CSCI 550: Adv. Data Mining

07- Clustering- Representative-based

Announcement

- ACM and AWC will be holding a joint meeting to prepare students for the career fair this Tuesday (9/19) 5:00-7:00 PM in Barnard 347. They'll be holding practice interviews, talking about the companies that are attending, teaching about opportunities and how to prepare, as well as doing some casual resume reviews.
- Mini project 1 is posted and due September 30th.

The Goal of Unsupervised Learning

- The goal of clustering is to discover interesting things about the unlabeled measurements
- Can we discover subgroups among the variables or among the observations?
- In Unsupervised learning, we observe only the features because we do not have an associated response variable Y.
- We already discuss principal components analysis, as an unsupervised learning tool used for reducing data dimensionality and visualization of data

The Challenge of Unsupervised Learning

- Unsupervised learning is more subjective as there is no simple goal for the analysis.
- But techniques for unsupervised learning are of growing importance in a number of fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements
  - groups of shoppers characterized by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie viewers.

Advantage of Unsupervised learning

- It is often easier to obtain unlabeled data from a lab instrument or a computer than labeled data, which can require human intervention.
- For example it is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other.
- The definition is broad and vague.
- To make this concrete, we must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

Clustering vs. PCA

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

Clustering for Market Segmentation

- Suppose we have access to a large number of measurements (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
- Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.
- The task of performing market segmentation amounts to clustering the people in the data set.

Clustering Techniques

- Representative-based methods (i.e. K-means)
- Density-based methods (i.e. DBSCAN)
- Hierarchical methods (i.e. Agglomerative)
- Graph-based methods (i.e. Spectral clustering)

Representative-based Clustering

- Given a dataset D with n points x_i in a d-dimensional space, and given the number of desired clusters k, the goal of representative-based clustering is to partition the dataset into k groups or clusters.
- For each cluster C_i there exists a representative point that
- summarizes the cluster, a common choice being the mean (also called the centroid) μ_i of all points in the cluster

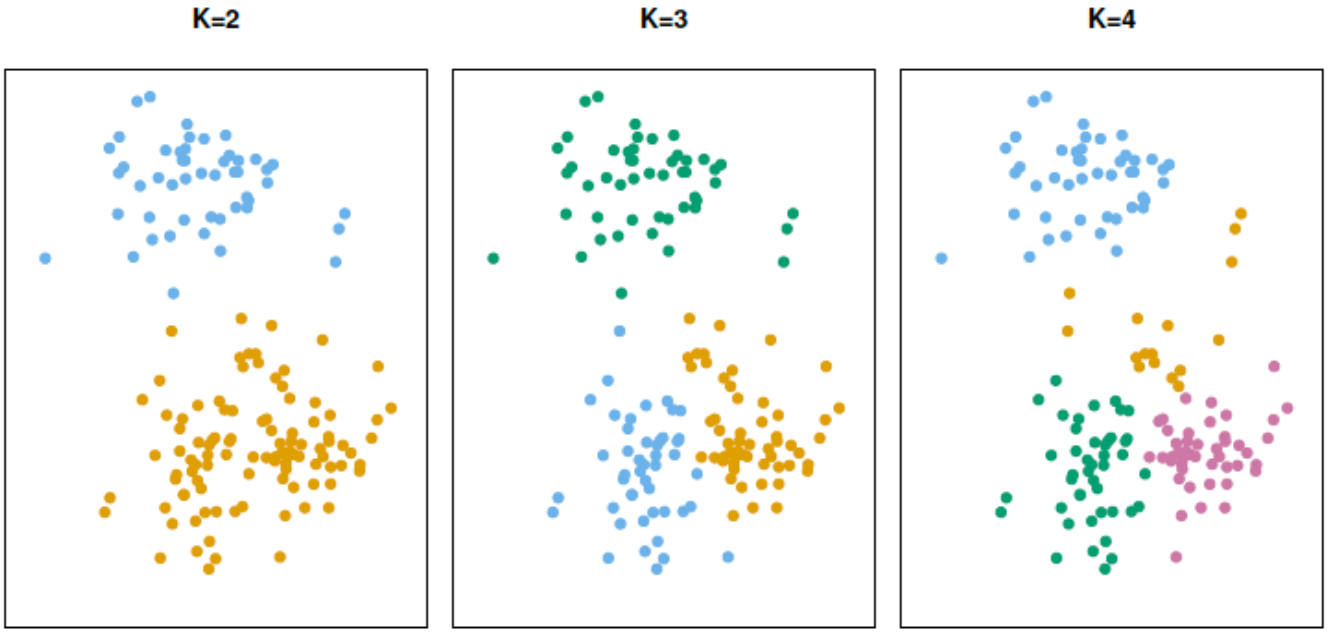
K-means clustering

- The idea behind K-means clustering is that a good clustering is one for which the *within-cluster variation* is as small as possible.
- Hence we want to solve the problem:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\}$$

- 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.

K-means clustering

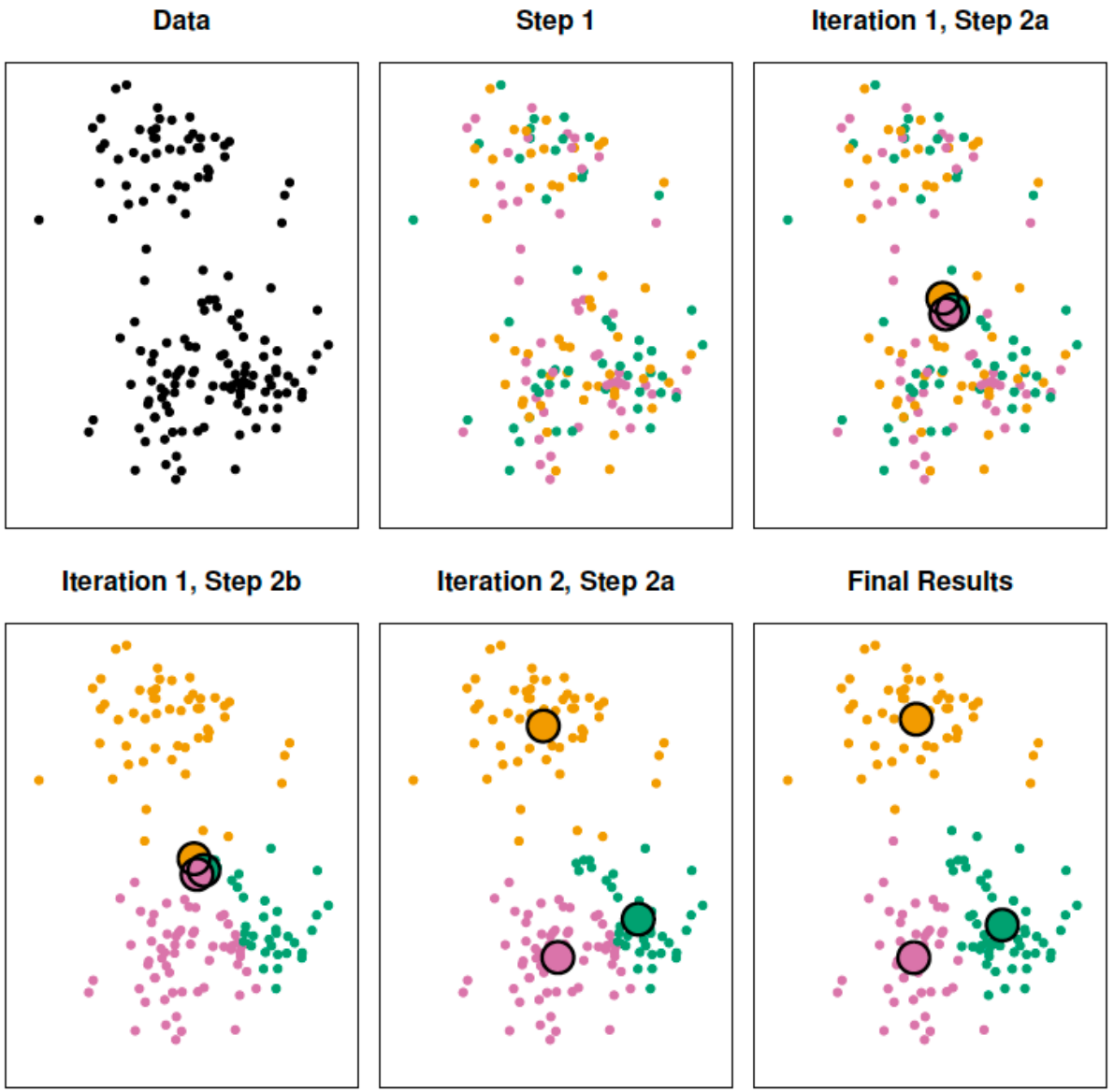


A simulated data set with 150 observations in 2-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm.

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 or when the updates to centroids are small enough:
   i. For each of the K clusters, compute the cluster centroid. The k^{th} cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   ii. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Example


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Details of Previous Figure

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

- Top left: The observations are shown.
- Top center: In Step 1 of the algorithm, each observation is randomly assigned to a cluster.
- Top right: In Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
- Bottom left: In Step 2(b), each observation is assigned to the nearest centroid.
- Bottom center: Step 2(a) is once again performed, leading to new cluster centroids.
- Bottom right: The results obtained after 10 iterations.

Example: different starting values

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 in Step 1

Details of Previous Figure

- K-means clustering performed six times on the data from previous figure with K = 3, each time with a different random assignment of the observations in Step 1 of the K-means algorithm.
- Above each plot is the value of the objective function
- Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation between the clusters.
- Those labeled in red all achieved the same best solution, with an objective value of 235.8