CS 6220 - Fall 2015 Olga Vitek

Homework 7

Each part of the problems 5 points
Due on Blackboard by **5pm on Wednesday November 11**.
JWHT is 'An Introduction to Statistical Learning' by James, Witten, Hastie and Tibshirani.

In this homework we will consider a genomic dataset, which is similar to the NCI60 dataset from JWHT (genomic datasets are nice for illustrating multivariate analysis, because all the variables are on a comparable scale).

A study by Tuch et al. (2010) recorded the activity of genes in 6 tissues from 3 human subjects. There are two types of tissues: healthy, denoted as 'N', and cancer (oral squamous cell carcinoma), denoted as 'T'. Each tissue is characterized by 10453 variables (i.e., genes). The numbers in the array are activity of each gene is each tissue (quantified as counts of messenger RNA molecules).

Two versions of the dataset are available on Piazza and on the course website: the full dataset with 10453 genes, and a subset of dataset of 'active' genes (i.e., genes that systematically change activity between healthy and tumor tissues).

1. **PCA:**

   (a) On the full dataset, perform the basic Principle Component Analysis of the tissues (i.e., tissues are the observations, and genes are the variables), without standardization (i.e. use option `scale.=FALSE`). Produce the score plot. On the plot, use colors of the points to indicate disease status, and annotate each sample with the subject id.

      i. How many principle components is there in this dataset, and what is the reason for this number?

      ii. What would be a desirable pattern of the score plot in this experiment? Does the score plot produce a desirable pattern?

   (b) Repeat the questions above, but after scaling each variable (i.e. use option `scale.=TRUE`). How does the scaling change the score plot? How does it affect the percentage of the variation explained by the first two principle components?

   (c) Repeat the questions 1 and 2 above, but for the subset of 'active' genes.

   (d) Summarize your findings. What is your recommendation for the analysis of similar dataset in the future?

2. **Clustering and heatmaps:** We will continue working with the dataset that contains a subset of 'active' genes.

   (a) We will first investigate the role of scaling the color. Use `heatmap` with the default distance for the dendrograms (Eucledian) and default scaling of the colors (rows), and compare the result with `scale="none"`. Explain the role of the scaling on

the appearance of the dendrograms and on the color of the heat map. We will be scaling the colors from now on.

(b) Center and standardize each row in the matrix of normalized counts. Draw the heatmap of the centered and scaled values using Eucledian and Correlation distances. Discuss the reason for the differences and the similarities between the heatmaps produced with these two distances.

(c) Summarize your findings. What is your recommendation for the visual representation of such data in the future?

3. **Data partition with K-means:**

(a) On the full dataset, perform K-means clustering of the tissues with K = 2. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

(b) Repeat the question above, but with the dataset that contains a subset of 'active' genes.

(c) Summarize your findings. What is your recommendation for clustering of such data in the future?