Applied computational intelligence

Homework 1

For this exercise set choose ONE of the dataset below[1]. Regardless of your choice, your submission must comply with the guidelines at the end of this document.

Data selection

You are given the possibility to choose one of following datasets:

1. Air Quality Index: The dataset contains daily reports of air quality index from EPA at various US Metro areas, as well as geographic data for the collection locations. Source: Kaggle.

2. Drinking water nitrate violations: These datasets contain the drinking water nitrate violations data summarized for catchments and the associated land cover, geology, climatic, nutrient input variables for each catchment. Source: Kaggle.

3. Deforestation: The dataset contains several datasets related to forest and deforestation from the website Our World in Data. Source: TidyTuesday.

4. Food Consumption and CO_2 Emissions: The dataset contains the annual CO_2 emissions per person for 130 nations worldwide published by Food and Agriculture Organization of the United Nations. Source: TidyTuesday.

5. 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.

6. Bone marrow transplant: The dataset describes paediatric patients with several haematological diseases. Source: UCI Machine learning repository.

7. Student performance: The dataset contains student achievement in secondary education of two Portuguese schools. Source: UCI Machine learning repository.

8. Wine quality: The dataset contains two datasets related to red and white wine samples from Portugal. Source: UCI Machine learning repository.

9. Chemical compound solubility: The dataset 1267 observations of chemical compounds with 228 predictor variables. Source: Applied predictive modelling book.

[1] Follow the links in the source to download the data.

10. Condition monitoring of hydraulic systems: The data set addresses the condition assessment of a hydraulic test rig based on multi sensor data. Source: UCI Machine learning repository.

11. Gas Sensor Array Drift: The dataset contains 13910 measurements from 16 chemical sensors exposed to 6 different gases at various concentration levels. Source: UCI Machine learning repository.

12. Optical Interconnection Network: The dataset contains 640 performance measurements from a simulation of 2-Dimensional Multiprocessor Optical Interconnection Network. Source: UCI Machine learning repository.

13. AirBNB in New York: The dataset contains an excerpt of AirBNBs in New York in 2019, including all needed information to find out more about hosts, geographical availability. Source: Kaggle.

14. UK Used Car: The dataset contains 100000 scraped used car listings, cleaned and split into car make. Source: Kaggle.

15. CNN-based stock market: The dataset contains daily features of S&P 500, NASDAQ Composite, Dow Jones Industrial Average, RUSSELL 2000 and NYSE Composite from 2010 to 2017. Source: UCI Machine learning repository.

Data analysis

Regardless of your choice, you must:

1 Define the goal of your goal and identify the output Y of your predictive model. Describe your data and their features in terms of number of observations N, number of predictor variables D. Make sure that predictors are numerical, not categorical.

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

Item 2 leads to D histograms, D means, D standard deviations and D 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?

3 Perform a bi-variate analysis of the predictors. Specifically, you must plot the scatter plots between all pairs of predictors. 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, or show the matrix as an image. Comment on the results.

4 Perform a 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.

Are the data well (or better) separated? What predictors show a high degree of overlap and thus are harder to separate?

Models for regression

Based on the selected predictors and output, you must

0 Split the data in training and test sets. Finalise the pre-processing step by, if needed, handling outliers, missing values or resampling the data.

1 Use the transformed predictors in the training set to learn an ordinary linear regression model and test the model using the test set (remember to apply the same pre-processing you used on the training set). Compare the model performance obtained on the test set with the estimates you would obtain using a resampling scheme as 5- or 10-fold cross validation: use both the $RMSE$ and R^2.

2 Use the transformed predictors in the training set to learn the penalised linear regression model that you believe being more adequate (justify/comment on your choice). Test the model using the test set (remember to apply the same pre-processing you used on the training set). Determine the optimal value of the penalising parameter λ using a 5- or 10-fold cross-validation based on the $RMSE$ (you can only use the training set in this phase, and your search space λ should consist of at least 10 values). Report on process (show the cross-validation profile, both on terms of the $RMSE$ and R^2). Report the accuracy ($RMSE$ and R^2) obtained on the test set.

3 Use the transformed predictors in the training set to learn either a PLS or a PCR regression model (justify/comment on your choice). Test the model using the test set (remember to apply the same pre-processing you used on the training set). Determine the optimal number of components using a 5- or 10-fold cross-validation based on the $RMSE$ (you can only use the training set in this phase). Report on process (show the cross-validation profile, both in terms of the $RMSE$ and R^2). Report the accuracy ($RMSE$ and R^2) obtained on the test set.

Finally compare the performances of the different predictive models. Is the a difference between training and test results? Why is the reason? Which model performs better? Which one would you suggest? Why would that be your modelling choice?

Guidelines

Regardless of your choice of the 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 IEEE webpage. Note that the article is max 6-page long and must include the following:
  - Title: Here, you summarise your paper in one sentence. [Spend time on it and try some alternatives. Avoid the obvious title 'Homework 1'☹. As part of the preparation, this helps 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 understand your chosen dataset and the regression models that can be used for it. Define how and why data need to be pre-processes. Discuss some possible application examples (if any) and provide the references.]
  - Methods: Here, you briefly describe your dataset and the methods used for analysing it. Provide a brief description of the methods, their pros and cons. [Also report and comment the main characteristics of the data. Each figure or table must be discussed in the text. Describe the features and the theoretical background of the methods you use for the regression.]
  - Results: Here, you explain and critically discuss the results of the data analysis and preprocessing task [Report and comment the main characteristics of the data. Plot the most representative histograms for the unconditional and class-conditional analysis. 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.] Also, you compare the models in terms of RMSE and R^2. Are there any difference between the models? Is there statistical difference between them? [Report and comment the main results of the analysis.]
  - References: Here, you provide bibliographic references. Use the bibliography style provided within the template [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 article as a zip file.

The work will be evaluated based on: adherence to the exercise instructions, adherence to the article instructions, clear and critical argumentation, formatting and orthography.

The work can be done individually or with max 4 co-authors. You can chose to write your article either in English or Portuguese[2]. You can base your work on the resources you might find on the web but you must adequately reference to them.

The work must be submitted by NOVEMBER 10, 2021. Extension on this deadline might be considered if unanimously requested at least 1 week prior the set date. Further note that delays will be penalised (<24h: 20% penalty; <48h: 40% penalty; etc.).

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