Assignment 8
Random Forests
Ensemble Classifiers

1 Assignment for those who are achieving projects

- `[X_b,Y_b]=bootstrap([X,Y]);`
  This function samples the data with replacement. More specifically each sample described in `X_b` and `Y_b` is randomly selected from `X` and `Y`, regardless of what are other samples in `X_b` and `Y_b`.
- `info=PNH_train1(V,displaying_function);`
  This function builds an ensemble classifier and stores the information in a structure called `info`. This structure has a field called `tree` which is an array of `trees`. The size of this array is `V`. `displaying_function` is a function called each time a new tree is being trained.
- `y_hat=PNH_predict1(info,x);`
  This function counts the number of trees predicting each possible class and predicts the class for which there is a maximum number of trees predicting that class.

  $$\widehat{y} = \arg\max_{c \leq C} \sum_{v=1}^{V} h_v(\mathbf{x}) \qquad (1)$$

  where $h_v(\mathbf{x})$ is the label predicted by the tree number v for sample $\mathbf{x}$.
- `[tab_train,tab_test]=PNH_displaying_function(info);`
  This function is called successively in `PNH_train1` each time a new tree is induced. Each time it is called, it tests `info` on the training and on the testing set. The two overall accuracy performances are stored respectively in `tab_train` and in `tab_test`.

1.1 Presenting the results

Plot two overall-accuracy curves as functions of the number of trees used in the ensemble classifier. These two overall accuracies are obtained when tested on the training set and on the testing set. Usually what is observed is that the overall-accuracy curve tested on the testing dataset is not going upward beyond a certain number of trees, meaning that generally no over-fitting is observed.

2 Assignment for those who are reviewing projects

The goal is to build Matlab functions that achieve some basic checks on the data provided along each project. Two files are to be delivered.

The first file is a `.pdf` document. Its name is `reviewer` followed by a number and a `H` indicating that it refers to the second assignment. The first part of this document explains what is tested by each test. The second part explains for each project what has passed and what has failed with precise values showing the problem. The third part is optional, it explains what supplementary information you would request from the projects and how this information could provide more valuable testing.

The second file is a `.m` script having the same name, it runs successively the different functions contained in this file that do the different testings.

3 Discussion

Your task is first of all to read all projects and check `Progress`. You should write a single `.pdf` document, named `discussionH.pdf` discussing how all projects have undergone this first step, the difficulties that have been overcome and those that remain challenging issues. You should then express your opinion as to whether I should come back on some specific issues. You may also add some specific comments to a specific project on *Discussions* [1] and some specific questions on *Questions*. You are also expected to write in *Questions* the answers to all other questions.

[1] Comments should be most respectful as any work needs attention, and regardless of it being possibly wrong, it is going to be useful to get a better understanding. So there can be no shame in being wrong.