Machine Learning & Neural Networks

Project: "Visual handwritten digits recognition"

Objective:

You have to do a research for a client who wants to know the pros and contras of the different ML techniques explained during the course when they are used for his/her problem of handwritten digits recognition.

Team Roles:

The work is to be done by stable teams of three people, randomly determined out of the students with "standard evaluation". Each team determines the role of each member at the very beginning and stable ever since. Roles are:

1. **Coordinator** is the last responsible of the whole work, who coordinates the work of the other two members, check for the overall quality of the results, including the presentation, also in charge of writing the document.
2. **Scientist** is the person responsible of the scientific soundness of the work, formulas and mathematical work. He/she has to check and verify scientific soundness of the obtained results
3. **Technician** is the person in charge of programing the different techniques to obtain desired results.

In case important problems are detected within a team that prevents from obtaining a good work, it has to be notified to the Professor a.s.a.p. so a solution can be determine on time.

The document hast to include the role of every team member, as well as the degree of work done by each member, in which 100% means an equally distributed amount of work, so that everyone has done his/her work accordingly. This percentage can vary if someone takes over the work of another team member so that there is an unbalanced distribution of work (total sum of the percentage of all three members hast to be 300%)

Note: The Project for those opting for "Evaluación Única" is the same project, but it has to be done fully from scratch, so that not any recycling of previous document or code is allowed at all (no matter if the student has collaborated in any previous project)

Deliverables:

1.- Document is a report in pdf not longer than 15 pages (front page, table of contents and annex with used code are counted separately) than is aimed to be read by the "client" who already knows about Pattern Recognition Techniques, but who wants to know how good they are for solving his/her specific problem. You have to convince him/her that you have done an exhaustive research and you have found the right solution for his/her problem. O course he/she would like to have a clear document and a good summary/conclusion. Your code is only included as an Annex.

2.- Results, evaluated in the recognition of new images of digits. You will be provided by a set of images in the file "Test_numbers.mat", that is going to be available during 24 hours indicated in Mooddle. It contains the data in the same format as the training data used in the work, but only containing the field "image".

You will have to upload the results for the classification into **one zip file** that contains four different files, one for each classification technique, with following names and contents:
File name: ######_technique.mat where:

- ###### are the 6digits of the student ID number (starting with "M" or " " (blank) depending on the case
- technique can be "bay","knn", "mlp" or "som" indicating the type of classifier

Each file has to contain the following variables:

- "name" is a cell with as many elements as people in the team. Each element is a chart string (maximum 8 char) with the name of each student in the team
- "PCA" is an integer that indicates the chosen dimensionality of the data reduced by PCA or Autoencoder (if used).
- "class" is a 1x10000 array where its i-th element is the class determined for the i-th digit in Test_number.mat

You have to be very careful and strict with the name of the file, the name of the variables and the type, size and contents of the variable, since **if they are not correct, the work cannot be evaluated.**

Due date: The mentioned zip file with the results has to be uploaded through Moodle any time during the same 24 hours the data is available in Moodle.

Database to be used: the handwritten digits

These images are originally downloaded from MNIST database: yann.lecun.com/exdb/mnist
There is an international competition using this data at www.haggle.com
The set of image to used in this Homework can be downloaded form the Moodle webpage and they are in the file Trainnumbers.mat, that contains the following data:

Trainnumbers.image is a 784x10000 matrix, where each column represents a 28x28 image of a handwritten digit.

Trainnumbers.label is a 10000 array where its i-th element is the number associated to the i-th image

Tasks breakdown:

Task 1.- Viability of PCA for reducing the dimensionality and further reconstruction. Quantify the reconstruction MSE versus the number of dimensionality of the reduced data and display the reconstructed images.

Viability of the following techniques for classification using previous dimensionality reduction with PCA. Obtain the confusion matrix. Compare the results with the other techniques

Task 2.- Using k-NN
Task 3.- Using Bayesian Classifiers.
Task 4.- Using MLP.
Task 5.- Using SOM.

Optional tasks for obtaining top marks above 7:

Task 6.- Use Linear Fischer Discriminant for reducing the dimensionality
Task 7.- Use K-means for clustering the images
Task 8.- Use Autoencoder for dimensionality reduction (first reduce with PCA to an intermediate reduction level). Compare the results with task 1.
Task 9.- Build a visual map using SOM where the pattern image is at the position of every neuron in the 2D map.
Task 10.- Please highlight any original improvement or combination of the above mentioned techniques that you have tried in order to improve the results

Factors to be considered

Here are some hints of key questions that you may want to research about and to give a good answer to.

General questions:

- How confident are you on the results based on the data used for training and for testing?
- What are the relative advantages of some algorithms vs. the others?
- How fast are your proposed algorithms?

Specific questions about the techniques:

Preprocessing:

- Is normalization improving the results in this case?
- What is a good number of dimensions to be obtained by the PCA?
- Is Fischer discriminant good in this case?
- Is Autoencoder good in this case?

Related to classical classifiers:

- What is a good value of k for the k-nn algorithm?
- Can be obtain better results when using more than one normal density functions for every class?

Related to Neural Classifiers

- What is a good number of hidden layers for the MLP?
- What is a good number of neurons for the MLP?
- What is the relevance of the training algorithm of the MLP?
- What is good dimensionality of the SOM?
- What is a good number of neurons for the SOM?
- There is a 2D visual map of the digits build using a SOM

Related to the use of available data

- How should be the data used for training and for testing?
- What is a good number of epochs for both MLP and SOM?
- What is the relevance of the samples order during training for both MLP and SOM?

General remarks

- The document is clear and convincing
- There is a clear comparative study of all the methods with a clear conclusion on which to apply
- There is a study of the computational time for classification and it is adequate compare to other presentations
- There are some additional points of views, comments or improvements in the provided solutions