Statistical Learning-Classification
STAT 441 / 841

Assignment 1
Department of Statistics and Actuarial Science
University of Waterloo

Due: Thursday September 28, by 1 pm
Drop your assignment to Drop Box 6 located across from MC4066.
The box has a ST/ACTSC label on it and your course has been divided alphabetically.
e.g., A-J, K-S, T-Z f
Instructor: Ali Ghodsi

Policy on Lateness: Slightly late assignments (up to 24 hs after due date) are accepted with 10% penalty. No assignment are accepted after 24 hs after the due date.

1. The Matlab data file 0_1_2.mat [1]contains 300 handwritten 0's, 1's and 2's images (one hundred each) scanned from postal envelopes, like the ones shown below.

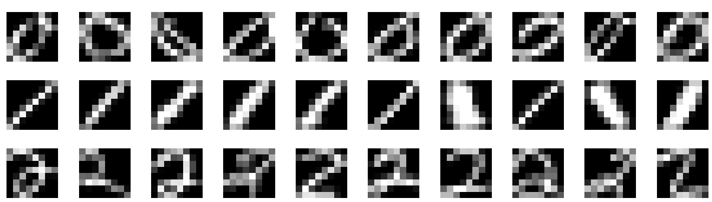

Figure 1:

These images are stored as a 64×300 matrix. Each column of the matrix is an 8×8 greyscale image (the pixel intensities are between 0 and 1). Figure 2 illustrates the two most significant dimensions found by PCA.

(*a*) Reproduce Figure 2. You need to find the low-dimensional mapping Y_{PCA} by PCA, and then call the function *plotimages* provided in the course web page. You need to implement PCA yourself. You may use *svd* , *eig*, and *eigs* functions in Matlab but you cannot use Matlab built-in functions *pca*, and *princomp*.

[1].mat file can be imported in Python with scipy:

https://docs.scipy.org/doc/scipy/reference/generated/scipy.io.loadmat.html

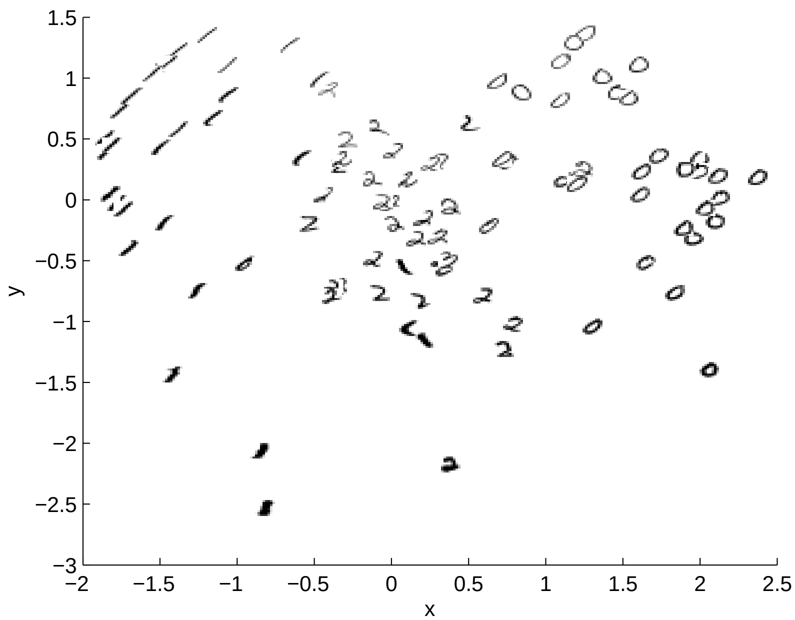

Figure 2: *A canonical dimensionality reduction problem from visual perception. The input consists of a sequence of 64-dimensional vectors, representing the brightness values of 8 pixel by 8 pixel images of digits 0, 1 and 2. Applied to* $n = 300$ *raw images. A two-dimensional projection is shown, with the original input images.*

(b) Produce a figure similar to Figure 2, but this time use FDA to map the data in 2-dimensional space.

(c) Use the two-dimensional data Y_{PCA} as the covariate. Use LDA and QDA to compute a linear and a quadratic decision boundary and report the analytic form of the boundaries.

(d) Plot the decision boundaries on the figure that you produced in part (a)

(e) Implement LDA as explained in class (Based on the Euclidian distance between points and the mean of each class). Report the error rates when LDA is applied to this data set.

(f) Can we implement QDA based on the Euclidian distance between points and the mean of each class? If your answer is yes explain how. Does this identical to the QDA that you computed in c

2. Prove that if $X \mid Y = 0 \sim N(\mu_0, \Sigma_0)$ and $X \mid Y = 1 \sim N(\mu_1, \Sigma_1)$, then the Bayes rule is

$$h^*(x) = \begin{cases} 1 & if\ r_1^2 < r_0^2 + 2\ log\left(\frac{\pi_1}{\pi_0}\right) + log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ 0 & otherwise \end{cases}$$

where

$$r_i^2 = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i),\ i = 0, 1$$

and $|A|$ denotes the determinant of a matrix A.

3. Consider a classifier with class conditional densities of the form $N(x|\mu_c, \sum_c)$. In LDA, we assume $\sum_c = \sum$ and in QDA, each $\sum_c$ is arbitrary. Here we consider the 2 class case in which $\sum_1 = k\sum_2$, for $k > 1$. That is, the Gaussian ellipsoids have the same shape, but the one for class 1 is wider. Derive an expression for the decision boundary.

Only for Grad Students

4. Suppose we have features $x \in R^d$, a two-class response, with class sizes n_1, n_2, and the target coded as $-n/n_1, n/n_2$.

   $a)$ Show that the LDA rule classifies to class 2 if

   $$x^T\hat{\Sigma}^{-1}(\hat{\mu_2} - \hat{\mu_1}) > \frac{1}{2}\hat{\mu_2}^T\hat{\Sigma}^{-1}\hat{\mu_2} - \frac{1}{2}\hat{\mu_1}^T\hat{\Sigma}^{-1}\hat{\mu_1} + log(\frac{n1}{n}) - log(\frac{n2}{n}),$$

   and class 1 otherwise.

   $b)$ Consider minimization of the least squares criterion

   $$\sum_{i=1}^{n}(y_i - \beta_0 - \beta^T x_i)^2.$$

   show that the solution $\hat{\beta}$ satisfies

   $$[(n-2)\hat{\Sigma} + \frac{n_1 n_2}{n}\hat{\Sigma_B}]\beta = n(\hat{\mu_2} - \hat{\mu_1})$$

   where $\hat{\Sigma_B} = (\hat{\mu_2} - \hat{\mu_1})(\hat{\mu_2} - \hat{\mu_1})^T$.

   $c)$ Show that $\hat{\Sigma_B}\beta$ is in the direction $(\hat{\mu_2} - \hat{\mu_1})$ and thus

   $$\hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu_2} - \hat{\mu_1}).$$

   Therefore the least squares regression coefficient is identical to the LDA coefficient, up to a scalar multiple.

5. The true error rate of a classifier h is

$$L(h) = P(\{h(X) \neq Y\})$$

Consider the special case where $Y \in \mathcal{Y} = \{0, 1\}$. Let

$$r(x) = P(Y = 1 \mid X = x)$$

In this case the Bayes classification rule h^* is

$$h^*(x) = \begin{cases} 1 & if\ r(x) > \frac{1}{2} \\ 0 & otherwise. \end{cases}$$

Prove that the Bayes classification rule is optimal, that is, if h is any other classification rule then $L(h^*) \leq L(h)$