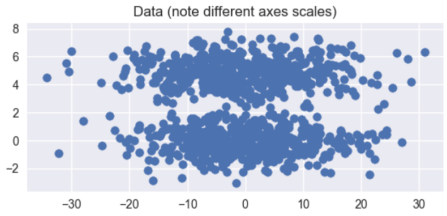
06 - Self Evaluation Questions

May 11, 2022

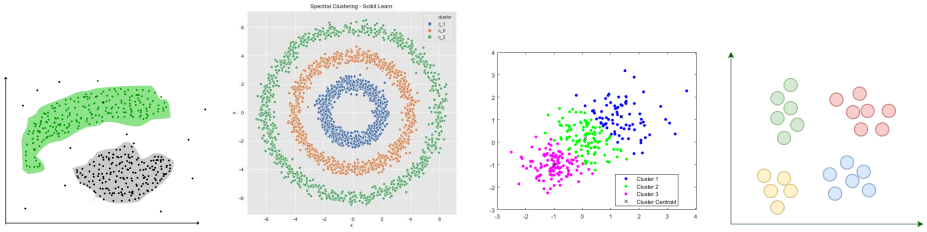
1. For solving an anomaly detection task, in which cases can you apply supervised learning techniques? And unsupervised learning techniques?
2. Define (with formula) the within-cluster variation
3. Can I apply K-means clustering to the following dataset? Do we need to do some pre-processing before? If yes, which one? Why?[1]



4. In k-means clustering, how do you initialize the centroids?
5. Consider a dataset with the following points $\mathcal{D} = \{A = (0,0), B = (2,2), C = (1,2), D = (4,1), E = (2,4)\}$. Perform K-means clustering with $K = 3$ clusters, initializing a triplet of centroids as you like. Use a paper, a pen and a calculator to do the exercise; At each iteration, annotate the total within cluster variation (or total inertia), as well as the updated position of the three centroids.
6. In K-means clustering, how do you fix the number K of clusters?
7. How can you judge the quality of clustering? Give at least two examples of appropriate metrics.
8. What is the best and the worst silhouette value?
9. Write the formula of the silhouette.
10. Explain how k-means clustering can be used for anomaly detection.

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

11. Consider the following clusterings:[2]



Which one(s) can be obtained via K-means algorithm?

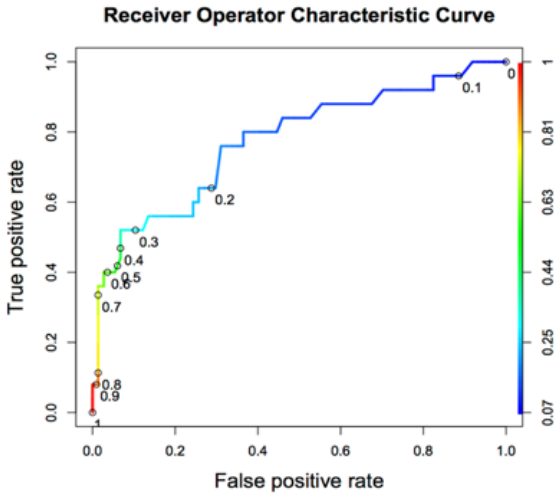
12. When doing anomaly detection with unsupervised learning, are labeled samples useful? Why?

13. Draw an example of precision and recall curve. Explain what's in the x-axis and y-axis. Write the corresponding formula.

14. Do the same for the ROC curve.

15. Suppose you have an anomaly detector and you vary the anomaly score threshold τ. If you increase τ, does the precision increase or decrease? And the recall? And the false positive rate? And the true positive rate?

16. Consider the following ROC curve[3]



Discuss in which scenario you would choose a threshold between 0.1 and 0.2 and in which other scenario you would choose a threshold of 0.3.

[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

17. If you have two anomaly detectors, in which way can you tell if one is better than the other?

18. Write the formula of the anomaly score of a sample, when using isolation forest.

19. How is the anomaly score calculated when using autoencoders.

20. Suppose you want to create an anomaly detector based on K-means. The first hyper-parameter you need to do is to decide the value of K. In order to do so, you can measure some appropriate metrics. Give some example of metrics. Is it correct to compute such metrics on the entire dataset, without dividing it into training/test data? Why?

21. Why cannot we use random forests instead of extra trees within isolation forests to perform unsupervised anomaly detection?