Assignment 7
Building new features to be used by linear classifiers
K Nearest Neighbor Algorithm
Using a validation data set to address the overfitting issue

1 Assignment for those who are achieving projects

1.1 Adding new features using polynoms

- `list_feature_l=PNG_list_features(F,degree);`
  This functions lists all possible ways of considering subsets of features with respect to the two following conditions.

$$1 \leq f_1 \leq f_2 \leq \ldots \leq f_k \leq F \quad \text{and} \quad k \leq \text{degree} \tag{1}$$

  `list_feature_l` is a matrix, each line is a possible subset of features. Each component is an integer indicating a feature.

- `Xp=PNG_poly(X,degree);`
  This function replaces the matrix `X` with a new matrix `Xp` having a larger number of features containing all possible products of features, the only constraint is that the number of features involved in each product should remain below `degree`. This function uses `list_features`.

1.2 Adding new features using a kernel

- Xf=PNG_rbf(X,subset,gamma);
  This function constructs a new dataset where the features are now obtained using the distance of each sample and each sample indicated by `subset`. The new number of features is equal to the number of samples listed in `subset`, so these samples are referred to as $\tilde{\mathbf{x}}_f = (x_{f,f'})_{f'}$. It uses the Radial Basis Function Kernel.

$$x_{n,f} = e^{-\gamma(\sum_{f'}(x_{n,f'}-x_{f,f'})^2)} \tag{2}$$

  where γ stands for `gamma`.

1.3 k Nearest Neighbor

We consider here the Euclidean distance defined as

$$d(\mathbf{x},\mathbf{x}') = \sqrt{\sum_f (x_f - x'_f)^2} \tag{3}$$

- `y_hat=PNG_predict(info,x);`
  `info` is a structure containing an integer denoted `k` and the training dataset `X` and `Y`. This function selects the k samples in the training set that are the closes to `x` in the sense of the Euclidean norm.

$$\forall \mathbf{x}', \quad d(\mathbf{x}_1,\mathbf{x}) \leq d(\mathbf{x}_2,\mathbf{x}) \leq \ldots \leq d(\mathbf{x}_k,\mathbf{x}) \leq d(\mathbf{x}',\mathbf{x}) \tag{4}$$

  The predicted label is set using the majority rule upon the labels of these k samples denoted as $y_{\Phi_1}, y_{\Phi_2}, \ldots, y_{\Phi_k}$.

$$\widehat{y} = \arg\max_{c \leq C} \sum_{k' \leq k} \mathbf{1}(y_{\Phi'_k} = c) \tag{5}$$

1.4 Using a validation dataset

- `function PNG_set_validation(alpha);`
  This function selects randomly $\lfloor \alpha N \rfloor$ samples from the training set. These samples are removed from the training set and moved to the validation set called `validation.mat`, $\alpha \in (0,1)$.

- `function info=PNG_train1(beta);`
  This function trains a multiclass linear classifier with an increased number of features obtained by multiplying features together. `beta` is used to set the size of the validation dataset. `degree` is set using the validation dataset.

- `function info=PNG_train2(beta);`
  This function trains a multiclass linear classifier with a new set of features obtained by randomly selecting a small set of samples and by applying the RBF kernel. γ is set using the validation dataset. `beta` is used to set the size of the validation dataset.

- `function info=PNG_train3(beta);`
  This function selects with the validation dataset `k` so as to be used when predicting with the `kNN` algorithm. `beta` is used to set the size of the validation dataset.

1.5 Presenting the results

Plot the overall accuracy obtained with the three different algorithms when tested on the testing dataset. Consider different values of β.

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 `G` 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 `discussionG.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.