DATA MINING – Project Guidelines

Salvatore Citraro
salvatore.citraro@phd.unipi.it

a.a. 2021/2022

UNIVERSITÀ DI PISA

General Information

Who?

- Groups of min 3, max 4 people (ideally, heterogeneously distributed… e.g., 2 DS + 1 InfoUma);
- Insert your group at the following link: https://docs.google.com/spreadsheets/d/1SuU8YLHKQcGvg4itG7xkpYKpyTJ77_bHQIVtsRN4_Hg/edit#gid=0

What? – Input + Exercises + …

- A Project:
  - Data Understanding & Preparation;
  - Clustering;
  - Classification;
  - Pattern Mining.

What? – … + Output

- A Report:
  - max 20 pages of text, including tables and figures;

only the report will be evaluated!

General Information

When?

- **[TBD]** 1 mid-draft (i.e., Data Understanding & Preparation + at least 2 Clustering algorithms) + complete project:
  - You will have a *Midtearm Deadline* (mid-draft) + a *Final Deadline* (complete project) ;
  - Deadlines will be provided (and maybe extended) on the main page of the course;
  - Delivering the mid-draft allows you to get 1 bonus point;
  - You will be provided with a new dataset, if you don't deliver the complete project by the final deadline;

Where?

- send the report to datamining.unipi@gmail.com & salvatore.citraro@phd.unipi.it
  - Object: [DM 1 21/22] Project
  - Report title: Project_Surname1_Surname2…

How?

- Exercises: Python, KNIME or a combination of them;
- Report: write in LaTex (Overleaf) (suggested);

The Project

- You can choose between two datasets to analyze:
  - **Glasgow Norms** ☆☆☆☆☆
    - *the mental shape of words (psycholinguistic information); suggested classification task: polysemy;*
  - **Seismic Bumps** ☆☆☆☆☆
    - *identifying hazardous seismic bumps in a coal mine;*

You can choose ONLY ONE dataset

Glasgow Norms: suggested for groups with at least 1 InfoUma;
Seismic Bumps: suggested for groups with at least 1 DS (or 1 Phys);

Links to the file datasets will be provided on the main page of the course!

The Report

Structure

- Title page and possible index not counted in the 20 pages limit;
- Only PDF are allowed, no python code, no KNIME workflows;
- It is better to use font size higher than 9pt;
- Multiple columns are allowed;

Content

- You must justify every choice (from the variables management to the parameters you tune);
- Discuss every result; even if some of them don’t convince you, be fair and try to discuss the possible limitations (they can be imputed to the dataset, to an algorithm that does not fit with the dataset, etc…);
- Plots and tables without any comment are useless;
- Nice and readable plots make your analysis more understandable ;
- Even if you find a top configuration for your algorithm (e.g., k-means, k=5) you must list which are the different parameters you tested and justify your choice;

The Exercises

- Data Understanding & Preparation;
- Clustering;
- Classification;
- Pattern Mining;

The next slides provide several analytical suggestions, but:

- You are allowed to organize the content of the complete project as you prefer;
- You are allowed to identify the classification task as you prefer;
- You are allowed to explore tools and methodologies not introduced during the lectures (e.g., feature selection methods, new plots, algorithms), but it is suggested to write me an email before;

Data Understanding & Preparation (30 pts)

- Data Semantics (3 pts)
  - Introduce the variables with their meaning and characteristics;
- Distribution of the variables and statistics (7 pts)
  - Explore (single, pairs of…) variables quantitatively (e.g., statistics, distributions);
- Assessing data quality (7 pts)
  - Are present errors, outliers, missing values, semantic inconsistencies, etc?
- Variable transformations (6 pts)
  - Is it better to use for further modules transformed variables (e.g., log-transformated)?
- Pairwise correlations and eventual elimination of variables (7 pts)
  - Matrix correlation (analyze high correlated variables);

Clustering (30 pts)

- Clustering analysis by K-Means (8 pts)
  - Choice the attributes, identify the best value of k, characterize the clusters (w.r.t. centroid analysis and variable distribution within);
- Analysis by density-based clustering (8 pts)
  - Choice the attributes, identify the best parameter configuration, characterize clusters;
- Analysis by hierarchical clustering (8 pts)
  - Choice the attributes, the distance function, analyze several dendrograms;
- Final discussion (6 pts)
  - Which is the *best* algorithm? Remember that *best* is studied w.r.t. several aggregate statistics, cluster distributions and w.r.t. the typology of algorithm used for that particular dataset;

Classification (30 pts)

- Classification by Decision Trees (18 pts)
  - Choice the attributes, identify the best parameter configuration(s), eg. gain criterion, then visualize, interpret the tree(s) - 9 pts;
  - Evaluate the performances of the algorithm(s) w.r.t. confusion matrix, accuracy, precision, recall, F1, ROC curve - 9 pts.
- Classification by Other Algorithms or Baselines (6 pts)
  - Try to use another classification algorithm (KNN or Random Forest, suggested) or use a baseline model for the comparison;
- Final discussion (6 pts)
  - Which is the *best* algorithm? *Best* can be studied w.r.t. the performance evaluation or other preferred point of view;

Pattern Mining (30 pts)

- Frequent Pattern extraction (6 pts)
  - Using different values of support, etc;
- Discuss Frequent Pattern (7 pts)
  - Including qualitative and quantitative analysis, e.g., how the number of patterns w.r.t k min_sup changes;
- Association Rules extraction (6 pts)
  - Using different values of confidence, etc;
- Discuss Association rules (7 pts)
  - Including qualitative and quantitative analysis, e.g., how the number of rules w.r.t k min_conf changes, histograms of rules' confidence and lift;
- Exploit the most useful extracted rules (4 pts)
  - E.g., use them to replace missing values or to predict the target variable;

Bonus & Other

- You can get 3 additional extra points in the final mark w.r.t. the following criteria:
  - Innovation (0.5 pts)
  - Experimentation (0.5 pts)
  - Performance (0.5 pts)
  - Appearance, Summary, Organization (0.5 pts)
  - Mid-draft within time (1 point)

- **Project Mark**: average of the previous modules + 3 bonus points;