Unsupervised Learning: Cluster Analysis

Machine Learning (BSMC-GA 4439)

Wenke Liu

09-10-2018

Outline

- Background
- Defining proximity
- Clustering methods
- Determining number of clusters
- Other approaches

Cluster analysis as unsupervised Learning

- Unsupervised learning (as opposed to supervised learning) seeks to identify structures in the usually high-dimensional data **X** without the help of a known label **Y**.
- Cluster analysis focuses on informative partition of observations into groups.
- Cluster analysis may lead to simpler and sometimes more meaningful representation of the data.

What is a cluster?

- Objects within each cluster are closer to one another than objects in different clusters.
- Internal cohesion: homogeneity.
- External isolation: separation.

Similarity, dissimilarity and distance

- Cluster analysis starts with defining a quantitative measure of proximity between each pair of objects (observations) x_i and x_j. It can be a similarity measure s_{ij} (the larger the closer) or a dissimilarity measure δ_{ij} (the smaller the closer).
- Proximity measures can be directly given or calculated from the data.
- When a dissimilarity measure fulfills the metric inequality $\delta_{ij} + \delta_{im} \geq \delta_{jm}$ for all (i, j, m) and $\delta_{ii} = 0$, it is a distance measure d_{ij}.
- Not all metric distances are Euclidean.

Proximity measures for continuous data

- Dissimilarity measures between continuous data points are usually general distance measures or correlation measures.
- Variables can be weighted by a nonnegative w_k.
- Pearson correlation and angular separation are also metric.

Measure	Formula
D1: Euclidean distance	$d_{ij} = \left[\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2\right]^{1/2}$
D2: City block distance	$d_{ij} = \sum_{k=1}^{p} w_k \lvert x_{ik} - x_{jk}\rvert$
D3: Minkowski distance	$d_{ij} = \left(\sum_{k=1}^{p} w_k^r \lvert x_{ik} - x_{jk}\rvert^r\right)^{1/r} \quad (r \geq 1)$
D4: Canberra distance (Lance and Williams, 1966)	$d_{ij} = \begin{cases} 0 & \text{for } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^{p} w_k \lvert x_{ik} - x_{jk}\rvert / (\lvert x_{ik}\rvert + \lvert x_{jk}\rvert) & \text{for } x_{ik} \neq 0 \text{ or } x_{jk} \neq 0 \end{cases}$
D5: Pearson correlation	$\delta_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\cdot})(x_{jk} - \bar{x}_{j\cdot}) \Big/ \left[\sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\cdot})^2 \sum_{k=1}^{p} w_k (x_{jk} - \bar{x}_{j\cdot})^2\right]^{1/2}$ where $\bar{x}_{i\cdot} = \sum_{k=1}^{p} w_k x_{ik} \Big/ \sum_{k=1}^{p} w_k$
D6: Angular separation	$\delta_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k x_{ik} x_{jk} \Big/ \left(\sum_{k=1}^{p} w_k x_{ik}^2 \sum_{k=1}^{p} w_k x_{jk}^2\right)^{1/2}$

Proximity measures for binary data

		Individual i		
	Outcome	1	0	Total
Individual j	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

Measure	Formula
S1: Matching coefficient	$s_{ij} = (a+d)/(a+b+c+d)$
S2: Jaccard coefficient (Jaccard, 1908)	$s_{ij} = a/(a+b+c)$
S3: Rogers and Tanimoto (1960)	$s_{ij} = (a+d)/[a+2(b+c)+d]$
S4: Sneath and Sokal (1973)	$s_{ij} = a/[a+2(b+c)]$
S5: Gower and Legendre (1986)	$s_{ij} = (a+d) \Big/ \left[a+\frac{1}{2}(b+c)+d\right]$
S6: Gower and Legendre (1986)	$s_{ij} = a \Big/ \left[a+\frac{1}{2}(b+c)\right]$

Proximity measures for mixed-type data

Gower's general similarity measure:

$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} \Big/ \sum_{k=1}^{p} w_{ijk}$$

For categorical variables, if the two objects have the same value for variable k, s_{ijk}=1, otherwise s_{ijk}=0.

For continuous variables, s_{ijk}=1-$|x_{ik}$-$x_{jk}|/R_k$, where R_k is the range of the kth variable.

The weights w can be very flexible to deal with missing values or specific requirements.

Clustering algorithms

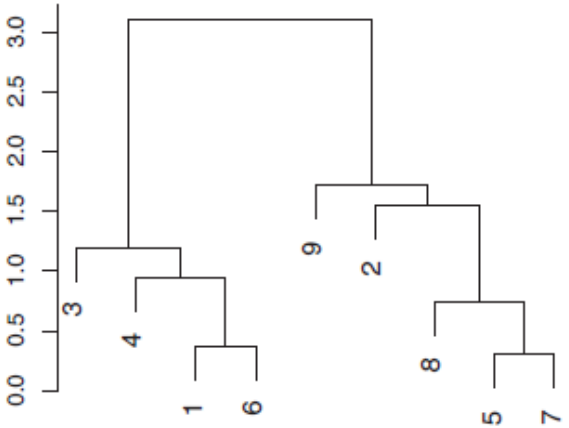
- Hierarchical clustering
- Optimization clustering
- Model-based clustering
- Density-based clustering

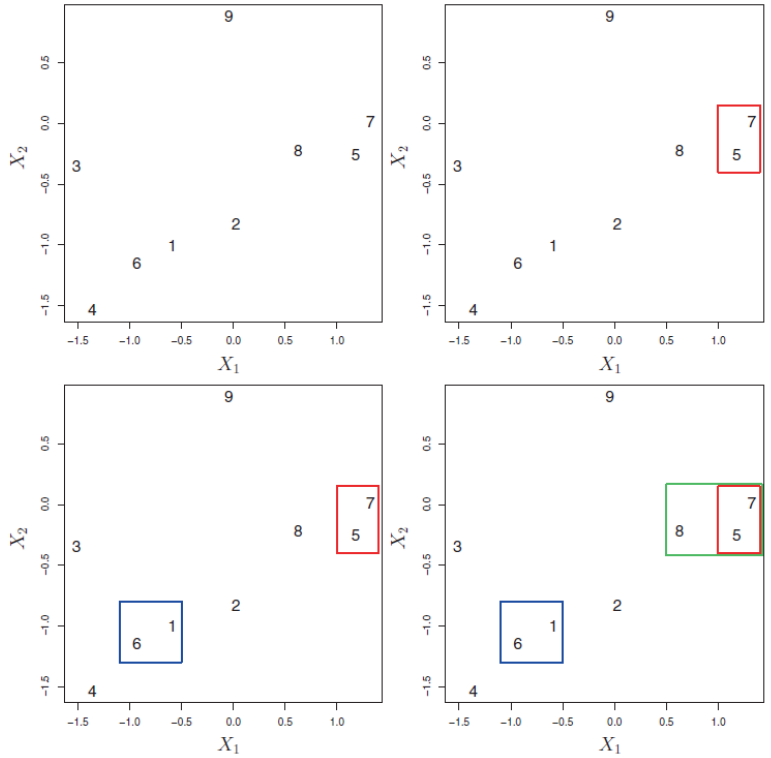
Hierarchical clustering

- Partition the data in a series of steps
- Agglomerative and divisive methods
- Produces a tree-shaped dendrogram

Agglomerative hierarchical clustering

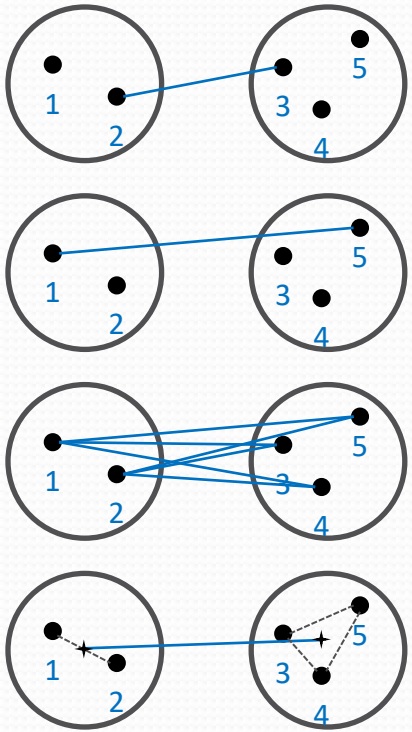
dendrogram





ITSL chapter 10

Distance between clusters: linkage



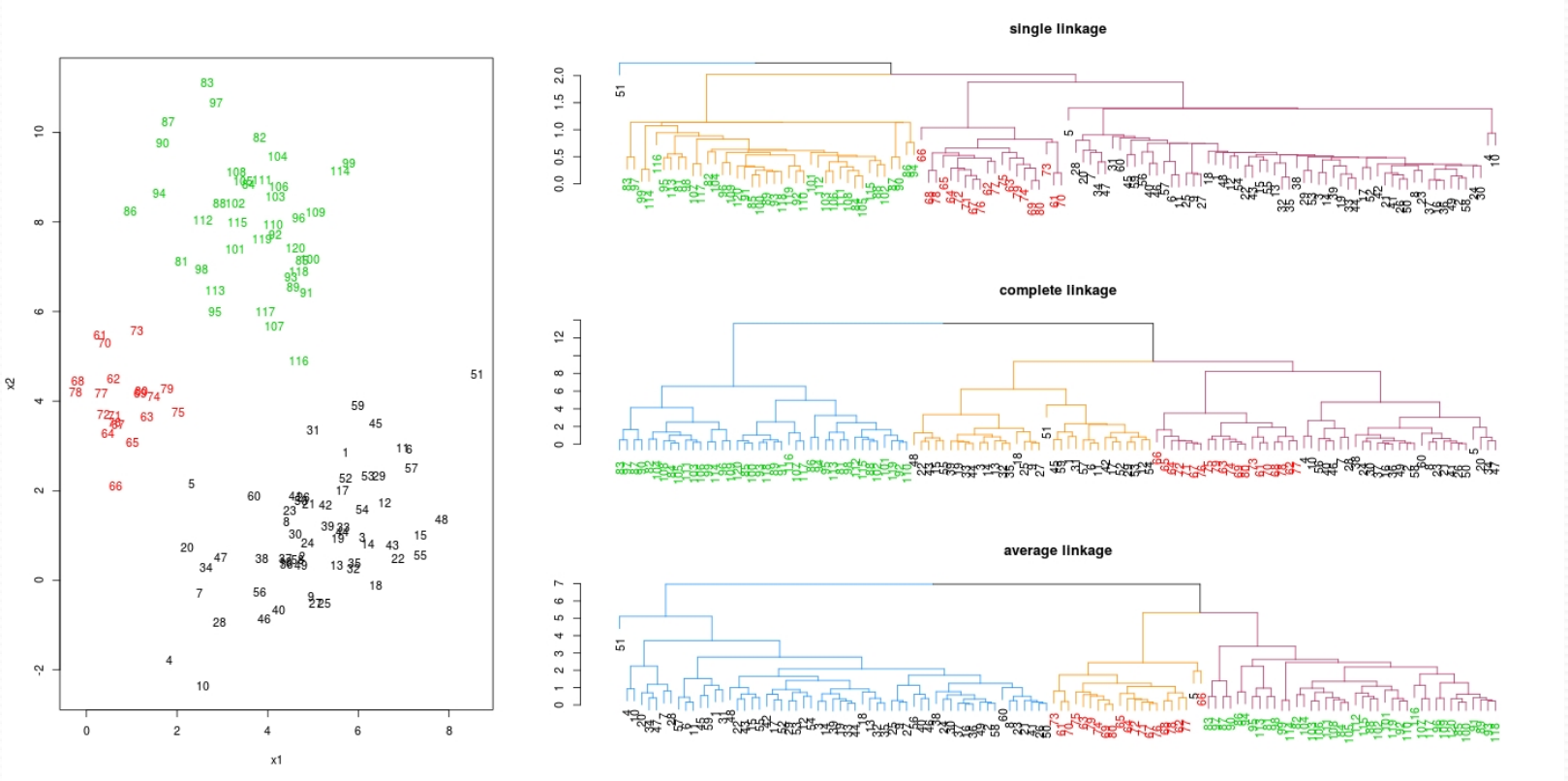
Single linkage (nearest neighbor):
d_{23}

Complete linkage (furthest neighbor):
d_{15}

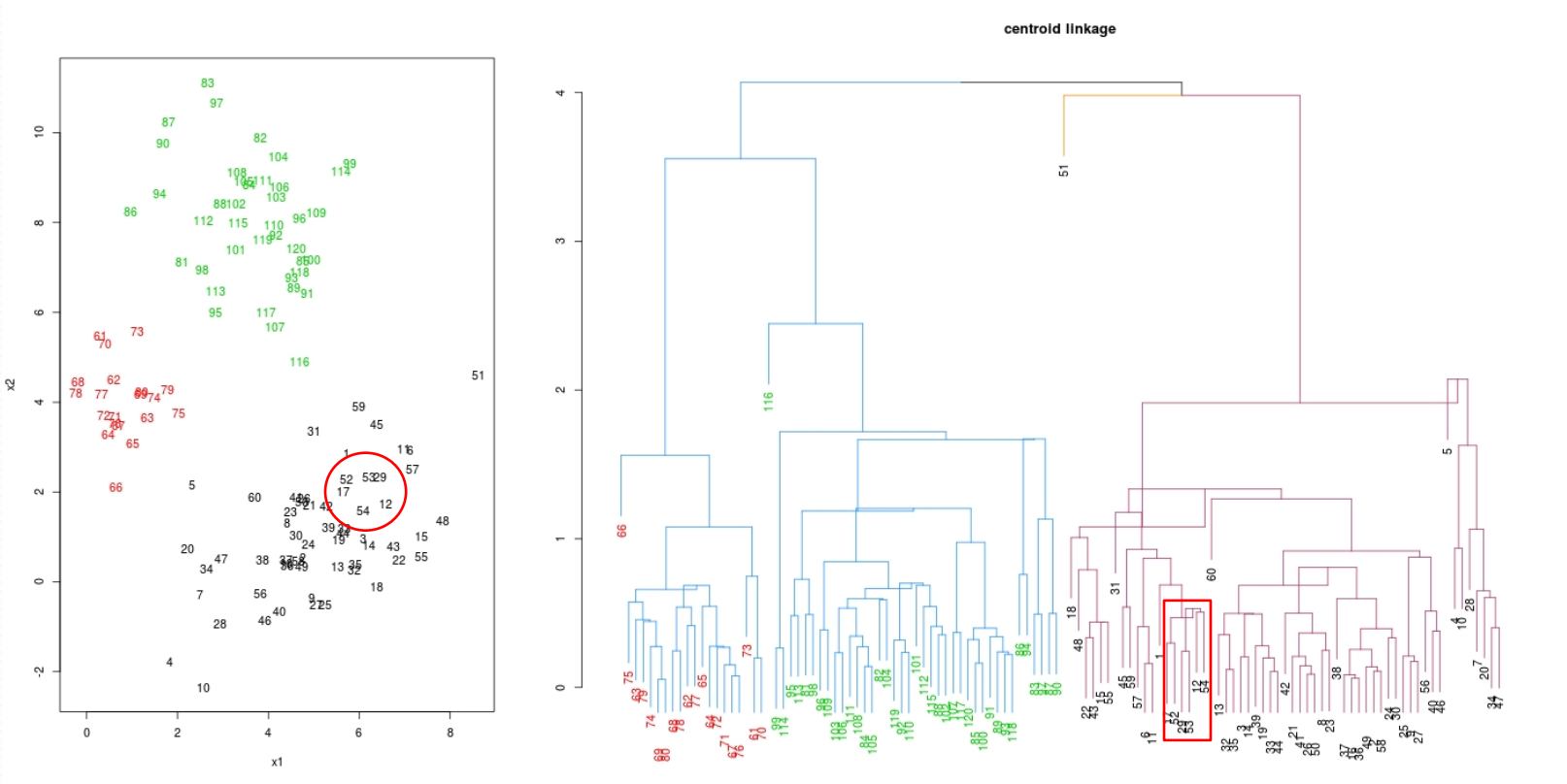
Average linkage (mean):
$(d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25})/6$

Centroid linkage:
Calculated from original data **X**

Linkage choice affects clustering results



Centroid linkage may lead to inversion



Summary of different linkage options

Method	Alternative name[a]	Usually used with:	Distance between clusters defined as:	Remarks
Single linkage Sneath (1957)	Nearest neighbour	Similarity or distance	Minimum distance between pair of objects, one in one cluster, one in the other	Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure.
Complete linkage Sorensen (1948)	Furthest neighbour	Similarity or distance	Maximum distance between pair of objects, one in one cluster, one in the other	Tends to find compact clusters with equal diameters (maximum distance between objects). Does not take account of cluster structure.
(Group) Average linkage Sokal and Michener (1958)	UPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	Tends to join clusters with small variances. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust.
Centroid linkage Sokal and Michener (1958)	UPGMC	Distance (requires raw data)	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged cluster. Subject to reversals.
Weighted average linkage McQuitty (1966)	WPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	As for UPGMA, but points in small clusters weighted more highly than points in large clusters (useful if cluster sizes are likely to be uneven).
Median linkage Gower (1967)	WPGMC	Distance (requires raw data)	Squared Euclidean distance between weighted centroids	Assumes points can be represented in Euclidean space for geometrical interpretation. New group is intermediate in position between merged groups. Subject to reversals.
Ward's method Ward (1963)	Minimum sum of squares	Distance (requires raw data)	Increase in sum of squares within clusters, after fusion, summed over all variables	Assumes points can be represented in Euclidean space for geometrical interpretation. Tends to find same-size, spherical clusters. Sensitive to outliers.

[a]U = unweighted; W = weighted; PG = pair group; A = average; C = centroid.

Summary of hierarchical clustering

- Performer has to choose proximity measure, clustering methods and the number of clusters (k).
- Being able to show hierarchy structures.
- Unsatisfactory grouping cannot be undone in later steps.
- May not perform well for large datasets.

Optimization clustering

- With any partition of the n data points into g groups, there could be an index $c(n, g)$ that measures the quality of the partition, which can be optimized.
- Usually $c(n, g)$ is related to within-group homogeneity or between-group separation.
- Produces the 'best' group assignment, but no hierarchical structure.

Optimization clustering

- In theory, one could find the ‘best’ clustering solution by going through all possible C and find the smallest within cluster scatter $W(C)$.
- In reality the number of possible C grows rapidly with increasing n, and this direct approach is not feasible.
- The best C will have to be approximated by some iterative algorithm.

K-means clustering

- If all *p* variables are quantitative and the dissimilarity is **squared** Euclidean distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{p} (x_{im} - x_{jm})^2 = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2$$

- Then $W(C)$ is the sum of distance between each member of the cluster and the mean vector of that cluster:

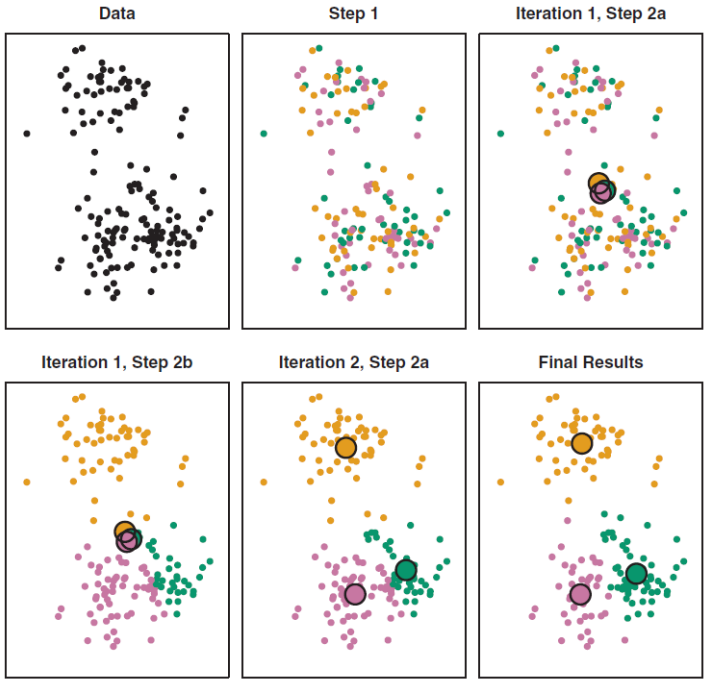
$$W(C) = \frac{1}{2} \sum_{g=1}^{k} \sum_{\substack{C(i)=g \\ C(j)=g}} \left\| \mathbf{x_i} - \mathbf{x_j} \right\|^2 = \sum_{g=1}^{k} n_k \sum_{C(i)=g} \left\| \mathbf{x_i} - \bar{\mathbf{x}}_k \right\|^2$$

K-means clustering

$W(C)$ can be iteratively improved by the following algorithm, but may converge on a local minimum. Different random initial assignments may produce different results.

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



ITSL chapter 10

K-medoids

- In the k-means algorithm, the problem of minimizing within group scatter is transformed into minimizing the distance between each member and a cluster center (in this case the centroid).
- This approach can be extended to other dissimilarity measure and other choice of cluster centers.
- K-medoids is a robust extension of the K-means, in which a representative data point is chosen as the cluster center. Optimization can be based directly on dissimilarity matrix.

K-medoids

Algorithm 14.2 *K-medoids Clustering.*

1. For a given cluster assignment C find the observation in the cluster minimizing total distance to other points in that cluster:

$$i_k^* = \underset{\{i:C(i)=k\}}{\operatorname{argmin}} \sum_{C(i')=k} D(x_i, x_{i'}). \tag{14.35}$$

   Then $m_k = x_{i_k^*}$, $k = 1, 2, \ldots, K$ are the current estimates of the cluster centers.

2. Given a current set of cluster centers $\{m_1, \ldots, m_K\}$, minimize the total error by assigning each observation to the closest (current) cluster center:

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} D(x_i, m_k). \tag{14.36}$$

3. Iterate steps 1 and 2 until the assignments do not change.

ESL chapter 14

Summary of optimization clustering

- Search for an optimization of a clustering criterion, which depends on dissimilarity and within cluster homogeneity or between cluster separation.
- Methods like K-means and K-medoids are approximations, no guarantee of finding the global maximum.
- Only produces cluster assignments.
- Initial values of cluster centers affect final results.

Model-based clustering

- A parametric solution (probability density estimation).
- Assume that data points are sampled from a family of probability density functions of the form:

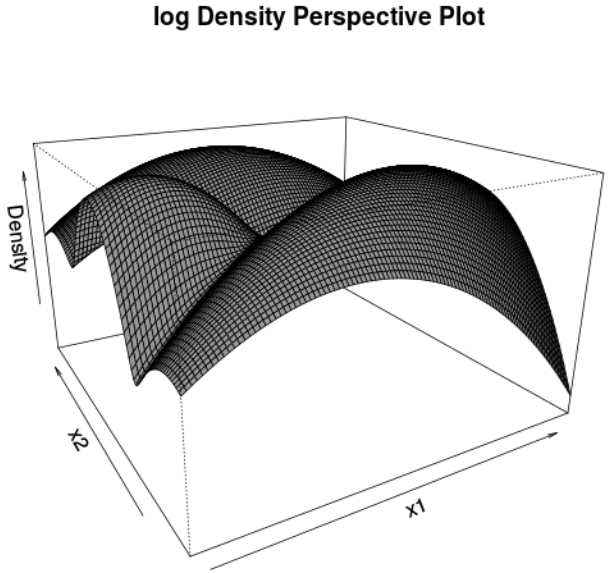
$$f(\mathbf{x}, \mathbf{p}, \boldsymbol{\theta}) = \sum_{g=1}^{k} p_g \, f_g(\mathbf{x}, \boldsymbol{\theta})$$

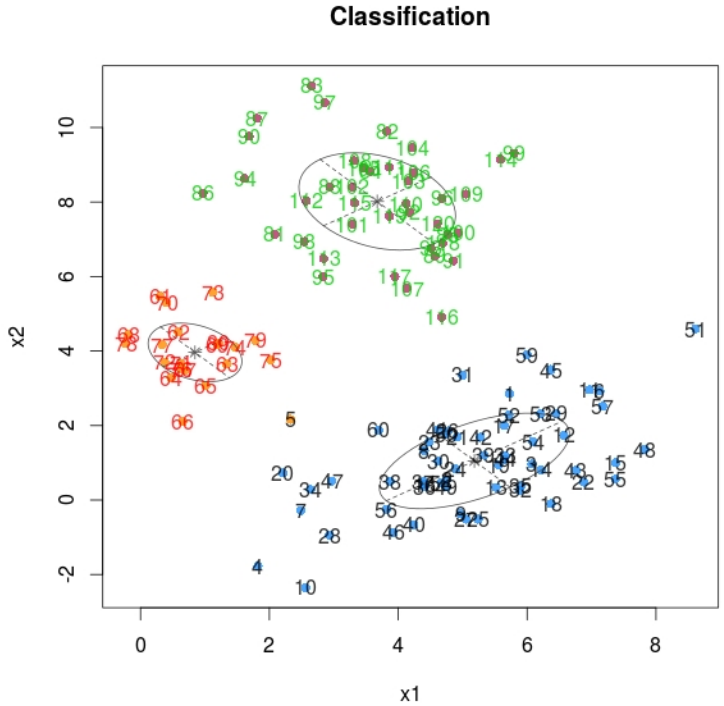
- Cluster assignment is based on the estimated posterior probability:

$$P(\text{cluster } g \mid \mathbf{x}_i) = \frac{\hat{p}_g \, f_g(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{f(\mathbf{x}_i, \hat{\mathbf{p}}, \hat{\boldsymbol{\theta}})}$$

- Parameters are estimated by maximum likelihood or Bayesian methods.

Gaussian mixture models example





How to determine *k*?

- Sometimes the number of clusters *k* is known *a priori*.
- Otherwise, there is no universal solution, but heuristics are available.
- There are a lot of indices that may help evaluating the 'quality' of clustering with different *k* (the `NbClust` package in R includes **30** of them…).

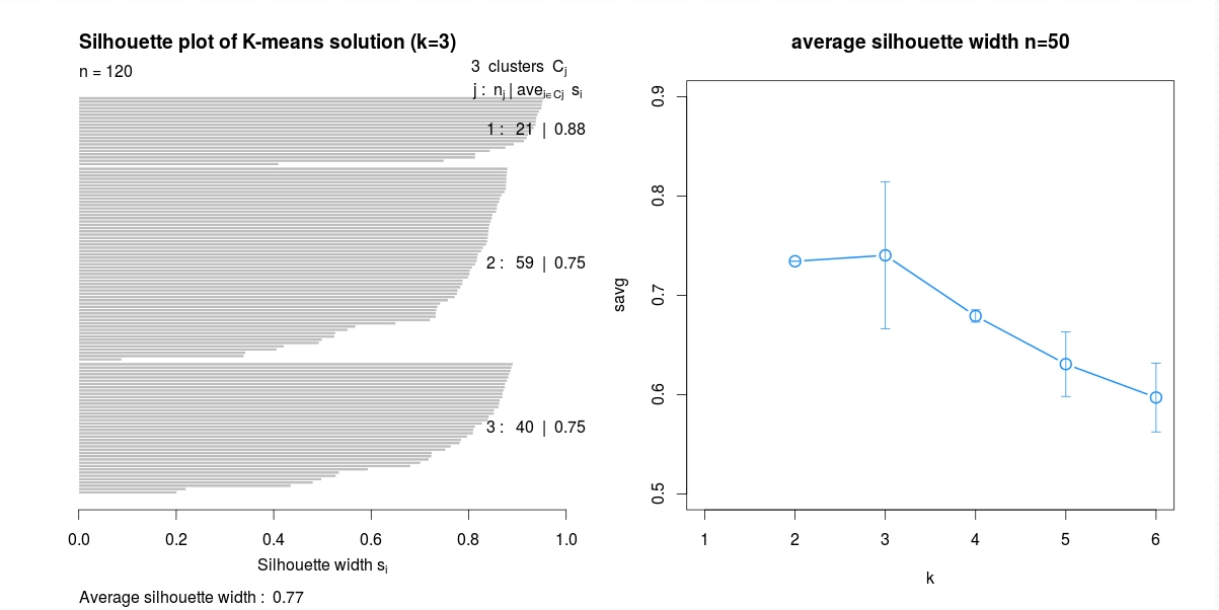
Silhouette analysis

- Silhouette measures the tightness and separation of the clustering solution.

$$a(i) = d(i, A)$$

$$b(i) = \min_{C \neq A} d(i, C)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

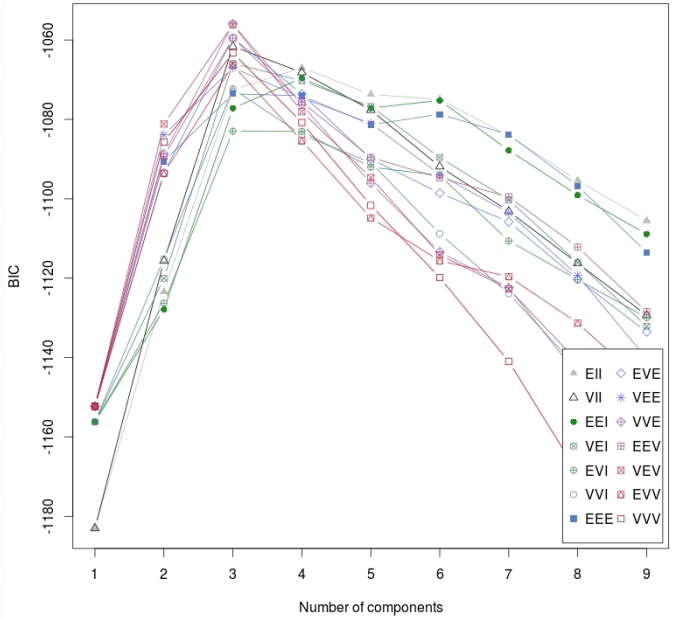


Determining *k* for finite mixture model

- Equivalent to a model selection problem.
- Akaike's information criterion (AIC) or Bayesian information criterion (BIC) are usually used.
- AIC and BIC are penalized likelihood.

$$\mathrm{AIC} = -2LL(x, \hat{\theta}) + 2p$$

$$\mathrm{BIC} = -2LL(x, \hat{\theta}) + p\ln(n)$$



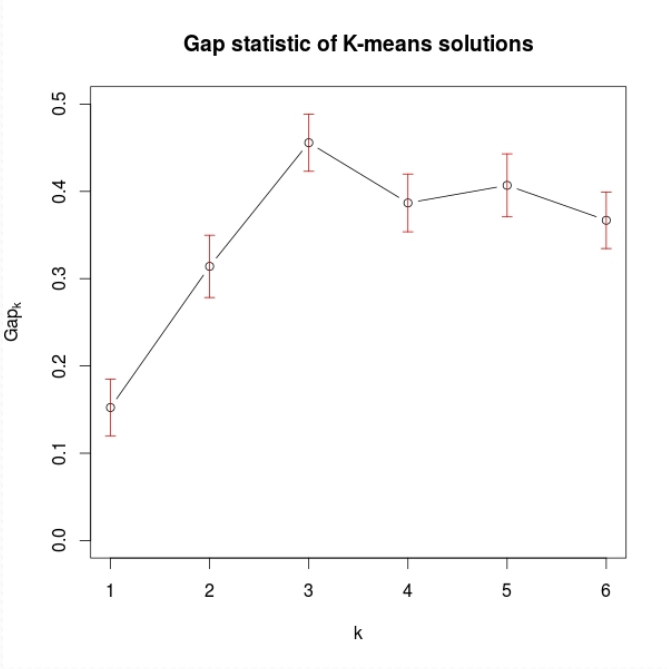
The Gap statistic

- Normalize the within cluster scatter criterion over a reference distribution.

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,j \in C_r} d_{ij}$$

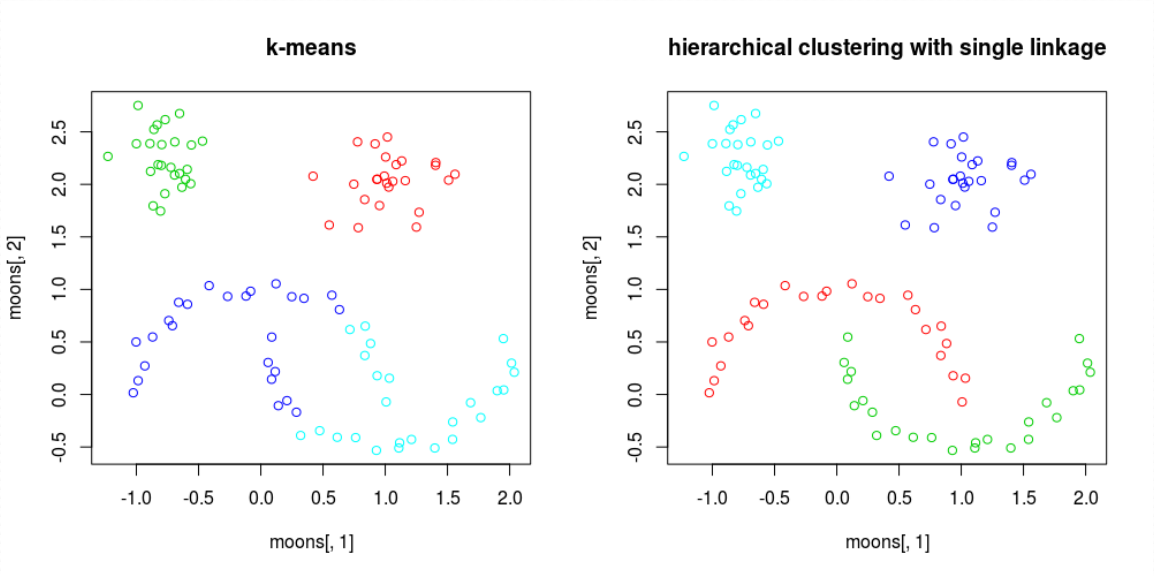
$$\mathrm{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

- Reference distribution obtained by Monte Carlo sampling of a uniform distribution over a box aligned with the principle component of the data.



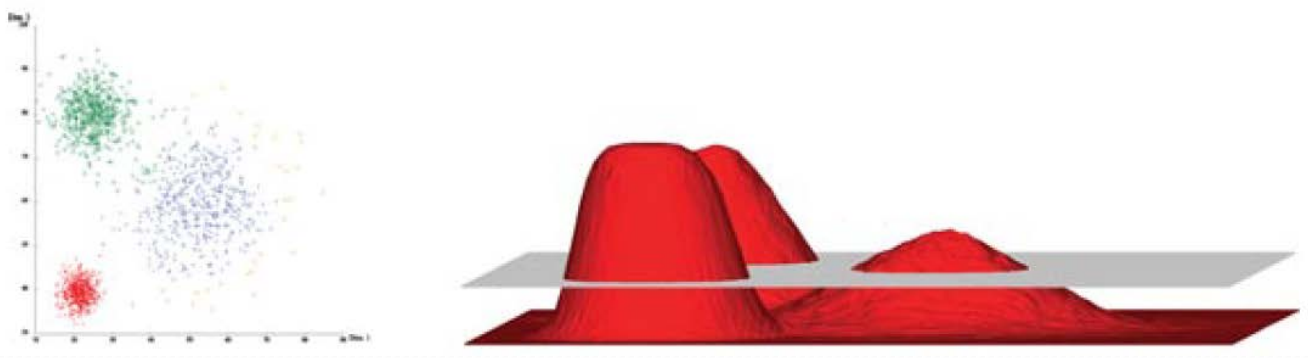
Density-based clustering

- Separation criterion tends to identify convex clusters
- Certain scenarios favor non-convex shaped clustering
- Chaining effect isn't always a bad thing!



Density-based clustering

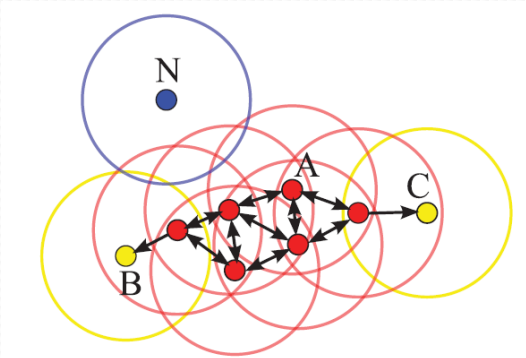
- Density-based clustering can be imaged as cutting through the probability density function.
- Points with low density are not assigned to any clusters.



Kriegel et al. (2011). Density-based clustering. *WIREs Data Mining and Knowledge Discovery*

Density-based spatial clustering of applications with noise (DBSCAN)

- Parameters to determine densities around each points: ε and *minPts*
- Core, border and noise points
- Clustering based on density-reachability

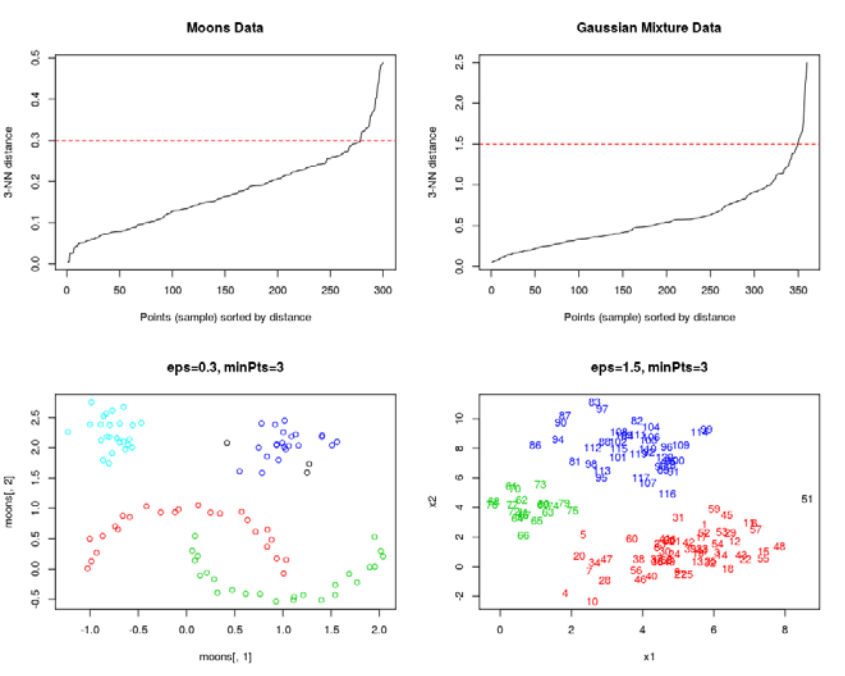


ALGORITHM 2: Abstract DBSCAN Algorithm

```
1 Compute neighbors of each point and identify core points      // Identify core points
2 Join neighboring core points into clusters                    // Assign core points
3 foreach non-core point do
4     Add to a neighboring core point if possible               // Assign border points
5     Otherwise, add to noise                                   // Assign noise points
```

Schubert et al. (2015). DBSCAN revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*.

DBSCAN clustering



- For low-dimensional data, *minPts* can be heuristically determined as some $k \geq p+1$, where p is the dimensionality of the data.
- ε can then be determined from the elbow in the *k*th-nearest neighbor distance plot

Hahsler et al. dbscan: Fast Density-based Clustering with R. https://cran.r-project.org/web/packages/dbscan/vignettes/dbscan.pdf

General remarks

- Proximity measure has to be defined with caution.
- Choice of clustering methods depends on proximity (e.g., K-means only for squared Euclidean).
- Determining the number of *k* is a hard problem, but can be guided by heuristics.
- Density-based clustering may help identify non-convex clusters, but one still have to be careful about parameter selection.

Recommended readings

- *An Introduction to Statistical Learning: With Applications in R*. James G, Witten D, Hastie T, Tibshirani R. Springer: 2013. Chapter 10.
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Hastie T, Tibshirani R, Friedman J. Springer: 2011. Chapter 14.
- *Cluster analysis* (5ed). Everitt BS, Landau S, Leese M, Stahl D. Wiley: 2011.