STATS 415: Introduction to Clustering

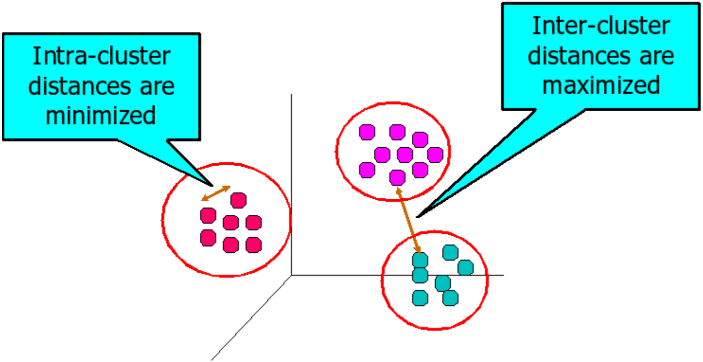
Prof. Liza Levina

Department of Statistics, University of Michigan

Goal of clustering

Given a data set, find meaningful groups of the objects such that

- The objects in one group are similar to one another
- The objects in one group are different from objects in other groups



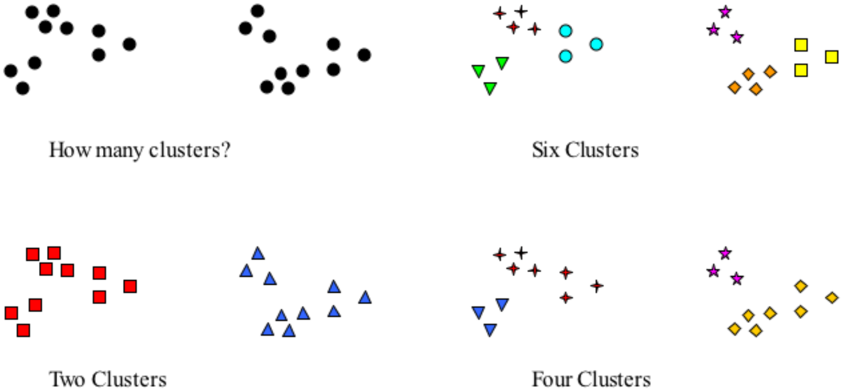
Some applications of clustering

- Understanding the data
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- Summarizing the data
  - Reduce the size of large data sets

What is NOT clustering

- Classification: classes pre-determined and training labels available
- Simple segmentation (e.g., divide the class roster into groups by last name alphabetically)
- Data querying (e.g. find all students in 415 who are data science majors): a result of an external specification

Difficulties in defining meaningful clusters



Evaluating clustering

- For supervised problems, such as classification, we have clear measures of success (e.g., classification error).
- For cluster analysis, which is unsupervised, there is no such measure, and generally clusters are ”in the eye of the beholder”.
- Why do we need measures of quality?
  - To avoid finding patterns in noise
  - To compare two sets of clusters

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis ... Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

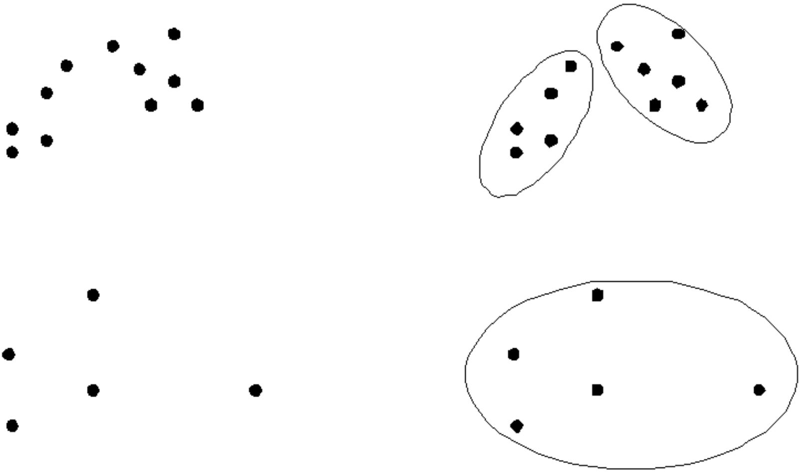
Algorithms for Clustering Data, by Jain and Dubes (1988).

Some measures exist, but in general there is still no good answer.

Types of clustering methods

- Partition methods:
  objects are partitioned into non-overlapping groups and each object is in exactly one group
- Hierarchical methods:
  objects are partitioned into nested groups organized as a hierarchical tree

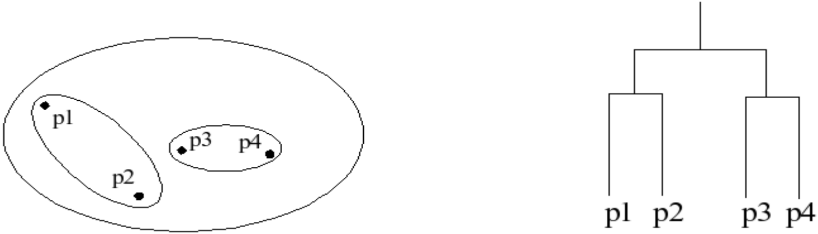
Toy example: partitional clustering



Original Points **A Partitional Clustering**

Toy example: hierarchical clustering

Traditional Hierarchical Clustering



Clustering Algorithms

- K-means
- Hierarchical clustering
- Density-based clustering (time permitting)
- Model-based clustering (time permitting)

Types of input for clustering

- The data matrix (n objects, p variables)
- Dissimilarity matrix ($n \times n$): only have information on how "dissimilar" objects are.
- Similarity matrix ($n \times n$): information on how "similar" the objects are.

Dissimilarity measures

- The lower the value, the more similar the objects are
- Non-negative: $\delta(x,y) \geq 0$
- The lowest value is 0, and $\delta(x,x) = 0$.
- Symmetric: $\delta(x,y) = \delta(y,x)$
- Distance, or metric: a dissimilarity measure that satisfies the triangle inequality,

$$\delta(x,y) \leq \delta(x,z) + \delta(y,z)$$

- Minkowski distance, or ℓ_q-distance: $\|x-y\|_q = \left[\sum_{k=1}^{p} |x_k - y_k|^q\right]^{1/q}$
  $q = 2$: Euclidean distance
  $q = 1$: ℓ_1, or Manhattan distance

Similarity measures

- The higher the value, the more similar the objects are; but not required to be positive.
- Symmetric: $s(x,y) = s(y,x)$.
- In many cases, normalized to have a [0,1] range
- Correlation coefficient, cosine, inner product, etc

Similarity/dissimilarity measures for mixed variables

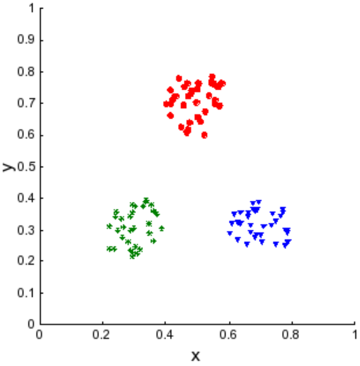
- Dissimilarity can be defined for categorical variables (usually 0/1).
- Many data sets have mixed variables (categorical, ordinal, numerical).
- Can define overall dissimilarity by scaling individual ones to take values in $[0, 1]$, and taking a (potentially weighted) average:

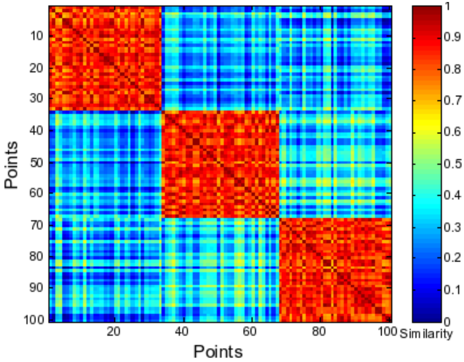
$$\delta(x,y) = \frac{\sum_{j=1}^{p} w_j \delta_j(x_j, y_j)}{\sum_{j=1}^{p} w_j}$$

- Allowing w_j to depend on the data can handle missing data by setting $w_j = 0$ if the j-th variable is missing

Visualizing the similarity matrix

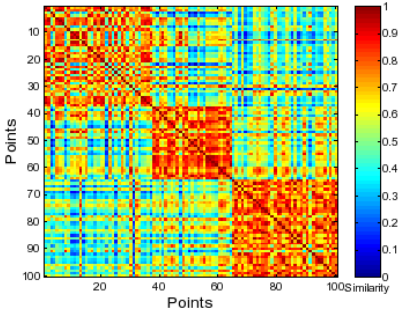
Order the nodes by cluster labels

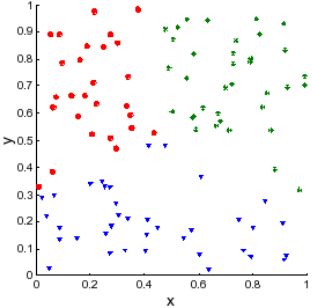




Comparing similarity matrices

Clusters obtained from data uniformly distributed in a square are not so “crisp”





Cohesion and separation

- Cluster cohesion: measures how closely related objects are within clusters
- Cluster separation: measures how distinct or well-separated different clusters are
- Usually measured by distance to centroids and between centroids, using the same distance measure that was used to construct clusters
- There is a trade-off between cohesion and separation (see p. 5)

Example: sum of squared errors

- C_k, $k = 1, \ldots, K$ is the set of points assigned to cluster k
- $n_k = |C_k|$ be the number of points in cluster k
- m_k is the centroid of class k (usually the mean)
- m be the centroid of the entire dataset
- Within cluster Sum of Squared errors, or WSS (cohesion)

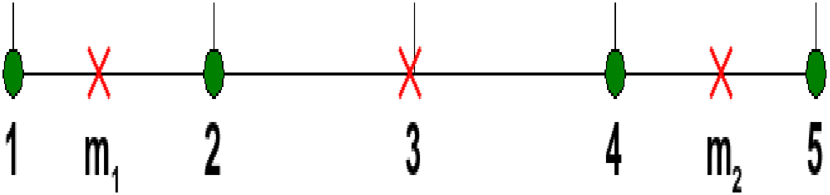
$$\text{WSS} = \sum_k \sum_{x \in C_k} \|x - m_k\|^2$$

- Between cluster Sum of Squared errors, or BSS (separation)

$$\text{BSS} = \sum_k n_k \cdot \|m - m_k\|^2$$

Toy example: Cohesion and separation

k=2
WSS=(1-1.5)^2+(2-1.5)^2+(4-4.5)^2+(5-4.5)^2=1
BSS=2(1.5-3)^2+2(4.5-3)^2=9
TSS=1+9



k=1
m1=3 m=3 n=4
WSS=(1-3)^2+(2-3)^2+(4-3)^2+(5-3)^2
=10
BSS=0

Trade-off between cohesion and separation

TSS = BSS + WSS = Constant

- $K = 1$ cluster

$$\begin{aligned} \text{WSS} &= (1-3)^2+(2-3)^2+(4-3)^2+(5-3)^2 = 10 \\ \text{BSS} &= 4\times(3-3)^2 = 0 \\ \text{TSS} &= 10 \end{aligned}$$

- $K = 2$ clusters

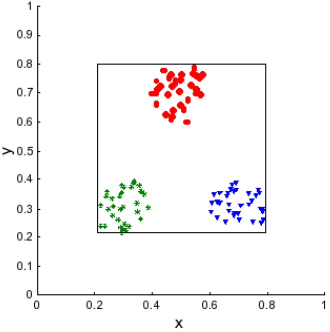
$$\begin{aligned} \text{WSS} &= (1-1.5)^2+(2-1.5)^2+(4-4.5)^2+(5-4.5)^2 = 1 \\ \text{BSS} &= 2\times(3-1.5)^2+2\times(4.5-3)^2 = 9 \\ \text{TSS} &= 10 \end{aligned}$$

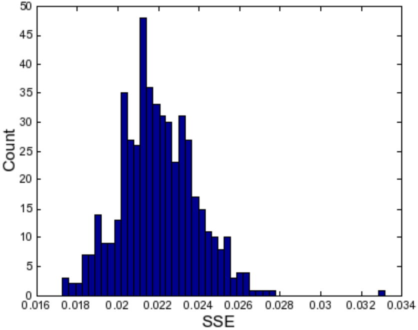
A possible framework for cluster validity

- How do we interpret these numbers? E.g. $\mathrm{WSS} = 9$: good, bad, average?
- The more "atypical" a clustering result is, the more likely it represents valid structure in the data.
- What is atypical? One possibility is to compare to clustering applied to uniformly distributed data.

Example: assessing "significance" of WSS

- Left: data, WSS = 0.005
- Right: draw a sample of the same size (100) uniformly over the range of 0.2 – 0.8 for x and y; repeat 500 times. Construct a histogram of 500 WSS values.
- Is 0.005 "typical"? Can formalize with a p-value.





The silhouette coefficient

- Combines cohesion and separation
- Let a_i be the average distance of object i to the other objects in the same cluster (cohesion)
- Let $d(i,k)$ be the average distance from object i to all objects in cluster k which does not contain i, and $b_i = \min_k d(i,k)$ (separation)
- The silhouette coefficient for point i is defined as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} = 1 - \frac{a_i}{b_i} \quad \text{if} \;\; a_i \leq b_i$$

Interpreting the silhouette coefficient

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- Always between -1 and 1
- The closer to 1, the better
- Large negative value $\Rightarrow$ poor clustering (“misclustered” point)
- Value close to $0 \Rightarrow$ ambiguous point
- Usually plotted as a bar chart grouped by cluster, ordered within cluster from best to worst

Example: silhouette plots

Two different methods applied to the same dataset