DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Lecture 19.3 - Clustering and Validation

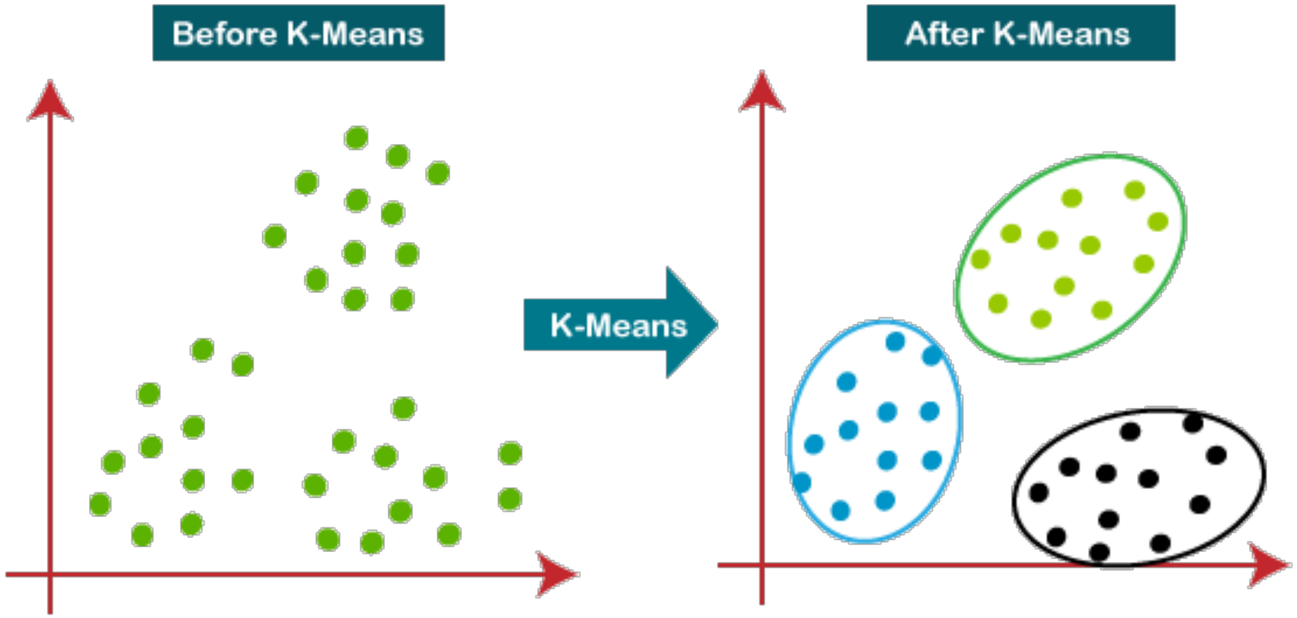
Clustering Analysis

Cluster: a collection of data objects

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters
- High Intra-class similarity
- Low Inter-class similarity

Cluster analysis

- Grouping a set of data objects into clusters



Source: https://leasetruk.medium.com/k-means-algorithm-with-example-b8994b3c4ef3

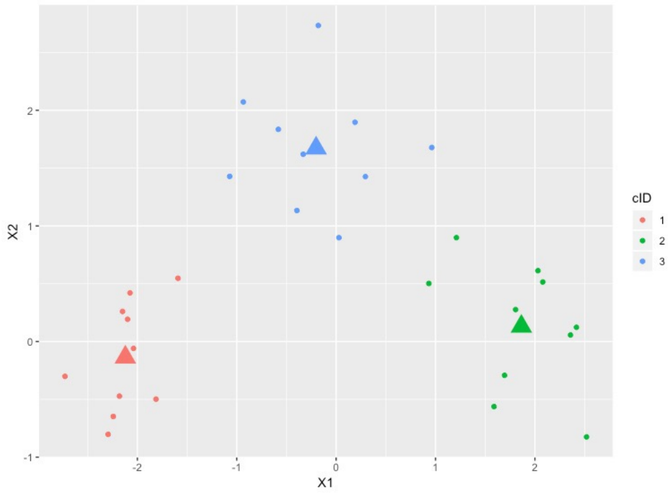
Clustering Validation

For supervised classification we have a variety of measures to evaluate how good the model is.

- Accuracy, precision, recall, F1-score

For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?

- To avoid finding patterns in noise
- To compare clustering algorithms
- To compare two sets of clusters
- To compare two clusters



Measures of Cluster Validity

Numerical measures to judge various aspects of cluster validity

Internal Index measures the goodness of a clustering structure without respect to external information.

- Sum of Squared Error (SSE)

External Index measures the extent to which cluster labels match externally supplied class labels.

- Entropy

Internal Measures: SSE

Sum of Squared Error (SSE) measures how closely related objects are in a cluster.

$$SSE = \sum_{i} \sum_{x \in Ci} (x - m_i)^2$$

where m_i is a centroid of a cluster C_i.

SSE is also known as the within cluster sum of squares (WSS)

Internal Measures: Cohesion and Separation

Cluster Cohesion measures how closely related objects are in a cluster, and the within cluster sum of square (WSS) can be used to quantify it.

$$WSS = \sum_{i} \sum_{x \in Ci} (x - m_i)^2$$

Cluster Separation measures how distinct or well-separated a cluster is from other clusters, and the between cluster sum of squares (BSS) can be used to quantify it.

$$BSS = \sum_{i} |C_i| \, (m - mi)^2$$

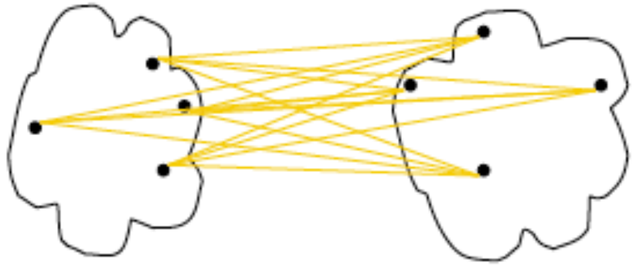
where $|C_i|$ is a size of a cluster C_i and m is is the centroid of all the samples.

Internal Measures: Cohesion and Separation

A proximity graph-based approach can be also used to measure cohesion and separation.

- Cluster cohesion is the sum of the weight of all links within a cluster.
- Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.

cohesion



separation

External Measures: Entropy and Purity

Entropy: For each cluster, the class *i* distribution of the data is calculated for cluster *j*, p_{ij} is the probability that a member of cluster *j* belongs to class *i*.

- Then the entropy of each cluster *j*,

$$e_j = \sum_i p_{ij} \, log_2(pij)$$

- The total entropy for a set of clusters is where m_j is the number of samples in cluster *j* and m is the total number of samples.

Purity: the purity of cluster *j*, $purity_j = max(p_{ij})$and the overall purity of a clustering,

$$purity = \sum_i \frac{m_i}{m} purity_j$$

DLI Accelerated Data Science Teaching Kit

Thank You