

Math 77C Project 3 (Due June 14)

Instructions: This project will require some MATLAB code. It will be convenient to answer the questions below and provide the accompanying MATLAB code in the same `.m` file. When turning in the project, please name your file `project3_yourlastname.m` and email it to `eesser@uci.edu`. If you are submitting multiple files, please zip them together in a file named `project3_yourlastname.zip`.

1. Let X_1 and X_2 be the lists of points in \mathbb{R}^2 shown in Figure 1 as the blue and red classes respectively. As matrices, each column of X_1 and X_2 corresponds to a point. We can think of these points as feature vectors for objects belonging to the two classes. We will use Fisher's linear discriminant to design a linear classifier that can be used to predict which class a given feature vector corresponds to. The data can be found in `X1.mat` and `X2.mat`. After downloading these `.mat` files, type `load X1; load X2;` to load the data.

1a. Find the number of points N_1 and N_2 in each class as well as the class means m_1 and m_2 .

1b. Use Fisher's linear discriminant to find a good direction w to project the points onto in order to maximize the ratio of the between-class variance to the within-class variance. Normalize w so that it is a unit vector.

1c. The projection of a point x onto the line spanned by w is given by $(w^T x)w$. Let $Z_1 = w^T X_1$ and $Z_2 = w^T X_2$. Assume Z_1 and Z_2 are normally distributed with means μ_1, μ_2 and variances σ_1^2, σ_2^2 . Estimate the parameters of these Gaussians by computing the maximum likelihood.

1d. Find w_0 such that the linear discriminant function $w^T x + w_0$ minimizes the classification error.

1e. Apply this linear classifier to the given data X_1 and X_2 . What fraction of points in class 1 are correctly classified? What fraction of points in class 2 are correctly classified?

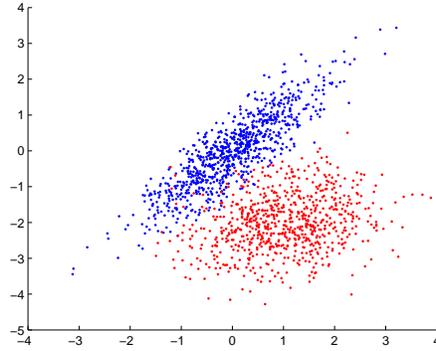


Figure 1: Class 1 (blue) and class 2 (red)

2. The goal of this project is to design a feature vector and a linear discriminant function to distinguish between images of handwritten numbers 3 and 5. Training and testing sets of 28 by 28 images of handwritten numbers are available at <http://www.cs.nyu.edu/~roweis/data.html>. The file `digits.mat` contains training images `X3image`, `X5image` and testing images `Y3image`, `Y5image`. Example images are shown in Figure 2.

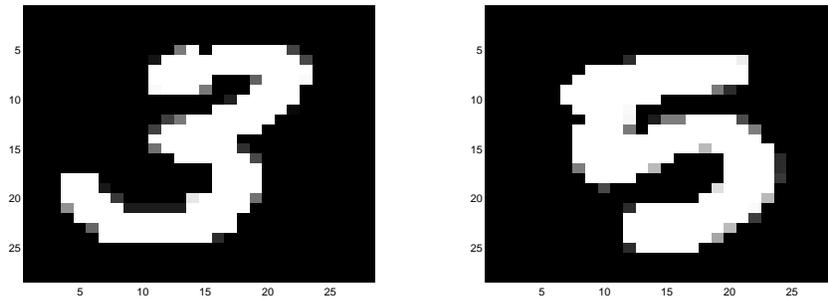


Figure 2: Example images of handwritten numbers

2a. Design at least two features (functions of the image) that you think could be used to distinguish between 3 and 5, and define a corresponding feature vector for each image. Higher dimensional feature vectors with more than just two features will probably be more effective. You may for example consider using the entire image itself as the feature vector.

2b. Use the training sets to define a linear discriminant function of the form $y(x) = w^T x + w_0$, where $y(x) < 0$ would classify the feature vector x as belonging to class 3

and $y(x) \geq 0$ would classify x as belonging to class 5. You may for example use Fisher's discriminant to determine a good choice of w .

2c. Use your linear discriminant function to classify the images in the testing set.

2d. What fractions of points in class 3 and in class 5 are correctly labeled? Show the corresponding confusion matrix.

2e. Classify the images in the testing set by randomly assigning a class to each image with equal probability. What is the resulting confusion matrix? How does it compare to the results in 2d?