Applied computational intelligence

Homework II

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

You are given a set of data[1] 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` ($\mathrm{N_{tr}}$=951) and `solTestX` ($\mathrm{N_{ts}}$=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.

2 Use the transformed predictors in the training set to learn a L_2-penalized 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

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

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.

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 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 Alternative 2), you must generate a report consisting of the following:

- A description of the steps performed in each task, the associated plot/tables of results and your comments.
- The code you used to perform for each task. 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.

Post-graduate students are encouraged to choose Alternative 2. The data can be, for instance, part of your own research work or somehow associated with it.

You can base your work on the book by M. Kuhn and K. Johnson, Applied predictive modeling, Springer (2014).

The report must be submitted by **November 06, 2017**. Note that delays will be penalized (<24h: 20% penalty; <48h: 40% penalty; etc.).