Lab 5: Descriptive Models

Background

Briefly, “gene expression” is typically a measure of the amount of mRNA (copies) of a particular gene (before translated to a protein) that has been produced by transcription in a particular tissue type at a particular time. Many genes are up-regulated or down-regulated together through complex regulatory networks and cascades based on the specific needs of a cell or tissue. This data is a small random subset taken from a larger group of 20000 genes that were measured. Gene expression data can be used for many different purposes. Suites of genes can be co-regulated in specific cancer types and not others. This can be used purely for prediction; it can also be used to find to find heterogeneity in cancer types (i.e. not all “types” of breast cancer are the same) which may help identify why individuals with different “types” of cancer respond to drugs differently. Gene expression can be used for the inference of “types” of specific cancers by identifying which particular genes or classes of genes have co-expression.

Instructions

Load the “Session9GeneExpression1000.csv” and “Session9labels.csv” files. “GeneExpression1000.csv” contains gene expression data for 801 tumor samples (rows) and 1000 genes (columns) it has a header row of gene names. “labels.csv” contains the “Class” variable (column 2) which identifies which type of cancer each of the 801 samples in “GeneExpression1000.csv” come from. Below are the labels for each tumor type in the “Class” variable of “labels.csv”

Abbreviation	Tumor type
COAD	Colon Adenocarcinoma
KIRC	Kidney Renal Clear Cell Carcinoma
LUAD	Lung Adenocarcinoma
PRAD	Prostate Adenocarcinoma
BRCA	Breast Invasive Carcinoma

Use set.seed(1389)

After centering and scaling the gene expression data (“Session9GeneExpression1000.csv”):

1. Choose either k-means or kohonoen SOM's to analyse the data.
2. Follow up your k-means or SOM model with heirarchical clustering using hclust.

Address the following questions depending on the model you used:

Question specific for a k-means analysis

1. Based on the Hartigan method, what k should be used for a k-means model with this data? (Note: compare k=2 to k=15 and this could take a little while)

Question specific for a SOM analysis:
1. What number of nodes is recommended kohonen SOM with this data?

Question/reports due for either analysis:

Fit your preferred model (kmeans or Kohonen SOM) with the appropriate number of k (k-means) or the square with the appropriate number of nodes (SOM) and then use hierarchical clustering to cluster the kmeans clusters (the centers) or SOM nodes (the codes). From the hierarchical cluster cut the tree in to 5 groups (use the cutree function with k=5 as additional argument in instead of h=5).

Before hclust, the individuals were labelled based on the groups/nodes from your previous model (modSOM$unit.classif or modkmean$cluster). Each "group" from your original model is now classified into 5 new groups from hclust & cutree. Now you need to relabel each individual based on the new groupings.

Make a table with the “Class” variable (cancer labels for each individual) from “Session9labels.csv” (the second column) and these new groupings. For example if your new variable describing for each individual which new group it is in is called "newID5groups" and your dataframe from session9 labels is called "labels" then you can make a table with the table function as table(labels[,2],newID5groups). This should make a 5x5 table with the five cancer types as rows and your 5 groups labels as columns and counts of individuals cross-classified.

2. Paste/report the table.

3. Is there a strong correspondence between cancer cell line and the 5 groups from the hierarchical clustering?

4. Which cancer line seems to be best represented by the groups from the clustering?