UNSUPERVISED LEARNING

SUPERVISED VS UNSUPERVISED LEARNING

- In supervised learning methods such as regression and classification we observe both a set of features $X_1, X_2, \ldots, X_p$ for each object, as well as a response or outcome variable Y. The goal is then to predict Y using $X_1, X_2, \ldots, X_p$.
- Now, we explore unsupervised learning, where we observe only the features $X_1, X_2, \ldots, X_p$. We are not interested in prediction, because we do not have an associated response variable Y.

THE GOALS OF UNSUPERVISED LEARNING

- The goal is statistical inference:
  - to discover interesting things about the measurements.
  - to find an informative way to visualize the data
  - can we discover subgroups among the variables or among the observations?

THE CHALLENGE OF UNSUPERVISED LEARNING

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
- But techniques for unsupervised learning are of growing importance in a number of fields:
  - subgroups of 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.

REALITY 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?

UNSUPERVISED LEARNING

- Unsupervised learning is important for understanding the variation and grouping structure of a set of unlabeled data, and can be a useful pre-processor for supervised learning
- It is more difficult than supervised learning because there is no gold standard (like an outcome variable) and no single objective (like test set accuracy)
- It is an active field of research, with many recently developed tools such as self-organizing maps, independent components analysis and spectral clustering

DATA CLUSTERING

CLUSTERING

- Clustering refers to a set of techniques for finding subgroups, or clusters, in a data set
- A good clustering is one when the observations within a group are similar but between groups are different
- 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 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

TWO CLUSTERING METHODS

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters
- In hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clustering obtained for each possible number of clusters, from 1 to n

PCA VS CLUSTERING

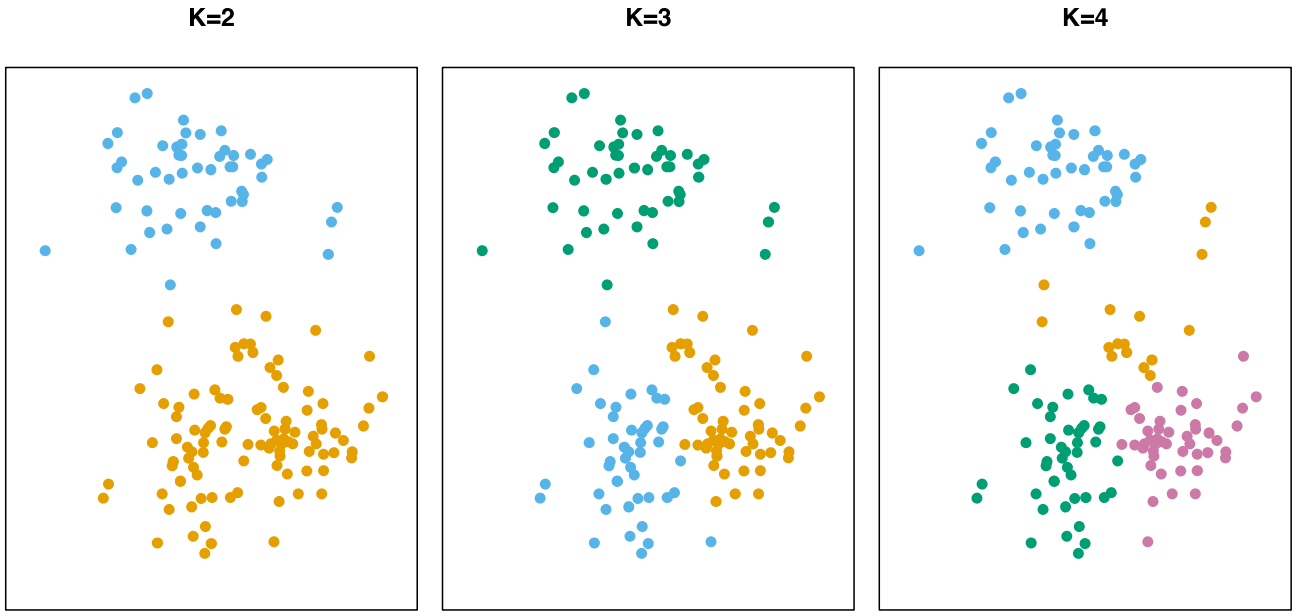
- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance
- Clustering looks for homogeneous subgroups among the observations

K-MEANS CLUSTERING

INTRODUCTION

- Determine K – the number of clusters
- By $C_1, \ldots, C_K$ denote the clusters satisfying two properties:
  - Each observation belongs to at least one of the K clusters
  - The clusters are non-overlapping: no observation belongs to more than one cluster

EXAMPLE



- A simulated data set with 150 observations in 2-dimensional space
- K-means algorithm will assign each observation to exactly one of the K clusters

WITHIN CLUSTER VARIATION

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

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{j=1}^{K} WCV(C_j) \right\}$$

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

WITHIN-CLUSTER VARIATION

- Typically we use Euclidean distance

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

- This gives the optimization problem that defines K-means clustering

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

K-MEANS ALGORITHM

1. Randomly assign each observation to one of K clusters
2. Iterate until the cluster assignments stop changing:
   a) For each of the K clusters, compute the cluster centroid, where the k^{th} cluster centroid is the mean of the observations assigned to the k^{th} cluster
   b) Assign each observation to the cluster whose centroid is the closest (where "closest" is defined using Euclidean distance)

K-MEANS ALGORITHM

- However it is not guaranteed to give the global minimum
- When the result no longer changes, a *local optimum* has been reached
- The results obtained will depend on the initial (random) cluster assignment of each observation in the initial step
- For this reason, it is important to run the algorithm multiple times from different random initial configurations
- Then one selects the *best* solution, i.e. that for which the objective is the smallest

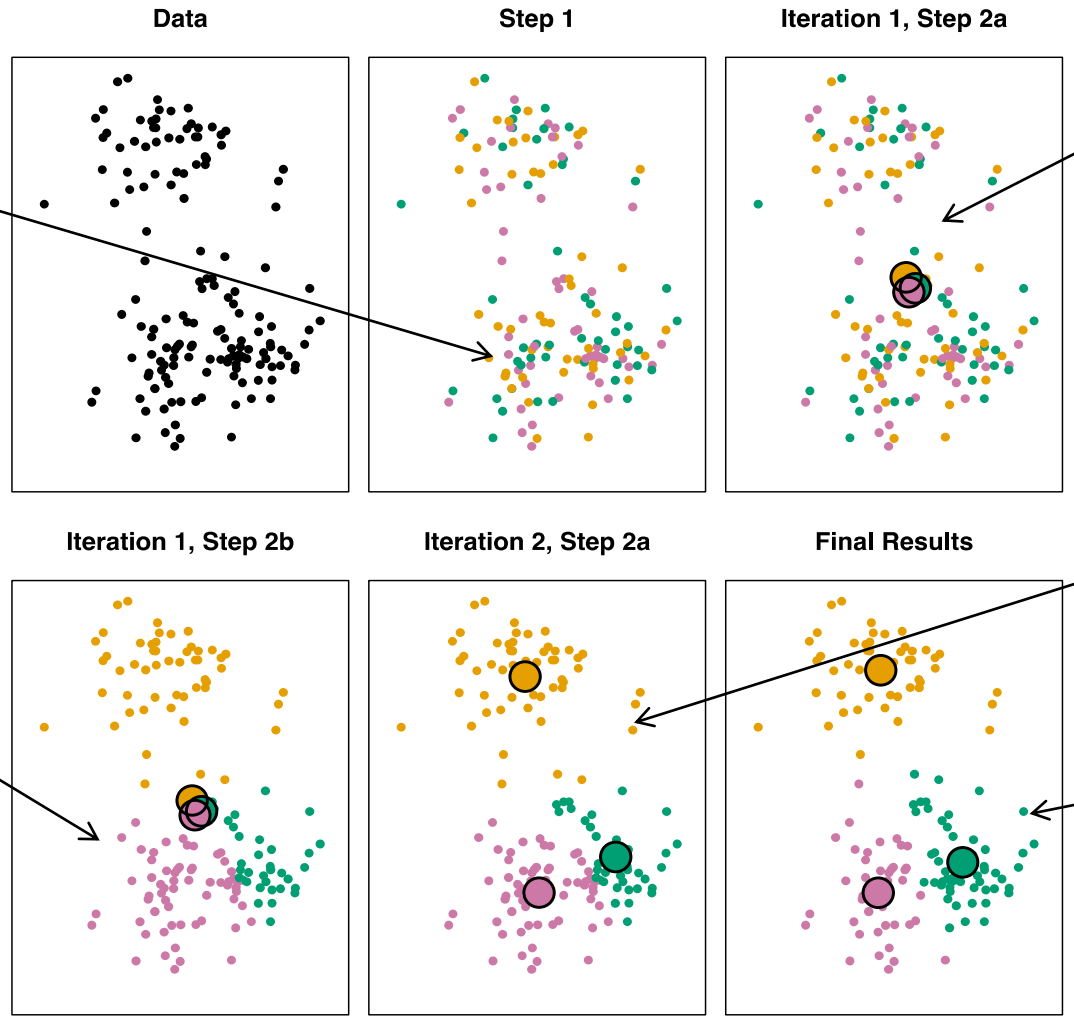
K-MEANS ALGORITHM

- This algorithm will decrease the value of the objective at each step
- Why?

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

$$\overline{x_{kj}} = \frac{1}{|C_k|}\sum_{i \in C_k} x_{ij}$$

AN ILLUSTRATION: K=3



LOCAL OPTIMUMS: DIFFERENT STARTING VALUES

- 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.
- 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 between the clusters.
- Those labeled in red all achieved the same best solution, with an objective value of 235.8.