DSO530 Statistical Learning Methods

Lecture 9 : Unsupervised Learning part II

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
xint@marshall.usc.edu

Introduction

- This is the second part of Lecture 9; it covers clustering
- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set
- Seek to partition observations into distinct groups so that the observations within each group are similar to each other, while observations in different groups are different from each other
- What means to be "different" and "similar"? Often it is a domain specific issue
- Think about why clustering is not classification; recall fraud detection in credit card transactions vs. fraud detection in medicare expense
- Sometimes, people do PCA first and then do clustering in exploratory data analysis (EDA)
  - *PCA* looks to find a low-dimensional representation of the observations that explain a good fraction of the variance
  - *Clustering* looks to find homogeneous subgroups among the observations

y is yes/no there is an aligned interest in predicting fraud. therefore labels are good

you don't care about cost since you are insured.

Different Clustering Approaches

- There exist a great number of clustering methods
- In this section, we focus on *K-means clustering* and *hierarchical clustering*

  required

  - K-means clustering: partition the observations into a pre-specified number of clusters k clusters required.
  - not required. hierarchical clustering (optional): not know in advance how many clusters we want; end up with a tree-like visual representation of the observations, called a *dendrogram*

K-means Clustering

n obs into k non-overlaping parts.

non-empty (atleast 1 obs)

- Let $C_1, \cdots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:
  - $C_1 \cup C_2 \cdots \cup C_K = \{1, \cdots, n\}$
  - C_k and $C_{k'}$ have empty intersection, for different k and k'

criteria for clustering.

- Idea for k-means clustering: a good clustering is one for which the within-cluster variation is as small as possible
- $W(C_k)$: a measure of the amount by which the observations within a cluster differ from each other, defined by

squared distance b/w the i^{th} obs and the i'^{th} observation

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

cardinality of set C_k: no. of observations.

- K-means algorithm solves

$$minimize_{C_1, \cdots, C_K} \underbrace{\left\{ \sum_{k=1}^{K} W(C_k) \right\}}_{\text{make this small}}$$

Computationally very difficult to find ideal k value.

K-means Clustering - an illustration

more clusters is not always necessary.

K=2 or 3 best.

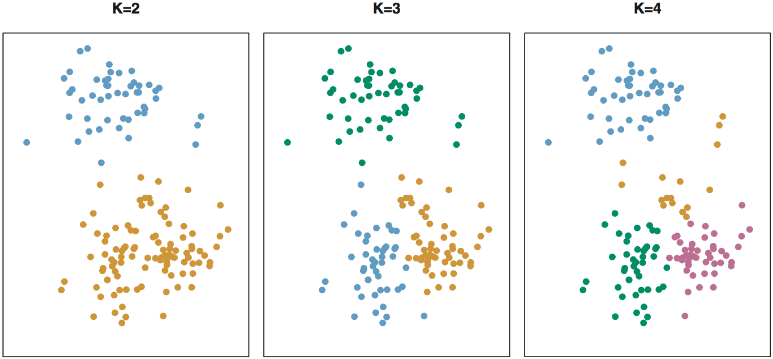


FIGURE 10.5. *A simulated data set with 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. 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 Algorithm

faulty initialization

this is a problem how to deal with it?

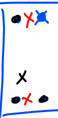
• multiple default assignment (10,20,etc)

• AKA multiple initialization.

Using kmeans:

2(a) { • black are randomly chosen to be in 1 cluster.
• 1st centroid is found for black & blue group.

example



red is the new & final centroids.

- Brutal force way is infeasible computationally
- The next algorithm finds an approximate solution

2(b) Assign pts to centroid. Therefore upper black turns blue and in next iteration centroid changes.

Algorithm 10.1 *K-Means Clustering*

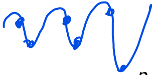
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of 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).

Why the Algorithm Works? (optional)

- Algorithm 10.1 is guaranteed to decrease the value of the following objective at each step

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

⇒ therefore, multiple initializations required



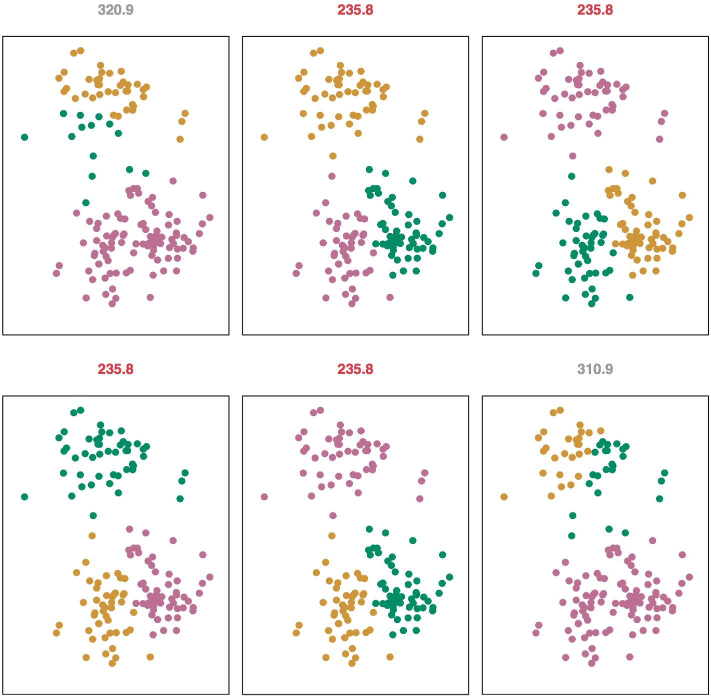
- To understand this, we need

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

kth cluster centroid.

ith obs

  - where $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$
- Because the K-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial (random) cluster assignment of each observation
- It is important to run the algorithm multiple times from different random initial configurations (which one do you choose in the end?)



320.9

235.8

235.8

235.8

235.8

310.9

Hierarchical Clustering (Optional)

- Hierarchical clustering is an alternative approach to K-means clustering which does not require that we commit to a particular choice of K
- In this section, we describe *bottom-up* or *agglomerative* clustering

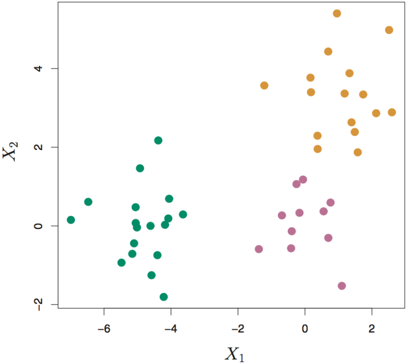


FIGURE 10.8. *Forty-five observations generated in two-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.*

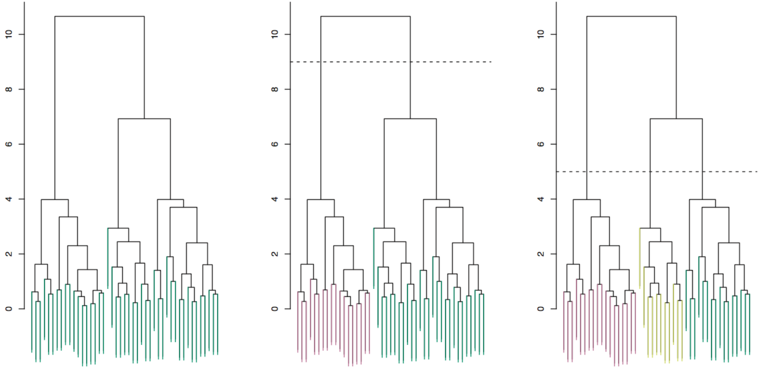


FIGURE 10.9. Left: *dendrogram obtained from hierarchically clustering the data from Figure 10.8 with complete linkage and Euclidean distance.* Center: *the dendrogram from the left-hand panel, cut at a height of nine (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.* Right: *the dendrogram from the left-hand panel, now cut at a height of five. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure.*

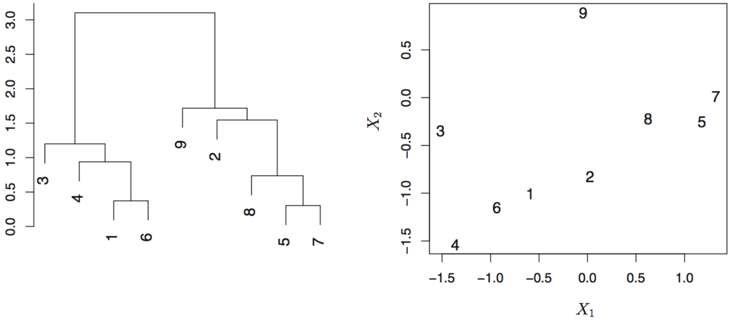


FIGURE 10.10. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6. *However, observation* 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. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* 8, 5, *and* 7.

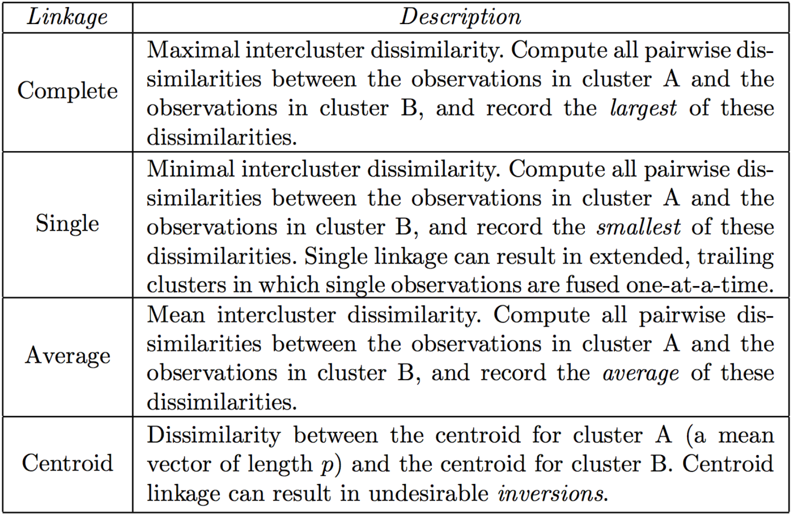
Algorithm 10.2 *Hierarchical Clustering*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.



Linkage	*Description*
Complete	Maximal intercluster 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	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster 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	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

TABLE 10.2. *A summary of the four most commonly-used types of linkage in hierarchical clustering.*

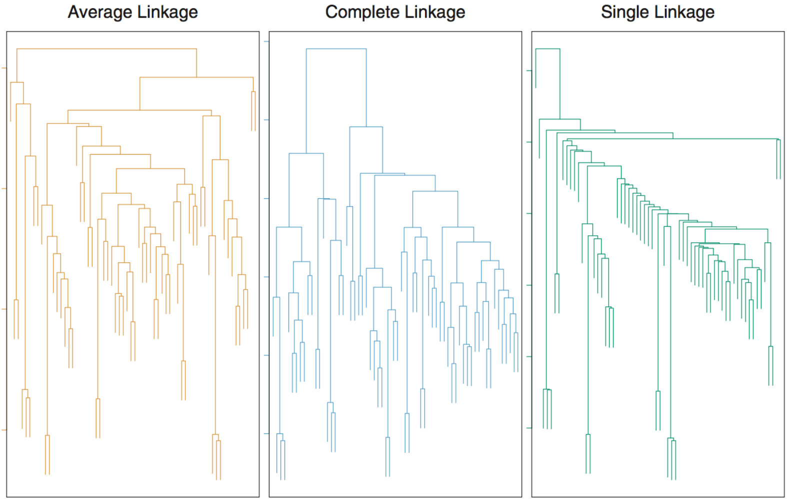


FIGURE 10.12. *Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.*

Practical Issues in Clustering

Depends on case by case

- Should the features first be rescaled in some way?
- In the case of K-means clustering, how many clusters should we look for in the data? Arguable
- In the case of hierarchical clustering (optional),
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendrogram in order to obtain clusters?
- Answers: With these methods, there is no single right answer–any solution that exposes some interesting aspects of the data should be considered
- In practice, we try several different choices, and look for the one with the most useful or interpretable solution
- Caution: clustering results should not be taken as the absolute truth about a data set. Should be used just for insights & interpretation

When k increases, inertia will always decrease.

This does not necessarily make a better model by simply ↑ k.