

Classification - Clustering

STAT3014
Week 4 - 1

Basic principles of clustering

Aim: to group observations that are “similar” based on predefined criteria.

Issues:

- Data types – counts, ratio, ordinal , categorical and interval scale (continuous).
- Missing data
- Scaling
- Metric:
 - Euclidean
 - Manhattan

2

Commonly used measure?

- A metric is a measure of the **similarity** or **dissimilarity** between two data objects and it's used to form data points into clusters
- Two main classes of distance:
 - **Correlation coefficients** (compares shape of expression curves)

$$\frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

S_x = Standard deviation of x
 S_y = Standard deviation of y

- **Distance metrics**

- City Block (Manhattan) distance: $d(X, Y) = \sum_i |x_i - y_i|$

- Euclidean distance: $d(X, Y) = \sqrt{\sum_i (x_i - y_i)^2}$

3

R functions

```
> X = 1:3
> Y = 1:3 + 1

> dist(cbind(X,Y))
      1      2
2 1.414214
3 2.828427 1.414214

> dist(cbind(X,Y), method="manhattan")
 1 2
2 2
3 4 2

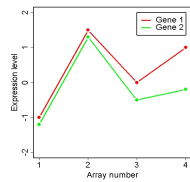
> cor(cbind(X,Y))
  X Y
X 1 1
Y 1 1

> as.dist(1-cor(cbind(X,Y)))
[1] 0
```

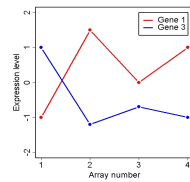
4

Correlation (a measure between -1 and 1)

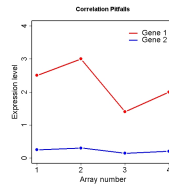
- Others include Spearman's ρ and Kendall's τ
- You can use **absolute correlation** to capture both positive and negative correlation



Positive correlation



Negative correlation



Potential pitfalls
Correlation = 1

5

Distance between clusters

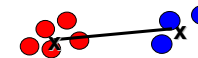
Between-cluster dissimilarity measures



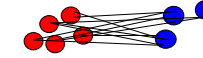
Single (minimum)



Complete (maximum)



Distance between centroids



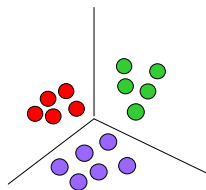
Average (Mean) linkage

6

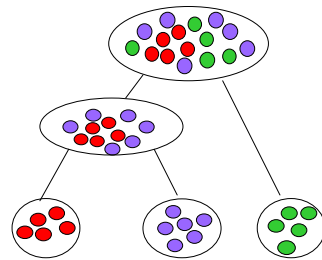
Clustering algorithms

- Clustering algorithm comes in 2 basic flavors

Partitioning



Hierarchical

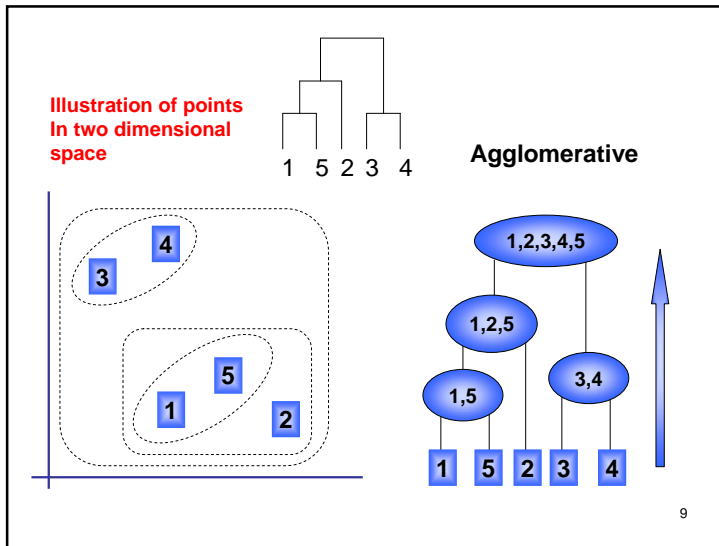


7

Hierarchical methods

- Hierarchical clustering methods produce a **tree** or **dendrogram**.
- They avoid specifying how many clusters are appropriate by providing a partition for each k obtained from cutting the tree at some level.
- The tree can be built in two distinct ways
 - bottom-up: **agglomerative** clustering.
 - top-down: **divisive** clustering.

8



Agglomerative Methods

- Start with n sample.
- At each step, **merge** the two closest clusters using a measure of between-cluster dissimilarity which reflects the shape of the clusters
- The distance between clusters is defined by the method used (e.g., if complete linkage, the distance is defined as the distance between furthest pair of points in the two clusters)

10

?hclust

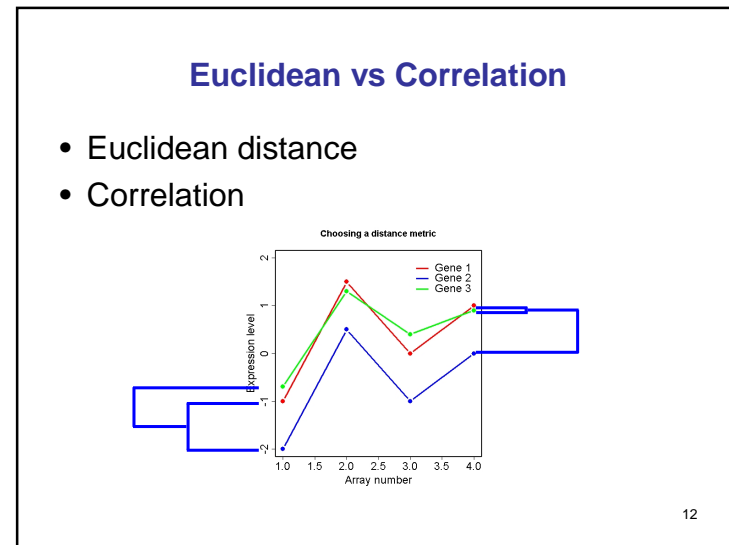
```

> par(mfrow=c(2,2))
> hc <- hclust(dist(USArrests), method="ave")
> plot(hc)
> hc <- hclust(dist(USArrests), method="single")
> plot(hc)
> hc <- hclust(dist(USArrests), method="complete")
> plot(hc)

```

Compare the trees using different agglomeration method.

11



Partitioning methods

- Partition the data into a **pre-specified** number k of mutually exclusive and exhaustive groups.
- Iteratively reallocate the observations to clusters until some criterion is met, e.g. minimize within cluster sums of squares.
- Examples:
 - k-means, self-organizing maps (SOM), PAM, etc.;
 - Fuzzy: needs stochastic model, e.g. Gaussian mixtures.

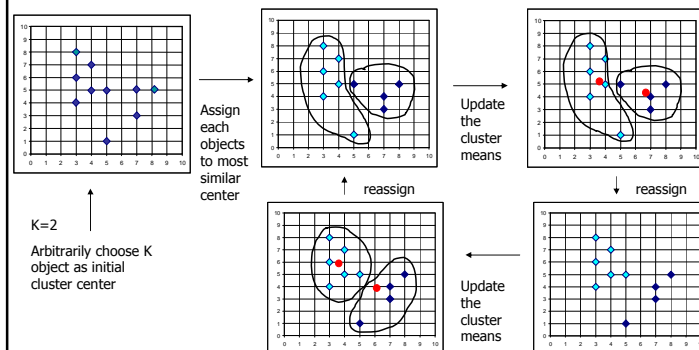
13

Example: K-means

- Arbitrarily choose k objects as the initial cluster centers
- Until no change, do
 - (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster
 - Update the cluster means, i.e., calculate the mean value of the objects for each cluster

14

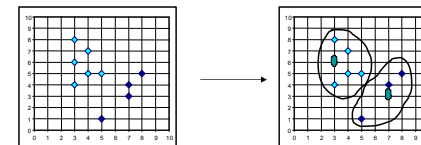
K-Means: Example



15

A Problem of K-means

- Sensitive to outliers
 - Outlier: objects with extremely large values
 - May substantially distort the distribution of the data
- K-medoids: the most centrally located object in a cluster



16

?kmeans

```
# a 2-dimensional example
> x <- rbind(matrix(rnorm(100, sd = 0.3), ncol = 2),
              matrix(rnorm(100, mean = 1, sd = 0.3), ncol = 2))
> colnames(x) <- c("x", "y")
> cl <- kmeans(x, 2)
> plot(x, col = cl$cluster)
> points(cl$centers, col = 1:2, pch = 8, cex=2)
```

17

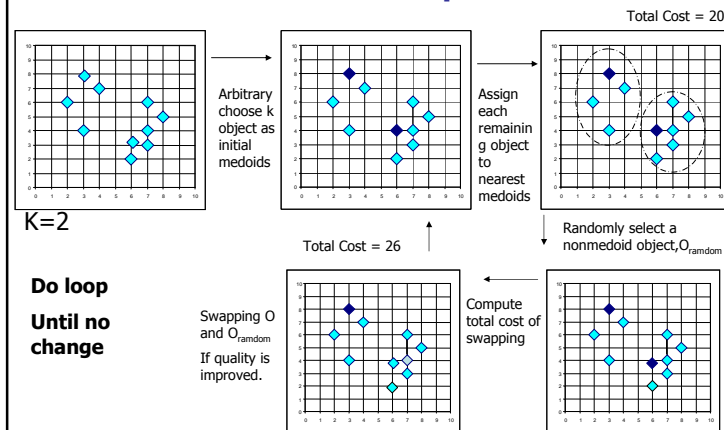
PAM: A K-medoids Method

- PAM: partitioning around Medoids
- Arbitrarily choose k objects as the initial medoids
- Until no change, do
 - (Re)assign each object to the cluster to which the nearest medoid
 - Randomly select a non-medoid object o' , compute the total cost, S , of swapping medoid o with o'
 - If $S < 0$ then swap o with o' to form the new set of k medoids

```
library(cluster)
?pam
```

18

PAM: Example



19

Pros and Cons of PAM

- PAM is more robust than k-means in the presence of noise and outliers
 - Medoids are less influenced by outliers
- PAM is efficient for small data sets but does not scale well for large data sets.

20

Partitioning vs. hierarchical

Partitioning: Advantages

- Optimal for certain criteria.
- Samples automatically assigned to clusters

Disadvantages

- Need initial k ;
- Often require long computation times.
- All samples are forced into a cluster.

Hierarchical Advantages

- Faster computation.
- Visual.

Disadvantages

- Unrelated genes are eventually joined
- Rigid, cannot correct later for erroneous decisions made earlier.
- Hard to define clusters.

21

R cluster analysis packages

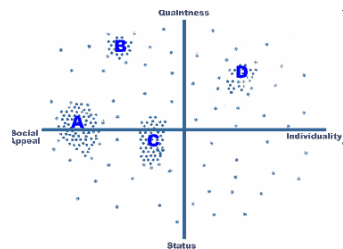
- **cclust**: convex clustering methods.
- **class**: self-organizing maps (SOM).
- **cluster**:
 - AGglomerative NESTing (**agnes**),
 - Clustering LARe Applications (**clara**),
 - DLvisive ANALysis (**diana**),
 - Fuzzy Analysis (**fanny**),
 - MONothetic Analysis (**mona**),
 - Partitioning Around Medoids (**pam**).
- **e1071**:
 - fuzzy C-means clustering (**cmeans**),
 - bagged clustering (**bc1ust**).
- **flexmix**: flexible mixture modeling.
- **fpc**: fixed point clusters, clusterwise regression and discriminant plots.
- **GeneSOM**: self-organizing maps.
- **mclust**, **mclust98**: model-based cluster analysis.
- **mva**:
 - hierarchical clustering (**hc1ust**),
 - k-means (**kmeans**).
- Specialized summary, plot, and print methods for clustering results.

Download
from CRAN

22

An application of clustering -- marketing

- Examples:
 - Segmenting the market and determining target markets.
 - Product positioning and new product development .



23

An application of clustering -- microarray data

Gene expression data on p genes for n samples

		mRNA samples					
		sample1	sample2	sample3	sample4	sample5	...
Genes	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene i in mRNA sample j

activity

24

