PROBLEMS

Section 2.1

1. In the two-category case, under the Bayes decision rule the conditional error is given by Eq. 7. Even if the posterior densities are continuous, this form of the conditional error virtually always leads to a discontinuous integrand when calculating the full error by Eq. 5.

   (a) Show that for arbitrary densities, we can replace Eq. 7 by $P(error|x) = 2P(\omega_1|x)P(\omega_2|x)$ in the integral and get an upper bound on the full error.

   (b) Show that if we use $P(error|x) = \alpha P(\omega_1|x)P(\omega_2|x)$ for $\alpha < 2$, then we are not guaranteed that the integral gives an upper bound on the error.

   (c) Analogously, show that we can use instead $P(error|x) = P(\omega_1|x)P(\omega_2|x)$ and get a lower bound on the full error.

   (d) Show that if we use $P(error|x) = \beta P(\omega_1|x)P(\omega_2|x)$ for $\beta > 1$, then we are not guaranteed that the integral gives an lower bound on the error.

Section 2.2

2. Suppose two equally probable one-dimensional densities are of the form $p(x|\omega_i) \propto e^{-|x-a_i|/b_i}$ for $i = 1, 2$ and $0 < b_i$.

   (a) Write an analytic expression for each density, that is, normalize each function for arbitrary a_i and positive b_i.

   (b) Calculate the likelihood ratio as a function of your four variables.

   (c) Sketch a graph of the likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the case $a_1 = 0$, $b_1 = 1$, $a_2 = 1$ and $b_2 = 2$.

Section 2.3

3. Consider minimax criterion for the zero-one loss function, that is, $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$.

   (a) Prove that in this case the decision regions will satisfy

$$\int\limits_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\, d\mathbf{x} = \int\limits_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\, d\mathbf{x}.$$

   (b) Is this solution always unique? If not, construct a simple counterexample.

4. Consider the minimax criterion for a two-category classification problem.

   (a) Fill in the steps of the derivation of Eq. 23.

   (b) Explain why the overall Bayes risk must be concave down as a function of the prior $P(\omega_1)$, as shown in Fig. 2.4.

   (c) Assume we have one-dimensional Gaussian distributions $p(x|\omega_i) \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$, but completely unknown prior probabilities. Use the minimax criterion to find the optimal decision point x^* in terms of μ_i and σ_i under a zero-one risk.

(d) For the decision point x^* you found in (c), what is the overall minimax risk? Express this risk in terms of an error function erf$(\cdot)$.

(e) Assume $p(x|\omega_1) \sim N(0,1)$ and $p(x|\omega_2) \sim N(1/2, 1/4)$, under a zero-one loss. Find x^* and the overall minimax loss.

(f) Assume $p(x|\omega_1) \sim N(5,1)$ and $p(x|\omega_2) \sim N(6,1)$. Without performing any explicit calculations, determine x^* for the minimax criterion. Explain your reasoning.

5. Generalize the minimax decision rule in order to classify patterns from three categories having triangle densities as follows:

$$p(x|\omega_i) = T(\mu_i, \delta_i) \equiv \begin{cases} (\delta_i - |x-\mu_i|)/\delta_i^2 & \text{for } |x-\mu_i| < \delta_i \\ 0 & \text{otherwise,} \end{cases}$$

where $\delta_i > 0$ is the half-width of a distribution ($i = 1, 2, 3$). Assume for convenience that $\mu_1 < \mu_2 < \mu_3$, and make some minor simplifying assumptions about the δ_i's as needed, to answer the following:

(a) In terms of the priors $P(\omega_i)$, means and half-widths, find the optimal decision points x_1^* and x_2^* under a zero-one (categorization) loss.

(b) Generalize the minimax decision rule to *two* decision points, x_1^* and x_2^*, for such triangular distributions.

(c) Let $\{\mu_i, \delta_i\} = \{0, 1\}, \{.5, .5\}$, and $\{1, 1\}$. Find the minimax decision rule (i.e., x_1^* and x_2^*) for this case.

(d) What is the minimax risk for part (c)?

6. Consider the Neyman-Pearson criterion for two univariate normal distributions: $p(x|\omega_i) \sim N(\mu_i, \sigma_i^2)$ and $P(\omega_i) = 1/2$ for $i = 1, 2$. Assume a zero-one error loss, and for convenience let $\mu_2 > \mu_1$.

(a) Suppose the maximum acceptable error rate for classifying a pattern that is actually in ω_1 as if it were in ω_2 is E_1. Determine the single-point decision boundary in terms of the variables given.

(b) For this boundary, what is the error rate for classifying ω_2 as ω_1?

(c) What is the overall error rate under zero-one loss?

(d) Apply your results to the specific case $p(x|\omega_1) \sim N(-1, 1)$ and $p(x|\omega_2) \sim N(1, 1)$ and $E_1 = 0.05$.

(e) Compare your result to the Bayes error rate (i.e., without the Neyman-Pearson conditions).

7. Consider Neyman-Pearson criteria for two Cauchy distributions in one dimension:

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

Assume a zero-one error loss, and for simplicity $a_2 > a_1$, the same "width" b, and equal priors.

(a) Suppose the maximum acceptable error rate for classifying a pattern that is actually in ω_1 as if it were in ω_2 is E_1. Determine the single-point decision boundary in terms of the variables given.

(b) For this boundary, what is the error rate for classifying ω_2 as ω_1?

(c) What is the overall error rate under zero-one loss?

(d) Apply your results to the specific case $b = 1$ and $a_1 = -1$, $a_2 = 1$ and $E_1 = 0.1$.

(e) Compare your result to the Bayes error rate (i.e., without the Neyman-Pearson conditions).

8. Let the conditional densities for a two-category one-dimensional problem be given by the Cauchy distribution described in Problem 7.

(a) By explicit integration, check that the distributions are indeed normalized.

(b) Assuming $P(\omega_1) = P(\omega_2)$, show that $P(\omega_1|x) = P(\omega_2|x)$ if $x = (a_1 + a_2)/2$, that is, the minimum error decision boundary is a point midway between the peaks of the two distributions, regardless of b.

(c) Plot $P(\omega_1|x)$ for the case $a_1 = 3$, $a_2 = 5$ and $b = 1$.

(d) How do $P(\omega_1|x)$ and $P(\omega_2|x)$ behave as $x \to -\infty$? $x \to +\infty$? Explain.

9. Use the conditional densities given in Problem 7, and assume equal prior probabilities for the categories.

(a) Show that the minimum probability of error is given by

$$P(error) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_2 - a_1}{2b}\right|.$$

(b) Plot this as a function of $|a_2 - a_1|/b$.

(c) What is the maximum value of $P(error)$ and under which conditions can this occur? Explain.

10. Consider the following decision rule for a two-category one-dimensional problem: Decide ω_1 if $x > \theta$; otherwise decide ω_2.

(a) Show that the probability of error for this rule is given by

$$P(error) = P(\omega_1)\int_{-\infty}^{\theta} p(x|\omega_1)\,dx + P(\omega_2)\int_{\theta}^{\infty} p(x|\omega_2)\,dx.$$

(b) By differentiating, show that a necessary condition to minimize $P(error)$ is that θ satisfies

$$p(\theta|\omega_1)P(\omega_1) = p(\theta|\omega_2)P(\omega_2).$$

(c) Does this equation define θ uniquely?

(d) Give an example where a value of θ satisfying the equation actually *maximizes* the probability of error.

11. Suppose that we replace the deterministic decision function $\alpha(\mathbf{x})$ with a *randomized rule*, namely, one giving the probability $P(\alpha_i|\mathbf{x})$ of taking action α_i upon observing $\mathbf{x}$.

(a) Show that the resulting risk is given by

$$R = \int \left[\sum_{i=1}^{a} R(\alpha_i|\mathbf{x}) P(\alpha_i|\mathbf{x}) \right] p(\mathbf{x})\, d\mathbf{x}.$$

(b) In addition, show that R is minimized by choosing $P(\alpha_i|\mathbf{x}) = 1$ for the action α_i associated with the minimum conditional risk $R(\alpha_i|\mathbf{x})$, thereby showing that no benefit can be gained from randomizing the best decision rule.

(c) Can we benefit from randomizing a suboptimal rule? Explain.

12. Let $\omega_{max}(\mathbf{x})$ be the state of nature for which $P(\omega_{max}|\mathbf{x}) \geq P(\omega_i|\mathbf{x})$ for all i, $i = 1, \ldots, c$.

(a) Show that $P(\omega_{max}|\mathbf{x}) \geq 1/c$.

(b) Show that for the minimum-error-rate decision rule the average probability of error is given by

$$P(error) = 1 - \int P(\omega_{max}|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}.$$

(c) Use these two results to show that $P(error) \leq (c-1)/c$.

(d) Describe a situation for which $P(error) = (c-1)/c$.

Section 2.4

13. In many pattern classification problems one has the option either to assign the pattern to one of c classes, or to *reject* it as being unrecognizable. If the cost for rejects is not too high, rejection may be a desirable action. Let

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \quad i, j = 1, \ldots, c \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise,} \end{cases}$$

where λ_r is the loss incurred for choosing the $(c+1)$th action, rejection, and λ_s is the loss incurred for making any substitution error. Show that the minimum risk is obtained if we decide ω_i if $P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x})$ for all j and if $P(\omega_i|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$, and reject otherwise. What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?

14. Consider the classification problem with rejection option.

(a) Use the results of Problem 13 to show that the following discriminant functions are optimal for such problems:

$$g_i(\mathbf{x}) = \begin{cases} p(\mathbf{x}|\omega_i)P(\omega_i) & i = 1, \ldots, c \\ \frac{\lambda_s - \lambda_r}{\lambda_s} \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j) & i = c+1. \end{cases}$$

(b) Plot these discriminant functions and the decision regions for the two-category one-dimensional case having

- $p(x|\omega_1) \sim N(1, 1)$,
- $p(x|\omega_2) \sim N(-1, 1)$,
- $P(\omega_1) = P(\omega_2) = 1/2$, and
- $\lambda_r/\lambda_s = 1/4$.

(c) Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1.

(d) Repeat for the case having

- $p(x|\omega_1) \sim N(1, 1)$,
- $p(x|\omega_2) \sim N(0, 1/4)$,
- $P(\omega_1) = 1/3$, $P(\omega_2) = 2/3$, and
- $\lambda_r/\lambda_s = 1/2$.

Section 2.5

15. Confirm Eq. 47 for the volume of a d-dimensional hypersphere as follows:

(a) Verify that the equation is correct for a line segment ($d = 1$).

(b) Verify that the equation is correct for a disk ($d = 2$).

(c) Integrate the volume of a line over appropriate limits to obtain the volume of a disk.

(d) Consider a general d-dimensional hypersphere. Integrate its volume to obtain a formula (involving the ratio of gamma functions, $\Gamma(\cdot)$) for the volume of a $(d + 1)$-dimensional hypersphere.

(e) Apply your formula to find the volume of a hypersphere in an odd-dimensional space by integrating the volume of a hypersphere in the lower even-dimensional space, and thereby confirm Eq. 47 for odd dimensions.

(f) Repeat the above but for finding the volume of a hypersphere in even dimensions.

16. Derive the formula for the volume of a d-dimensional hypersphere in Eq. 47 as follows:

(a) State by inspection the formula for V_1.

(b) Follow the general procedure outlined in Problem 15 and integrate twice to find V_{d+2} as a function of V_d.

(c) Assume that the functional form of V_d is the same for all odd dimensions (and likewise for all even dimensions). Use your integration results to determine the formula for V_d for d odd.

(d) Use your intermediate integration results to determine V_d for d even.

(e) Explain why we should expect the functional form of V_d to be different in even and in odd dimensions.

17. Derive the formula (Eq. 46) for the volume V of a hyperellipsoid of constant Mahalanobis distance r (Eq. 45) for a Gaussian distribution having covariance $\mathbf{\Sigma}$.

18. Consider two normal distributions in one dimension: $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Imagine that we choose two random samples x_1 and x_2, one from each of the normal distributions and calculate their sum $x_3 = x_1 + x_2$. Suppose we do this repeatedly.

(a) Consider the resulting distribution of the values of x_3. Show that x_3 possesses the requisite statistical properties and thus its distribution is normal.

(b) What is the mean, μ_3, of your new distribution?

(c) What is the variance, σ_3^2?

(d) Repeat the above with two distributions in a multi-dimensional space, i.e., $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

19. Starting from the definition of entropy (Eq. 37), derive the general equation for the maximum-entropy distribution given constraints expressed in the general form

$$\int b_k(x)p(x)\,dx = a_k, \quad k = 1, 2, \ldots, q$$

as follows:

(a) Use Lagrange undetermined multipliers $\lambda_1, \lambda_2, \ldots, \lambda_q$ and derive the synthetic function:

$$H_s = -\int p(x)\left[\ln p(x) - \sum_{k=0}^{q} \lambda_k b_k(x)\right] dx - \sum_{k=0}^{q} \lambda_k a_k.$$

State why we know $a_0 = 1$ and $b_0(x) = 1$ for all x.

(b) Take the derivative of H_s with respect to $p(x)$. Set the integrand to zero, and thereby prove that the minimum-entropy distribution obeys

$$p(x) = \exp\left[\sum_{k=0}^{q} \lambda_k b_k(x) - 1\right],$$

where the $q + 1$ parameters are determined by the constraint equation.

20. Use the final result from Problem 19 for the following.

(a) Suppose we know solely that a distribution is nonzero only in the range $x_l \leq x \leq x_u$. Prove that the maximum entropy distribution is uniform in that range, that is,

$$p(x) \sim U(x_l, x_u) = \begin{cases} 1/|x_u - x_l| & x_l \leq x \leq x_u \\ 0 & \text{otherwise.} \end{cases}$$

(b) Suppose we know solely that a distribution is nonzero only for $x \geq 0$ and that its mean is μ. Prove that the maximum entropy distribution is

$$p(x) = \begin{cases} \frac{1}{\mu}e^{-x/\mu} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(c) Now suppose we know solely that the distribution is normalized, has mean μ, and standard deviation σ^2, and thus from Problem 19 our maximum entropy distribution must be of the form

$$p(x) = \exp[\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2].$$

Write out the three constraints and solve for λ_0, λ_1, and λ_2 and thereby prove that the maximum entropy solution is a Gaussian, that is,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right].$$

21. Three distributions—a Gaussian, a uniform distribution, and a triangle distribution (cf. Problem 5)—each have mean zero and standard deviation σ^2. Use Eq. 37 to calculate and compare their entropies.

22. Calculate the entropy of a multidimensional Gaussian $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

23. Consider the three-dimensional normal distribution $p(\mathbf{x}|\omega) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix}.$$

(a) Find the probability density at the point $\mathbf{x}_0 = (.5, 0, 1)^t$.

(b) Construct the whitening transformation $\mathbf{A}_w$ (Eq. 44). Compute the matrices representing eigenvectors and eignvalues, $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$. Next, convert the distribution to one centered on the origin with covariance matrix equal to the identity matrix, $p(\mathbf{x}|\omega) \sim N(\mathbf{0}, \mathbf{I})$.

(c) Apply the same overall transformation to $\mathbf{x}_0$ to yield a transformed point $\mathbf{x}_w$.

(d) By explicit calculation, confirm that the Mahalanobis distance from $\mathbf{x}_0$ to the mean $\boldsymbol{\mu}$ in the original distribution is the same as for $\mathbf{x}_w$ to $\mathbf{0}$ in the transformed distribution.

(e) Does the probability density remain unchanged under a general linear transformation? In other words, is $p(\mathbf{x}_0|N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = p(\mathbf{T}^t\mathbf{x}_0|N(\mathbf{T}^t\boldsymbol{\mu}, \mathbf{T}^t\boldsymbol{\Sigma}\mathbf{T}))$ for some linear transform $\mathbf{T}$? Explain.

(f) Prove that a general whitening transform $\mathbf{A}_w = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$ when applied to a Gaussian distribution ensures that the final distribution has covariance proportional to the identity matrix $\mathbf{I}$. Check whether normalization is preserved by the transformation.

24. Consider the multivariate normal density with mean $\boldsymbol{\mu}$, $\sigma_{ij} = 0$ and $\sigma_{ii} = \sigma_i^2$, that is, the covariance matrix is diagonal: $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$.

(a) Show that the evidence is

$$p(\mathbf{x}) = \frac{1}{\prod\limits_{i=1}^{d}\sqrt{2\pi}\sigma_i}\exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right].$$

(b) Plot and describe the contours of constant density.

(c) Write an expression for the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$.

Section 2.6

25. Fill in the steps in the derivation from Eq. 59 to Eqs. 60–65.

26. Let $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1, 2$ in a two-category d-dimensional problem with the same covariances but arbitrary means and prior probabilities. Consider the squared Mahalanobis distance

$$r_i^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i).$$

(a) Show that the gradient of r_i^2 is given by

$$\boldsymbol{\nabla} r_i^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i).$$

(b) Show that at any position on a given line through $\boldsymbol{\mu}_i$ the gradient $\boldsymbol{\nabla} r_i^2$ points in the same direction. Must this direction be parallel to that line?

(c) Show that $\boldsymbol{\nabla} r_1^2$ and $\boldsymbol{\nabla} r_2^2$ point in opposite directions along the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$.

(d) Show that the optimal separating hyperplane is tangent to the constant probability density hyperellipsoids at the point that the separating hyperplane cuts the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$.

(e) True or False: For a two-category problem involving normal densities with arbitrary means and covariances, and $P(\omega_1) = P(\omega_2) = 1/2$, the Bayes decision boundary consists of the set of points of equal Mahalanobis distance from the respective sample means. Explain.

27. Suppose we have two normal distributions with the same covariances but different means: $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. In terms of their prior probabilities $P(\omega_1)$ and $P(\omega_2)$, state the condition that the Bayes decision boundary *not* pass between the two means.

28. Two random variables $\mathbf{x}$ and $\mathbf{y}$ are called statistically independent if $p(\mathbf{x}, \mathbf{y}|\omega) = p(\mathbf{x}|\omega)p(\mathbf{y}|\omega)$.

(a) Prove that if $x_i - \mu_i$ and $x_j - \mu_j$ are statistically independent (for $i \neq j$), then σ_{ij} as defined in Eq. 43 is 0.

(b) Prove that the converse is true for the Gaussian case.

(c) Show by counterexample that this converse is *not* true in the general case.

29. Figure 2.15 shows that it is possible for a decision boundary for two three-dimensional Gaussians to be a line segment. Explain how this can arise by analyzing a simpler one-dimensional case as follows.

(a) Consider two one-dimensional Gaussians whose means differ and whose variances differ. Explain why for this case we can always find prior probabilities such that the decision boundary is a single point.

(b) Use your result to explain how the three-dimensional two-Gaussian case can yield a line segment decision boundary.

30. Consider the Bayes decision boundary for two-category classification in d dimensions.

(a) Prove that for any arbitrary hyperquadric in d dimensions, there exist normal distributions $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and priors $P(\omega_i)$, $i = 1, 2$, that possess this hyperquadric as their Bayes decision boundary.

(b) Is your answer to part (a) true if the priors are held fixed and nonzero, e.g., $P(\omega_1) = P(\omega_2) = 1/2$?

Section 2.7

31. Let $p(x|\omega_i) \sim N(\mu_i, \sigma^2)$ for a two-category one-dimensional problem with $P(\omega_1) = P(\omega_2) = 1/2$.

(a) Show that the minimum probability of error is given by

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2}\, du,$$

where $a = |\mu_2 - \mu_1|/(2\sigma)$.

(b) Use the inequality

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-t^2/2}\, dt \le \frac{1}{\sqrt{2\pi}a} e^{-a^2/2}$$

to show that P_e goes to zero as $|\mu_2 - \mu_1|/\sigma$ goes to infinity.

32. Let $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \sigma^2\mathbf{I})$ for a two-category d-dimensional problem with $P(\omega_1) = P(\omega_2) = 1/2$.

(a) Show that the minimum probability of error is given by

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2}\, du,$$

where $a = \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|/(2\sigma)$.

(b) Let $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = (\mu_1, \ldots, \mu_d)^t \neq \mathbf{0}$. Use the inequality from Problem 31 to show that P_e approaches zero as the dimension d approaches infinity.

(c) Express the meaning of this result in words.

33. Suppose we know exactly two arbitrary distributions $p(\mathbf{x}|\omega_i)$ and priors $P(\omega_i)$ in a d-dimensional feature space.

(a) Prove that the true error cannot decrease if we first project the distributions to a lower-dimensional space and then classify them.

(b) Despite this fact, suggest why in an actual pattern recognition application we might not want to include an arbitrarily high number of feature dimensions.

Section 2.8

34. Show that if the densities in a two-category classification problem differ significantly from Gaussian, the Chernoff and Bhattacharyya bounds are not likely to be informative by considering the following one-dimensional examples. Consider a number of problems in which the mean and variance are the same (and thus the Chernoff bound and the Bhattacharyya bound remain the same), but nevertheless have a wide range in Bayes error. For definiteness, assume $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions have means at $\mu_1 = -\mu$ and $\mu_2 = +\mu$, and $\sigma_1^2 = \sigma_2^2 = \mu^2$.

(a) Use the equations in the text to calculate the Chernoff and the Bhattacharyya bounds on the error.

(b) Suppose the distributions are both Gaussian. Calculate explicitly the Bayes error. Express it in terms of an error function erf$(\cdot)$ and as a numerical value.

(c) Now consider another case, in which half the density for ω_1 is concentrated at a point $x = -2\mu$ and half at $x = 0$; likewise (symmetrically) the density for ω_2 has half its mass at $x = +2\mu$ and half at $x = 0$. Show that the means and variance remain as desired, but that now the Bayes error is 0.25.

(d) Now consider yet another case, in which half the density for ω_1 is concentrated near $x = -2$ and half at $x = -\epsilon$, where ϵ is an infinitessimally small positive distance; likewise (symmetrically) the density for ω_2 has half its mass near $x = +2\mu$ and half at $+\epsilon$. Show that by making ϵ sufficiently small, the means and variances can be made arbitrarily close to μ and μ^2, respectively. Show, too, that now the Bayes error is zero.

(e) Compare your errors in (b), (c), and (d) to your Chernoff and Bhattacharyya bounds of (a) and explain in words why those bounds are unlikely to be of much use if the distributions differ markedly from Gaussians.

35. Show for nonpathological cases that if we include more feature dimensions in a Bayesian classifier for multidimensional Gaussian distributions then the Bhattacharyya bound decreases. Do this as follows: Let $P_d(P(\omega_1), \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, P(\omega_2), \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, or simply P_d, be the Bhattacharyya bound if we consider the distributions restricted to d dimensions.

(a) Using general properties of a covariance matrix, prove that $k(1/2)$ of Eq. 77 must increase as we increase from d to $d+1$ dimensions, and hence the error bound must decrease.

(b) Explain why this general result does or does not depend upon *which* dimension is added.

(c) What is a "pathological" case in which the error bound does *not* decrease, that is, for which $P_{d+1} = P_d$?

(d) Is it ever possible that the *true* error—that is, not just the *bound*—could *increase* as we go to higher dimension?

(e) Prove that as $d \to \infty$, $P_d \to 0$ for nonpathological distributions. Describe pathological distributions for which this infinite limit does not hold.

(f) Given that the Bhattacharyya bound decreases for the inclusion of a particular dimension, does this guarantee that the *true* error will decrease? Explain.

36. Derive Eqs. 74 and 75 from Eq. 73 by the following steps:

(a) Substitute the normal distributions into the integral and gather the terms dependent upon **x** and those that are not dependent upon **x**.

(b) Factor the term independent of **x** from the integral.

(c) Integrate explicitly the term dependent upon **x**.

37. Consider a two-category classification problem in two dimensions with

$$p(\mathbf{x}|\omega_1) \sim N(\mathbf{0}, \mathbf{I}),\ p(\mathbf{x}|\omega_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \mathbf{I}\right),\ \text{and } P(\omega_1) = P(\omega_2) = 1/2.$$

(a) Calculate the Bayes decision boundary.

(b) Calculate the Bhattacharyya error bound.

(c) Repeat the above for the same prior probabilities, but

$$p(\mathbf{x}|\omega_1) \sim N\left(\mathbf{0}, \begin{pmatrix}2 & .5\\ .5 & 2\end{pmatrix}\right) \text{ and } p(\mathbf{x}|\omega_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}5 & 4\\ 4 & 5\end{pmatrix}\right).$$

38. Derive the Bhattacharyya error bound without the need for first examining the Chernoff bound. Do this as follows:

(a) If a and b are nonnegative numbers, show directly that min[a,b] $\leq \sqrt{ab}$.

(b) Use this to show that the error rate for a two-category Bayes classifier must satisfy

$$P(error) \leq \sqrt{P(\omega_1)P(\omega_2)}\ \rho \leq \rho/2,$$

where ρ is the so-called *Bhattacharyya coefficient*

$$\rho = \int \sqrt{p(\mathbf{x}|\omega_1)\ p(\mathbf{x}|\omega_2)}\ d\mathbf{x}.$$

39. Use signal detection theory, as well as the notation and basic Gaussian assumptions described in the text, to address the following.

(a) Prove that $P(x > x^*|x \in \omega_2)$ and $P(x > x^*|x \in \omega_1)$, taken together, uniquely determine the discriminability d'.

(b) Use error functions erf$(\cdot)$ to express d' in terms of the hit and false alarm rates. Estimate d' if $P(x > x^*|x \in \omega_1) = 0.8$ and $P(x > x^*|x \in \omega_2) = 0.3$. Repeat for $P(x > x^*|x \in \omega_1) = 0.7$ and $P(x > x^*|x \in \omega_1) = 0.4$.

(c) Given that the Gaussian assumption is valid, calculate the Bayes error for both the cases in (b).

(d) Using a trivial one-line computation determine which case has the higher d':

Case A: $P(x > x^*|x \in \omega_1) = 0.8$, $P(x > x^*|x \in \omega_2) = 0.3$ or

Case B: $P(x > x^*|x \in \omega_1) = 0.3$, $P(x > x^*|x \in \omega_2) = 0.7$.

Explain your logic.

40. Suppose in our signal detection framework we had two Gaussians, but with different variances (cf. Fig. 2.20)—that is, $p(x|\omega_1) \sim N(\mu_1, \sigma_1^2)$ and $p(x|\omega_2) \sim N(\mu_2, \sigma_2^2)$ for $\mu_2 > \mu_1$ and $\sigma_2^2 \neq \sigma_1^2$. In that case the resulting ROC curve would no longer be symmetric.

(a) Suppose in this asymmetric case we modified the definition of the discriminability to be $d'_a = |\mu_2 - \mu_1|/\sqrt{\sigma_1\sigma_2}$. Show by nontrivial counterexample or analysis that one cannot determine d'_a uniquely based on a single pair of hit and false alarm rates.

(b) Assume we measure the hit and false alarm rates for two different, but unknown, values of the threshold x^*. Derive a formula for d'_a based on such measurements.

(c) State and explain all pathological values for which your formula does not give a meaningful value for d'_a.

(d) Plot several ROC curves for the case $p(x|\omega_1) \sim N(0, 1)$ and $p(x|\omega_2) \sim N(1, 2)$.

41. Consider two one-dimensional triangle distributions having different means, but the same width:

$$p(x|\omega_i) = T(\mu_i, \delta) = \begin{cases} (\delta - |x - \mu_i|)/\delta^2 & \text{for } |x - \mu_i| < \delta \\ 0 & \text{otherwise,} \end{cases}$$

with $\mu_2 > \mu_1$. We define a new discriminability here as $d'_T = (\mu_2 - \mu_1)/\delta$.

(a) Write an analytic function, parameterized by d'_T, for the operating characteristic curves.

(b) Plot these novel operating characteristic curves for $d'_T = \{.1, .2, \ldots, 1.0\}$. Interpret your answer for the case $d'_T = 1.0$ and 2.0.

(c) Suppose we measure $P(x > x^*|x \in \omega_2) = 0.4$ and $P(x > x^*|x \in \omega_1) = 0.7$. What is d'_T? What is the Bayes error rate?

(d) Infer the decision rule employed in part (c). That is, express x^* in terms of the variables given in the problem.

(e) Suppose we measure $P(x > x^*|x \in \omega_2) = 0.3$ and $(x > x^*|x \in \omega_1) = 0.9$. What is d'_T? What is the Bayes error rate?

(f) Infer the decision rule employed in part (e). That is, express x^* in terms of the variables given in the problem.

42. Equation 72 can be used to obtain an upper bound on the error. One can also derive tighter analytic bounds in the two-category case—both upper and lower bounds—analogous to Eq. 73 for general distributions. If we let $p \equiv p(x|\omega_1)$, then we seek tighter bounds on $\min[p, 1-p]$ (which has discontinuous derivative).

(a) Prove that

$$b_L(p) = \frac{1}{\beta}\ln\left[\frac{1+e^{-\beta}}{e^{-\beta p} + e^{-\beta(1-p)}}\right]$$

for any $\beta > 0$ is a lower bound on $\min[p, 1-p]$.

(b) Prove that one can choose β in (a) to give an arbitrarily tight lower bound.

(c) Repeat (a) and (b) for the upper bound given by

$$b_U(p) = b_L(p) + [1 - 2b_L(0.5)]b_G(p)$$

where $b_G(p)$ is any upper bound that obeys

$$\begin{aligned} b_G(p) &\geq \min[p, 1-p] \\ b_G(p) &= b_G(1-p) \\ b_G(0) &= b_G(1) = 0 \\ b_G(0.5) &= 0.5. \end{aligned}$$

(d) Confirm that $b_G(p) = 1/2\sin[\pi p]$ obeys the conditions in (c).

(e) Let $b_G(p) = 1/2\sin[\pi p]$, and plot your upper and lower bounds as a function of p, for $0 \leq p \leq 1$ and $\beta = 1, 10$, and 50.

Section 2.9

43. Let the components of the vector $\mathbf{x} = (x_1, \ldots, x_d)^t$ be binary-valued (0 or 1), and let $P(\omega_j)$ be the prior probability for the state of nature ω_j and $j = 1, \ldots, c$. Now define

$$p_{ij} = \Pr[x_i = 1|\omega_j] \qquad \begin{matrix} i = 1, \ldots, d \\ j = 1, \ldots, c, \end{matrix}$$

with the components of x_i being statistically independent for all $\mathbf{x}$ in ω_j.

(a) Interpret in words the meaning of p_{ij}.

(b) Show that the minimum probability of error is achieved by the following decision rule: Decide ω_k if $g_k(\mathbf{x}) \geq g_j(\mathbf{x})$ for all j and k, where

$$g_j(\mathbf{x}) = \sum_{i=1}^{d} x_i \ln \frac{p_{ij}}{1 - p_{ij}} + \sum_{i=1}^{d} \ln(1 - p_{ij}) + \ln P(\omega_j).$$

44. Let the components of the vector $\mathbf{x} = (x_1, \ldots, x_d)^t$ be ternary valued (1, 0 or -1), with

$$\begin{aligned} p_{ij} &= \Pr[x_i = \ 1\ |\omega_j] \\ q_{ij} &= \Pr[x_i = \ 0\ |\omega_j] \\ r_{ij} &= \Pr[x_i = -1|\omega_j], \end{aligned}$$

and with the components of x_i being statistically independent for all $\mathbf{x}$ in ω_j.

(a) Show that a minimum probability of error decision rule can be derived that involves discriminant functions $g_j(\mathbf{x})$ that are quadratic function of the components x_i.

(b) Suggest a generalization to more categories of your answers to this and Problem 43.

45. Let $\mathbf{x}$ be distributed as in Problem 43 with $c = 2$, d odd, and

$$\begin{aligned} p_{i1} &= p > 1/2 \qquad i = 1, \ldots, d \\ p_{i2} &= 1 - p \qquad\quad i = 1, \ldots, d, \end{aligned}$$

and $P(\omega_1) = P(\omega_2) = 1/2$.

(a) Show that the minimum-error-rate decision rule becomes

$$\text{Decide } \omega_1 \text{ if } \sum_{i=1}^{d} x_i > d/2 \text{ and } \omega_2 \text{ otherwise.}$$

(b) Show that the minimum probability of error is given by

$$P_e(d, p) = \sum_{k=0}^{(d-1)/2} \binom{d}{k} p^k (1 - p)^{d-k}.$$

where $\binom{d}{k} = d!/(k!(d-k)!)$ is the binomial coefficient.

(c) What is the limiting value of $P_e(d, p)$ as $p \to 1/2$? Explain.

(d) Show that $P_e(d, p)$ approaches zero as $d \to \infty$. Explain.

46. Under the natural assumption concerning losses, i.e., that $\lambda_{21} > \lambda_{11}$ and $\lambda_{12} > \lambda_{22}$, show that the general minimum risk discriminant function for the independent binary case described in Section 2.9.1 is given by $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$, where $\mathbf{w}$ is unchanged, and

$$w_0 = \sum_{i=1}^{d} \ln \frac{1 - p_i}{1 - q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)} + \ln \frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}}.$$

47. The Poisson distribution for a discrete variable $x = 0, 1, 2, \ldots$ and real parameter λ is

$$P(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

(a) Prove that the mean of such a distribution is $\mathcal{E}[x] = \lambda$.

(b) Prove that the variance of such a distribution is $\mathcal{E}[x - \bar{x}] = \lambda$.

(c) The *mode* of a distribution is the value of x that has the maximum probability. Prove that the mode of a Poisson distribution is the greatest integer that does not exceed λ. That is, prove that the mode is $\lfloor\lambda\rfloor$, read "floor of lambda." (If λ is an integer, then both λ and $\lambda - 1$ are modes.)

(d) Consider two equally probable categories having Poisson distributions but with differing parameters; assume for definiteness $\lambda_1 > \lambda_2$. What is the Bayes classification decision?

(e) What is the Bayes error rate?

Section 2.10

48. Suppose we have three categories in two dimensions with the following underlying distributions:

- $p(\mathbf{x}|\omega_1) \sim N(\mathbf{0}, \mathbf{I})$
- $p(\mathbf{x}|\omega_2) \sim N\left(\binom{1}{1}, \mathbf{I}\right)$
- $p(\mathbf{x}|\omega_3) \sim \frac{1}{2}N\left(\binom{.5}{.5}, \mathbf{I}\right) + \frac{1}{2}N\left(\binom{-.5}{.5}, \mathbf{I}\right)$

with $P(\omega_i) = 1/3, i = 1, 2, 3$.

(a) By explicit calculation of posterior probabilities, classify the point $\mathbf{x} = \binom{.3}{.3}$ for minimum probability of error.

(b) Suppose that for a particular test point the first feature is missing. That is, classify $\mathbf{x} = \binom{*}{.3}$.

(c) Suppose that for a particular test point the second feature is missing. That is, classify $\mathbf{x} = \binom{.3}{*}$.

(d) Repeat all of the above for $\mathbf{x} = \binom{.2}{.6}$.

49. Show that Eq. 95 reduces to Bayes rule when the true feature is $\boldsymbol{\mu}_i$ and $p(\mathbf{x}_b|\mathbf{x}_t) \sim N(\mathbf{x}_t, \boldsymbol{\Sigma})$. Interpret this answer in words.

Section 2.11

50. Use the conditional probability matrices in Example 4 to answer the following separate problems.

(a) Suppose it is December 20—the end of autumn and the beginning of winter—and thus let $P(a_1) = P(a_4) = 0.5$. Furthermore, it is known that the fish was caught in the north Atlantic, that is, $P(b_1) = 1$. Suppose the lightness has not been measured but it is known that the fish is thin, that is, $P(d_2) = 1$. Classify the fish as salmon or sea bass. What is the expected error rate?

(b) Suppose all we know is that a fish is thin and medium lightness. What season is it now, most likely? What is your probability of being correct?

(c) Suppose we know a fish is thin and medium lightness and that it was caught in the north Atlantic. What season is it, most likely? What is the probability of being correct?

51. Consider a Bayesian belief net with several nodes having unspecified values. Suppose that one such node is selected at random, with the probabilities of its nodes computed by the formulas described in the text. Next, another such node is chosen at random (possibly even a node already visited), and the probabilities are similarly updated. Prove that this procedure will converge to the desired probabilities throughout the full network.

Section 2.12

52. Suppose we have three categories with $P(\omega_1) = 1/2$, $P(\omega_2) = P(\omega_3) = 1/4$ and the following distributions

- $p(x|\omega_1) \sim N(0, 1)$
- $p(x|\omega_2) \sim N(.5, 1)$
- $p(x|\omega_3) \sim N(1, 1)$,

and that we sample the following four points: $x = 0.6, 0.1, 0.9, 1.1$.

(a) Calculate explicitly the probability that the sequence actually came from $\omega_1, \omega_3, \omega_3, \omega_2$. Be careful to consider normalization.

(b) Repeat for the sequence $\omega_1, \omega_2, \omega_2, \omega_3$.

(c) Find the sequence having the maximum probability.

COMPUTER EXERCISES

Several of the computer exercises will rely on the following data.

	ω_1			ω_2			ω_3		
sample	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
1	−5.01	−8.12	−3.68	−0.91	−0.18	−0.05	5.35	2.26	8.13
2	−5.43	−3.48	−3.54	1.30	−2.06	−3.53	5.12	3.22	−2.66
3	1.08	−5.52	1.66	−7.75	−4.54	−0.95	−1.34	−5.31	−9.87
4	0.86	−3.78	−4.11	−5.47	0.50	3.92	4.48	3.42	5.19
5	−2.67	0.63	7.39	6.14	5.72	−4.85	7.11	2.39	9.21
6	4.94	3.29	2.08	3.60	1.26	4.36	7.17	4.33	−0.98
7	−2.51	2.09	−2.59	5.37	−4.63	−3.65	5.75	3.97	6.65
8	−2.25	−2.13	−6.94	7.18	1.46	−6.66	0.77	0.27	2.41
9	5.56	2.86	−2.26	−7.39	1.17	6.30	0.90	−0.43	−8.71
10	1.03	−3.33	4.33	−7.50	−6.32	−0.31	3.52	−0.36	6.43

Section 2.5

1. You may need the following procedures for several exercises below.

(a) Write a procedure to generate random samples according to a normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in d dimensions.

(b) Write a procedure to calculate the discriminant function (of the form given in Eq. 49) for a given normal distribution and prior probability $P(\omega_i)$.

(c) Write a procedure to calculate the Euclidean distance between two arbitrary points.

(d) Write a procedure to calculate the Mahalanobis distance between the mean $\boldsymbol{\mu}$ and an arbitrary point $\mathbf{x}$, given the covariance matrix $\boldsymbol{\Sigma}$.

2. Refer to Computer exercise 1 (b) and consider the problem of classifying 10 samples from the table above. Assume that the underlying distributions are normal.

(a) Assume that the prior probabilities for the first two categories are equal ($P(\omega_1) = P(\omega_2) = 1/2$ and $P(\omega_3) = 0$) and design a dichotomizer for those two categories using only the x_1 feature value.

(b) Determine the empirical training error on your samples, that is, the percentage of points misclassified.

(c) Use the Bhattacharyya bound to bound the error you will get on novel patterns drawn from the distributions.

(d) Repeat all of the above, but now use *two* feature values, x_1 and x_2.

(e) Repeat, but use all *three* feature values.

(f) Discuss your results. In particular, is it ever possible for a finite set of data that the empirical error might be *larger* for more data dimensions?

3. Repeat Computer exercise 2 but for categories ω_1 and ω_3.

4. Consider the three categories in Computer exercise 2, and assume $P(\omega_i) = 1/3$.

(a) What is the Mahalanobis distance between each of the following test points and each of the category means in Computer exercise 2: $(1, 2, 1)^t$, $(5, 3, 2)^t$, $(0, 0, 0)^t$, $(1, 0, 0)^t$.

(b) Classify those points.

(c) Assume instead that $P(\omega_1) = 0.8$, and $P(\omega_2) = P(\omega_3) = 0.1$ and classify the test points again.

5. Illustrate the fact that the average of a large number of independent random variables will approximate a Gaussian by the following:

(a) Write a program to generate n random integers from a uniform distribution $U(x_l, x_u)$. (Some computer systems include this as a single, compiled function call.)

(b) Now write a routine to choose x_l and x_u randomly, in the range $-100 \leq x_l < x_u \leq +100$, and n (the number of samples) randomly in the range $0 < n \leq 1000$.

(c) Generate and plot a histogram of the accumulation of 10^4 points sampled as just described.

(d) Calculate the mean and standard deviation of your histogram, and plot it

(e) Repeat the above for 10^5 and for 10^6. Discuss your results.

Section 2.8

6. Explore how the empirical error does or does not approach the Bhattacharyya bound as follows:

(a) Write a procedure to generate sample points in d dimensions with a normal distribution having mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

(b) Consider the normal distributions

$$p(\mathbf{x}|\omega_1) \sim N\left(\begin{pmatrix}1\\0\end{pmatrix}, \mathbf{I}\right) \text{ and } p(\mathbf{x}|\omega_2) \sim N\left(\begin{pmatrix}-1\\0\end{pmatrix}, \mathbf{I}\right)$$

with $P(\omega_1) = P(\omega_2) = 1/2$. By inspection, state the Bayes decision boundary.

(c) Generate $n = 100$ points (50 for ω_1 and 50 for ω_2) and calculate the empirical error.

(d) Repeat for increasing values of n, $100 \leq n \leq 1000$, in steps of 100 and plot your empirical error.

(e) Discuss your results. In particular, is it ever possible that the empirical error is greater than the Bhattacharyya or Chernoff bound?

7. Consider two one-dimensional normal distributions $p(x|\omega_1) \sim N(-.5, 1)$ and $p(x|\omega_2) \sim N(+.5, 1)$ and $P(\omega_1) = P(\omega_2) = 0.5$.

(a) Calculate the Bhattacharyya bound for the error of a Bayesian classifier.

(b) Express the true error rate in terms of an error function, erf($\cdot$).

(c) Evaluate this true error to four significant figures by numerical integration (or other routine).

(d) Generate 10 points each for the two categories and determine the empirical error using your Bayesian classifier. (You should recalculate the decision boundary for each of your data sets.)

(e) Plot the empirical error as a function of the number of points from either distribution by repeating the previous part for 50, 100, 200, 500 and 1000 sample points from each distribution. Compare your asymptotic empirical error to the true error and the Bhattacharyya error bound.

8. Repeat Computer exercise 7 with the following conditions:

(a) $p(x|\omega_1) \sim N(-.5, 2)$ and $p(x|\omega_2) \sim N(.5, 2)$, $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$.

(b) $p(x|\omega_1) \sim N(-.5, 2)$ and $p(x|\omega_2) \sim N(.5, 2)$ and $P(\omega_1) = P(\omega_2) = 1/2$.

(c) $p(x|\omega_1) \sim N(-.5, 3)$ and $p(x|\omega_2) \sim N(.5, 1)$ and $P(\omega_1) = P(\omega_2) = 1/2$.

Section 2.11

9. Write a program to evaluate the Bayesian belief net for fish in Example 3, including the information in $P(x_i|a_j)$, $P(x_i|b_j)$, $P(c_i|x_j)$, and $P(d_i|x_j)$. Test your program on the calculation given in the Example. Apply your program to the following cases, and state any assumptions you need to make.

(a) A dark, thin fish is caught in the north Atlantic in summer. What is the probability it is a salmon?

(b) A thin, medium fish is caught in the north Atlantic. What is the probability it is winter? spring? summer? autumn?

(c) A light, wide fish is caught in the autumn. What is the probability it came from the north Atlantic?