Introduction to clustering

Gianluca Campanella

Contents

Clustering

k-means

Clustering

Classification versus clustering

Classification

- Data are ‘labelled’ → supervised

→ Find a ‘rule’ that assigns labels to new observations

Clustering

- Data are ‘unlabelled’ → unsupervised

→ Identify structure and patterns

Classification versus clustering

Idea

- Group observations that are ‘close’ (high intra-cluster similarity)
- Identify ‘natural’ groupings (low inter-cluster similarity)

Types of clustering

- **Hard**: each observation belongs to exactly one cluster
- **Soft** (or **fuzzy**): observations may belong to multiple clusters
- **Hierarchical**: observations belong to ‘concentric’ clusters

k-means

k-means

Given the number of clusters *k*…

- Select *k* centroids (e.g. *k* observations at random)
- For each observation:
    - Determine distances to the centroids
    - Reassign to the closest centroid
- Recompute the centroids
- Repeat until no observations move group

k-means

Questions

- How do we define similarity?
- How many clusters do we use?

Curse of dimensionality

As the number of variables (coordinates) increases…

- The volume of the space increases
- Pairwise distances become more similar → sparsity
- Some samples have huge neighbourhoods → ‘hubs’