Machine Learning

Applied Project: Topic 3

Submission

- You may carry the project out in groups of two students.
- Must upload the deliverables to Moodle before the deadline.

Notes:

- **A single submission per group**.
- **Do not use Moodle comments** to deliver any sort of important information (e.g., group members). Include all relevant information in the deliverables.

Deliverables

1. A PDF file with a project report (maximum 15 pages). Must contain the names of group members as well as the name of the used CSV file.
2. The python notebook used to carry out the analysis: The notebook must run, out of the box, on Google Colab.
3. An mp4 video with a presentation as a narrated Power Point. The video length has to be between 5 and 10 minutes.
4. A pdf with the presentation.
5. Optional: An mp4 video telling your learning experience in the module (what you have learned, what was hard/easy, what you wish you had learned etc.). The video can be recorded from a smartphone.

Notes:

- Both students need to appear in the presentation video. Failure to do so can lead to failing the activity.
- If the size of the videos exceeds the maximum size allowed in Moodle, links to services such as One Drive (1 TB available for UPM students), Google Drive o Dropbox are allowed. Make sure that:
  - Accessible links are provided,
  - Videos can be downloaded in an mp4 file,
  - Files do not exceed 200 MB.
- Note that you can compress video easily with VLC.
- See https://goo.gl/SKXWSE on how to turn your presentation into a video (also links to how to record any timing, narrations, and laser pointer gestures).

Phases

The project will be delivered in **two phases**. Regarding the first phase:

- look for three clusters with K-means;
- corresponds to 10% of the project's grade (no minimum grade, and it is therefore not strictly necessary);
- deadline: 09/10/2025 at midnight.

The second phase corresponds to the delivery of the final and complete report and corresponds to 90% of the project's grade.

What to do?

The aim is to try and discover meaningful clusters in the data and describe, that is, interpret them. In order to do this, several actions need to be carried out:

1. Preprocess the data;
2. Consider different clustering algorithms;
3. Assess their output;
4. Choose a single, final clustering and describe it.

In order to carry out the above actions, you need to, at least, do the following:

1. Preprocessing
    1. Consider different preprocessing steps and assess how do they affect the outcome.
    2. Discuss the pros and cons, for this particular data set, of converting categorical variables to numeric ones with OneHotEncoding, OrdinalEncoder or a custom transformation.
2. Hierarchical Clustering
    1. Vary linkage (single, complete, etc.) and then choose one; justify.
    2. Plot and analyze dendrograms, in order to answer:
    - What is the effect of linkage?
    - What seems to be a reasonable clustering?
    - What are the clusters in that clustering?
    - Is there are meaningfull hierarchy of clusters?
3. Partitional Clustering
    1. Apply K-means varying the number of clusters
    2. Choose a number of clusters and justify the choice
    3. Analyze the Silhouette scores for the chosen clustering
4. DBSCAN
    1. Consider different neighbourhood sizes (`eps` parameter)
5. Gaussian Mixture Models
    1. Choose the number of clusters with the AIC or the BIC score.

The data

The task is to cluster individuals based on their demographic and economic attributes to identify groups with similar socioeconomic profiles.

Variables:

- **age** – age in years.
- **education** – highest level of education attained.
- **education_num** – years of education (numeric version of `education`).
- **marital_status** – marital status (e.g., married, single, divorced).
- **relationship** – family role within the household (e.g., husband, not-in-family).
- **gender** – gender.
- **capital_gain** – income from capital gains.
- **hours_per_week** – hours worked per week.

There are two csv files uploaded, data-even.csv and data-odd.csv, both with the same structure (i.e., columns). Each group will use only one csv; in particular, group with odd group numbers will use data-odd.csv whereas groups with even group numbers will use data-even.csv. That is, groups 1, 3, 5, etc. will use data-odd.csv wheras groups 2, 4, 6, etc. will use data-even.csv.

Report structure

The report may be written in either Spanish or English. The structure of the report is as follows:

1. Cover page. Include a cover page with title, authors' names, email, course, and date. **Make sure to list all authors (group members)**.
2. Introduction. Briefly explain the problem, the data, your overall approach and give overview of your results.
3. Method. Indicate your strategy and reasoning for the different clustering approaches, justifying the choices of parameters. You may discuss why certain options might make sense and while others might not.
4. Results. Clustering evaluation as well as the interpretation and analysis of a single clustering.
5. Conclusion. Provide conclusions regarding the project, your approach and your results. Discuss possible additional steps towards improvement.
6. References.

Grading

Achieving a grade of 5 requires:

- Delivering all mandatory content, as specified above;
- Absence of major conceptual or formal mistakes.

- Pass the in-class presentation.

Secondary criteria include:

- The quality and clarity of the explanations, justifications, the writing and the presentation. Balancing clarity with rigor.
- Use of plots and other aids that help understanding the report.
- Extent of originality. Use of own data processing, methods, analysis.
- Applying good sklearn practices, such as the Pipeline.

Important: There will be penalties for not following submission instructions, such as exceeding the page limit, not providing the code, etc. Additional recommendations:

- Not everything that you do needs to appear in the report; keep the report concise and to the point. I can check the notebook to see all necessary details.
- Keep in mind that the intended audience is the course instructor. Thus, for example, you should not spend paragraphs explaining what clustering is.
- Avoid plain copy-paste of code and output in the report; present the results in a clear and concise way. In particular, **do not deliver Jupyter notebooks rendered as pdf**.
- Watch out with figures: for example, ensure the axis' scales are discernible.
- I encourage discussion of the achieved results, the motivation for trying some options and not others, and the reasoning behind the choices made, etc.

In-class presentation

There will be five minutes per group, mainly consisting of questions by the instructor and discussion. Failing to show up results in failing the activity.