

Ma 322: Biostatistics

Homework Assignment 12

Prof. Wickerhauser

Due Friday, April 27th, 2012

Read Chapter 16, “Working with Multivariate Data,” pages 288–318 of our text.

1. Load the NCI data set from the class web site and write R codes to do the following:
 - (a) Print the list of cancers that appear 7 or more times in the 64 rows.
 - (b) Count the total number of rows of data from cancers appearing 7 or more times.
 - (c) Print the column names that contain gene expression data with the top 5 variances.
 - (d) Plot the classification tree for the rows in part a and the columns in part c.
 - (e) Print the misclassification rate for the tree in part d.
2. Use the Fisher `iris` data set with the functions in the `cluster` library.
 - (a) Use `agnes()` to compute, and then plot, the dendrogram for the iris data decomposed by agglomerative hierarchical clustering.
 - (b) Use `diana()` to compute, and then plot, the dendrogram for the iris data decomposed by divisive hierarchical clustering.
 - (c) Use `kmeans()` to find 3 clusters in the iris data and determine the number of misclassifications of the 150 plants.
 - (d) Use `pam()` to find 3 clusters in the iris data and determine the number of misclassifications of the 150 plants.
3. Use the data set of 57 cancers with at least 3 repetitions, with the top 12 expressed genes, produced by the commands in <http://www.math.wustl.edu/~victor/classes/ma322/r-eg-27.txt>,
This data may also be found in <http://www.math.wustl.edu/~victor/classes/ma322/nci57x13.R> and, after copying that file into the folder being used by your R session, can be read into data frame `nci` with the command

```
load("nci57x13.R")
```

Install and load the `vegan` library into R and use `isomap()` for the following:

- (a) Find the Euclidean distances between samples. Using $k = 2$ nearest neighbors, apply multidimensional scaling to these Euclidean distances and plot the result.
- (b) Find the Manhattan distances between samples. Using $k = 2$ nearest neighbors, apply multidimensional scaling to these Manhattan distances and plot the result.