INFOB2DA | Practical Assignment 2

Clustering methods and distance functions (Total 100 Points)

Utrecht University | Visualization and Graphics Group
Dr. Michael Behrisch

Alister Machado, Mohamed Ali Gaidi, Ahlam Abdelkhalki, Najoua Ouaali, Sita Newer

Submission deadline:
Sunday, 29.09.2024, 23:59.

General information

You must form groups of 3 students. Individual submissions are only accepted in special cases. Each group member must understand the entire assignment, including the code which you will create!

- Submission deadline: Sunday, 29.09.2024, 23:59pm. You will have to present your submissions on Monday, 30.09.2024 in the regular exercise/werkcollege slot. If you are unable to present your work there, send us an email.
- If you have questions or need more information, you can always use the Questions channel on Teams, or our Office Hours on Friday the 20th and 27th of October.
- You are only allowed to use the Python programming language in your code.
- For this practical assignment, you can achieve 100 points in total. Your overall practical assignments grade will be determined by the sum of your points in PA1-4.

Handing in your assignment

Hand in the Jupyter Notebook, dataset, notebook checkpoints (the folder which is automatically created in your project by Jupyter) and presentation slides as a ZIP-file on MS Teams. Make sure you press the submit/inleveren button after uploading your files in Teams, otherwise the submission is not completed.

Introduction

The aim of this practical assignment is to make use of unsupervised machine learning algorithms and specifically clustering algorithms for answering the following type of question: "*Which data-imposed groupings exist in the dataset at hand.*"

After this assignment you will be able to …

1) … reflect on the importance of distance functions for clustering algorithms
2) … select an appropriate clustering algorithm given a dataset
3) … judge the data quality's impact on the ML result
4) … judge the performance of clustering algorithms.
5) … compare the performance of various clustering approaches
6) … have a structured approach to solve clustering problems.

Dataset overview

The online_shoppers_intention.csv file contains shopping intention metrics for 12330 different shoppers.

Feature	Description
Administrative	Int, the number of pages of this type visited by the visitor in that session.
Administrative_Duration	Float, the total time spent in this page category.
Informational	Int, the number of pages of this type visited by the visitor in that session.
Informational_Duration	Float, the total time spent in this page category.
ProductRelated	Int, the number of pages of this type visited by the visitor in that session.
ProductRelated_Duration	Float, the total time spent in this page category.
BounceRates	Float, the percentage of visitors who enter the site from that page and then leave.
ExitRates	Float, the percentage that a specific page was the last in the shopper's session.
PageValues	Float, the average value for a web page that a user visited before completing a transaction.
SpecialDay	Float, the closeness of a special day, such as Valentine's.
Month	Str, which month they visited.
OperatingSystem	Int, encoded, the kind of operating system used.
Browser	Int, encoded, the kind of browser used.
Region	Int, encoded, the region of the shopper.
TrafficType	Int, encoded, the type of traffic.
VisitorType	Str, returning or new visitor.
Weekend	Bool, whether it was the weekend or not.

Revenue	**Class label**, bool, if they spend money after visiting.

Task 0: Setup Environment (0 Points)

Result: a ready to go Jupyter notebook.
Recommend software and libraries: Python, Visual Studio Code, Jupyterlab

This assignment uses the same setup as specified in the first assignment.

Task 1: Get dataset on screen (11 Points)

Goal: After you have achieved this task, you are able to explore the dataset with basic functions and visualize important features.
Graded result: The result of question 1.1 must only be included in the presentation. The result of question 1.2 must be visible in both the notebook and the presentation.
Recommended libraries: Pandas, Plotly

- Use Pandas to import the online shopper's intention dataset.

1.1 Explore the dataset as you did in the first assignment, so you have a good understanding of its content and features. Come up with an interesting one-minute story about the dataset. **(5 points)**

1.2 Create a visualization which shows the characteristics of online shoppers, who use ‘browser 13’ compared to online shoppers who use other browsers. **(6 points)**

Task 2: Preprocessing (0 Points)

Goal: After you have achieved this task, you are able to transform the data into a suitable input for clustering algorithms.
Result: A preprocessed dataset.
Recommended libraries: Pandas, Sklearn

- Think about the data and its shape which should be used as the input for clustering algorithms. Use preprocessing techniques which you think are necessary (hint: look at the upcoming algorithms and their documentation).

Task 3: Clustering algorithms (9 Points)

Goal: After you have achieved this task, you are able to successfully use clustering algorithms.
Graded result: The results of questions 3.1-3.3 must be visible in both the notebook and the presentation.
Recommended libraries: Sklearn, Plotly

- For the next questions, use the preprocessed dataset of task 2.

3.1 Apply Sklearn's Affinity Propagation clustering. Visualize the created clusters. **(3 points)**

3.2 Apply Sklearn's DBSCAN clustering. Visualize the created clusters. **(3 points)**

3.3 Apply Sklearn's Birch clustering. Visualize the created clusters. **(3 points)**

Task 4: Evaluation of clustering methods (20 Points)

Goal: After you have achieved this task, you are able to evaluate clustering models using different evaluation measures and reason on their performance.
Graded result: The results of questions 4.1-4.3 must be visible in both the notebook and the presentation. Code submission and documentation of the main parts.
Recommended libraries: Sklearn, Plotly

4.1 Manually implement the Silhouette score to evaluate the results of all clustering models. You are not allowed to use any predefined functions or libraries, except for small functionalities, such as the factorial function of the Math library. **(15 points)**

4.2 Use Sklearn's Davies Bouldin Score to evaluate the results of all clustering models. **(2,5 points)**

4.3 Use Sklearn's Calinski-Harabasz Index to evaluate the results of all clustering models. **(2,5 points)**

Task 5: Distance functions (30 Points)

Goal: After you have achieved this task, you are able to understand distance functions and be able to reason on the impact for the different clustering algorithms.
Graded result: The results of questions 5.1-5.4 must be visible in both the notebook and the presentation. Code submission and documentation of the main parts.
Recommended libraries: Sklearn, Plotly

- For the next questions, use only the DBSCAN clustering algorithm. In preparation for the following sub questions, explore the different parameters DBSCAN has to offer to create a clustering which separates the main data from the outliers.

5.1 Manually implement a Euclidean distance function and use it as a parameter of the DBSCAN algorithm. When creating the distance function, you are not allowed to use any predefined functions or libraries. (7,5 points)

5.2 Manually implement a Manhattan distance function and use it a parameter of the DBSCAN algorithm. You are not allowed to use any predefined functions or libraries, except for core functionalities, such as the factorial function of the Math library. (7,5 points)

5.3 Manually implement a Cosine similarity distance function and use it as a parameter of the DBSCAN algorithm. You are not allowed to use any predefined functions or libraries, except for core functionalities, such as the factorial function of the Math library. (7,5 points)

5.4 Evaluate the effect which the different distance functions have on the DBSCAN algorithm. To evaluate, use any of the evaluation methods of the previous task. Visualize the results to support your arguments. (7,5 points)

Task 6: Reflection (30 Points)

Goal: Practice to communicate your findings to domain-experts and domain-novices.

Graded result: Presentation (*max 15 minutes per group including Q&A from TAs; we accompany more presentation slots to achieve the full 15min/group*) and presentation quality/structure, PDF with screenshots along with descriptive text (e.g., annotations, captions, bullet points to highlight specific findings), data and performance tables, code quality and cleanness, knowledge of each team member of the code (make sure to have the code running during the presentation/Q&A).

Guiding Questions (meant solely as a loose guideline on the kind of questions you should answer in your presentation; list is non-exclusive; does not impose any order; answering all questions is not obligatory for full points; we intentionally give less guidelines from PA2 onwards):

∉ What defines a good clustering?
∉ Compare the different algorithm parameter choices.
∉ How to effectively evaluate clustering algorithms? Compare the different choices.
∉ What is the difference between the distance functions?