TELECOM ParisTech **SD-TSIA204**
2018/2019 **Pavlo Mozharovskyi, François Portier**

TP 2 : Linear regression

For this lab, you have to upload a **single** `ipynb` file. Please use the following script to format your filename (bad name will lead to a 1 point penalty):

```
# Change here using YOUR own first and last names
fn1 = "pavlo"
ln1 = "mozharovskyi"
filename = "_".join(map(lambda s: s.strip().lower(),
                        ["SD-TSIA204_lab2", ln1, fn1])) + ".ipynb"
```

You have to upload it on EOLE (site pédagogique / TP) before Wednesday 18/01/2019, 23h59 in the folder corresponding to your lab group. Out of 20 points, 5 are specifically dedicated to:

- Presentation quality: writing, clarity, no typos, visual efforts for graphs, titles, legend, colorblindness, etc. (2 points).
- Coding quality: indentation, PEP8 Style, readability, adapted comments, brevity (2 points)
- No bug on the grader's machine (1 point)

Note: you can use https://github.com/agramfort/check_notebook to check your notebook is fine, and also use https://github.com/kenkoooo/jupyter-autopep8 to enforce `pep8` style.

Beware: labs submitted late, by email or uploaded in a wrong group folder will be graded 0/20.

Exercice 1. (Data set contaminated by the robot)

We work with the data set `diabetes` accessible in python. The initial data consists of $n = 442$ patients and $p = 10$ covariates. The output variable Y is a score reflecting the disease progressing. For fun, a bad robot has contaminated the data set by adding 200 inappropriate exploratory variables. Since simple noising the data was not sufficient for the robot, he arbitrarily permuted the variables. To complete the picture, the robot has erased any trace of his villainous act and thus we do not know which variables are relevant. The new data set contains $n = 442$ patients and $p = 210$ covariates denoted by X. Are you capable to resolve the enigma created by the playful machine and retrieve the relevalnt variables?

1) Import the data set `data_dm3.csv` accessible by the link https://bitbucket.org/portierf/shared_files/downloads/data_dm3.csv. The last column is the output variable Y. The other columns are the exploratory variables. Provide the number of the exploratory variables and the number of the observations.

2) Are the exploratory variables centered? Normalized? And the output variable? Provide a scatter plot of four randomly chosen exploratory variables and the output variable (a scatter plot or a bi-plot/pairwise-plot plots all possible pairs of variables). Comment the obtained graphs.

3) Train and test sample. Crate two samples: one to learn the model X_{train} and one to test it X_{test}. Put 20% of the data set in the test sample. Provide the size of each of the 2 samples. Note that the new sample of the covariates X_{train} is not normalized. In what follows, please pay attention to include the intercept in the regression models.

4) Provide the covariance matrix for X_{train}. Plot the eigenvalues of the covariance (or correlation) matrix in descending order. Explain why does it make sense to keep only first PCA variables. In what follows, we will keep 60 variables.

5) Following the observations of the question (Q4), apply the method "PCA before OLS" that consists in applying OLS with Y and $X_{\text{train}}V_{(1:60)}$, where $V_{(1:60)}$ contains the eigenvectors (associated with the 60 largest eigenvalues) of the covariance matrix. Run linear regression (with intercept), then plot the values of the coefficients (but not for the intercept). On another graph, do the same using the classical OLS.

6) Provide the intercept values for the 2 regressions from the previous question. Also, provide the mean value of the output variable Y (for the train set). Are the two intercepts equal? Comment. Exceptionally for this question, center and normalize the variables after PCA (the low dimensional ones). Run the regression and verify that the intercept is equal to the average of Y on the train set.

7) For the two methods (OLS and PCA before OLS): Plot the residuals of the prediction for the test sample. Plot their density (one can use a histogram for example). Calculate the determination coefficient for the test sample. Calculate the prediction risk for the test sample.

8) Program the method of the forward variable selection. You can use the test statistics of the test for nullity (as seen during the course). For the moment, do not define the stop criterion for the method, i.e. add a variable at each time until all the variables are selected. Provide the order of the variable selection.

9) Stop criterion: We choose to stop if the p-value is larger than 0.1. Illustrate the method providing (i) the 3 graphs of the test statistics obtained when selecting the 1st, 2nd and 3rd variables (in abscissa: the indices of the variables; in the ordinate: the value of the test statics), (ii) the graphs of the first 50 p-values (each associated to a selected variable). On the same plot, trace the horizontal line with the ordinate 0.1. Finally, provide the list of the selected variables.

10) Run OLS on the selected variables. Provide the prediction risk for the test sample and compare with those for OLS and PCA before OLS.

11) To prepare for the cross-validation, split randomly the train sample in 4 equal parts (called “folds”). Provide the numbers of the observations falling into each fold.

12) Apply the ridge regression method. For the choice of the regularization parameter, run the cross-validation on the “folds” defined in the previous question. Each “fold” is used to calculate the prediction risk while the resting ones are used for estimating the model. Then the 4 risks are averaged. Plot the estimated risk curve as a function of the regularization parameter (pay attention when choosing the range for the values of the regularization parameter). Provide the optimal regularization parameter and the corresponding risk for the test sample.

13) Using the function `lassoCV` of the library `sklearn`, choose the regularization parameter for the LASSO. Provide the corresponding risk.

14) Provide the variables selected by the LASSO. How many are they? Apply the OLS method to the selected variables. This method is called Least-square LASSO.

15) This last question is an outlook for the non-linear approach. Using the variables selected by the LASSO, (Q13) or by the forward selection method (Q9), develop the method of the non-linear regression. Learn the different parameters using the cross-validation and provide the risk for the test sample. It is possible to have worse performance than the OLS method. Comment.