1. **What is clustering in machine learning?**
2. A supervised learning technique
3. **A type of unsupervised learning technique**
4. A reinforcement learning technique
5. A regression technique
6. **Which of the following is an example of unsupervised learning?**
7. Linear regression
8. Decision tree classification
9. **K-Means clustering**
10. Support vector machines
11. **Which statement best describes the elbow method in K-Means clustering?**
12. **It helps determine the optimal number of clusters**
13. It calculates the distance between centroids and data points
14. It measures the silhouette score of each cluster
15. It finds the centroids that are farthest from each other
16. **What is a drawback of the K-Means clustering algorithm?**
17. It requires labeled data for training.
18. It doesn't work well with large datasets.
19. **It is sensitive to the initial choice of centroids.**
20. All of the above
21. **Which clustering algorithm uses the concept of medoids instead of centroids?**
22. K-Means
23. DBSCAN
24. Agglomerative Hierarchical Clustering
25. **K-Medoids**
26. **In hierarchical clustering, what does the "linkage criterion" determine?**
27. The number of clusters
28. The maximum distance between data points in a cluster
29. **The order in which clusters are merged**
30. The distance metric used to measure similarity
31. **In hierarchical clustering, what does the dendrogram represent?**
32. The final cluster assignments
33. The distance between data points
34. **The hierarchical merging of clusters**
35. The cluster centroids
36. In the context of clustering, what does the term "centroid" refer to?
37. **A data point at the center of a cluster**
38. The distance between two clusters
39. The density of a cluster
40. The number of data points in a cluster
41. Which clustering algorithm assigns each data point to the cluster with the nearest centroid?
42. **K-Means**
43. DBSCAN
44. Agglomerative Hierarchical Clustering
45. Gaussian Mixture Model
46. What is the primary drawback of hierarchical clustering when dealing with large datasets?
47. It produces inaccurate results.
48. **It is computationally expensive.**
49. It requires the number of clusters to be specified in advance.
50. It cannot handle non-numeric data.
51. Which algorithm is used for density-based clustering?
52. PCA
53. Agglomerative
54. **DBSCAN**
55. K-Means
56. Calculate Euclidean distance (7,5) and (9,8)
57. **3.61**
58. 3.32
59. 3.59
60. 3.46
61. Which of the following is required by K-means clustering?
62. defined distance metric
63. number of clusters
64. initial guess as to cluster centroids
65. **all of the mentioned**
66. Which of the following clustering requires merging approach?
67. Partitional
68. **Hierarchical**
69. Naive Bayes
70. None of the mentioned
71. In hard clustering, how many clusters does each data point belong to?
72. **One**
73. Two
74. Multiple
75. None
76. What is the primary objective of anomaly detection in data analysis?
77. To identify common patterns in the data
78. To classify data into predefined categories
79. **To detect unusual or rare data points**
80. To perform regression analysis
81. In DBSCAN, what is the significance of the "epsilon" parameter?
82. The minimum number of points required to form a cluster
83. The number of clusters to be formed
84. **The maximum distance between data points to be considered neighbors**
85. The number of iterations for convergence
86. Which clustering algorithm uses a proximity matrix to merge clusters?
87. K-Means
88. DBSCAN
89. **Agglomerative Hierarchical Clustering**
90. Gaussian Mixture Model
91. The parameter "minPts" in DBSCAN refers to:
92. The minimum distance between two points to be considered neighbors
93. **The minimum number of points required to form a cluster**
94. The maximum distance between two points to be considered neighbors
95. The density of the dataset
96. Which clustering metric is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)?
97. **Silhouette Score**
98. Davies-Bouldin Index
99. Calinski-Harabasz Index
100. Dunn Index
101. Identify the correct formula for calculating Fowlkes-Mallows Score(FMI):
102. FMI = TP / mean((TP + FP) \* (TP + FN))
103. **FMI = TP / sqrt((TP + FP) \* (TP + FN))**
104. FMI = TP / avg((TN + FP) \* (TP + FN))
105. FMI = TP / sum((TN + FN) \* (TP + FN))
106. Select the true about Fowlkes-Mallows Index:
107. Measures the silhouette of a clustering result.
108. Compares the density of points in different clusters.
109. Evaluates the stability of a clustering algorithm.
110. **Measures the similarity between two clustering solutions obtained after applying different clustering algorithms**
111. Measures the similarity between two clustering solutions obtained after applying different clustering algorithms
112. **Collective Anomalies**
113. Point anomalies
114. Contextual anomalies
115. Fundamental anomalies
116. Anomalies require fewer partitions to be separated from the rest of the data is concept of:
117. **Isolation Forest**
118. Decision tree
119. hierarchical clustering
120. DBSCAN
121. What is a dendrogram in hierarchical clustering?
122. A visualization of cluster centroids
123. A statistical test for cluster significance
124. **A tree-like diagram showing the hierarchy of clusters**
125. A measure of cluster purity
126. What is the Silhouette score used for in clustering?
127. Measuring the density of clusters
128. Calculating cluster centroids
129. Determining the number of clusters
130. **Evaluating the quality of clustering results**
131. Which clustering technique is suitable for finding clusters with irregular shapes and varying densities?
132. K-Means
133. Agglomerative Clustering
134. **DBSCAN**
135. PCA
136. What is the purpose of the "elbow method" in clustering?
137. **To determine the optimal number of clusters**
138. To calculate the average cluster size
139. To visualize the hierarchical structure of clusters
140. To measure the distance between centroids
141. What is the purpose of the "K" in K-Means clustering?
142. To represent the number of data points
143. To represent the number of features
144. **To represent the number of clusters**
145. To represent the number of iterations
146. Agglomerative hierarchical clustering starts with:
147. **One cluster per data point**
148. One cluster for all data points
149. A random cluster assignment
150. No initial clusters
151. In divisive hierarchical clustering, how does the process start?
152. By randomly dividing the data into two clusters
153. **With a single cluster containing all data points**
154. With each data point in its own cluster
155. By selecting the cluster with the highest variance
156. The Fowlkes-Mallows Index (FMI) is based on which statistical concept?
157. **Precision and recall**
158. Entropy
159. Variance
160. Principal component analysis
161. What does the Normalized Mutual Information (NMI)=0 means:
162. Correlation
163. **no mutual information**
164. Mutual sharing of information
165. Normalization
166. What does "k" represent in k-Means clustering?
167. **The number of clusters to form**
168. The number of features in the dataset
169. The distance metric used for clustering
170. The number of iterations for convergence
171. In k-Means clustering, what is the objective function that the algorithm tries to minimize?
172. Entropy
173. **Sum of squared distances (inertia)**
174. Cross-entropy
175. Mutual information
176. Which sklearn function is used to perform k-Means clustering?
177. sklearn.cluster.kMeans()
178. **sklearn.cluster.KMeans()**
179. sklearn.clustering.kmeans()
180. sklearn.cluster.kmeans()
181. What is the significance of the "inertia" attribute in sklearn's KMeans object?
182. It represents the number of clusters.
183. It represents the cluster centroids.
184. It represents the distance between data points.
185. **It measures the quality of clustering by sum of squared distances.**
186. How is the number of clusters (k) typically determined in k-Means clustering?
187. **By using the "Elbow method"**
188. By selecting the largest possible value
189. By using hierarchical clustering
190. By setting it to 1
191. What is the role of the cluster centroids in k-Means clustering?
192. They represent the data points in each cluster.
193. They are the initial guesses for cluster assignments.
194. **They represent the means of data points in each cluster.**
195. They are used to calculate the number of iterations.
196. Which parameter in sklearn's KMeans allows you to specify the maximum number of iterations for convergence?
197. n\_clusters
198. init
199. **max\_iter**
200. random\_state
201. In k-Means clustering, what does the "random\_state" parameter control?
202. The number of clusters
203. **The initial randomization of cluster centroids**
204. The distance metric used for clustering
205. The number of iterations for convergence
206. What is the Elbow Method primarily used for in machine learning?
207. **Clustering**
208. Regression
209. Classification
210. Dimensionality reduction
211. In the context of the Elbow Method, what is the "elbow"?
212. A curve in the data
213. **The bend or inflection point in a plot of the cost function**
214. A data point in the dataset
215. The angle between two vectors
216. What does a "sharp elbow" in the Elbow Method plot typically indicate?
217. The optimal number of clusters is not clear.
218. The data is not suitable for clustering.
219. T**here is a clear and distinct optimal number of clusters.**
220. The clustering algorithm is unstable.
221. What can happen if there is no clear "elbow" in the Elbow Method plot?
222. It means the clustering algorithm has a bug.
223. It indicates that the data is noisy and unreliable.
224. It suggests that there is no good clustering solution for the data.
225. It **implies that any number of clusters can be chosen**.
226. What is the main limitation of the Elbow Method?
227. It requires a large amount of computational resources.
228. It can only be used with supervised learning algorithms.
229. **It may not always produce a clear "elbow" point.**
230. It can only be applied to datasets with a small number of features.
231. In the context of clustering, what is a "medoid"?
232. The centroid of a cluster
233. The mode of a dataset
234. The median value of a feature
235. **The most representative data point within a cluster**
236. What is the primary advantage of using medoid-based clustering over centroid-based clustering like K-Means?
237. Medoids are easier to calculate.
238. **Medoids are more robust to outliers.**
239. Medoids can handle high-dimensional data.
240. Medoids produce smaller clusters.
241. What is the primary goal of anomaly detection?
242. To identify normal patterns in data
243. **To find outliers or anomalies in data**
244. To classify data into predefined categories
245. To perform dimensionality reduction
246. In anomaly detection, what are anomalies also known as?
247. Inliers
248. Normals
249. **Outliers**
250. Patterns
251. What is Isolation Forest primarily used for in anomaly detection?
252. Detecting contextual anomalies
253. **Detecting point anomalies**
254. Detecting both point and contextual anomalies
255. Clustering data
256. Which of the following best describes a point anomaly?
257. An anomaly that only occurs in specific contexts or situations
258. **An anomaly that deviates significantly from the majority of data points**
259. An anomaly that is only detectable when analyzing multiple attributes together
260. An anomaly that is not present in the dataset
261. In the context of Isolation Forest, how are anomalies identified?
262. By identifying data points with the highest density
263. **By isolating data points that require the fewest splits in the tree**
264. By comparing each data point to a predefined threshold
265. By clustering data points based on their attributes
266. Which of the following is an example of a contextual anomaly?
267. **A sudden spike in website traffic during a marketing campaign**
268. A fraudulent credit card transaction in a dataset of legitimate transactions
269. An outlier data point in a univariate dataset
270. A data point with a value significantly different from the mean
271. What is agglomerative clustering in machine learning?
272. **A clustering technique that starts with single data points and merges them into clusters iteratively.**
273. A technique that splits data into multiple clusters based on the similarity of data points.
274. A dimensionality reduction technique.
275. A technique for classifying data points into predefined categories.
276. In agglomerative clustering, which measure is used to determine the similarity between clusters during merging?
277. **Euclidean distance**
278. Cosine similarity
279. Jaccard coefficient
280. Mean squared error
281. In scikit-learn, what function is used to perform agglomerative clustering?
282. KMeans
283. **AgglomerativeClustering**
284. DBSCAN
285. PCA
286. Which linkage criterion in agglomerative clustering tends to produce spherical clusters?
287. Single linkage
288. Complete linkage
289. Average linkage
290. **Ward linkage**
291. Which agglomerative clustering linkage criterion is based on the average pairwise distance between data points in two clusters?
292. Single linkage
293. Complete linkage
294. **Average linkage**
295. Ward linkage
296. In agglomerative clustering, what is the initial state of each data point before merging?
297. **Each data point is a separate cluster.**
298. All data points are in a single cluster.
299. Data points are randomly assigned to clusters.
300. The clustering starts with predefined clusters.
301. What is the primary characteristic of single linkage clustering?
302. It merges clusters based on the average distance between their data points.
303. It merges clusters based on the maximum distance between their data points.
304. **It merges clusters based on the minimum distance between their data points**
305. It merges clusters randomly.
306. What is the primary characteristic of complete linkage clustering?
307. It merges clusters based on the average distance between their data points.
308. **It merges clusters based on the maximum distance between their data points.**
309. It merges clusters based on the minimum distance between their data points.
310. It merges clusters randomly.
311. What is the primary purpose of a dendrogram?
312. To display the individual data points in a dataset.
313. To show the distribution of data points within a cluster.
314. **To visualize the hierarchical structure of clusters.**
315. To calculate the average linkage between clusters.
316. What is DBSCAN primarily used for in clustering?
317. Clustering data based on the least-squares error.
318. **Identifying outliers and noise in the data.**
319. Classifying data into predefined categories.
320. Performing dimensionality reduction.
321. What does DBSCAN stand for?
322. **Density-Based Spatial Clustering of Applications with Noise**
323. Distance-Based Silhouette Coefficient Analysis Network
324. Dimension-Based Spectral Clustering with Anomalies
325. Density-Based Segmentation and Clustering of Attributes
326. In DBSCAN, how is a core point defined?
327. A data point that is farthest from all other data points.
328. A data point that is closest to the centroid of its cluster.
329. **A data point with at least a specified number of neighbors within a specified radius.**
330. A data point that is located at the center of the dataset.
331. In DBSCAN, what is the purpose of the epsilon (ε) parameter?
332. It specifies the maximum number of clusters.
333. It defines the minimum number of neighbors required to form a cluster.
334. **It sets the maximum distance between data points for them to be considered neighbors.**
335. It determines the dimensionality of the data.
336. What is the role of the 'labels\_' attribute in scikit-learn's DBSCAN implementation?
337. It stores the original dataset.
338. **It stores the cluster assignments for each data point.**
339. It stores the centroids of the clusters.
340. It stores the density values for each data point.
341. What does the '-1' label signify in the cluster assignments produced by DBSCAN?
342. **Noise or outliers**
343. Core points
344. Border points
345. Undefined cluster
346. What is the Silhouette score used for in clustering?
347. **To measure the quality of a clustering algorithm by assessing the compactness and separation of clusters.**
348. To calculate the total variance within a cluster.
349. To determine the number of clusters in a dataset.
350. To evaluate the performance of a classification algorithm.
351. The Dunn index measures:
352. Cluster compactness
353. Cluster separation
354. **Both compactness and separation of clusters**
355. The number of data points in each cluster
356. Which clustering metric assesses the average distance between data points within the same cluster?
357. Fowlkes-Mallows index
358. Silhouette score
359. **Davies-Bouldin index**
360. Rand index
361. The Rand index is used for:
362. Measuring the quality of a clustering algorithm
363. Evaluating the performance of a classification algorithm
364. **Comparing the similarity between two different clusterings**
365. Assessing the compactness of clusters
366. Which clustering metric ranges from -1 to 1, with higher values indicating better cluster quality?
367. **Silhouette score**
368. Dunn index
369. Davies-Bouldin index
370. Rand index
371. In the context of clustering metrics, what does "inertia" refer to?
372. The tendency of clusters to overlap
373. **The total variance within a cluster**
374. The number of clusters in a dataset
375. The density of data points within a cluster
376. What does a lower Davies-Bouldin Index value indicate?
377. **Better cluster quality with well-separated and compact clusters.**
378. Poor cluster quality with clusters that are close together and not compact.
379. The number of clusters in the dataset.
380. The total variance within a cluster.
381. What is the range of values for the Davies-Bouldin Index?
382. **0 to 1**
383. 0 to infinity
384. -1 to 1
385. -infinity to infinity
386. What does it signify if the Davies-Bouldin Index value is close to 1?
387. The clusters are well-separated and compact.
388. **The clusters are poorly separated and not compact.**
389. There are too many clusters in the dataset.
390. The clustering algorithm is not functioning correctly.