ECE 411 Homework 9 (Fall 2022)
Instructor: Dr. Chau-Wai Wong
Material Covered: Linear/Quadratic Discriminant Analysis, False Positive/Negative Rate, ROC Curve, Equal Error Rate

Problem 1 (20 points) [Decision Making and Associated Probability] Complete *ISLR-4.7.6–7.*

Problem 2 (20 points) [Quadratic Discriminant Analysis (QDA)]

a) Random variables X_1 and X_2 are two predictors/features that will be used later to generate a dataset for classification. They are related by a first-order autoregressive model defined as follows:

$$X_2 = \rho X_1 + e, \tag{1}$$

where $\rho \in (0, 1)$ is a fixed constant, both X_1 and X_2 are of mean μ and variance σ^2, and $e \sim \mathcal{N}\left(0, (1-\rho^2)\sigma^2\right)$ is independent of X_1. Prove that $\mu = 0$ and $\text{cov}(X_1, X_2) = \rho\sigma^2$. Justify each entry of the variance-covariance matrix for random vector $[X_1, X_2]^T$ of the following form:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{2}$$

b) Using Eq. (1), simulate 500 data points for class 1 and class 2 with following multivariate Gaussian distributions respectively:

$$\begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \tag{3}$$

$$\begin{bmatrix} X_1^{(2)} \\ X_2^{(2)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} d \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right), \tag{4}$$

where $\sigma^2 = 1$, $\rho = 0.8$, and $d = 4$. Plot all data points in plane. Use "o" for class 1 and "*" for class 2. Note that for class 2, the center is $(d, 0)$.

c) The decision boundary is a set of 2nd-order curves shown as follows:

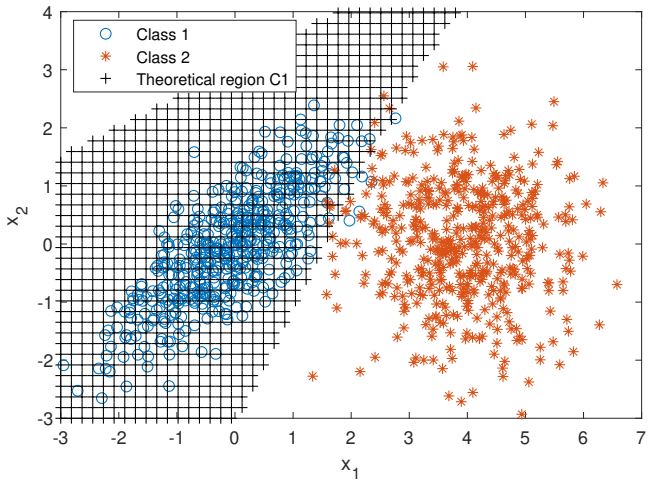
$$\rho c(x_1^2 + x_2^2) - 2cx_1x_2 + 2dx_1 - d^2 + \sigma^2 \ln(1-\rho^2) = 0, \tag{5}$$

where $c = \rho/(1-\rho^2)$. Use a brute-force method to draw on the data plot the theoretical decision region for class 1. Your result will be similar to the example shown on the next page.

d) (Bonus, 5 points) Start from the density function of the multivarite Gaussian, prove the expression for decision boundary given in c). Assume equal prior.

Problem 3 (20 points) [Two-Class Discrimination]

a) Complete *ISLR-4.7.3* to obtain a discriminant score/function $\delta_k(x)$ for class k. (Hint: After taking $\ln[\pi_k f_k(x)]$, remove all additive terms that do not contain x AND k.)



b) Prove that the decision threshold between class 1 and class 2 is one of the roots of the following quadratic equation:

$$\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)x^2-2\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)x+\left(\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}\right)+2\ln\left(\frac{\pi_2\sigma_1}{\pi_1\sigma_2}\right)=0. \tag{6}$$

c) Let $\mu_1=0$, $\mu_2=5$, $\sigma_1=1$, $\sigma_2=2$, and assume equal prior for both classes. Simulate 5,000 points for each class. Draw histograms for the two classes in the same plot. What is the theoretical decision threshold according to b)? Use this decision threshold to calculate a confusion matrix. Clearly indicate the dimensions for ground truth and for predicted results, respectively. What are the False Positive Rate (FPR) and False Negative Rate (FNR)?

d) Generate an ROC curve by varying threshold from the small value to the largest value of the overall dataset. (Hint: Your ROC curve should tell you that at 10% false positive, the true positive rate should be from 96%–98%.)

Problem 4 (Bonus, 20 points) [Equal Error Rate for Gaussian] Assume data points are generated from two distributions, namely, $\mathcal{N}(\mu_0,\sigma_0^2)$ for Class 0 and $\mathcal{N}(\mu_1,\sigma_1^2)$ for Class 1, where $\mu_0<\mu_1$ and $\sigma_0>\sigma_1$.

a) Draw by hand an illustration that contains the PDFs for both classes. Your illustration must reflect the relative relations of the means and standard deviations. Pick an arbitrary decision threshold on the horizontal axis and denote it as η. Use shades to label the areas under curves corresponding to the false positive rate (FPR) and the false negative rate (FNR), respectively. What are the analytic expressions for FPR(η) and FNR(η)? Express them using CDFs $F_0(\cdot)$ and $F_1(\cdot)$, where the CDF for class i is defined as $F_i(x)=\mathbb{P}[X_i\le x]$, $i=1,2$. (Reviewing your undergraduate statistics textbook on CDF will significantly help.)

b) Use the relationship between $F_i(\cdot)$ and $\Phi(\cdot)$ after standardizing the Gaussian random variable X_i into a standard Gaussian random variable Z_i

$$F_i(x)=\mathbb{P}[X_i\le x]=\mathbb{P}\left[Z_i=\frac{X_i-\mu_i}{\sigma_i}\le\frac{x-\mu_i}{\sigma_i}\right]=\Phi\left(\frac{x-\mu_i}{\sigma_i}\right), \tag{7}$$

where $\Phi(\cdot)$ is the CDF for the standard Gaussian random variable. Show that

$$\text{FPR}(\eta) = 1 - \Phi\left(\frac{\eta - \mu_0}{\sigma_0}\right) \text{ and } \text{FNR}(\eta) = \Phi\left(\frac{\eta - \mu_1}{\sigma_1}\right). \tag{8}$$

c) Prove that a decision threshold leading to the equal error rate (EER) is of the following form:

$$\eta^* = \frac{\sigma_1\mu_0 + \sigma_0\mu_1}{\sigma_0 + \sigma_1}. \tag{9}$$

(Hints: Set $\text{FPR}(\eta^*) = \text{FNR}(\eta^*)$. Exploit this property: The left and right tails of a *standard* Gaussian distribution have the same probability.)

d) Prove that

$$\text{EER} = \text{FPR}(\eta^*) = \text{FNR}(\eta^*) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma_0 + \sigma_1}\right). \tag{10}$$

Problem 5 (20 points) [Empirical ROC Curves for Gaussian] This problem directly reuses some theoretical result of Problem 4. Let $\mu_0 = 0$, $\mu_1 = 5$, $\sigma_0 = 2$, and $\sigma_1 = 1$. Simulate $N = 5000$ data points for each class.

a) Generate an empirical ROC curve by varying threshold from the smallest value to the largest value of the overall dataset.

b) What is the EER that can be read from the empirical ROC curve? What is the theoretical EER given by Problem 4d? Are they close?

c) Draw the theoretical ROC curve using the results in Problem 4b. Is the empirical ROC curve consistent with theoretical one?

d) Plot another two empirical ROC curves for $N = 500$ and $N = 50000$. Describe their differences in visual appearance, and explain why.

Problem 6 (Bonus, 10 points) [ClassEval] (Try it on or after mid-November) Have you completed ClassEval? It can be found at:

`http://go.ncsu.edu/cesurvey`

Grading: (a) Yes = 10 points, thank you! Please attach the screenshot of the confirmation page. (b) I promise to do it soon = 2 point for good intentions. (c) No = 0 points, a possibly honest answer, but why not spend 5 minutes and get 10 points?