

Solution to Problem Set 2

MAS 622J/1.126J: Pattern Recognition and Analysis

Handed in Wednesday, September 24th, 2008

Problem 1:

In a particular binary hypothesis testing application, the conditional density for a scalar feature y given class w_1 is

$$p_{y|w_1}(y|w_1) = k_1(\exp(-y^2/18))$$

Given class w_2 the conditional density is

$$p_{y|w_2}(y|w_2) = k_2(\exp(-(y-3)^2/8))$$

- a. Find k_1 and k_2 , and plot the two densities on a single graph using Matlab/Python.

We solve for the parameters k_1 and k_2 by recognizing that the two equations are in the form of the normal Gaussian distribution.

$$\begin{aligned} k_1 e^{-\frac{y^2}{18}} &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ \frac{-y^2}{18} &= \frac{-(y-\mu)^2}{2\sigma^2} \\ \mu &= 0; \sigma^2 = 9 \\ k_1 &= \frac{1}{\sqrt{18\pi}} \end{aligned}$$

In a similar fashion, we find that

$$k_2 = \frac{1}{\sqrt{8\pi}}$$

These distributions are plotted in Figure 1.

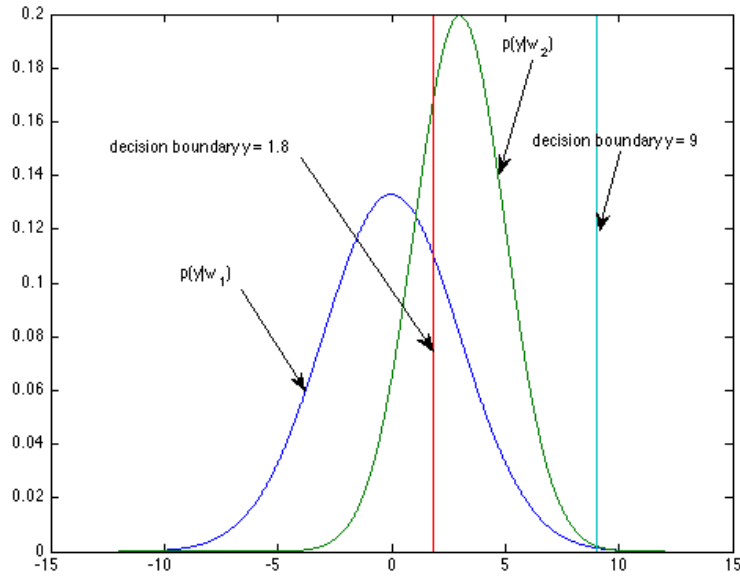


Figure 1: The pdf and decision boundaries

- b. Assume that the prior probabilities of the two classes are equal, and that the cost for choosing correctly is zero. If the costs for choosing incorrectly are $C_{12} = 1$ and $C_{21} = 1.5$, what is the expression for the Bayes risk?

The conditional risk for *Two-Category Classification* is discussed in Section 2.2.1 in D.H.S. Where we see that the formulas for the risk associated with the action, α_i , of classifying a feature vector, x , as class, ω_i , is different for each action:

$$R(\alpha_1|x) = \lambda_{11}P(\omega_1|x) + \lambda_{12}P(\omega_2|x)$$

$$R(\alpha_2|x) = \lambda_{21}P(\omega_1|x) + \lambda_{22}P(\omega_2|x)$$

The prior probabilities of the two classes are equal:

$$P(\omega_1) = P(\omega_2) = .5$$

The assumption that the cost for choosing correctly is zero:

$$\lambda_{11} = 0, \lambda_{22} = 0$$

The costs for choosing incorrectly are given as C_{12} and C_{21} :

$$\lambda_{12} = C_{12} = 1$$

$$\lambda_{21} = C_{21} = 1.5$$

Thus the expression for the conditional risk of α_1 is:

$$\begin{aligned} R(\alpha_1|y) &= \lambda_{11}P(\omega_1|y) + \lambda_{12}P(\omega_2|y) \\ R(\alpha_1|y) &= 0P(\omega_1|y) + 1P(\omega_2|y) \\ R(\alpha_1|y) &= P(\omega_2|y) \end{aligned}$$

And the expression for the conditional risk of α_2 :

$$\begin{aligned} R(\alpha_2|y) &= \lambda_{21}P(\omega_1|y) + \lambda_{22}P(\omega_2|y) \\ R(\alpha_2|y) &= 1.5P(\omega_1|y) + 0P(\omega_2|y) \\ R(\alpha_2|y) &= 1.5P(\omega_1|y) \end{aligned}$$

- c. Find the decision regions which minimize the Bayes risk, and indicate them on the plot you made in part (a)

The Bayes Risk is the integral of the conditional risk when we use the optimal decision regions, R_1 and R_2 . So, solving for the optimal decision regions is a matter of solving for the roots of the inequality:

$$\begin{aligned} R(\alpha_1|y) &< R(\alpha_2|y) \\ P(\omega_2|y) &< 1.5P(\omega_1|y) \\ \frac{P(y|\omega_2)P(\omega_2)}{P(y)} &< \frac{1.5P(y|\omega_1)P(\omega_1)}{P(y)} \end{aligned}$$

Given the priors are equal this simplifies to:

$$P(y|\omega_2) < 1.5P(y|\omega_1)$$

Next, using the values, k_1 and k_2 , from part (a), we have expressions for $p_{y|\omega_1}$ and $p_{y|\omega_2}$.

$$\begin{aligned} \frac{1}{\sqrt{8\pi}}e^{-\frac{(y-3)^2}{8}} &< 1.5\frac{1}{\sqrt{18\pi}}e^{-\frac{y^2}{18}} \\ e^{-\frac{(y-3)^2}{8}} &< e^{-\frac{y^2}{18}} \\ \frac{-(y-3)^2}{8} &< \frac{-y^2}{18} \\ 18(y-3)^2 &> 8y^2 \\ 5y^2 - 54y + 81 &> 0 \end{aligned}$$

The decision boundary is found by solving for the roots of this quadratic, $y = 9$ and $y = 1.8$.

d. For the decision regions in part (c), what is the numerical value of the Bayes risk?

For $y < 1.8$ the decision rule will choose ω_1

For $1.8 < y < 9$ the decision rule will choose ω_2

For $y > 9$ the decision rule will choose ω_1

Thus the decision region χ_1 is $y < 1.8$ and $y > 9$, and the decision region χ_2 is $1.8 < y < 9$.

$$\begin{aligned}
 Risk &= \int_{\chi_1} \lambda_{11}P(\omega_1|y) + \lambda_{12}P(\omega_2|y) \\
 &+ \int_{\chi_2} \lambda_{21}P(\omega_1|y) + \lambda_{22}P(\omega_2|y) \\
 &= \int_{\chi_1} \lambda_{12}P(y|\omega_2)P(\omega_2) + \int_{\chi_2} \lambda_{21}P(y|\omega_1)P(\omega_1) \\
 &= \int_{y=-\text{inf}}^{y=1.8} (N(3, 4))(.5) + \int_{y=9}^{y=\text{inf}} (N(3, 4))(.5) \\
 &+ \int_{y=1.8}^{y=9} 1.5(N(0, 9))(.5) \\
 &= \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{(3 - 1.8)}{\sqrt{8}}\right)\right) + \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{(9 - 3)}{\sqrt{8}}\right)\right) \\
 &+ \frac{1.5}{2}\left(\frac{1}{2}\text{erf}\left(\frac{9}{\sqrt{18}}\right) - \frac{1}{2}\text{erf}\left(\frac{1.8}{\sqrt{18}}\right)\right) \\
 &= .3425
 \end{aligned}$$

Problem 2:

Let's consider a simple communication system. The transmitter sends out messages $\mathbf{m} = 0$ or $\mathbf{m} = 1$, occurring with a priori probabilities $\frac{1}{4}$ and $\frac{3}{4}$ respectively. The message is contaminated by a noise \mathbf{n} , which is independent from \mathbf{m} and takes on the values $-1, 0, 1$ with probabilities $\frac{1}{8}, \frac{3}{4}, \frac{1}{8}$ respectively. The received signal, or the observation, can be represented as $\mathbf{r} = \mathbf{m} + \mathbf{n}$. From \mathbf{r} , we wish to infer what the transmitted message \mathbf{m} was (estimated state), denoted using $\hat{\mathbf{m}}$. $\hat{\mathbf{m}}$ also takes values on 0 or 1. When $\mathbf{m} = \hat{\mathbf{m}}$, the detector correctly receives the original message, otherwise an error occurs.

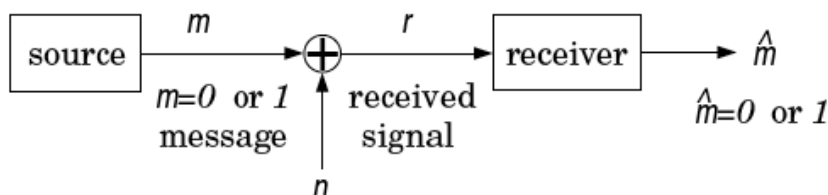


Figure 2: A simple receiver

- a. Find the decision rule that achieves the maximum probability of correct decision. Compute the probability of error for this decision rule.

It is equivalent to find the decision rule that achieves the minimum probability of error. The receiver decides the transmitted message is 1, i.e., $\hat{m} = 1$ if

$$\begin{aligned} P(r|\mathbf{m} = 1) \cdot \Pr(\mathbf{m} = 1) &\geq P(r|\mathbf{m} = 0) \cdot \Pr(\mathbf{m} = 0) \\ 3P(r|\mathbf{m} = 1) &\geq P(r|\mathbf{m} = 0) \end{aligned} \quad (1)$$

Otherwise the receiver decides $\hat{m} = 0$. The likelihood functions for these two cases are

$$P(r|\mathbf{m} = 1) = \begin{cases} 1/8 & \text{if } r = 0 \\ 3/4 & \text{if } r = 1 \\ 1/8 & \text{if } r = 2 \end{cases} \quad P(r|\mathbf{m} = 0) = \begin{cases} 1/8 & \text{if } r = -1 \\ 3/4 & \text{if } r = 0 \\ 1/8 & \text{if } r = 1 \end{cases}$$

Eq. (1) holds only when r is 1 or 2. Therefore, the decision rule can be summarized as

$$\hat{\mathbf{m}} = \begin{cases} 1 & \text{if } r = 1 \text{ or } 2 \\ 0 & \text{if } r = -1 \text{ or } 0 \end{cases}$$

The probability of error is

$$\begin{aligned} \Pr(e) &= \Pr(\hat{\mathbf{m}} = 1|\mathbf{m} = 0) \Pr(\mathbf{m} = 0) + \Pr(\hat{\mathbf{m}} = 0|\mathbf{m} = 1) \Pr(\mathbf{m} = 1) \\ &= \Pr(r = 1|\mathbf{m} = 0) \Pr(\mathbf{m} = 0) + \Pr(r = 0|\mathbf{m} = 1) \Pr(\mathbf{m} = 1) \\ &= 1/8 * 1/4 + 1/8 * 3/4 = 1/8 \end{aligned} \quad (2)$$

- b. Let's have the noise \mathbf{n} be a continuous random variable. \mathbf{n} is uniformly distributed between $-\frac{3}{4}$ and $\frac{3}{4}$, and still statistically independent of \mathbf{m} . First, plot the pdf of \mathbf{n} . Then, find a decision rule that achieves the minimum probability of error, and compute the probability of error.

The decision rule is still determined by using Eq. (1). The likelihood functions become continuous, instead of discrete in (a):

$$p(r|\mathbf{m} = 1) = \begin{cases} 2/3 & \text{if } 1/4 < r \leq 7/4 \\ 0 & \text{otherwise} \end{cases} \quad P(r|\mathbf{m} = 0) = \begin{cases} 2/3 & \text{if } -3/4 < r \leq 3/4 \\ 0 & \text{otherwise} \end{cases}$$

The interesting region is where the two pdf's overlap with each other, namely when $1/4 < r \leq 3/4$. From Eq.(1), we know we should decide $\hat{\mathbf{m}} = 1$ for this range.

The decision rule can be summarized as

$$\hat{\mathbf{m}} = \begin{cases} 1 & \text{if } 1/4 < r \leq 7/4 \\ 0 & \text{if } -3/4 < r \leq 1/4 \end{cases}$$

Note that at the decision boundaries, there is ambiguity on which decision we should make. Again, either decision won't change the probability of error, so it is acceptable to decide both ways.

The probability of error is

$$\begin{aligned} \Pr(e) &= \Pr(\hat{\mathbf{m}} = 1|\mathbf{m} = 0) \Pr(\mathbf{m} = 0) + \Pr(\hat{\mathbf{m}} = 0|\mathbf{m} = 1) \Pr(\mathbf{m} = 1) \\ &= \Pr(1/4 < r \leq 3/4|\mathbf{m} = 0) \Pr(\mathbf{m} = 0) \\ &= (3/4 - 1/4) * 2/3 * 1/4 = 1/12 \end{aligned} \tag{3}$$

Problem 3:

[Note: Use Matlab or Python for the computations, but make sure to explicitly construct every transformation required, that is either type it or write it. Do not use Matlab or Python if you are asked to explain/show something.]

Consider the three-dimensional normal distribution $p(\mathbf{x}|w)$ with mean μ and covariance matrix Σ where

$$\mu = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 3 \\ 0 & 3 & 5 \end{pmatrix}.$$

Compute the matrices representing the eigenvectors and eigenvalues Φ and Λ to answer the following:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 8 \end{pmatrix}, \Phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

a. Find the probability density at the point $\mathbf{x}_0 = (9 \ 0 \ 3)^T$

$$p(x_0|\omega) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} e^{-\frac{(x_0-\mu)^T \Sigma^{-1} (x_0-\mu)}{2}} \quad (4)$$

$$|\Sigma| = 16; \quad \Sigma^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5/16 & -3/16 \\ 0 & -3/16 & 5/16 \end{pmatrix}$$

The squared Mahalanobis distance from the mean to x_0 is:

$$\begin{aligned} (x_0 - \mu)^T \Sigma^{-1} (x_0 - \mu) &= \\ \left[\begin{pmatrix} 9 \\ 0 \\ 3 \end{pmatrix} - \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \right]^T &\begin{pmatrix} 1 & 0 & 0 \\ 0 & 5/16 & -3/16 \\ 0 & -3/16 & 5/16 \end{pmatrix} \left[\begin{pmatrix} 9 \\ 0 \\ 3 \end{pmatrix} - \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \right] \\ &= 51.3125 \end{aligned}$$

We plug these values in to find that the density at x_0 is:

$$p(x_0|\omega) = 1.1437e - 13$$

b. Construct an orthonormal transformation $\mathbf{y} = \Phi^T \mathbf{x}$. Show that for orthonormal transformations, Euclidean distances are preserved (i.e., $\|y\|^2 = \|x\|^2$).

$$\begin{aligned} y &= \Phi^T x \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^T x \end{aligned}$$

To prove that for orthonormal transformations, Euclidean distances are preserved we have to show that $\|y\|^2 = \|x\|^2$.

Proof: Let y be a random variable such that $y = A^T x$ where A^T is an orthonormal transformation (i.e., $A^T A = I$)

$$\begin{aligned} \|y\|^2 &= y^T y \\ &= (A^T x)^T A^T x \\ &= x^T A A^T x \\ &= x^T x \\ &= \|x\|^2 \end{aligned}$$

- c. After applying the orthonormal transformation add another transformation $\Lambda^{-1/2}$ and convert the distribution to one centered on the origin with covariance matrix equal to the identity matrix. Show that $\mathbf{A}_w = \Phi\Lambda^{-1/2}$ is a linear transformation (i.e., $\mathbf{A}_w(a\mathbf{x} + b\mathbf{y}) = a\mathbf{A}_w\mathbf{x} + b\mathbf{A}_w\mathbf{y}$)

$$\begin{aligned} u &= \Lambda^{-1/2}y \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{2\sqrt{2}} \end{pmatrix} y \end{aligned}$$

Where u can also be expressed as $u = \Lambda^{-1/2}\Phi^T x$

Let $u = \mathbf{A}_w^T x$ where $\mathbf{A}_w = \Phi\Lambda^{-1/2}$.

Given $p(x|w) \sim N(\mu, \Sigma)$ then $p(u|w) \sim N(\mathbf{A}_w^T \mu, \mathbf{A}_w^T \Sigma \mathbf{A}_w)$

Since u is a result of a whitening transform to x , then the covariance matrix of u is proportional to the Identity matrix I . In this case, since Φ is a matrix of normalized eigenvectors of Σ and Λ is the eigenvalue matrix, the transformation $A^T = \Lambda^{-1/2}\Phi^T$ makes the covariance matrix equal to I .

Therefore, to convert the distribution to one centered on the origin with the covariance equal to the Identity matrix, it is enough to define a new variable z such that $z = \mathbf{A}_w^T x - \mathbf{A}_w^T \mu$. In that way $p(z|w) \sim N(0, 1)$

To prove that $\mathbf{A}_w = \Phi\Lambda^{-1/2}$ is a linear transformation we need to prove that $\mathbf{A}_w(a\mathbf{x} + b\mathbf{y}) = a\mathbf{A}_w\mathbf{x} + b\mathbf{A}_w\mathbf{y}$.

Proof.

$$\begin{aligned} \mathbf{A}_w(a\mathbf{x} + b\mathbf{y}) &= \Phi\Lambda^{-1/2}(a\mathbf{x} + b\mathbf{y}) \\ &= \Phi(a\Lambda^{-1/2}\mathbf{x} + b\Lambda^{-1/2}\mathbf{y}) \\ &= (a\Phi\Lambda^{-1/2}\mathbf{x} + b\Phi\Lambda^{-1/2}\mathbf{y}) \\ &= (a\mathbf{A}_w\mathbf{x} + b\mathbf{A}_w\mathbf{y}) \end{aligned}$$

- d. Apply the same overall transformation to \mathbf{x}_0 to yield a transformed point \mathbf{x}_w

$$\begin{aligned}
\mathbf{x}_w &= \mathbf{A}_w^T \mathbf{x}_0 - \mathbf{A}_w^T \mu \\
&= \mathbf{A}_w^T (\mathbf{x}_0 - \mu) \\
&= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{4} \end{pmatrix}^T \begin{pmatrix} 7 \\ -1 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} 7 \\ 3/2 \\ 1/4 \end{pmatrix}
\end{aligned}$$

- e. Calculate the Mahalanobis distance from \mathbf{x}_0 to the mean μ and from \mathbf{x}_w to $\mathbf{0}$. Are they different or are they the same? Why?

The squared Mahalanobis distance from \mathbf{x}_0 to the mean μ , we found in part (a), is 51.3125

The squared Mahalanobis distance from \mathbf{x}_w to 0 is:

$$\begin{aligned}
\mathbf{x}_w^T \mathbf{x}_w &= \begin{pmatrix} 7 \\ 3/2 \\ 1/4 \end{pmatrix}^T \begin{pmatrix} 7 \\ 3/2 \\ 1/4 \end{pmatrix} \\
&= 51.3125
\end{aligned}$$

The two distances are the same, as expected under any linear transformation.

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

Proof.

Let \mathbf{x} be a random variable such that $p(\mathbf{x}|\mu, \Sigma) \sim N(\mu, \Sigma)$.

Let $\mathbf{y} = \mathbf{Z}^T \mathbf{x}$, where $p(\mathbf{y}|\mathbf{Z}^T \mu, \mathbf{Z}^T \Sigma \mathbf{Z}) \sim N(\mathbf{Z}^T \mu, \mathbf{Z}^T \Sigma \mathbf{Z})$, and

\mathbf{Z} is a general linear transformation where $|\mathbf{Z}| \neq 0$.

Let $p_x(\mathbf{x}) = p(\mathbf{x}|\mu, \Sigma)$ and $p_y(\mathbf{y}) = p(\mathbf{y}|\mathbf{Z}^T \mu, \mathbf{Z}^T \Sigma \mathbf{Z})$.

$$\begin{aligned} P_Y(y) &= Pr(Y \leq \mathbf{y}) \\ &= Pr(\mathbf{Z}^T X \leq \mathbf{y}) \\ &= Pr(X \leq (\mathbf{Z}^T)^{-1} \mathbf{y}) \text{ note: } (\mathbf{Z}^T)^{-1} \text{ exists since } |\mathbf{Z}| \neq 0. \\ &= P_X((\mathbf{Z}^T)^{-1} \mathbf{y}) \end{aligned}$$

Taking the derivative of both sides

$$\begin{aligned} \frac{d}{dy} P_Y(y) &= \frac{d}{dy} P_X((\mathbf{Z}^T)^{-1} \mathbf{y}) \\ &= \frac{d}{dx} P_X((\mathbf{Z}^T)^{-1} \mathbf{y}) \Big| \frac{d}{dy} (\mathbf{Z}^T)^{-1} \mathbf{y} \\ &= \frac{d}{dx} P_X(\mathbf{x}) |(\mathbf{Z}^T)^{-1}| \\ &= p_X(\mathbf{x}) |(\mathbf{Z}^T)^{-1}| \\ &= p_X(\mathbf{x}) |(\mathbf{Z}^{-1})^T| \\ &= p_X(\mathbf{x}) |\mathbf{Z}^{-1}| \\ &= \frac{p_X(\mathbf{x})}{|\mathbf{Z}|} \end{aligned}$$

Therefore the density will be change unless $|\mathbf{Z}| = 1$.

Thus, if the determinant of \mathbf{Z} is equal to 1, then the probability density remains unchanged under a general linear transformation. Notice that \mathbf{Z} is not necessarily an orthonormal transformation.

A key point of this problem is that depending on the transformation applied to a Gaussian some things will remain unchanged while others will change.

- The Euclidean distance is invariant under an orthonormal transformation (i.e., if \mathbf{Y} is an orthonormal transformation of \mathbf{X} then $\sqrt{(x_1 - x_2)^2} = \sqrt{(y_1 - y_2)^2}$).
- Linear transformation will usually change the probability density, except when the determinant is equal to one.

- Under linear transformations the Mahalanobis distance is preserved. Consider the transformation $y = A^T x$. Then we have:

$$\begin{aligned} \sqrt{(y_1 - y_2)^T \Sigma_y^{-1} (y_1 - y_2)} &= \sqrt{(x_1 - x_2)^T A \Sigma_y^{-1} A^T (x_1 - x_2)} \\ \text{Note: } \Sigma_y^{-1} &= A^{-1} \Sigma_x^{-1} A^{-T} \\ &= \sqrt{(x_1 - x_2)^T A A^{-1} \Sigma_x^{-1} A^{-T} A^T (x_1 - x_2)} \\ &= \sqrt{(x_1 - x_2)^T \Sigma_x^{-1} (x_1 - x_2)} \end{aligned}$$

Problem 4:

Let \mathbf{x} be an observation vector. You would like to determine whether \mathbf{x} belongs to w_1 or w_2 based on the following decision rule, namely *decision rule 1*.

Decide w_1 if $-\ln \mathbf{p}(\mathbf{x}|w_1) + \ln \mathbf{p}(\mathbf{x}|w_2) < \ln\{\mathbf{P}(w_1)/\mathbf{P}(w_2)\}$; otherwise decide w_2 .

You know that this rule does not lead to perfect classification therefore you must calculate the probability of error. Let χ_1 and χ_2 be the region in the domain of \mathbf{x} such that

$\mathbf{p}(\mathbf{x}|w_1)\mathbf{P}(w_1) > \mathbf{p}(\mathbf{x}|w_2)\mathbf{P}(w_2)$ and $\mathbf{p}(\mathbf{x}|w_1)\mathbf{P}(w_1) < \mathbf{p}(\mathbf{x}|w_2)\mathbf{P}(w_2)$, respectively.

Then if $\mathbf{x} \in \chi_i$, for $i = 1, 2$ assign the sample to class w_i . Use excruciating detail to answer the following:

- Show that the $\text{Pr}[\text{error}]$ for this rule is given by:

$$\text{Pr}[\text{error}] = \mathbf{P}(w_1)\epsilon_1 + \mathbf{P}(w_2)\epsilon_2$$

$$\text{where } \epsilon_1 = \int_{\chi_2} \mathbf{p}(\mathbf{x}|w_1) \mathbf{d}\mathbf{x} \text{ and } \epsilon_2 = \int_{\chi_1} \mathbf{p}(\mathbf{x}|w_2) \mathbf{d}\mathbf{x}$$

$$\begin{aligned} \text{Pr}[\text{error}] &= P(\text{error}|w_1)P(w_1) + P(\text{error}|w_2)P(w_2) \\ &= P(x \in \chi_2|w_1)P(w_1) + P(x \in \chi_1|w_2)P(w_2) \\ &= P(w_1) \int_{\chi_2} p(x|w_1) dx + P(w_2) \int_{\chi_1} p(x|w_2) dx \\ &= P(w_1)\epsilon_1 + P(w_2)\epsilon_2 \end{aligned}$$

- b. Describe what the previous equation says about the total error. (hint: identify what ϵ_1 and ϵ_2 mean)

There are two ‘types’ of errors that we can make: one when we misclassify samples from w_1 and the other when we misclassify samples from w_2 . The total error is a weighted sum each of these.

- c. Suppose that for a given decision, you must pay a cost depending on the true class of the sample based on *decision rule 1*. Assume that a wrong decision is more expensive than a correct one, where $\lambda_{ij} = \lambda(\text{deciding } w_i | w_j)$ is the loss incurred for deciding w_i when the state of nature is w_j . Write an expression for the expected cost, namely risk, R , such that

$$E[\text{cost}] = E[\text{fixed costs}] + E[\text{variable costs}]$$

Decision rule 1 is the same we used for part (1): Decide w_1 if $-\ln \mathbf{p}(\mathbf{x}|w_1) + \ln \mathbf{p}(\mathbf{x}|w_2) < \ln\{\mathbf{P}(w_1)/\mathbf{P}(w_2)\}$; otherwise decide w_2 . Any decision has one of four costs associated with it, so the $E[\text{cost}]$ or Risk is:

$$\begin{aligned} E[\text{cost}] &= \sum_{i=1}^2 \sum_{j=1}^2 \lambda_{ij} \Pr[x \in \chi_i | w_j] \Pr[w_j] \\ &= \int_{\chi_1} \lambda_{11} p(x|w_1) p(w_1) + \lambda_{12} p(x|w_2) p(w_2) dx \\ &\quad + \int_{\chi_2} \lambda_{21} p(x|w_1) p(w_1) + \lambda_{22} p(x|w_2) p(w_2) dx \end{aligned} \quad (1)$$

Note that each conditional density must sum to one over all the entire χ_1 and χ_2 , since χ_1 and χ_2 do not overlap and cover the entire domain:

$$\int_{\chi_1} p_{x|w_i}(x|w_i) dx + \int_{\chi_2} p_{x|w_i}(x|w_i) dx = 1 \quad (2)$$

Then applying (2) to (1) we have:

$$\begin{aligned} E[\text{cost}] &= \int_{\chi_1} \lambda_{11} p(x|w_1) p(w_1) + \lambda_{12} p(x|w_2) p(w_2) dx \\ &\quad + \lambda_{21} p(w_1) \left(1 - \int_{\chi_1} p(x|w_1) dx\right) + \lambda_{22} p(w_2) \left(1 - \int_{\chi_1} p(x|w_2) dx\right) \\ &= \int_{\chi_1} -(\lambda_{21} - \lambda_{11}) p(w_1) p(x|w_1) + (\lambda_{12} - \lambda_{22}) p(w_2) p(x|w_2) dx \\ &\quad + \{\lambda_{21} p(w_1) + \lambda_{22} p(w_2)\} \end{aligned} \quad (3)$$

Where $E[\text{variable cost}] = \int_{\chi_1} -(\lambda_{21} - \lambda_{11})p(w_1)p(x|w_1) + (\lambda_{12} - \lambda_{22})p(w_2)p(x|w_2)dx$
and $E[\text{fixed cost}] = \{\lambda_{21}p(w_1) + \lambda_{22}p(w_2)\}$

- d. Suppose that for a given value of \mathbf{x} , the integrand in the risk function is positive. How can you decrease the risk? (hint: think about where you would assign \mathbf{x} to and why you would make that decision)

From equation (3) we see that the risk R is to be minimized by appropriately choosing the region χ_1 . Since the $E[\text{fixed cost}]$ is not a function of χ_1 , minimizing R is equal to minimizing the $E[\text{variable cost}]$.

First consider that an arbitrary function $g(y)$ with both positive and negative values, if one integrates $g(y)$ over regions where the function is positive, the integral increases; if one integrates over regions where $g(y)$ is negative, the integral decreases.

Then, to minimize $E[\text{variable cost}]$ one should define the region χ_1 to include all of the points and only the points where $E[\text{variable cost}]$ is negative. In other words, since χ_1 is the region where one chooses w_1 , then one should choose w_1 whenever $g(y)$ as defined by the integrand of the $E[\text{variable cost}]$ is negative. For the current problem we are given a value of \mathbf{x} that makes the integrand of the risk function positive. Then we can decrease the risk by decreasing the $E[\text{variable cost}]$ by assigning \mathbf{x} to χ_2 .

- e. Show that for a zero-one cost function $\lambda_{12} - \lambda_{22} = \lambda_{21} - \lambda_{11}$, $E[\text{cost}] = \Pr[\text{error}]$.

Zero-one cost function: $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$

$$\begin{aligned} E[\text{cost}] &= \int_{\chi_1} \lambda_{11}p(x|w_1)p(w_1) + \lambda_{12}p(x|w_2)p(w_2)dx \\ &\quad + \int_{\chi_2} \lambda_{21}p(x|w_1)p(w_1) + \lambda_{22}p(x|w_2)p(w_2)dx \\ &= p(w_2) \int_{\chi_1} p(x|w_2)dx + p(w_1) \int_{\chi_2} p(x|w_1)dx \\ &= p(w_2)\epsilon_1 + p(w_1)\epsilon_2 \end{aligned}$$

Problem 5:

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

- a. Prove that $\mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_2)$ and $\mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_1)$, taken together, uniquely determine the discriminability \mathbf{d}'

Let $\mathbf{x} = x$ and $\mathbf{x}^* = x^*$. Based on the Gaussian assumption, we see that $P(x > x^* | x \in w_i) = P\left(\frac{x - \mu_i}{\sigma_i} > \frac{x^* - \mu_i}{\sigma_i} | x \in w_i\right) \sim N(0, 1)$ for $i = 1, 2$. Thus, we know $\left(\frac{x - \mu_2}{\sigma_2}\right)$ from the hit rate $P(x > x^* | x \in w_2)$ and $\left(\frac{x - \mu_1}{\sigma_1}\right)$ from the false alarm rate $P(x > x^* | x \in w_1)$, and these let us calculate the discriminability.

$$\begin{aligned} \left| \frac{x^* - \mu_1}{\sigma_1} - \frac{x^* - \mu_2}{\sigma_2} \right| &= \left| \frac{x^* - \mu_1}{\sigma} - \frac{x^* - \mu_2}{\sigma} \right| \\ &= \left| \frac{\mu_2 - \mu_1}{\sigma} \right| \\ &= \frac{|\mu_2 - \mu_1|}{\sigma} \\ &= d' \end{aligned}$$

Therefore d' is uniquely determined by the hit and false alarm rates.

- b. Use error functions $erf(*)$ to express \mathbf{d}' in terms of the hit and false alarm rates. Estimate \mathbf{d}' if $\mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_1) = .65$ and $\mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_2) = .5$. Repeat for \mathbf{d}' if $\mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_1) = .95$ and $\mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_2) = .15$.

There are a couple of ways in which you can proceed. This document will detail one of them.

Let y be a random variable such that $P(y > y^* | y \in w_i) \sim N(0, 1)$ for $i = 1, 2$.

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_0^{y^*} e^{-\frac{t^2}{2}} dt &= \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}} \int_{-y^*}^{y^*} e^{-\frac{t^2}{2}} dt \right) \\ \text{let } u &= \frac{t^2}{2}; du = \frac{dt}{\sqrt{2}} \\ &= \frac{1}{2} \left(\frac{1}{\sqrt{\pi}} \int_{-\frac{y^*}{\sqrt{2}}}^{\frac{y^*}{\sqrt{2}}} e^{-u} du \right) \\ &= \frac{1}{2} \left(\frac{2}{\sqrt{\pi}} \int_0^{\frac{y^*}{\sqrt{2}}} e^{-u} du \right) \\ &= \frac{1}{2} erf \left(\frac{y^*}{\sqrt{2}} \right) \end{aligned}$$

Then, given that $P(x > x^* | x \in w_i) = P\left(\frac{x - \mu_i}{\sigma_i} > \frac{x^* - \mu_i}{\sigma_i} | x \in w_i\right) \sim N(0, 1)$ for $i = 1, 2$.

$$\begin{aligned}
P\left(\frac{x - \mu_i}{\sigma_i} > \frac{x^* - \mu_i}{\sigma_i} \mid x \in w_i\right) &= 1 - P\left(\frac{x - \mu_i}{\sigma_i} < \frac{x^* - \mu_i}{\sigma_i} \mid x \in w_i\right) \\
&= \begin{cases} 1 - \left(\frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x^* - \mu_i}{\sqrt{2}\sigma_i}\right)\right) & \text{if } \frac{x^* - \mu_i}{\sigma_i} > 0 \\ 1 - \left(\frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(\frac{\mu_i - x^*}{\sqrt{2}\sigma_i}\right)\right) & \text{if } \frac{x^* - \mu_i}{\sigma_i} < 0 \end{cases} \\
&= \begin{cases} \frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(\frac{x^* - \mu_i}{\sqrt{2}\sigma_i}\right) & \text{if } \frac{x^* - \mu_i}{\sigma_i} > 0 \\ \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{\mu_i - x^*}{\sqrt{2}\sigma_i}\right) & \text{if } \frac{x^* - \mu_i}{\sigma_i} < 0 \end{cases} \\
&= \begin{cases} \frac{1}{2} \left(1 - \operatorname{erf}\left(\frac{x^* - \mu_i}{\sqrt{2}\sigma_i}\right)\right) & \text{if } \frac{x^* - \mu_i}{\sigma_i} > 0 \\ \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{\mu_i - x^*}{\sqrt{2}\sigma_i}\right)\right) & \text{if } \frac{x^* - \mu_i}{\sigma_i} < 0 \end{cases}
\end{aligned}$$

Therefore:

$$\frac{x^* - \mu_i}{\sigma_i} = \sqrt{2} \left(\operatorname{erf}^{-1}[1 - 2P(x > x^* \mid x \in w_i)]\right) \text{ if } P(x > x^* \mid x \in w_i) < .5$$

$$\frac{\mu_i - x^*}{\sigma_i} = \sqrt{2} \left(\operatorname{erf}^{-1}[2P(x > x^* \mid x \in w_i) - 1]\right) \text{ if } P(x > x^* \mid x \in w_i) > .5$$

Then to find the values of $\operatorname{erf}^{-1}()$ use a table of erf , a table of the cumulative normal distribution, or use Matlab erf and erfinv functions.

$$\begin{aligned}
d_1^* &= \left| -\frac{\mu_1 - x^*}{\sigma} - \frac{x^* - \mu_2}{\sigma} \right| \\
&= \left| -\sqrt{2} \left(\operatorname{erf}^{-1}[2(.65) - 1]\right) - \sqrt{2} \left(\operatorname{erf}^{-1}[2(.5) - 1]\right) \right| \\
&= \left| -\sqrt{2} \left(\operatorname{erf}^{-1} [.3]\right) - \sqrt{2} \left(\operatorname{erf}^{-1} [0]\right) \right| \\
&= .3853
\end{aligned}$$

$$\begin{aligned}
d_2^* &= \left| -\frac{\mu_1 - x^*}{\sigma} - \frac{x^* - \mu_2}{\sigma} \right| \\
&= \left| -\sqrt{2} \left(\operatorname{erf}^{-1}[2(.95) - 1]\right) - \sqrt{2} \left(\operatorname{erf}^{-1}[1 - 2(.15)]\right) \right| \\
&= \left| -\sqrt{2} \left(\operatorname{erf}^{-1} [.9]\right) - \sqrt{2} \left(\operatorname{erf}^{-1} [.7]\right) \right| \\
&= 2.6813
\end{aligned}$$

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

According to Problem 3

$$\Pr[\text{error}] = \mathbf{P}(w_1)\epsilon_1 + \mathbf{P}(w_2)\epsilon_2$$

$$\text{where } \epsilon_1 = \int_{\chi_2} \mathbf{p}(\mathbf{x}|w_1)\mathbf{d}\mathbf{x} \text{ and } \epsilon_2 = \int_{\chi_1} \mathbf{p}(\mathbf{x}|w_2)\mathbf{d}\mathbf{x}$$

Since the regions χ_1 and χ_2 are defined by our decision boundary x^* . We can see that

$$\epsilon_1 = \int_{\chi_2} \mathbf{p}(\mathbf{x}|w_1)\mathbf{d}\mathbf{x} = P(x < x^* | x \in w_1) = 1 - P(x > x^* | x \in w_1)$$

Similarly

$$\epsilon_2 = \int_{\chi_1} \mathbf{p}(\mathbf{x}|w_2)\mathbf{d}\mathbf{x} = P(x > x^* | x \in w_2)$$

Therefore,

$$\begin{aligned} \Pr[\text{error1}] &= (1 - .65)P(w_1) + (.5)P(w_2) \\ \Pr[\text{error2}] &= (1 - .95)P(w_1) + (.15)P(w_2) \end{aligned}$$

And if the priors are equally likely,

$$\begin{aligned} \Pr[\text{error1}] &= .425 \\ \Pr[\text{error2}] &= .1 \end{aligned}$$

- d. Using a trivial one-line computation or a graph determine which case has the higher \mathbf{d}' , and explain your logic:

$$\text{Case A: } \mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_1) = .75 \text{ and } \mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_2) = .35.$$

$$\text{Case B: } \mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_1) = .8 \text{ and } \mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_2) = .25.$$

From part (b) we can calculate the discriminability given the hit and false alarm rates, and see that Case B has a higher discriminability.

$$\begin{aligned}
d_1^* &= \left| -\frac{\mu_1 - x^*}{\sigma} - \frac{x^* - \mu_2}{\sigma} \right| \\
&= \left| -\sqrt{2} (\operatorname{erf}^{-1}[2(.75) - 1]) - \sqrt{2} (\operatorname{erf}^{-1}[1 - 2(.35)]) \right| \\
&= \left| -\sqrt{2} (\operatorname{erf}^{-1} [.5]) - \sqrt{2} (\operatorname{erf}^{-1} [.3]) \right| \\
&= 1.0598
\end{aligned}$$

$$\begin{aligned}
d_2^* &= \left| \frac{x^* - \mu_1}{\sigma} + \frac{\mu_2 - x^*}{\sigma} \right| \\
&= \left| \sqrt{2} (\operatorname{erf}^{-1}[1 - 2(.25)]) + \sqrt{2} (\operatorname{erf}^{-1}[2(.8) - 1]) \right| \\
&= \left| \sqrt{2} (\operatorname{erf}^{-1} [.5]) + \sqrt{2} (\operatorname{erf}^{-1} [.6]) \right| \\
&= 1.5161
\end{aligned}$$

A better way to argue about this is to start from part (c). For case A, the error of probability is $(1-.75 + .35) / 2 = .3$; for case B, the probability of error is $(1-.8+.25) / 2 = 0.225$. While B's probability of error is smaller, there must be a larger discriminability.

Key point of this problem: The error function erf is related to the cumulative normal distribution. Assuming that the probability of x belonging to one of two classes is Gaussian, then knowing the values of the false alarm and hit rates for an arbitrary x^* is enough to compute the discriminability d' . Moreover, if the Gaussian assumption holds, a determination of the discriminability allows us to calculate the Bayes error rate.

Problem 6:

- Show that the maximum likelihood (ML) estimation of the mean for a Gaussian is unbiased but the ML estimate of variance is biased (i.e., slightly wrong). Show how to correct this variance estimate so that it is unbiased.

Sample mean is unbiased:

$$\begin{aligned} E[\hat{\mu}] &= E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[x_i] \\ &= \frac{1}{n} \sum_{i=1}^n \mu \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned}$$

This document will show two ways to prove that the sample variance is biased.
Proof 1:

$$\begin{aligned} E\left[\sum_{i=1}^n (x_i - \hat{\mu})^2\right] &= E\left[\sum_{i=1}^n (x_i)^2\right] - nE[\hat{\mu}^2] \\ &= nE[(x_i)^2] - \frac{1}{n}E\left[\left(\sum_{i=1}^n x_i\right)^2\right] \\ &= n(\text{var}(x_i) + (E[x_i])^2) - \frac{1}{n}E\left[\left(\sum_{i=1}^n x_i\right)^2\right] \\ &= n\sigma^2 + \frac{1}{n}(nE[x_i])^2 - \frac{1}{n}E\left[\left(\sum_{i=1}^n x_i\right)^2\right] \\ &= n\sigma^2 - \frac{1}{n}\left[E\left[\left(\sum_{i=1}^n x_i\right)^2\right] - \left(E\left[\sum_{i=1}^n x_i\right]\right)^2\right] \\ &= n\sigma^2 - \frac{1}{n}\text{var}\left(\sum_{i=1}^n x_i\right) \\ &= n\sigma^2 - \frac{1}{n}n\sigma^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

Proof 2:

$$\begin{aligned} E[\sigma_{ML}^2] &= E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2] \\ &= \frac{1}{n} \sum_{i=1}^n E[x_i^2] - 2\hat{\mu}E\left[\sum_{i=1}^n x_i\right] + \sum_{i=1}^n E[\hat{\mu}^2] \\ &= \frac{1}{n} [n(\sigma^2 + \mu^2) - 2nE[\hat{\mu}^2] + nE[\hat{\mu}^2]] \\ &= \frac{1}{n} [n\sigma^2 + n\mu^2 - nE[\hat{\mu}^2]] \\ &= \frac{1}{n} [n\sigma^2 + n\mu^2 - n(\frac{\sigma^2}{n} + \mu^2)] \\ &= \frac{n-1}{n} \sigma^2 \end{aligned}$$

Therefore

$$\begin{aligned} \text{unbiased: } \hat{\sigma}^2 &= \frac{n}{n-1} \sigma_{ML}^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2 \end{aligned}$$

b. For this part you'll write a program with Matlab/Python to explore the biased and unbiased ML estimations of variance for a Gaussian distribution. Find the data for this problem on the class webpage as ps2.dat. This file contains n=5000 samples from a 1-dimensional Gaussian distribution.

- (a) Write a program to calculate the ML estimate of the mean, and report the output.
- (b) Write a program to calculate both the biased and unbiased ML estimate of the variance of this distribution. For n=1 to 5000, plot the biased and unbiased estimates of the variance of this Gaussian. This is as if you are being given these samples sequentially, and each time you get a new sample you are asked to re-evaluate your estimate of the variance. Give some interpretation of your plot.

The dataset was generated from a Gaussian normal distribution $\sim N(6, 1)$. The following Matlab code calculates the ML estimate of the mean = 6, and produces Figure 3 indicating that the variance = 1.0.

```
data = load('ps2.dat');
```

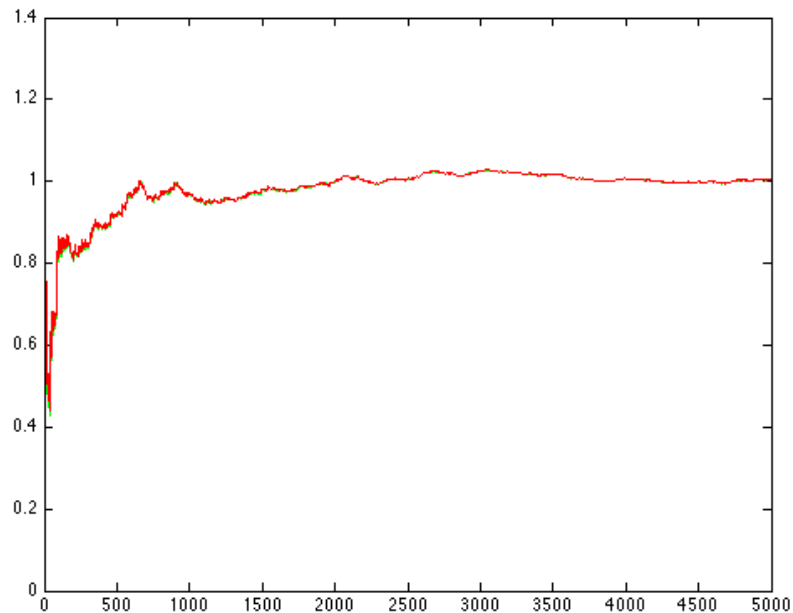


Figure 3: Though the two estimators have quite different estimates in the first few samples, the bias loses significance as n is large, and both estimators give the true variance of the data $\sigma = 1.0$.

```

n = length(data)
%ML estimate of mean:
MLmean = sum(data) / n
MLvar = [];
unMLvar = [];
%ML estimate of variance:
for(i=2:n)
    varML = 0;
    tempMean = sum(data(1:i)) / i;
    for (j=1:i)
        varML = varML + (data(j) - tempMean)^2;
    end

    MLvar(i) = varML/i;
    x = data(1:i);
    unMLvar(i) = cov(data(1:i));
end
figure
plot(1:1:n, MLvar(1:n), 'g', 1:1:n, unMLvar(1:n), 'r')

```

Problem 7:

Suppose \mathbf{x} and \mathbf{y} are random variables. Their joint density, depicted below, is constant in the shaded area and 0 elsewhere.

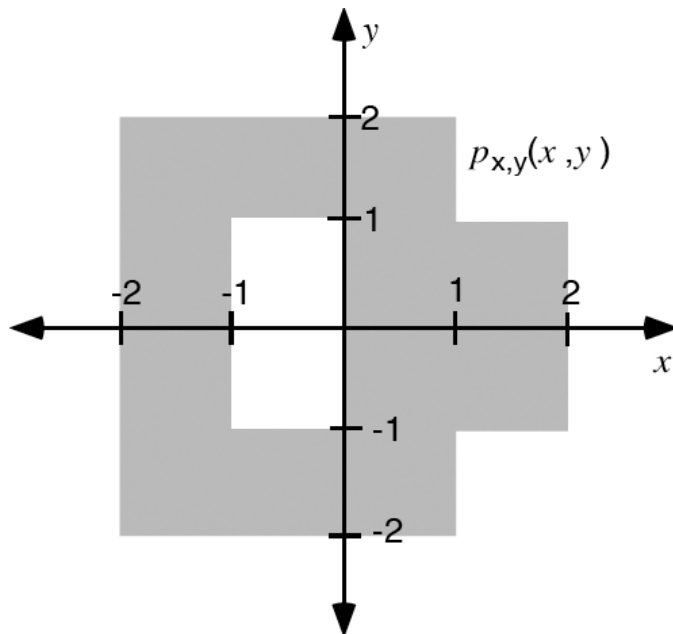


Figure 4: The joint distribution of \mathbf{x} and \mathbf{y}

- a. Let ω_1 be the case when $\mathbf{x} \leq 0$, and ω_2 be the case when $\mathbf{x} > 0$. Determine the *a priori* probabilities of the two classes $P(\omega_1)$ and $P(\omega_2)$. \mathbf{y} is the observation, from which we infer whether ω_1 or ω_0 happens. Find the likelihood functions, namely, the two conditional distributions $p(y|\omega_1)$ and $p(y|\omega_2)$.

By simply counting the number of unit squares in the shaded areas on the left and right sides of the line $x = 0$, we can directly find out that there are 6 unit squares on each side. Thus, the two cases are equally likely, i.e. $P(\omega_1) = P(\omega_2) = 0.5$.

It's also pretty straightforward to obtain the likelihood functions $p(y|\omega_1)$ and $p(y|\omega_2)$ by counting the number of unit squares for different ranges of y . We just need to be careful with normalizing them such that the integral of the distribution is 1.

- b. Find the decision rule that minimizes the probability of error, and calculate what the probability of error is. Please note that there will be ambiguities at decision boundaries, but how you classify when y falls on the decision boundary doesn't affect the probability of error.

As shown in (a), the *a priori* probabilities of ω_1 and ω_2 are identical. The minimum probability of error decision rule simply relies on the comparison of the two likelihood functions. In other words, it becomes a ML decision rule.

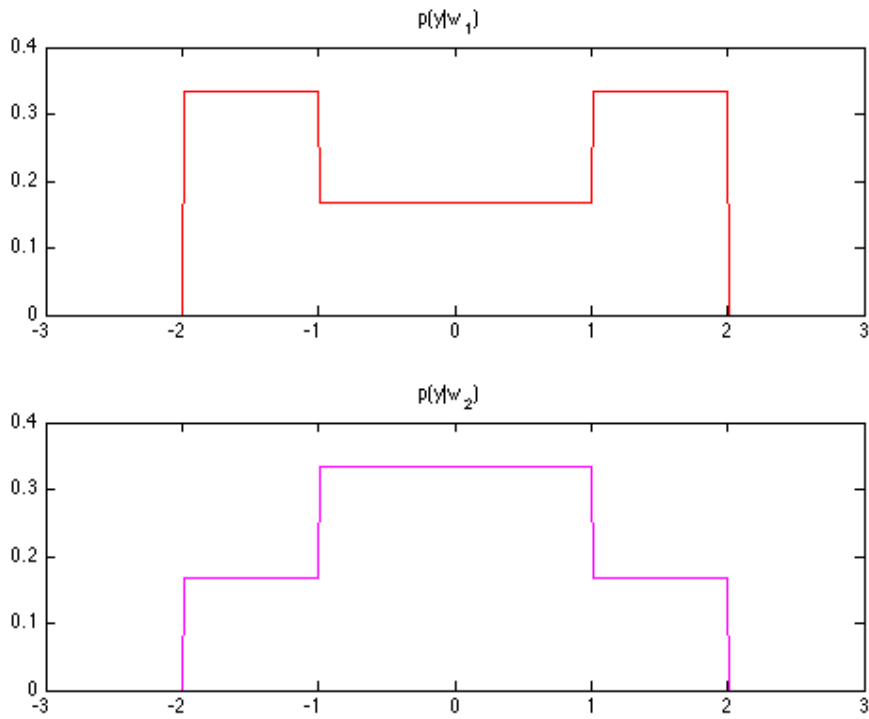


Figure 5: The likelihood functions $p(y|\omega_1)$ and $p(y|\omega_2)$

The decision rule can be summarized as

$$\hat{\omega} = \begin{cases} \omega_1 & \text{if } -2 < y \leq -1 \text{ or } 1 < y \leq 2 \\ \omega_2 & \text{if } -1 < y \leq 1 \end{cases}$$

The probability of error is thus

$$\begin{aligned} \Pr(e) &= \Pr(\hat{\omega} = \omega_1|\omega_2) \Pr(\omega_2) + \Pr(\hat{\omega} = \omega_2|\omega_1) \Pr(\omega_1) \\ &= 1/2 \Pr(-2 < y \leq -1, 1 < y \leq 2|\omega_2) + 1/2 \Pr(-1 < y \leq 1|\omega_1) \\ &= 1/2 * 1/3 + 1/2 * 1/3 = 1/3 \end{aligned} \tag{5}$$