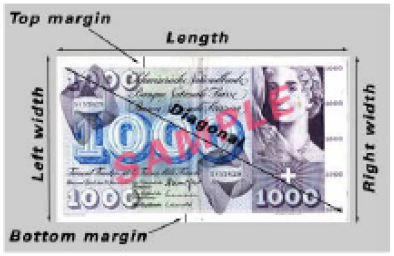
Data Mining II
Homework 3

1) (10 points) Consider the tumor microarray data in the package library(ElemStatLearn).
   >library(ElemStatLearn)
   >data(nci)
   >head(nci)
   The data consists of several different types of tumor samples. We observed that it many clustering algorithms there are often found to be 2-3 groups/clusters in this well-studied data, although there are 14 subtypes of tumor cells (unique(colnames(nci))).
   Run a SOM algorithm and present the results (e.g., U-matrix, phase plots if appropriate, hclust on prototypes). How well does the SOM method characterize the tumor cells into groups?

2) (10 points) Consider the USArrests data. I suggest scaling the data first.
   > library(ISLR)
   > data(USArrests)
   > head(USArrests)
   a) Perform hierarchical clustering with complete linkage and Euclidean distance to cluster the states. Cut the dendrogram at a height that results in three clusters. Is this what you would expect?
   b) Fit a SOM to the data and present the results (e.g., U-matrix, phase plots if appropriate, hclust on prototypes). Is this what you would expect? Does this result generally support your results in Part A.
   c) Comment on the advantages and limitations of hierarchical clustering to SOM, and discuss when one would be preferred over the other.

3) (10 points) Access the SwissBankNotes data (posted with assignment). 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. Generate some score plots (use colors for the combined). Do you notice any differences in the results? Show your work, and ustify the selection of Principal Components, including diagnostic plots.