Chapter 2

Bayesian decision theory

Problem Solutions

Section 2.1

1. Equation 7 in the text states

$$P(error|x) = \min[P(\omega_1|x), P(\omega_2|x)].$$

(a) We assume, without loss of generality, that for a given particular x we have $P(\omega_2|x) \geq P(\omega_1|x)$, and thus $P(error|x) = P(\omega_1|x)$. We have, moreover, the normalization condition $P(\omega_1|x) = 1 - P(\omega_2|x)$. Together these imply $P(\omega_2|x) > 1/2$ or $2P(\omega_2|x) > 1$ and

$$2P(\omega_2|x)P(\omega_1|x) > P(\omega_1|x) = P(error|x).$$

This is true at every x, and hence the integrals obey

$$\int 2P(\omega_2|x)P(\omega_1|x)dx \geq \int P(error|x)dx.$$

In short, $2P(\omega_2|x)P(\omega_1|x)$ provides an upper bound for $P(error|x)$.

(b) From part (a), we have that $P(\omega_2|x) > 1/2$, but in the current conditions not greater than $1/\alpha$ for $\alpha < 2$. Take as an example, $\alpha = 4/3$ and $P(\omega_1|x) = 0.4$ and hence $P(\omega_2|x) = 0.6$. In this case, $P(error|x) = 0.4$. Moreover, we have

$$\alpha P(\omega_1|x)P(\omega_2|x) = 4/3 \times 0.6 \times 0.4 < P(error|x).$$

This does not provide an upper bound for all values of $P(\omega_1|x)$.

(c) Let $P(error|x) = P(\omega_1|x)$. In that case, for all x we have

$$\begin{aligned} P(\omega_2|x)P(\omega_1|x) &< P(\omega_1|x)P(error|x) \\ \int P(\omega_2|x)P(\omega_1|x)dx &< \int P(\omega_1|x)P(error|x)dx, \end{aligned}$$

and we have a lower bound.

Therefore we have

$$\begin{aligned} R_{ran} &= \int \left[\sum_{i=1}^{a} R(\alpha_i(\mathbf{x})|\mathbf{x}) P(\alpha_i|\mathbf{x}) \right] p(\mathbf{x}) d\mathbf{x} \\ &\geq \int R(\alpha_{max}|\mathbf{x}) \left[\sum_{i=1}^{a} P(\alpha_i|\mathbf{x}) \right] p(\mathbf{x}) d\mathbf{x} \\ &= \int R(\alpha_{max}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= R_B, \end{aligned}$$

the Bayes risk. Equality holds if and only if $P(\alpha_{max}(\mathbf{x})|\mathbf{x}) = 1$.

12. We first note the normalization condition

$$\sum_{i=1}^{c} P(\omega_i|\mathbf{x}) = 1 \text{ for all } \mathbf{x}.$$

(a) If $P(\omega_i|\mathbf{x}) = P(\omega_j|\mathbf{x})$ for all i and j, then $P(\omega_i|\mathbf{x}) = 1/c$ and hence $P(\omega_{max}|\mathbf{x}) = 1/c$. If one of the $P(\omega_i|\mathbf{x}) < 1/c$, then by our normalization condition we must have that $P(\omega_{max}|\mathbf{x}) > 1/c$.

(b) The probability of error is simply 1 minus the probability of being correct, i.e.,

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

(c) We simply substitute the limit from part (a) to get

$$\begin{aligned} P(error) &= 1 - \int \underbrace{P(\omega_{max}|\mathbf{x})}_{=g \geq 1/c} p(\mathbf{x})\, d\mathbf{x} \\ &= 1 - g \int p(\mathbf{x})\, d\mathbf{x} = 1 - g. \end{aligned}$$

Therefore, we have $P(error) \leq 1 - 1/c = (c-1)/c$.

(d) All categories have the same prior probability and each distribution has the same form, i.e., the distributions are indistinguisable.

13. If we choose the category ω_{max} that has the maximum posterior probability, our risk at a point $\mathbf{x}$ is:

$$\lambda_s \sum_{j \neq max} P(\omega_j|\mathbf{x}) = \lambda_s[1 - P(\omega_{max}|\mathbf{x})],$$

whereas if we reject, our risk is λ_r. If we choose a non-maximal category ω_k (where $k \neq max$), then our risk is

$$\lambda_s \sum_{j \neq k} P(\omega_j|\mathbf{x}) = \lambda_s[1 - P(\omega_k|\mathbf{x})] \geq \lambda_s[1 - P(\omega_{max}|\mathbf{x})].$$

This last inequality shows that we should never decide on a category other than the one that has the maximum posterior probability, as we know from our Bayes analysis.

Consequently, we should either choose ω_{max} or we should reject, depending upon which is smaller: $\lambda_s[1 - P(\omega_{max}|\mathbf{x})]$ or λ_r. We reject if $\lambda_r \leq \lambda_s[1 - P(\omega_{max}|\mathbf{x})]$, that is, if $P(\omega_{max}|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$.

Section 2.4

14. Consider the classification problem with rejection option.

(a) The minimum-risk decision rule is given by:

$$\begin{aligned} \text{Choose } \omega_i \text{ if } P(\omega_i|\mathbf{x}) &\geq P(\omega_j|\mathbf{x}), \text{for all } j \\ \text{and if } P(\omega_i|\mathbf{x}) &\geq 1 - \frac{\lambda_r}{\lambda_s}. \end{aligned}$$

This rule is equivalent to

$$\begin{aligned} \text{Choose } \omega_i \text{ if } p(\mathbf{x}|\omega_i)P(\omega_i) &\geq p(\mathbf{x}|\omega_j)P(\omega_j) \text{ for all } j \\ \text{and if } p(\mathbf{x}|\omega_i)P(\omega_i) &\geq \left(1 - \frac{\lambda_r}{\lambda_s}\right) p(\mathbf{x}), \end{aligned}$$

where by Bayes' formula

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}.$$

The optimal discriminant function for this problem is given by

$$\text{Choose } \omega_i \text{ if } g_i(\mathbf{x}) \geq g_j(\mathbf{x}) \text{ for all } i = 1, \ldots, c, \text{ and } j = 1, \ldots, c+1.$$

Thus the discriminant functions are:

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

(b) Consider the case $p(x|\omega_1) \sim N(1,1), p(x|\omega_2) \sim N(-1,1), P(\omega_1) = P(\omega_2) = 1/2$ and $\lambda_r/\lambda_s = 1/4$. In this case the discriminant functions in part (a) give

$$\begin{aligned} g_1(x) &= p(x|\omega_1)P(\omega_1) = \frac{1}{2}\frac{e^{-\frac{1}{2}(x-1)^2}}{\sqrt{2\pi}} \\ g_2(x) &= p(x|\omega_2)P(\omega_2) = \frac{1}{2}\frac{e^{-\frac{1}{2}(x+1)^2}}{\sqrt{2\pi}} \\ g_3(x) &= \left(1 - \frac{\lambda_r}{\lambda_s}\right)[p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)] \\ &= \left(1 - \frac{1}{4}\right)\left[\frac{1}{2}\frac{e^{-\frac{1}{2}(x-1)^2}}{\sqrt{2\pi}} + \frac{1}{2}\frac{e^{-\frac{1}{2}(x+1)^2}}{\sqrt{2\pi}}\right] \\ &= \frac{3}{8\sqrt{2\pi}}\left[e^{-\frac{1}{2}(x-1)^2} + e^{-\frac{1}{2}(x+1)^2}\right] = \frac{3}{4}[g_1(x) + g_2(x)]. \end{aligned}$$

as shown in the figure.

Under a general linear transformation $\mathbf{T}$, we have that $\mathbf{x}' = \mathbf{T}^t\mathbf{x}$. The transformed mean is

$$\boldsymbol{\mu}' = \sum_{k=1}^{n} \mathbf{x}'_k = \sum_{k=1}^{n} \mathbf{T}^t\mathbf{x}_k = \mathbf{T}^t \sum_{k=1}^{n} \mathbf{x}_k = \mathbf{T}^t\boldsymbol{\mu}.$$

Likewise, the transformed covariance matrix is

$$\begin{aligned}
\boldsymbol{\Sigma}' &= \sum_{k=1}^{n} (\mathbf{x}'_k - \boldsymbol{\mu}')(\mathbf{x}'_k - \boldsymbol{\mu}')^t \\
&= \mathbf{T}^t \left[\sum_{k=1}^{n} (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu}) \right] \mathbf{T} \\
&= \mathbf{T}^t\boldsymbol{\Sigma}\mathbf{T}.
\end{aligned}$$

We note that $|\boldsymbol{\Sigma}'| = |\mathbf{T}^t\boldsymbol{\Sigma}\mathbf{T}| = |\boldsymbol{\Sigma}|$, and thus

$$p(\mathbf{x}_o|N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = p(\mathbf{T}^t\mathbf{x}_o|N(\mathbf{T}^t\boldsymbol{\mu}, \mathbf{T}^t\boldsymbol{\Sigma}\mathbf{T})).$$

(f) Recall the definition of a whitening transformation given by Eq. 44 in the text: $\mathbf{A}_w = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$. In this case we have

$$\mathbf{y} = \mathbf{A}_w^t\mathbf{x} \sim N(\mathbf{A}_w^t\boldsymbol{\mu}, \mathbf{A}_w^t\boldsymbol{\Sigma}\mathbf{A}_w),$$

and this implies that

$$\begin{aligned}
\mathrm{Var}(\mathbf{y}) &= \mathbf{A}_w^t(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t\mathbf{A}_w \\
&= \mathbf{A}_w^t\boldsymbol{\Sigma}\mathbf{A} \\
&= (\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})^t\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^t(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}) \\
&= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Phi}^t\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^t\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2} \\
&= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} \\
&= \mathbf{I},
\end{aligned}$$

the dentity matrix.

24. Recall that the general multivariate normal density in d-dimensions is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right].$$

(a) Thus we have if $\sigma_{ij} = 0$ and $\sigma_{ii} = \sigma_i^2$, then

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2) \\
&= \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_d^2 \end{pmatrix}.
\end{aligned}$$

Thus the determinant and inverse are particularly simple:

$$\begin{aligned}
|\boldsymbol{\Sigma}| &= \prod_{i=1}^{d} \sigma_i^2, \\
\boldsymbol{\Sigma}^{-1} &= \mathrm{diag}(1/\sigma_1^2, \ldots, 1/\sigma_d^2).
\end{aligned}$$

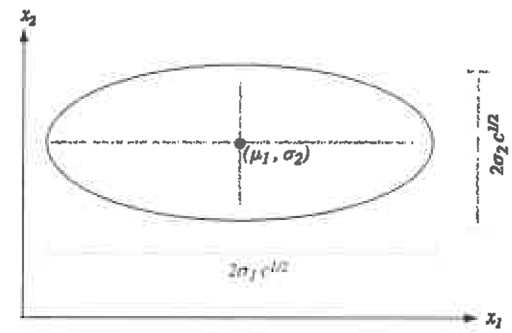
This leads to the density being expressed as:

$$\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathrm{diag}(1/\sigma_1^2, \ldots, 1/\sigma_d^2)(\mathbf{x}-\boldsymbol{\mu})\right] \\
&= \frac{1}{\prod_{i=1}^{d}\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right].
\end{aligned}$$

(b) The contours of constant density are concentric ellipses in d dimensions whose centers are at $(\mu_1, \ldots, \mu_d)^t = \boldsymbol{\mu}$, and whose axes in the ith direction are of length $2\sigma_i\sqrt{c}$ for the density $p(\mathbf{x})$ held constant at

$$\frac{e^{-c/2}}{\prod_{i=1}^{d}\sqrt{2\pi}\sigma_i}.$$

The axes of the ellipses are parallel to the coordinate axes. The plot in 2 dimensions ($d = 2$) is shown:



(c) The squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ is:

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}-\boldsymbol{\mu})^t \begin{pmatrix} 1/\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_d^2 \end{pmatrix} (\mathbf{x}-\boldsymbol{\mu}) \\
&= \sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2.
\end{aligned}$$

Section 2.6

25. A useful discriminant function for Gaussians is given by Eq. 52 in the text,

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i).$$

We expand to get

$$\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}\left[\mathbf{x}^t\mathbf{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_i^t\mathbf{\Sigma}^{-1}\mathbf{x} - \mathbf{x}^t\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^t\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \\
&= -\frac{1}{2}\Big[\underbrace{\mathbf{x}^t\mathbf{\Sigma}^{-1}\mathbf{x}}_{\text{indep. of } i} - 2\boldsymbol{\mu}_i^t\mathbf{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i\Big] + \ln P(\omega_i).
\end{aligned}$$

We drop the term that is independent of i, yields the equivalent discriminant function:

$$\begin{aligned} g_i(\mathbf{x}) &= \boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) \\ &= \mathbf{w}_i^t\mathbf{x} + w_{io}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{w}_i &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \\ w_{io} &= -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i). \end{aligned}$$

The decision boundary for two Gaussians is given by $g_i(\mathbf{x}) = g_j(\mathbf{x})$ or

$$\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) = \boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln P(\omega_j).$$

We collect terms so as to rewrite this as:

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln\frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}\left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{\ln\,[P(\omega_i)/P(\omega_j)](\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\right]$$

$$\underbrace{-\frac{1}{2}\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j}_{=0} = 0.$$

This is the form of a linear discriminant

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_o) = 0,$$

where the weight and bias (offset) are

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

and

$$\mathbf{x}_o = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\,[P(\omega_i)/P(\omega_j)](\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)},$$

respectively.

26. The densities and Mahalanobis distances for our two distributions with the same covarance obey

$$\begin{aligned} p(\mathbf{x}|\omega_i) &\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \\ r_i^2 &= (\mathbf{x} - \boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \end{aligned}$$

for $i = 1, 2$.

(a) Our goal is to show that $\nabla r_i^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$. Here ∇r_i^2 is the gradient of r_i^2, that is, the (column) vector in d-dimensions given by:

$$\begin{pmatrix} \frac{\partial r_i^2}{\partial x_1} \\ \vdots \\ \frac{\partial r_i^2}{\partial x_d} \end{pmatrix}.$$