Hide menu

Case Study

Reading: Clustering Analysis Case Study - Demo
2h

Discussion Prompt: Clustering Analysis Exploration Exercise
2h

Graded Assignment: Self Reflection
Started

Self Reflection

Review Learning Objectives

Assignment details

Due
Mar 31, 11:59 PM IST

Attempts
Unlimited

Your grade

You haven't submitted this yet. We keep your highest score.

--

Like Dislike Report an issue

Instructions

1. Reflecting on the case study, what was the most challenging aspect of applying clustering analysis to solve the real-world problem? How did you overcome this challenge, and what did you learn from it? **1 point**

Moreover, choosing threshold (epsilon) for DBSCAN is also a challenge.

Here the data contains the ground-truth labels, hence matching the cluster-labels output by the algorithm with the ground-truth, we can intuitively create most probable mapping between them and then compute the accuracy of clustering. Although Silhouette coefficient gives some idea about the quality of the clusters, whether different / same species belong to different / same cluster respectively, we can compute it using the GT labels.

Your answer cannot be more than 10000 characters.

Resume

2. Describe a situation where you encountered technical problems while applying clustering analysis to the case study data. How did you troubleshoot and resolve these issues to ensure accurate results? **1 point**

While selecting DBSCAN neighborhood threshold (eps) there was the following problem: firstly, the threshold itself is very much dependent on whether or not the features were z-score-normalized and secondly, choosing slightly higher threshold may result in outputting a single cluster, leading to an error. Also, Visualization on two dimensions / using principal components were not enough in separating the clusters, we used TSNE instead, which produced much better result. Also, since DBSCAN does not take an input parameter specifying the number of output clusters to be produced, it generates arbitrary numbers of clusters, and by controlling eps the number of clusters is brought down to 3, equaling number of species.

Your answer cannot be more than 10000 characters.

3. Reflect on the interpretation of the clustering model results. How did you derive actionable insights from the model outcomes to make data-driven decisions for the real-world problem? **1 point**

This is a challenging part. The clustering algorithms kmeans and agglomerative clustering accepts the number of clusters to be produced and specifying 3 output clusters and comparing them with the GT labels, we can create a mapping between the cluster label and the GT species name (e.g., comparing the cluster sizes), and then use the mapping to compute the number of datapoints correctly labeled.

Your answer cannot be more than 10000 characters.

4. How did the application of clustering analysis in the case study scenario enhance your critical thinking and problem-solving skills? Provide specific examples of how clustering analysis aided you in making informed decisions [Practice this question as if you were in an interview!] **1 point**

Comparing the GT species names with the cluster labels outputs produced by the clustering algorithms itself requires some thinking in terms of best possible mapping of the labels with the species names (based on comparing the respective cluster sizes and the GT species size, finding the best match). Silhouette coefficient indicates the cluster quality but we still can use the GT to compute the accuracy of clustering.

Your answer cannot be more than 10000 characters.

5. What were the most valuable lessons you learned from completing the case study? How do you plan to apply these insights to further develop your clustering analysis skills and grow as a data analyst? **1 point**

Evaluation of unsupervised learning techniques such as clustering is more difficult than supervised ones, because of the absence of ground-truth labels. We need to rely on the measures such as Silhouette coefficient, which may not be always the best possible measure. Whenever we have the GT available (at least for the training dataset), we can always use the GT labels to check the accuracy of clustering and then try to improve by using additional / less features or applying some feature transformation.

Your answer cannot be more than 10000 characters.

Coursera Honor Code Learn more

I, **SANDIPAN DEY**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Submit **Save draft**

Last saved on Feb 24, 2:54 AM IST

Like Dislike Report an issue