Applied computational intelligence

Homework 1

The goal is to get a good insight into a dataset by mean of summary statistics and visualisations. For this exercise set choose one alternative below. Regardless of your choice, your submission must comply with the guidelines at the end of this document.

Data selection

You are give the possibility to choose one set of data attached to the HW assignment:

- Alternative 1 - Abalone: The dataset contains the characteristics of abalones. The data can be either i) retrieved from UC Irvine Machine Learning Repository, or ii) within R using the commands: `library(AppliedPredictiveModeling); data(abalone)`.
- Alternative 2 - Air Quality Index: The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. Source: UCI Machine learning repository.
- Alternative 3 - Wine quality: The dataset contains two datasets related to red and white wine samples from Portugal. The data can be retrieved from UCI Machine learning repository.
- Alternative 4 - Heart Rate Prediction to Monitor Stress Level: The dataset contains attributes taken from signals measured using ECG recorded for different individuals having different heart rates at the time the measurement was taken. Source: Kaggle.
- Alternative 5 - Environment data in Melbourne: Energy consumption, climate, and wastewater characteristics of Melbourne wastewater treatment plant for period of six years (2014-2019). The data can be retrieved from Mendeley data.
- Alternative 6 - Concrete mixture data: Data on experiments designed to find concrete formulations that maximize compressive strength. The data used here consists of separate experiments from 17 sources with common experimental factors were combined into one "meta-experiment". The data can be retrieved from UCI Machine learning repository
- Alternative 7 - Your choice: You have a set of data of your own interest. The dataset should comprise of a certain number of observations, each observation consists of a certain number of predictors and corresponding class label.

Data analysis

Regardless of your choice, you must:

1 Describe your dataset and its features, identifying the number of observations N, number of predictor variables D, number of classes L and class-distribution (that is, the number of observations for each of the classes).

2 Perform an unconditional mono-variate analysis of each of the D predictors. Specifically, you must plot their (unconditional) histograms and box-plots, calculate their (unconditional) mean μ_d, standard deviation σ_d and skewness γ_d, with $d = 1, \ldots, D$, using all the N observations.

3 Perform a class-conditional mono-variate analysis of each of the predictors. Again, you must plot their (class-conditional) histograms and box-plots, calculate their (class-conditional) mean $\mu_{d|l}$, standard deviation $\sigma_{d|l}$ and skewness $\gamma_{d|l}$, with $d = 1, \ldots, D$, now using only the N_l observations of class l, for each the L classes.

Item 2 leads to D histograms, D means, D standard deviations and D skewness values. Item 3 leads to $D \times L$ histograms, $D \times L$ means, $D \times L$ standard deviations and $D \times L$ skewness values. Tabulate all means, standard deviations and values of skewness, for both items. Comment on the results, highlight any remarkable fact that emerge from this exploratory analysis. Are there predictors that seem to show any discriminative power (as in, 'are they, alone, capable to separate the classes')?

Then, you must

4 Perform an unconditional bi-variate analysis of the predictors. Specifically, you must plot the scatter plots between all pairs of predictors. For each point (observation), use colours or symbols to indicate the associated class label. Investigate the existence of potential relationships between pairs of predictors and the presence of potential outliers.

Are there any relevant relationships between pairs of predictors? If yes, are these relationships linear? Quantify linear dependence between predictors using pair-wise correlation coefficients ρ_{d_i,d_j}, with $d_i, d_j = 1, \ldots, D$. Either tabulate the correlation coefficients as a correlation matrix $\boldsymbol{\rho}$ with $\boldsymbol{\rho}(i,j) = \rho_{d_i,d_j}$, or show the matrix as an image. Comment on the results.

As final task, you must

5 Perform an unconditional multi-variate analysis of the predictors. Specifically, you must implement the principal components analysis (PCA) yourself without using pre-made PCA functions or libraries. For visualisation purposes, retain only the first two principal components (those associated with the two largest eigenvalues) and plot the scatter plot of the projected observations. Again, for each projected point (observation) you must use colours or symbols to indicate the associated class label. [Remember to perform the necessary pre-processing of the data].

Are the classes well (or better) separated? Are the boundaries between classes linear? What classes show a high degree of overlap and thus are harder to separate?

Guidelines

Regardless of your choice of data, you must generate the following:

- Article: You must generate a report in the format of a conference paper following the template from the IEEE conference proceedings available at the Manuscript Templates for Conference Proceedings. The paper should not be longer than 6 pages and must include the following:
  - Title (5pt): Provide a concise, one-sentence summary that captures the main contribution of your paper [Hint: Spend time on it and try some alternatives[1]. As part of the preparation, this will help both you to write a clear abstract and the reader to grasp the content of the work].
  - Abstract (10pt): Here, you introduce the main objective and overview of the work [Hint: Briefly summarise the goal, scope, methods, and key results].
  - Introduction (20pt): Here, you provide some context and background [Hint: Briefly review relevant literature to understand the dataset, and explain how and justify the need for exploration. Discuss application examples and provide the references].
  - Methods (30pt): Here, you briefly describe your data set and the methods used for analysing it [Hint: Report the main characteristics of the data. Include representative visualisations (e.g., histograms, box-plots) from both unconditional and class-conditional analyses. All figures and tables must be referenced and discussed in the main text. Summarise the features and theoretical background of the analytical methods used].
  - Results (35pt): Here, you explain and critically discuss the results of the preprocessing task [Hint: Report and comment the main results of the analysis].
  - References: Here, you provide bibliographic references [Hint: Highlight the main findings, patterns, or issues discovered through your exploratory work].
- Code listing: Submit the code used to perform the analysis. Regardless of your choice programming, your code must be executable/functioning. The code (and the relevant functions, when needed) can be either pasted at the end of the 6-page article (for instance as an appendix) or packaged together with the paper as a zip file. Make sure to include any dependencies or instructions needed to run the code.
- Repository: You must create and share with the professor and course assistants a Git-based repository to host your project (e.g., on GitHub)[2]. The repository must

[1] Avoid the obvious title "Homework 1: Data pre-processing".

[2] Alternatively, you may use another Git-based platform, provided that the repository is accessible without restrictions to the professor and course assistants.

include your final report in PDF format, all source code used for the analysis, and any supporting files (such as data access scripts or configuration files). It must also contain a `README` file that provides a brief description of the project, clear instructions on how to run the code (including any required dependencies), and a clear statement of each co-author contribution. Submissions without an accessible repository will not be evaluated.

The work can be done individually or in group of maximum 4 co-authors. You can chose to write your paper either in English or Portuguese[3].

You are allowed to consult external resources; however, all sources must be properly cited. If you choose to use AI tools while working on your HW, do so responsibly, for example, to clarify concepts or check your reasoning, not to generate complete solutions. Using AI to produce full answers undermines your learning and violates academic integrity policies. You must acknowledge the use of such tools by citing them appropriately. Include both the prompt and the AI-generated output in an appendix of your homework report or in a separate file in your Git repository.

The work must be submitted by October 19, 2025. Extensions to this deadline may be granted only if unanimously requested at least one week prior to the deadline. Please note that late submissions will be penalised as follows: up to 24 hours late incurs a 20% penalty; up to 48 hours late incurs a 40% penalty; and further delays may lead to additional penalties.

- Draft submission deadline: You are advised to submit a draft version of your paper by October 5, 2025. The draft should include a complete outline of your report, preliminary results and figures, and a clear statement of each team member's contribution. This draft is mandatory in order to receive feedback before the final submission. Failure to submit the draft by the deadline may negatively affect your final evaluation.

[3] In LaTeX, specify \usepackage[portuguese]{babel} in the preamble to change the language.