



Homework I

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 data¹ 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 $d = 1, \dots, D$, 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 $d = 1, \dots, D$, now using only the N_l observations of class l , for each of 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 emerges 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

¹The 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 $d_i, d_j = 1, \dots, D$. Either tabulate the correlation coefficients as a correlation matrix $\boldsymbol{\rho}$ with $\rho(i, j) = \rho_{d_i, d_j}$, 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 $d = 1, \dots, D$, 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 $d = 1, \dots, D$, now using only the N_l observations of class l , for each of 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 $d_i, d_j = 1, \dots, D$. 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 Alternative 2), you must generate a report consisting of the following:

- A description of the steps performed in your analysis, the associated plot/tables of results and your comments.
- 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 in the report (for instance, as an appendix) or packaged together with the report as a zip file.

You can base your analysis on the work by Max Kuhn in <https://github.com/topepo>.

The report must be submitted by **October 01, 2017**. Note that **delays will be penalized** (<24h: 20% penalty; <48h: 40% penalty; etc.).