

# ML Non Supervisionato

## ▼ Clustering

- Hard clustering
- Soft Clustering

## ▼ K-Means

### ▼ Algoritmo

- ▼ Scegliere numero K cluster
  - Elbow criteria
  - indice di Davies-Bouldin
- Generare K centroidi che rappresentano i cluster
- Evaluation Step (E-Step)
- Minimization Step (M-Step)

### ▼ funzione obiettivo da minimizzare

- Somma dei Quadrati nel Cluster (Within-Cluster Sum of Squares - WCSS)

## ▼ Clusterign Gerarchico

- Agglomerativo
- Divisivo
- ▼ Algoritmo
  - Single Linkage
  - Average Linkage
  - Complete Linkage
  - Centroid Linkage

## ▼ PCA

- Standardizzazione dei dati in input  $X$   $n$  (osservazioni) \*  $m$  (numero di variabili)
- Calcolo della matrice di covarianza  $S$   $m \times m$  di  $X$
- ▼ Calcolo di  $m$  autovettori e  $m$  autovalori della matrice di covarianza di  $S$

- La varianza dei dati proiettati lungo le componenti principali decresce all'aumentare delle componenti, la maggior parte della varianza complessiva è concentrata nelle prime componenti principali  $var PC1 > var PC2 > var PC3 > \dots > var PCm$
- Trasformazione dei dati passando alle coordinate nello spazio delle componenti principali:  $Y=X \cdot V$
- ▼ Scelta delle prime  $p$  componenti principali che rappresentano gran parte della varianza delle variabili originali (scree plot o plot della frazione di varianza spiegata dalle prime  $p$  componenti).
  - La scelta delle  $p$  componenti principali si può effettuare rappresentando graficamente gli autovalori (screen plot) e scegliamo il numero che corrisponde al punto di gomito