Assignment-1

Apply discriminant analysis to recognize the digits in the MNIST

Problem Set 1

In this problem we will apply discriminant analysis to recognize the digits in the MNIST data set (http://yann.lecun.com/exdb/mnist/). As a bonus problem we will construct "Fisher digits". We will train our model using the training data sets ("train-images-idx3-ubyte.gz" and "train-labels-idx1-ubyte.gz") and test the performance using the test data set ("t10k-images-idx3-ubyte.gz" and "t10k-labels-idx1-ubyte.gz").

1. The images are 28 x 28 pixels in gray-scale. The categories are 0, 1, ... 9. We concatenate the image rows into a 28 x 28 vector and treat this as our feature, and assume the feature vectors in each category in the training data "train-images-idx3-ubyte.gz") have Gaussian distribution. Draw the mean and standard deivation of those features for the 10 categories as 28 x 28 images using the training images ("train-images-idx3-ubyte.gz"). There should be 2 images for each of the 10 digits, one for mean and one for standard deviation. We call those "mean digits" and "standard deviation digits" in CSE455/555.

2. Classify the images in the testing data set ("t10k-images-idx3-ubyte.gz") using 0-1 loss function and Bayesian decision rule and report the performance. Why it doesn't perform as good as many other methods on LeCuns web page? Before coding the discriminant functions, review Section 2.6.

3. [Optional] Construct the "Fisher digits" from the MNIST data set according to Sections 3.8.2 and 3.8.3. This web page on Fisher faces (http://www.scholarpedia.org/article/Fisherfaces) and this web page (https://www.bytefish.de/blog/fisherfaces/) might be helpful. Answer two questions about these sections: (a) Why should the vector w minimizing Eq. (103) satisfy Eq. (104)? (b) Why should the between-class scatter matrix in Eq. (115) is $n_1 * n_2 / n$ times the one in Eq. (102) in two-class case (i.e., c=2)? In addition, convince ourselves that Eq. (125) is the quotient between two "volumes" by referring the Wikipedia page on determinant (https://en.wikipedia.org/wiki/Determinant).

Please submit your code, documentation and solutions electronically through UBLearns. Please typeset your mathematics in LaTeX or Word.

I strongly recommend you to use the problems after each chapters to check your understanding of the learning materials. But I will avoid using those problems in problem sets since the solution manual (http://home.iitk.ac.in/~crkrish/MLT/PCDudaHartStorkSlotions.pdf) is widely accessible. I will also avoid using the "stock" problems common to machine learning, pattern recognition and signal processing.