Lecture 25 Cluster Analysis II

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

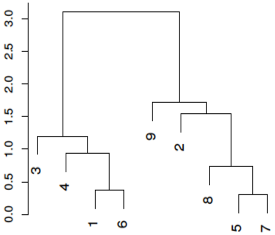
Department of Electrical and Computer Engineering
University of Alberta

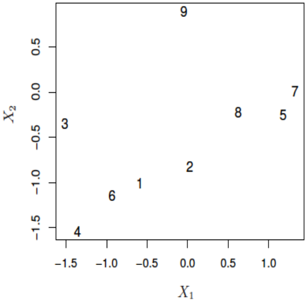
April 8, 2021

Outline

Another Example

- An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space. The raw data on the right was used to generate the dendrogram on the left.



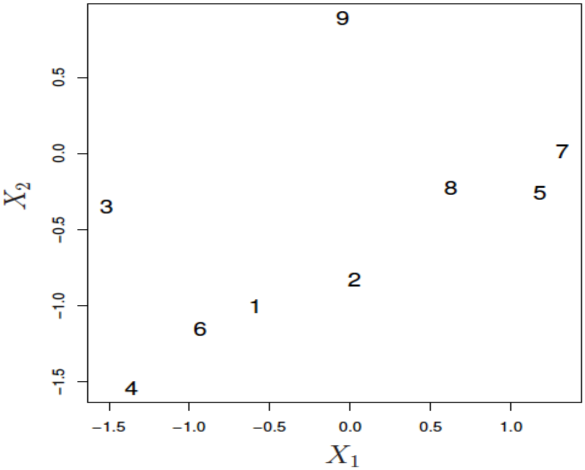


Another Example

- Observations 5 and 7 are quite similar to each other, as are observations 1 and 6.
- However, observation 9 is no more similar to observation 2 than it is to observations 8, 5, and 7, even though observations 9 and 2 are close together in terms of horizontal distance.
- This is because observations 2, 8, 5, and 7 all fuse with observation 9 at the same height, approximately 1.8.

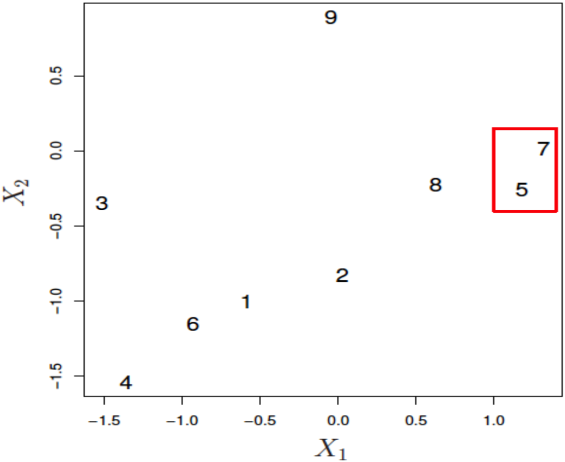
Another Example

Merges in previous example



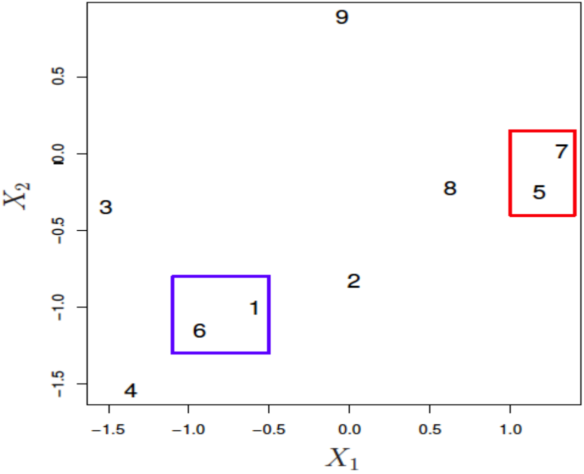
Another Example

Merges in previous example



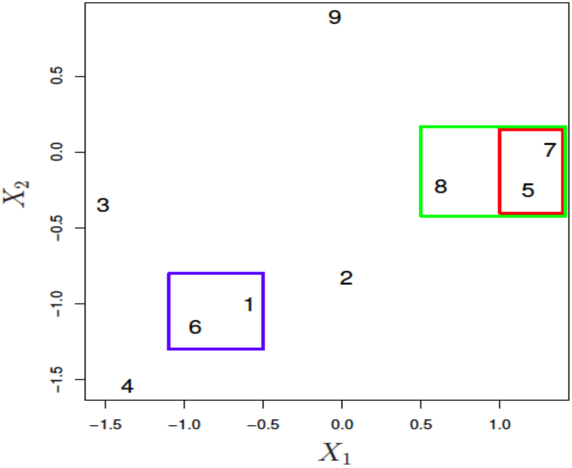
Another Example

Merges in previous example



Another Example

Merges in previous example

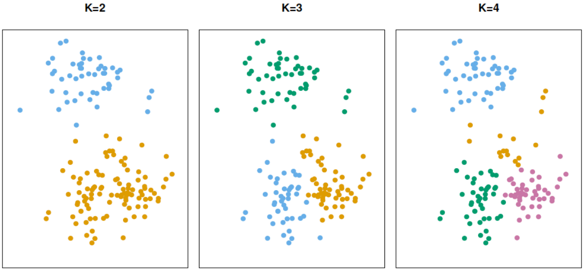


Types of Linkage

- Complete Linkage: Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
- Single Linkage: Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.
- Average Linkage: Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
- Centroid Linkage: Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B.

K-means Clustering

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- A simulated data set with 150 observations in 2-dimensional space.



K-means Clustering

- Panels show the results of applying K-means clustering with different values of K, the number of clusters.
- The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm.
- Note that there is no ordering of the clusters, so the cluster coloring is arbitrary.
- These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

K-means Clustering

- Let $C_1, \cdots, C_K$ denote sets containing the indices of the observations in each cluster. Theses satisfy two properties:
- 1. $C_1 \cup \cdots \cup C_K = \{1, \cdots, n\}$. In other words, each observation belongs to at least one of the K clusters.
- 2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.
- For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

K-means Clustering

- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- The within-cluster variation for cluster C_k is a measure $\mathrm{WCV}(C_k)$ of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

$$\min_{C_1,\cdots,C_K} \sum_{k=1}^{K} \{WCV(C_k)\}.$$

- In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

Within-cluster variation

- Typically we use Euclidean distance

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

where $|C_k|$ denotes the number of observations in the kth cluster.

- The optimization problem that defines K-means clustering is of the form

$$\min_{C_1,\cdots,C_K} \sum_{k=1}^{K} \left\{ \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

K-Means Clustering Algorithm

- 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).

K-Means Clustering Algorithm

- This algorithm is guaranteed to decrease the value of the objective at each step. Why?
- Note that

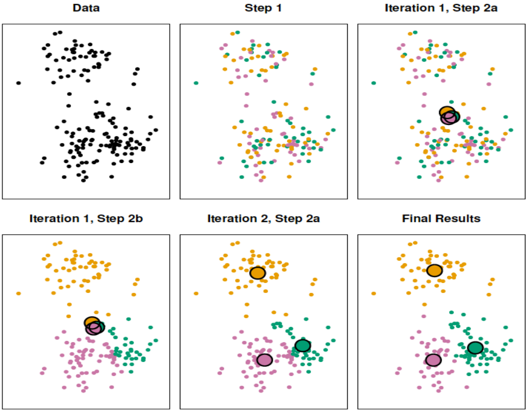
$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2,$$

where $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ is the mean for feature j in cluster C_k.

- In Step 2(a) the cluster means for each feature are the constants that minimize the sum-of-squared deviations.
- In Step 2(b), reallocating the observations can only reduce the objective value.
- However, K-Means is not guaranteed to produce the global minimum. Why not?

Example

The progress of the K-means algorithm with $K = 3$ with 10 iterations.



Example

Different starting values and above each plot is the value of the objective. Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation.

Practical issues

- Should the observations or features first be standardized in some way? For instance, maybe the variables should be centered to have mean zero and scaled to have standard deviation one.
- In the case of hierarchical clustering, What dissimilarity measure should be used? What type of linkage should be used?
- How many clusters to choose? (in both K-means or hierarchical clustering). Difficult problem. No agreed-upon method.

Summary and Remark

- Hierarchical clustering
- K-means clustering
- Read textbook Chapter 14 and R code
- Do R lab