**Data visualization using MDS**

**DATASET**

This data is a subset of OCR data taken from

http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

**VIEW AN IMAGE**

Each row ‘ocr.txt’ consist of images of the digit 2,3 or 4. Each image is a 16x16 image,

stored in one row as a 256 dimensional vector. You can view any images as follows:

X = matrix(read.table(‘ocr.txt’))

r = X[1,] #first digit image, i.e., image in row 1

im = matrix(r,nrow=16,byrow=TRUE) #convert vector to image

image(im[,ncol(im):1]) #view image

Similarly you can view any image you like for any row of train and test matrices

**TASK 1: MULTI-DIMENSIONAL SCALING (MDS) For DATA VISUALIZATION**

Ypu can read about MDS from Alpaydin’s book. Traditionally this algorithm was used to

create a 2D map of all the countries. When making an atlas, the challenge is to map all

the points on a sphere (earth) to a 2D surface (paper). MDS algorithm is given below:

Given a data matrix X (mxn) without the labels, proceed as follows:

1. Center the data so that the mean of all columns is zero. Subtract from each element of

the matrix X, the mean of its corresponding column. The new centered data matrix is

! .

2. Get B matrix, which denotes the pair wise dot product between each image. In other

words, it is the dot product between different instances. B = ! . B is an mxm

matrix.

3. Find the Eigen values λ and the corresponding Eigen vectors **e** of matrix B. Sort the

Eigen values in descending order.

4. Retain only the first two highest Eigen values and only the first two corresponding

Eigen vectors.

5. Get a new transformed dataset as Z = VS, where V is a matrix with columns as eigen

values (only mx2) and S is diagonal matrix (2x2) with entries sii=sqrt(λi). Z is two

dimensional data because we retained only the 2 highest eigen values. If we had

retained k eigen values then we would be left with an mxk matrix Z.

6. Now take the Z matrix. Each row of this matrix corresponds to an image and we have

each image represented by only two features, which can be plotted.

7. Make a scatter plot in 2D space for each digit in a **different color** (the digits are 2,3,4).

You’ll need to find all rows of Z that correspond to 2 and then plot them in one color.

Repeat the same for 3 and 4. It is important to plot each digit in a different color so

that you can see the relationship of each digit which is the same and its relationship

with each digit, which is different.

*XC*

*XCXT*

*C*

**TASK 2: DISPLAY EIGEN DIGITS**

Eigen digits are successfully used for building OCR systems. Every digit has a unique

signature which can be captured very well using eigen values and eigen vectors of the

covariance matrix of data.

Implement the following algorithm:

1. Get a subset of data matrix ! , which denotes only those rows which correspond to

label = 2.

2. Get the covariance matix ! of data matrix ! .

3. Get the eigen values and eigen vectors of ! .

4. Reshape the the first 4 eigen vectors as 4x4 matrices corresponding to the highest

eigen values and display all four of them using image in R.

5. Reshape the the last 4 eigen vectors corresponding to the lowest eigen values and

display them using image in R.

6. Repeat the above for digit 3

**R FUNCTIONS**

1. **Eigen:** for finding Eigen values and Eigen vectors. Read its documentation. R would

return a sorted list of Eigen values.

2. **Plot** and **points**: for making the scatter plot of different digits on the same graph.