Unsupervised Learning

Enrique Alifio Ditya

Questions and Answers

What is Unsupervised Learning?

Unsupervised learning is a type of machine learning branch that learns patterns from unlabeled data. The goal of unsupervised learning is to find patterns in the data that can be used to group similar data points together. Unsupervised learning is often used for exploratory data analysis, and is also used as a preprocessing step for supervised learning tasks. These can be done using clustering algorithms, which are the focus of this repository.

How does each of the implemented algorithms work?

K-Means

1. Initialize the hyperparameters k (number of clusters) and max_iter (maximum number of iterations).
2. Randomly initialize k cluster centroids.
3. Repeat until convergence or max_iter is reached:
   (a) Assign each data point to the nearest cluster centroid.
   (b) Update the cluster centroids by taking the mean of all data points assigned to that cluster.
4. Return the cluster centroids and the cluster assignments.

K-Medoids

1. Initialize the hyperparameters k (number of clusters) and max_iter (maximum number of iterations).
2. Randomly initialize k cluster medoids.
3. Repeat until convergence or max_iter is reached:
   (a) Assign each data point to the nearest cluster medoid.
   (b) Update the cluster medoids by taking the medoid of all data points assigned to that cluster.
4. Return the cluster medoids and the cluster assignments.

DBSCAN

1. Initialize the hyperparameters eps (maximum distance between two data points to be considered as neighbors) and min_pts (minimum number of data points to form a dense region).
2. Randomly select a data point that has not been visited.
3. Retrieve all data points that are within eps distance from the selected data point.
4. If the number of data points retrieved is less than min_pts, mark the selected data point as noise and return to step 2.
5. Otherwise, mark the selected data point as a core point and assign all retrieved data points as neighbors of the selected data point.
6. Repeat until all data points have been visited.
7. Return the cluster assignments.

What are the advantages and disadvantages of each algorithm?

K-Means

Advantages

- Simple and easy to implement.
- Fast and efficient.
- Works well with large datasets.
- Works well with data that is well separated.

Disadvantages

- Requires the number of clusters to be specified beforehand (can use methods such as elbow method to find the sub-optimal number of clusters).
- Sensitive to outliers.
- Sensitive to initial cluster centroids.
- Struggles with non-linear or irregularly shaped clusters.

K-Medoids

Advantages

- Simple and easy to implement.
- Fast and efficient.
- Works well with large datasets.
- Works well with data that is well separated.

Disadvantages

- Requires the number of clusters to be specified beforehand (can use methods such as elbow method to find the sub-optimal number of clusters).
- Sensitive to outliers.
- Sensitive to initial cluster medoids.
- Struggles with non-linear or irregularly shaped clusters.

DBSCAN

Advantages

- Does not require the number of clusters to be specified beforehand.
- Can find arbitrarily shaped clusters.
- Robust to outliers.

Disadvantages

- Sensitive to the hyperparameters *eps* and *min_pts*.
- Does not work well with high-dimensional data.
- Does not work well with data that is not well separated.

K-Means vs K-Medoids vs DBSCAN

KMeans is often preferred when dealing with:

- Large datasets: Due to its faster computation, KMeans is more suitable for larger datasets.
- Gaussian-shaped clusters: KMeans performs well when clusters are relatively spherical and evenly distributed.
- Quick prototyping: When you need a fast and reasonably good clustering result, KMeans can be a good choice.

KMedoids is generally preferred when dealing with:

- Non-linear or irregularly shaped clusters: KMedoids is more robust in handling clusters with complex shapes and densities.
- Outliers: KMedoids is less affected by outliers because it uses actual data points as medoids.
- Guarantees convergence: If you need an algorithm that guarantees convergence to a local minimum, KMedoids is a better option.

DBSCAN is often preferred when dealing with:

- Arbitrary shaped clusters: DBSCAN can find clusters of any shape. It can even find a cluster completely surrounded by a different cluster.
- Outliers: DBSCAN is less affected by outliers because it uses the concept of density reachability.
- Guarantees convergence: If you need an algorithm that guarantees convergence to a local minimum, DBSCAN is a better option.

What are the applications of unsupervised learning in real life?

- Customer Segmentation: Grouping customers based on their purchasing behavior.
- Market Segmentation: Grouping customers based on their purchasing behavior.
- Document Clustering: Grouping documents based on their content.
- Anomaly Detection: Identifying fraudulent transactions.
- Image Segmentation: Grouping pixels based on their color.
- Topic Modeling: Grouping documents based on their content.
- And many more!