**60 questions sur l'optimisation et le tuning de l'apprentissage non supervisé**

1. Quelle technique est utilisée pour déterminer le nombre optimal de clusters dans K-means ?

A. Méthode du coude

B. Analyse en composantes principales

C. Validation croisée

D. Boosting

2. Quel paramètre de DBSCAN définit la distance maximale entre deux points pour qu'ils soient considérés comme voisins ?

A. MinPts

B. Epsilon (ε)

C. Rayon

D. Densité

3. Quelle métrique évalue la qualité des clusters en mesurant la cohésion intra-cluster et la séparation inter-cluster ?

A. Indice de Rand ajusté

B. Score de silhouette

C. Indice de Davies-Bouldin

D. Information mutuelle normalisée

4. Quelle technique de réduction de dimensionnalité est souvent utilisée avant le clustering pour visualiser les données en 2D ou 3D ?

A. Analyse discriminante linéaire

B. t-SNE

C. Analyse en composantes indépendantes

D. Factorisation en matrices non-négatives

5. Quel algorithme de clustering ne nécessite pas de spécifier le nombre de clusters à l'avance ?

A. K-means

B. Espérance-maximisation

C. DBSCAN

D. K-medoids

6. Quelle approche peut être utilisée pour optimiser les hyperparamètres d'un algorithme de clustering ?

A. Descente de gradient stochastique

B. Recherche par grille

C. Rétropropagation

D. Boosting

7. Quelle métrique mesure la similarité entre deux partitions de données, souvent utilisée pour comparer un clustering obtenu à une vérité terrain ?

A. Score de silhouette

B. Indice de Calinski-Harabasz

C. Indice de Rand ajusté

D. Coefficient de détermination

8. Quelle technique peut être utilisée pour gérer les valeurs aberrantes avant d'appliquer un algorithme de clustering ?

A. Normalisation min-max

B. Winsorisation

C. One-hot encoding

D. Augmentation de données

9. Quel algorithme de clustering est particulièrement sensible à l'initialisation des centroïdes ?

A. DBSCAN

B. Clustering hiérarchique

C. K-means

D. OPTICS

10. Quelle méthode peut être utilisée pour déterminer la valeur optimale de epsilon (ε) dans DBSCAN ?

A. Méthode du coude

B. Graphique des k plus proches voisins

C. Validation croisée

D. Boostrap aggregating

11. Quelle technique de visualisation peut aider à évaluer la qualité d'un clustering ?

A. Histogramme

B. Diagramme en boîte

C. Graphique de dispersion en couleur

D. Courbe ROC

12. Quel indice mesure le ratio entre la somme des distances intra-cluster et la somme des distances inter-cluster ?

A. Indice de Dunn

B. Indice de Davies-Bouldin

C. Indice de Calinski-Harabasz

D. Coefficient de silhouette

13. Quelle approche peut être utilisée pour déterminer le nombre optimal de composantes dans une analyse en composantes principales (ACP) ?

A. Méthode du coude

B. Critère de Kaiser

C. Analyse parallèle

D. Toutes les réponses ci-dessus

14. Quel algorithme de clustering est basé sur la densité et peut détecter des clusters de formes arbitraires ?

A. K-means

B. DBSCAN

C. Gaussian Mixture Model

D. Agglomerative Clustering

15. Quelle technique peut être utilisée pour réduire l'impact des valeurs aberrantes sur le clustering ?

A. Normalisation Z-score

B. Clustering spectral

C. Utilisation de la médiane au lieu de la moyenne

D. Augmentation de la dimensionnalité

16. Quel paramètre de K-means++ améliore l'initialisation des centroïdes par rapport à K-means standard ?

A. Nombre d'itérations

B. Tolérance de convergence

C. Méthode de sélection des centroïdes initiaux

D. Métrique de distance

17. Quelle approche peut être utilisée pour gérer les clusters de tailles très différentes ?

A. Augmentation de la dimensionnalité

B. Utilisation de K-means++

C. Normalisation des données

D. Clustering hiérarchique agglomératif

18. Quelle technique peut être utilisée pour évaluer la stabilité des clusters obtenus ?

A. Validation croisée

B. Bootstrapping

C. Analyse de sensibilité

D. Toutes les réponses ci-dessus

19. Quel algorithme de clustering est particulièrement adapté aux données de grande dimension ?

A. DBSCAN

B. K-means

C. Clustering hiérarchique

D. Subspace clustering

20. Quelle approche peut être utilisée pour déterminer l'importance relative des features dans un clustering ?

A. Analyse en composantes principales

B. Sélection de features basée sur la variance

C. Clustering avec pondération des features

D. Toutes les réponses ci-dessus

21. Quelle technique peut être utilisée pour visualiser la structure hiérarchique des clusters ?

A. Dendrogramme

B. Heatmap

C. t-SNE

D. Graphique de dispersion

22. Quel algorithme de clustering est basé sur la maximisation de l'espérance ?

A. K-means

B. DBSCAN

C. Gaussian Mixture Model

D. Agglomerative Clustering

23. Quelle approche peut être utilisée pour gérer les données catégorielles dans le clustering ?

A. One-hot encoding

B. Encodage ordinal

C. Utilisation de métriques de distance adaptées (ex: distance de Gower)

D. Toutes les réponses ci-dessus

24. Quelle technique peut être utilisée pour accélérer K-means sur de grands ensembles de données ?

A. Mini-batch K-means

B. K-means++

C. Elkan's K-means

D. Toutes les réponses ci-dessus

25. Quel paramètre de DBSCAN détermine le nombre minimum de points pour former un cluster dense ?

A. Epsilon (ε)

B. MinPts

C. Rayon

D. Densité

26. Quelle approche peut être utilisée pour déterminer le nombre optimal de clusters dans un clustering hiérarchique ?

A. Méthode du coude

B. Critère d'information bayésien (BIC)

C. Analyse du dendrogramme

D. Toutes les réponses ci-dessus

27. Quelle technique de réduction de dimensionnalité préserve les distances locales entre les points ?

A. Analyse en composantes principales (ACP)

B. t-SNE

C. UMAP

D. Analyse factorielle

28. Quel algorithme de clustering est particulièrement sensible à l'échelle des features ?

A. DBSCAN

B. K-means

C. Clustering hiérarchique avec liaison complète

D. Toutes les réponses ci-dessus

29. Quelle approche peut être utilisée pour évaluer la qualité d'un clustering en l'absence de vérité terrain ?

A. Indice de Rand ajusté

B. Score de silhouette

C. Information mutuelle normalisée

D. F1-score

30. Quelle technique peut être utilisée pour gérer les données manquantes dans le clustering ?

A. Suppression des observations avec données manquantes

B. Imputation par la moyenne/médiane

C. Utilisation d'algorithmes tolérants aux données manquantes

D. Toutes les réponses ci-dessus

31. Quel algorithme de clustering est basé sur la connectivité des points ?

A. K-means

B. DBSCAN

C. Gaussian Mixture Model

D. Agglomerative Clustering

32. Quelle approche peut être utilisée pour déterminer l'importance des clusters obtenus ?

A. Analyse de la taille des clusters

B. Évaluation des centroïdes des clusters

C. Analyse des silhouettes des clusters

D. Toutes les réponses ci-dessus

33. Quelle technique peut être utilisée pour visualiser des clusters dans des espaces de haute dimension ?

A. Projection sur les deux premières composantes principales

B. t-SNE

C. UMAP

D. Toutes les réponses ci-dessus

34. Quel paramètre de K-means contrôle la convergence de l'algorithme ?

A. Nombre d'itérations maximum

B. Tolérance

C. Nombre de redémarrages

D. Métrique de distance

35. Quelle approche peut être utilisée pour gérer les clusters de formes non sphériques ?

A. Utilisation de K-means avec noyau

B. DBSCAN

C. Clustering spectral

D. Toutes les réponses ci-dessus

36. Quelle technique peut être utilisée pour évaluer la stabilité des résultats de clustering face à de petites perturbations des données ?

A. Validation croisée

B. Bootstrapping

C. Ajout de bruit gaussien

D. Toutes les réponses ci-dessus

37. Quel algorithme de clustering est particulièrement adapté aux données de forme sphérique ?

A. DBSCAN

B. K-means

C. Clustering hiérarchique avec liaison simple

D. OPTICS

38. Quelle approche peut être utilisée pour déterminer le nombre optimal de composantes dans un modèle de mélange gaussien ?

A. Critère d'information d'Akaike (AIC)

B. Critère d'information bayésien (BIC)

C. Validation croisée

D. Toutes les réponses ci-dessus

39. Quelle technique peut être utilisée pour accélérer DBSCAN sur de grands ensembles de données ?

A. Utilisation d'structures de données d'indexation spatiale

B. HDBSCAN

C. Échantillonnage des données

D. Toutes les réponses ci-dessus

40. Quel paramètre de t-SNE contrôle l'équilibre entre la préservation de la structure locale et globale ?

A. Perplexité

B. Nombre d'itérations

C. Taux d'apprentissage

D. Dimension de sortie

41. Quelle approche peut être utilisée pour gérer les outliers dans le clustering ?

A. Utilisation de DBSCAN

B. Prétraitement des données pour supprimer les outliers

C. Utilisation de méthodes robustes (ex: K-medoids)

D. Toutes les réponses ci-dessus

42. Quelle technique peut être utilisée pour évaluer la qualité d'un clustering hiérarchique ?

A. Coefficient de corrélation cophénétique

B. Indice de Calinski-Harabasz

C. Score de silhouette

D. Toutes les réponses ci-dessus

43. Quel algorithme de clustering est basé sur la décomposition spectrale de la matrice de similarité ?

A. K-means

B. DBSCAN

C. Clustering spectral

D. Agglomerative Clustering

44. Quelle approche peut être utilisée pour déterminer l'importance relative des features dans un modèle de mélange gaussien ?

A. Analyse des variances des composantes

B. Analyse des poids des features dans la log-vraisemblance

C. Sélection de features basée sur l'information mutuelle

D. Toutes les réponses ci-dessus

45. Quelle technique peut être utilisée pour visualiser la distribution des distances entre les points dans un ensemble de données ?

A. Histogramme des distances

B. Graphique de dispersion des distances

C. Courbe des k plus proches voisins

D. Toutes les réponses ci-dessus

46. Quel paramètre de l'algorithme OPTICS contrôle la taille minimale des clusters ?

A. Epsilon (ε)

B. MinPts

C. Xi

D. Angle maximum

47. Quelle approche peut être utilisée pour gérer les données temporelles dans le clustering ?

A. Utilisation de métriques de distance adaptées (ex: DTW)

B. Extraction de caractéristiques temporelles

C. Clustering de séries temporelles

D. Toutes les réponses ci-dessus

48. Quelle technique peut être utilisée pour évaluer la stabilité des résultats de clustering face à différentes initialisations ?

A. Consensus clustering

B. Bootstrapping

C. Validation croisée

D. Toutes les réponses ci-dessus

49. Quel algorithme de clustering est particulièrement adapté aux données de grande dimension et faible densité ?

A. DBSCAN

B. K-means

C. Subspace clustering

D. Agglomerative Clustering

50. Quelle approche peut être utilisée pour déterminer le nombre optimal de clusters dans un clustering spectral ?

A. Analyse des valeurs propres de la matrice laplacienne

B. Méthode du coude sur l'inertie

C. Optimisation du score de silhouette

D. Toutes les réponses ci-dessus

51. Quelle technique peut être utilisée pour accélérer le clustering hiérarchique sur de grands ensembles de données ?

A. Clustering hiérarchique par échantillonnage

B. Utilisation d'algorithmes de liaison rapide

C. Approximation des distances

D. Toutes les réponses ci-dessus

52. Quel paramètre de UMAP contrôle l'équilibre entre la préservation de la structure locale et globale ?

A. Nombre de voisins

B. Distance minimale

C. Métrique

D. Dimension de sortie

53. Quelle approche peut être utilisée pour gérer les données mixtes (numériques et catégorielles) dans le clustering ?

A. Utilisation de métriques de distance adaptées (ex: distance de Gower)

B. Encodage des variables catégorielles

C. Clustering sur des sous-ensembles homogènes de features

D. Toutes les réponses ci-dessus

54. Quelle technique peut être utilisée pour évaluer la qualité d'un clustering basé sur la densité ?

A. DBCV (Density-Based Clustering Validation)

B. Score de silhouette

C. Indice de Calinski-Harabasz

D. Indice de Davies-Bouldin

55. Quel algorithme de clustering est basé sur la maximisation de la marge entre les clusters ?

A. K-means

B. Support Vector Clustering

C. DBSCAN

D. Gaussian Mixture Model

56. Quelle approche peut être utilisée pour déterminer l'importance des features dans un clustering spectral ?

A. Analyse des vecteurs propres de la matrice laplacienne

B. Sélection de features basée sur la variance

C. Analyse de sensibilité des résultats du clustering

D. Toutes les réponses ci-dessus

57. Quelle technique peut être utilisée pour visualiser la structure des clusters dans des espaces de haute dimension ?

A. Projection sur les composantes principales

B. t-SNE

C. UMAP

D. Toutes les réponses ci-dessus