Chapter 6

Clustering

Contents

6.1 Clustering

Clustering is an unsupervised learning technique which automatically partitions unlabeled data into groups of similar datapoints. It is useful for:

Segmentation Segmenting a large set of cases into small subsets that can be treated similarly.

- e.g., image segmentation.

Compression Generate a more compact description of a dataset.

- e.g., handwritten digit recognition.

Representation Model an underlying process that generates the data as a mixture of different, localized processes.

6.2 Clustering Applications

- Cluster news articles or web pages or search results by topic.
- Cluster protein sequences by function of genes according to expression profile.

- Cluster users of social networks by interest.
- Cluster galaxies or nearby stars.

6.3 Clustering Algorithms

6.3.1 Flat clustering

No inter-cluster structure.

- k-means algorithm.
- Gaussian mixture models (GMM).
- Spectral clustering.

6.3.2 Hierarchical clustering

Clusters for a hierarchy.

- Bottom-up (agglomerative clustering).
- Top-down (divisive clustering).

6.3.3 Hard clustering

Items are assigned to a unique cluster.

- k-means algorithm.
- Spectral clustering.

6.3.4 Soft (fuzzy) clustering

Cluster membership is a real-valued function, distributed across several clusters.

- Soft k-means.
- Gaussian mixture models.

6.3.5 Centroid-based clustering

This type of clustering algorithm forms around the centroids of the data points. E.g., $k-means$, $k-modes$.

6.3.6 Distribution-based clustering

Clustering algorithm is modeled using statistical distributions. It assumes that the data points in a cluster are generated from a particular probability distribution, and the algorithm aims to estimate the parameters of the distribution. E.g., GMM.

6.3.7 Density-based clustering

This type of clustering algorithm groups together data points that are in high-density concentration and separates points in low-concentration regions. E.g., DBSCAN.

6.4 k-Means Clustering

- k-means algorithm is an iterative clustering algorithm, based on the Euclidean distance. It is a non-parametric learning algorithm.
- A greedy algorithm (Lloyd's algorithm) locally optimizes the cluster quality measure:
  - The cluster quality measure is computed based on the cluster centroid.
  - Find the closest cluster center for each item and assign it to that cluster.
  - Recompute the cluster centroid as the mean of items, for the newly-assigned items in the cluster.
- Initialize: Pick k points as cluster centers.
- Repeat:
  - Assign data points to closest cluster center.
  - Change the cluster center to the average of its assigned points.
- Stop when the assignments of data points do not change.
- Input: A set of n datapoints $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ with k clusters.
- Output: k representatives $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$.
- Objective: choose $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$ such that:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{k} ||x_i - c_j||^2 \tag{6.1}$$

- Initialize cluster centers $c_1, c_2, \ldots, c_k$ and clusters $C_1, C_2, \ldots, C_k$.
- Repeat until there is no further change:
  - For each $j : C_j \leftarrow \{x$ whose closest center is $c_j\}$.
  - For each $j : c_j \leftarrow$ mean of C_j.