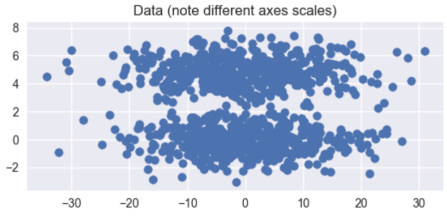
Fiche d'autoévaluation - 06

May 11, 2022

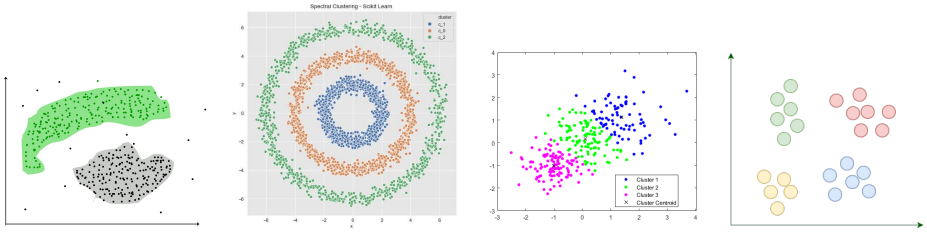
1. Pour faire de la détection d'anomalies, dans quel cas faut-il utiliser les techniques d'apprentissage supervisé et dans quel cas non supervisé?
2. Définissez (avec une formule) la variation within-cluster
3. Peux-je appliquer le partitionnement des données (clustering) par l'algorithme des K-moyennes (K-means) sur le dataset suivant? Faut-il faire du prétraitement d'abord? Si oui, lequel? Pourquoi?[1]



4. Comment intialise-t-on les centroides dans la méthodes des K moyennes?
5. Considérez le dataset suivant $\mathcal{D} = \{A = (0,0), B = (2,2), C = (1,2), D = (4,1), E = (2,4)\}$. Réalisez les passages de l'algorithme des K-moyennes, avec $K = 3$ clusters. Initialisez les centroides comme vous préférez. Utilisez un papier, un stylo et une calculette; à chaque itération, annotez l'inertie totale (total within cluster variation or total inertia), ainsi que la position des trois centroides.
6. Comment peut-on évaluer la qualité d'un partitionnement des données (clustering)? Donnez au moins deux exemples de métriques.
7. Quel est la meilleure et la pire valeur de silhouette?
8. Écrivez la formule pour calculer la silhouette.
9. Expliquez comment peut-on utiliser la méthodes des K-moyennes pour faire de la détection d'anomalies.

[1] Picture from `https://stats.stackexchange.com/a/283941/161064`

10. Considérez les partitionnements de données (clusterings) suivants :[2]



Lequel peut-on obtenir par l'algorithme des K-moyennes?

11. Quand on fait de la détection d'anomalies à l'aide de techniques d'apprentissage non supervisées, est-il utile d'avoir des exemples étiquettes (labeled samples)? Pourquoi?

12. Dessinez un exemple de courbe précision-rappel (precision-recall curve). Expliquez ce qu'on a dans l'axe des x et des y. Écrivez la formula correspondante.

13. Répétez l'exercice pour la courbe ROC.

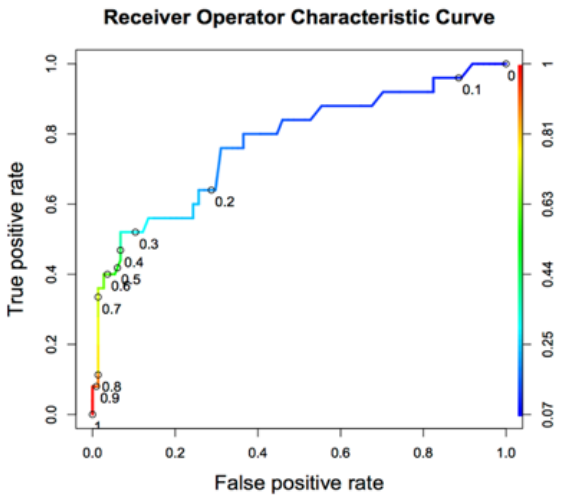
14. La courbe ROC et la courbe précision-rappel (precision-recall), ont-elles une métrique en commun?

15. Considérez un détecteur d'anomalies et faites varier la seuil de score d'anomalie (anomaly score threshold) τ. Si vous augmentez τ, est-ce que la précision augment ou diminue? Et le rappel (recall)? Et le tau de faux positifs (false positive rate)? Et le tau de vrai positif (true positive rate)?

16. Considérez la courbe ROC suivante[3]

[2] Pictures from `https://www.geeksforgeeks.org/clustering-in-machine-learning/`, `https://www.kdnuggets.com/2020/05/getting-started-spectral-clustering.html`, `https://fr.mathworks.com/help/stats/kmeans.html`, `https://www.baeldung.com/java-k-means-clustering-algorithm`

[3] Picture from `https://www.chegg.com/homework-help/questions-and-answers/given-roc-curve-threshold-would-pick-wanted-correctly-identify-small-group-patients-receiv-q34274988`



Discutez dans quelle situation vous choisiriez un seuil (threshold) τ entre 0.1 et 0.2 and dans quelle autre situation vous choisiriez un seuil de 0.3.

17. Si vous devez comparer deux détecteurs d'anomalies, comment pouvez-vous dire si l'un est meilleur que l'autre.

18. Écrivez la formule du score d'anomalie (anomaly score) quand on utilise Isolation Forest.

19. Comment calcule-t-on le score d'anomalie quand on utilise un auto-encoder?

20. Supposes de réaliser un détecteur d'anomalies basé sur K-moyennes. Le premier hyper-paramètre à décider est la valeur de K. Pour ce faire, vous pouvez vous baser sur des métriques. Donnez des exemples. Est-il correct de calculer ces métriques sur tout le jeu de données, sans le partitionner en données d'entraînement et de test? Pourquoi?

21. Pourquoi on ne peut pas utiliser les forêts aléatoires (random forests) à la place des extra trees dans les forêts d'isolation (random forests), pour faire de la détection d'anomalies non-supervisée?