02582 Computational Data Analysis, spring 2019

Hand in on 26th of March 2019 at 23.59 at the latest (DTU inside)

Case 1

The data for this exercise consist of 100 observations (y, x), of *response Y* (univariate), *features X* (100-dimentional). Further, we have 1000 additional observations of the features, here denoted x_{new}. Data are presented in the text file *case1Data.csv* which is found on the course page on DTU inside under *Assignments*.

You can use any programming language you prefer e.g. `R`, `Pyton` or `matlab`. This exercise is open and you can approach the task as you find suited. You do not have to restrict yourselves to methods presented in the course. 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 for your choices and assess the quality of the chosen model. Apart from your predictions, $\hat{y}_{new}$, you should also give an estimate of your expected prediction error.

To complete this case you have to hand in four documents.

1. A report on the case.
2. Your predictions $\hat{y}_{new}$
3. Your estimated prediction error $\widehat{RMSE}$.
4. Your answers to some additional evaluation questions.

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

The report

Your report should be short (1-3 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.
- How did you handle missing data?
- How did you handle factors in the features?

- Argue for your choice of model. How did you do model selection and validation?
- Estimate the predictive performance of your model. 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, you cannot just calculate the error, you need to estimate it. Your estimate will be denoted $\widehat{RMSE}$. Describe what you did.

The predictions and estimated prediction error

Your predictions $\hat{y}_{new}$ and your estimated prediction error $\widehat{RMSE}$ should be uploaded to DTU inside in two text files. $\hat{y}_{new}$ in a file named **predictions.csv** and $\widehat{RMSE}$ in **estimatedRMSE.csv**. The formats are illustrated in *predictionsTEMPLATE.csv* and *estimatedRMSEtemplate.csv*. Please do not include headers in the file.

Your predictions $\hat{y}_{new}$ and $\widehat{RMSE}$ will be evaluated by the teachers.

Additional questions

After completing the case, you should answer some evaluation questions and fill them into the provided *questionnaire.txt*, please rename the questionnaire by appending your study number to the filename ie *questionnaireSxxxxxx.txt*. Your answers will not impact the evaluation of your performance on the case. It is important that you follow the template and write your answer as illustrated in red (in this example):

```
Q: What is your favourite color?
A: Petrol blue

Q: What do you think an average baby elephant weighs at birth
(in kilos)?
A: 71
```

The competition

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