ECEN 649 Pattern Recognition – Spring 2017

Problem Set 2

Due on: Mar 2

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

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

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.

Solution: We need to solve the equation $P(Y=0|X=x^*) = P(Y=1|X=x^*)$. Since the classes are equally-likely, this is equivalent to $p(x^*|Y=0) = p(x^*|Y=1)$. By inspection of the Cauchy class-conditional densities, it is clear that this will happen if and only if

$$(x^*-a_0)^2 = (x^*-a_1)^2 \;\Leftrightarrow\; x^* = \frac{a_0+a_1}{2}$$

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

Solution:

$$\begin{aligned}
\epsilon^* &= \int_{-\infty}^{x^*} p(x|Y=1)\,P(Y=1)\,dx + \int_{x^*}^{\infty} p(x|Y=0)\,P(Y=0)\,dx \\
&= 2\,\frac{1}{2}\int_{x^*}^{\infty} p(x|Y=0)\,P(Y=0)\,dx \;\text{ (by symmetry)} \\
&= \int_{\frac{a_0+a_1}{2}}^{\infty} \frac{1}{\pi b}\,\frac{1}{1+\left(\frac{x-a_0}{b}\right)^2}\,dx
\end{aligned}$$

By making the substitution $u = \frac{x-a_0}{b}$, we obtain

$$\epsilon^* = \frac{1}{\pi}\int_{\frac{a_1-a_0}{2b}}^{\infty} \frac{1}{1+u^2}\,du \;=\; \frac{1}{\pi}\arctan|u|\,\Big|_{u=\frac{a_1-a_0}{2b}}^{u\to\infty} \;=\; \frac{1}{2} - \frac{1}{\pi}\arctan\left|\frac{a_1-a_0}{2b}\right|$$

(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?

Solution: We have

$$\epsilon^*(w) = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{w}{2}\right)$$

where, by definition, $w = \left|\frac{a_1 - a_0}{b}\right| > 0$. The plot of this function can be seen in Figure 1. We can see that the Bayes error decays monotonically with increased "standard separation" between the classes, i.e. with larger values of $\left|\frac{a_1 - a_0}{b}\right|$. For example, the Bayes error is halved (equal to 0.25) when $w = 2$, that is, $|a_1 - a_0|$ is equal to $2b$ units (b plays here a similar role to the standard deviation of Gaussian densities). From Figure 1, we can see that the maximum value of ϵ^* is 0.5, which occurs for $w = 0$, that is, $a_0 = a_2$. This corresponds to the case where the class-conditional densities are equal, so that there is maximal confusion between the classes — a Bayes error of 0.5 means that the best one can do is equivalent to flipping a coin. Conversely, as the distance $\left|\frac{a_1 - a_0}{2b}\right|$ becomes infinitely large compared to b, then the class-conditional densities are maximally separated and the Bayes error tends to its minimum (infimum, in this case) value of zero.

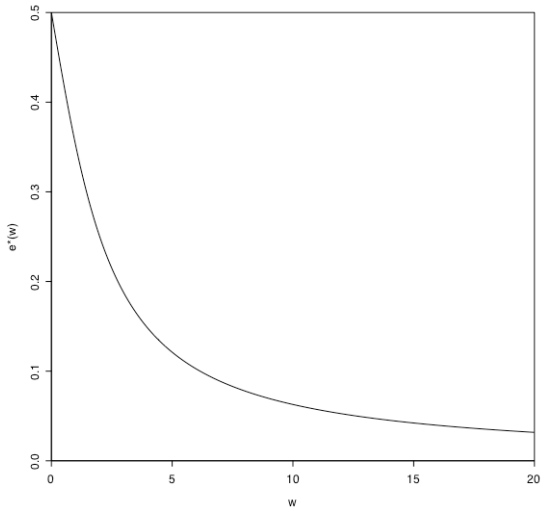

Figure 1: Bayes error as a function of standard separation between classes in the Cauchy case.

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 $N \sim \text{Gaussian}(0, \sigma^2)$. 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 $F^{-1}(1/2)$.)

Solution: This problem was solved in class.

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,

Solution: If variables T and B are available, we use the fact that

$$P(E \leq a) = \begin{cases} \frac{a}{4}, & 0 \leq a \leq 4 \\ 1, & a > 4 \end{cases}$$

to obtain:

$$\eta(T,B) = P(Y=1|T,B) = P(TBE \leq 8|T,B) = P\left(E \leq 8/TB|T,B\right)$$
$$= \begin{cases} \frac{2}{TB}, & TB \geq 2 \\ 1, & TB < 2 \end{cases}$$

Therefore, the Bayes decision is given by:

$$\psi^*(T,B) = \begin{cases} 1, & \eta(T,B) \geq \frac{1}{2} \\ 0, & \text{otw} \end{cases} = \begin{cases} 1, & TB \leq 4 \\ 0, & \text{otw} \end{cases}$$

There are two ways of computing the Bayes error. The first is via direct integration:

$$\begin{aligned}\epsilon^* &= E\left[\min\{\eta(T,B), 1-\eta(T,B)\}\right] \\ &= \int\limits_{\eta(T,B)\geq\frac{1}{2}} (1-\eta(T,B))f(T,B)\,dBdT + \int\limits_{\eta(T,B)<\frac{1}{2}} \eta(T,B)f(T,B)\,dBdT\end{aligned}$$

In our case, $f(T,B) = \frac{1}{16} I_{\{0\leq T\leq 4, 0\leq B\leq 4\}}$. Therefore

$$\epsilon^* = \frac{1}{16}\left[\sum_{i=1}^{3} \iint\limits_{R_i} (1-\eta(T,B))\,dBdT + \iint\limits_{R_4} \eta(T,B)\,dBdT\right] \tag{1}$$

where the regions R_i (see Figure 2) are given by:

$$\begin{aligned}R_1 &= \{0\leq T\leq 4,\ 0\leq B\leq \min\{4, 2/T\}\} \\ R_2 &= \{1/2\leq T\leq 1,\ 2/T\leq B\leq 4\} \\ R_3 &= \{1\leq T\leq 4,\ 2/T\leq B\leq 4/T\} \\ R_4 &= \{1\leq T\leq 4,\ 4/T\leq B\leq 4\}\end{aligned}$$

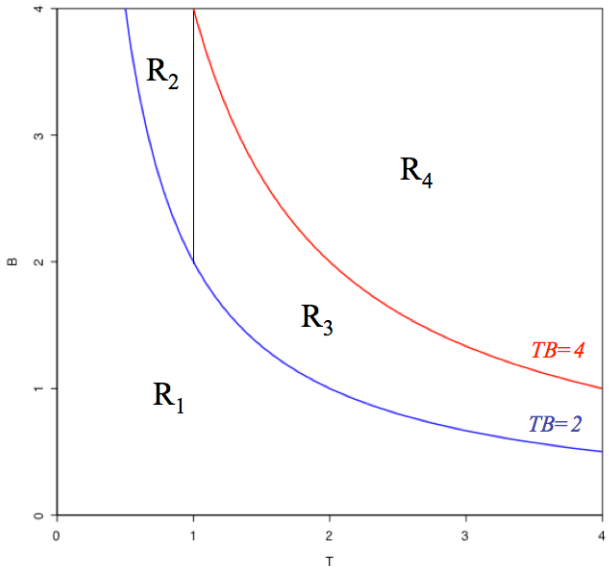

Figure 2: Integration regions in equation 1.

Notice that $1-\eta(T,B) = 0$ over R_1, $1-\eta(T,B) = 1-2/TB$ over R_2 and R_3, and $\eta(T,B) = 2/TB$ over R_4. Therefore, equation (1) becomes:

$$\epsilon^* = \frac{1}{16}\left[\sum_{i=2}^{3} \iint\limits_{R_i} 1\,dBdT - \sum_{i=2}^{3} \iint\limits_{R_i} \frac{2}{TB}\,dBdT + \iint\limits_{R_4} \frac{2}{TB}\,dBdT\right] \tag{2}$$

The special form of the regions R_2,R_3, and R_4 allows one to readily compute the double integrals in (2), to get:

$$\iint\limits_{R_2} 1\, dBdT = 2 - 2\ln 2 \quad \iint\limits_{R_3} 1\, dBdT = 4\ln 2$$

$$\iint\limits_{R_2} \frac{2}{TB}\, dBdT = (\ln 2)^2 \quad \iint\limits_{R_3} \frac{2}{TB}\, dBdT = \iint\limits_{R_4} \frac{2}{TB}\, dBdT = 4(\ln 2)^2$$

The integrals above are all straightforward to compute, needing only knowledge of the anti-derivative $\int \frac{1}{u}\, du = \ln u$. Substituting these into (2) gives:

$$\epsilon^* = \frac{1}{16}\left[2 + 2\ln 2 - (\ln 2)^2\right] \cong 0.1816\,.$$

The second way to compute ϵ^* is to use the approach that was used in class:

$$\begin{aligned}
\epsilon^* &= E[P(\psi(T,B) \neq Y|T,B)] \\
&= E[P(\psi^*(T,B) = 0, Y = 1|T,B) + P(\psi^*(T,B) = 1, Y = 0|T,B)] \\
&= E[I_{\psi^*(T,B)=0}\, \eta(T,B) + I_{\psi^*(T,B)=1}(1 - \eta(T,B))] \\
&= \int_4^{16} \eta(T,B) f(T,B)\, d(T,B) + \int_0^4 (1 - \eta(T,B)) f(T,B)\, d(T,B) \qquad (3) \\
&= -\frac{1}{8}\int_4^{16} \frac{\ln 16u}{u}\, du - \frac{1}{16}\int_2^4 \left(1 - \frac{2}{u}\right) \ln 16u\, du \\
&= \frac{1}{16}\left[2 + 2\ln 2 - (\ln 2)^2\right] \cong 0.1816\,.
\end{aligned}$$

This method of solution is overall simpler, as it requires only one-dimensional integration. Note, however, that it requires knowledge of the slightly more complex anti-derivative $\int \frac{\ln u}{u}\, du = \frac{1}{2}(\ln u)^2$.

(b) only T is available.

Solution: If only T is available, we use the hint to get

$$f_{E\times B}(a) = \begin{cases} \frac{1}{16}\ln\frac{16}{a}, & 0 \leq a \leq 16 \\ 0, & \text{otw} \end{cases}$$

and from this

$$P(EB \leq a) = \int_0^a f_{E\times B}(u)\, du = \begin{cases} \frac{a}{16}\left(1 + \ln\frac{16}{a}\right), & 0 \leq a \leq 16 \\ 1, & a > 16 \end{cases}$$

to obtain

$$\eta(T) = P(Y=1|T) = P(TBE \le 8|T) = P\left(EB \le 8/T|T\right)$$
$$= \begin{cases} \frac{1}{2T}(1+\ln 2T), & T \ge \frac{1}{2} \\ 1, & T < \frac{1}{2} \end{cases}$$

Since $\frac{1}{2T}(1+\ln 2T) = 1/2$ and $T \ge 1/2$ imply $T \cong 2.678$, the Bayes decision is given by:

$$\psi^*(T) = \begin{cases} 1, & \eta(T) \ge \frac{1}{2} \\ 0, & \text{otw} \end{cases} \cong \begin{cases} 1, & T \le 2.678 \\ 0, & \text{otw} \end{cases}$$

The Bayes error is given by:

$$\epsilon^* = \int\limits_{\eta(T)\ge\frac{1}{2}} (1-\eta(T))f(T)\,dT + \int\limits_{\eta(T)<\frac{1}{2}} \eta(T)f(T)\,dT$$

Since $f(T) = \frac{1}{4}I_{\{0\le T\le 4\}}$, and recalling the result of part a), we can write:

$$\epsilon^* \cong \frac{1}{4}\left[\int_{1/2}^{2.678}\left(1-\frac{1}{2T}-\frac{\ln 2T}{2T}\right)\,dT \;+\; \int_{2.678}^{4}\left(\frac{1}{2T}+\frac{\ln 2T}{2T}\right)\,dT\right] \cong 0.3031$$

The Bayes error for one variable is therefore larger than for two variables, as expected.

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

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

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

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

Solution: We have that

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

(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

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

Solution: Directly from the previous item,

$$\begin{aligned} \epsilon \;=\; \int P(\psi(X) \neq Y | X = x)\, p(x)\, dx \;&=\; 1 \;-\; \sum_{i=0}^{c-1} \int I_{\psi(x)=i}\, \eta_i(x)\, p(x)\, dx \\ &= \;1 \;-\; \sum_{i=0}^{c-1} \int\limits_{\{x|\psi(x)=i\}} \eta_i(x)\, p(x)\, dx\,. \end{aligned}$$

(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 $P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x)$.

Solution: Again using the result of item (a),

$$\begin{aligned} &P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x) \\ &\quad = \;\eta_{\psi^*(x)}(x) - \eta_{\psi(x)}(x) \;=\; \left[\max_{i=0,1,\ldots,c-1} \eta_i(X)\right] \;-\; \eta_{\psi(x)}(x) \geq 0\,, \end{aligned}$$

by definition of $\psi^*(x)$. Integration over the feature space yields

$$\epsilon - \epsilon^* \;=\; E\left[P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x)\right] \;\geq\; 0\,.$$

(d) Show that the Bayes error is given by

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

Solution: Using the result of item (a) and the definition of ψ^*,

$$\epsilon^* \;=\; E\left[P(\psi^*(X) \neq Y \mid X = x)\right] \;=\; 1 - E\left[\eta_{\psi^*(x)}(x)\right] \;=\; 1 - E\left[\max_{i=0,1,\ldots,c-1} \eta_i(X)\right] .$$

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.

Solution: The optimal decision function minimizes the conditional risk

$$R(\alpha(x) = \alpha_i | X = x) \;=\; \sum_{j=0}^{c-1} \lambda_{ij} \eta_j(x)$$

at each point $x \in R^d$. For $i = 0, 1, \ldots, c-1$, this gives

$$R(\alpha(x) = \alpha_i | X = x) \;=\; \lambda_m (1 - \eta_i(x))\,,$$

while for $i = c$ (rejection), one obtains

$$R(\alpha(x) = \alpha_c | X = x) \;=\; \lambda_r\,.$$

It becomes clear then that the optimal decision is

$$\alpha^*(x) \;=\; \begin{cases} \text{classify into class } i\,, & \text{if } i = \arg\max\limits_{j=0,1,\ldots,c-1} \eta_j(x) \text{ and } 1 - \max\limits_{j=0,1,\ldots,c-1} \eta_j(X) \leq \frac{\lambda_r}{\lambda_m}\,, \\ \text{reject}\,, & \text{if } 1 - \max\limits_{j=0,1,\ldots,c-1} \eta_j(X) > \frac{\lambda_r}{\lambda_m}\,. \end{cases}$$

Rejection depends on the magnitude of the ratio λ_r/λ_m in comparison to the "margin" $1 - \max \eta_j(X)$. The larger the latter is, the more confidence one has in choosing class $i = \arg\max \eta_j(x)$. If $\lambda_r/\lambda_m = 0$, then one will always reject, i.e., one is always unwilling to classify because the cost of rejection is too small (a degenerate case). If $\lambda_r/\lambda_m = 0.5$, then one will reject classification unless $\max \eta_j(X)$ is at least 0.5 (a reasonable case, if $c \geq 3$). If $\lambda_r/\lambda_m = 1$ then one will never reject, because the cost of rejection is too high (this corresponds to the classical case).

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,

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

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.

Solution: Using the properties of multivariate Gaussian distributions (see Lecture 2), we know that, if $Z \sim N_d(\mu, \Sigma)$, $a \in R^d$, and $b \in R$, then $a^T Z + b$ is a univariate Gaussian random variable with mean $a^T\mu + b$ and variance $a^T\Sigma a$. Therefore, following the hint,

$$\begin{aligned}\epsilon^0 \;&=\; P(g(X) \geq 0 \mid X \in \Pi_0) = P(a^t Z + b \geq 0) \\ &= 1 - F_{a^t Z+b}(0) \;=\; 1 - \Phi\left(-\frac{a^T\mu_0 + b}{\sqrt{a^T\Sigma_0 a}}\right) \;=\; \Phi\left(\frac{a^T\mu_0 + b}{\sqrt{a^T\Sigma_0 a}}\right)\,,\end{aligned}$$

since $Z \sim N_d(\mu_0, \Sigma_0)$. Similarly, we have

$$\epsilon^1 \;=\; P(g(X) < 0 \mid X \in \Pi_1) = P(a^t Z + b < 0) \;=\; F_{a^t Z+b}(0) \;=\; \Phi\left(-\frac{a^T\mu_1 + b}{\sqrt{a^T\Sigma_1 a}}\right)\,,$$

since $Z \sim N_d(\mu_1, \Sigma_1)$. Using the hint, one obtains

$$\epsilon \;=\; c_0\epsilon^0 + c_1\epsilon^1 \;=\; c_0\,\Phi\left(\frac{a^T\mu_0 + b}{\sqrt{a^T\Sigma_0 a}}\right) \;+\; c_1\,\Phi\left(-\frac{a^T\mu_1 + b}{\sqrt{a^T\Sigma_1 a}}\right)\,. \tag{4}$$

(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?

Solution: By plugging in the optimal values of a and b, with the assumption that $c_0 = c_1 = \frac{1}{2}$ (see Lecture 3),

$$\begin{aligned}a \;&=\; \Sigma^{-1}(\mu_1 - \mu_0) \\ b \;&=\; -\frac{1}{2}(\mu_1 - \mu_0)^T\Sigma^{-1}(\mu_1 + \mu_0)\end{aligned}$$

in eq. (4) and performing some algebraic simplifications, one obtains the desired result. The maximum value of the Bayes error $\epsilon^* = 0.5$ happens when $\delta = 0$, i.e., the class-conditional densities are equal to each other (maximum confusion), whereas the minimum (infimum, in this case) value $\epsilon^* = 0$ happens as $\delta \to \infty$, i.e., the classes are infinitely separated.

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.

   Solution:

   From equation (65) in DHS, we have that

$$x_0 = \frac{1}{2}(\mu_1 + \mu_0) - t(\mu_1 - \mu_0)$$

   where

$$t = \frac{1}{\Delta^2} \ln \frac{P(Y=1)}{P(Y=0)} \tag{5}$$

   Here, $\Delta = (\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)$ is the Mahalanobis distance between the classes. If $P(Y = 1) = P(Y = 0)$, then $t = 0$ and the decision hyperplane passes through the midpoint between μ_1 and μ_0. On the other hand, if $P(Y = 1) \neq P(Y = 0)$ (without loss of generality, let us assume $P(Y = 1) > P(Y = 0)$, so that $t > 0$), then we can see that x_0 moves along the line defined by μ_1 and μ_0, towards μ_0, according to the bias given by t. The critical point $x_0 = \mu_0$ corresponds to

$$t(\mu_1 - \mu_0) = \frac{1}{2}(\mu_1 - \mu_0) \Rightarrow t = \frac{1}{2}$$

   For $t > 2$, the decision hyperplane will not pass between the means. From (5), we can see that this is equivalent to

$$\ln \frac{P(Y=1)}{P(Y=0)} > \frac{1}{2}\Delta^2 \Rightarrow \frac{P(Y=1)}{P(Y=0)} > e^{\frac{1}{2}\Delta^2}$$

   The skewness of the situation is due to a large difference (large ratio) between the a-priori probabilities.

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

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

   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.

Solution:

We need to compute $\eta(x)$. First note that

$$p(x) = p(x|Y=0)P(Y=0) + p(x|Y=1)P(Y=1) = \frac{1}{2}\left[p(x|Y=0) + p(x|Y=1)\right]$$
$$= \begin{cases} 1\,, & 0 \le x \le \frac{1}{2} \\ \frac{1}{2}\,, & 1 \le x \le \frac{3}{2} \text{ and } 2 \le x \le \frac{5}{2} \\ 0\,, & \text{otherwise.} \end{cases}$$

Therefore,

$$\eta(x) = \frac{p(x|Y=1)P(Y=1)}{p(x)} = \frac{1}{2}\frac{p(x|Y=1)}{p(x)} = \begin{cases} \frac{1}{2}, & 0 \le x \le \frac{1}{2} \\ 1, & 2 \le x \le \frac{5}{2} \\ 0, & 1 \le x \le \frac{3}{2} \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

It follows that

$$\min\{\eta(x), 1-\eta(x)\} = \begin{cases} \frac{1}{2}\,, & 0 \le x \le \frac{1}{2} \\ 0\,, & 1 \le x \le \frac{3}{2} \text{ and } 2 \le x \le \frac{5}{2} \\ \text{undefined}\,, & \text{otherwise.} \end{cases}$$

Therefore, the Bayes error is given by

$$\epsilon^* = E[\min\{\eta(x), 1-\eta(x)\}] = \int \min\{\eta(x), 1-\eta(x)\}p(x)\,dx = \int_0^{\frac{1}{2}} \frac{1}{2}\,dx = \frac{1}{4}\,,$$

whereas the asymptotic 1-NN error rate is given by

$$\epsilon_{\text{NN}} = E[2\eta(x)(1-\eta(x))] = \int 2\eta(x)(1-\eta(x))p(x)\,dx = \int_0^{\frac{1}{2}} \frac{1}{2}\,dx = \frac{1}{4}\,.$$

Therefore, $\epsilon_{NN} = \epsilon^*$, even though $\epsilon^* \neq 0$ and $\epsilon^* \neq \frac{1}{2}$. This is made possible by the ingenious way of picking the class-conditional densities $p(x|Y=0)$ and $p(x|Y=0)$ (the ability to come up with such counter-examples often comes in handy in showing facts about pattern recognition).