Case 1

02582 Computational Data Analysis

February 2024

Case 1

The data for this exercise consist of 100 observations (y, x), of response Y (vector), features X (100-dimensional feature matrix). Further, we have 1000 additional observations, here denoted x_{new}. Data are presented in the text file case1Data.txt and case1Data_Xnew.txt which are found on the course page on DTU learnunder Assignments. You can use any programming language you prefer e.g. R, Pyton or matlab. You can choose the methods you find suited to solve the case, please argue for your choices in the report. You should work in groups of no more than two people. In short your task is to build a predictive model of Y based on X. Argue your choices and assess the quality of the chosen model. Apart from your predictions, $\hat{y}_{new}$, you should also estimate your prediction error. To complete this case you have to hand in three documents.

- A report on the case (max 5 pages, all included).
- Your predictions $\hat{y}_{new}$ (in a file called predictions _YourStudentNos.txt - please insert your student numbers as a replacement for YourStudentNos)
- Your estimated prediction error RMSE (in a file called estimatedRMSE _YourStudentNos.txt).

The requirements for the documents are described in greater detail in the following sections.

The report

Your report should be short (max 5 pages), in pdf format. You can choose to use the provided latex template for your report. The report should answer to the following items:

- Describe your model and method (including model selection and validation).
- Argue for your choices of model, model selection and validation.
- Describe how you handled missing data.

- Describe how you handled factors in the features (catergorical variables).
- Estimate the predictive performance of your model on x_{new}. We are interested in the root mean squared error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. As you do not know the true values y_{new}, you cannot just calculate the error, you need to estimate it. Your estimate will be denoted $\hat{RMSE}$. Describe what you did.

The predictions and estimated prediction error

Your predictions $\hat{y}_{new}$ and your estimated prediction error $\hat{RMSE}$ should be uploaded to DTU inside in two text files. $\hat{y}_{new}$ in a file named predictions _YourStudentNos.txt and $\hat{RMSE}$ in estimatedRMSE _YourStudentNos.txt. The formats are illustrated in predictions _YourStudentNo.txt and estimatedRMSE _YourStudentNo.txt. Please do not include headers in the file. Your predictions $\hat{y}_{new}$ and $\hat{RMSE}$ will be evaluated by the teachers.

The competition

There is no case study without a great competition - actually we have two. There will be a prize for the group who submits the best predictions $\hat{y}_{new}$ in terms of their $RMSE$ (calculated by the teacher). The other prize goes to the group who gives the closest estimate $\hat{RMSE}$ to their actual $RMSE$ (measured in percent deviation and again calculated by the teacher). The winner will be announced at the lectures.