ECE 492-45 Homework 7

Material Covered: ROC Curve, Equal Error Rate, Maximum Likelihood, Generalized Linear Models, Cross-Validation

Problem 1 (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 $F_i(x) = \mathbb{P}[X_i \le x],\ i = 1, 2$.

(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

$$F_i(x) = \mathbb{P}[X_i \le x] = \mathbb{P}\left[Z_i = \frac{X_i - \mu_i}{\sigma_i} \le \frac{x - \mu_i}{\sigma_i}\right] = \Phi\left(\frac{x - \mu_i}{\sigma_i}\right), \tag{1}$$

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

$$\text{FPR}(\eta) = 1 - \Phi\left(\frac{\eta - \mu_0}{\sigma_0}\right) \ \text{ and } \ \text{FNR}(\eta) = \Phi\left(\frac{\eta - \mu_1}{\sigma_1}\right). \tag{2}$$

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

$$\eta^* = \frac{\sigma_1\mu_0 + \sigma_0\mu_1}{\sigma_0 + \sigma_1}. \tag{3}$$

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

(d) Prove that

$$\text{EER} = \text{FPR}(\eta^*) = \text{FNR}(\eta^*) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma_0 + \sigma_1}\right). \tag{4}$$

Problem 2 (20 points) [Empirical ROC Curves for Gaussian] This problem is a continuation of Problem 1. 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 1(d). Are they close?

(c) Draw the theoretical ROC curve using the results in Problem 1(b). 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 3 (20 points) [Maximum Likelihood Estimator (MLE)]

(a) Calculate the MLE for variance θ (or σ^2, a more common notation if you prefer) for a random sample $X_1, \ldots, X_n$ drawn from the normal distribution with PDF shown as follows:

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-(x-\mu)^2/2\theta}, \quad -\infty < x < \infty. \tag{5}$$

(b) Calculate the MLE for parameter b for a random sample $X_1, \ldots, X_n$ drawn from an exponential distribution with PDF of the following form:

$$f(x;b) = \frac{1}{b}\, e^{-x/b}, \quad x \geq 0. \tag{6}$$

(c) The exponential distribution is more often parameterized using the rate parameter λ with PDF of the following form:

$$f(x;\lambda) = \lambda\, e^{-\lambda x}, \quad x \geq 0. \tag{7}$$

Use the invariance principle of MLE, show that $\hat{\lambda}_{\text{MLE}} = 1/\bar{X}$.

Problem 4 (20 points) [Generalized Linear Model (GLM)] Response $Y_i \sim \mathrm{B}(n, p_i)$ is a binomial random variable in which n is known. The (conditional) PDF is shown as follows:

$$\mathbb{P}[Y_i = k | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = \binom{n}{k} p_i^k (1-p_i)^{n-k}, \quad k \in \{0, 1, \ldots, n\}. \tag{8}$$

(a) Explain why the linear regression may not the best fit to find the relation between Y_i and a set of predictors $X_{i,1}, \ldots, X_{i,q}$.

(b) One proposes to link the conditional mean μ_i and the predictors $\underset{\sim}{x}_i$ using a generalized linear model shown as follows:

$$g(\mu_i) = \underset{\sim}{\beta}^T \underset{\sim}{x}_i \tag{9}$$

where $g(u) = \log(\frac{u}{n-u})$ and $\mu_i = \mathbb{E}[Y_i | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = np_i$. From the variable transformation viewpoint, show that $g(\cdot)$ matches the ranges for the two sides of Eq. (9).

(c) Rewrite the PDF into an exponential family form shown as follows:

$$f_Y(y;\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{10}$$

where θ is the natural parameter. Show that $g(\cdot)$ in (b) is the canonical link function when taking μ_i as the input.

Problem 5 (20 points) [Cross-Validation] Complete ISLR-5.3.1–3.