IML Course Project 2013

Submission: Date: Jan 17th 2013. **Groups**: up to 3 students.

Background:

In this project you will be solving a face recognition problem. Each image is represented by a column vector of length 6222. These features are composed out of 102 groups of 61 elements each. The elements of each group appear consecutively, i.e., the first 61 elements belong to the first group, the next 61 elements belong to the second, etc. Each such group is a histogram (a result of counting).

Main dataset -- train and test:

The face recognition problem is represented as a same/not same problem. Given two vectors: v1 and v2, you need to decide whether the two face images they encode belong to the same person (label +1) or not (-1). The pairs of training vectors are ordered as corresponding columns of the matrices X1train and X2train. For example, the 5th training example is the pair (X1train(:,5), X2train(:,5)). The label of this specific training pair is ytrain(5) == 1.

X1train -- 6222 x 3600 -- the first image of each training pair
X2train -- 6222 x 3600 -- the second image of each training pair
ytrain -- 3600 x 1 -- training labels

Note: There are repeated persons and even repeated images among the columns of X1train and X2train. This may cause problems when using cross validation during the development of your solution. In order to partly resolve this issue, the training set is split into three parts. The persons are not shared among the parts; Each part has a different group of individuals and the groups do not overlap. Naturally, no images are shared among the three parts. The ids of each part (1,2, or 3) are given as the vector gidtrain.

gidtrain -- 1 x 3600 -- the group of each training pair [1,2,3]

You are tested on the matching columns of X1test and X2test, a total of 2400 pairs; Each test pair is of the form (X1test(:,i), X2test(:,i)) for i = 1:2400. Your task is to predict the labels [-1,1] of each test pair. These are stored in the vector ytest.

X1test -- 6222 x 2400 -- the first image of each test pair
X2test -- 6222 x 2400 -- the second image of each test pair

ytest -- 2400 x 1 -- test labels "and to you they're gold, and you don't get them. Why?". Your work will be scored mostly by the accuracy of predicting this vector.
The persons in the test set do not appear in the training set.

Additional training data:

To facilitate more learning we provide you with additional supervised information. You may choose to use this additional dataset, which is provided as the matrices XXextratrain and YYextratrain; However, you are not obligated to do so.

This information is given as a multiclass classification problem and not as same/not-same pairs. If yyextratrain(i)==yyextratrain(j) then XXextratrain(:,i) and XXextratrain(:,j) encode face images of the same person; if yyextratrain(i)~=yyextratrain(j) then XXextratrain(:,i) and XXextratrain(:,j) encode face images of two different persons.

XXextratrain -- 6222 x 71843 - each column is the encoding of one image
yyextratrain -- 71843 x 1 -- contains the ids of the persons in the images

Important: the persons in this additional dataset do not overlap with the persons in the same/not-same dataset above.

Data files:

All data is available as a Matlab data file:
dataforproject.mat 339.0 MB
https://mega.co.nz/#!eAclBT5K!ey8UBgJPh18BXcvfa0fUyCwhfpAKt-YYEo85swOohog

A local copy exists here:
http://www.cs.tau.ac.il/~wolf/foriml/dataforproject.mat

```
>> load dataforproject.mat
>> whos
  Name                Size                  Bytes  Class     Attributes

  X1test            6222x2400             119462400  double
  X1train           6222x3600             179193600  double
  X2test            6222x2400             119462400  double
  X2train           6222x3600             179193600  double
  XXextratrain      6222x71843           3576057168  double
  gidtrain             1x3600                 28800  double
  ytrain            3600x1                    28800  double
  yyextratrain     71843x1                   574744  double
```

Deliverables:

A single zip file called X.Y.Z.zip (where X, Y, and Z are the id numbers of the submitting students) containing the following:

I. The vector ytest of size 2400 x 1, saved as ytest.mat

II. A fully documented Matlab code, which includes the script gogo.m that starts with loading the data (load dataforproject.mat) and ends with saving your solution (save ytest.mat ytest).

III. A file called readme.txt or readme.pdf that explains your solution and also documents the other alternatives you have tried during the development process. For each alternative (including the submission one), please provide: (1) a script of the form gogo_alt1.m (2) the performance score you used during development in order to compare the alternatives. You can use a table as the one below:

script name	**description summary** *[I'm just making this up, "my report" also refers to a fictional readme.pdf file]*	**score 1 (Section 3.1 of my report)**	**score 2 (Section 3.2 of my report)**
gogo_alt1.m	kernel pca (RBF, gamma = 0.1, dim=20) followed by AdaBoost. No use of Additional training data (ATD). See section 2.1 of my report.	0.73	0.73
gogo_alt2.m	feature normalization to the range [0..1] followed by polynomial svm d=4. ATD was used for …. See section 2.2 of my report.	0.84	0.84
...			

You should submit according to the submission guidelines of the HW as published in the course web site: A hard copy of the files submitted to the TA's mailbox in addition to the email submission to ml.intro.2013@gmail.com (The subject of the email should be "final project X.Y.Z").

You should also follow carefully the 'programming assignment' guidelines - any official and publicly available software package may be used as long as an exact reference is indicated and usage instructions are detailed. All other software must be original and included in your submission.