Clustering problem

Procheta Sen

UNIVERSITY OF LIVERPOOL

Clustering problem

- Given a dataset, partition its objects into sets (**clusters**) $C_1, C_2, \ldots, C_k$ such that the objects in each cluster are "**similar**" to one another.
- Specific definitions depend on how the notion of **similarity** is defined.

Why cluster data?

- Data summarization
- Topic detection
- Visualisation
- Outlier detection
- Community detection

Clustering is unsupervised learning

- Supervised learning

  Labels for training instances are provided

- Unsupervised learning

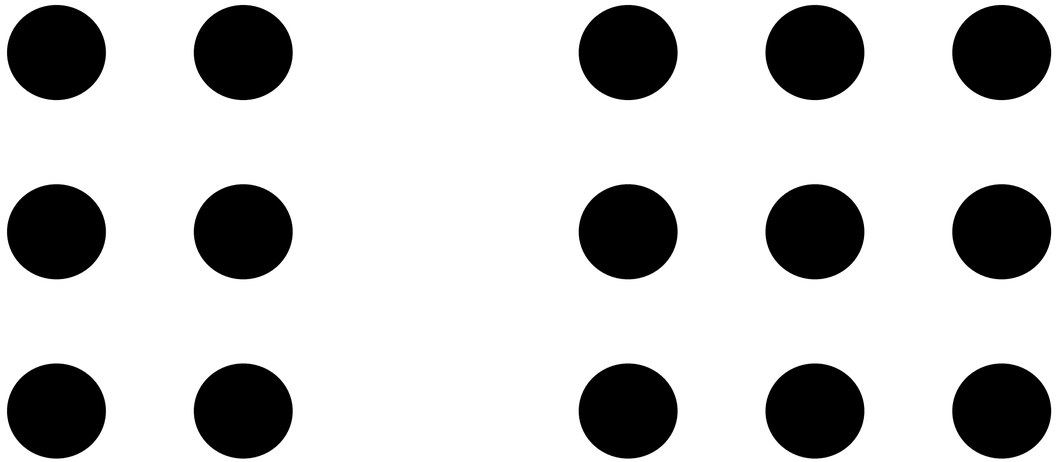
  No labels for training instances are provide

- Semi-supervised learning

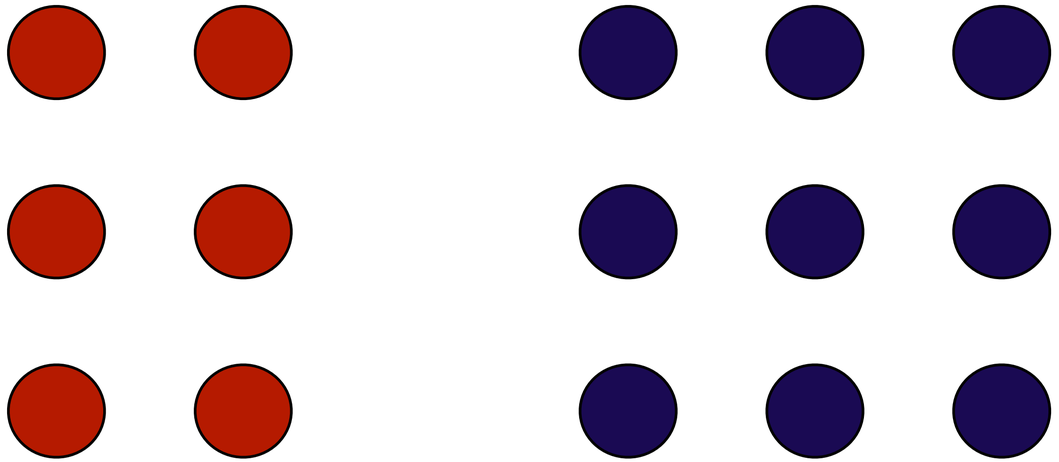
  Both labeled and unlabeled training instances are provided

- What can we learn about training data if we do not have any labels?
  - The similarity and distribution of the features can still be learned and this can be used to create rich feature spaces for supervised learning (if required)

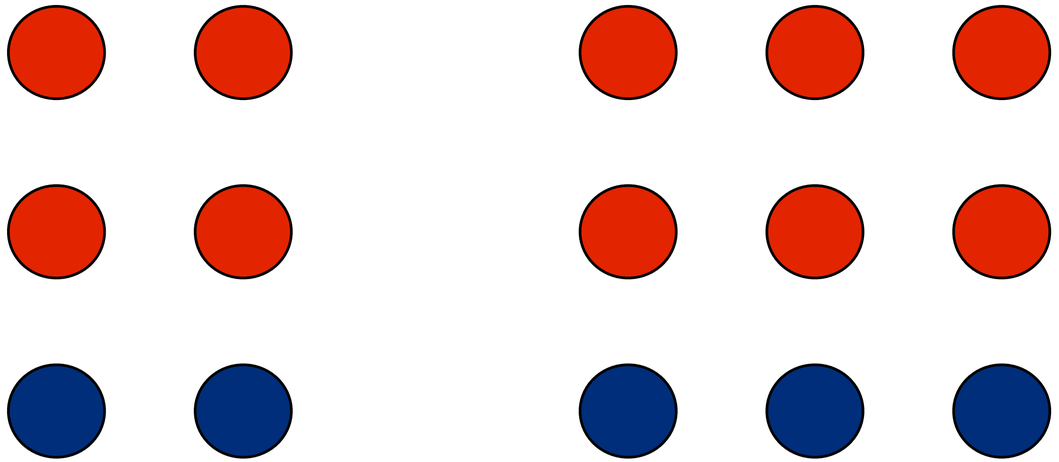
How to cluster the following data



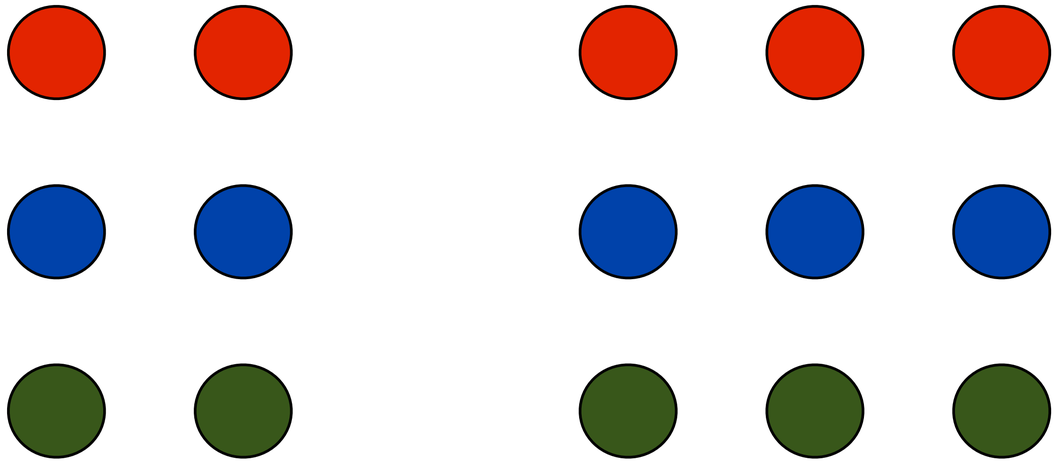
How to cluster the following data



How to cluster the following data



How to cluster the following data



How many clusters?

General Remarks

- A single dataset can be clustered in several ways
- There is no single right or wrong clustering
  - Simply different views on the same data
- Then how can we measure the quality of a clustering algorithm?
  - **Extrinsic methods**: Compare the clusters produced by a clustering algorithm against some reference (gold standard or ground truth) set of clusters
  - **Intrinsic methods**: only the partition of objects into clusters is used

Clustering Algorithms

- **Representative-based**
  - Choose k representatives, assign each element in the dataset to a representative, and iteratively update the partition
    - k-Means, k-Medoids
- **Hierarchical**
  - Create a hierarchy of clusters (dendrogram)
    - Agglomerative clustering (bottom-up)
    - Conglomerative clustering (top-down)
- **Graph-based clustering**
  - Community detection (Modularity optimisation)
  - Graph-cut algorithms (Spectral Clustering)
- **Many other types**