

# SPLEX TME 1

## Exploratory Analysis with Significance Tests. Multiple Hypothesis Testing. Handling Missing Data. AdaBoost

### Data

We explore two data sets downloadable from the Machine Learning Repository (<http://archive.ics.uci.edu/ml/index.php>)

- Breast Cancer Wisconsin (Diagnostic) Data Set ([https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)))
- Mice Protein Expression Data Set (<https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression>)

### Exploratory Analysis

We will perform an exploratory analysis with Python. You can use the **Spyder environment** which is already installed on the machines.

You will need to load the following packages:

```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
import statsmodels.sandbox.stats.multicomp as sm
```

1. Load the Breast Cancer data set using `pd.read.table()` (for the Mice Data Set you will need `pd.ExcelFile()`)
2. Some data sets have missing data. You can impute them by replacing the missing values by median values with `fillna(data, inplace=True)` (you will need to impute data in the Mice Data Set)
3. Both the Mice data set and the Breast Cancer are binary classification tasks ( $M$  and  $B$  are two classes in the Breast Cancer, and *Ts65Dn* and *Control* for the Mice)
4. Find the correlation coefficients between variables with `stats.pearsonr()`. Are there a lot of variables which are strongly correlated? What is the meaning of the sign of the correlation coefficient?
5. Run the Wilcoxon test (if you have two classes or Kruskal-Wallis test if you have more than two classes) `stats.wilcoxon()` to find variables which are significant to discriminate two classes.
6. We perform a multiple hypothesis testing (since we have a lot of variables), and we need to adjust the p-values. You can adjust the p-values with `sm.multipletests()`

Consider different adjustment methods:

<http://jpktd.blogspot.fr/2013/04/multiple-testing-p-value-corrections-in.html>

What is the most and the least stringent methods of adjustment? Can you explain why?

7. Compare the distributions of variables in two classes with `stats.ttest_ind()`
8. Boxplot the distributions of the variables of the observations from class 1 and from class 2 using `boxplot()`. Are the plotted distributions coherent with the results obtained by `stats.ttest_ind()`?

9. What are your conclusions? What variables are the most significant? What significance threshold (significance level) should be chosen? What multiple hypothesis adjustment method would you use? Why?

## AdaBoost

1. Test the AdaBoost classifier

```
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
```

2. Test various number of weak learners

3. Run the AdaBoost on a 2-dimensional problem and analyse the class boundary

```
http://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_twoclass.html#sphx-glr-auto-examples-ensemble-plot-adaboost-twoclass-py
```

4. Is AdaBoost a linear or a non-linear classifier?