Unsupervised learning

Week 8

Daniele Bianchi[1]

whitesphd.com

[1] School of Economics and Finance
Queen Mary, University of London

Summary

This week we discuss the use of machine learning methods which do not use the target variables for the model calibration. These are called "unsupervised" learning methods. We discuss both the theory and the applications.

Contents

Unsupervised learning

Unsupervised learning

Unsupervised vs Supervised Learning:

- Most of this course focuses on **supervised learning** methods such as regression and classification.
- In that setting 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 $x_1, x_2, \ldots, x_p$.
- Here we instead focus on **unsupervised learning**, where we observe only the features $x_1, x_2, \ldots, x_p$. We are not interested in prediction any more, because we do not have an associated response variable y.

The Goals of Unsupervised Learning

- The goal is to discover interesting things about measurements: is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?
- We discuss two methods:
  - **Principal component analysis**, a tool used for data visualization or data pre-processing before supervised techniques are applied, and
  - **Clustering**, a broad class of methods for discovering unknown subgroups in data.

The Challenge of Unsupervised Learning

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediciton of a response.
- 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.

Principal Component Analysis

Principal Component Analysis

- Principal Component Analysis (PCA) produces a low-dimensional representation of a dataset.
- PCA finds a sequence of linear combinations of the variables that have a maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervissed learning problems, PCA also serves as a tool for data visualization.

Principal Component Analysis: Details

- The **first principal component** of a set of features $x_1, x_2, \ldots, x_p$ is the normalised linear combination of the features

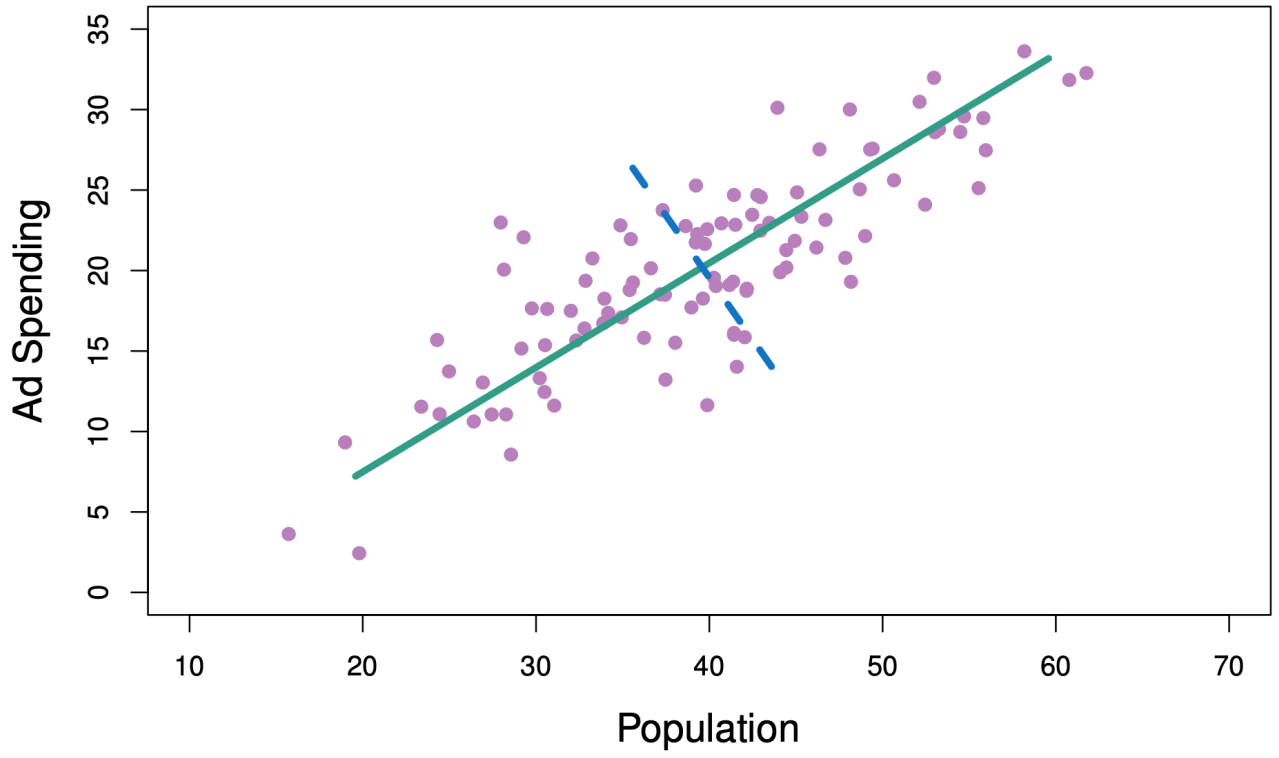
$$Z_1 = \phi_{11}x_1 + \phi_{21}x_2 + \ldots + \phi_{p1}x_p$$

that has the largest variance. By **normalized**, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

- We refer to the elements $\phi_{11}, \ldots, \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})'$.

- We constrain the loadings so that their sum of squares is equal to one.

PCA: Example



The population size (pop) and advertisement 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.

Computation of Principal Components

- Suppose we have a $n \times p$ data set $\mathbf{X}$. Since we are only interested in variance, we assume that each of the variables in $\mathbf{X}$ has been centered to have mean zero (that is, the column means of $\mathbf{X}$ are zero).

- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \ldots + \phi_{p1} x_{ip}$$

  for $i = 1, \ldots, n$ that has largest sample variance, subject to the constraint $\sum_{j=1}^{p} \phi_{j1}^{1} = 1$.

- Since each of the x_{ij} has mean zero, then so does z_{i1}. Hence the sample variance of the z_{i1} can be written as $1/n \sum_{i=1}^{n} z_{i1}^{2}$.

Further principal components

- The second principal component is the linear combination of $x_1, x_2, \ldots, x_p$ that has maximal variance among all linear combinations that are **uncorrelated** with z_1.
- The second principal component scores $z_{12}, z_{22}, \ldots, z_{n2}$ take the form

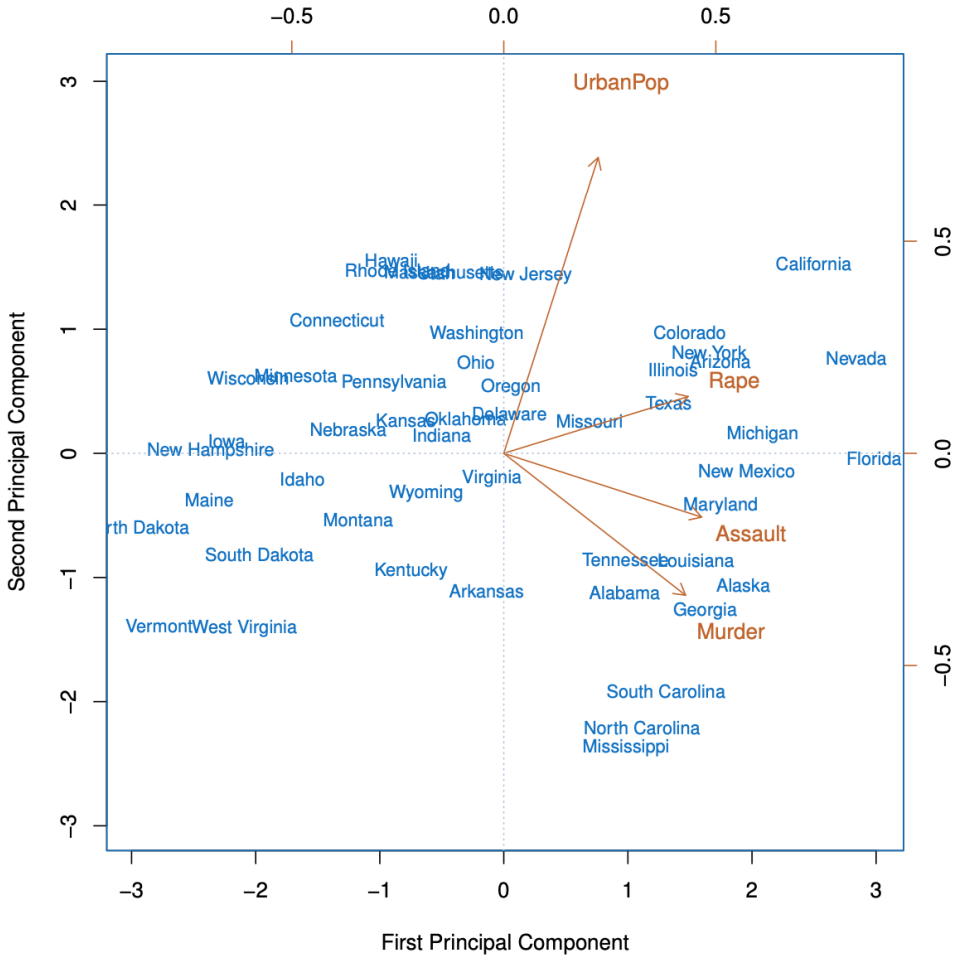
$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \ldots + \phi_{p2} x_{ip}$$

  where ϕ_2 is the second principal component loading vector with elements $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$.
- It turns out that constraining z_2 to be uncorrelated with z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction ϕ_1. And so on.

Illustration

- **USArrests** data: For each of the 50 states in the USA, the data contains the number of arrests per 100,000 residents for three crimes, i.e., assualt, murder, and rape. The data also record the % population as UrbanPop.
- The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

Illustration



Source: "An introduction to Statistical Learning", James et al. (2013).

Illustration: Figure details

The first two principal components for the USArrests data.

- The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors. For example, the loading for *rape* on the first component is 0.54, and its loading on the second principal component 0.17.
- This figure is known as a **biplot**, because it displays both the principal component scores and the principal component loadings.

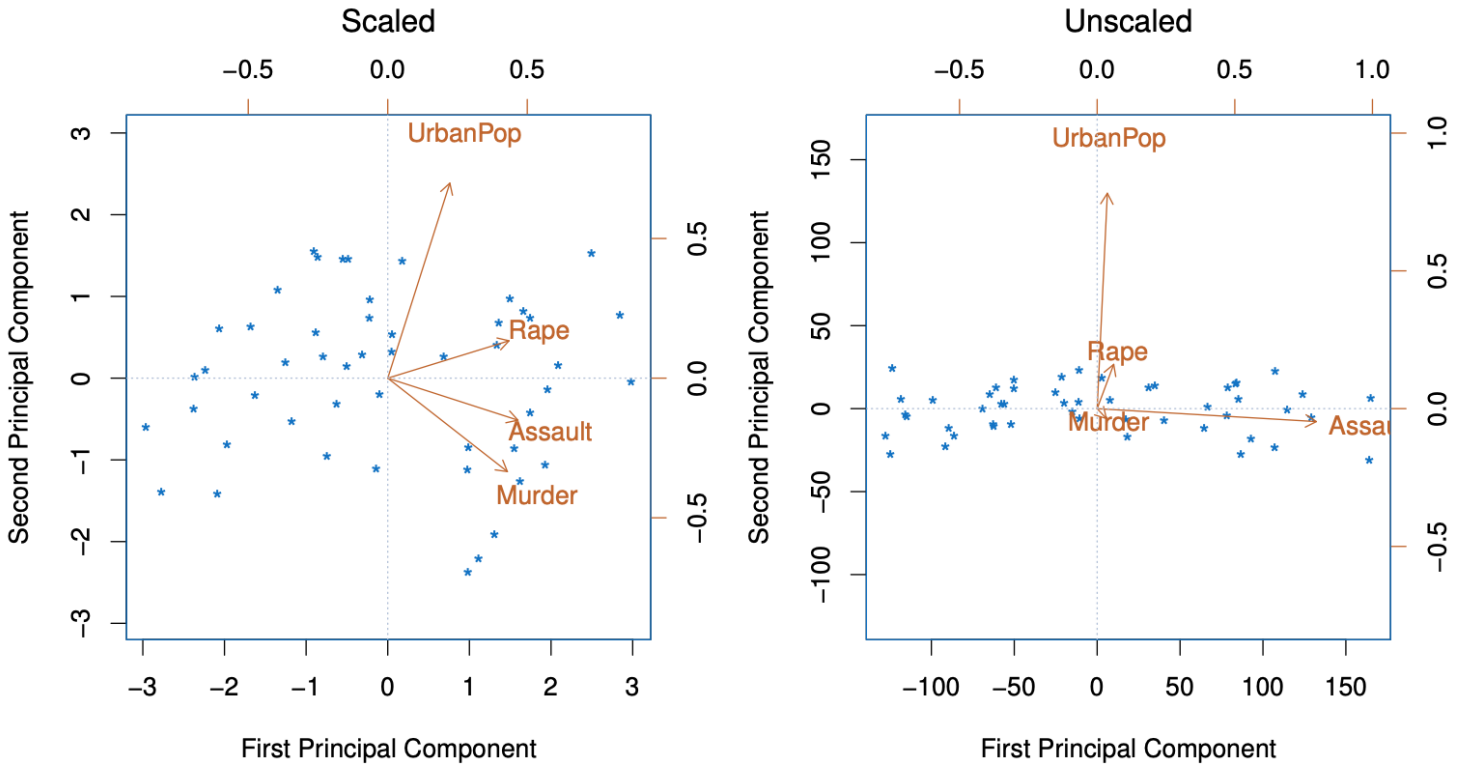
Illustration

Table 1: PCA loadings

	PC1	PC2
Murder	0.536	-0.418
Assault	0.583	-0.188
UrbanPop	0.278	0.872
Rape	0.543	0.167

Scaling variables

- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.
- If they are in the same units, you might or might not scale the variables.



Proportion variance explained

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.
- The **total variance** present in a data set is defined as

$$\sum_{j=1}^{p} \text{Var}\left(x_j\right) = \sum_{j=1}^{p} \frac{1}{n} x_{ij}^2,$$

  ans the variance explained by the mth principal component is

$$\text{Var}\left(Z_m\right) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2,$$

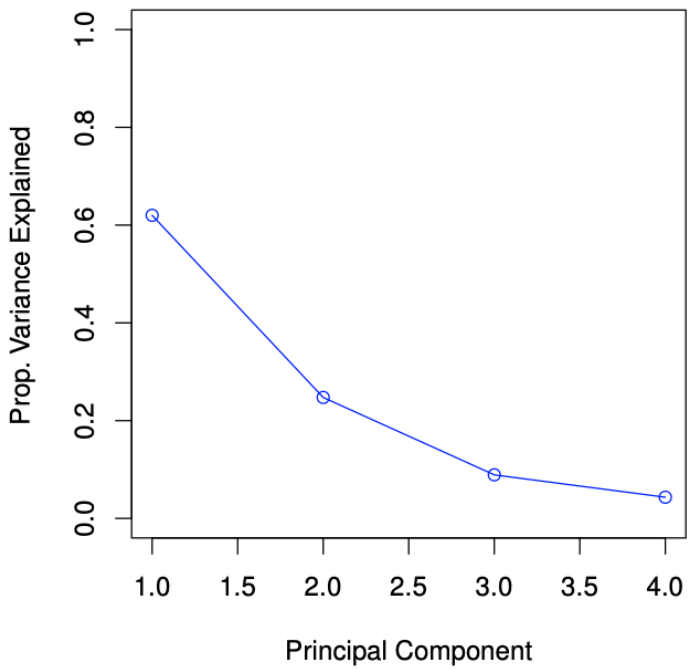
- It can be shown that $\sum_{j=1}^{p} \text{Var}\left(x_j\right) = \sum_{m=1}^{M} \text{Var}\left(Z_m\right)$ with $M = \min\left(n-1, p\right)$.

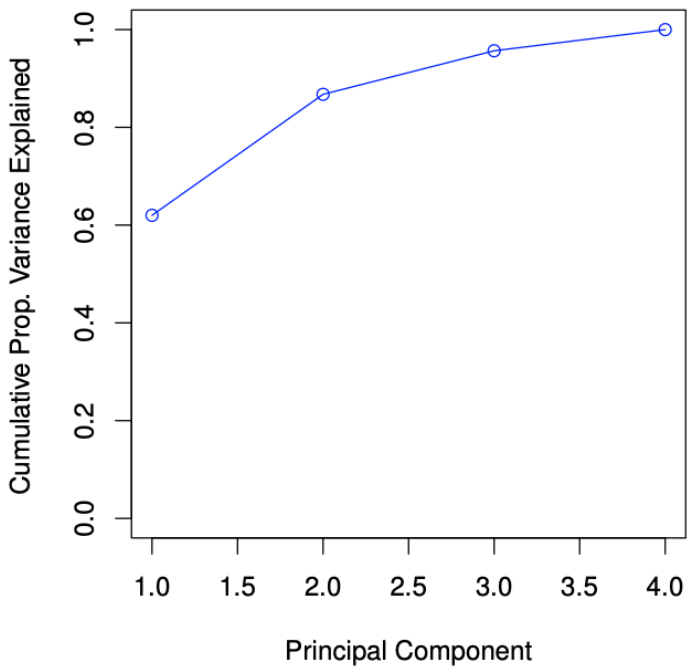
Proportion variance explained

– Therefore, the PVE of the mth principal component is given by the positive quantity between 0 and 1

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2},$$

– The PVEs sum to one. We sometimes display the cumulative PVEs.





Clustering

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.
- It 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 based on knowledge of the data being studied.

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.

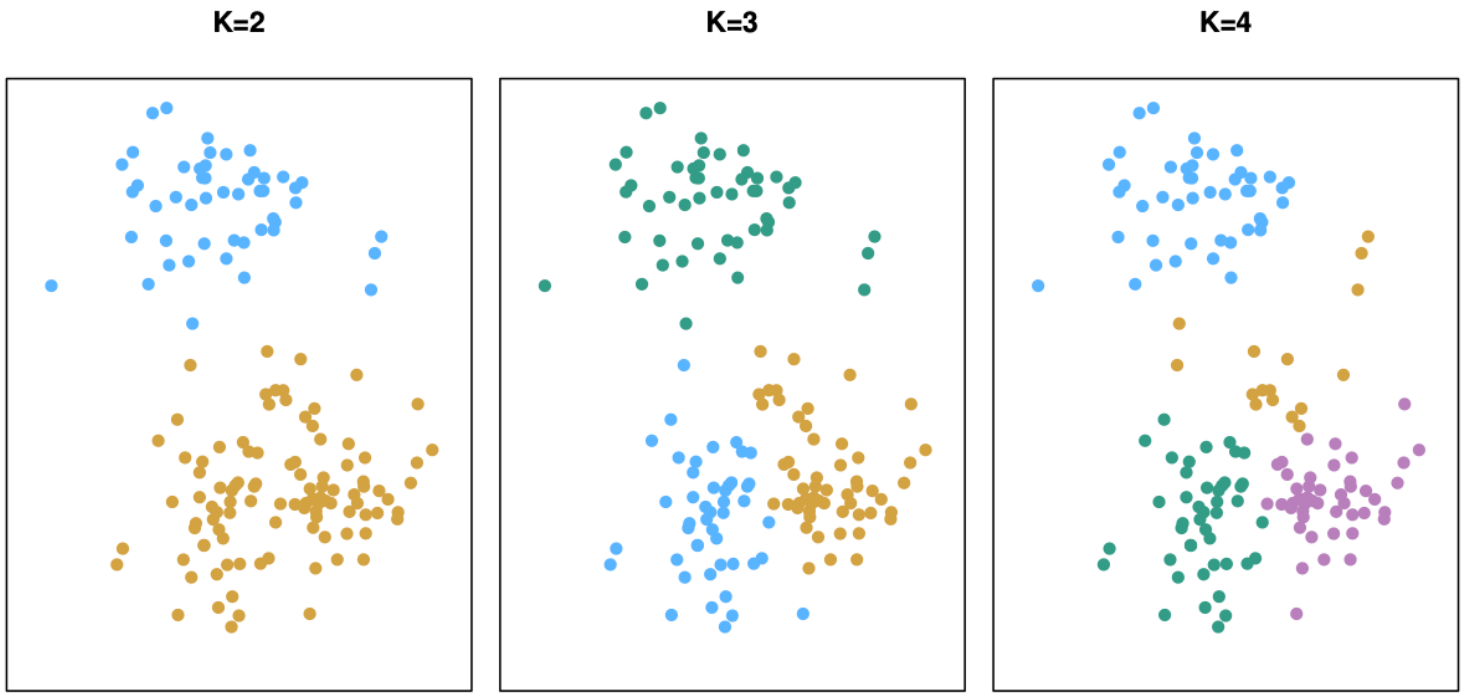
Clustering for market segmentation

- 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 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 clusterings obtained for each possible number of clusters, from 1 to n.

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. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

Details of K-means clustering

Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \ldots C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

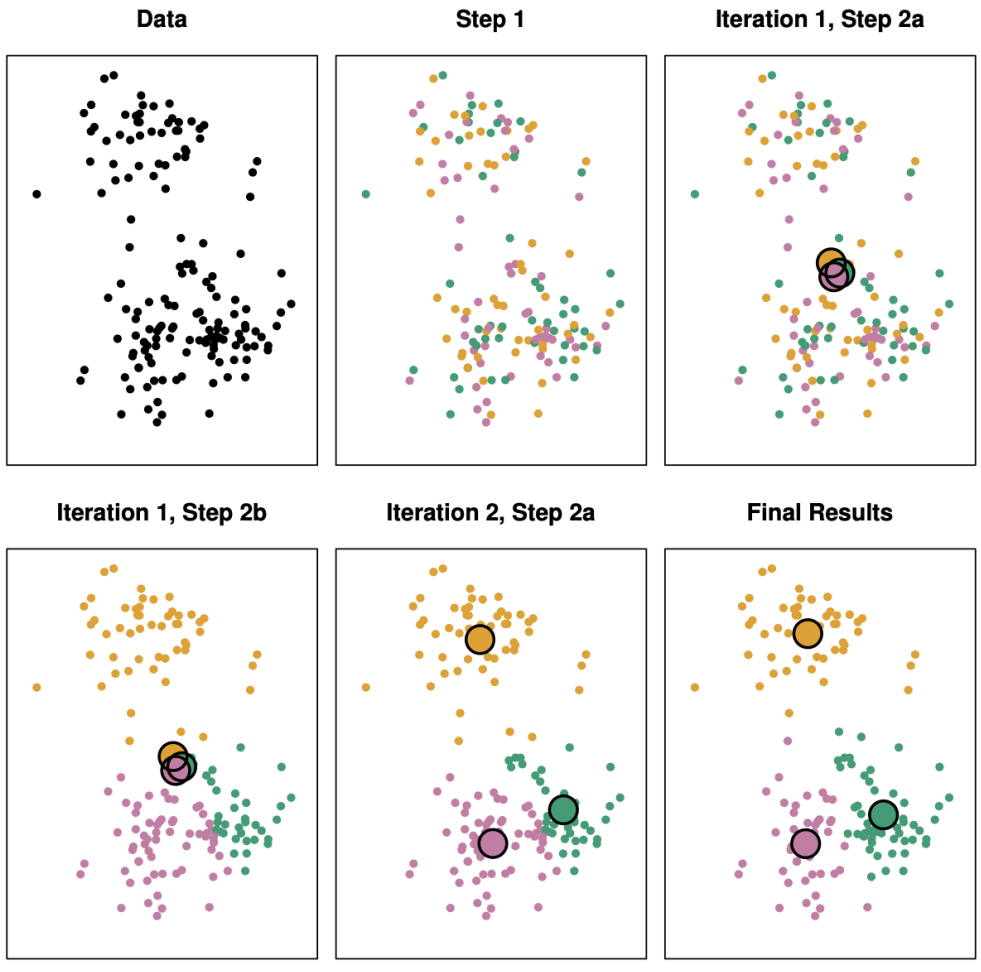
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

K-means clustering algorithm

- Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
- Iterate until the cluster assignments stop changing:
  - 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.
  - Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

K-means clustering algorithm



Source: "An introduction to Statistical Learning", James et al. (2013).

Details of the 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.

Conclusions

- 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.

References

1. "An Introduction to Statistical Learning: With Applications in R By Gareth James, Trevor Hastie, Robert Tibshirani, Daniela Witten." (2014): 556-557, chapter 14.