Northeastern University
College of Science

Module 12-Homework

Problem 1 (60 points) Analysis of the ALL data set

(a) Define an indicator variable ALL.fac such that ALL.fac=1 for T-cell patients and ALL.fac=2 for B-cell patients.

(b) Plot the histograms for the first three genes' expression values in one row.

(c) Plot the pairwise scatterplots for the first five genes.

(d) Do a 3D scatterplot for the genes "39317_at", "32649_at" and "481_at", and color according to ALL.fac (give different colors for B-cell versus T-cell patients). Can the two patient groups be distinguished using these three genes?

(e) Do K-means clustering for K=2 and K=3 using the three genes in (d). Compare the resulting clusters with the two patient groups. Are the two groups discovered by the clustering analysis?

(f) Carry out the PCA on the ALL data set with scaled variables. What proportion of variance is explained by the first principal component? By the second principal component?

(g) Do a biplot of the first two principal components. Observe the pattern for the loadings. What info is the first principal component summarizing?

(h) For the second principal component PC2, print out the three genes with biggest PC2 values and the three genes with smallest PC2 values.

(i) Find the gene names and chromosomes for the gene with biggest PC2 value and the gene with smallest PC2 value. (Hint: review Module 10 on searching the annotation.)

Problem 2 (40 points) Variables scaling and PCA in the iris data set

In this module and last module, we mentioned that the variables are often scaled before doing the PCA or the clustering analysis. By "scaling a variable", we mean to apply a linear transformation to center the observations to have mean zero and standard deviation one. In last module, we also mentioned using the correlation-based dissimilarity measure versus using the Euclidean distance in clustering analysis. It turns out that the correlation-based dissimilarity measure is proportional to the squared Euclidean distance on the scaled variables. We check this on the iris data set. And we compare the PCA on scaled versus unscaled variables for the iris data set.

(a) Create a data set consisting of the first four numerical variables in the iris data set (That is, to drop the last variable Species which is categorical). Then make a scaled data set that centers each of the four variables (columns) to have mean zero and variance one.

(b) Calculate the correlations between the columns of the data sets using the cor() function. Show that these correlations are the same for scaled and the unscaled data sets.

(c) Calculate the Euclidean distances between the columns of the scaled data set using dist() function. Show that the squares of these Euclidean distances are proportional to the (1-correlation)s. What is the value of the proportional factor here?

(d) Show the outputs for doing PCA on the scaled data set and on the unscaled data set. (Apply PCA on the two data sets with option "scale=FALSE". Do NOT use option "scale=TRUE", which will scale data no matter which data set you are using.) Are they the same?

(e) What proportions of variance are explained by the first two principle components in the scaled PCA and in the unscaled PCA?

(f) Find a 90% confidence interval on the proportion of variance explained by the second principal component, in the scaled PCA.