Applied computational intelligence

Homework 1: Data pre-processing

For this exercise set, choose either Alternative 1 or Alternative 2, below. Regardless of your choice, your submission must comply with the guidelines at the end of this document.

Alternative 1

You are given a set of data[1] consisting of $N = 214$ observations of glass samples. For each sample, there are $D = 9$ predictor variables (the refractive index and the percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe) and the corresponding class label. In total there are $L = 7$ classes.

You must

1 Perform an unconditional mono-variate analysis of each of the D predictors. Specifically, you must plot their (unconditional) histogram, calculate their (unconditional) mean μ_d, standard deviation σ_d and skewness γ_d, with $d = 1, \ldots, D$, using all the N observations. [To calculate means, standard deviations and skewness, you can either use native functions or implement appropriate expression yourself]

2 Perform a class-conditional mono-variate analysis of each of the predictors. Again, you must plot their (class-conditional) histogram, 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 eac the L classes.

Item 1 leads to D histograms, D means, D standard deviations and D skewness values. Item 2 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

3 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

[1]The data can be either i) downloaded from the UC Irvine Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/glass+identification`, or ii) retrieved within R using the commands: `library(mlbench); data(Glass)`.

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

4 Perform an unconditional multi-variate analysis of the predictors. Specifically, you must perform a principal components analysis of the predictors, 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?

Alternative 2

Assuming you have at your disposal a set of data of your own interest and this dataset consists of a certain number of observations, each observation consists of a certain number of predictors (make sure the predictors are numerical, not categorical) and corresponding class label, you might prefer to investigate the characteristics of your own data.

In this case, you must first describe your data and their features in terms of 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).

Then, you must perform the analysis as defined in Exercise 1. That is:

You must

1 Perform an unconditional mono-variate analysis of each of the D predictors. Specifically, you must plot their (unconditional) histogram, calculate their (unconditional) mean μ_d, standard deviation σ_d and skewness γ_d, with $d = 1, \ldots, D$, using all the N observations. [To calculate means, standard deviations and skewness, you can either use native functions or implement appropriate expression yourself]

2 Perform a class-conditional mono-variate analysis of each of the predictors. Again, you must plot their (class-conditional) histogram, 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 eac the L classes.

Item 1 leads to D histograms, D means, D standard deviations and D skewness values. Item 2 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

3 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

4 Perform an unconditional multi-variate analysis of the predictors. Specifically, you must perform a principal components analysis of the predictors, 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 (Alternative 1 or 2), you must generate the following:

- Article: You must generate a report in the format of a conference paper following the template adapted from the ESANN (European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning) conferences[2] that is attached to the homework assignment. The paper should not be longer than 6 pages and must include the following:
  - Title: Here, you summarize your paper in one sentence. [Spend time on it and try some alternatives[3]. 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: Here, you introduce the main objective and overview of the work [Provide a short and informative view of the work, its scope and results].
  - Introduction: Here, you provide some context and background [Briefly, explore the literature in order to define how and why data need to be pre-processes. Discuss some examples of application and provide the references.]
  - Methods: Here, you briefly describe your data set and the methods used for analysing it [Report and comment the main characteristics of the data. Plot the most representative histograms for the unconditional and class-conditional analysis (here or in the next session). Each figure or table must be discussed in the text. Describe the features and the theoretical background of the methods you use for the analysis.]
  - Results: Here, you explain and critically discuss the results of the preprocessing task [Report and comment the main results of the analysis.]
  - References: Here, you provide bibliographic references [Report the books and/or articles that you used for studying the methods and perform the analysis. Each reference reported in this section must be cited in the main text].
- Code listing: The code you used to perform the analysis. Regardless of your choice programming, your code must be executable/functioning. The code (and the relevant functions, if 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.

The work can be done individually or in group of maximum 5 co-authors. You can chose to write your paper either in English or Portuguese[4] and base your analysis on the work by Max Kuhn in `https://github.com/topepo`.

The work must be submitted by **September 9, 2018**. Note that **delays will be penalized** (<24h: 20% penalty; <48h: 40% penalty; etc.).

[2] The original template has been slightly modified to wider the page margins and include page numbers.
[3] Avoid the obvious title "Homework 1: Data pre-processing".
[4] In LaTeX, specify \usepackage[portuguese]{babel} in the preamble to change the language.