Assignment 3

1 Assignment for those who are achieving projects

You are in charge of programming seven Matlab scripts.

- `name_of_file=PNC_BinaryDataset(class);` [1] This function creates a simplified binary dataset whose name is `name_of_file` followed by `.mat`. It contains the following Matlab objets.
  - `X` is a matrix. Each line refers to a pixel of an hyperspectral image. Each column refers to a wavelength or an index such as NDVI. There should be two, three or four features.
  - `Y` is a column-vector. Each component refers to a pixel. Each component is either 0 or 1. 1 indicates that the pixel belongs truly to a chosen class indicated by `class` and 0 indicates that is does not belong to that class.
  - `metadata` is a structure with supplementary information. It contains the following fields.
    - ∗ `size` is the size of the image, it is a row-vector.
    - ∗ `features_s` is a structure containing information on the different wavelengths. Each field of this structure has the name of a set of features, and is equal to a row vector indicating the corresponding column numbers.
    - ∗ `F` is the number of features.
    - ∗ `Void_pixels` is the number of pixels whose class is unknown in the original hyperspectral image.
    - ∗ `classes_l` is a cell containing a list of the names of each of the two classes.
    - ∗ `C` is equal to 2.

The following functions are adapted to the binary context.

- `info=PNC_train1();` [2] Note that we do not need the name of the data file which is `training.mat` created by

  `PNB_createSEt(name_of_file,K);`

  This is a dummy function and `info` is a structure with two fields.
  - `feature_number` is an integer denoted here as f.
  - `threshold` is a number denoted here λ
  - `sign` is either 1 or -1 denoted here as s.

  The values of these functions are randomly selected.

- `y=PNC_predict1(info,x);` [3] This achieves a split, it predicts the class

$$\widehat{x} = \mathbf{1}(sx_f \leq s\lambda) \tag{1}$$

  where x_f is the f component of sample x which is a line of matrix X.

- `score=PNC_score1(info,name_of_file,predict_function);` This function computes the score obtained when using the function stored in `predit_function` and with parameter values contained in `info` when testing all samples contained in the dataset called `name_of_file`. `score` is a structure with the following metrics.
  - recall
  - precision
  - overall accuracy
  - confusion matrix

- `score=PNC_score2(name_of_file,train_function,predict_function,K);` [4] This function computes the following metrics on the dataset called `name_of_file` using the training and predicting functions stored in `train_function` and `predict_function`.
  - recall
  - precision

[1] `N` stands for the project number.
[2] `N` stands for the project number.
[3] `N` stands for the project number.
[4] `N` stands for the project number.

  - overall accuracy
  - confusion matrix

  All of these metrics are averaged over K randomly chosen training and testing sets. These metrics are stored in the following fields of a structure named `score`. The name of the fields are respectively `recall`, `precision`, `overall_accuracy` and `confusion_matrix`. The basic content of this function in section A.

- `[N,X,N1,X1,N0,X0]=PNC_show_hist(name_of_file,feature_number)` computes three histograms on the values of one feature. The first one makes no restriction and considers all samples of the dataset, the second one considers only samples whose true class is 1 and the third one considers only samples whose true class is 0.

- `info=PNC_train2();` [5] This functions uses `PNC_predict1` and `PNC_score1` with `'training.mat'` as `name_of_file` to find the best value for `info` by testing randomly drawn values, a large number of times. Note that `PNC_predict1` can be used for this task. Here *best* means getting the highest value of `overall accuracy`. `info` is a structure with three fields.

  - `feature_number` is an integer denoted here as f.
  - `threshold` is a number denoted here λ
  - `sign` is either 1 or -1 denoted here as s.

The `.pdf` document is named `project_NC.pdf` and contains any relevant information. The following issues are to be described.

1. Why did you choose this or these features, why did you use this vegetation index or why not?
2. After showing on a same graph three histograms computed with `PNC_show_hist`, how did you choose the threshold and the sign in `info`. What are the performances computed with `PNC_score`.
3. Discuss the differences between the predictor found with `train2` and the predictor found using the two histograms.

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 an `C` 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.

As in the previous assignments, the checks could check the consistency of the different informations and it can build a simplified fake new database and check the obtained results with respect to that specific database.

3 Discussion

Your task is first of all to read all projects and check `Progress`. You should write a single `.pdf` document, named `discussionC.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* [6] and some specific questions on *Questions*. You are also expected to write in *Questions* the answers to all other questions.

A Content of `PNC_score2`

```
function score=PNC_score2(name_of_file,train_function,predict_function,K)
% score=PNC_score2('ProjectNA',@PNC_train,@PNC_predict,5);
  score.confusion_matrix=zeros(2);
  score.OA=0;
  score.recall=0;
  score.precision=0;
  for k=1:K
```

[5] `N` stands for the project number.

[6] 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.

```
    PNB_createSet(name_of_file,K);
    info=train_function();
    score1=PNC_score1(info,'testing.mat',predict_function,K);
    score.confusion_matrix=score.confusion_matrix+score1.confusion_matrix;
    score.overall_accuracy=score.overall_accuracy+score1.overall_accuracy;
    score.recall=score.recall+score1.recall;
    score.precision=score.precision+score1.precision;
  end
  score.confusion_matrix=score.confusion_matrix/K;
  score.overall_accuracy=score.overall_accuracy/K;
  score.recall=score.recall/K;
  score.precision=score.precision/K;
end
```