Uppsala University

Introduction to Machine Learning, Big Data, and AI

Assignment 2

General information

- The recommended tool in this course is R (with the IDE R-Studio). You can download R **here** and R-Studio **here**.
- Report all results in a single, *.pdf -file. *Other formats, such as Word, Rmd, or similar, will automatically be failed.*
- The report should be submitted to the Studium.
- A report that do not contain the general information (see template) will be automatically rejected.
- When working with R, we recommend writing the reports using R markdown and the provided **R markdown template**. The template includes the formatting instructions and how to include code and figures.
- If you have a problem with creating a PDF file directly from R markdown, start by creating an HTML file, and then just print the HTML to a PDF.
- Instead of R markdown, you can use other software to make the pdf report, but the same instructions for formatting should be used. These instructions are also available in **the PDF produced from the R markdown template**.
- The course has its own R package `uuml` with data and functionality to simplify coding. To install the package just run the following:

  1. `install.packages("remotes")`
  2. `remotes::install_github("MansMeg/IntroML", subdir = "rpackage")`

- We collect common questions regarding installation, and technical problems in a course Frequently Asked Questions (FAQ). This can be found **here**.
- Deadline for all assignments is **Sunday at 23.59**. See the course page for dates.
- If you have any suggestions or improvements to the course material, please post in the course chat feedback channel, create an issue, or submit a pull request to the public repository!

1 Regularized Regression

The datasets `prob2_train` and `prob2_test` contains simulated data with 240 explanatory variables (`V1-V240`) and 1 numerical response variable (`y`). As per the dataset names, the first dataset contains training data and the second contains test data for this problem. To access the data, just run:

```
library(uuml)
data("prob2_train")
data("prob2_test")
dim(prob2_train)


## [1] 200 241
```

You should do the following and present the results in your report:

1. Fit a linear model to the training data. What are the results? Why does this happen?
2. Use `cv.glmnet` from the `glmnet` package to fit a linear lasso regression to the training data. Describe what the function does. Include the plot showing the MSE for different values of λ in your report and describe how to interpret it.
3. Have a look at the coefficients you get from the λ_{min} and λ_{1se} models. Describe the resulting models in the report, e.g., if any variables have been removed from the model. But please do not print all 240 coefficients!
4. Use `cv.glmnet` from the `glmnet` package to fit a linear ridge regression to the training data.
5. Use the models (lasso/ridge, λ_{min} or λ_{1se}) to make predictions for the test data. Present the MAE and RMSE of the four models in a table in your report. Discuss the results, comparing the interpretability and predictive performance of the models.

2 A simple Spam filter

The purpose of this task is to design an email filter to classify email messages as either *spam* or *not spam* (i.e., ham). The data consists of variables derived from 4601 emails, available in the dataset `Email`. There are 58 columns in the file. The first 57 columns are predictors, while the last column is a binary response variable with one (1) indicating *spam* and zero (0) *not spam*. The 57 predictors can be divided into three types. Predictors 1-48 are the proportion of certain words. For example, the first column is `make`, which shows the proportion of the word "make" in the email. Predictors 49-54 are the proportion of characters. For example, the 49th column is `semi.colon`, which shows the proportion of the character ";" in the email. Predictors 55-57 are concerned with capital letters in the email. The 55th column is `average`, which shows the average length of uninterrupted sequences of capital letters. The 56th column is `longest`, which shows the length of the longest uninterrupted sequences of capital letters. The 57th column is `sum`, which shows the total number of capital letters in the email.

To access the data, use:

```
library(uuml)
data("Email")
```

1. First, split the data into a training set (90 % of the data) and a test set (10 % of the data) by random assignment. Set the test data aside so it will be used only for model assessment.
2. You should fit a logistic regularized regression model to the training data and compare the performance of the models using the test data. Handcraft features you might think would be good, based on the training data. Use the training set to produce a good predictive model as possible.
3. Identify the observations that are misclassified in the training set. Can you find a way to improve the model by handling the misclassified observations?
4. Evaluate your final trained model on the test data *once*. In your report, you should present the following:
   (a) Confusion matrices for train and test data.
   (b) Estimate the accuracy, precision, recall, and F_1 for your classifier.
   (c) Is it important that the spam filter is interpretable, or is it OK to use a black box-model in this case?