Contents

Contents

Contents

Contents

Contents

Contents

Outline

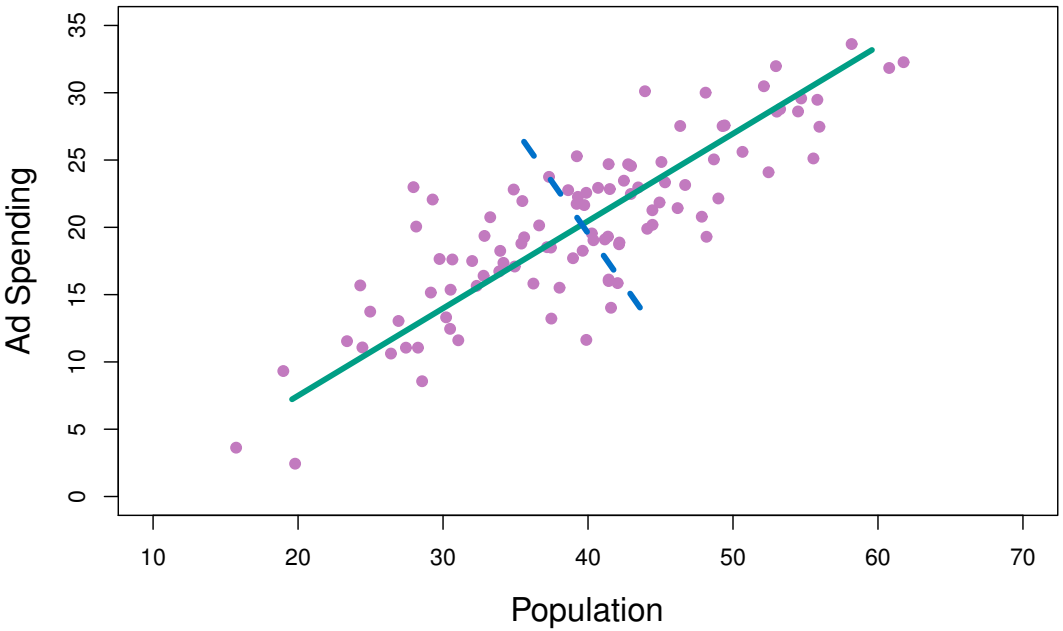
The Goals of Unsupervised Learning

- The goal is to discover interesting things about the measurements : is there an informative way to visualize the data ? Can we discover subgroups among the variables or among the observations ?
- Historical example John Snow map of cholera epidemy case (wikipedia-en)
- We discuss two methods :
  1. Principal Components Analysis (**PCA**) , a tool used for data visualization or data pre-processing before supervised techniques are applied, see model-selection course ;
  2. Clustering, a broad class of methods for discovering unknown subgroups in data. We present **k-means** algorithm and **hierarchical clustering** to obtain dendrogram.

The Challenges 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. e.g. no cross-validation to measure the error.
- 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 ;
  - Sentimental analysis.

Outline



- The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles.
- The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

- The first direction defines a direction in feature space along which the data vary the most. It is the reason why it is used in feature engineering (pre-processing)
- The second direction is the second one with this properties.
- Maths background : Gram-Schmidt process to build an orthonormal matrix.

PCA : Details

- The first principal component of a set of features $X_1, X2, ..., X_p$ is the normalized linear combination of the features
  $Z_1 = \phi_{11} X1 + \phi_{21} X_2 + ... + \phi_{p1} X_p$
  That has the largest variance. By normalized, we mean that $\sum_{j=1}^{p} \phi_{jp}^2 = 1$
- We refer to the elements $\phi_{11}, ..., \phi_{p1}$ as the loadings of the first principal component ; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, ..., \phi_{p1})^T$
- Since each of the x_{ij} has mean zero, then so does z_{i1} (for any values of z_{j1}). Hence the sample variance of the z_{i1} can be written as $\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$
- This is related to the properties of the "maximum variance explained".

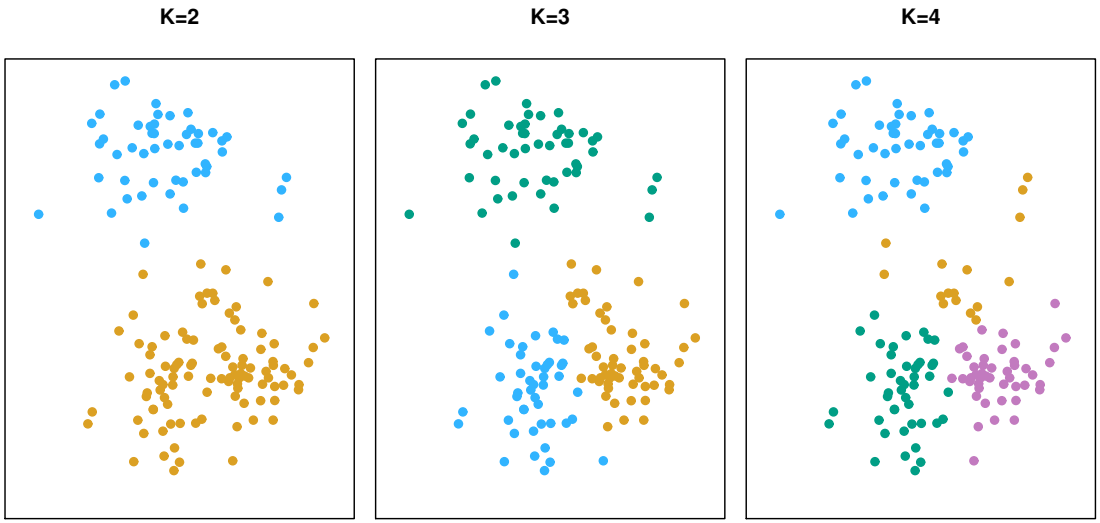
PCA : Details

- Consequently, this is translated in an optimization problem : $maximize_{(\phi_{11},\ldots,\phi_{p1})} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2$ subject to $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
- This problem can be solved via a singular-value decomposition **SVD** of the matrix X, a standard technique in linear algebra. This technique is also used in collaborative filtering algo.
- How many Principal Components must we use ? Elbow trick (fig 10.4).

Outline

- Clustering refers to a very broad set of techniques for finding homogeneous subgroups, or clusters, in a data set ;
- To make this concrete, we must define what it means for two or more observations to be similar or different. We must find a metric ;
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

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. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary.

Details of K-means clustering

- These cluster labels were not used in clustering ; instead, they are the outputs of the clustering procedure.
- Settings :
  Let $C_1, ..., C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties :
  1. $C_1 \cup C_2 \cup ... C_K = \{1, ..., n\}$. 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 $i \in C_k$.

Details of K-means clustering : continued

- 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 $WCV(C_k)$ of the amount by which the observations within a cluster differ from each other.
- 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. It is translated in the following optimisation problem, if we use the euclidean distance : $WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$

Details of K-means clustering : Algorithm

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

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** :
  - 2.1 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.
  - 2.2 Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

3 (option) Stop criterion when cost doesn't decrease "too much". Global minimum is not guarantee.

Outline

- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage (later we discuss strategies for choosing K)
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- In this section, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a **dendrogram** is built starting from the leaves and combining clusters up to the trunk.

Hierarchical Clustering Algorithm : Description

The approach in words :

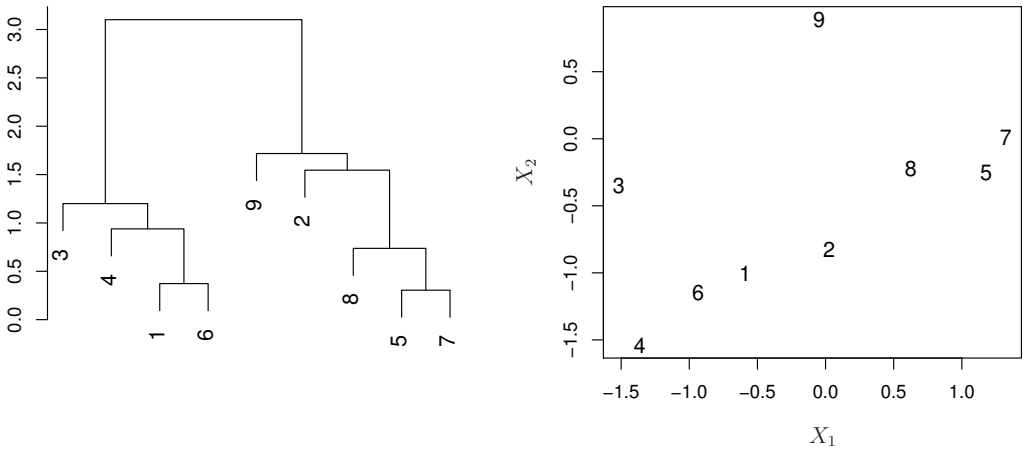
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster. strategies for choosing K)

Hierarchical Clustering Algorithm : Details

The approach in words :

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster. strategies for choosing K)

Hierarchical Clustering Algorithm : Example 1

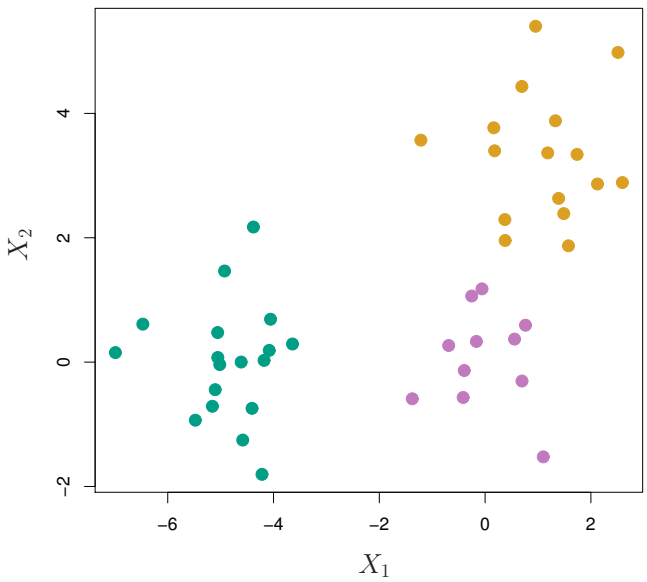


- An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space. The raw data on the right was used to generate the dendrogram on the left.
- Observations 5 and 7 are quite similar to each other, as are observations 1 and 6 (continued)

Hierarchical Clustering Algorithm : Example 1

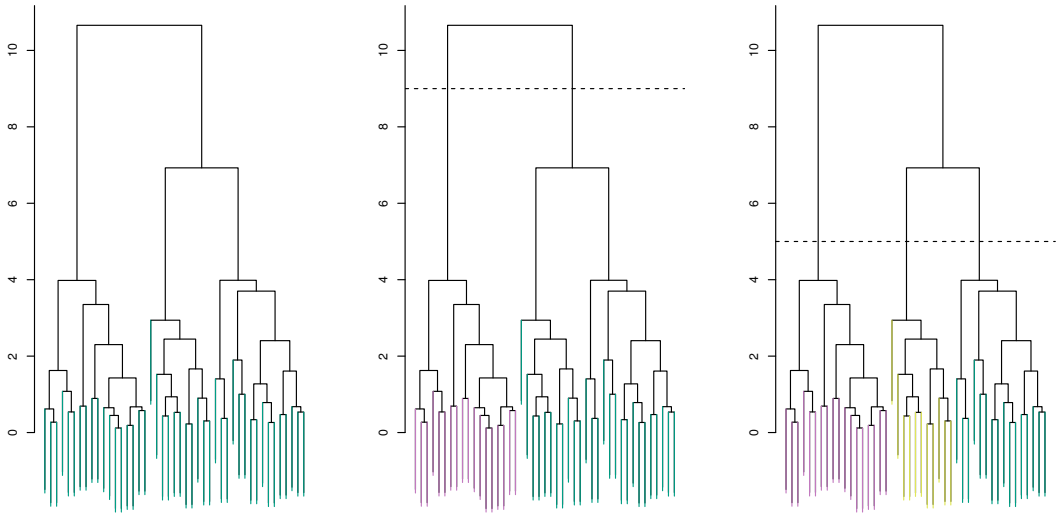
- However, observation 9 is no more similar to observation 2 than it is to observations 8 ; 5 ; and 7, even though observations 9 and 2 are close together in terms of horizontal distance.
- This is because observations 2 ; 8 ; 5 ; and 7 all fuse with observation 9 at the same height, approximately 1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

Hierarchical Clustering Algorithm : Example 2



- 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.

Hierarchical Clustering Algorithm : Example 2



- Left : dendrogram obtained from hierarchically clustering the data from previous Figure 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.

Outline

- Basket Analysis, frequent itemset discovery => A Priori algorithm
- k-medoids Clustering
(ESL Chap. 14)

Outline

- Both clustering and PCA seek to simplify the data via a small number of summaries, but their mechanisms are different :
  - 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.
- 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 intrinsically 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.
- It is indirectly responsible for the renewal of neural network by deep learning. See The Elements of Statistical Learning, chapter 14.