Assignment 2
Cluster Analysis

Submission Instructions

- This assignment is due before **11:59pm** on **Monday, 10th April 2023.**
- Your submission should be a written report (.**pdf** only, word docs not accepted)
- Please note, incorrect file types will not be opened or graded.
- Your submission should use the following name convention:
    - {student_number}_assignment_2_report.pdf
- If you submit multiple times, only the most recent submission is kept by Brightspace. The latest submission will be the graded attempt.
- You should **not** upload a new attempt after the deadline as this will override your previous submission and your assignment will be considered late.
- This assignment is ***individual*** and thus your submission should be unique. The UCD plagiarism policy applies, see here for the student guide: https://www.ucd.ie/secca/t4media/plagiarism_studentguide.pdf

Video Submissions

You are welcome to submit a video instead of a written report. Your video should begin with you saying your name and student number. You should then say the question name, along with the question answer.

The same grading rubric will be used as for the written reports.

Late Submission

- Late submissions will be graded according to UCD's late submission policy which you can find here:
  - https://hub.ucd.ie/usis/!W_HU_MENU.P_PUBLISH?p_tag=GD-DOCLAND&ID=137
  - Assignments received any time within one week after the due date will have the grade awarded reduced by two grade points (e.g. An A- becomes a B). Work received after that but within two weeks of the due date will have the grade reduced by four grade points (e.g. An A- becomes a C+). Work submitted later than two weeks after the due date cannot be accepted and a student will not pass that particular assessment.
- For information on extenuating circumstances, please refer to the UCD site: https://www.ucd.ie/students/studentdesk/extenuatingcircumstances/

Grading

- Grading scheme used is the *Standard Conversion Grade Scale* 40% Pass (70% = A-)*
- This assignment will be worth 20% of your overall grade for the module.

Assignment Instructions

Begin by downloading the *assignment_2.zip* file. You should see the following files:

Filename	**Description**
clustering_examples.ipynb	This notebook contains some code examples of different clustering algorithms using sklearn.
assignment_2.ipynb	This notebook contains the code you should run (or edit) to complete your assignment.
salary_data.csv	This is the dataset you should use for your assignment.

Task 1: K-means

→ *See Task 1 in the assignment_2 notebook*

Your first task is to run k-means on two small datasets (50 points each), one should be randomly generated and the other is provided (*salary_data.csv*).

Use the `run_kmeans` function in the notebook and run it 10 different times on each dataset.

1.1 Report the results of these runs and discuss the clusters that are found. Include screenshots of the outputs to aid your discussion.

1.2 Describe your understanding of these three parameters in the `run_kmeans` function and discuss how they impact the results:

```
kmeans = KMeans(n_clusters=k, init='random', n_init=1)
```

Create a new function that runs k-means on your datasets but this time using k-means++. Refer to the sklearn documentation as needed. Re-run k-means using k-means++ on the salary dataset.

1.3 Report and discuss the clusters that are found. Compare them to the previous output where kmeans++ was not used.

Create your own dataset with 30 data points that are somewhat clearly clustered. Generate an elbow plot for a reasonable range of values of k for this dataset.

1.4 Report the Elbow plot and discuss how you would interpret and use such information.

Task 2: Compare Algorithms

→ See Task 2 in the assignment_2 notebook

Generate a dataset with 3,000 samples (points) using the `sklearn.datasets.make_classification` method. Run it until you find a dataset you think is interesting to cluster. You can also use multiple datasets.

2.1 Run the k-means algorithm on this dataset. Report and discuss your results. Also discuss how you selected a value of k and any other relevant parameters.

2.2 Run the DBSCAN algorithm on this dataset. Report and discuss your results. Discuss your approach to parameter estimation in this case.

3.3 Implement one other clustering algorithm on this dataset. Report and discuss your results.

3.4 Compare and contrast the results of all three algorithms (k-means, DBSCAN, your choice) on the same dataset.

Additional Report Instructions

- Your written report should be appropriately formatted and the writing should be coherent and concise.
- Include your name and student number at the top of the first page.
- Screenshots should be **legible**. It is usually better to copy & paste your code than to use screenshots.
- Discussion should go beyond the descriptive and should be critical in nature.

Comments

This assignment has been designed to evaluate:

- Understanding of clustering analysis.
- Ability to research and learn new algorithms/topics/technologies.
- Clarity of communication.
- Attention to detail and ability to follow instructions.