Bayesian Decision Theory

Computer Science- Pattern Recognition

Prof. Dr. Dhahir A. Abdullah

Bayesian Decision Theory

- Bayesian Decision Theory is a **fundamental statistical approach** that quantifies the trade offs between various decisions using **probabilities** and **costs** that accompany such decisions.
- First, we will assume **that all probabilities are known**.
- Then, we will study the cases where the **probabilistic structure is not completely known.**

Fish Sorting Example

- **State of nature** is a random variable.
- Define **w** as the type of fish we observe (state of nature, class) where
  - w = w1 for sea bass,
  - w = w2 for salmon.
  - P(w1) is the **a priori probability** that the next fish is a sea bass.
  - P(w2) is the a priori probability that the next fish is a salmon.

Prior Probabilities

- Prior probabilities reflect our knowledge of how likely each type of fish will appear before we actually see it.
- How can we choose **P(w1)** and **P(w2)?**
  - Set P(w1) = P(w2) if they are equiprobable (**uniform priors**).
  - May use different values depending on the fishing area, time of the year, etc.
- Assume there are no other types of fish
  - P(w1) + P(w2) = 1

Making a Decision

- How can we make a decision with only the prior information?

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1) > P(w_2) \\ w_2 & \text{otherwise} \end{cases}$$

- What is the **probability of error** for this decision?
  - P(error) = min{P(w1), P(w2)}

Making a Decision

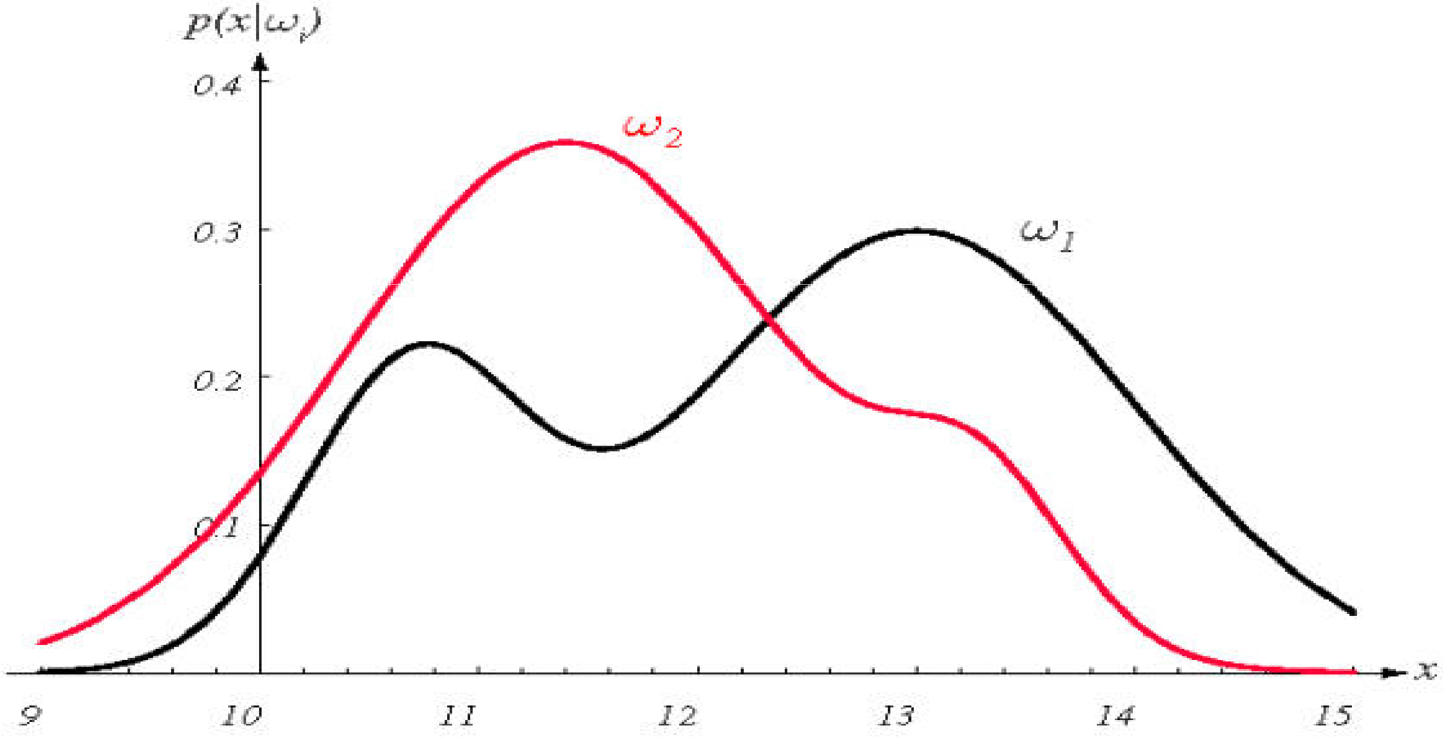
- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2

- **Make further measurement and compute the class –conditional densities**

Class-Conditional Probabilities

- Let’s try to improve the decision using the lightness measurement x.
- Let x be a continuous random variable.
- Define **p(x|wj)** as **the class-conditional probability density** (probability of x given that the state of nature is wj for j = 1, 2).
- **p(x|w1**) and **p(x|w2)** describe the difference in lightness between populations of sea bass and salmon

Class-Conditional Probabilities



Hypothetical class-conditional probability density functions for two classes.

Posterior Probabilities

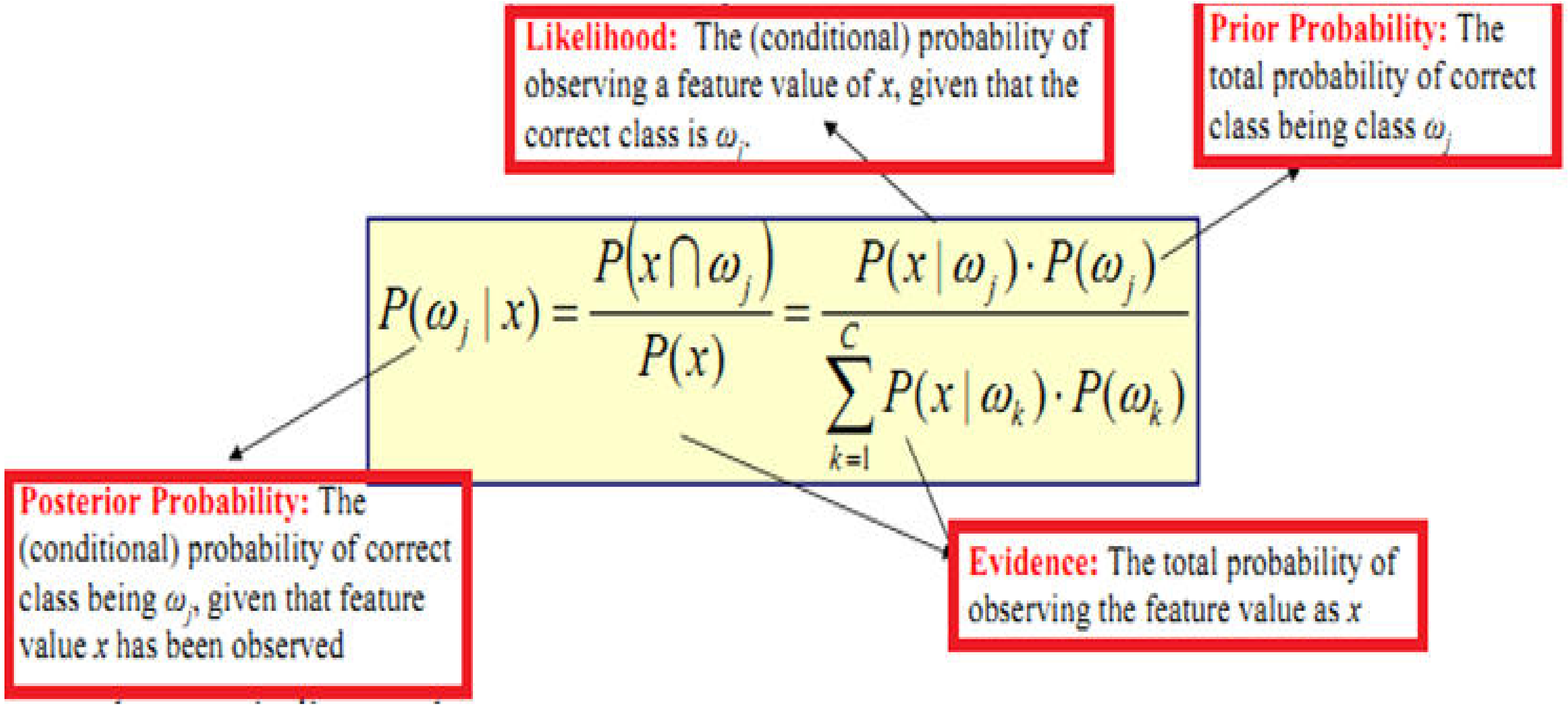
- Suppose we know **P(wj)** and **p(x|wj)** for j = 1, 2, and measure the lightness of a fish as the value x.
- Define **P(wj |x)** as the **a posteriori probability** (probability of the state of nature being wj given the measurement of feature value x).
- We can use the **Bayes** formula to convert the prior probability to the posterior probability

$$P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)}$$

- where

$$p(x) = \sum_{j=1}^{2} p(x|w_j)P(w_j)$$

Posterior Probabilities (remember)



Likelihood: The (conditional) probability of observing a feature value of x, given that the correct class is ω_j.

Prior Probability: The total probability of correct class being class ω_j

$$P(\omega_j \mid x) = \frac{P(x \cap \omega_j)}{P(x)} = \frac{P(x \mid \omega_j) \cdot P(\omega_j)}{\sum_{k=1}^{C} P(x \mid \omega_k) \cdot P(\omega_k)}$$

Posterior Probability: The (conditional) probability of correct class being ω_j, given that feature value x has been observed

Evidence: The total probability of observing the feature value as x

Posterior Probabilities (remember)

Posterior = (Likelihood . Prior) / Evidence

Making a Decision

- p(x|wj) is called the **likelihood** and p(x) is called the **evidence.**
- How can we make a decision after observing the value of x?

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

Making a Decision

- **Decision strategy:** Given the posterior probabilities for each class

  X is an observation for which:

  if $P(\omega_1 / x) > P(\omega_2 / x)$ ⇒ True state of nature = ω_1
  if $P(\omega_1 / x) < P(\omega_2 / x)$ ⇒ True state of nature = ω_2

Making a Decision

- p(x|wj) is called the **likelihood** and p(x) is called the **evidence.**
- How can we make a decision after observing the value of x?

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

- Rewriting the rule gives

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} \\ w_2 & \text{otherwise} \end{cases}$$

- Note that, at every x, **P(w1|x) + P(w2|x) = 1.**

Probability of Error

- What is the probability of error for this decision?

$$P(error|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases}$$

- What is the probability of error?

$$P(error) = \int_{-\infty}^{\infty} p(error, x)\, dx = \int_{-\infty}^{\infty} P(error|x)\, p(x)\, dx$$

Probability of Error

- **Decision strategy for Minimizing the probability of error**
- Decide ω_1 *if* $P(\omega_1 / x) > P(\omega_2 / x)$; otherwise decide ω_2

  Therefore:

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

  (Bayes decision)

Probability of Error

- What is the probability of error for this decision?

$$P(error|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases}$$

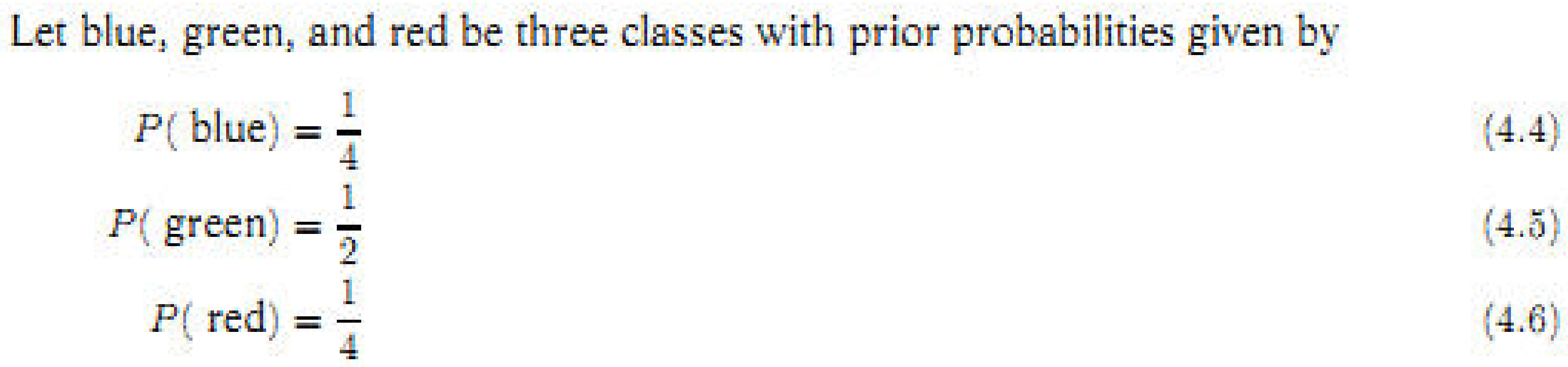
- What is the probability of error?

$$P(error) = \int_{-\infty}^{\infty} p(error, x)\, dx = \int_{-\infty}^{\infty} P(error|x)\, p(x)\, dx$$

- **Bayes decision rule** minimizes this error because

$$P(error|x) = \min\{P(w_1|x), P(w_2|x)\}$$

Example



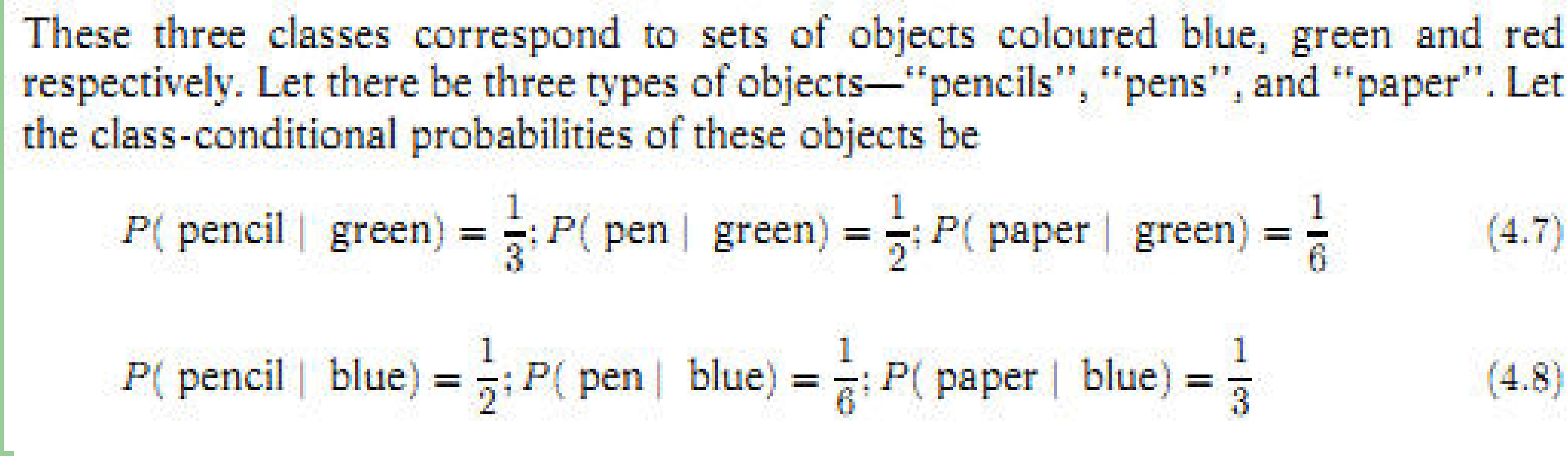
Let blue, green, and red be three classes with prior probabilities given by

$$P(\text{ blue}) = \frac{1}{4} \tag{4.4}$$

$$P(\text{ green}) = \frac{1}{2} \tag{4.5}$$

$$P(\text{ red}) = \frac{1}{4} \tag{4.6}$$

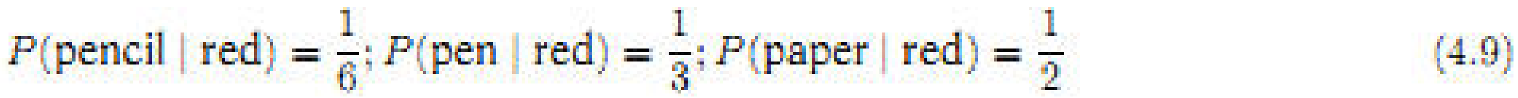
Example (cont.)



These three classes correspond to sets of objects coloured blue, green and red respectively. Let there be three types of objects—"pencils", "pens", and "paper". Let the class-conditional probabilities of these objects be

$$P(\text{ pencil} \mid \text{ green}) = \frac{1}{3}; P(\text{ pen} \mid \text{ green}) = \frac{1}{2}; P(\text{ paper} \mid \text{ green}) = \frac{1}{6} \tag{4.7}$$

$$P(\text{ pencil} \mid \text{ blue}) = \frac{1}{2}; P(\text{ pen} \mid \text{ blue}) = \frac{1}{6}; P(\text{ paper} \mid \text{ blue}) = \frac{1}{3} \tag{4.8}$$

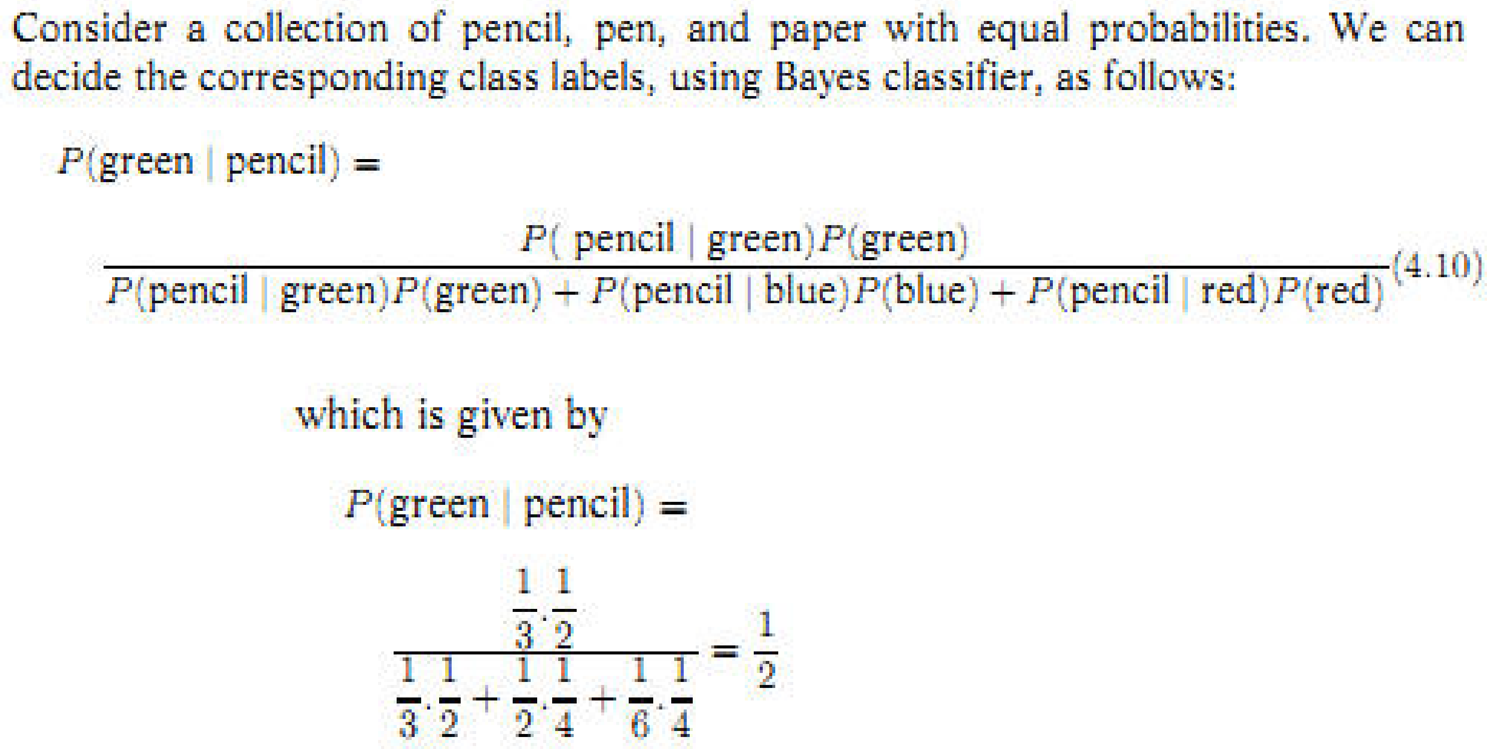


$$P(\text{pencil} \mid \text{red}) = \frac{1}{6}; P(\text{pen} \mid \text{red}) = \frac{1}{3}; P(\text{paper} \mid \text{red}) = \frac{1}{2} \tag{4.9}$$

Example (cont.)

Assign colours to objects.

Example (cont.)



Consider a collection of pencil, pen, and paper with equal probabilities. We can decide the corresponding class labels, using Bayes classifier, as follows:

$$P(\text{green} \mid \text{pencil}) =$$

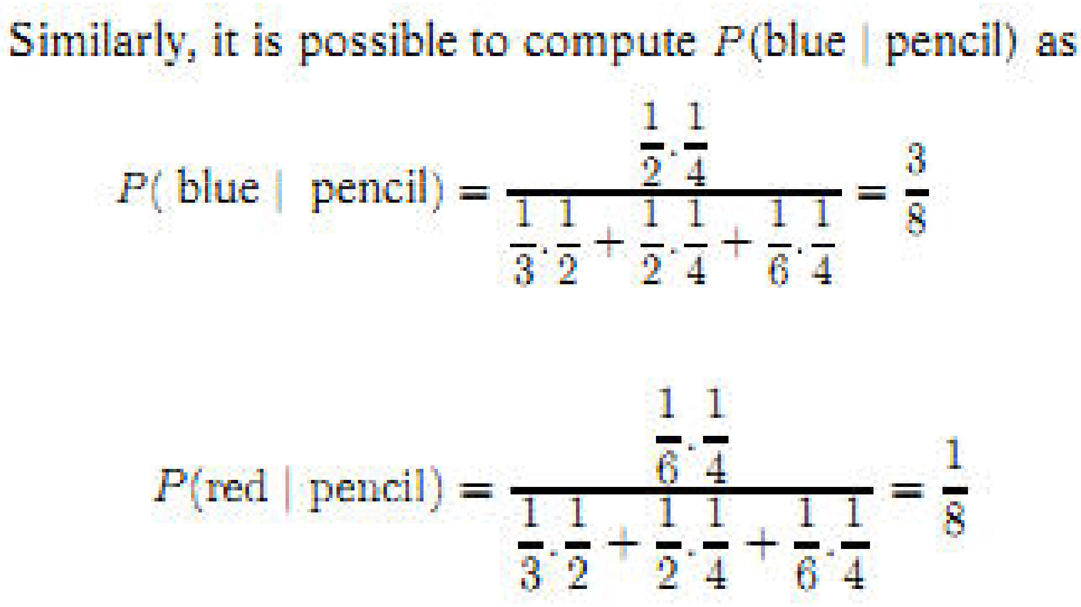
$$\frac{P(\text{ pencil} \mid \text{green})P(\text{green})}{P(\text{pencil} \mid \text{green})P(\text{green}) + P(\text{pencil} \mid \text{blue})P(\text{blue}) + P(\text{pencil} \mid \text{red})P(\text{red})} \quad (4.10)$$

which is given by

$$P(\text{green} \mid \text{pencil}) =$$

$$\frac{\frac{1}{3}.\frac{1}{2}}{\frac{1}{3}.\frac{1}{2} + \frac{1}{2}.\frac{1}{4} + \frac{1}{6}.\frac{1}{4}} = \frac{1}{2}$$

Example (cont.)

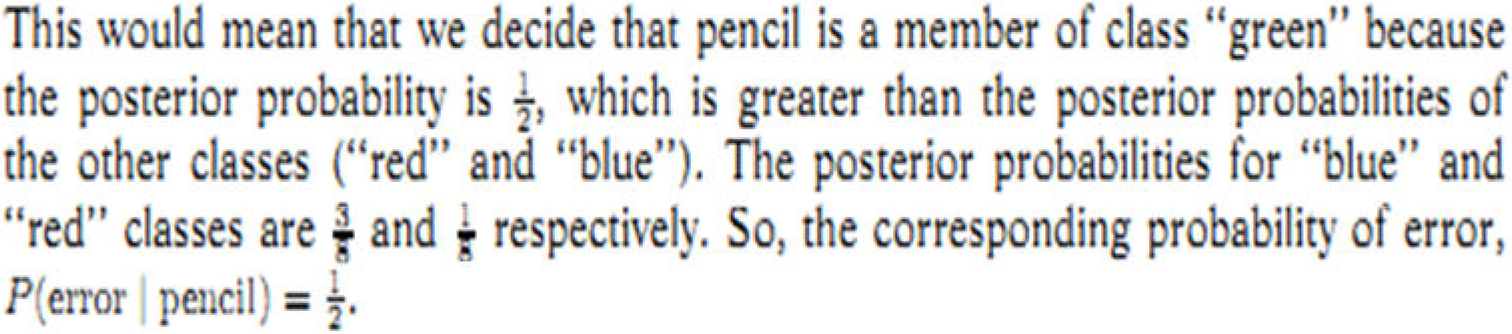


Similarly, it is possible to compute $P(\text{blue} \mid \text{pencil})$ as

$$P(\text{ blue} \mid \text{ pencil}) = \frac{\frac{1}{2}\cdot\frac{1}{4}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{4}+\frac{1}{6}\cdot\frac{1}{4}} = \frac{3}{8}$$

$$P(\text{red} \mid \text{pencil}) = \frac{\frac{1}{6}\cdot\frac{1}{4}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{4}+\frac{1}{6}\cdot\frac{1}{4}} = \frac{1}{8}$$

Example (cont.)



This would mean that we decide that pencil is a member of class "green" because the posterior probability is $\frac{1}{2}$, which is greater than the posterior probabilities of the other classes ("red" and "blue"). The posterior probabilities for "blue" and "red" classes are $\frac{3}{8}$ and $\frac{1}{8}$ respectively. So, the corresponding probability of error, $P(\text{error} \mid \text{pencil}) = \frac{1}{2}$.

$$P(\text{red} \mid \text{pencil}) = \frac{1}{8}$$

$$P(\text{green} \mid \text{pencil}) = \frac{1}{2}$$

$$P(\text{blue} \mid \text{pencil}) = \frac{3}{8}$$

Example (cont.)

Assign colour to pen objects.

$$P(\text{ blue}) = \frac{1}{4}$$

$$P(\text{ green}) = \frac{1}{2}$$

$$P(\text{ red}) = \frac{1}{4}$$

$$P(\text{ pencil} \mid \text{ green}) = \frac{1}{3}; P(\text{ pen} \mid \text{ green}) = \frac{1}{2}; P(\text{ paper} \mid \text{ green}) = \frac{1}{6}$$

$$P(\text{ pencil} \mid \text{ blue}) = \frac{1}{2}; P(\text{ pen} \mid \text{ blue}) = \frac{1}{6}; P(\text{ paper} \mid \text{ blue}) = \frac{1}{3}$$

$$P(\text{pencil} \mid \text{red}) = \frac{1}{6}; P(\text{pen} \mid \text{red}) = \frac{1}{3}; P(\text{paper} \mid \text{red}) = \frac{1}{2}$$

Example (cont.)

In a similar manner, for pen, the posterior probabilities are

$$P(\text{green} \mid \text{pen}) = \frac{2}{3}; P(\text{blue} \mid \text{pen}) = \frac{1}{9}; P(\text{red} \mid \text{pen}) = \frac{2}{9} \tag{4.14}$$

This enables us to decide that pen belongs to class "green" and $P(\text{error} \mid \text{pen}) = \frac{1}{3}$.

Example (cont.)

Assign colour to paper objects.

$$P(\text{ blue}) = \frac{1}{4}$$

$$P(\text{ green}) = \frac{1}{2}$$

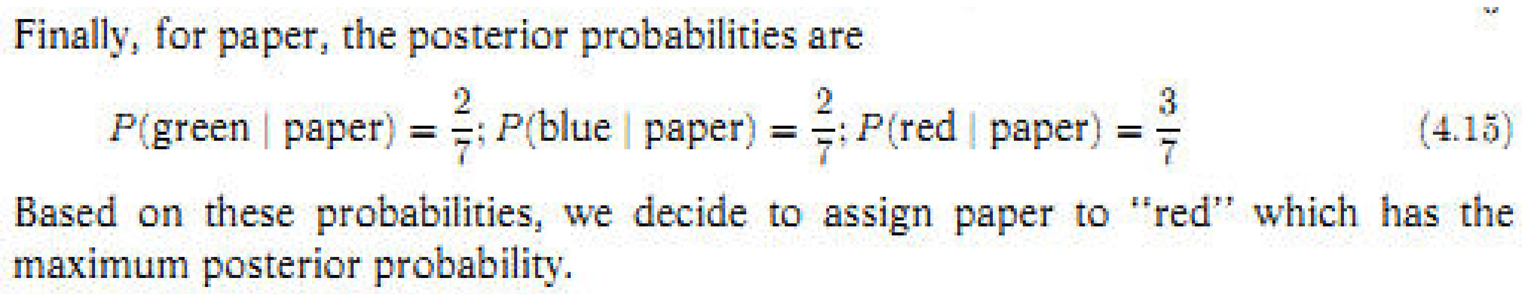
$$P(\text{ red}) = \frac{1}{4}$$

$$P(\text{ pencil} \mid \text{ green}) = \frac{1}{3}; P(\text{ pen} \mid \text{ green}) = \frac{1}{2}; P(\text{ paper} \mid \text{ green}) = \frac{1}{6}$$

$$P(\text{ pencil} \mid \text{ blue}) = \frac{1}{2}; P(\text{ pen} \mid \text{ blue}) = \frac{1}{6}; P(\text{ paper} \mid \text{ blue}) = \frac{1}{3}$$

$$P(\text{pencil} \mid \text{red}) = \frac{1}{6}; P(\text{pen} \mid \text{red}) = \frac{1}{3}; P(\text{paper} \mid \text{red}) = \frac{1}{2}$$

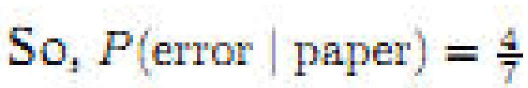
Example (cont.)



Finally, for paper, the posterior probabilities are

$$P(\text{green} \mid \text{paper}) = \frac{2}{7}; P(\text{blue} \mid \text{paper}) = \frac{2}{7}; P(\text{red} \mid \text{paper}) = \frac{3}{7} \tag{4.15}$$

Based on these probabilities, we decide to assign paper to "red" which has the maximum posterior probability.



So, $P(\text{error} \mid \text{paper}) = \frac{4}{7}$

Example (cont.)

Average probability of error =

$$P(\text{error} \mid \text{pencil}) \times \frac{1}{3} + P(\text{error} \mid \text{pen}) \times \frac{1}{3} + P(\text{error} \mid \text{paper}) \times \frac{1}{3} \qquad (4.16)$$

As a consequence, its value is

$$\text{Average probability of error} = \frac{1}{2}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{4}{7} = \frac{59}{126} \qquad (4.17)$$

Bayesian Decision Theory

How can we generalize to:

- more than one feature?
  - replace the scalar x by the feature vector x
- more than two states of nature?
  - just a difference in notation
- allowing actions other than just decisions?
  - allow the possibility of rejection
- different risks in the decision?
  - define how costly each action is

Bayesian Decision Theory

- Let **{w1, . . . ,wc}** be the finite set of **c** states of nature (classes, categories).
- Let **{α1, . . . , αa}** be the finite set of a possible actions.
- Let **λ(αi|wj)** be the loss incurred for taking action **αi** when the state of nature is **wj** .
- Let **x** be the **d-component** vector-valued random variable called the feature vector .

Bayesian Decision Theory

- p(x|wj) is the class-conditional probability density function.
- P(wj) is the prior probability that nature is in state wj .
- The posterior probability can be computed as

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

- where

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|w_j)P(w_j)$$

Loss function

- Allow actions and not only decide on the state of nature. How costly an action is?
- Introduce a loss function which is more general than the probability of error
- The loss function states how costly each action taken is
- Allowing actions other than classification primarily allows the possibility of rejection
- Refusing to make a decision in close or bad cases

Loss function

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or “categories”)

Let , $\alpha(x)$ *maps a pattern x into one of the actions from* $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$, the set of possible actions

Let, $\lambda(\alpha_i / \omega_j)$ be the loss incurred for taking action α_i when the category is ω_j

Conditional Risk

- Suppose we observe **x** and take action **αi**.
- If the true state of nature is **wj** , we incur the loss **λ(αi|wj).**
- The expected loss with taking action αi is

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|\mathbf{x})$$

which is also called the conditional risk.

Ex. Target Detection

Actual class	ω_1 *(var)*	ω_2 *(yok)*
Choose ω_1	$\lambda(\alpha_1 / \omega 1)$ hit	$\lambda(\alpha_1 / \omega 2)$ false alarm
Choose ω_2	$\lambda(\alpha_2 / \omega_1)$ miss	$\lambda(\alpha_2 / \omega 2)$ do nothing

Minimum-Risk Classification

- The general decision rule **α(x)** tells us which action to take for observation **x**.
- We want to find the decision rule that minimizes the overall risk

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

- Bayes decision rule minimizes the overall risk by selecting the action **αi** for which **R(αi|x)** is **minimum**.
- The resulting minimum overall risk is called the Bayes risk and is the best performance that can be achieved.

Two-Category Classification

- Define

  α_1: deciding w_1,

  α_2: deciding w_2,

  $\lambda_{ij} = \lambda(\alpha_i | w_j)$.

- Conditional risks can be written as

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}\, P(w_1|\mathbf{x}) + \lambda_{12}\, P(w_2|\mathbf{x}),$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21}\, P(w_1|\mathbf{x}) + \lambda_{22}\, P(w_2|\mathbf{x}).$$

Two-Category Classification

- The minimum-risk decision rule becomes

$$\text{Decide} \begin{cases} w_1 & \text{if } (\lambda_{21} - \lambda_{11})P(w_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(w_2|\mathbf{x}) \\ w_2 & \text{otherwise} \end{cases}$$

- This corresponds to deciding **w1** if

$$\frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(w_2)}{P(w_1)}$$

⇒ comparing the **likelihood ratio** to a threshold that is independent of the observation x.

Optimal decision property

“If the likelihood ratio exceeds a threshold value T, independent of the input pattern x, we can take optimal actions”

Minimum-Error-Rate Classification

- Actions are decisions on classes (**αi** is deciding **wi**).
- If action **αi** is taken and the true state of nature is **wj** , then the decision is correct if i = j and in error if i ≠ j.
- We want to find a decision rule that minimizes the probability of error

Minimum-Error-Rate Classification

- Define the zero-one loss function

$$\lambda(\alpha_i|w_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

(all errors are equally costly).

- I Conditional risk becomes

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|w_j)\, P(w_j|\mathbf{x}) \\ &= \sum_{j\neq i} P(w_j|\mathbf{x}) \\ &= 1 - P(w_i|\mathbf{x}). \end{aligned}$$

Minimum-Error-Rate Classification

- Minimizing the risk requires maximizing $P(wi|x)$ and results in the minimum-error decision rule
  - **Decide wi if $P(wi|x) > P(wj\ |x)\ \forall j = i$.**

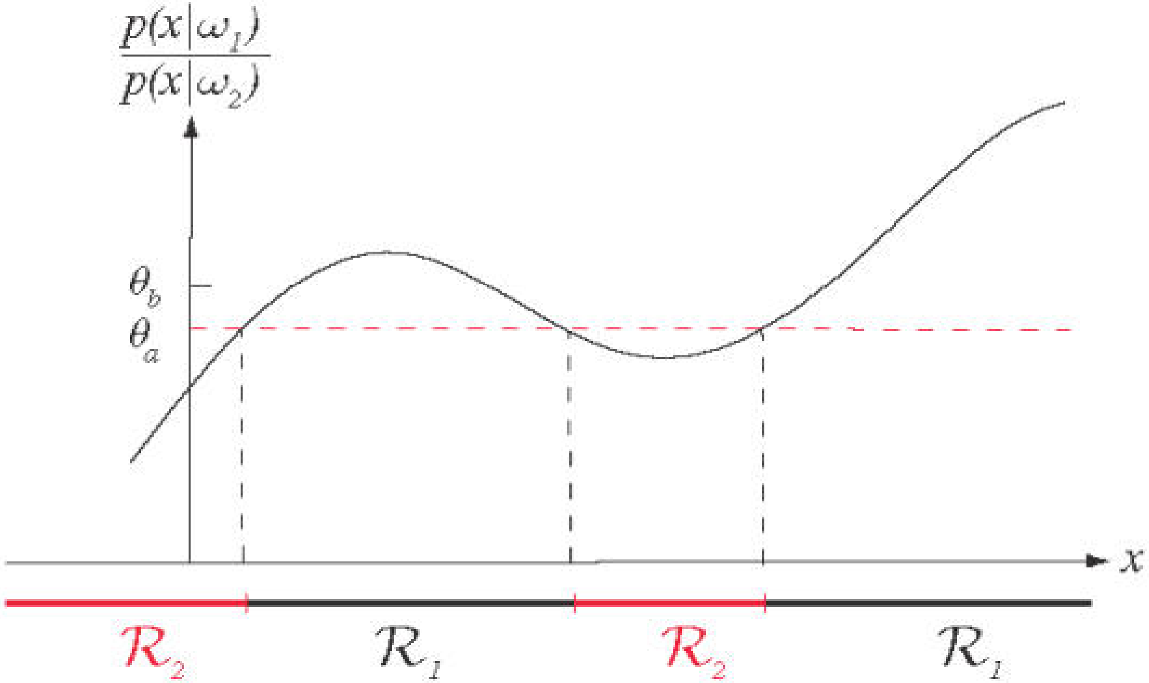
- The resulting error is called the Bayes error and is the best performance that can be achieved.

Minimum-Error-Rate Classification

Regions of decision and zero-one loss function, therefore:

$$Let\ \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}.\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda\ then\ decide\ \omega_1\ \ if:\ \frac{P(x\,|\,\omega_1)}{P(x\,|\,\omega_2)}>\theta_\lambda$$

Minimum-Error-Rate Classification



The likelihood ratio $p(\mathbf{x}|w_1)/p(\mathbf{x}|w_2)$. The threshold θ_a is computed using the priors $P(w_1) = 2/3$ and $P(w_2) = 1/3$, and a zero-one loss function. If we penalize mistakes in classifying w_2 patterns as w_1 more than the converse, we should increase the threshold to θ_b.

Discriminant Functions

- A useful way of representing classifiers is through discriminant functions **gi(x),** i = 1, . . . , c, where the classifier assigns a feature vector x to class wi if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$$

- For the classifier that minimizes conditional risk

$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$$

- For the classifier that minimizes error

$$g_i(\mathbf{x}) = P(w_i|\mathbf{x})$$

Discriminant Functions

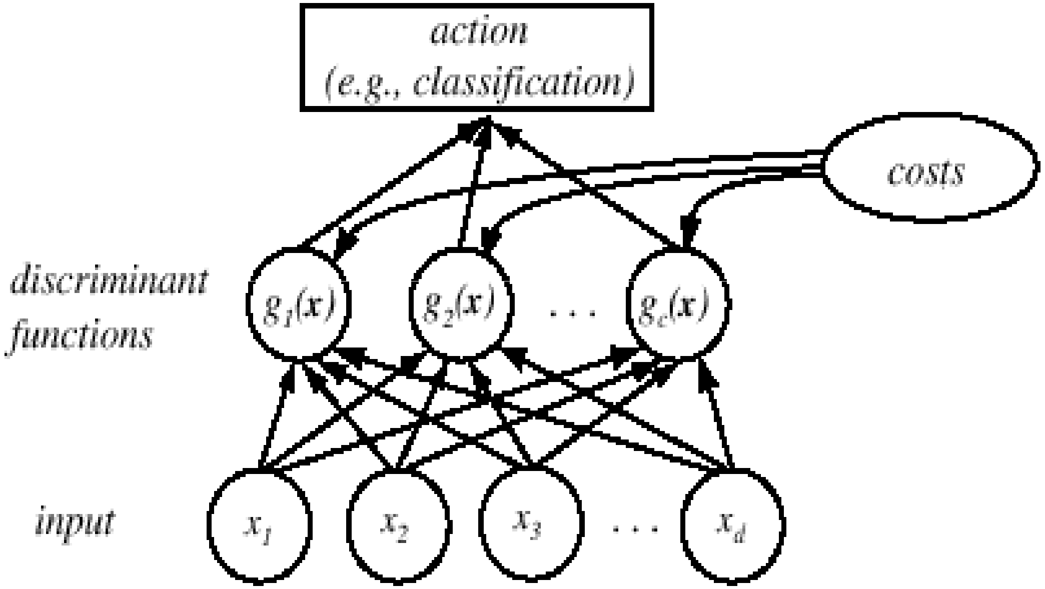
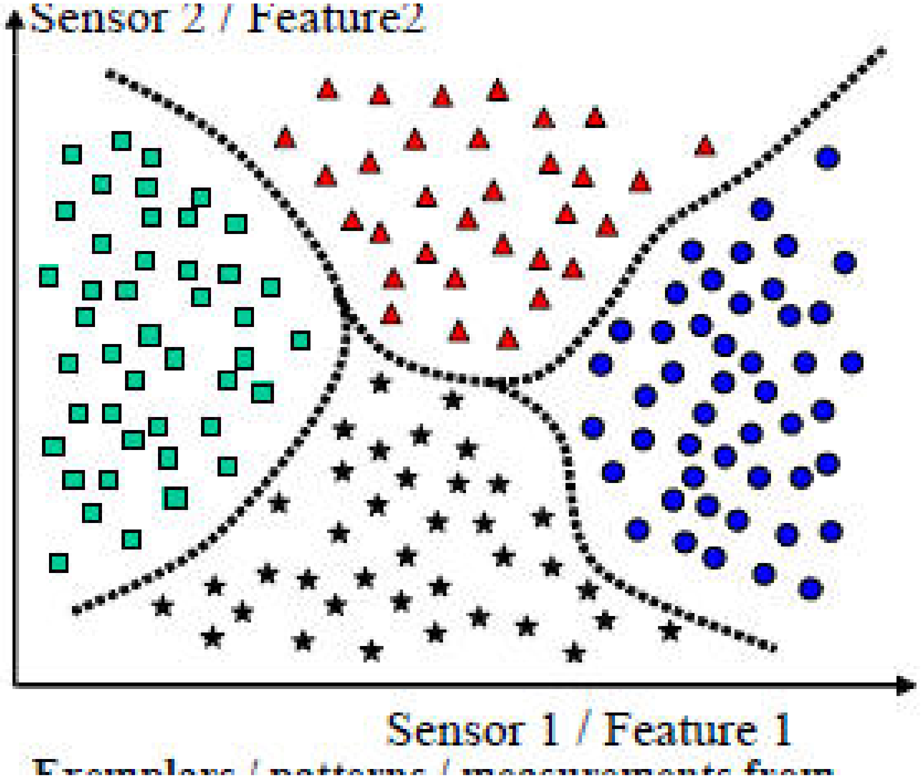


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions

- These functions divide the feature space into **c** decision regions **(R1, . . . , Rc)**, separated by decision boundaries.



Discriminant Functions

$g_i(x)$* can be any monotonically increasing function of *$P(\omega_i / x)$

$g_i(x) \equiv f\,(P(\omega_i / x)) = P(x / \omega_i)\, P(\omega_i)$

or natural logarithm of any function of $P(\omega_i / x)$

$g_i(x) = \ln P(x / \omega_i) + \ln P(\omega_i)$

Discriminant Functions

- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions g_1 and g_2

    Let $g(x) \equiv g_1(x) - g_2(x)$

Decide ω_1 *if* $g(x) > 0$; Otherwise decide ω_2

Discriminant Functions

- The two-category case
  - The computation of g(x)

$$g(x) = P(\omega_1 | x) - P(\omega_2 | x)$$

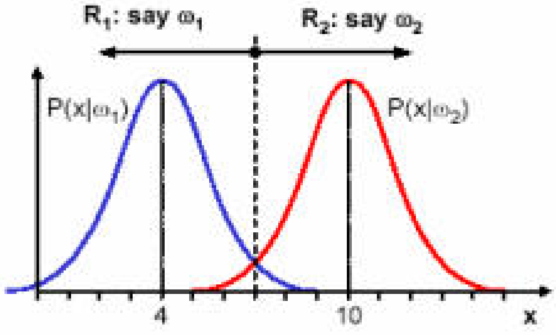
$$= ln \frac{P(x | \omega_1)}{P(x | \omega_2)} + ln \frac{P(\omega_1)}{P(\omega_2)}$$

Example

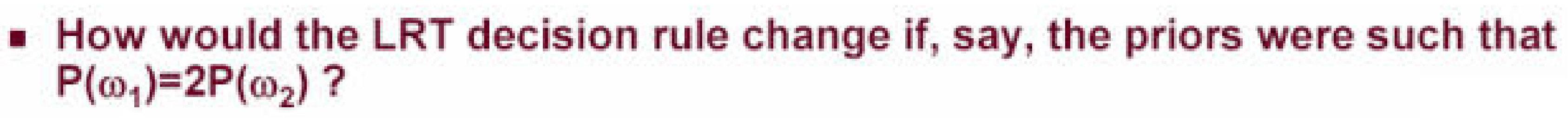
- **Given a classification problem with the following class conditional densities, derive a decision rule based on the Likelihood Ratio Test (assume <u>equal priors</u>)**

$$P(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2} \qquad P(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}$$

- **Solution**
  - Substituting the given likelihoods and priors into the LRT expression: $\Lambda(x) = \dfrac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \dfrac{1}{1}$
  - Simplifying the LRT expression: $\Lambda(x) = \dfrac{e^{-\frac{1}{2}(x-4)^2}}{e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$
  - Changing signs and taking logs: $(x-4)^2 - (x-10)^2 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0$
  - Which yields: $x \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 7$
  - This LRT result makes sense from an intuitive point of view since the likelihoods are identical and differ only in their mean value



Exercise



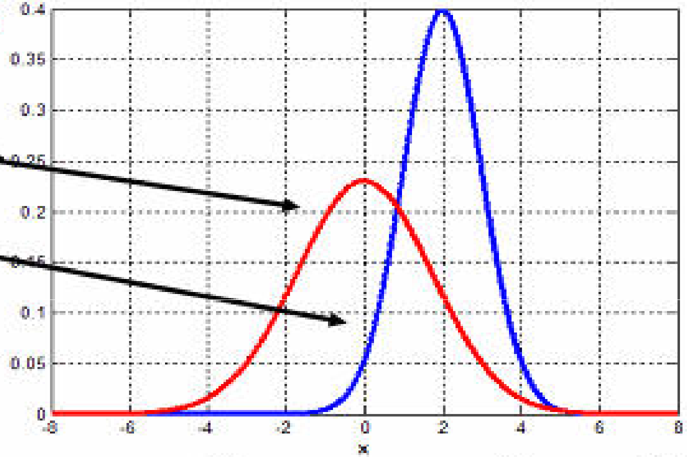
- How would the LRT decision rule change if, say, the priors were such that $P(\omega_1)=2P(\omega_2)$?

Example

- Consider a classification problem with two classes defined by the following likelihood functions

$$P(x\,|\,\omega_1)=\frac{1}{\sqrt{2\pi}\sqrt{3}}e^{-\frac{1}{2}\frac{x^2}{3}}$$

$$P(x\,|\,\omega_2)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-2)^2}$$



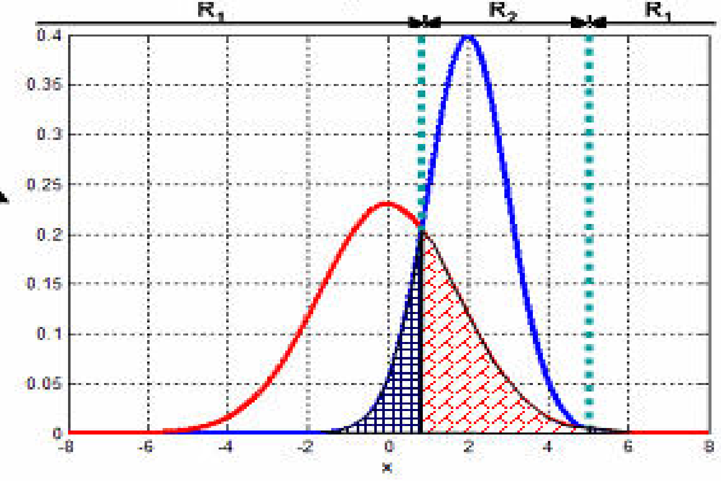
  - Sketch the two densities
  - What is the likelihood ratio?
  - Assume $P[\omega_1]=P[\omega_2]=0.5$, $\lambda_{11}=\lambda_{22}=0$, $\lambda_{12}=1$ and $\lambda_{21}=3^{1/2}$. Determine a decision rule that minimizes the probability of error

$$\Lambda(x)=\frac{\dfrac{1}{\sqrt{2\pi}\sqrt{3}}e^{-\frac{1}{2}\frac{x^2}{3}}}{\dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-2)^2}}\underset{\omega_2}{\overset{\omega_1}{\gtrless}}\frac{1}{\sqrt{3}}$$

$$\frac{e^{-\frac{1}{2}\frac{x^2}{3}}}{e^{-\frac{1}{2}(x-2)^2}}\underset{\omega_2}{\overset{\omega_1}{\gtrless}}1$$

$$-\frac{1}{2}\frac{x^2}{3}+\frac{1}{2}(x-2)^2\underset{\omega_2}{\overset{\omega_1}{\gtrless}}0$$

$$2x^2-12x+12\underset{\omega_2}{\overset{\omega_1}{\gtrless}}0\Rightarrow x=4.73,1.27$$



Exercise

Select the optimal decision where:

$\Omega = \{\omega_1, \omega_2\}$

$P(x / \omega_1) \longrightarrow$ N(2, 0.5) (Normal distribution)

$P(x / \omega_2) \longrightarrow$ N(1.5, 0.2)

$P(\omega_1) = 2/3$

$P(\omega_2) = 1/3$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

The Gaussian Density

- Gaussian can be considered as a model where the feature vectors for a given class are continuous-valued, randomly corrupted versions of a single typical or prototype vector.
- Some proper ties of the Gaussian:
- Analytically tractable.
- Completely specified by the 1st and 2nd moments.
- Has the maximum entropy of all distributions with a given mean and variance.
- Many processes are asymptotically Gaussian (Central Limit Theorem).
- Linear transfor mations of a Gaussian are also Gaussian.

Univariate Gaussian

For $x \in \mathbb{R}$:

$$\begin{aligned} p(x) &= N(\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \end{aligned}$$

where

$$\begin{aligned} \mu &= E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx, \\ \sigma^2 &= E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, p(x)\, dx. \end{aligned}$$

Univariate Gaussian

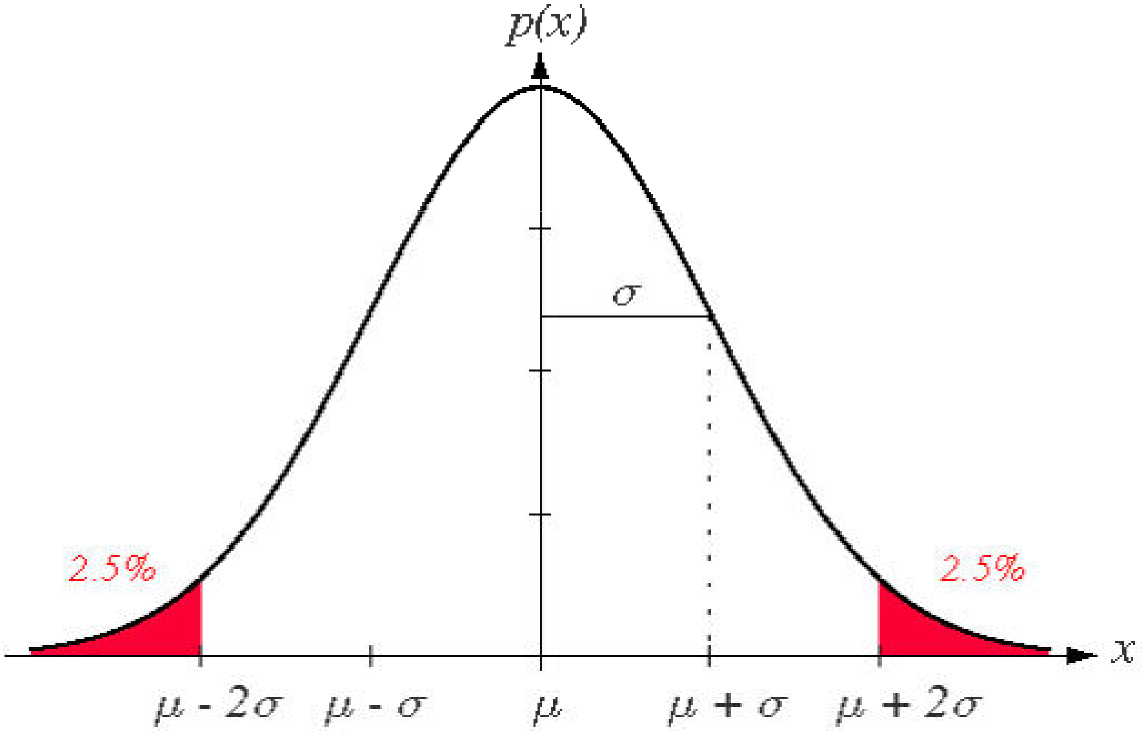


Figure 3: A univariate Gaussian distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$.

Multivariate Gaussian

For $\mathbf{x} \in \mathbb{R}^d$:

$$\begin{aligned} p(\mathbf{x}) &= N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \end{aligned}$$

where

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x},$$

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\, p(\mathbf{x})\, d\mathbf{x}.$$

Linear Transformations

- Recall that, given $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{d \times k}$, $\mathbf{y} = \mathbf{A}^T\mathbf{x} \in \mathbb{R}^k$, if $x \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $y \sim N(\mathbf{A}^T\boldsymbol{\mu}, \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A})$.
- As a special case, the *whitening transform*

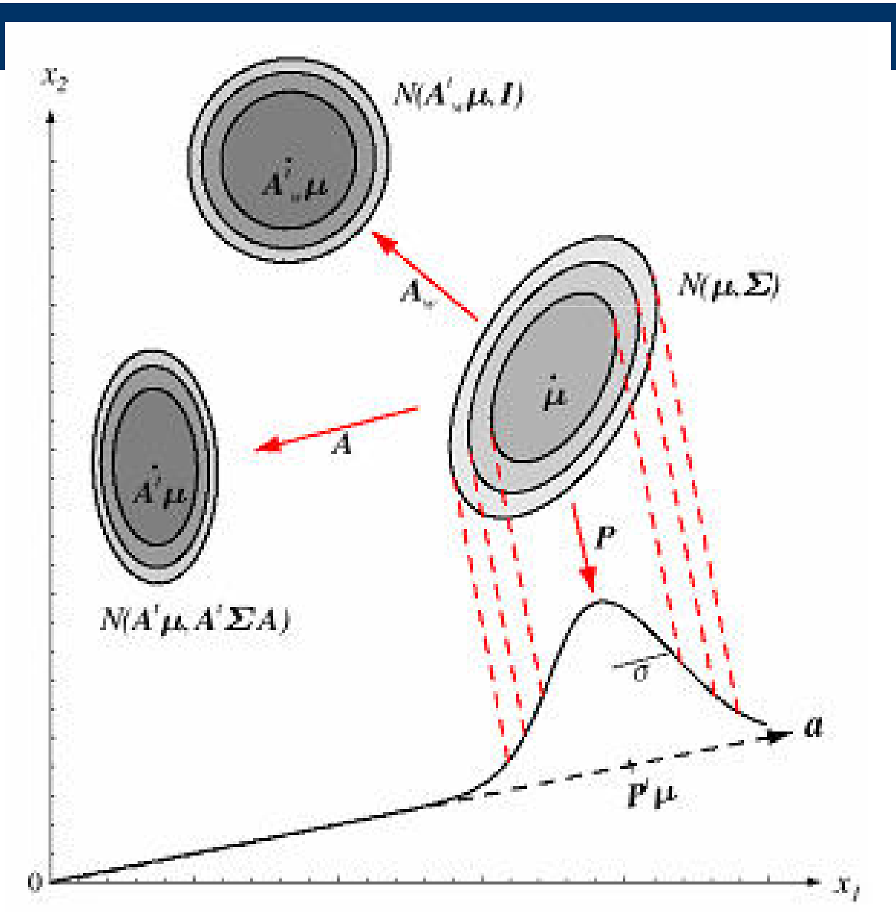
$$\mathbf{A}_\mathrm{w} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$$

where

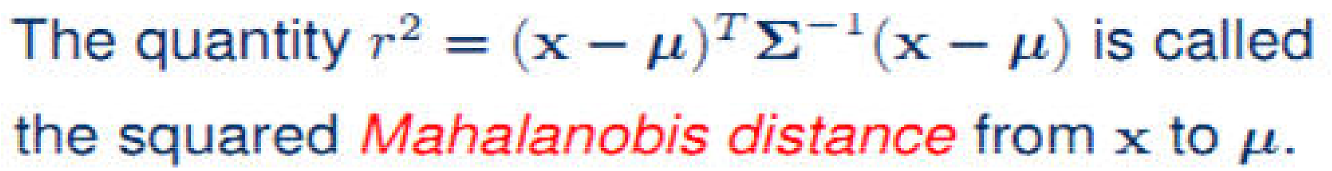
- $\boldsymbol{\Phi}$ is the matrix whose columns are the orthonormal eigenvectors of $\boldsymbol{\Sigma}$,
- $\boldsymbol{\Lambda}$ is the diagonal matrix of the corresponding eigenvalues,

gives a covariance matrix equal to the identity matrix $\mathbf{I}$.

Linear Transformations



Mahalanobis Distance



The quantity $r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is called the squared *Mahalanobis distance* from $\mathbf{x}$ to $\boldsymbol{\mu}$.

Mahalanobis distance takes into account the covariance among the the variables in calculating distance.

Mahalanobis Distance

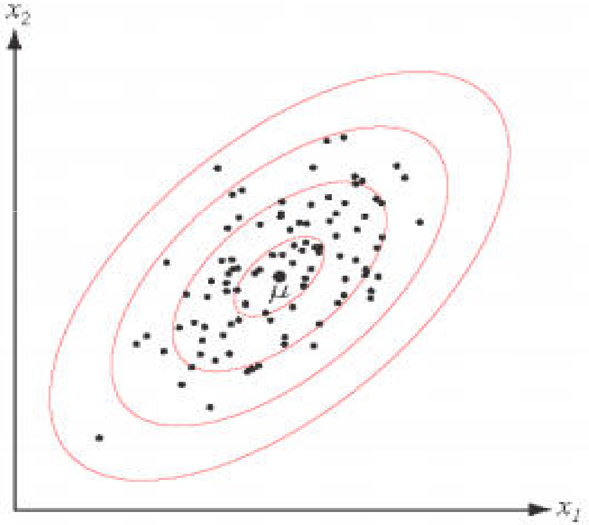


Figure 4: Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The loci of points of constant density are the ellipses for which $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is constant, where the eigenvectors of $\boldsymbol{\Sigma}$ determine the direction and the corresponding eigenvalues determine the length of the principal axes. The quantity $r^2=(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is called the squared *Mahalanobis distance* from $\mathbf{x}$ to $\boldsymbol{\mu}$.

Discriminant Functions for the Gaussian Density

Assume that class conditional density $p(x / \omega_i)$ is multivariate normal

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

Discriminant Functions for the Gaussian Density

- Discriminant functions for minimum-error-rate classification can be written as

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|w_i) + \ln P(w_i).$$

- For $p(\mathbf{x}|w_i) = N(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i^{-1}}(\mathbf{x} - \boldsymbol{\mu_i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma_i}| + \ln P(w_i).$$

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

- the simplest case,
- the features are statistically independent,
- each feature has the same variance.

Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

- the determinant of the ∑:

  $\det \Sigma = \sigma^{2d}$

- Because:

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \quad \det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii}$$

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

- the inverse of of the ∑:

$$\Sigma^{-1} = (1/\sigma^2)\ \mathbf{I}$$

- Because:

$$A = \begin{pmatrix} a_{11} & 0 & \cdots 0 \\ 0 & a_{22} & \cdots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \cdots a_{nn} \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} 1/a_{11} & 0 & \cdots 0 \\ 0 & 1/a_{22} & \cdots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \cdots 1/a_{nn} \end{pmatrix}$$

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

- by using:

det $\sum = \sigma^{2d}$ $\sum^{-1} = (1/\sigma^2)$ **I**

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln|\Sigma_i| + \ln P(\omega_i)$$

$$\Downarrow$$

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}{2\sigma^2} + \ln P(\omega_i) \Rightarrow g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}\|^2}{2\sigma^2} + \ln P(\omega_i)$$

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}{2\sigma^2} + \ln P(\omega_i)$$

$$\Downarrow$$

$$g_i(x) = -\frac{1}{2\sigma^2}[x^t x - 2\mu_i{}^t x + \mu_i^t \mu_i \ ] + lnP(w_i)$$

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

$$g_i(x) = -\frac{1}{2\sigma^2}\left[x^t x - 2\mu_i{}^t x + \mu_i^t \mu_i \right] + lnP(w_i)$$

- The quadratic term is same for all functions, so we can omit the quadratic term.

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

Discriminant functions are

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad \text{(linear discriminant)}$$

where

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \ln P(w_i)$$

(w_{i0} is the threshold or bias for the i'th category).

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

Decision boundaries are the hyperplanes $g_i(\mathbf{x}) = g_j(\mathbf{x})$, and can be written as

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

where

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(w_i)}{P(w_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ passes through the point $\mathbf{x}_0$ and is orthogonal to the vector $\mathbf{w}$.

Case 1: $\Sigma_i = \sigma^2 I$

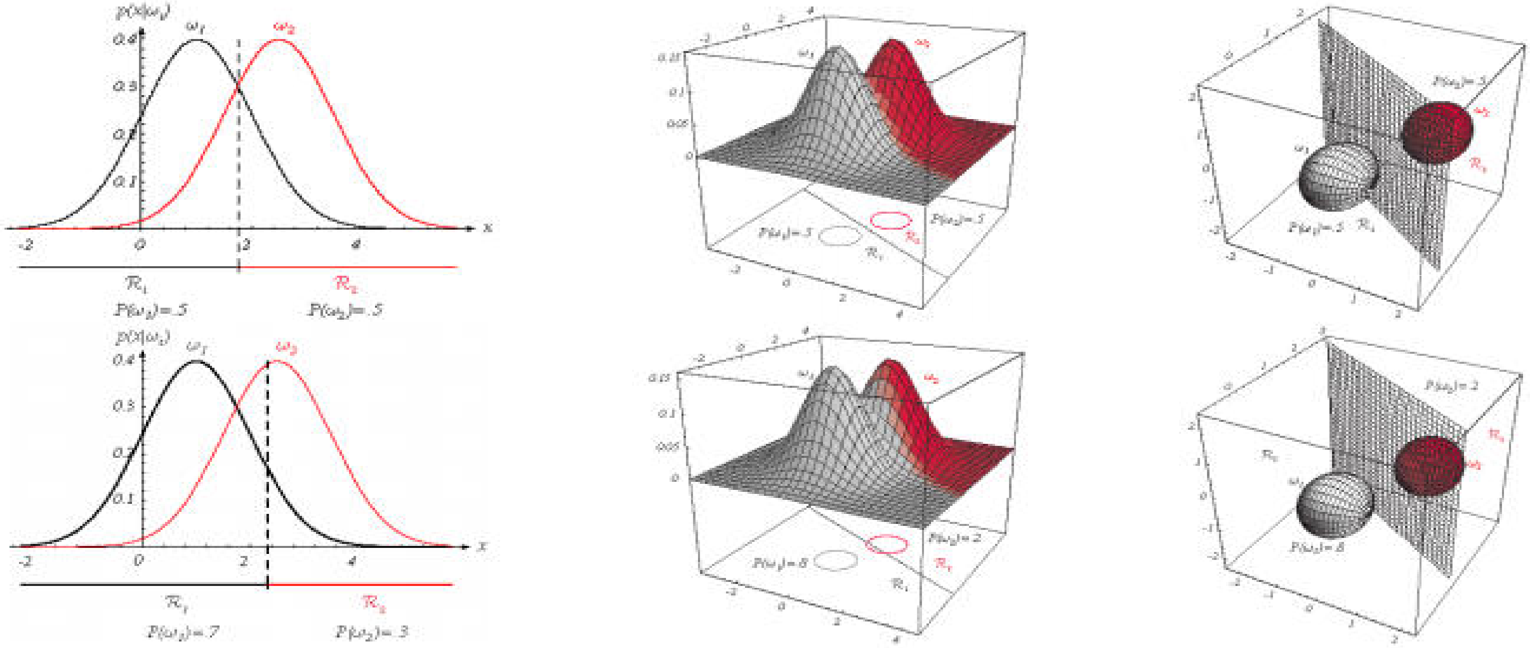


Figure 5: If the covariance matrices of two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. The decision boundary shifts as the priors are changed.

Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

Special case when $P(w_i)$ are the same for $i = 1, \ldots, c$ is the *minimum-distance classifier* that uses the decision rule

$$\text{assign } \mathbf{x} \text{ to } w_{i^*} \text{ where } i^* = \arg \min_{i=1,\ldots,c} \|\mathbf{x} - \boldsymbol{\mu}_i\|.$$