Project in Data Intensive Systems

4DV652
Lab Lecture 7
Welf Löwe

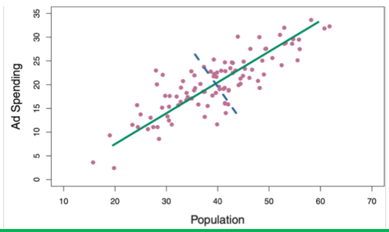
Agenda

- Principle component analysis
- Clustering
- Lab 7 task descriptions

Principle component analysis (PCA)

- Dimensionality reduction:
  - Unsupervised learning
  - Finds a low-dimensional representation of the observations that explain a good fraction of the variance and/or preserves distance of the observations
- PCA is a method of dimensionality reduction
- Each of the components/dimensions found by PCA is a linear combination of the p features.
- Once we have computed the principal components, we can plot them
  - Geometrically, this amounts to projecting the original data down onto the subspace spanned by ϕ_1, ϕ_2, and ϕ_3, and plotting the projected points

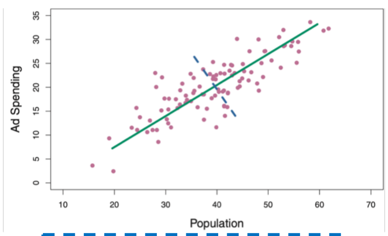
Simple example, p=2 : population size (pop) and ad spending for a particular company (ad)



$$\boxed{z_{i1} = 0.839 \times (\mathrm{pop}_i - \overline{\mathrm{pop}}) + 0.544 \times (\mathrm{ad}_i - \overline{\mathrm{ad}})}$$

4

Simple example, p=2: population size (pop) and ad spending for a particular company (ad)

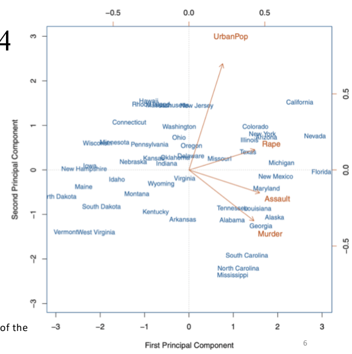


$$\boxed{\phi_{12} = 0.544 \text{ and } \phi_{22} = -0.839}$$

5

Larger example, p=4 (USArrests data set)

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186



Assault, Murder, Rape, and UrbanPop (the percent of the population in each state living in urban areas).

6

Agenda

- Principle component analysis
- Clustering
- Lab 7 task descriptions

Clustering

- Find homogeneous or self-similar subgroups among the observations
- Approaches
  - K-means clustering,
    - number of clusters K is input
  - Hierarchical clustering,
    - Does not require that we commit to a particular choice of K
    - Suggests many possible clusterings visualized in a tree representation called dendrogram
    - Visual analytics (human insight) to decide on a clustering that "makes sense"
  - …

Clustering

A sets of clusters C_1 , … , C_K each a set of observations such that

- each observation belongs to at least one of the K clusters.

$$C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$$

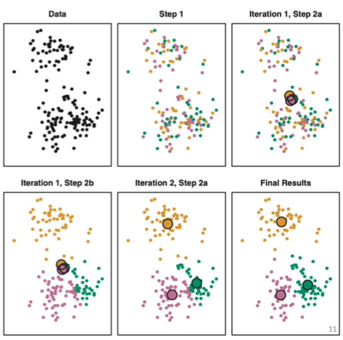
- the clusters are non-overlapping: no observation belongs to more than one cluster.

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'.$$

K-means clustering algorithm (heuristic)

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 k-th cluster centroid is the vector of the p feature means for the observations in the k-th cluster.
   b. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Example iterations

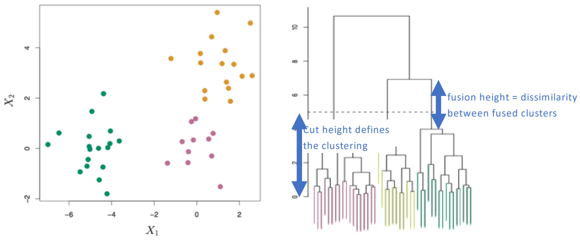


K=3

Hierarchical Clustering Algorithm (bottom-up)

1. Begin with i=n observations and treat each as its own cluster.
2. Until i=2:
   a. Compute all $i(i-1)/2$ pairwise inter-cluster (Euclidian) dissimilarities among the i clusters as maximal, minimal, or mean observation dissimilarities, or centroid dissimilarities (linkage)
   b. Identify the pair of clusters that are least dissimilar, i.e., most similar.
   c. Fuse these two clusters.
      - The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
   d. There are i−1 remaining clusters.

Dendrograms as cluster representation
(Example on random data)



13

Lab assignment 8: PCA and Clustering

- ML
  - Perform a clustering of the AIMO observations (exclude the scores)
  - Interpret the clusters
  - Perform a PCA of the AIMO observations
  - Draw the first two principle component scores of the data points and their clusters (colored 2D scatterplot)
  - Interpret the plot
- Software development
  - Maintenance sprint
- Reporting:
  - In a eights notebook, document the ML steps
- Deadline: 2021-03-31