Freie Universität Berlin
Institut für Informatik
AG Künstliche Intelligenz und Maschinelles Lernen
Prof. Dr. Eirini Ntoutsi
TA: Manuel Heurich

9th December 2021

Machine Learning for Data Science - Project 2

Winter Term 2021/2022

1 Goal

The goal is to perform cluster analysis on the dataset used for the first project, the Adult dataset. In particular, you have to choose at least two clustering algorithms among the ones covered in the lecture and report on its performance as well as on the resulting clusters. Experimental setup, parameter selection, quality measures and conclusions have to be described in the report.

2 Dataset

For the analysis, we will use the Adult dataset[1], which contains census data from the US. The class feature, i.e., information on whether a person's income exceeds 50K dollars per year or not, should not be used for clustering. However, you can use this feature for the labeling of the clusters and for the evaluation of the clustering results.

3 Tasks

3.1 Feature selection - Dataset preparation

Decide on which features to use for clustering. Note that the dataset consists of features of different types. Justify your design choices. What does your final dataset for the clustering task look like? Please list and describe the selected features.

3.2 Clustering

- **Choice of algorithms**
  Choose at least two clustering algorithms from the ones covered in the lecture.

[1] https://archive.ics.uci.edu/ml/datasets/adult

- **Parameter Tuning**
  How are the parameters for your algorithms set? Justify your decisions.

- **Evaluation measures**
  Evaluate the clustering quality using both internal and external measures (if applicable).

- **Model interpretability/visualization**
  Describe the resulting clusters in a human comprehensible way (labeling).

3.3 Protected Attributes

The dataset contains protected attributes, e.g. gender. Decision-making based on these attributes can result in unethical outcomes at the expense of certain subgroups, e.g. women. Gaining insight regarding those protected attributes and their importance for potential decision-making is the first step to avoiding models exploiting the protected features we do not want any decision based on, neither human nor computation based.

Examine your results from the previous tasks with regard to the distribution of protected attributes within the resulting clusters. Report on your observations regarding similarities and differences between the results for both, different algorithms and different parameter settings.

4. Project Logistics

- Programming language: Python

- Format: Jupyter notebooks (+ supporting python files if needed)
  - Comment on your code that we understand what you did
  - Explain your main results (and plots) in a project report (.pdf file)

- Working in groups: We recommend you work in groups of 3

  Note: All group members will receive the same grade.

- If you reuse code, make sure to name your sources and be able to explain the code (if asked).

5. Submission

- Deadline: 20/01/22; 23:55 (Berlin time)

- Submission through Whiteboard in the matching Assignment entry (‘Project 2’)

- Deliverables:
  i) Project report as .pdf
  ii) Code as .zip or .tar.gz
  iii) Readme file guiding us to reproduce your results using your implementation

- The names of all group members on top of your Project Report and in the submission text field in Whiteboard