Machine Learning and Computational Intelligence
Lecture 6

Sanjeeb Prasad Panday, PhD
Associate Professor
Dept. of Electronics and Computer Engineering
Director (ICTC)
IOE, TU

Bayesian Decision Theory

Chapter 2 (Duda et al.) – Sections 2.1-2.10
& Case Studies

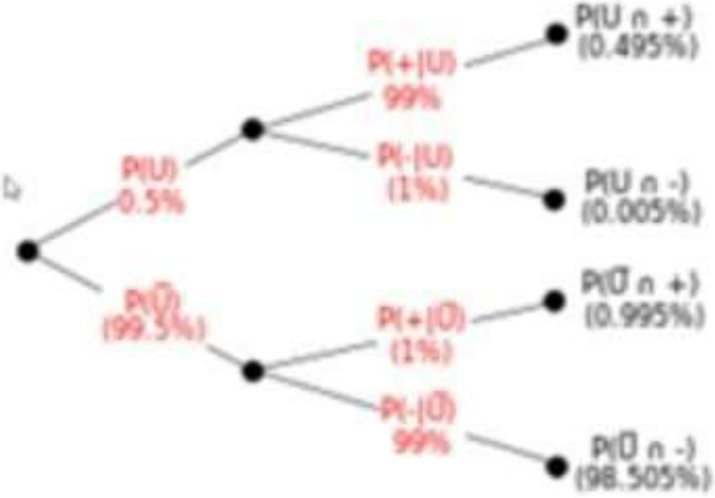
CS479/679 Pattern Recognition
Dr. George Bebis

Probability

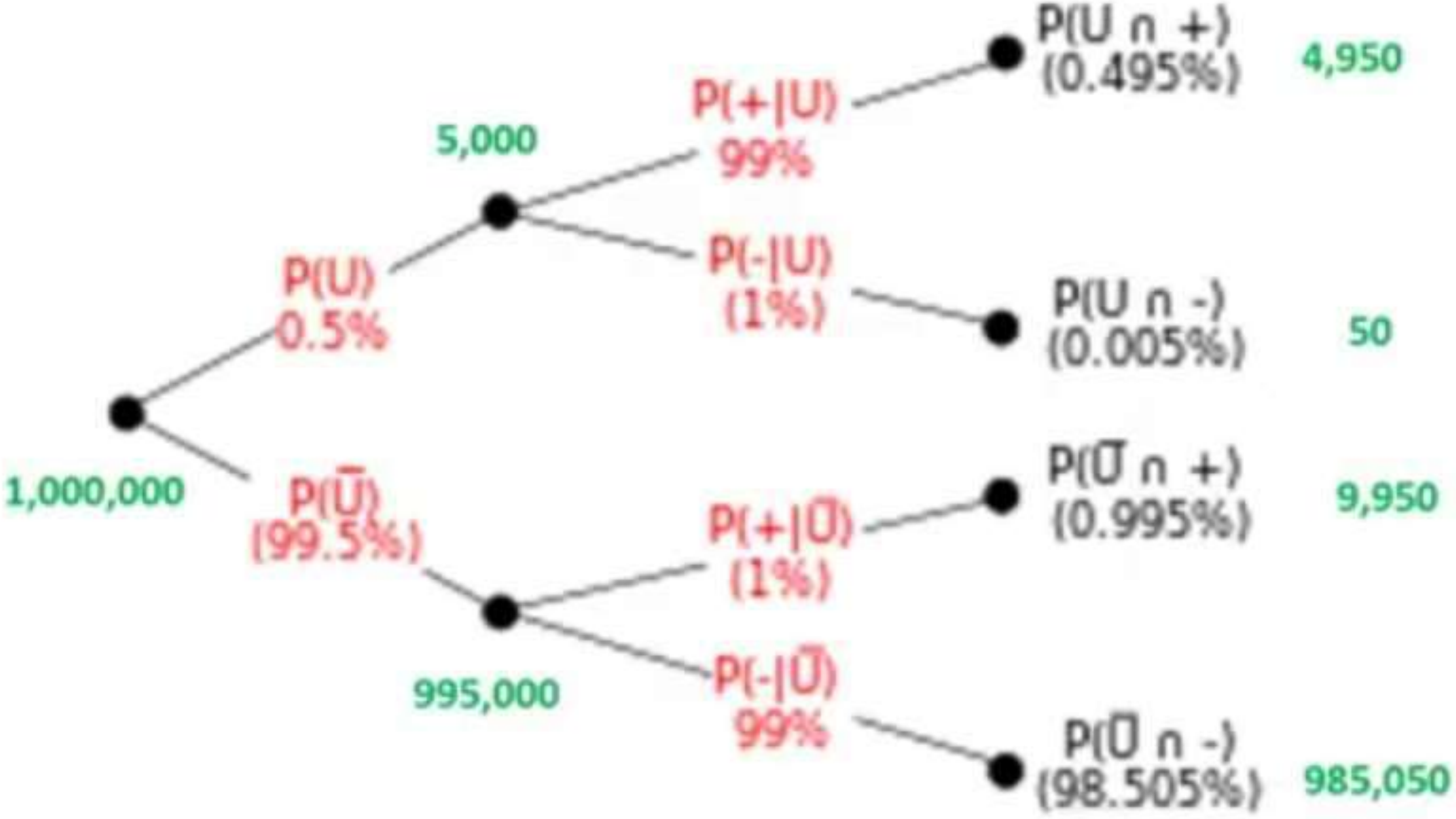
- Joint probability
  - P(A&B), P(A,B), $P(A \cap B)$
- Conditional probability
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - $P(A \cap B) = P(A|B) * P(B)$
- Independent variables
  - $P(A|B) = P(A)$
  - $P(A \cap B) = P(A) * P(B)$
- Bayes Rule
  - $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

Bayes' Rule

- Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual tests positive, what is the probability that he is a user?

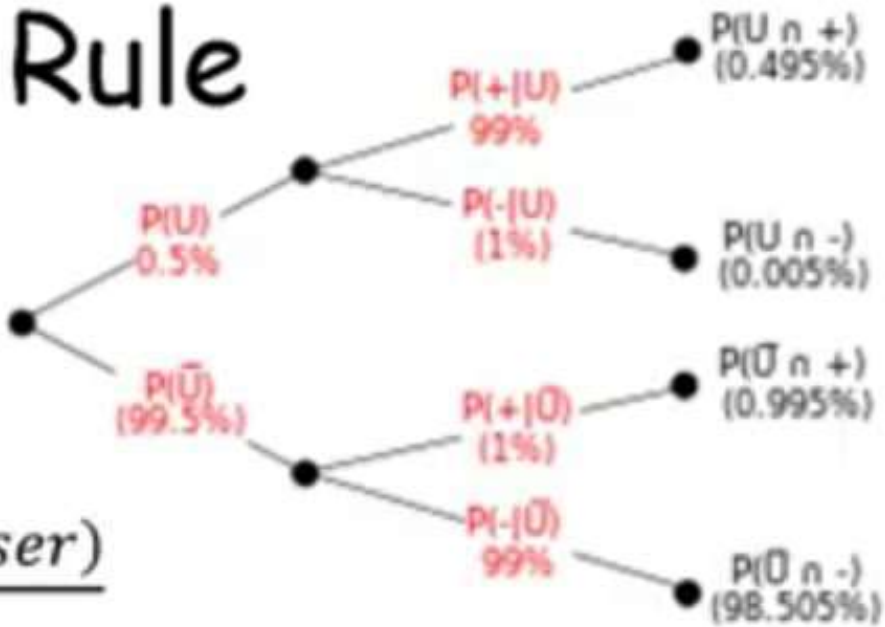


Bayes' Rule



$$P(User|+) = \frac{4,950}{4,950 + 9,950} = \frac{4,950}{14900} = 0.33 = 33\%$$

Bayes' Rule



$$P(User|+) = \frac{P(+|User)P(User)}{P(+)}$$

$$= \frac{P(+|User)P(User)}{P(+,User)+P(+|NonUser)}$$

$$= \frac{P(+|User)P(User)}{P(+|User)P(User) + P(+|NonUser)P(NonUser)}$$

$$= \frac{0.99*0.005}{0.99*0.005 + 0.01*0.995}$$

$$\approx 33\%$$

Bayesian Decision Theory

- Design classifiers to make decisions subject to minimizing an expected ”risk”.
  - The simplest risk is the classification error.
  - When misclassification errors are not equally important, the risk can include the cost associated with different misclassification errors.

Terminology

- State of nature ω *(class label):*
  - e.g., ω_1 for sea bass, ω_2 for salmon

- Probabilities $P(\omega_1)$ and $P(\omega_2)$ *(priors):*
  - e.g., prior knowledge of how likely is to get a sea bass or a salmon

- Probability density function $p(x)$ *(evidence):*
  - e.g., how frequently we will measure a pattern with feature value x (e.g., x corresponds to lightness)

Terminology (cont'd)

- Conditional probability density *$p(x/\omega_j)$* (*likelihood*) :
  - e.g., how frequently we will measure a pattern with feature value x given that the pattern belongs to class ω_j

e.g., lightness distributions between salmon/sea-bass populations

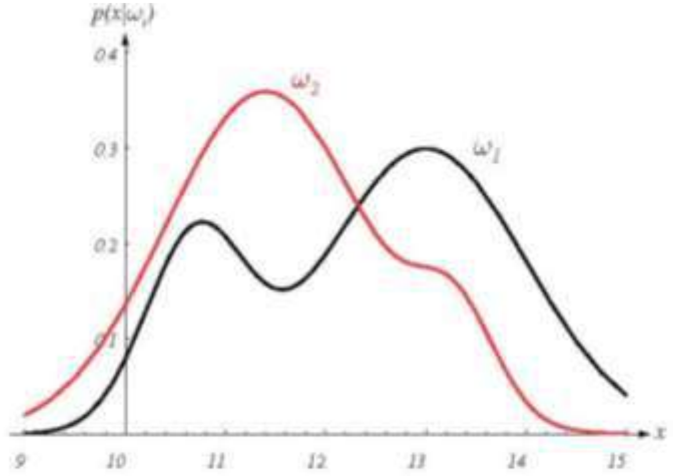


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_j. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons,

Terminology (cont'd)

- Conditional probability $P(\omega_j/x)$ *(posterior)* :
  - e.g., the probability that the fish belongs to class ω_j given feature x.

Decision Rule Using Prior Probabilities Only

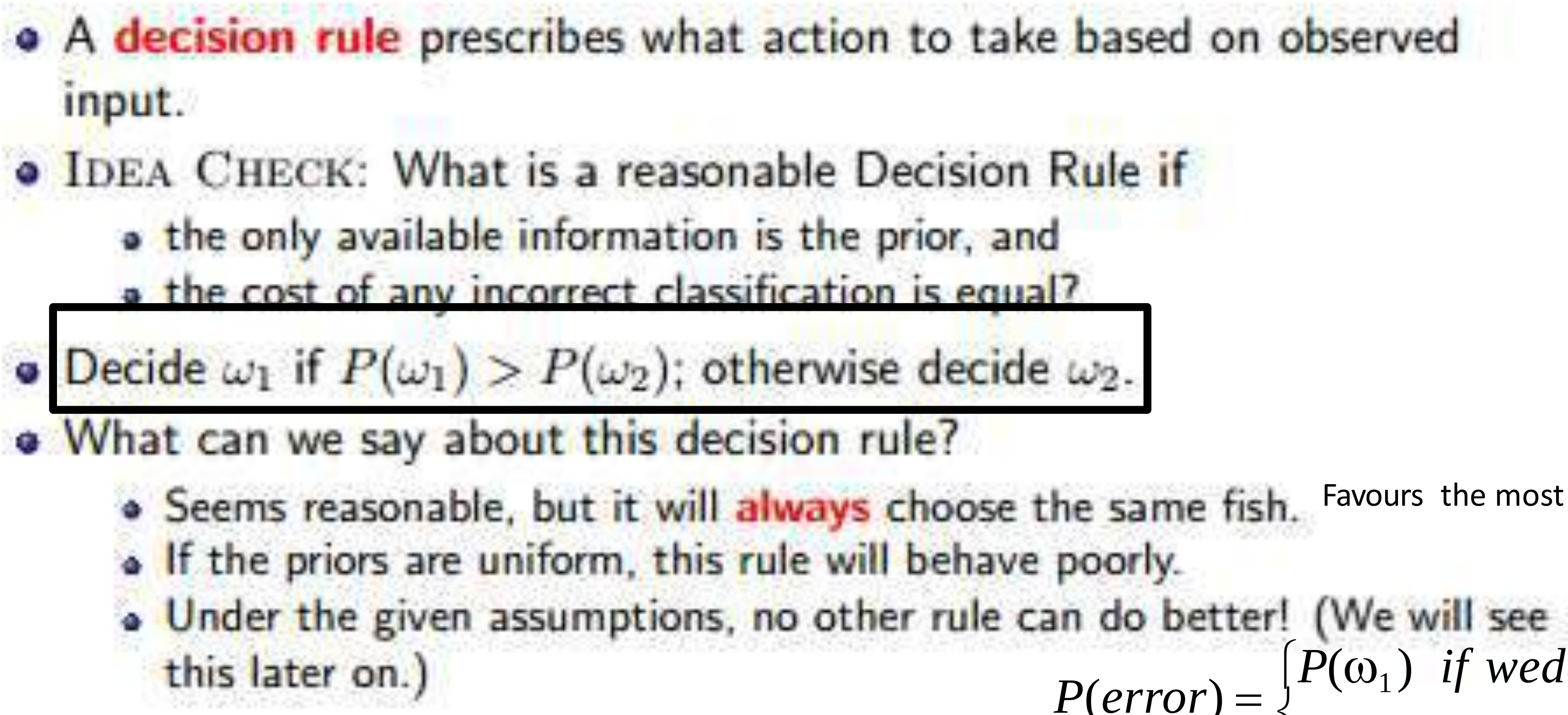
Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise **decide** ω_2

$$P(error) = \begin{cases} P(\omega_1) \ \ if \ \ we\, decide\, \omega_2 \\ P(\omega_2) \ if \ \ we\, decide\, \omega_1 \end{cases}$$

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

- Favours the most likely class.
- This rule will be making the same decision all times.
  - i.e., optimum if no other information is available

Decision Rule from only Priors



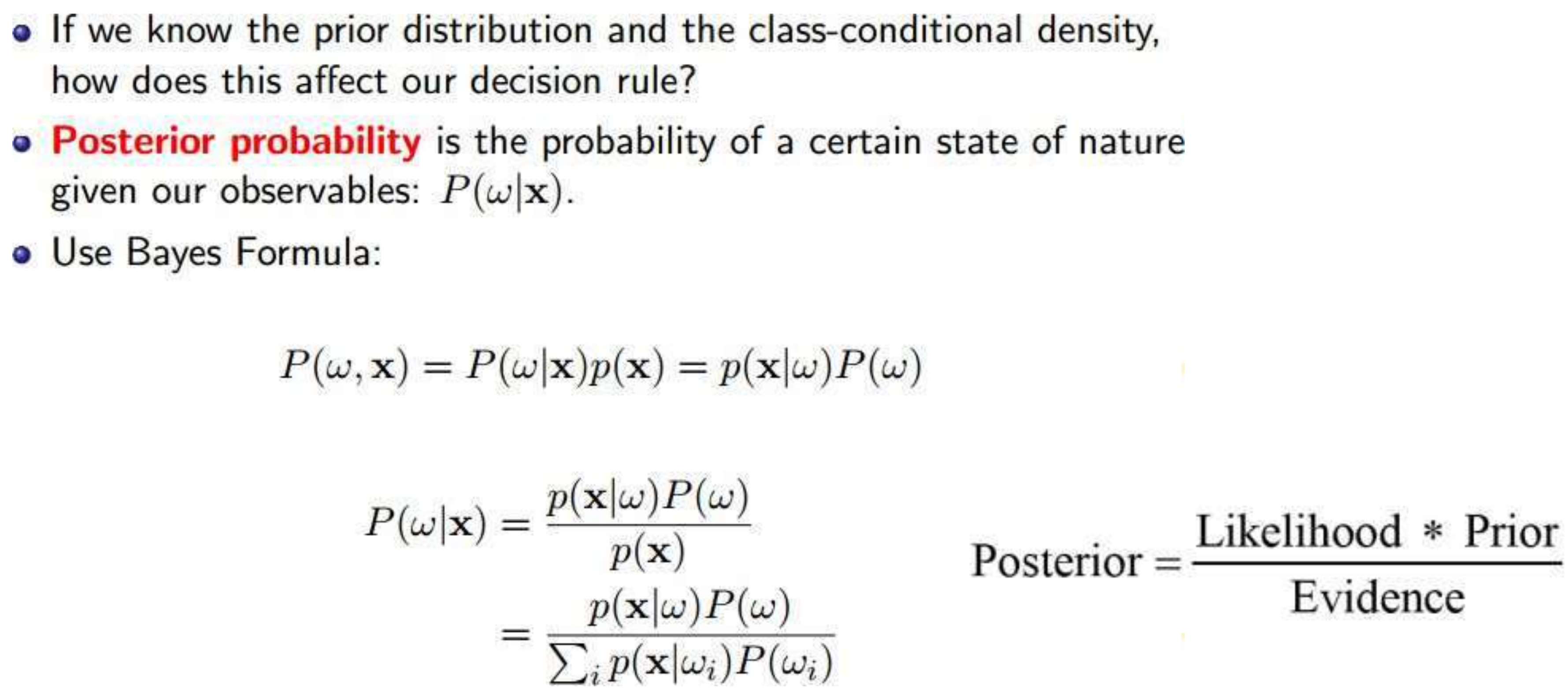
- A **decision rule** prescribes what action to take based on observed input.
- IDEA CHECK: What is a reasonable Decision Rule if
  - the only available information is the prior, and
  - the cost of any incorrect classification is equal?
- Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2.
- What can we say about this decision rule?
  - Seems reasonable, but it will **always** choose the same fish.
  - If the priors are uniform, this rule will behave poorly.
  - Under the given assumptions, no other rule can do better! (We will see this later on.)

Favours the most likely class.

$$P(error) = \begin{cases} P(\omega_1) & if\ we decide \omega_2 \\ P(\omega_2) & if\ we decide \omega_1 \end{cases}$$

or *P(error) = min[P(ω_1), P(ω_2)]*

Conditional probability *$P(\omega_j/x)$* *(posterior) :*



- If we know the prior distribution and the class-conditional density, how does this affect our decision rule?
- **Posterior probability** is the probability of a certain state of nature given our observables: $P(\omega|\mathbf{x})$.
- Use Bayes Formula:

$$P(\omega, \mathbf{x}) = P(\omega|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\omega)P(\omega)$$

$$\begin{aligned} P(\omega|\mathbf{x}) &= \frac{p(\mathbf{x}|\omega)P(\omega)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|\omega)P(\omega)}{\sum_i p(\mathbf{x}|\omega_i)P(\omega_i)} \end{aligned}$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

- e.g., the probability that the fish belongs to class ω_j given feature x.

Decision Rule Using Conditional Probabilities

- Using Bayes' rule:

$$\boxed{P(\omega_j / x) = \frac{p(x/\omega_j)P(\omega_j)}{p(x)} = \frac{likelihood \times prior}{evidence}}$$

where $p(x) = \sum_{j=1}^{2} p(x/\omega_j)P(\omega_j)$ (i.e., scale factor – sum of probs = 1)

Decide ω_1 if $P(\omega_1 /x) > P(\omega_2/x)$; otherwise **decide** ω_2

or

Decide ω_1 if $p(x/\omega_1)P(\omega_1)>p(x/\omega_2)P(\omega_2)$; otherwise **decide** ω_2

or

Decide ω_1 if $\underbrace{p(x/\omega_1)/p(x/\omega_2)}_{\text{likelihood ratio}} > \underbrace{P(\omega_2)/P(\omega_1)}_{\text{threshold}}$; otherwise **decide** ω_2

Decision Rule Using Conditional Probabilities (cont'd)

$$P(\omega_1) = \frac{2}{3} \qquad P(\omega_2) = \frac{1}{3}$$

$p(x/\omega_j)$

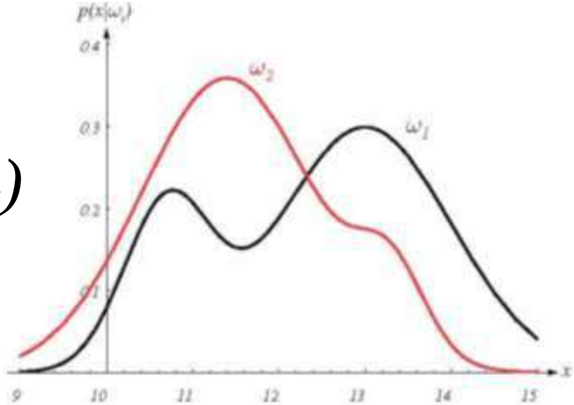


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons,

$P(\omega_j/x)$

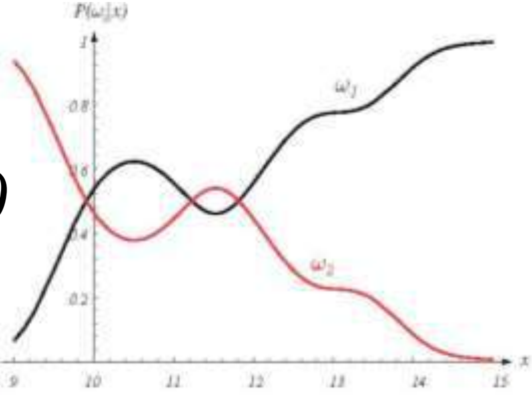
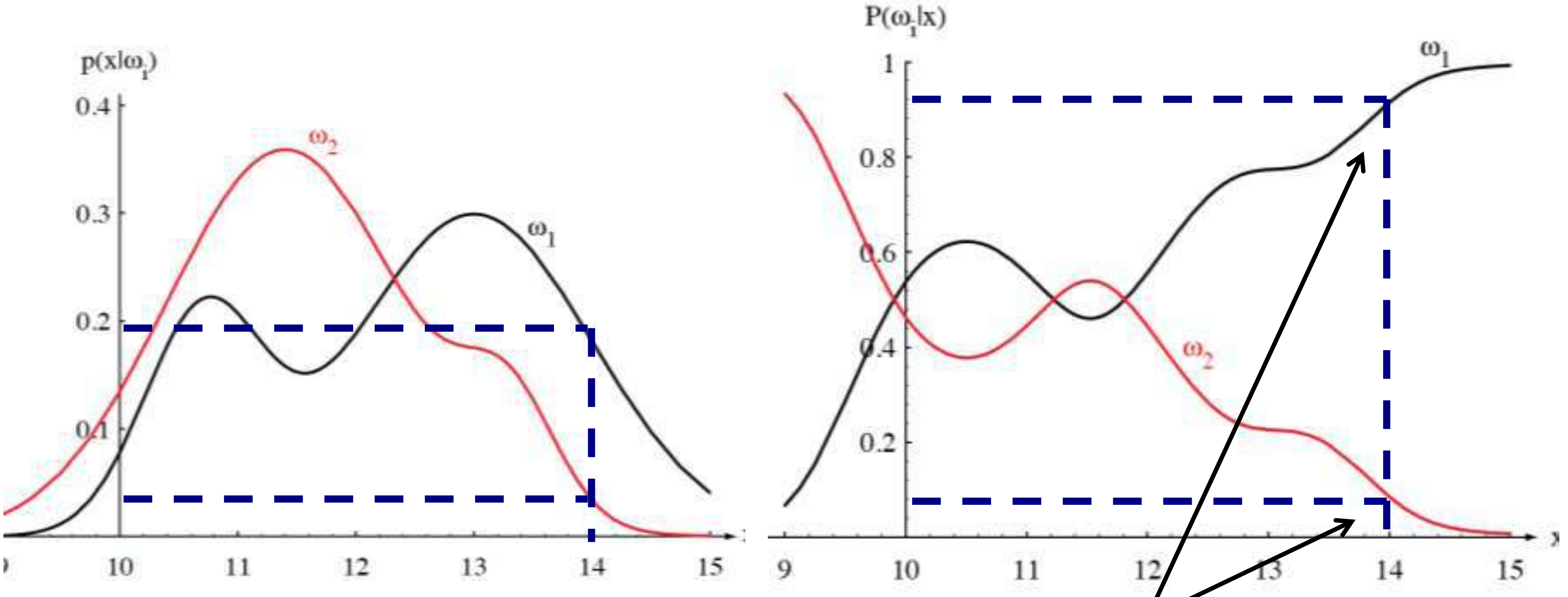


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Posteriors Sum To 1.0

- Two-class fish sorting problem $(P(\omega_1) = 2/3, P(\omega_2) = 1/3)$:



- For every value of x, the posteriors sum to 1.
- At $x = 14$, the probability x is in category ω_1 is 0.92.
- The probability x is in ω_2 is 0.08.
- Likelihoods and posteriors are related via Bayes Rule.

Probability of Error

- The probability of error is defined as:

$$P(error/x) = \begin{cases} P(\omega_1/x) \ \ if \ \ we\,decide\,\omega_2 \\ P(\omega_2/x) \ if \ \ we\,decide\,\omega_1 \end{cases}$$

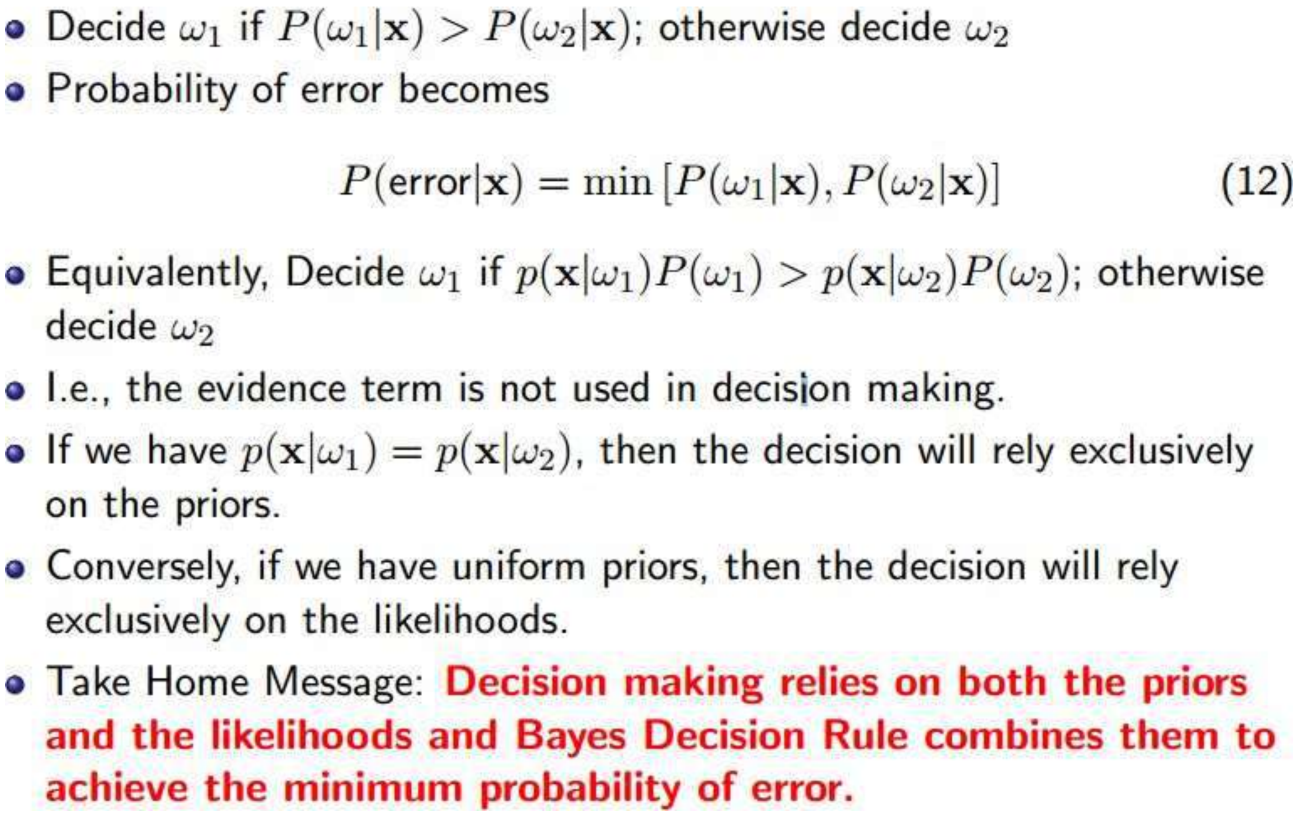
or $P(error/x) = min[P(\omega_1/x), P(\omega_2/x)]$

- What is the average probability error?

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

- The Bayes rule is optimum, that is, it minimizes the average probability error!

Bayes Decision Rule (with equal costs)



- Decide ω_1 if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$; otherwise decide ω_2
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min\left[P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})\right] \tag{12}$$

- Equivalently, Decide ω_1 if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$; otherwise decide ω_2
- I.e., the evidence term is not used in decision making.
- If we have $p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$, then the decision will rely exclusively on the priors.
- Conversely, if we have uniform priors, then the decision will rely exclusively on the likelihoods.
- Take Home Message: **Decision making relies on both the priors and the likelihoods and Bayes Decision Rule combines them to achieve the minimum probability of error.**

Where do Probabilities come from?

- There are two competitive answers:

  (1) **Relative frequency** (objective) approach.
  - Compute probabilities from experiments.

  (2) **Bayesian** (subjective) approach.
  - Compute probabilities from models.

Example (objective approach)

- Classify cars whether they are more or less than \$50K:
  - Classes: C_1 if price > \$50K, C_2 if price <= \$50K
  - Feature: x, the height of a car

- Use the Bayes' rule to compute the posterior probabilities:

$$P(C_i / x) = \frac{p(x / C_i) P(C_i)}{p(x)}$$

- We need to estimate $p(x/C_1)$, $p(x/C_2)$, $P(C_1)$, $P(C_2)$

Example (cont'd)

- Collect data
  - Ask drivers how much their car was and measure height.

- Determine prior probabilities $P(C_1), P(C_2)$
  - e.g., 1209 samples: $\#C_1=221$ $\#C_2=988$

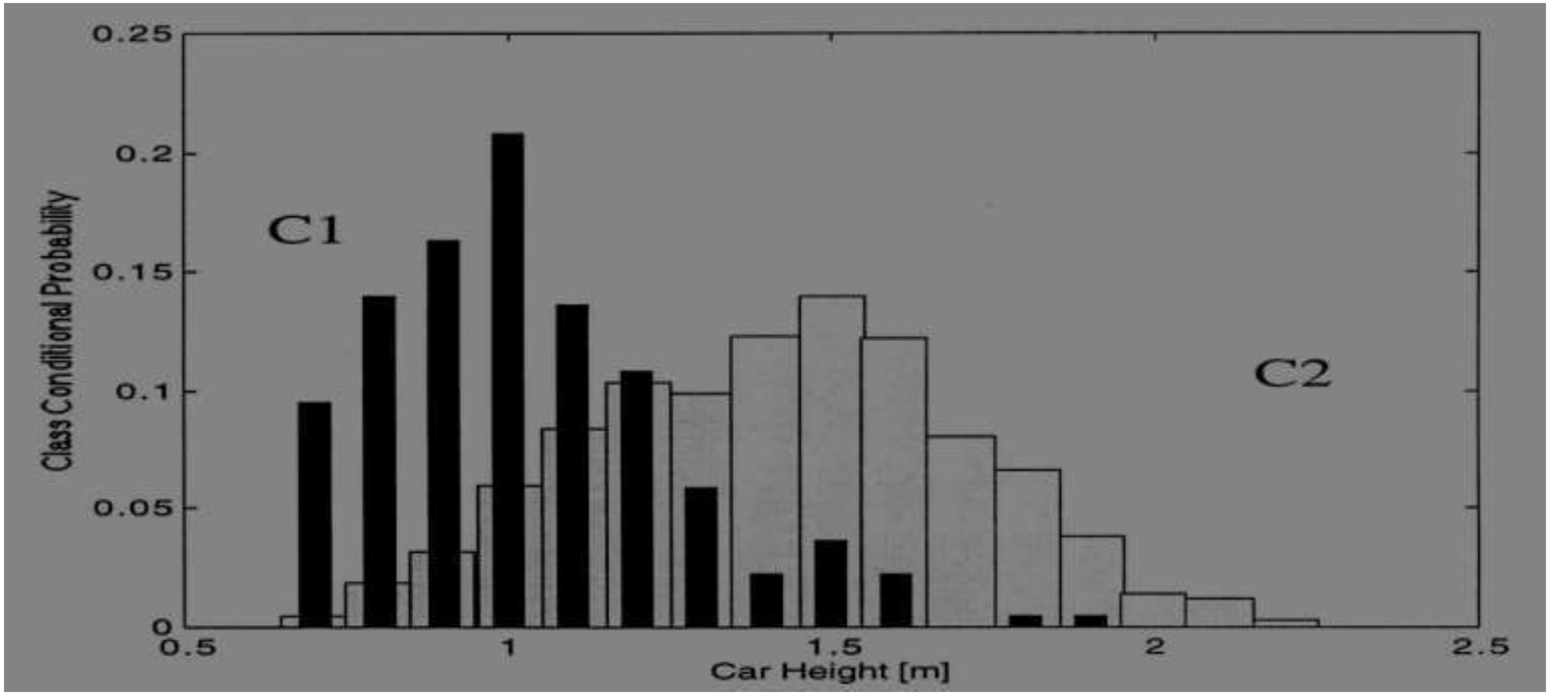
$$P(C_1) = \frac{221}{1209} = 0.183$$

$$P(C_2) = \frac{988}{1209} = 0.817$$

Example (cont'd)

- Determine class conditional probabilities (*likelihood*)
  - Discretize car height into bins and compute normalized histogram.

$p(x / C_i)$



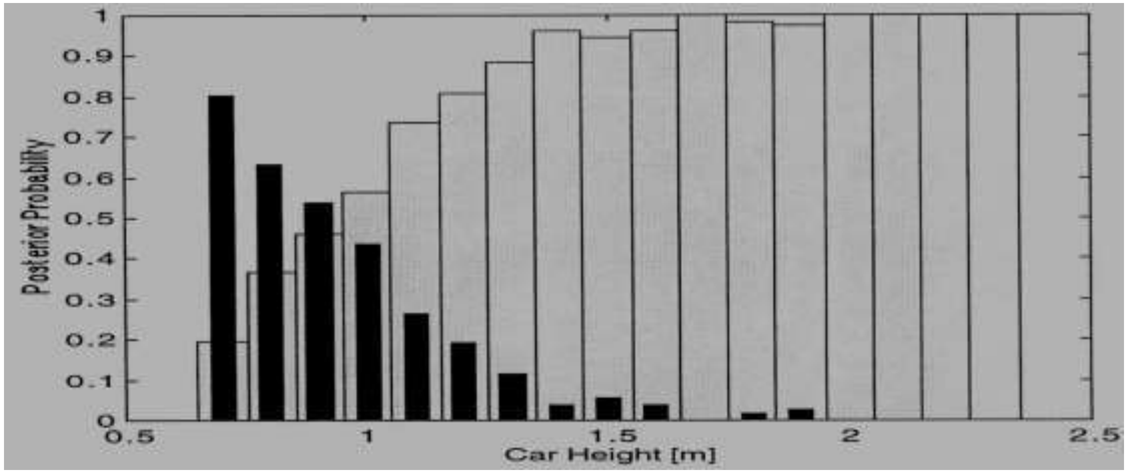
Example (cont'd)

- Calculate the posterior probability for each bin, e.g.:

$$P(C_1 / x = 1.0) = \frac{p(x = 1.0 / C_1)\,P(C_1)}{p(x = 1.0 / C_1)\,P(C_1) + p(x = 1.0 / C_2)\,P(C_2)} =$$

$$= \frac{0.2081 * 0.183}{0.2081 * 0.183 + 0.0597 * 0.817} = 0.438$$

$P(C_i / x)$



Example (subjective approach)

- Associate a model with each class, e.g., Gaussian

$$N(\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\mathbf{x}-\mu)^t\Sigma^{-1}(\mathbf{x}-\mu)]$$

$p(x/C_1)$ ~ $N(\mu_2,\Sigma_2)$

$p(x/C_2)$ ~ $N(\mu_1,\Sigma_1)$

$P(C_1) = P(C_2) = 0.5$

- Use Bayes rule to compute posterior probabilities:

$$P(C_i/x) = \frac{p(x/C_i)P(C_i)}{p(x)}$$

A More General Theory

- Use more than one features.
- Allow more than two categories.
- Allow actions other than classifying the input to one of the possible categories (e.g., rejection).
- Employ a more general error function (i.e., conditional "risk") by associating a "cost" (based on a "loss" function) with different errors.

Terminology

- Features form a vector $\mathbf{x} \in R^d$
- A set of *c* categories $\omega_1, \omega_{2, \ldots,} \omega_c$
- A finite set of *l* actions $\alpha_{1,} \alpha_{2, \ldots,} \alpha_l$ (typically $l \geq$ c)
- A *loss* function $\lambda(\alpha_i / \omega_j)$
  - the cost associated with taking action α_i when the correct classification category is ω_j
- Conditional risk $R(\alpha_i /\mathbf{x})$ – expected loss of taking action α_i given **x**
- Classification is now performed using the $R(\alpha_i /\mathbf{x})$ instead of $P(\omega_i /\mathbf{x})$

Conditional Risk

- Suppose we take action α_i when **x** is observed.
- The **conditional risk** (or **expected loss**) with taking action α_i is defined as:

$$R(a_i / \mathbf{x}) = \sum_{j=1}^{c} \lambda(a_i / \omega_j) P(\omega_j / \mathbf{x})$$

where $$P(\omega_j / \mathbf{x}) = \frac{p(\mathbf{x} / \omega_j) P(\omega_j)}{p(\mathbf{x})}$$

Overall Risk

- The overall risk is defined as:

$$R = \int R(a(\mathbf{x}) / \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

where *α(**x**)* is a general decision rule that determines which action $\alpha_{1,}$ $\alpha_{2, \ldots,}$ α_l to take for every **x**.

- Use the Bayes rule to minimize R.

Overall Risk (cont'd)

- The *Bayes rule* minimizes *R* by:

  (i) Computing *$R(\alpha_i/\mathbf{x})$* for every α_i given an **x**

  (ii) Choosing the action α_i with the minimum conditional risk *$R(\alpha_i/\mathbf{x})$*

- The resulting minimum R^* is called *Bayes risk* and is the best performance that can be achieved:

$$R^* = \min R$$

Example: Two-category classification

- Define
  - α_1: decide ω_1
  - $\alpha_{2:}$ decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i / \omega_j)$

- The conditional risks are:

$$R(a_i / \mathbf{x}) = \sum_{j=1}^{c} \lambda(a_i / \omega_j) P(\omega_j / \mathbf{x})$$

⇩

$$R(a_1/\mathbf{x}) = \lambda_{11} P(\omega_1/\mathbf{x}) + \lambda_{12} P(\omega_2/\mathbf{x})$$
$$R(a_2/\mathbf{x}) = \lambda_{21} P(\omega_1/\mathbf{x}) + \lambda_{22} P(\omega_2/\mathbf{x})$$

Example: Two-category classification (cont'd)

- Minimum risk decision rule:

$$\textbf{Decide } \omega_1 \text{ if } R(a_1/\mathbf{x}) < R(a_2/\mathbf{x}) \text{; otherwise decide } \omega_2$$

or, using

$$R(a_1/\mathbf{x}) = \lambda_{11}P(\omega_1/\mathbf{x}) + \lambda_{12}P(\omega_2/\mathbf{x})$$
$$R(a_2/\mathbf{x}) = \lambda_{21}P(\omega_1/\mathbf{x}) + \lambda_{22}P(\omega_2/\mathbf{x})$$

$$\textbf{Decide } \omega_1 \text{ if } (\lambda_{21} - \lambda_{11})P(\omega_1/\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2/\mathbf{x}) \text{; otherwise decide } \omega_2$$

or, using

$$P(\omega_j/\mathbf{x}) = \frac{p(\mathbf{x}/\omega_j)P(\omega_j)}{p(\mathbf{x})}$$

$$\textbf{Decide } \omega_1 \text{ if } \frac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)} \text{; otherwise decide } \omega_2$$

likelihood ratio threshold

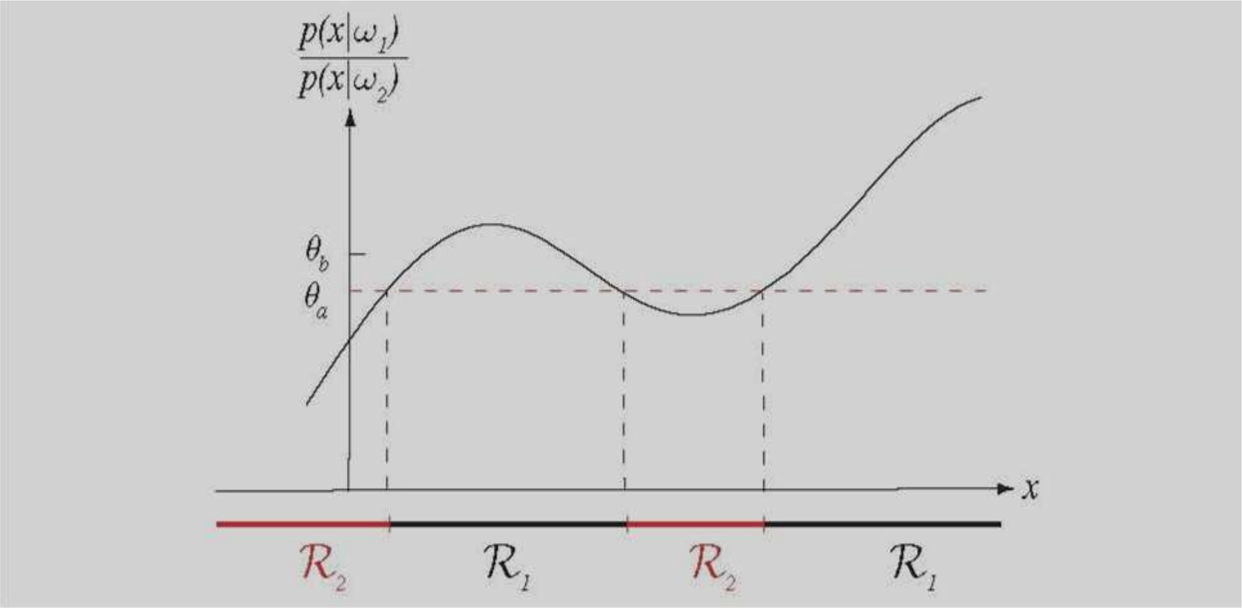
Example

Assuming general loss:

$$\textbf{Decide } \omega_1 \text{ if } \frac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)}; \text{ otherwise decide } \omega_2$$

Assuming zero-one loss:

Decide ω_1 if $p(x/\omega_1)/p(x/\omega_2)>P(\omega_2)/P(\omega_1)$ otherwise **decide** ω_2



(decision regions)

$$\theta_a = P(\omega_2)/P(\omega_1)$$

$$\theta_b = \frac{P(\omega_2)(\lambda_{12}-\lambda_{22})}{P(\omega_1)(\lambda_{21}-\lambda_{11})}$$

Example: Likelihood Ratio

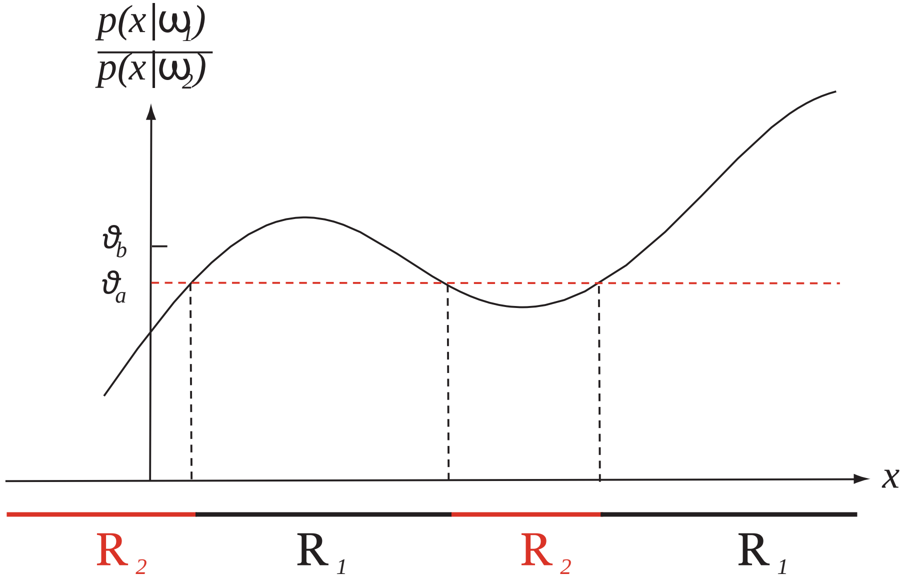


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold ϑ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold ϑ_b, and hence R_1 becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright ⓧc 2001 by John Wiley & Sons, Inc.

R·I·T

Special Case:
Zero-One Loss Function

- Assign the same loss to all errors:

$$\lambda(a_i/\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

- What is the conditional risk in this case?

$$R(a_i/\mathbf{x}) = \sum_{j=1}^{c} \lambda(a_i/\omega_j) P(\omega_j/\mathbf{x}) = \sum_{i \neq j} P(\omega_j/\mathbf{x}) = 1 - P(\omega_i/\mathbf{x})$$

Special Case: Zero-One Loss Function (cont'd)

- The decision rule becomes:

$$\textbf{Decide } \omega_1 \text{ if } R(a_1/\mathbf{x}) < R(a_2/\mathbf{x})\text{; otherwise decide } \omega_2$$

or $$\textbf{Decide } \omega_1 \text{ if } 1 - P(\omega_1/\mathbf{x}) < 1 - P(\omega_2/\mathbf{x})\text{; otherwise decide } \omega_2$$

or $$\textbf{Decide } \omega_1 \text{ if } P(\omega_1/\mathbf{x}) > P(\omega_2/\mathbf{x})\text{; otherwise decide } \omega_2$$

- The overall risk becomes the average probability error!

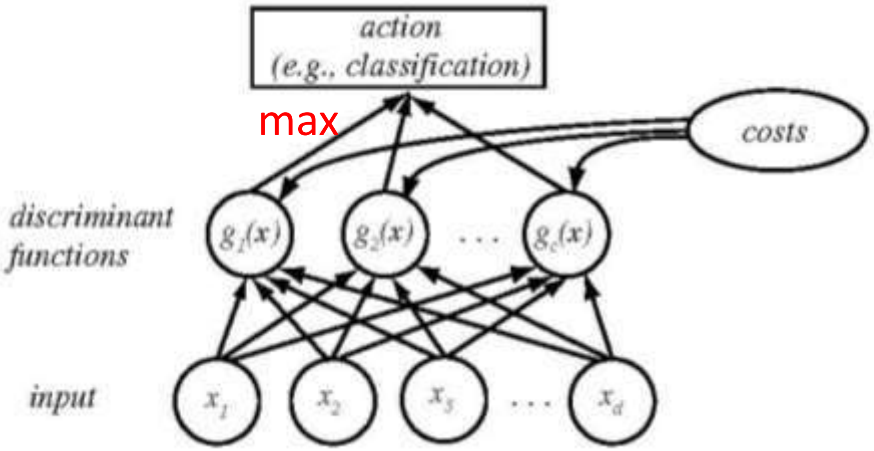
Discriminant Functions

- Represent a classifier through discriminant functions

$$g_i(x),\ i = 1, \ldots, c$$

- A feature vector **x** is assigned to class ω_i if:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ for all } j \neq i$$



Discriminants for Bayes Classifier

- Assuming a general loss function:

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

- Assuming the zero-one loss function:

$$g_i(\mathbf{x})=P(\omega_i / \mathbf{x})$$

Discriminants for Bayes Classifier (cont'd)

- Replacing $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$, where $f()$ is monotonically increasing, does not change the classification results.

$g_i(x)=P(\omega_i/x)$ ⇒

$$g_i(\mathbf{x}) = \frac{p(\mathbf{x}/\omega_i)P(\omega_i)}{p(\mathbf{x})}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}/\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}/\omega_i) + \ln P(\omega_i)$$

we'll use this extensively!

Case of two categories

- More common to use a single discriminant function (*dichotomizer*) instead of two:

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

$$\textbf{Decide } \omega_1 \text{ if } g(\mathbf{x}) > 0\text{; otherwise decide } \omega_2$$

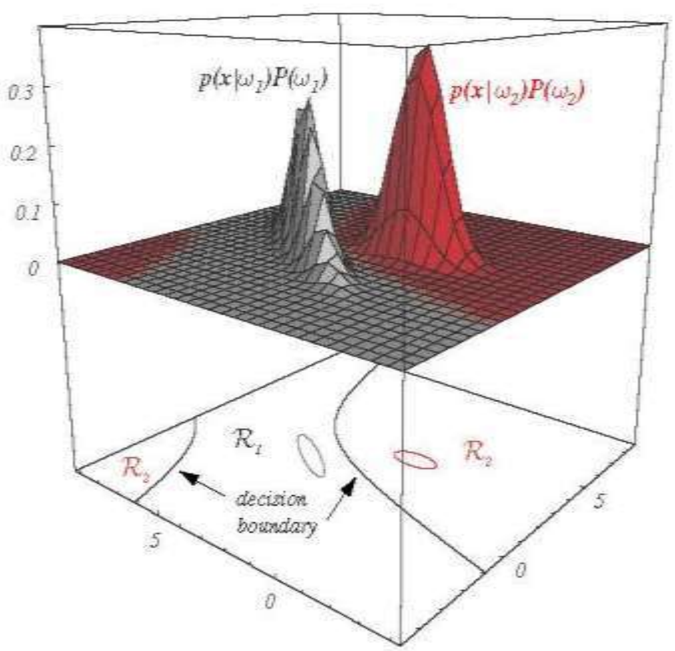
Examples:

$$g(\mathbf{x}) = P(\omega_1 / \mathbf{x}) - P(\omega_2 / \mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} / \omega_1)}{p(\mathbf{x} / \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Decision Regions and Boundaries

- Discriminants divide the feature space in *decision regions* R_1, R_2, …, R_c, separated by *decision boundaries*.



Decision boundary

is defined by:

$$g_1(\mathbf{x})=g_2(\mathbf{x})$$

Example: Decision Regions for Binary Classifier

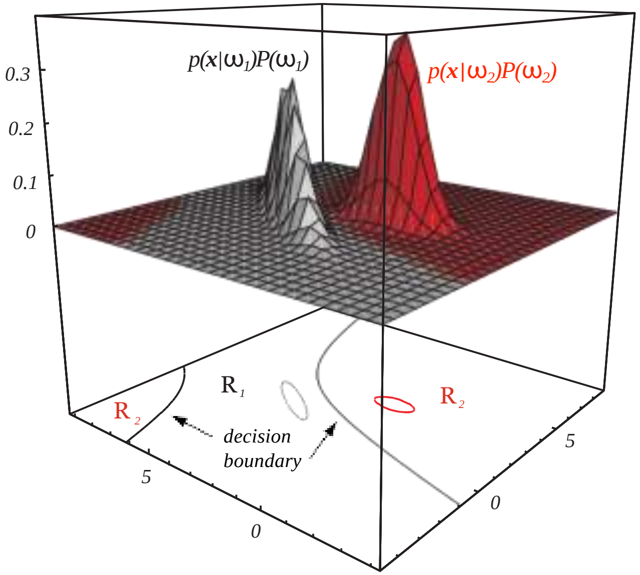


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region R_2 is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright ⓒ 2001 by John Wiley & Sons, Inc.

The (Univariate) Normal Distribution

Why are Gaussians so Useful?

They represent many probability distributions in nature quite accurately. In our case, when patterns can be represented as random variations of an ideal prototype (represented by the mean feature vector)

- Everyday examples: height, weight of a population

Univariate Normal Distribution

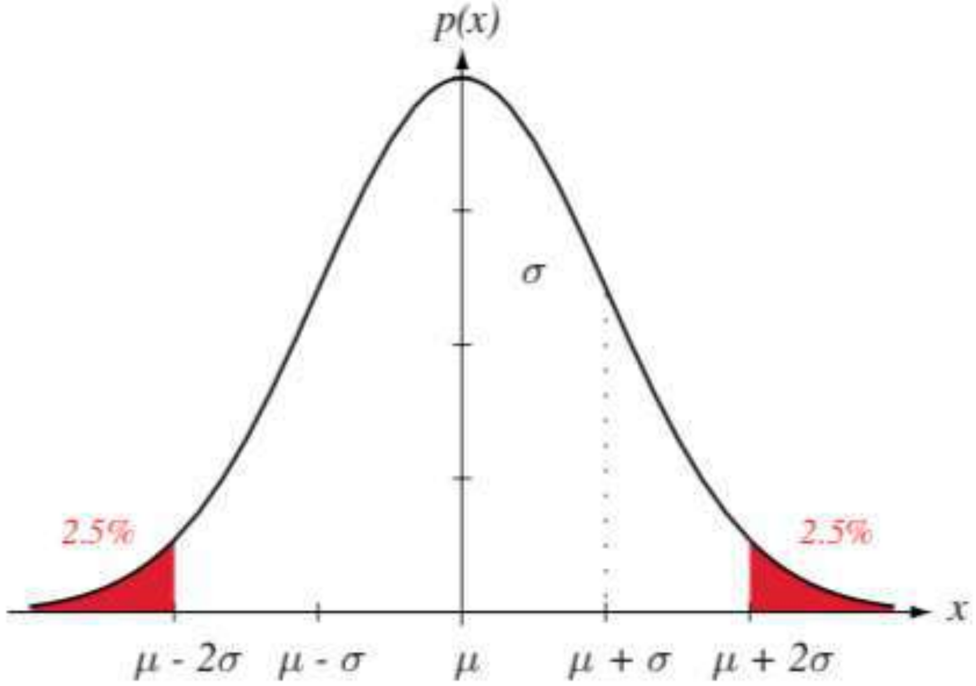


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Formal Definition

Peak of the Distribution (the mean)

Has value: $\frac{1}{\sqrt{2\pi}\sigma}$

Definition for Univariate Normal

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

Def. for mean, variance

$$\mu = \int_{-\infty}^{\infty} x \, p(x) \, dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, dx$$

Multivariate Normal Density

Informal Definition

A normal distribution over two or more variables (*d* variables/dimensions)

Formal Definition

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

$$\mu = \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x}$$

$$\mathbf{\Sigma} = \int (\mathbf{x}-\mu)(\mathbf{x}-\mu)^t p(\mathbf{x}) d\mathbf{x}$$

The Covariance Matrix (Σ)

For our purposes...

Assume matrix is positive definite, so the determinant of the matrix is always positive

Matrix Elements

- Main diagonal: variances for each individual variable
- Off-diagonal: covariances of each variable pairing i & j (note: values are repeated, as matrix is symmetric)

Independence and Correlation

For multivariate normal covariance matrix

- Off-diagonal entries with a value of 0 indicate uncorrelated variables, that are *statistically* independent (variables likely do not influence one another)

- Roughly speaking, covariance positive if two variables increase together (positive correlation), negative if one variable decreases when the other increases (negative correlation)

A Two-Dimensional Gaussian Distribution, with Samples Shown

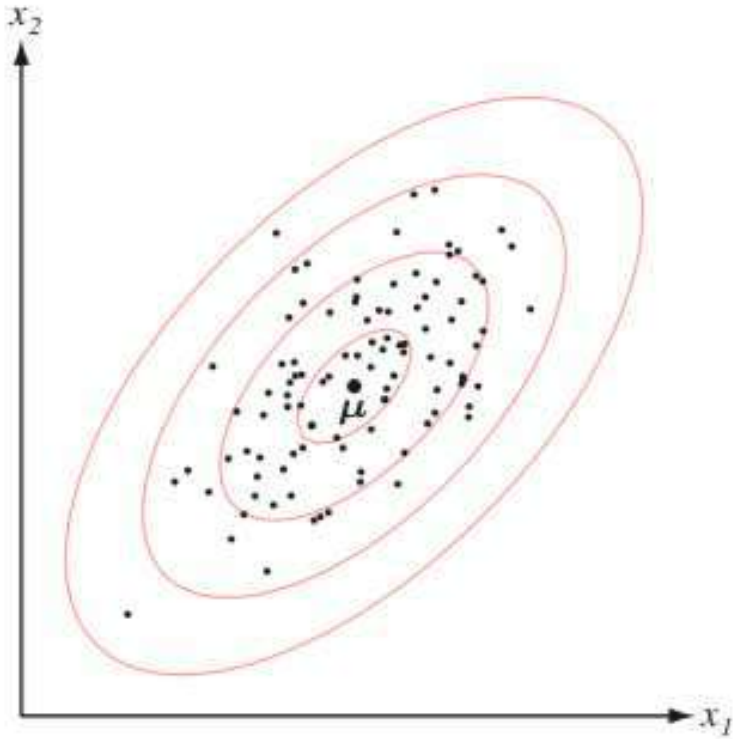
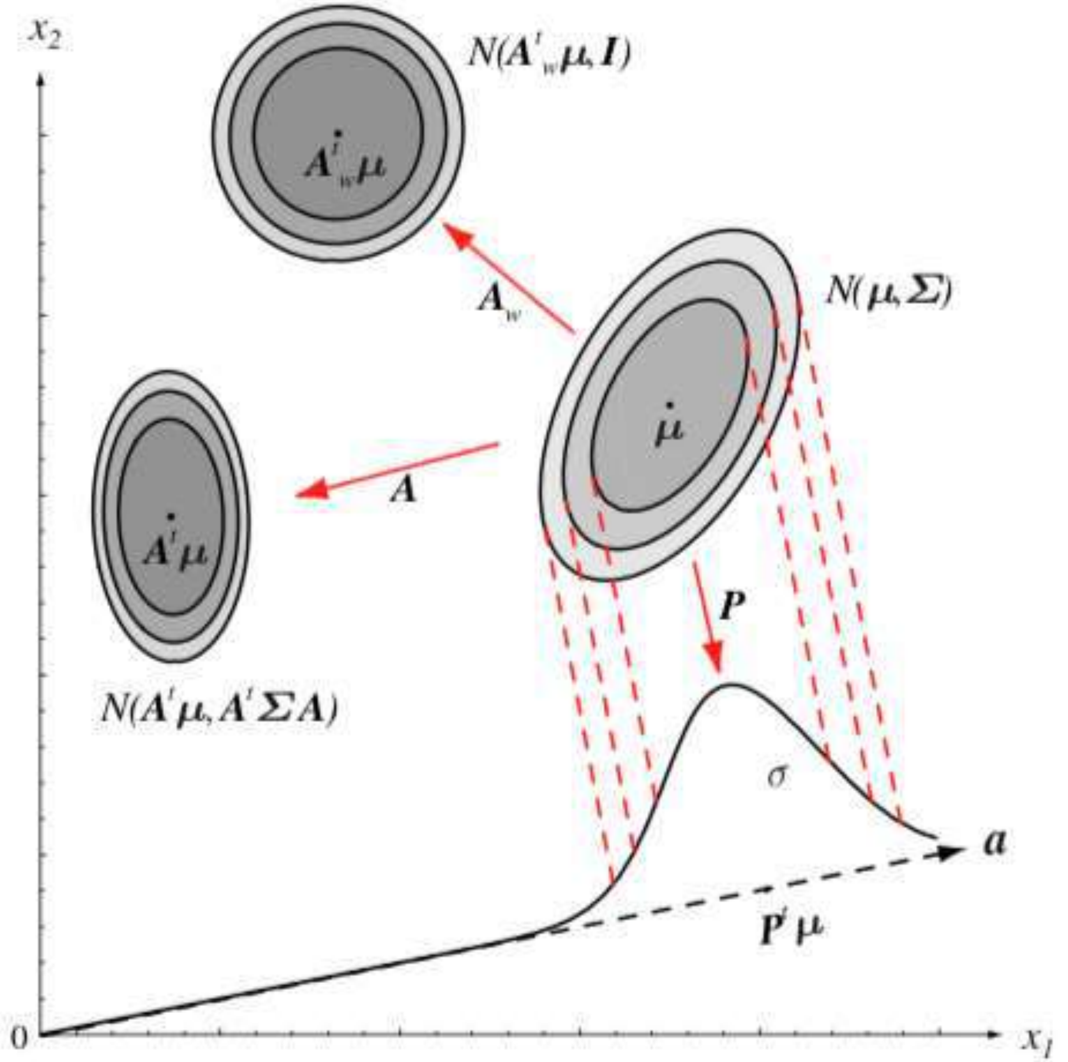


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Linear Transformations in a 2D Feature Space



Discriminant Functions ($g_i(x)$) for the Normal Density

Discriminant Functions

We will consider three special cases for:

- normally distributed features, and
- minimum-error-rate classification (0-1 loss)

Recall: $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$

if $p(x|\omega_i) \sim N(\mu_\mathbf{i}, \Sigma_\mathbf{i})$ then approx. $p(\mathbf{x}|\omega_i)$

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

Minimum Error-Rate Discriminant Function for Multivariate Gaussian Feature Distributions

ln (natural log) of

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

gives a general form for our discriminant functions:

$$g_i(\mathbf{x}) = \underline{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{i}})^t {\boldsymbol{\Sigma}_{\mathbf{i}}}^{-1}(\mathbf{x}-\mu_{\mathbf{i}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathbf{i}}|} + \ln P(\omega_i)$$

Special Cases for Binary Classification

Purpose

Overview of commonly assumed cases for feature likelihood densities, $p(\mathbf{x}|\omega_i)$

- Goal: eliminate common additive constants in discriminant functions. These do not affect the classification decision (i.e. define $g_i(x)$ providing "just the differences")

- Also, look at resulting decision surfaces (defined by $g_i(x) = g_j(x)$)

Three Special Cases

1. Statistically independent features, identically distributed Gaussians for each class

2. Identical covariances for each class

3. Arbitrary covariances

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

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

Remove:

Items in red: same across classes (“unimportant additive constants”)

Inverse of Covariance Matrix: $\mathbf{\Sigma_i^{-1}} = (1/\sigma^2)\mathbf{I}$

Only effect is to scale vector product by $1/\sigma^2$

Discriminant function:

$$g_i(x) = -\frac{(\mathbf{x} - \mu_\mathbf{i})^t(\mathbf{x} - \mu_\mathbf{i})}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(x) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\mu_\mathbf{i}^\mathbf{t}\mathbf{x} + \mu_\mathbf{i}{}^t\mu_\mathbf{i}] + \ln P(\omega_i)$$

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

Linear Discriminant Function

Produced by factoring the previous form

$$g_i(x) = \boxed{\mathbf{w_i^t}}\mathbf{x} + \boxed{\omega_{i0}}$$

$$g_i(x) = \boxed{\frac{1}{\sigma^2}\mu_{\mathbf{i}}{}^{t}}\mathbf{x}\boxed{-\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i)}$$

Threshold or Bias for Class i: ω_{i0}

Change in prior translates decision boundary

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

Decision Boundary: $g_i(x) = g_j(x)$

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

$$(\mu_i - \mu_j)^t(x - \left(\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{(\mu_i - \mu_j)^t(\mu_i - \mu_j)} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)\right))$$

- Decision boundary goes through x0 along line between means, orthogonal to this line
- If priors equal, x0 between means (*minimum distance classifier*), otherwise x0 shifted
- If variance small relative to distance between means, priors have limited effect on boundary location

Case 1: Statistically Independent Features with Identical Variances

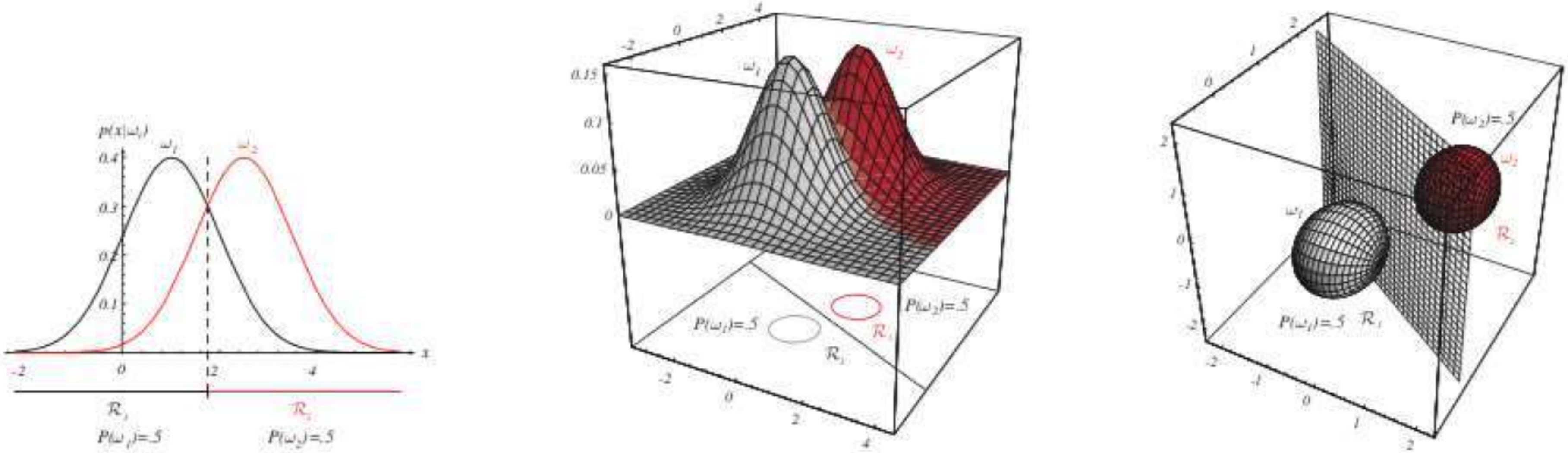
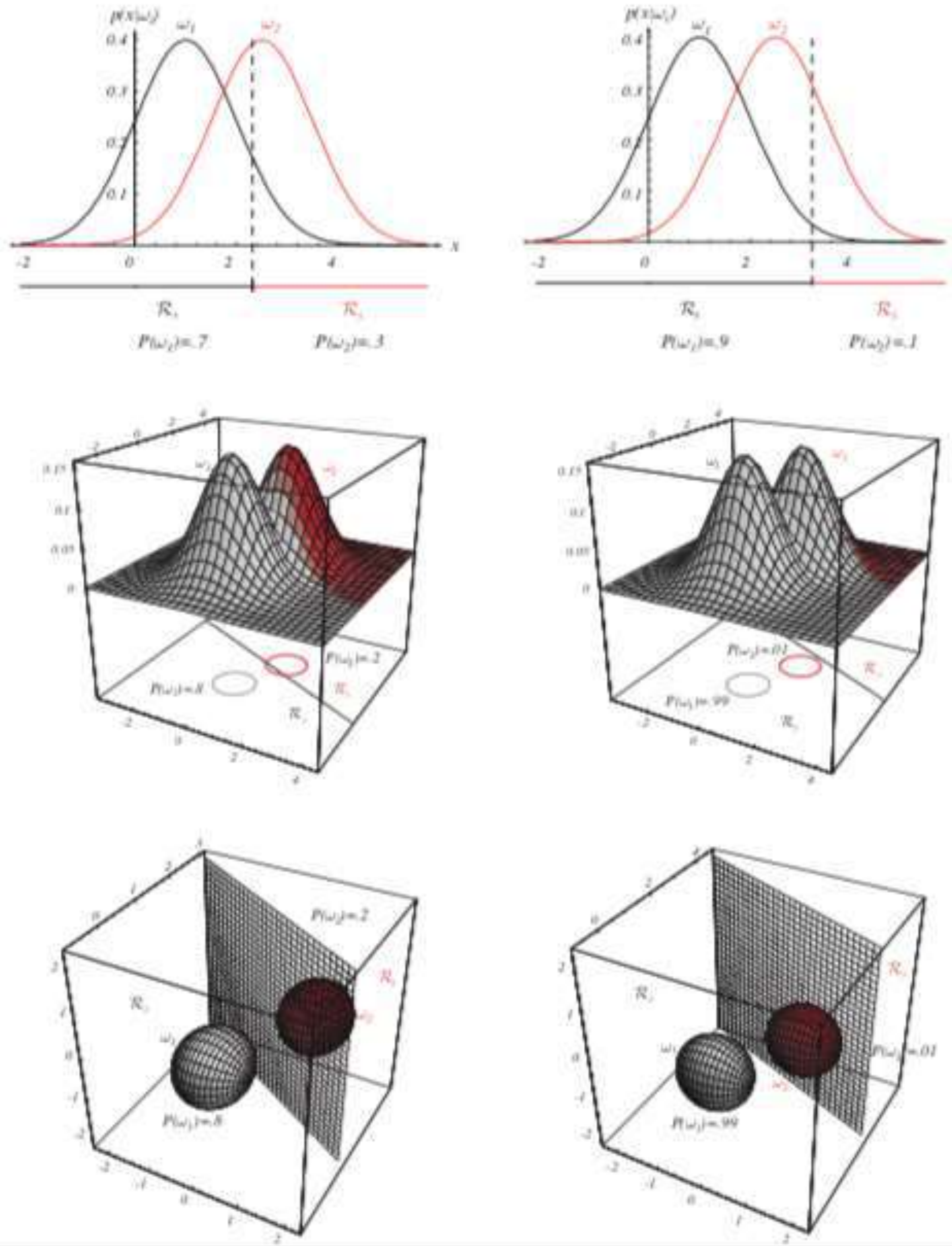


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Example: Translation of Decision Boundaries Through Changing Priors



Case II: Identical Covariances, $\mathbf{\Sigma_i = \Sigma}$

$$g_i(\mathbf{x}) = -\frac{1}{2}\boxed{(\mathbf{x}-\mu_{\mathrm{i}})^t \mathbf{\Sigma}_{\mathrm{i}}{}^{-1}(\mathbf{x}-\mu_{\mathrm{i}})} \boxed{-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_{\mathrm{i}}|} + \ln P(\omega_i)$$

Remove

Terms in red; as in Case I these can be ignored (same across classes)

Squared Mahalanobis Distance (yellow)

Distance from x to mean for class i, taking covariance into account; defines contours of fixed density

Case II: Identical Covariances, $\mathbf{\Sigma_i = \Sigma}$

Expansion of squared Mahalanobis distance

$$
\begin{aligned}
(\mathbf{x}-\mu_i)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_{\mathbf{i}}) \\
&= \mathbf{x^t\Sigma^{-1}x} - \mathbf{x^t}\underline{\mathbf{\Sigma^{-1}}\mu_{\mathbf{i}}} - \underline{\mu_{\mathbf{i}}^{\mathbf{t}}\mathbf{\Sigma^{-1}}}\mathbf{x} + \mu_{\mathbf{i}}^{\mathbf{t}}\mathbf{\Sigma^{-1}}\mu_{\mathbf{i}} \\
&= \boxed{\mathbf{x^t\Sigma^{-1}x}} - 2(\mathbf{\Sigma^{-1}}\mu_{\mathbf{i}})^{\mathbf{t}}\mathbf{x} + \mu_{\mathbf{i}}^{\mathbf{t}}\mathbf{\Sigma^{-1}}\mu_{\mathbf{i}}
\end{aligned}
$$

the last step comes from symmetry of the covariance matrix and thus its inverse:

$$\Sigma^t = \Sigma, (\Sigma^{-1})^t = \Sigma^{-1}$$

Once again, term above in red is an additive constant independent of class, and can be removed

Case II: Identical Covariances, $\boldsymbol{\Sigma}_{\mathbf{i}} = \boldsymbol{\Sigma}$

Linear Discriminant Function

$$g_i(x) = \mathbf{w}_\mathbf{i}^\mathbf{t}\mathbf{x} + \omega_{i0}$$

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

Decision Boundary: $g_i(x) = g_j(x)$

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

$$(\boldsymbol{\Sigma}^{-\mathbf{1}}(\mu_\mathrm{i} - \mu_\mathrm{j}))^\mathbf{t}(\mathbf{x} - \left(\frac{\mathbf{1}}{\mathbf{2}}(\mu_\mathrm{i} + \mu_\mathrm{j}) - \frac{\mathbf{ln}[\mathbf{P}(\omega_\mathrm{i})/\mathbf{P}\omega_\mathrm{j})}{(\mu_\mathrm{i} - \mu_\mathrm{j})\boldsymbol{\Sigma}^{-\mathbf{1}}(\mu_\mathrm{i} - \mu_\mathrm{j})}(\mu_\mathrm{i} - \mu_\mathrm{j})\right))$$

$$= 0$$

Case II: Identical Covariances, $\Sigma_i = \Sigma$

Notes on Decision Boundary

- As for Case I, passes through point x0 lying on the line between the two class means. Again, x0 in the middle if priors identical
- Hyperplane defined by boundary generally not orthogonal to the line between the two means

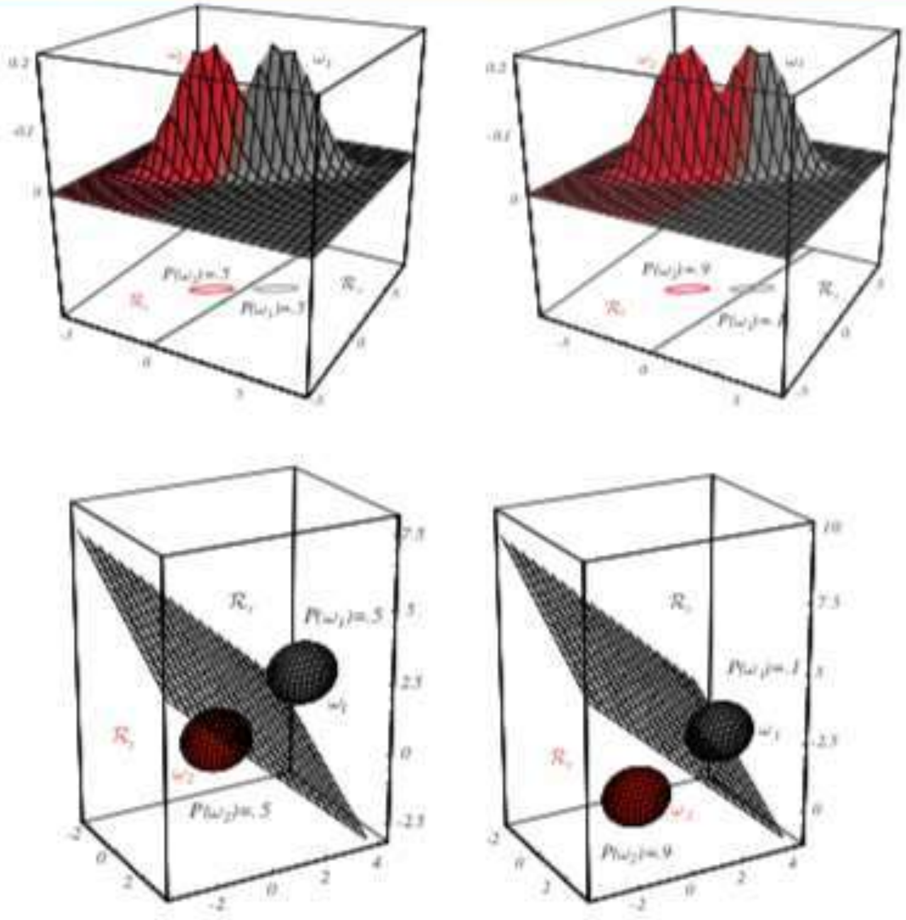


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions a ellipsoidal surfaces in three dimensions) and decision regions for equal but asymm ric Gaussian distributions. The decision hyperplanes need not be perpendicular to t line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. St *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Case III: arbitrary $\mathbf{\Sigma_i}$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_\mathbf{i})^t \mathbf{\Sigma_i}^{-1}(\mathbf{x} - \mu_\mathbf{i}) \boxed{- \frac{d}{2}\ln 2\pi} - \frac{1}{2}\ln|\mathbf{\Sigma_i}| + \ln P(\omega_i)$$

Remove

Can only remove the one term in red above

Discriminant Function (quadratic)

$$g_i(x) = x^t \boxed{W_i} x + \boxed{w_i^t} x + \boxed{\omega_{i0}}$$

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

Case III: arbitrary Σ_i

Decision Boundaries

Are hyperquadrics: can be hyperplanes, hyperplane pairs, hyperspheres, hyperellipsoids, hyperparabaloids, hyperhyperparabaloids

Decision Regions

Need not be simply connected, even in one dimension (next slide)

Case 3: Arbitrary Covariances

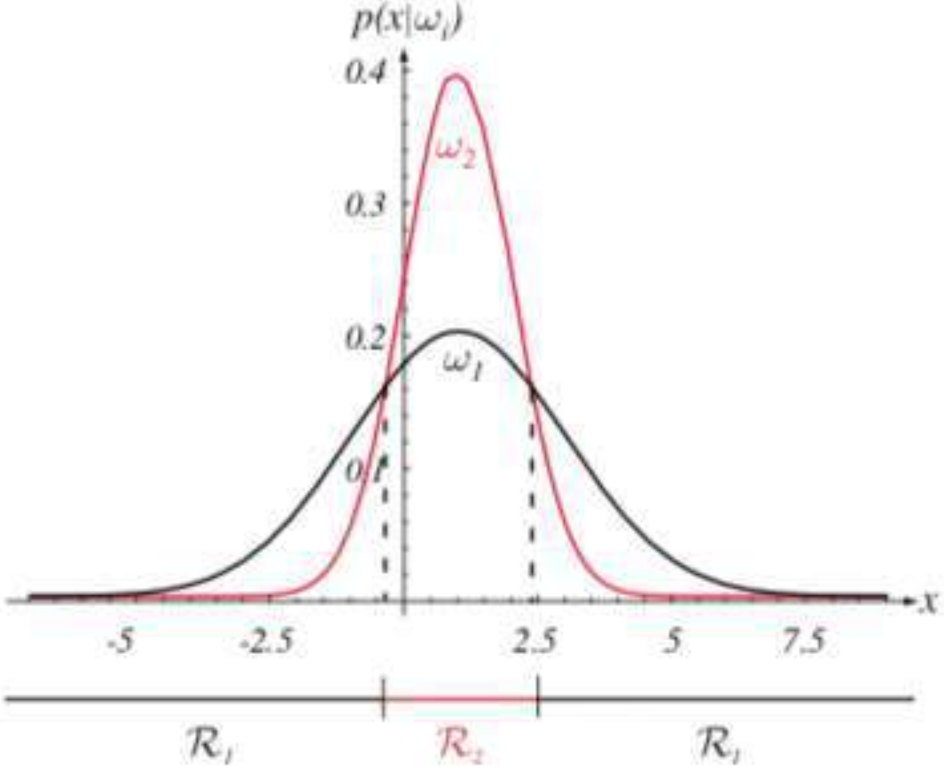
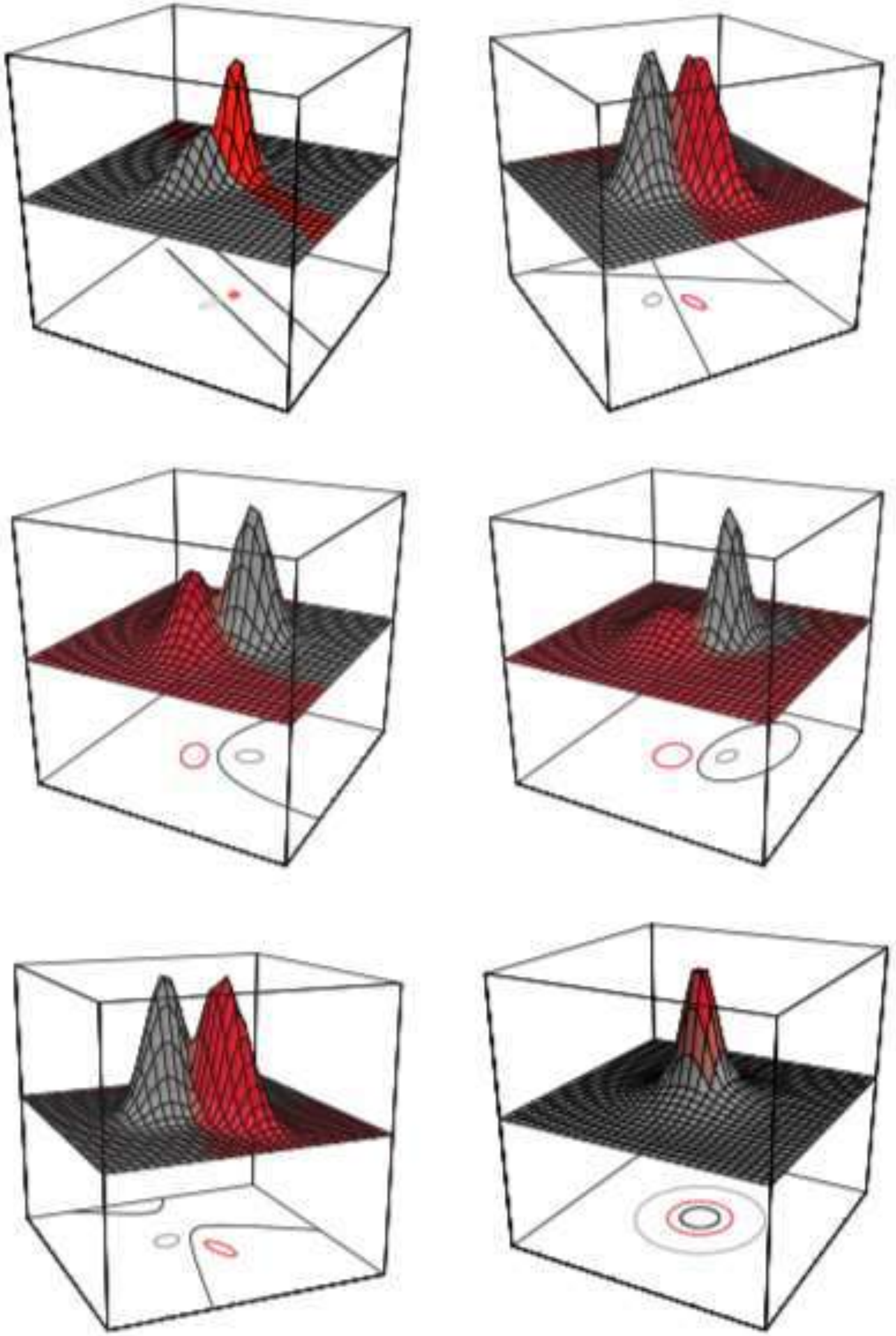


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



See Fig. 2.15 for 3D cases

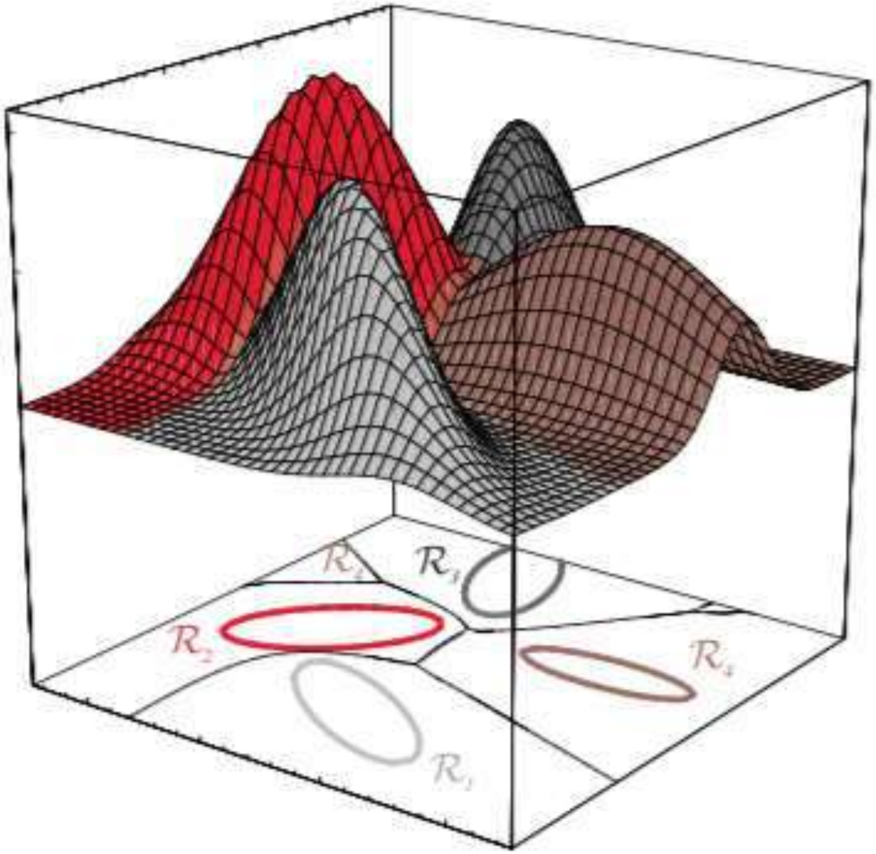
More than Two Categories

Decision Boundary

Defined by two most likely classes for each segment

Other Distributions

Possible; underlying Bayesian Decision Theory is unmodified, however



Discrete Features

Roughly speaking...

Replace probability densities by probability mass functions. Expressions using integrals are changed to use summations, e.g.

$$\int p(\mathbf{x}|\omega_j)\ d\mathbf{x} \qquad\qquad \sum_{x} P(\mathbf{x}|\omega_j)$$

Bayes Formula

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

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

Example: Independent Binary Features

Binary Feature Vector

x = {x1, ..., xd} of 0/1 -valued features, where each xi is 0/1 with probability: $p_i = Pr[x_i = 1|\omega_1]$

Conditional Independence

Assume that *given a class*, the features are independent

Likelihood Function

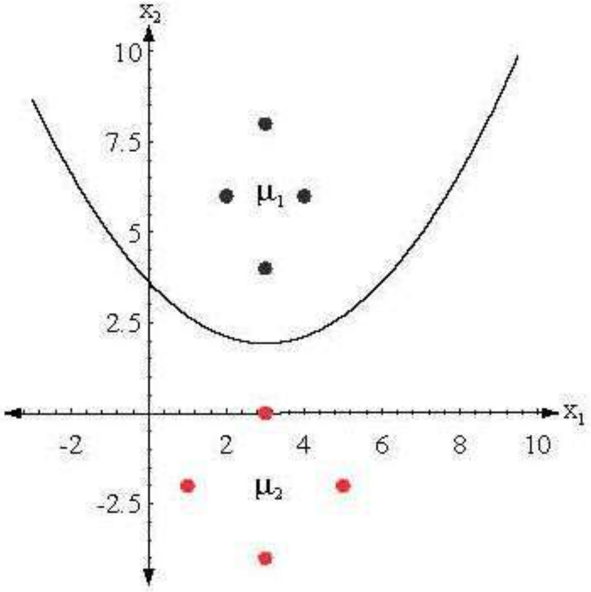
$$P(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

Example

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Decision boundary: $x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$

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



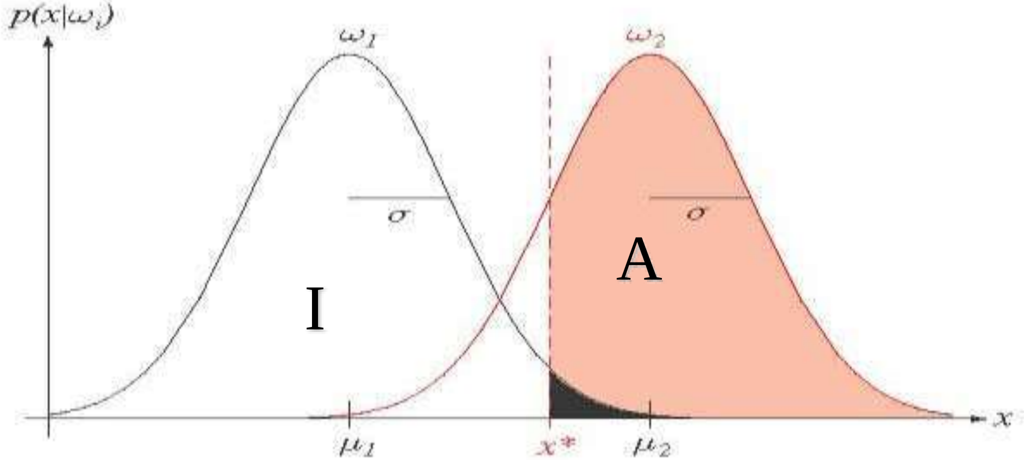
boundary does

not pass through

midpoint of μ_1,μ_2

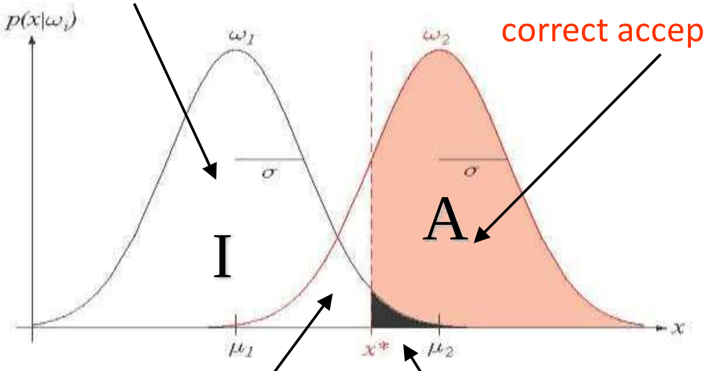
Example: Person Authentication

- Authenticate a person using biometrics (e.g., fingerprints).
- There are two possible distributions (i.e., classes):
  - *Authentic* (A) and *Impostor* (I)



Example: Person Authentication (cont'd)

- Possible decisions:
  - (1) **correct acceptance** (true positive):
    - X belongs to A, and we decide A
  - (2) **incorrect acceptance** (false positive):
    - X belongs to I, and we decide A
  - (3) **correct rejection** (true negative):
    - X belongs to I, and we decide I
  - (4) **incorrect rejection** (false negative):
    - X belongs to A, and we decide I



correct rejection

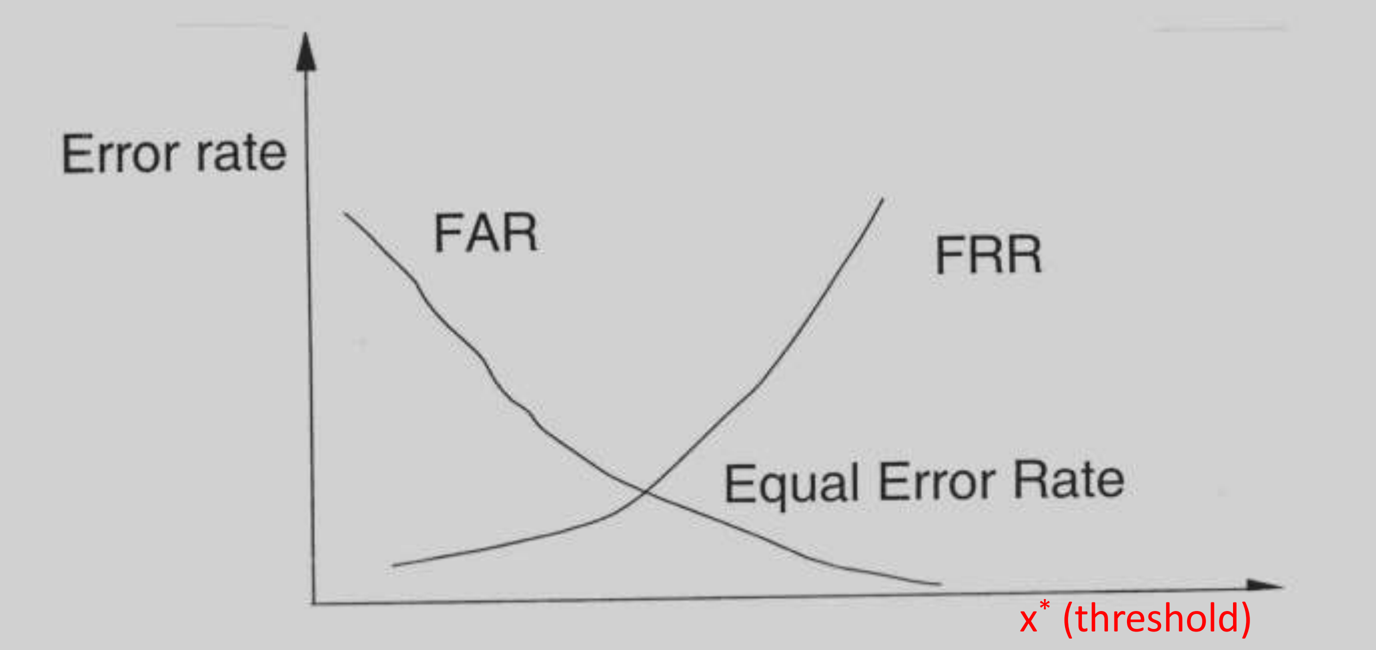
correct acceptance

false negative

false positive

Error vs Threshold

ROC Curve

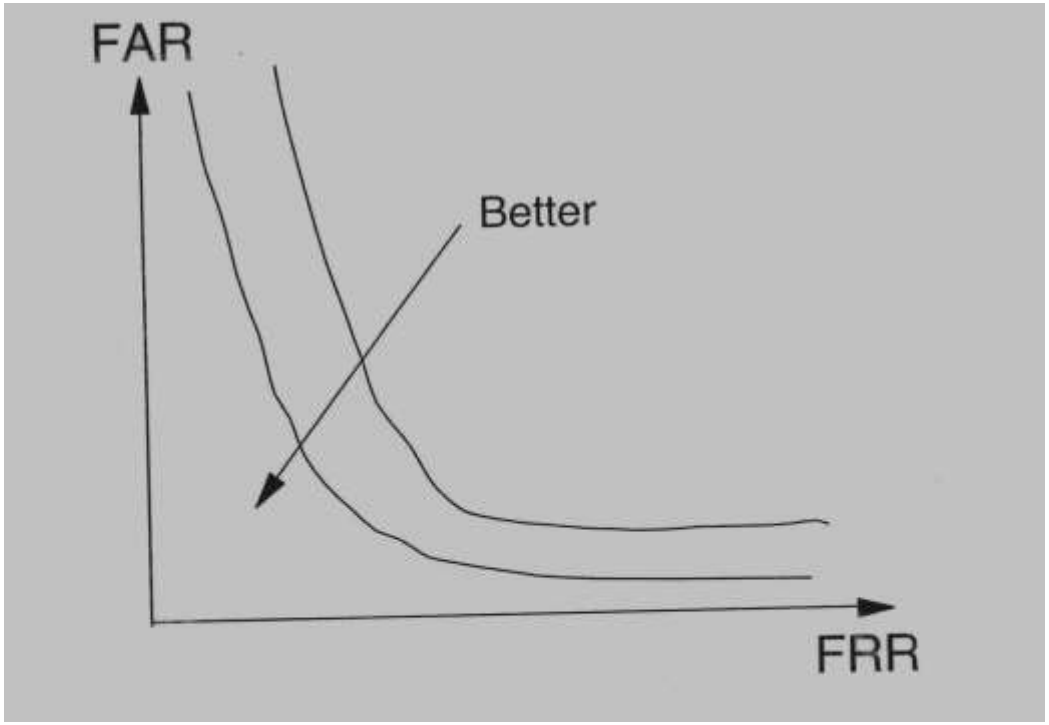


FAR: False Accept Rate (False Positive)

FRR: False Reject Rate (False Negative)

False Negatives vs False Positives

ROC Curve



More common to plot FRR vs FAR

FAR: False Accept Rate (False Positive)

FRR: False Reject Rate (False Negative)

Thank you for your attention