

Data Mining: Lecture 3

Cluster Analysis

S. Nõmm

¹Department of Software Science, Tallinn University of Technology

19.09.2018

Introduction

Given a set of data points, partition them into groups with respect to chosen similarity criteria.

- Data summarization.
- Discover the structure of the set.
- Part of pre processing

Feature selection

Given a set of data points, partition them into groups with respect to chosen similarity criteria.

- Filter Models

- ▶ Predictive Attribute Dependence
- ▶ Entropy

$$E = - \sum_{i=1}^m [p_i \log(p_i) + (1 - p_i) \log(1 - p_i)]$$

- ▶ Hopkins Statistic

$$H = \frac{\sum_{i=1}^r \beta_i}{\sum_{i=1}^r (\alpha_i + \beta_i)}.$$

- Wrapper models

Representative-Based Algorithms

- The k -Means Algorithm.
- The Kernel k -Means Algorithm
- The k -Medians Algorithm
- The k -Medoids Algorithm

The k -Means Algorithm

Let data set \mathcal{D} to be clustered into K clusters.

Generate K centroids randomly. **Repeat**

- For Each point of \mathcal{D} computed distances to each of K centroids.
- The point is assigned the class label of the closest centroid.
- Update centroid coordinates for each class by computing the mean values.

Until converge.

Hierarchical Clustering Algorithms

- Bottom-Up Agglomerative Methods
- Top-Down Divisive Methods

Bottom-Up Agglomerative Methods

Initialize $n \times n$ distance matrix M **Repeat**

- Find the pair of closest clusters and merge them
- Update matrix M

Until termination criterion

Group-Based Statistics

- Best (single) linkage
- Worst (complete) linkage
- Group-average linkage
- Closest centroid
- Variance based criterion
- Ward's method

EM-algorithm

Let us consider K-Means from the probabilistic point of view.

- (E-step) Each data point of the set \mathcal{D} has a probability belonging to cluster j , which is proportional to the scaled and exponentiated Euclidean distance to each representative Y_j . In the k-means algorithm, this is done in a "hard" way, by choosing the smallest Euclidean distance to the representative of Y_j .
- (M-step) The center Y_j is the weighted mean over all the data points where the weight is defined by the probability of assignment to cluster j . The hard version of this is used in k-means, where each data point is either assigned to a cluster or not assigned to a cluster (i.e., 0-1 probabilities).

EM-algorithm

Assumption: the data was generated from a mixture of k distributions with probability distributions $\mathcal{G}_1 \dots \mathcal{G}_k$. Each distribution \mathcal{G}_i represents a cluster and is also referred to as a mixture component.

- (E-Step) Given the current value of the parameters in Θ , estimate the posterior probability $P(\mathcal{G}_i|X_j, \Theta)$ of the component \mathcal{G}_i having been selected in the generative process, given that we have observed data point X_j . The quantity $P(\mathcal{G}_i|X_j, \Theta)$ is also the soft cluster assignment probability that we are trying to estimate. This step is executed for each data point X_j and mixture component \mathcal{G}_i .
- (M-Step) Given the current probabilities of assignments of data points to clusters, use the maximum likelihood approach to determine the values of all the parameters in Θ that maximize the log-likelihood fit on the basis of current assignments.

Grid- and density- based methods

One of the major problems with distance-based and probabilistic methods is that the shape of the underlying clusters is already defined implicitly by the underlying distance function or probability distribution. Possible solutions:

- Grid- based methods
- Density- based methods
- Graph- based algorithms
- Nonnegative matrix factorization

Grid-based method

- Discretize each dimension of data \mathcal{D} into r ranges;
- Determine dense grid cells at a given density level;
- Create graph in which dense grids are connected if they are adjacent;
- Determine connected components of graph;
- return points in each connected component as a cluster;

Density-based methods

Definition

*Data point is defined as a **core point**, if it contains at least τ data points.*

Where τ is the density parameter.

Definition

*A data point is defined as a **border point**, if it contains less than τ points, but it also contains at least one core point within the radius ϵ .*

Definition

*A data point that is neither a core point nor a border point is defined as a **noise point**.*

DBSCAN

- Determine core, border and noise points of \mathcal{D} at level (ϵ, τ) ;
- Create graph in which core points are connected if they are within ϵ of one another;
- Determine connected components in graph;
- Assign each border point to connected component with which it is best connected;
- Return points in each connected component as a cluster;

Cluster Validation

- Internal Cluster Validation
 - ▶ Sum of square distances to centroids;
 - ▶ Intracluster to intercluster distance ratio;
 - ▶ Silhouette coefficient;
 - ▶ Probabilistic measure;
- External Cluster Validation, used when ground truth information is available.
 - ▶ Confusion matrix;
 - ▶ Cluster purity;
 - ▶ Gini index;

Internal Validation Criteria

- Intracluster to Intercluster distance ratio:

$$S_{intra} = \sum_{(\bar{X}_i, \bar{X}_j) \in P} \frac{s(\bar{X}_i, \bar{X}_j)}{|P|}$$
$$S_{inter} = \sum_{(\bar{X}_i, \bar{X}_j) \in Q} \frac{s(\bar{X}_i, \bar{X}_j)}{|Q|}$$

where, P is the set of all point pairs of given cluster, Q is the set of all possible pairs where one point belong to the cluster and the other not, s is a distance function. Smaller values of S_{intra}/S_{inter} indicate better clustering behaviour.

- Silhouette coefficient:

$$S_i = \frac{D_{min_i}^{out} - D_{avg_i}^{in}}{\max\{D_{min_i}^{out}, D_{avg_i}^{in}\}}$$

where, $D_{min_i}^{out}$ is the minimal distance to data points outside of the given cluster, and $D_{avg_i}^{in}$ is the average distance to the points within the given cluster. Larger positive values indicate highly separated clusters.

Cluster Purity. NB! Not unsupervised any more!!!

- Let m_{ij} represent the number of data points from class (ground-truth cluster) i that are mapped to (algorithm determined) cluster j .
- Denote number of data points in true cluster i are by N_i , the number of data points in algorithm-determined cluster j by M_j .

$$N_i = \sum_{j=1}^{k_d} m_{ij}; \quad M_j = \sum_{i=1}^{k_t} m_{ij};$$

- For a given algorithm-determined cluster j , the number of data points P_j in its dominant class is: $P_j = \max_i m_{ij}$.
- Purity index is defined

$$P_a = \frac{\sum_{j=1}^{k_d} P_j}{\sum_{j=1}^{k_d} M_j}.$$

Gini index

- Gini index for algorithm determined cluster j is defined:

$$G_j = 1 - \sum_{i=1}^{k_t} \left(\frac{m_{ij}}{M_j} \right)^2.$$

- Average Gini index is defined as follows:

$$G = \frac{\sum_{j=1}^{k_d} G_j M_j}{\sum_{j=1}^{k_d} M_j}.$$