Clustering

Exploratory Data Analysis with **R**

Herakles and Iolaos battle the Hydra. Classical (450-400 BCE)

Boris Steipe

UNIVERSITY OF TORONTO

DEPARTMENT OF BIOCHEMISTRY
DEPARTMENT OF MOLECULAR GENETICS

This module includes some material originally developed by Raphael Gottardo, FHCRC and by Sohrab Shah, UBC

OBJECTIVES

Understand clustering principles;

Be able to apply basic hierarchical and partitioning methods to data;

Know how to interpret the results;

Understand cluster quality control;

Know about alternatives.

Outline

Principles

Hierarchical clustering

Partitioning methods

- Centroid based clustering (K-means *etc.*)
- Model based clustering

Examples

Complexes in interaction data

Domains in protein structure

Proteins of similar function
(based on measured similar properties, e.g. coregulation)

...

Introduction to clustering

Clustering ...

- ... is an example of unsupervised learning
- ... is useful for the analysis of patterns in data
- ... can lead to class discovery.

Clustering is the partitioning of a data set into groups of elements that are more similar to each other than to elements in other groups.

Clustering is a completely general method that can be applied to genes, samples, or both.

HIERARCHICAL CLUSTERING

Given N items and a distance metric...

1. Assign each item to its own "cluster".
   Initialize the distance matrix between clusters as the distance between items.
2. Find the closest pair of clusters and merge them into a single cluster.
3. Compute new distances between clusters.
4. Repeat 2-3 until all clusters have been merged into a single cluster.

HIERARCHICAL CLUSTERING

"Given N items and a **distance metric** ..."

What is a **metric**? $d : X \times X \rightarrow \mathbf{R}$

A metric has to fulfill three conditions:

$d(x,y) = 0 \Leftrightarrow x = y$ "identity"

$d(x,y) = d(y,x)$ "symmetry"

$d(x,y) \leq d(x,z) + d(z,y)$ "triangle inequality"

Distance metrics

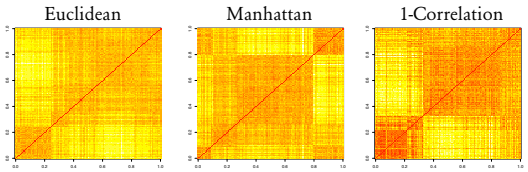
Common metrics include:

Manhattan distance: $d_{xy} = \sum_{i=1}^{p} |x_i - y_i|$

Euclidean distance: $\|x - y\| = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$

1-correlation:
(proportional to Euclidean distance, but invariant to range of measurement from one sample to the next).

Distance metrics compared



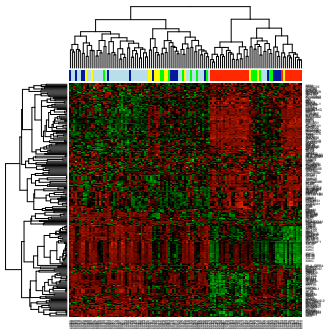
Distance matters!

Other distance metrics

- Hamming distance for ordinal, binary or categorical data:

$$d_{xy} = \sum_{i=1}^{p} I(x_i \neq y_i)$$

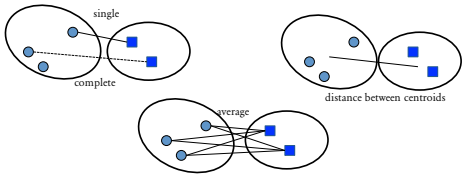
AGGLOMERATIVE HIERARCHICAL CLUSTERING



HIERARCHICAL CLUSTERING

- Anatomy of hierarchical clustering
  - distance matrix
  - linkage method
- Output
  - dendrogram
    - a tree that defines the relationships between objects and the distance between clusters
    - a nested sequence of clusters

Linkage methods



Hierarchical clustering analyzed

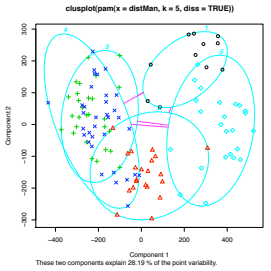
Advantages	Disadvantages
There may be small clusters nested inside large ones	Clusters might not be naturally represented by a hierarchical structure
No need to specify number groups ahead of time	Its necessary to 'cut' the dendrogram in order to produce clusters
Flexible linkage methods	Bottom up clustering can result in poor structure at the top of the tree. Early joins cannot be 'undone'

Partitioning methods

- Anatomy of a partitioning based method
  - data matrix
  - distance function
  - **number of groups**
- Output
  - group assignment of every object

Partitioning based methods

- Choose K groups
  - initialise group centers
    - aka centroid, medoid
  - assign each object to the nearest centroid according to the distance metric
  - reassign (or recompute) centroids
  - repeat last 2 steps until assignment stabilizes



K-means vs. K-medoids

K-means	K-medoids
Centroids are the 'mean' of the clusters	Centroids are an actual object that minimizes the total within cluster distance
Centroids need to be recomputed every iteration	Centroid can be determined from quick look up into the distance matrix
Initialisation difficult as notion of centroid may be unclear before beginning	Initialisation is simply K randomly selected objects
`kmeans`	`pam`

Partitioning based methods

Advantages	Disadvantages
Number of groups is well defined	Have to choose the number of groups
A clear, deterministic assignment of an object to a group	Sometimes objects do not fit well to any cluster
Simple algorithms for inference	Can converge on locally optimal solutions and often require multiple restarts with random initializations

K-means

N items, assume K clusters

Goal is to minimize $\sum_{k=1}^{K} \sum_{\{\boldsymbol{x}_i : l(\boldsymbol{x}_i)=k\}} \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2$

over the possible assignments and centroids . c_k

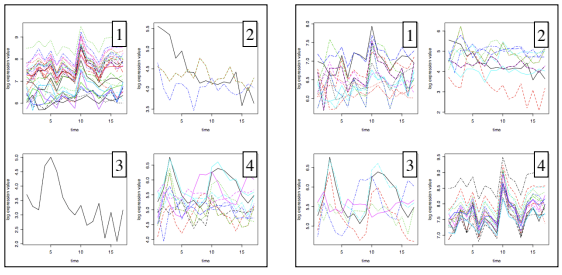
c_k represents the location of the cluster.

K-means

1. Divide the data into K clusters
   Initialize the centroids with the mean of the clusters
2. Assign each item to the cluster with closest centroid
3. When all objects have been assigned, recalculate the centroids (mean)
4. Repeat 2-3 until the centroids no longer move

K-means

Single linkage, k=4 K-means, k=4



Summary

K-means and hierarchical clustering methods are simple, fast and useful techniques

Beware of memory requirements for HC

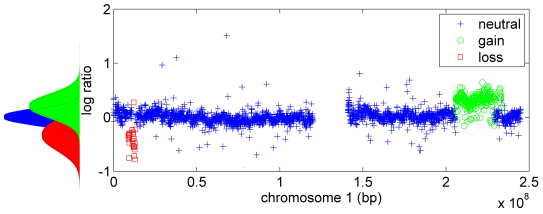
Both are bit "*ad hoc*":

- Number of clusters?
- Distance metric?
- Good clustering?

MODEL BASED APPROACHES

- Assume the data are 'generated' from a mixture of K distributions
  - What cluster assignment and parameters of the K distributions best explain the data?
- 'Fit' a model to the data
- Try to get the best fit
- Classical example: mixture of Gaussians (mixture of normals)
- Take advantage of probability theory and well-defined distributions in statistics

MODEL BASED CLUSTERING: ARRAY CGH



Advantages of model based approaches

- In addition to clustering patients into groups, we output a 'model' that best represents the patients in a group
- We can then associate each model with clinical variables and simply output a classifier to be used on new patients
- Choosing the number of groups becomes a model selection problem (cf. the Bayesian Information Criterion)
  - see Yeung et al Bioinformatics (2001)

Advanced topics in clustering

- Top down clustering
- Bi-clustering or 'two-way' clustering
- Principal components analysis
- Choosing the number of groups
  - model selection
    - AIC, BIC
    - Silhouette coefficient
    - The Gap curve
- Joint clustering and feature selection

Best method?

What is the **best** clustering method?

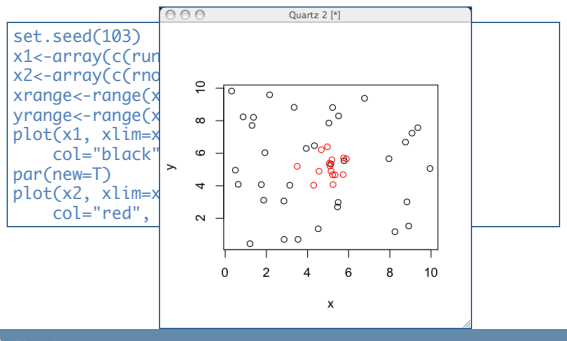
That depends on what you want to achieve. And sometimes clustering is not the best approach to begin with.

Density estimation

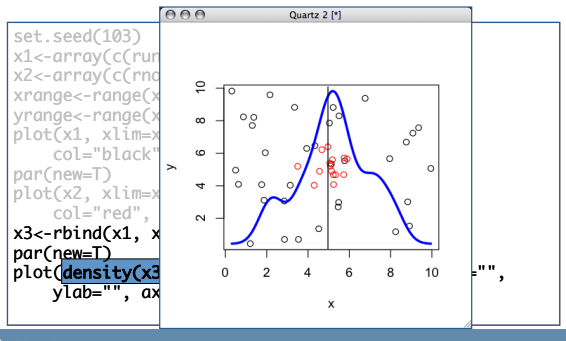
Clustering is a **partition** method. However, consider the following data:

```
set.seed(103)
x1<-array(c(runif(70, 0,10)), c(35,2))
x2<-array(c(rnorm(30, 7, 0.7)), c(15,2))
xrange<-range(x1[,1], x2[,1])
yrange<-range(x1[,2], x2[,2])
plot(x1, xlim=xrange, ylim=yrange,
    col="black", xlab="x", ylab="y")
par(new=T)
plot(x2, xlim=xrange, ylim=yrange,
    col="red", axes=F,  xlab="", ylab="")
```


Density estimation



(univariate) Density estimation



boris.steipe@utoronto.ca

TITLE

Text

Highlight