TELECOM ParisTech **MDI720**
2017/2018 **François Portier, Joseph Salmon**

TP : Python, Numpy, Pandas and 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 = "joseph"
ln1 = "salmon"
filename = "_".join(map(lambda s: s.strip().lower(),
                        ["SD204_lab1", ln1, fn1])) + ".ipynb"
```

You have to upload it on EOLE (site pédagogique / TP) before Wednesday 06/12/2017, 23h59 in the folder corresponding to your 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. **(Analysis electricity consumption)** If needed, a tutorial on pandas can be helpful : http://pandas.pydata.org/pandas-docs/stable/tutorials.html Let us use the dataset[1] **Individual household electric power consumption Data Set**.

First, execute the following commands :

```
# download part if needed.
url = u'https://archive.ics.uci.edu/ml/machine-learning-databases/00235/'
filename = 'household_power_consumption'
zipfilename = filename + '.zip'
Location = url + zipfilename

# testing existence of file:
if sys.version_info >= (3, 0):
    if not(path.isfile('zipfilename')):
        urllib.request.urlretrieve(Location, zipfilename)
else:
    if not(path.isfile('zipfilename')):
        urllib.urlretrieve(Location, zipfilename)
# unzip part
zip = zipfile.ZipFile(zipfilename)
zip.extractall()
# Detect and count lines with missing values.
na_values = ['?', '']
fields = ['Date', 'Time', 'Global_active_power', 'Sub_metering_1']

df = pd.read_csv(filename + '.txt', sep=';', nrows=200000,
                 na_values=na_values, usecols=fields)
```

[1] https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption; if this website is too slow use http://josephsalmon.eu/enseignement/TELECOM/MDI720/datasets/household_power_consumption.zip

We only focus on the Global_active_power and Sub_metering_1 features for the moment.

1) Count the number of rows where Global_active_power or Sub_metering_1 are missing (represented by a "nan"). Remove these rows.
2) Read the "Attribute Information" in https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption#. Now scale the variable Sub_metering_1 to have the same unit as Global_active_power.
3) Use to_datetime and set_index to create a Time Series (beware of the international dates format that is different from the French standard) and index your dataframe by timestamps.
4) Display the graphic of daily averages, between January 1 2007 and April 30 2007, with the variables Global_active_power and Sub_metering_1 on a same figure. Propose an explanation for the consumption behavior between February 23 and March 3 ? between April 10 and April 15 ?
   Rem : On top of matplotlib you could use the seaborn package for nicer display.
5) Display a barplot of the Sub_metering_1 by weekdays. Interpret the evolution of consumption throughout the week.

Let us now add some temperature information for our study. Such information can be found at http://josephsalmon.eu/enseignement/TELECOM/MDI720/datasets/TG_STAID011249.txt. Here the temperatures available are the one in the city of Orly (note that in the previous dataset the location were the consumption was recorded in France is unspecified).

6) Load the dataset with pandas, and keep only the DATE and TG columns. Divide by 10 the TG column to get Celsius temperature. Treat missing values as NaNs.
7) Create a pandas Time Series of the daily temperatures between January 1 2007 and April 30 2007. Display on the same graph the temperature and the Global_active_power Time Series.

Exercice 2. (Analysis of the auto-mpg dataset)

Here, we consider the auto-mpg.data. We aim at predicting cars consumption based on several characteristics : cylinders, displacement, horsepower, weight, acceleration, year, country and cars name. The output coding cars consumption (more precisely the “mpg”, *i.e.*, the distance ridden in miles for a gallon of oil) is written $\mathbf{y}$; For the first questions we do not use the qualitative feature origin and car name.

8) Import the dataset from https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data-original with Pandas. Add columns name using the option 'name' de read_csv and consulting : https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.names. You can check the impact of using sep=r"\s+". Is there a marker for missing values in this dataset ? If needed remove the corresponding lines.
9) Encode the three origins ('origin' feature) with meaningful labels such that 1 stands for USA, 2 for Europe and 3 for Japan[2].
10) Get the least-squares estimator $\hat{\boldsymbol{\theta}}$ (with intercept) the prediction vector $\hat{\mathbf{y}}$ considering only the 9 first line of the dataset. What do you observe (in particular for cylinders and model year) ?.
11) Now, get the least-squares estimator $\hat{\boldsymbol{\theta}}$ and the prediction vector $\hat{\mathbf{y}}$ (with intercept) over the whole dataset, after performing scaling/centering (the columns must have unit standard deviation and be zero mean). Which variables seem to best explain gasoline consumption according to your model ?[3]
12) Compute $\|\mathbf{r}\|^2$ (the square norm of the residual vector). Check numerically that, using for instance np.isclose :
$$\|\mathbf{y} - \bar{y}_n \mathbf{1}_n\|^2 = \|\mathbf{r}\|^2 + \|\hat{\mathbf{y}} - \bar{y}_n \mathbf{1}_n\|^2.$$
where $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ and $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^d$
13) Assume you observe a new car with the following values features :

cylinders	displacement	horsepower	weight	acceleration	year
6	225	100	3233	15.4	2017

Can you predict its consumption in this model ? Beware of the year encoding. Use a pipeline http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html for performing the rescaling and the least-squares step again.

[2] *cf.* http://lib.stat.cmu.edu/datasets/cars.desc

[3] Note that a more refined answer should rely on t-tests.