ECEN 649 Pattern Recognition – Spring 2018

Problem Set 2

Due on: Mar 1

1. Consider 1-dimensional Cauchy class-conditional densities:

$$p(x|Y=i) \;=\; \frac{1}{\pi b}\,\frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}\,, \qquad i=0,1\,,$$

where $-\infty < a_0 < a_1 < \infty$ are location parameters (there are no means for Cauchy distributions), and $b > 0$ is a dispersion parameter. Assume that the classes are equally likely, i.e., $P(Y=0) = P(Y=1) = \frac{1}{2}$.

   (a) Determine the Bayes classifier.

   (b) Determine the Bayes error as a function of the parameters a_0, a_1, and b.

   (c) Plot the Bayes error as a function of $(a_1 - a_0)/b$ and explain what you see. In particular, what are the maximum and minimum (infimum) values of the curve and what do they correspond to?

2. Suppose that it has been determined that the success of a student in a certain pass/fail class, which is coded by a binary variable Y, depends on the number of hours watching TV/day (T) and the number of beers ingested/day (B) as

$$Y \;=\; \begin{cases} 1 \text{ (pass)}, & \text{if } T+B+N \leq 7\,, \\ 0 \text{ (fail)}, & \text{otherwise.} \end{cases}$$

where $T, B \sim$ Exponential($\lambda = 1$), and N is noise that accounts for the uncertainty in the model.

   (a) Assuming that $N \sim$ Exponential(λ), where $\lambda > 0$ is not necessarily equal to 1, find the Bayes classifier and the Bayes error as a function of λ when T and B are used as features. Plot the Bayes classifier and the Bayes error as a function of λ and explain what you see.

   (b) Now consider that $N \sim$ Gaussian$(0, \sigma^2)$. Plot the Bayes classifier as a function of σ. If the Bayes error cannot be computed in closed form, write the answer as an integral.

   (c) Prove that the general condition for the Bayes classifier to be equal to the noiseless classifier:

$$\psi^\star(T,B) \;=\; \begin{cases} 1 \text{ (pass)}, & \text{if } T+B \leq 7\,, \\ 0 \text{ (fail)}, & \text{otherwise.} \end{cases}$$

   is that the median of the noise distribution be equal to zero. (If F is a CDF, then the median of the distribution is $F^{-1}(1/2)$.)

3. Consider a variation of the pass/fail classification problem, where the variables T, B, and E are still independent and identically distributed, but now are each distributed uniformly on the interval $[0, 4]$, and the model for Y is

$$Y \;=\; \begin{cases} 1\,, & TBE \leq 8\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

Find the Bayes classifier and the Bayes error when

(a) T, B are available,

(b) only T is available.

Hint: The probability density function for the product of two independent uniform r.v.'s defined on the interval $[0, L]$ is given by:

$$f(x) = \frac{1}{L^2}\,\ln\frac{L^2}{x}\,, \qquad 0 < x < L^2,$$

with $f(x) = 0$ outside the interval $[0, L^2]$. In addition, note that $\int \ln x \;=\; x \ln x - x$.

4. This problem concerns the extension to the multiple-class case of concepts derived in class for the two-class case. Let $Y \in \{0, 1, \ldots, c-1\}$, where c is the number of classes, and let

$$\eta_i(x) \;=\; P(Y = i \mid X = x)\,, \quad i = 0, 1, \ldots, c-1\,,$$

for each $x \in R^d$. We need to remember that these probabilities are not independent, but satisfy $\eta_0(x) + \eta_1(x) + \cdots + \eta_{c-1}(x) = 1$, for each $x \in R^d$, so that one of the functions is redundant. In the two-class case, this is made explicity by using a single $\eta(x)$, but using the redundant set above proves advantageous in the multiple-class case, as seen below. Hint: you should answer the following items in sequence, using the previous anwers in the solution of the following ones.

(a) Given a classifier $\psi : R^d \to \{0, 1, \ldots, c-1\}$, show that its conditional error $P(\psi(X) \neq Y \mid X = x)$ is given by

$$P(\psi(X) \neq Y \mid X = x) \;=\; 1 - \sum_{i=0}^{c-1} I_{\psi(x)=i}\, \eta_i(x) \;=\; 1 - \eta_{\psi(x)}(x)\,.$$

(b) Assuming that X has a density (i.e., X is a continuous feature vector), show that the classification error of ψ is given by

$$\epsilon \;=\; 1 - \sum_{i=0}^{c-1} \int_{\{x \mid \psi(x)=i\}} \eta_i(x) p(x)\, dx\,.$$

(c) Prove that the Bayes classifier is given by

$$\psi^*(x) = \arg\max_{i=0,1,\ldots,c-1} \eta_i(x)\,, \quad x \in R^d\,.$$

Hint: Start by considering the difference between conditional expected errors $P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x)$.

(d) Show that the Bayes error is given by

$$\epsilon^* \,=\, 1 - E\left[\max_{i=0,1,\ldots,c-1} \eta_i(X)\right].$$

5. This problem concerns classification with a rejection option. Assume that there are c classes and $c+1$ "actions" $\alpha_0, \alpha_1, \ldots, \alpha_c$. For $i = 0, \ldots, c-1$, action α_i is simply to classify into class i, whereas action α_c is to reject, i.e., abstain from committing to any of the classes, for lack of enough evidence. This can be modeled as a Bayes decision theory problem, where the cost λ_{ij} of taking action α_i when true state of nature is j is given by:

$$\lambda_{ij} \;=\; \begin{cases} 0\,, & i = j, \text{ for } i,j = 0,\ldots,c-1 \\ \lambda_r\,, & i = c \\ \lambda_m\,, & \text{otherwise,} \end{cases}$$

where λ_r is the cost associated with a rejection, and λ_m is the cost of misclassifying a sample. Determine the optimal decision function $\alpha^* : R^d \to \{\alpha_0, \alpha_1, \ldots, \alpha_c\}$ in terms of the posterior probabilities $\eta_i(x)$ — see the previous problem — and the cost parameters. As should be expected, the occurrence of rejections will depend on the relative cost λ_r/λ_m. Explain what happens when this ratio is zero, 0.5, and greater or equal than 1.

6. Consider the general two-class Gaussian model, where

$$p(x|Y=i) \sim N_d(\mu_i, \Sigma_i)\,, \quad i = 0, 1\,.$$

In Discriminant Analysis, it is common to say that each class defines a *population* Π_i, for $i = 0, 1$, and that a sample (e.g., patient, fish, metal) X comes from population Π_i, which is denoted by $X \in \Pi_i$, if $Y = i$.

(a) Given a *linear discriminant* $g(x) = a^t x + b$, where $a \in R^d$ and $b \in R$ are arbitrary parameters (these are not the optimal parameters), compute the classification error of the associated classifier

$$\psi(x) \;=\; \begin{cases} 1\,, & g(x) = a^t x + b \geq 0 \\ 0\,, & \text{otherwise.} \end{cases}$$

in terms of Φ (the c.d.f. of a standard normal random variable), and the parameters $a, b, \mu_0, \mu_1, \Sigma_0, \Sigma_1, c_0$ and c_1, where μ_i and Σ_i are the parameters of the Gaussian populations and $c_i = P(X \in \Pi_i)$ are the prior probabilitites, for $i = 0, 1$.

Hint: The classification error is given by

$$\begin{aligned}\epsilon &= P(\psi(X) \neq Y) \\ &= P(\psi(X) = 1 \mid Y = 0)P(Y = 0) + P(\psi(X) = 0 \mid Y = 1)P(Y = 1)\,.\end{aligned}$$

In the language of Discriminant Analysis, this becomes:

$$\begin{aligned}\epsilon &= P(g(X) \geq 0 \mid X \in \Pi_0)P(X \in \Pi_0) + P(g(X) < 0 \mid X \in \Pi_1)P(X \in \Pi_1) \\ &= c_0\epsilon^0 + c_1\epsilon^1\,,\end{aligned}$$

where $c_i = P(X \in \Pi_i)$ and ϵ^i is the error *conditional* to class i, for $i = 0, 1$. The overall error ϵ is thus a convex combination of the conditional errors ϵ^0 and ϵ^1, where the weights are given by the prior probabilities. To compute the conditional error ϵ^i, one would have, in principle, to solve the multidimensional integral of a Gaussian density over a half space; for example, for class 0,

$$\epsilon^0 = \int_{\{x \mid g(x) \geq 0\}} p(x|Y = 0)\,dx = \int_{\{x \mid g(x) \geq 0\}} N_d(\mu_0, \Sigma_0)\,dx\,.$$

This integral can be solved using some tricks (see Prob 2.32 in DHS), but there is a much easier, "pattern-recognition" way of computing this. Notice that

$$\epsilon^0 = P(g(X) \geq 0 \mid X \in \Pi_0) = P(a^t Z + b \geq 0)\,, \text{ where } Z \sim N_d(\mu_0, \Sigma_0)\,.$$

Use the properties of the Gaussian distribution to write this in terms of Φ.

(b) Using the result from the previous item, show that if $\Sigma_0 = \Sigma_1 = \Sigma$ and $c_0 = c_1 = \frac{1}{2}$, then the Bayes error for the problem is given by

$$\epsilon^* = \Phi\left(-\frac{\delta}{2}\right)\,,$$

where $\delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0)}$ is the Mahalanobis distance between the classes. Therefore, in this case, there is a tight relationship (in fact, one-to-one) between the Mahalanobis distance and the Bayes error. What is the maximum and minimum (infimum) Bayes errors and when do they happen?

7. This problem shows that the a-priori probabilities can have a huge impact on the optimal classifier. We showed that in the Gaussian model with equal covariance matrices, the optimal classifier is a hyperplane that passes through the midpoint between μ_0 and μ_1, provided that

the classes are equally likely. State the condition on the prior probabilities $P(Y = 0)$ and $P(Y = 1)$ such that the hyperplane not ony does not pass through the midpoint between μ_0 and μ_1, but it does not pass between μ_0 and μ_1 at all.

8. We pointed out in class that $\epsilon_{\text{NN}} = 0 \Leftrightarrow \epsilon^* = 0$ and $\epsilon_{\text{NN}} = \frac{1}{2} \Leftrightarrow \epsilon^* = \frac{1}{2}$. The question is whether it is possible to find a problem where $\epsilon_{\text{NN}} = \epsilon^* = \delta$ with $0 < \delta < \frac{1}{2}$, i.e., an intermediate value not at the extremes 0 and $\frac{1}{2}$. Show that this is so, by considering a one-dimensional problem with class-conditional densities

$$p(x \mid Y = i) = \begin{cases} 1\,, & 0 \le x \le \frac{1}{2} \\ 1\,, & i + 1 \le x \le i + \frac{3}{2} \\ 0\,, & \text{otherwise,} \end{cases}$$

for $i = 0, 1$. Assuming that $P(Y = 0) = P(Y = 1) = \frac{1}{2}$, show that $\epsilon_{\text{NN}} = \epsilon^* = \frac{1}{4}$.

Hint: Plot the probability densities and posterior probabilities.