LAB 3: Sparsity-based learning

- This lab is about feature selection within the framework of sparsity based regularization, using elastic net regularization.
- The aim of the lab is to play with the libraries and to get a practical grasp of what we have discussed in class.
- Follow the instructions below.

Goal:
This lab is divided in two parts depending of their level of complexity (**Beginner**, **Intermediate**). Your goal is to complete entirely, at least, one of the two parts.

Toy problem

We focus on a regression problem where the target function is linear. We will consider synthetic data generated (randomly sampled) according to a given probability distribution and affected by noise. You will have the possibility of controlling size of training and test sets, data dimension and number of relevant features.

NOTE:
In the code we use a different notation from what you have seen in the classes. The functional minimized is:

$$\min_{\beta \in \mathbb{R}^p} \|X\beta - Y\|^2 + \texttt{L1\_par}\|\beta\|_1 + \texttt{L2\_par}\|\beta\|_2^2.$$

In addition, in some MATLAB scripts the sparsity parameter `L1_par` can be also referenced as `tau`, and the smoothing parameter `L2_par` as `smooth_par`.

Download the file `regml2014_lab3.zip`, extract it and add all the sub-folders to the path. This file includes all the code you need!

PART I: Beginner

Overture: Warm up

Run the file `gui_l1l2.m` and the GUI will start. Have a look at the various components.

- Generate a training set with the default parameters.
- Press the run button to start a training phase with the selected `L1_par` and `L2_par` parameters and perform testing.
- Change values for `L1_par` and `L2_par` and have a look at test error and number of selected variables:
  - First set `L2_par` $= 0$ and vary `L1_par` trying to obtain a sparser or denser solution. What do you notice?
  - Repeat the experiment with a `L2_par` > 0. How do test error and number of selected features vary?
- Now select KCV for `L1_par` tuning and observe the KCV error curve.

Interlude: The Geek Part

Back on the MATLAB shell, have a look to the content of directory `./PROXIMAL_TOOLBOXES/L1L2_TOOLBOX`. There you will find, among the others, the code for command `l1l2_algorithm` (used for variable selection), `l1l2_kcv` (used for model selection with kcv or loo), `l1l2_pred` (for prediction on a test set).

For more information about the parameters and the usage of those scripts, type:

```
>> help l1l2_algorithm
>> help l1l2_kcv
>> help l1l2_pred
```

Finally, you may want to have a look at file `l1l2_demo_simple.m` for a complete example of analysis.

Allegro con brio: Analysis

Carry out the following experiments either with the GUI, when it is possible, by personalizing the file `demo_l1l2.m` or by writing appropriate scripts.

i) *Prediction*: Considering elastic net regularization, observe how the training and test error change

- When we change (increase or decrease) the regularization parameter associated with the ℓ_1-norm.

- When we change (increase or decrease) the correlation parameter associated with the ℓ_2-norm.
- The training set size grows (for instance, `[10,100,1000]` as long as MATLAB supports you!).
- The amount of noise on the generated data grows (the test set is generated with the same parameter of the training).

Change one parameter at a time!

ii) *Selection*: Considering elastic net regularization, observe how the number and values of non zero coefficients in the solution change

- When we change (increase or decrease) the regularization parameter associated with the ℓ_1-norm.
- When we change (increase or decrease) the correlation parameter associated with the ℓ_2-norm.
- The training set size grows.
- The amount of noise on the generated data grows.

iii) *Large p and small n*: Perform experiments similar to those above changing p (dimension of points), n (number of training points) and s (number of relevant variables)

- Set $p \ll n$ and $s > n$ (is it possible?)
- Set $p \gg n$ and $s > n$
- Set $p \gg n$ and $s < n$

PART II: Intermediate

Crescendo: Data standardization

Consider the classification dataset given in `part3-data.mat` (use the scripts and NOT the gui):

- Use `l1l2_algorithm` to analyze the feature selected with different values of the regularization parameters.
- Tune the sparsity parameter `tau` to select only one variable. Is there another variable that can provide a better solution? (*hint*: only the first ten column of `X` are correlated with `Y`).

- Can you figure out why the selected variable is not the one that you would expect? (*hint*: analyse the correlation between the columns of `X` and `Y` and the ranges of the columns of `X`, e.g. with `imagesc(X(:,1:10)),colorbar)`.