Applied computational intelligence

Homework II: Models for regression

This exercise set applies linear regression methods to a set of data containing a number of predictors D and a single outcome Y for a number of observations N.

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[1]

You are given a set of data[2] consisting of $N = 1267$ observations of chemical compounds. For each observation, there are $D = 228$ predictor variables (208 binary "fingerprints" FP that indicate the presence or absence of a particular chemical substructure, 16 count descriptors that indicate the number of bonds or the number of bromine atoms, and 4 continuous descriptors that indicate molecular weights and surface area). The outcome of the regression model is the solubility of the compound.

The predictor observations are split between training and test sets and contained in the given set of data as `solTrainX` (N_{tr}=951) and `solTestX` (N_{ts}=316). Analogously, the solubility values for each compound are contained in `solTrainY` and `solTestY`.

You must

0 Perform an exploratory analysis of the data and pre-process the predictors to remove potential skewness in their distribution. Based on the transformed predictors, comment on the presence of relationships between pairs of predictors (estimate the correlation matrix). Also, are the relationships between the predictors and outcome individually linear (estimate all predictor-outcome correlations)?

Then, you must

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.

[1] Mainly for undergrad students, post-grad students are strongly advised to use their own dataset.

[2] The data can be i) found enclosed to the homework assignment, or ii) retrieved within R using the commands: `library(AppliedPredictiveModeling); data(solubility)`.

2 Use the transformed predictors in the training set to learn a L_2-penalised linear regression model and 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 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 and 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.

4 Bonus Task: Use the transformed predictors in the training set to build a neural network model for regression and test the model using the test set (remember to apply the same pre-processing you used on the training set). Report the accuracy ($RMSE$ and R^2) obtained on the test set. Do the nonlinear model outperform the linear model you previously developed? If so, what might this tell you about the underlying relationship between the predictors and the response?

Alternative 2

You have at your disposal the sets of data given for HW1 (or of your own interest). The data consists of a certain number of observations, each observation consists of a certain number of predictors and an outcome that you wish to predict, 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 and outcome. Then, you must split the set of data into training and test set and perform the steps defined in Alternative 1.

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 IEEE conference proceedings[3]. The paper should be not be longer than 6 pages and must include the following:
  - Title: Here, you summarise your paper in one sentence. [Spend time on it and try some alternatives[4]. 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 regression and the models that can be used for it. Discuss some examples of application and provide the references.]
  - Methods: Here, you briefly describe your data set and the methods you use for regression. Provide a brief description of the methods, theirs 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 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 [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 programming choice, 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), packaged together with the paper as a zip file or provided the link of the notebook environment you used.

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[5] You can base your analysis on the resources you might find on the web but you must adequately reference to them.

The work must be submitted by October 9, 2022. 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.).

[3] Also available at: https://www.ieee.org/conferences/publishing/templates.html.

[4] Avoid the obvious title "Homework 2: Regression models".

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