**BJ: to Lab-5, PUM: CLUSTERING**

**CLUSTERING** in Orange

**1. Clustering Definition:**

Clustering is *an unsupervised ML technique* that aims to group data points into clusters (sets) based on their similarities. These clusters share common characteristics that differentiate them from points in other clusters. Clustering helps uncover hidden patterns or structures within unlabeled data.

**2. Clustering vs. Classification:**

* **Clustering:**
  + Deals with unlabeled data.
  + Discovers inherent groupings without prior knowledge of classes.
  + Useful for exploratory data analysis and identifying potential categories.
* **Classification:**
  + Deals with labeled data (data points have predefined class labels).
  + Learns a model to predict class labels for new unseen data.
  + Used for tasks like spam detection or customer segmentation based on existing categories.

**3. Key Clustering Algorithms:**

* **K-Means Clustering:** A popular partitioning method that divides data points into a predefined number (k) of clusters. It iteratively assigns data points to the nearest cluster centroid (mean) and recomputes the centroids until a convergence criterion is met.
* **Hierarchical Clustering:** This method builds a hierarchy of clusters, either in a top-down (divisive) or bottom-up (agglomerative) fashion. Divisive clustering starts with all data points in one cluster and iteratively splits them into smaller subclusters. Agglomerative clustering starts with individual data points as clusters and iteratively merges the most similar clusters until a desired hierarchy is formed.
* **Density-Based Spatial Clustering of Applications with Noise (DBSCAN):** This method identifies clusters based on areas of high density (many data points close together) separated by areas of low density. It doesn't require specifying the number of clusters beforehand and can handle outliers (data points significantly different from others).
* **Expectation Maximization (EM) Clustering:** This method assumes that the data belongs to a mixture of statistical distributions and uses an iterative approach to identify these distributions and assign data points to them. It's particularly useful for clustering data with complex underlying structures.

**4. Clustering Algorithms in Orange:**

Orange offers several widgets for clustering tasks:

* **K-Means:** Implements the K-Means algorithm for partitioning data into user-specified k clusters.
* **Hierarchical Clustering:** Provides options for both divisive and agglomerative hierarchical clustering with various linkage criteria (methods for measuring similarity between clusters).
* **DBSCAN:** Implements the DBSCAN algorithm for density-based clustering.
* **Self-Organizing Maps (SOM):** While not strictly a clustering algorithm, SOM projects data points onto a lower-dimensional grid, allowing visualization of potential clusters based on their proximity on the grid.
* **Louvain Modularity Clustering**: Purpose: This algorithm identifies communities (clusters) within networks. A network can represent relationships between entities (e.g., friendships in a social network, connections between webpages). Communities in a network often represent groups of nodes (entities) with denser connections within the group compared to connections with nodes outside the group.
* Method: Louvain Modularity works by iteratively optimizing a modularity score. This score measures how well the network is divided into communities, with higher scores indicating better separation. The algorithm starts with each node in its own community and then iteratively moves nodes to different communities to improve the modularity score.

**5. Orange Widgets for Clustering:**

**5.1. Widgets:**

* **K-Means:**
  + **Parameters:**
    - **Number of Clusters (k):** Specify the desired number of clusters.
    - **Initialization:** Choose between random initialization (centroids placed randomly) and k-means++ (improved initialization for better convergence).
    - **Distance:** Select the distance metric (e.g., Euclidean, Manhattan) for measuring similarity between data points and centroids.
    - **Max Iterations:** Set the maximum number of iterations for the K-Means algorithm to run.
* **Hierarchical Clustering:**
  + **Parameters:**
    - **Method:** Choose between "Divisive" or "Agglomerative" clustering.
    - **Linkage:** Select the linkage criterion (e.g., Single, Complete, Average) for measuring similarity between clusters during merging or splitting.
    - **Number of Clusters:** Optionally specify the desired number of terminal clusters in the hierarchy.
* **DBSCAN:**
  + **Parameters:**
    - **Eps (Minimum Points Radius):** Define the minimum radius of a cluster (minimum distance between data points within a cluster).
    - **Min Samples:** Set the minimum number of data points required to form a cluster.

**5.2. Note:**

These are just some of the common parameters for each widget. Additional parameters might be available depending on the specific widget version.

**6. Suitable Datasets for Clustering in Orange:**

Orange can handle various data types for clustering, including numerical and categorical data. However, some considerations apply:

* **Data Preprocessing:** Ensure data is appropriately scaled or normalized if features have different scales. This ensures features contribute equally to the clustering process.
* **Feature Selection:** For high-dimensional data, consider feature selection techniques to focus on the most relevant features for clustering.
* **Data Exploration:** Techniques like visualization (scatter plots, histograms) can help identify potential clusters and inform the choice of clustering algorithm.

***Datasets for Unsupervised Clustering:***

Here are some Orange datasets specifically designed for unsupervised clustering exploration:

* **Aggregation:** This dataset represents data points generated from a mixture of Gaussian distributions. Clustering here can help identify these underlying distributions.
* **Blobs:** This dataset contains data points forming well-separated clusters of different shapes and sizes. It's a good example for testing and comparing different clustering algorithms.
* **Circles:** This dataset includes data points arranged in distinct circles. It's another helpful example for visualizing and understanding cluster separation.

**Clustering-v2** (c)

1. **Definition of Clustering paradigm in ML**: Clustering is an unsupervised learning technique in machine learning where the goal is to group similar data points together based on some similarity or distance metric. The aim is to discover inherent structures or patterns in the data without any predefined labels. Clustering algorithms partition the dataset into groups or clusters, with data points within the same cluster being more similar to each other than to those in other clusters.
2. **Key distinctions between Clustering and Classification in ML**:
   * Clustering is unsupervised learning, while classification is supervised learning.
   * In clustering, there are no predefined labels or target variables, whereas in classification, the algorithm learns to predict the labels of new instances based on labeled training data.
   * Clustering aims to group similar data points together, whereas classification aims to assign class labels to data points based on their features.
   * Clustering algorithms typically do not require a training phase, while classification algorithms require training on labeled data.
3. **Key Algorithms/methods used for Clustering**: Some common clustering algorithms/methods include:
   * K-means clustering
   * Hierarchical clustering (agglomerative and divisive)
   * DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
   * Gaussian Mixture Models (GMM)
   * Mean Shift clustering
   * Spectral clustering
   * *Louvain Clustering:* The Louvain Clustering algorithm is a method for detecting communities in large networks. It is primarily used for network analysis rather than traditional clustering of feature-based data. This algorithm is implemented in Orange as a widget for community detection and network analysis rather than general clustering of datasets. It's worth noting that community detection is a related but distinct problem from traditional clustering, and Louvain Clustering is tailored for this specific task.
4. **Algorithms implemented in Orange system**: Orange supports several clustering algorithms, including K-means, hierarchical clustering, DBSCAN, and spectral clustering.

5.1 **Widgets in Orange for clustering problems**: Orange provides several widgets for clustering analysis, including:

* K-Means
* Hierarchical Clustering
* DBSCAN
* Spectral Clustering

5.2 **Parameters used in these Widgets**:

* For K-Means: Number of clusters (k), initialization method, maximum iterations, convergence threshold, random seed.
* For Hierarchical Clustering: Linkage method (e.g., single, complete, average), distance metric (e.g., Euclidean, Manhattan).
* For DBSCAN: Epsilon (neighborhood radius), minimum samples (minimum number of points in a neighborhood to be considered a core point), distance metric.
* For Spectral Clustering: Number of clusters, affinity (nearest neighbors, nearest centroids), similarity measure (e.g., Euclidean, cosine), eigen solver.

1. **Datasets in Orange for clustering problems**: Not-labeled datasets for clustering in Orange:

For clustering purposes, Orange provides several datasets that are not labeled and can be used for unsupervised learning. Some of these datasets include:

* + Animals dataset: Contains features describing various animals.
  + Animals (Multi-instance) dataset: Similar to the Animals dataset but in a multi-instance format.
  + Mushroom dataset: Contains features describing mushroom samples.
  + Titanic dataset: Contains passenger information from the Titanic.
  + Votes dataset: Contains information about votes by members of the U.S. House of Representatives.

These datasets are suitable for clustering analysis in Orange because they do not have predefined labels, allowing for unsupervised exploration of the data's structure and patterns.