

Analysis of High Dimensional Data

Homework 2

March 2020

For this homework you must use the dataset in the file *DataSet.RData*.

```
load("DataSet.RData")  
dim(DataSet)
```

```
## [1] 775 997
```

The outcome Y is binary, and there are 996 potential predictors.

1. Read the next tasks, and think about how you want to use the 775 observations in the dataset. You may want to split the dataset into parts. You do not necessarily have to proceed as in the examples of the course notes.
2. Build a logistic regression model with the lasso. Use the cross-validated area under the curve (AUC) for the selection of an appropriate value for the penalty parameter. (you need the function *cv.glmnet* for this task)
3. With the selected penalty parameter, fit again the logistic regression model with the lasso (using the function *glmnet*), and report the fitted model (selected predictors with parameter estimates).
4. Construct the ROC curve. Select an appropriate threshold c that gives a good compromise between sensitivity and specificity.
5. Give an estimate of the sensitivity and specificity for your final selected model. Here are the definitions of the sensitivity and specificity that need to be estimated:

$$\text{sensitivity} = P_{X^*} \{ \hat{\pi}(\mathbf{X}^*) > c \mid Y^* = 1, \mathcal{T} \}$$

$$\text{specificity} = P_{X^*} \{ \hat{\pi}(\mathbf{X}^*) \leq c \mid Y^* = 0, \mathcal{T} \}$$

You should write a report of max 2 pages (excl. R code, R output and graphs), and it should contain

- a short description of the procedure you followed for building the prediction model (no need to repeat the theory from the course notes)
- R code for building and evaluating the prediction model
- a list with the selected predictors and the corresponding parameter estimates of the logistic regression model, and estimates of sensitivity and specificity of your final selected model

A final note: this homework is intended to be a small assignment. There is no need to do more than what is asked. Of course there are a few decisions you have to make along the way. It is more important that you can motivate your decisions, rather than the decision itself. There is not a single best solution (as is often the case in data science applications); your arguments and motivation are thus important. I estimate that this assignment will take you at most 3 hours to complete.

Later, I will evaluate your final selected model based on a very large validation dataset, and I will compute unbiased estimates of the sensitivity and specificity of your final model.

I recommend to prepare your report in R markdown.

You must submit your report both as a pdf file and as a Rmd file (if you use markdown) or R file. Please use the following format:

- HW2_Name.pdf
- HW2_Name.Rmd (or HW2_Name.R)

The deadline for submission is April 6.