Machine Learning and Data Mining

26. May 2021.

Clustering

different types of clustering

- Exclusive (partitioning) (Example: K-means)
- Agglomerative (Example: Hierarchical clustering)
- Overlapping (Example: Fuzzy C-Means)
- Probabilistic (Example: Following keywords: "man's shoe." "women's shoe." "women's glove." "man's glove." can be clustered into two categories "shoe" and "glove" or "man" and "women.")

Clustering method

- K-means clustering
- K-NN (k nearest neighbors)
- Hierarchical clustering
- Principal Component Analysis (PCA)
- Independent Component Analysis

Understanding the K-Means Algorithm

Conventional k-means

- randomly select k centroids, where k is equal to the number of clusters
- Centroids are data points representing the center of a cluster.
- each cluster is associated with a centroid.
- a two-step process called expectation-maximization
- The expectation step assigns each data point to its closest centroid.
- the maximization step computes the mean of all the points for each cluster and sets the new centroid.

Conventional k-means

Algorithm 1 Conventional k-means

Input: k the number of clusters k

1: **repeat**
2: Step 1 (Assignment step): assigns each data point to its nearest centroid.
3: Step 2 (Update step): computes the mean of all the points for each cluster and define the new centroid as the computed mean value.
4: **until** the centroid positions stay unchanged during the step 2.

measure of error

- sum of the squared error (SSE)
- SSE is defined as the sum of the squared Euclidean distances of each point to its closest centroid

K-means

- aims K-means aims to partition n observations $X = \{x_1, ..., x_n\}$ into K clusters $S = \{S_1, S_2, ..., S_k\}$ in which each observation belongs to the cluster with the nearest mean
- μ_i is the mean of the points in the cluster S_i
- Given an initial set of K means $\mu_1, ..., \mu_k$ the algorithm proceeds by alternating between two steps
- each cluster is associated with a centroid.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Choosing the Appropriate Number of Clusters

methods for evaluating the appropriate number of clusters

- The elbow method ($WCSS = \sum_{i=1}^{m}(x_i - c_i)^2$)
- The silhouette coefficient