Clustering and Data Mining in R

Workshop Supplement

Thomas Girke

October 3, 2010

Outline

What is Clustering?

- Clustering is the classification of data objects into similarity groups (clusters) according to a defined distance measure.
- It is used in many fields, such as machine learning, data mining, pattern recognition, image analysis, genomics, systems biology, etc.

Why Clustering and Data Mining in R?

- Efficient data structures and functions for clustering.
- Efficient environment for algorithm prototyping and benchmarking.
- Comprehensive set of clustering and machine learning libraries.
- Standard for data analysis in many areas.

Outline

Data Transformations

Choice depends on data set!

- Center & standardize
  1. Center: subtract from each vector its mean
  2. Standardize: devide by standard deviation

  $\Rightarrow$ $Mean = 0$ and $STDEV = 1$
- Center & scale with the `scale()` fuction
  1. Center: subtract from each vector its mean
  2. Scale: divide centered vector by their root mean square (rms)

$$x_{rms} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} {x_i}^2}$$

  $\Rightarrow$ $Mean = 0$ and $STDEV = 1$
- Log transformation
- Rank transformation: replace measured values by ranks
- No transformation

Distance Methods

List of most common ones!

- Euclidean distance for two profiles X and Y

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Disadvantages: not scale invariant, not for negative correlations

- Maximum, Manhattan, Canberra, binary, Minowski, ...
- Correlation-based distance: $1 - r$
  - Pearson correlation coefficient (PCC)

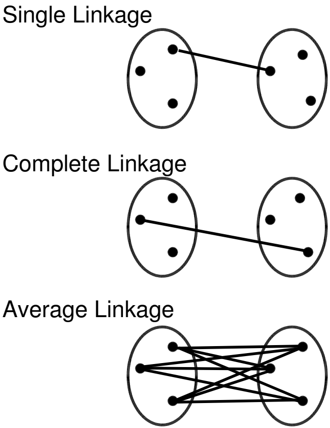
$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2)(\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}}$$

Disadvantage: outlier sensitive

  - Spearman correlation coefficient (SCC)

Same calculation as PCC but with ranked values!

Cluster Linkage



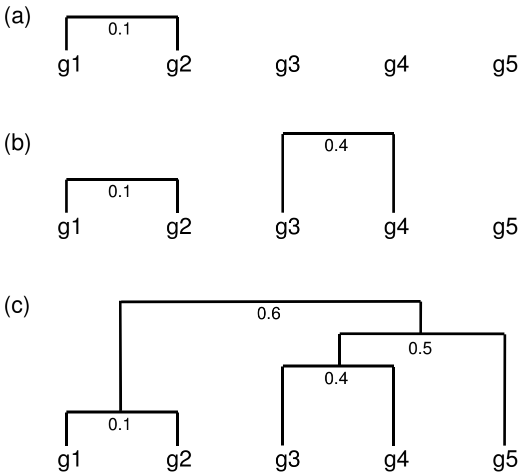
Outline

Hierarchical Clustering Steps

1. Identify clusters (items) with closest distance
2. Join them to new clusters
3. Compute distance between clusters (items)
4. Return to step 1

Hierarchical Clustering

Agglomerative Approach



Hierarchical Clustering Approaches

1. Agglomerative approach (bottom-up)

   `hclust()` and `agnes()`

2. Divisive approach (top-down)

   `diana()`

Tree Cutting to Obtain Discrete Clusters

1. Node height in tree
2. Number of clusters
3. Search tree nodes by distance cutoff

Outline

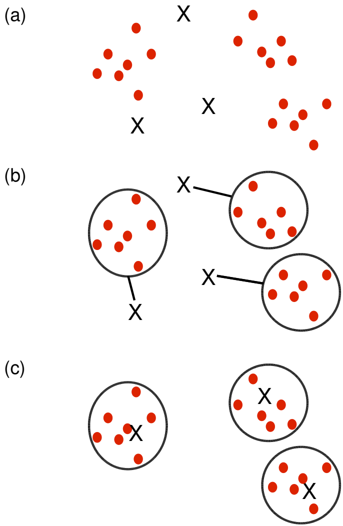
Non-Hierarchical Clustering

Selected Examples

K-Means Clustering

1. Choose the number of k clusters
2. Randomly assign items to the k clusters
3. Calculate new centroid for each of the k clusters
4. Calculate the distance of all items to the k centroids
5. Assign items to closest centroid
6. Repeat until clusters assignments are stable

K-Means



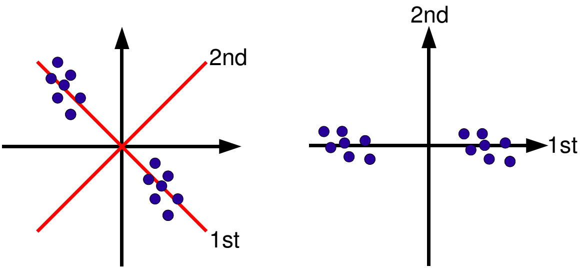
Principal Component Analysis (PCA)

Principal components analysis (PCA) is a data reduction technique that allows to simplify multidimensional data sets to 2 or 3 dimensions for plotting purposes and visual variance analysis.

Basic PCA Steps

- Center (and standardize) data
- First principal component axis
  - Accross centroid of data cloud
  - Distance of each point to that line is minimized, so that it crosses the maximum variation of the data cloud
- Second principal component axis
  - Orthogonal to first principal component
  - Along maximum variation in the data
- 1^{st} PCA axis becomes x-axis and 2^{nd} PCA axis y-axis
- Continue process until the necessary number of principal components is obtained

PCA on Two-Dimensional Data Set



Identifies the Amount of Variability between Components

Example

Principal Component	**1^{st}**	**2^{nd}**	**3^{rd}**	**Other**
Proportion of Variance	62%	34%	3%	rest

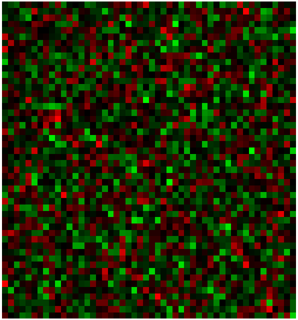
1^{st} and 2^{nd} principal components explain 96% of variance.

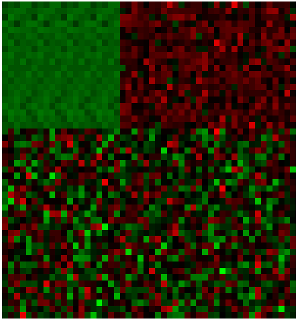
Multidimensional Scaling (MDS)

- Alternative dimensionality reduction approach
- Represents distances in 2D or 3D space
- Starts from distance matrix (PCA uses data points)

Biclustering

Finds in matrix subgroups of rows and columns which are as similar as possible to each other and as different as possible to the remaining data points.





Unclustered $\Longrightarrow$ Clustered

Remember: There Are Many Additional Techniques!

Continue with R manual section:
"Clustering and Data Mining"