Enabling People
Education for Future Generations

AI Course

Chapter 6. Quiz

For students

Samsung Innovation Campus

1. Which of the following statements about learning is incorrect?

   ① Supervised learning uses both input data and learning data given a result, but unsupervised learning is used in the form of learning using only input data without learning data.

   ② Unsupervised learning can be said to be a learning method that informs the problem but does not tell the answer.

   ③ As an unsupervised learning algorithm, hierarchical clustering analysis and K-means clustering can be used for clustering.

   ④ K-means clustering has the advantage of being easy to determine the initial number of clusters and easy to interpret the results.

ANSWER: 4
Option 4 is incorrect. K-means clustering has the DISADVANTAGE of requiring the number of clusters (k) to be specified in advance, which is often NOT easy to determine. Determining the optimal number of clusters requires methods like the Elbow method or Silhouette analysis.

2. Which of the following statements about clustering is incorrect?

   ① The most commonly used clustering technique is K-Means.

   ② Hierarchical Clustering is decided after clustering without determining the number of groups.

   ③ The algorithm of hierarchical clustering has K-Medoids.

   ④ Clustering is a technique that allocates and analyzes groups based on the similarity or dissimilarity between each entity.

ANSWER: 3
Option 3 is incorrect. K-Medoids is a partitioning clustering algorithm (like K-Means), NOT a hierarchical clustering algorithm. Hierarchical clustering algorithms include agglomerative methods (bottom-up) and divisive methods (top-down).

3. The EM algorithm belongs to the greedy algorithm. Please explain the reason.

ANSWER:
The EM (Expectation-Maximization) algorithm can be considered a greedy algorithm because at each iteration, it makes locally optimal choices that maximize the expected log-likelihood. In the E-step, it computes the expected value of the log-likelihood given the current parameters. In the M-step, it finds parameters that maximize this expected log-likelihood. While these local improvements guarantee convergence to a local maximum, they do not guarantee finding the global optimum, which is characteristic of greedy algorithms.

4. Write three or more examples of clustering algorithm. (ex. Data analysis)

ANSWER: Three or more examples of clustering algorithm applications:
1. Customer Segmentation - Grouping customers based on purchasing behavior, demographics, or preferences for targeted marketing campaigns.
2. Image Segmentation - Partitioning digital images into multiple segments to identify objects, boundaries, or regions of interest in computer vision applications.
3. Document Clustering - Organizing large collections of documents or articles into topic groups for information retrieval and recommendation systems.
4. Anomaly Detection - Identifying outliers or unusual patterns in network traffic, fraud detection, or quality control in manufacturing.

5. Explain about the two methods of choosing optimal number of clusters when using k-means. Describe how to find the optimal number of clusters using the two methods.

ANSWER: Two methods for choosing optimal number of clusters in k-means:
1. Elbow Method: Plot the within-cluster sum of squares (WCSS) or inertia against different values of k. As k increases, WCSS decreases. The optimal k is at the "elbow" point where the rate of decrease sharply changes, indicating diminishing returns from adding more clusters.
2. Silhouette Analysis: Calculate the silhouette coefficient for different values of k. This coefficient measures how similar an object is to its own cluster compared to other clusters, ranging from -1 to 1. Higher average silhouette scores indicate better-defined clusters. The optimal k is where the silhouette score is highest.

6. Cluster and visualize sample data from the code below using the KMeans, AgglomerativeClustering, and DBSCAN algorithms. However, for K-Means and agglomerative clustering, set the number of clusters to 3.

```
from sklearn.datasets import make_blobs, make_moons
from sklearn.cluster import KMeans
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN
import pandas as pd
import matplotlib.pyplot as plt

X, label = make_blobs(n_samples=300, n_features=2, centers=3, cluster_std=5,
random_state=123)
plt.scatter(X[:,0], X[:,1], c=label, alpha=0.7)
plt.title('Dataset #1 : Original')
plt.show()
```

ANSWER: Code solution:

```
# 1. K-Means Clustering
kmeans = KMeans(n_clusters=3, random_state=123)
labels_kmeans = kmeans.fit_predict(X)
plt.scatter(X[:,0], X[:,1], c=labels_kmeans, alpha=0.7)
plt.title('K-Means Clustering')
plt.show()
# 2. Agglomerative Clustering
agg = AgglomerativeClustering(n_clusters=3)
labels_agg = agg.fit_predict(X)
plt.scatter(X[:,0], X[:,1], c=labels_agg, alpha=0.7)
plt.title('Agglomerative Clustering')
plt.show()
```

```
# 3. DBSCAN Clustering
dbscan = DBSCAN(eps=2, min_samples=5)
labels_dbscan = dbscan.fit_predict(X)
plt.scatter(X[:,0], X[:,1], c=labels_dbscan, alpha=0.7)
plt.title('DBSCAN Clustering')
plt.show()
```