Machine Learning

Lecture 7: Bayesian

杨思蓓

SIST

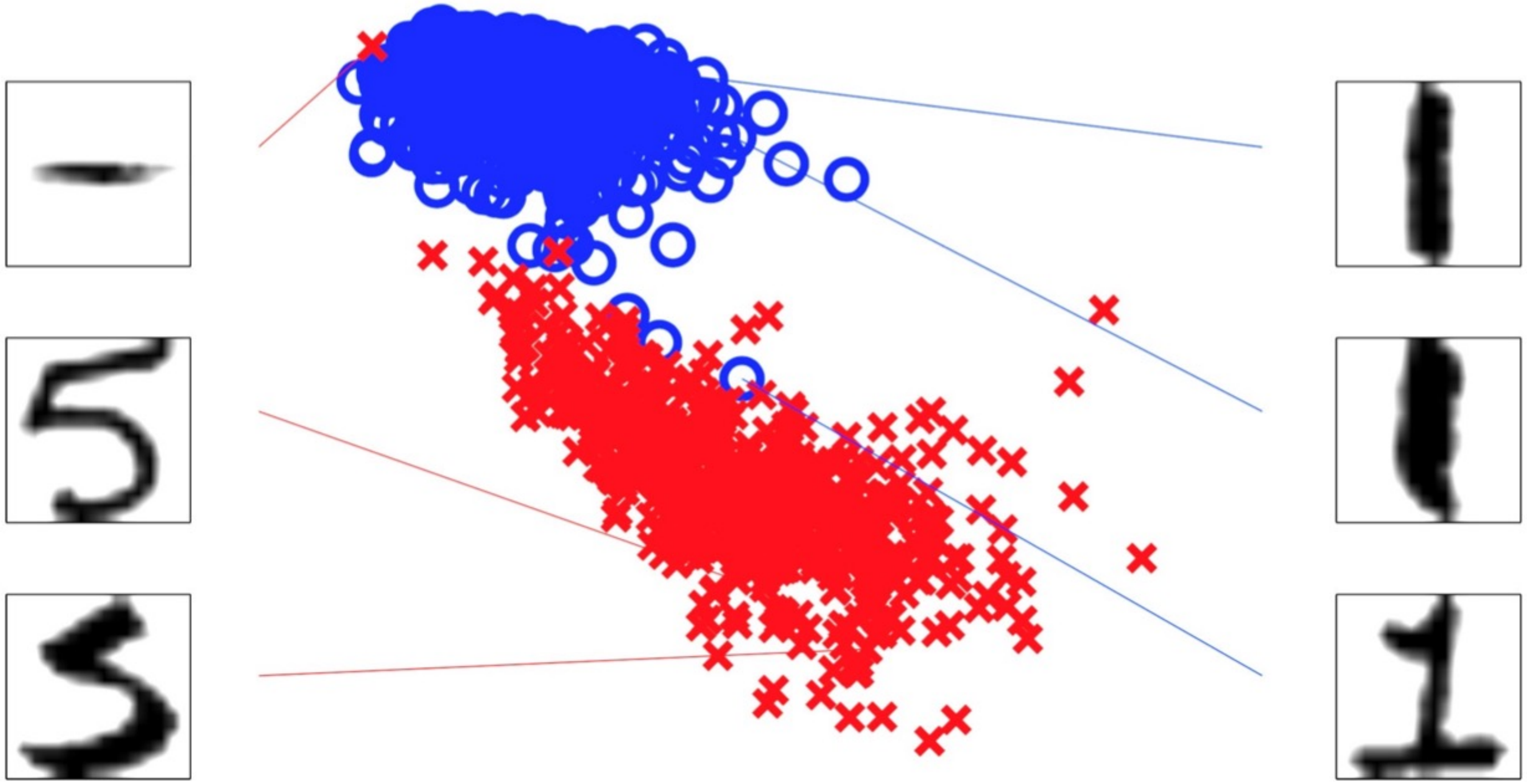
Email: yangsb@shanghaitech.edu.cn

Content

- Bayesian
- Discriminant Function

Basic Concepts of Classification

$\mathbf{x} = (x_0 \;\; x_1, x_2)$ x_1: intensity x_2: symmetry



Bayesian Decision Theory

Sea bass / Salman Example
Decision problem posed in probabilistic terms

x: sample
y: state of the nature, class
$P(y|x)$: given x , what is the probability of the state of the nature.

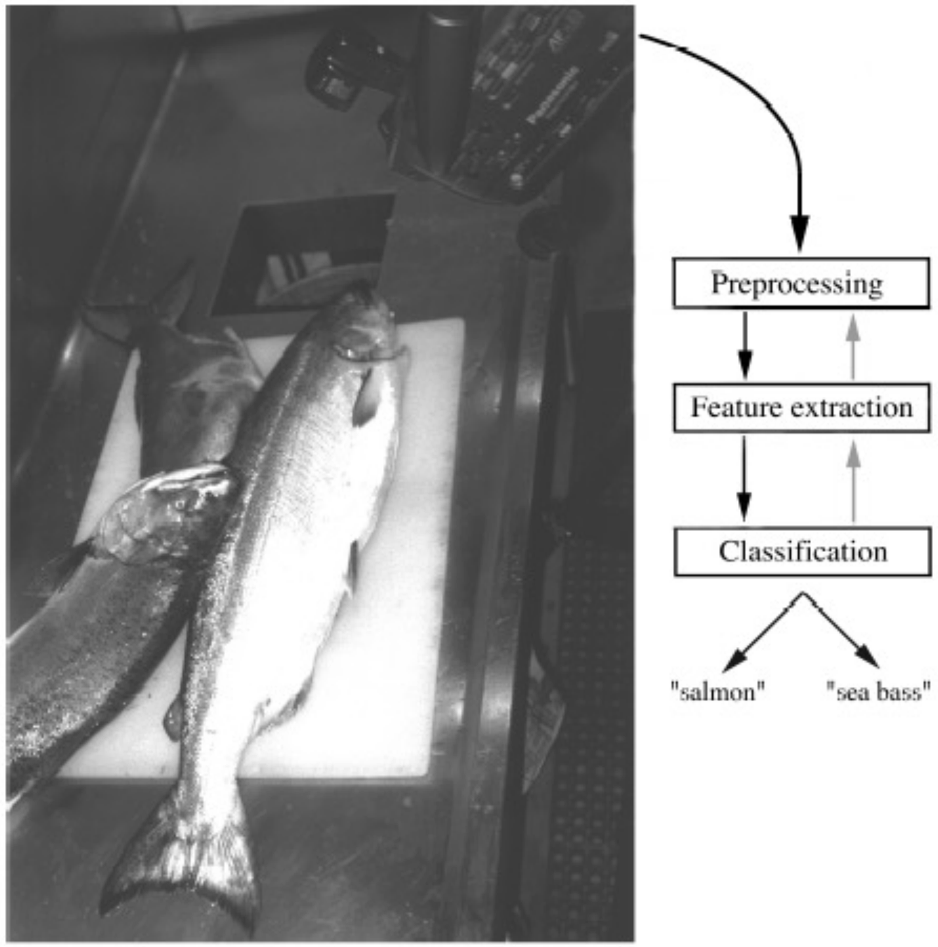


FIGURE 1.1. The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed. Next the features are extracted and finally the classification is emitted, here either "salmon" or "sea bass." Although the information flow is often chosen to be from the source to the classifier, some systems employ information flow in which earlier levels of processing can be altered based on the tentative or preliminary response in later levels (gray arrows). Yet others combine two or more stages into a unified step, such as simultaneous segmentation and feature extraction. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Basics of Probability

An experiment is a well-defined process with observable outcomes.

The set or collection of all outcomes of an experiment is called the sample space, S.

An event E is any subset of outcomes from S.

Probability of an event, P(E) is P(E) = number of outcomes in E / number of outcomes in S.

Bayes’ Theorem

Conditional probability: $P(A|B) = P(A, B)/P(B)$.

Test of Independence: A and B are said to be independent if and only if $P(A, B) = P(A) P(B)$.

Bayes' Theorem: $P(A|B) = P(B|A) P(A)/P(B)$.

Illustration

A	0	0	1	1	1	0
B	0	1	1	0	1	1

$P(A=1) = 3/6 = 1/2$, $P(A=0) = 3/6 = 1/2$.
$P(B=1) = 4/6 = 2/3$, $P(B=0) = 2/6 = 1/3$.
$P(A=1, B = 1) = 2/6 = 1/3$.

$P(A=1 \mid B = 1) = P(A=1, B = 1) / P(B=1) = 1/2$.
$P(B=1 \mid A = 1) = P(B=1, A = 1) / P(A=1) = 2/3$.
$P(A=1 \mid B = 1) P(B=1)/P(A=1) = 2/3 = P(B=1 \mid A = 1)$

Prior

State of nature, a priori (prior) probability

- Random variable (State of nature is unpredictable)
- The catch of salmon and sea bass is equiprobable
  - $P(y_1) = P(y_2)$ (uniform priors)
  - $P(y_1) + P(y_2) = 1$ (exclusivity and exhaustivity)
- Reflects our prior knowledge about how likely we are to observe a sea bass or salmon

Decision rule with only the prior information

- Decide y_1 if $P(y_1) > P(y_2)$, otherwise decide y_2

Likelihood

Suppose now we have a measurement or feature on the state of nature - say the fish lightness value

$P(x|y_1)$ and $P(x|y_2)$ describe the difference in lightness feature between populations of sea bass and salmon

$P\left(x|y_j\right)$ is called the **likelihood** of y_j *with respect to x; the category* y_j *for which* $P(x \mid y_j)$ is large is more likely to be the true category

Maximum likelihood decision

Assign input pattern x to class y_1 if

$$P(x \mid y_1) > P(x \mid y_2), \text{ otherwise } y_2$$

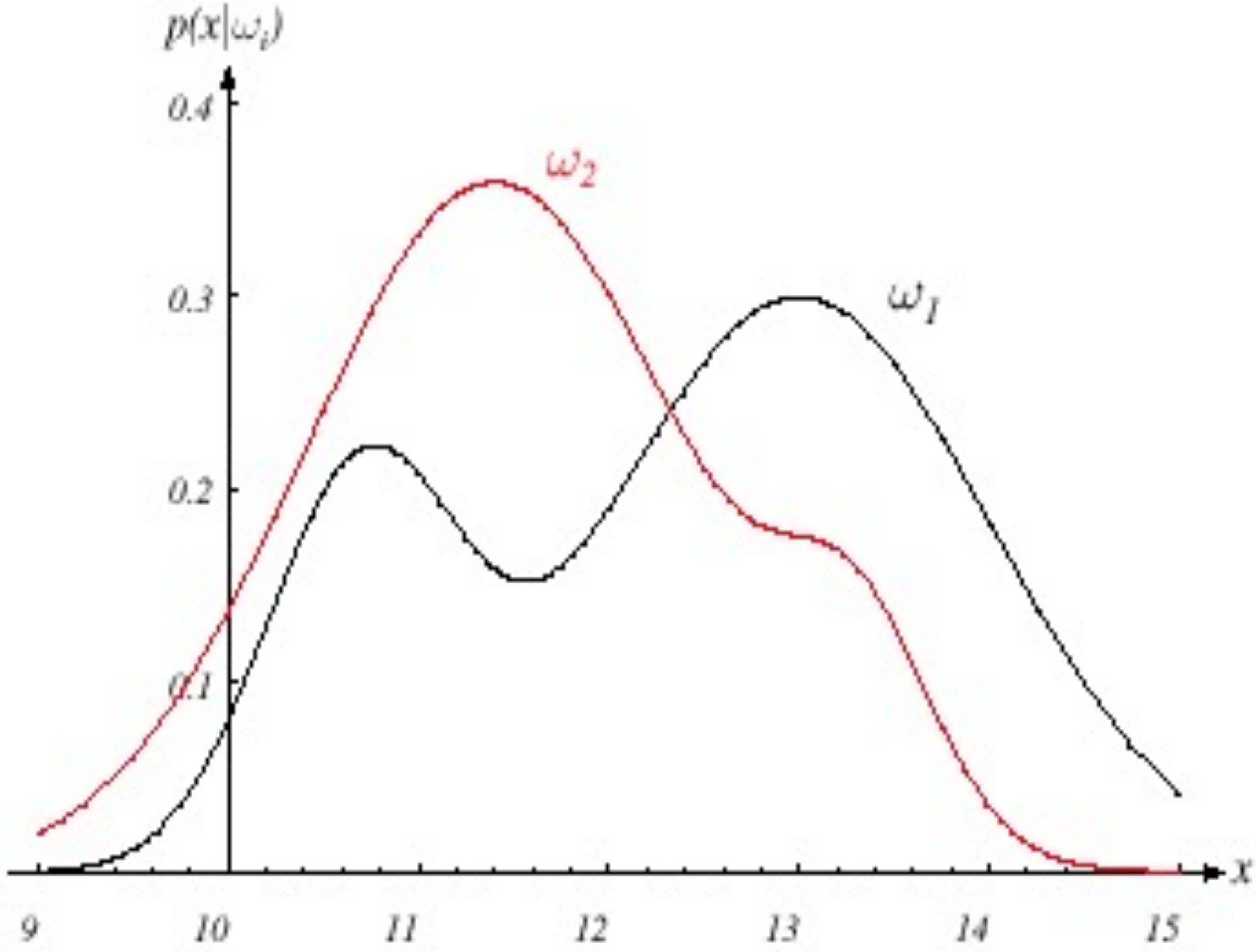


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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, Inc.

Posterior

Bayes formula

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

$$P(x) = \sum_{i=1}^{k} P(x|y_i)P(y_i)$$

Posterior = (Likelihood × Prior) / **Evidence**

- Evidence $P(\boldsymbol{x})$ can be viewed as a scale factor that guarantees that the posterior probabilities sum to 1

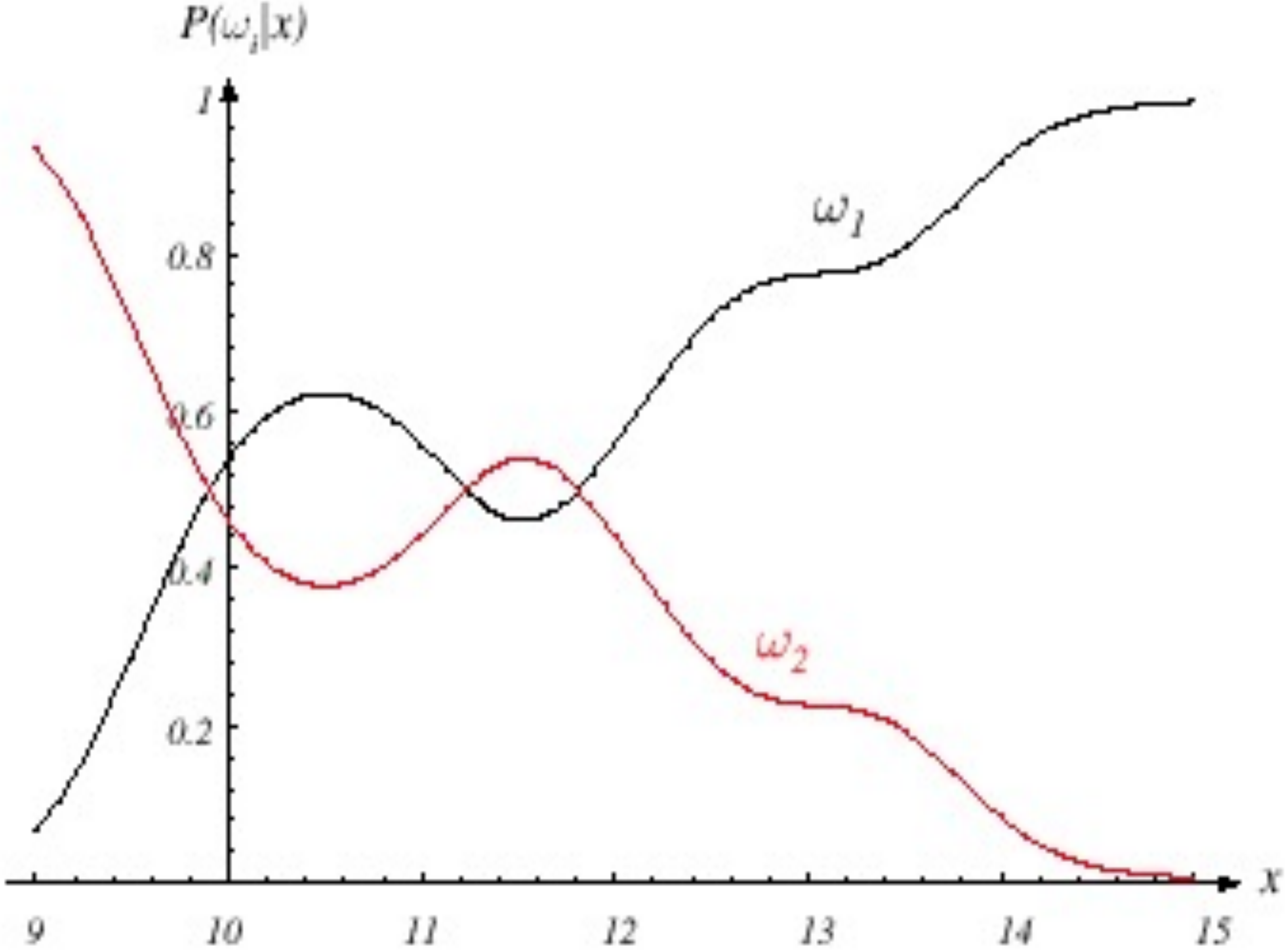


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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.

$P(y_1 \mid x)$ is the probability of the state of nature being y_1 given that feature value x has been observed

Decision given the posterior probabilities, Optimal Bayes Decision rule

X is an observation for which:

if $P(y_1 \mid x) > P(y_2 \mid x)$ ➔ True state of nature = y_1

if $P(y_1 \mid x) < P(y_2 \mid x)$ ➔ True state of nature = y_2

Therefore, whenever we observe a particular x, the probability of error is:

$$P(error \mid x) = P(y_1 \mid x) \text{ if we decide } y_2$$

$$P(error \mid x) = P(y_2 \mid x) \text{ if we decide } y_1$$

Bayes decision rule minimizes the probability of error

Decide y_1 *if* $P(y_1 \mid x) > P(y_2 \mid x)$;
otherwise decide y_2

Therefore:

$$P(error \mid x) = min\ [P(y_1 \mid x), P(y_2 \mid x)]$$

Unconditional error, *P(error) obtained by integration over all x w.r.t. p(x)*

Optimal Bayes Decision Rule

Decide y_1 *if* $P(y_1 \mid x) > P(y_2 \mid x)$;
otherwise decide y_2

Special cases:

(i) $P(y_1) = P(y_2)$; *Decide* y_1 if
$P(x \mid y_1) > P(x \mid y_2)$, otherwise y_2

(ii) $P(x \mid y_1) = P(x \mid y_2)$; *Decide* y_1 if
$P(y_1) > P(y_2)$, otherwise y_2

Bayesian Decision Theory – Generalization

Generalization of the preceding ideas

Use of more than one feature (*p* features)
Use of more than two states of nature (*c* classes)
Allowing other actions besides deciding on the state of nature
Introduce a loss function which is more general than the probability of error

Let $\{y_1, y_2, \ldots, y_c\}$ be the set of c states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of *a* possible actions

Let $\lambda(\alpha_i \mid y_j)$ be the loss incurred for taking action α_i when the true state of nature is y_j

General decision rule $\alpha(\boldsymbol{x})$ specifies which action to take for every possible observation $\boldsymbol{x}$

Bayes Risk

Conditional risk

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} \lambda\left(\alpha_i|y_j\right)P\left(y_j|\boldsymbol{x}\right)$$

Select the action for which the conditional risk $R(\alpha_i|\boldsymbol{x})$ *is minimum*

$$R = \sum_{over\ \boldsymbol{x}} R(\alpha_i|\boldsymbol{x})$$

Risk *R* is minimum and *R* in this case is called the

Bayes risk = best performance that can be achieved!

Two-category classification

α_1 : *deciding* y_1

α_2 : *deciding* y_2

$\lambda_{ij} = \lambda(\alpha_i \mid y_j)$

loss incurred for deciding y_i when the true state of nature is y_j

Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(y_1 \mid x) + \lambda_{12}P(y_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21}P(y_1 \mid x) + \lambda_{22}P(y_2 \mid x)$$

Bayes rule is the following:

if $R(y_1 \mid x) < R(y_2 \mid x)$

action y_1: “decide y_1” is taken

This results in the equivalent rule:

decide y_1 if:

$$(\lambda_{21} - \lambda_{11})\, P(x \mid y_1)\, P(y_1) > (\lambda_{12} - \lambda_{22})\, P(x \mid y_2)\, P(y_2)$$

and decide y_2 otherwise

Likelihood Ratio

The preceding rule is equivalent to the following rule:

If $\frac{P(\boldsymbol{x}|y_1)}{P(\boldsymbol{x}|y_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \times \frac{P(y_2)}{P(y_1)}$

Then take action α_1 (decide *y*$_1$)
Otherwise take action α_2 (decide *y*$_2$)

Optimal Decision Property

$$\frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \times \frac{P(\omega_2)}{P(\omega_1)}$$

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

Minimum Error Rate Classification

Actions are decisions on classes

- If action α_i is taken and the true state of nature is y_j then: the decision is correct if $i = j$ and in error if $i \neq j$

Seek a decision rule that minimizes the **probability of error** or the **error rate**

Zero-one (0-1) loss function: no loss for correct decision and a unit loss for any error

$$\lambda(\alpha_i|y_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Conditional risk:

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

$$= \sum_{j \neq i} P(y_j|\boldsymbol{x}) = 1 - P(y_i|\boldsymbol{x})$$

The risk corresponding to this loss function is the average probability of error

Minimum Error Rate Classification

$$R(\alpha_i|\boldsymbol{x}) = 1 - P(y_i|\boldsymbol{x})$$

Minimizing the risk ➔ Maximizing the posterior $P(y_i|\boldsymbol{x})$

For minimum error rate

Decide y_i if $P\ (y_i\ |\ x) > P(y_j\ |\ x)\ \forall j \neq i$

Decision Boundaries and Regions

Likelihood ratio rule

If $\frac{P(\boldsymbol{x}|y_1)}{P(\boldsymbol{x}|y_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \times \frac{P(y_2)}{P(y_1)}$, then decide *$y_1$*, otherwise decide *$y_2$*

Let $\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \times \frac{P(y_2)}{P(y_1)} = \theta_\lambda$, then decide *$y_1$* if $\frac{P(\boldsymbol{x}|y_2)}{P(\boldsymbol{x}|y_2)} > \theta_\lambda$

The threshold only involves the priors:

If $\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, then $\theta_\lambda = \frac{P(y_2)}{P(y_1)} \equiv \theta_a$

If $\lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$, then $\theta_\lambda = \frac{2P(y_2)}{P(y_1)} \equiv \theta_b$

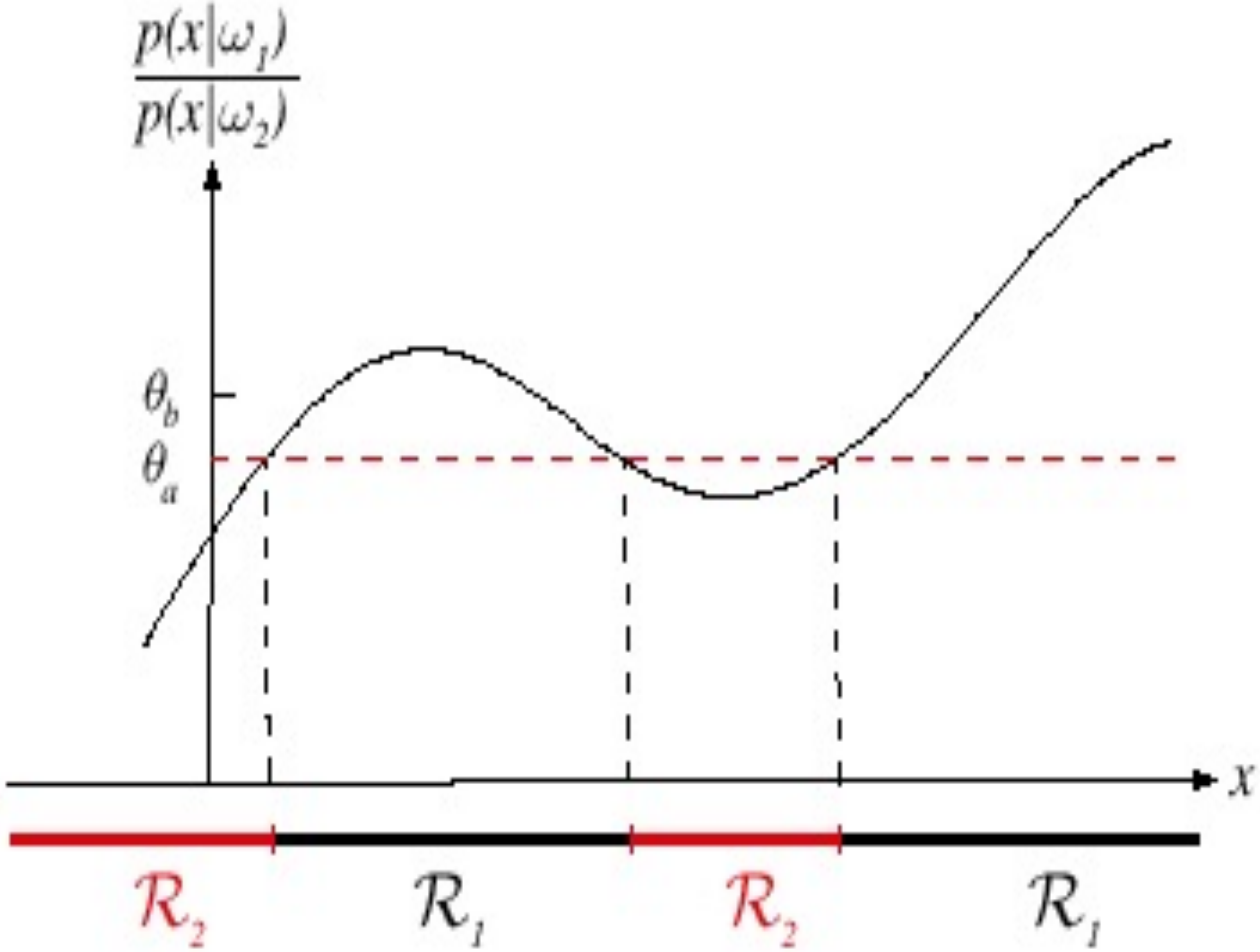


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classifiers, Discriminant Functions and Decision Surfaces

Many different ways to represent pattern classifiers; one of the most useful is in terms of discriminant functions

The multi-category case

Set of discriminant functions: $g_i(\boldsymbol{x}),\ i = 1, \cdots, c$

Classifier assigns a feature vector $\boldsymbol{x}$ to class y_i if:

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \qquad \forall j \neq i$$

Bayes Classifier

What is discriminant functions for Bayes classifier?

Minimize conditional risk $R(\alpha_i|\boldsymbol{x})$

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

(max. discriminant corresponds to min. risk!)

For the minimum error rate, we take

$$g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x})$$

(max. discrimination corresponds to max. posterior!)

$$g_i(\boldsymbol{x}) = P(\boldsymbol{x}|\omega_i)P(\omega_i)$$

$$g_i(\boldsymbol{x}) = \ln P(\boldsymbol{x}|\omega_i) + \ln P(\omega_i)$$

Decision Regions and Surfaces

Effect of any decision rule is to divide the feature space into *c* decision regions

If $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})\ \forall j \neq i$, then $\boldsymbol{x} \in \boldsymbol{\mathcal{R}}_i$
(Region $\boldsymbol{\mathcal{R}}_i$ means assign $\boldsymbol{x}$ to y_i)

The two-class case
Here 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 y_1 *if* $g(x) > 0$; Otherwise decide y_2

The Two-Class Case

$$g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x})$$

$$g(\boldsymbol{x}) = P(y_1|\boldsymbol{x}) - P(y_2|\boldsymbol{x})$$

$$g_i(\boldsymbol{x}) = \ln P(\boldsymbol{x}|y_i) + \ln P(y_i)$$

$$g(\boldsymbol{x}) = \ln \frac{P(\boldsymbol{x}|y_1)}{P(\boldsymbol{x}|y_2)} + \xcancel{\ln \frac{P(y_1)}{P(y_2)}}$$

The Normal Distribution

Normal density is analytically tractable
Continuous density
A number of processes are asymptotically Gaussian
Handwritten characters, speech signals and other patterns can be viewed as randomly corrupted versions of a single typical or prototype (Central Limit theorem)

Univariate density: $N(\mu, \sigma^2)$

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

μ = mean (or expected value) of *x*
σ^2 = variance (or expected squared deviation) of x

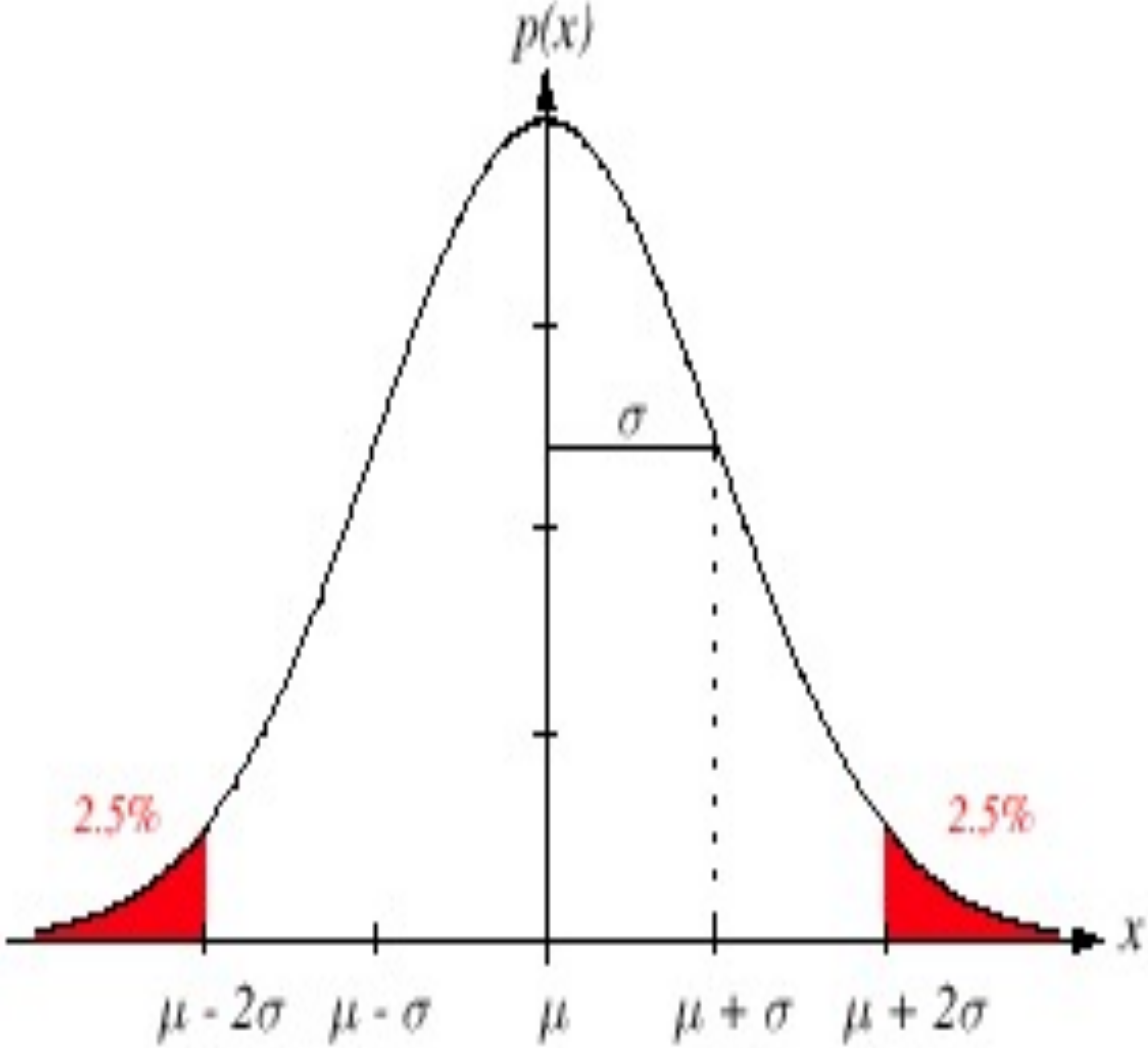


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

Normal Distribution

Multivariate density: $N(\boldsymbol{\mu}, \Sigma)$ (with dimension d)

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

$$\boldsymbol{x} = [x_1, \cdots, x_d]^T$$

$$\boldsymbol{\mu} = [\mu_1, \cdots, \mu_d]^T$$

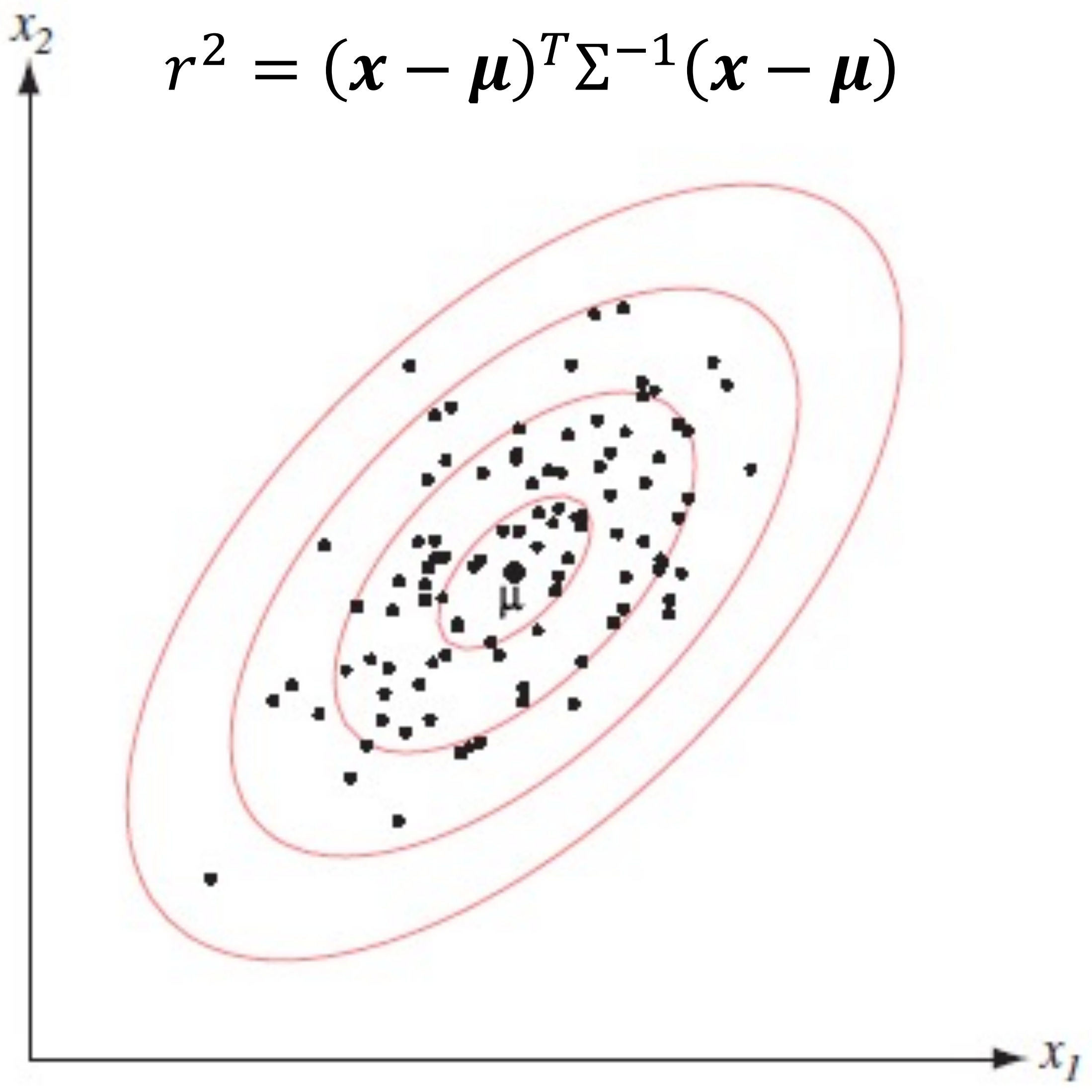
Σ: $d \times d$ covariance matrix, $|\cdot|$: determinant

The covariance matrix is always symmetric and positive semidefinite; we assume Σ is positive definite so the determinant of Σ is strictly positive

The multivariate normal density is completely specified by d + d(d+1)/2 parameters

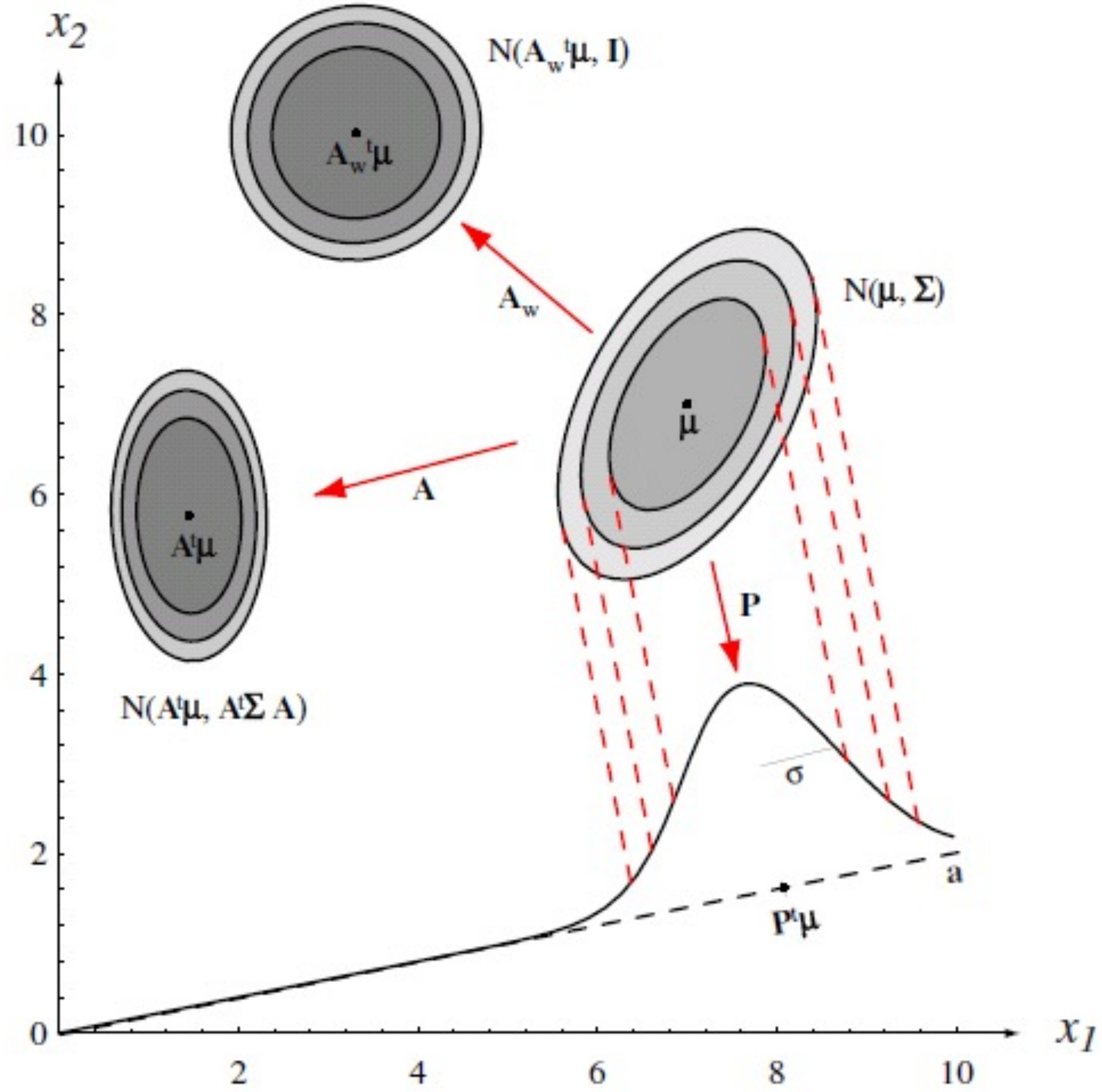
If x_1 and x_2 are statistically independent then the covariance of x_1 and x_2 is zero.

Multivariate Normal density



$$r^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

Transformation of Normal Variable



Discriminant Functions for the Normal Density

The minimum error-rate classification can be achieved by the discriminant function

$$g_i(\boldsymbol{x}) = \ln P(\boldsymbol{x}|\omega_i) + \ln P(\omega_i)$$

In case of multivariate normal densities

$$P(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]$$

$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$

Case $\boldsymbol{\Sigma_i = \sigma^2 I}$

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Features are statistically independent and each feature has the same variance

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

$$= -\frac{1}{2\sigma^2}\left(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\boldsymbol{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}\right) + \ln P(\omega_i)$$

Case $\boldsymbol{\Sigma_i = \sigma^2 I}$

$$g_i(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\left(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\boldsymbol{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right) + \ln P(\omega_i)$$

Equivalent to

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

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

Linear discriminant function

Case $\boldsymbol{\Sigma_i = \sigma^2 I}$

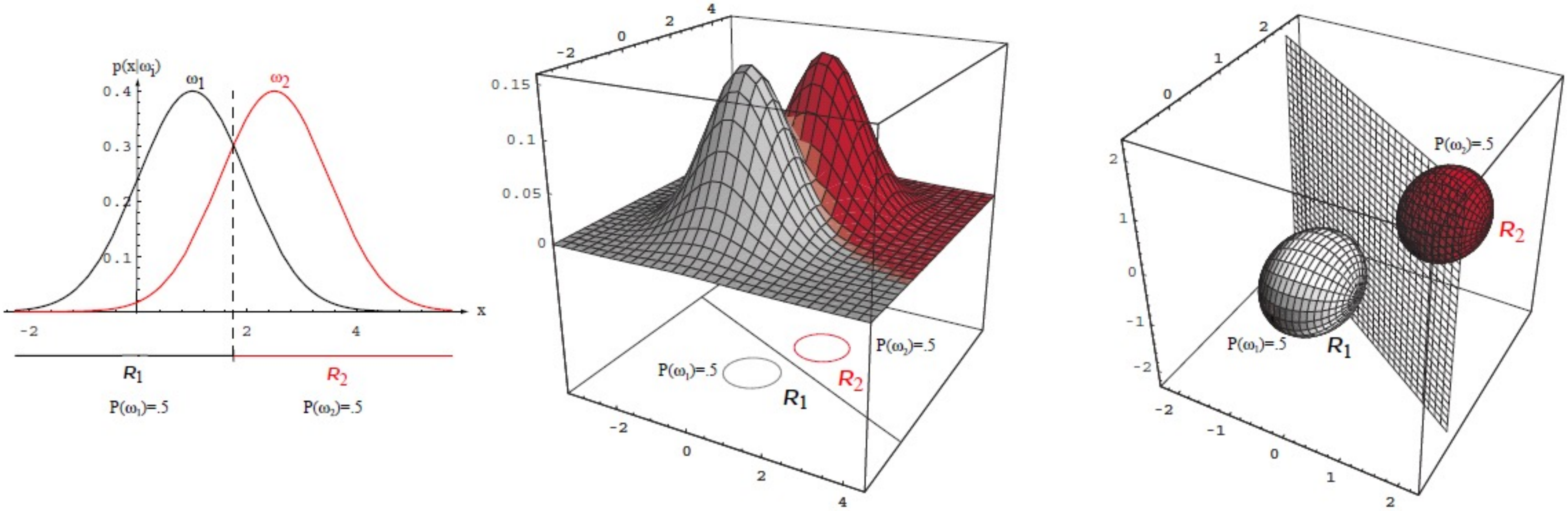
The decision surfaces for a linear machine are pieces of hyperplanes defined by the linear equations:

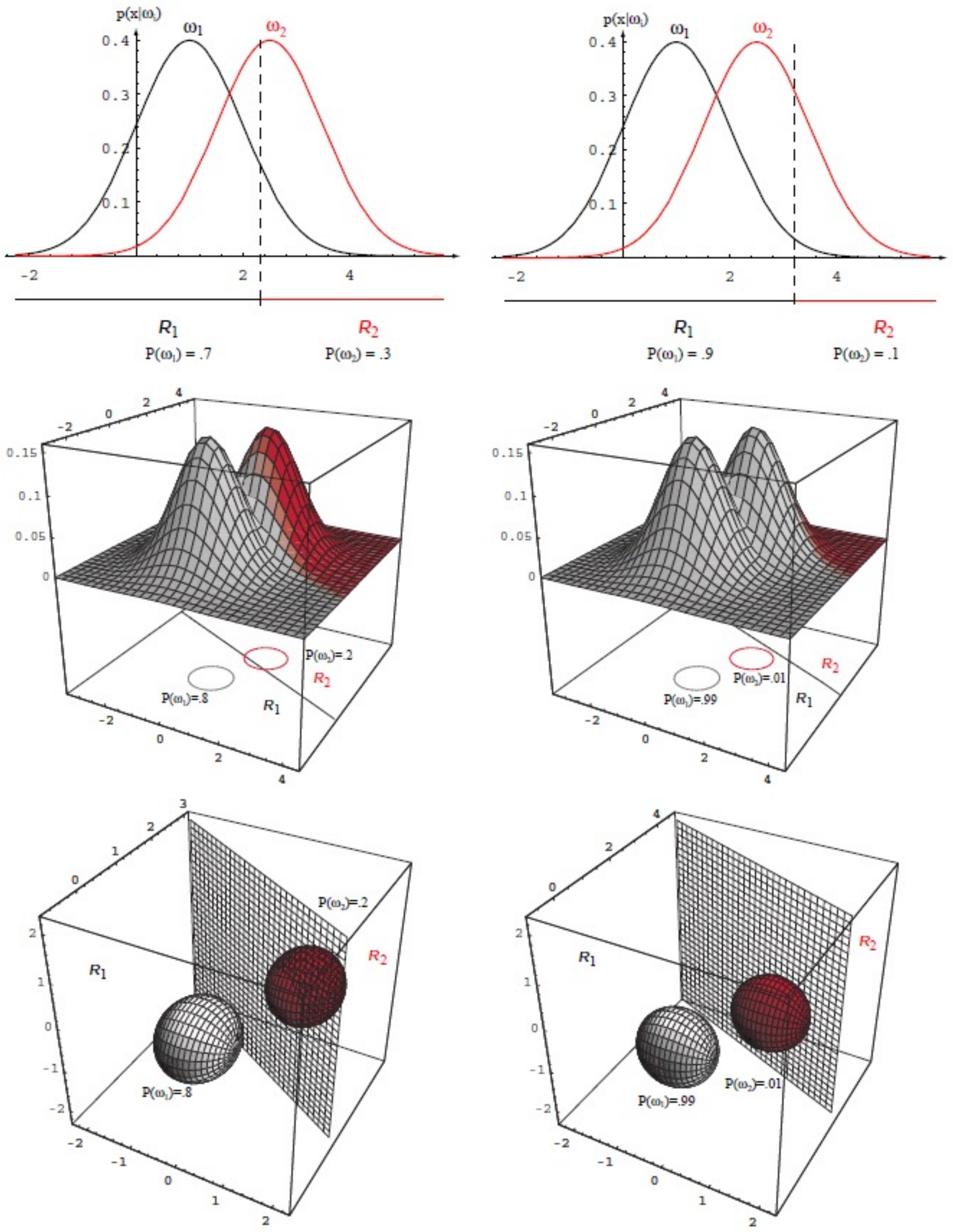
$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

$$0 = \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}\right)^T \boldsymbol{x} - \frac{\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j}{2\sigma^2} + \ln \frac{P(\omega_i)}{P(\omega_j)}$$

If $P(\omega_i) = P(\omega_j)$

$$\boldsymbol{x}_0 = \frac{1}{2}\left(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j\right)$$


p(x|ω_i)
ω_1
ω_2
0.4
0.3
0.2
0.1
x
-2
2
4
R_1
R_2
P(ω_1)=.5
P(ω_2)=.5
0.15
0.1
0.05
0
-2
0
2
4
P(ω_1)=.5
P(ω_2)=.5
R_1
R_2
-2
0
2
4
2
1
0
-1
-2
2
1
0
P(ω_2)=.5
R_2
P(ω_1)=.5
R_1
-2
-1
0
1
2


p(x|ω_i)
ω1
ω2
0.4
0.3
0.2
0.1
-2
2
4
R1
R2
P(ω1) = .7
P(ω2) = .3
p(x|ω_i)
ω1
ω2
0.4
0.3
0.2
0.1
-2
2
4
R1
R2
P(ω1) = .9
P(ω2) = .1
0.15
0.1
0.05
0
-2
0
2
4
P(ω2)=.2
R2
P(ω1)=.8
R1
-2
0
2
4
0.15
0.1
0.05
0
-2
0
2
4
R2
P(ω2)=.01
P(ω1)=.99
R1
-2
0
2
4
3
2
1
0
2
1
0
-1
-2
P(ω2)=.2
R2
R1
P(ω1)=.8
-2
-1
0
1
2
4
2
0
2
1
0
-1
-2
R2
R1
P(ω2)=.01
P(ω1)=.99
-2
-1
0
1
2

Case $\boldsymbol{\Sigma_i = \Sigma}$:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Covariance matrices of all classes are identical but can be arbitrary

$$g_i(\boldsymbol{x}) = -\frac{1}{2}\left(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i\right) + \ln P(\omega_i)$$

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

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

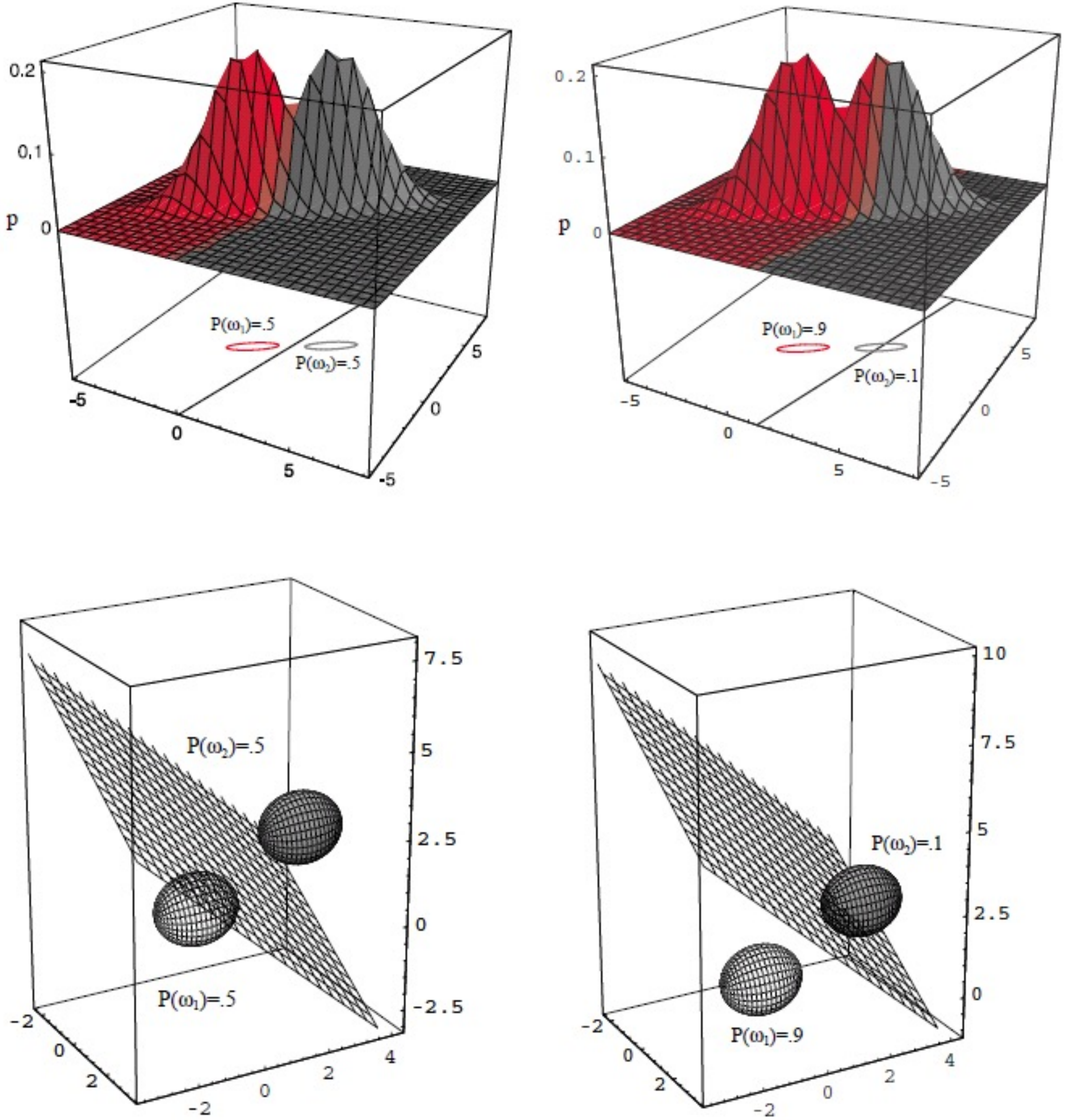
Linear Discriminant Analysis

Case $\boldsymbol{\Sigma_i = \Sigma}$: Linear Discriminant Analysis

Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

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


0.2
0.1
p
0
P(ω1)=.5
P(ω2)=.5
-5
0
5
5
0
-5
0.2
0.1
p
0
P(ω1)=.9
P(ω2)=.1
-5
0
5
5
0
-5
7.5
P(ω2)=.5
5
2.5
0
P(ω1)=.5
-2.5
-2
0
2
-2
0
2
4
10
7.5
5
P(ω2)=.1
2.5
0
P(ω1)=.9
-2
0
2
-2
0
2
4

Case $\Sigma_i = \Sigma$: Linear Discriminant Analysis

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

Estimating Parameters

$\boldsymbol{\mu}_i$

$$\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{j \in \omega_i} \boldsymbol{x}_j$$

$P(\omega_i)$

$$P(\omega_i) = \frac{N_i}{N}$$

Σ

$$\Sigma = \sum_{i=1}^{c} \sum_{j \in \omega_i} \frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T}{N - c}$$

Case Σ_i = arbitrary

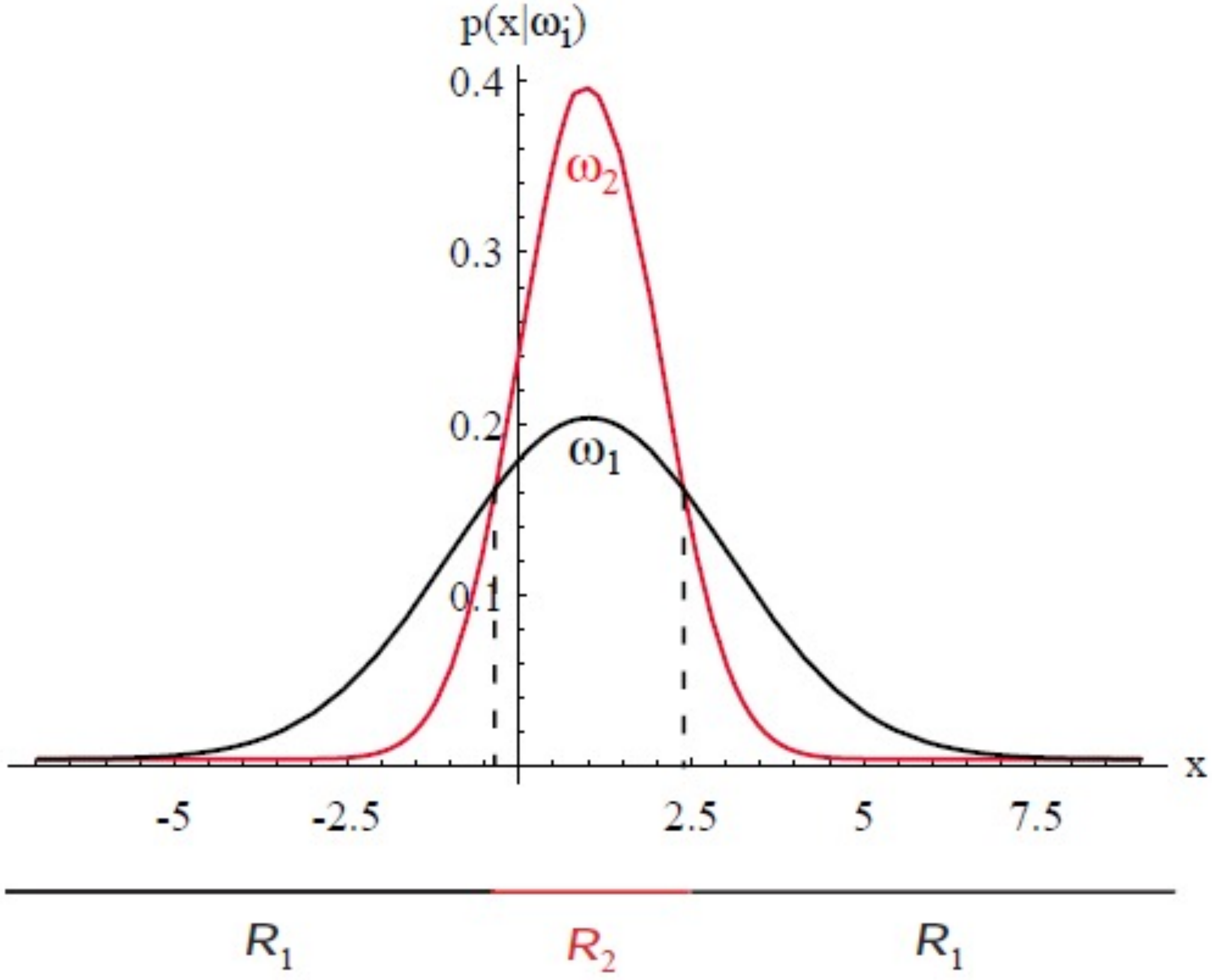
$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

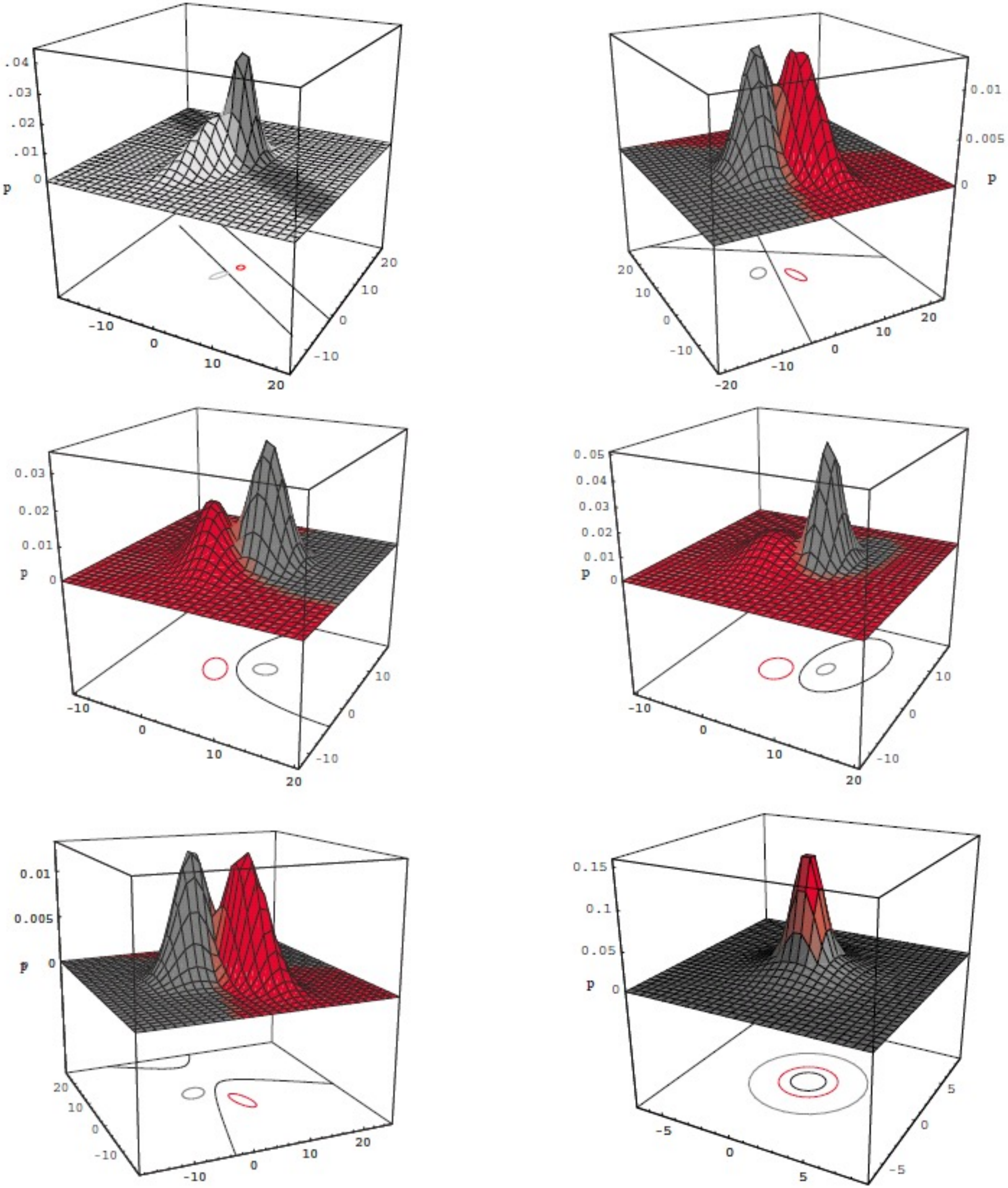
The covariance matrices are different for each category

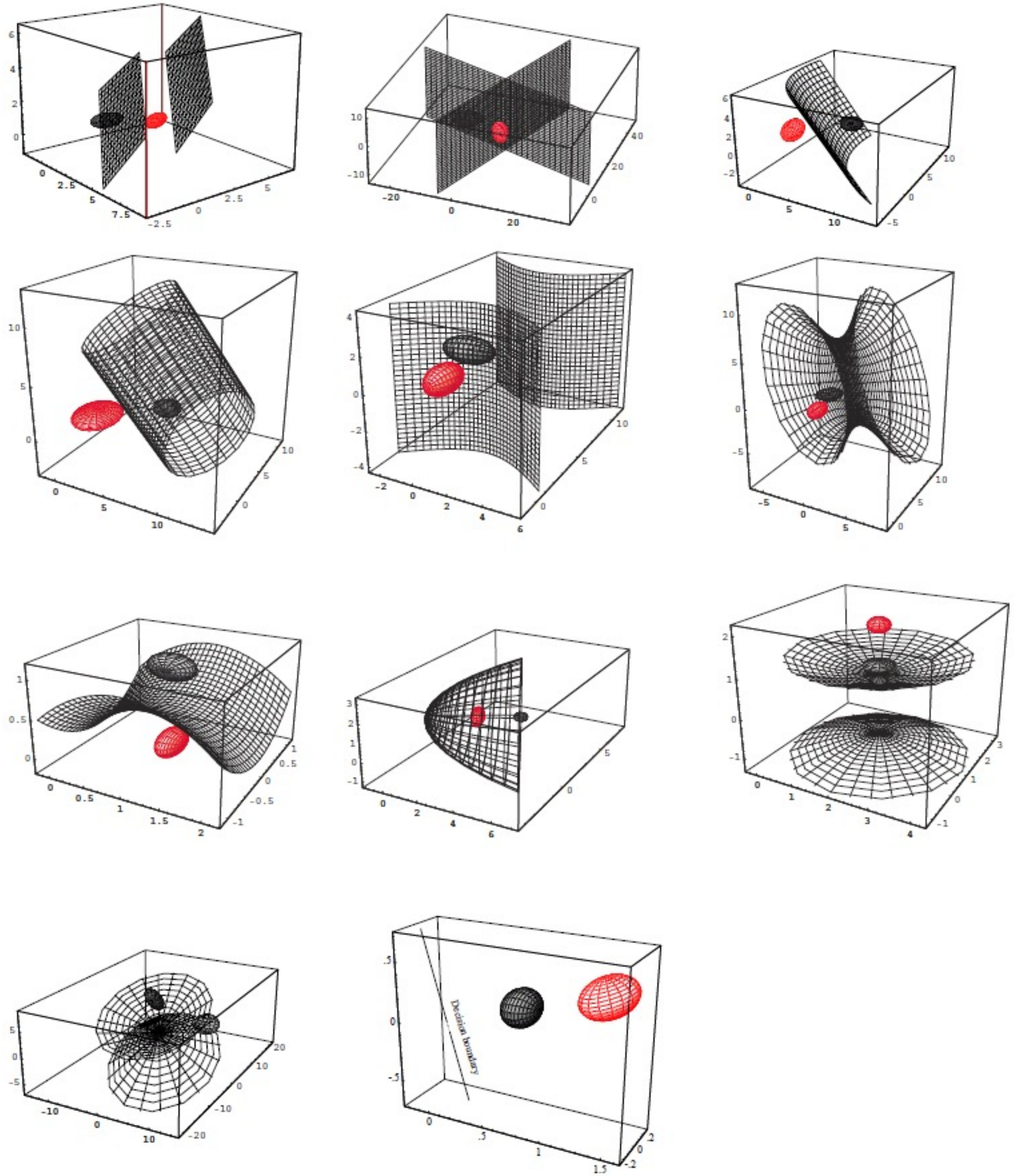
$$g_i(\boldsymbol{x}) = -\frac{1}{2}\left(\boldsymbol{x}^T\Sigma_i^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i\right) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^T W_i \boldsymbol{x} + \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

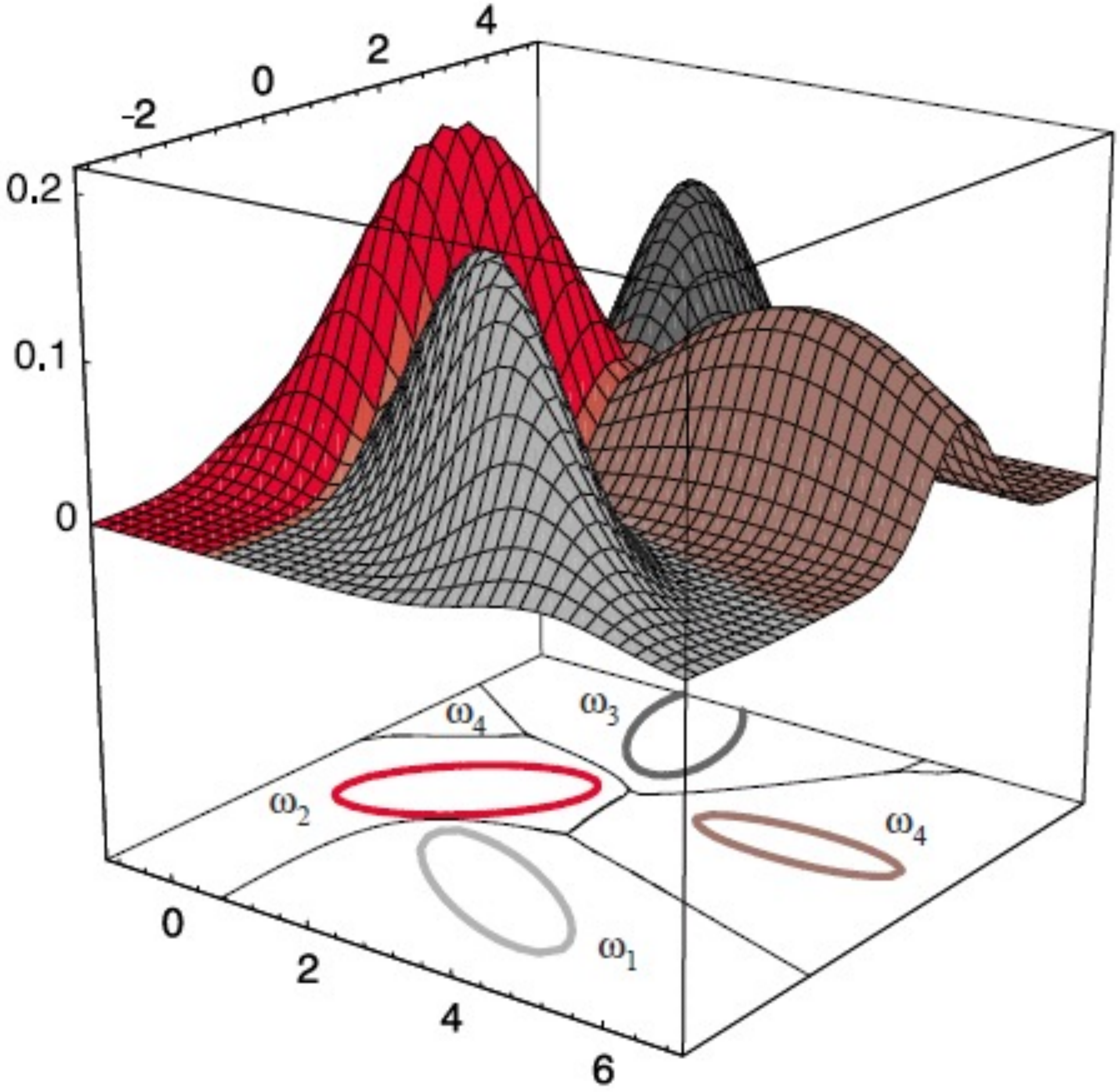
Quadratic Discriminant Analysis


$p(x|\omega_i)$
0.4
0.3
0.2
0.1
ω_2
ω_1
x
-5
-2.5
2.5
5
7.5
R_1
R_2
R_1


.04
.03
.02
.01
0
p
-10
0
10
20
20
10
0
-10
0.01
0.005
0
p
20
10
0
-10
-20
-10
0
10
20
0.03
0.02
0.01
p
0
-10
0
10
20
10
0
-10
0.05
0.04
0.03
0.02
0.01
p
0
-10
0
10
20
10
0
-10
0.01
0.005
p
0
20
10
0
-10
-10
0
10
20
0.15
0.1
0.05
p
0
-5
0
5
5
0
-5


Decision boundary

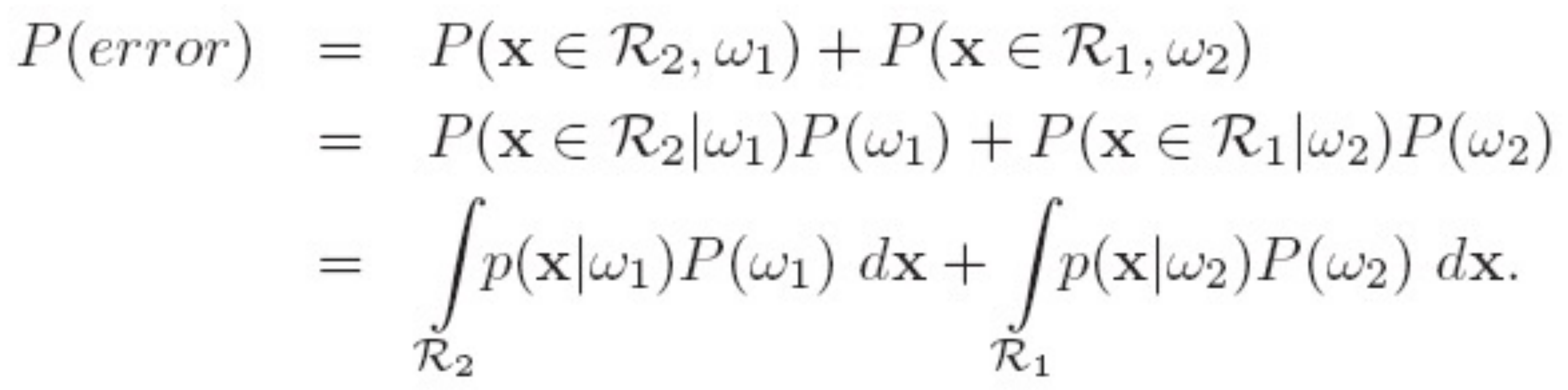
Discriminant Functions for the Normal Density



Error Probabilities and Integrals

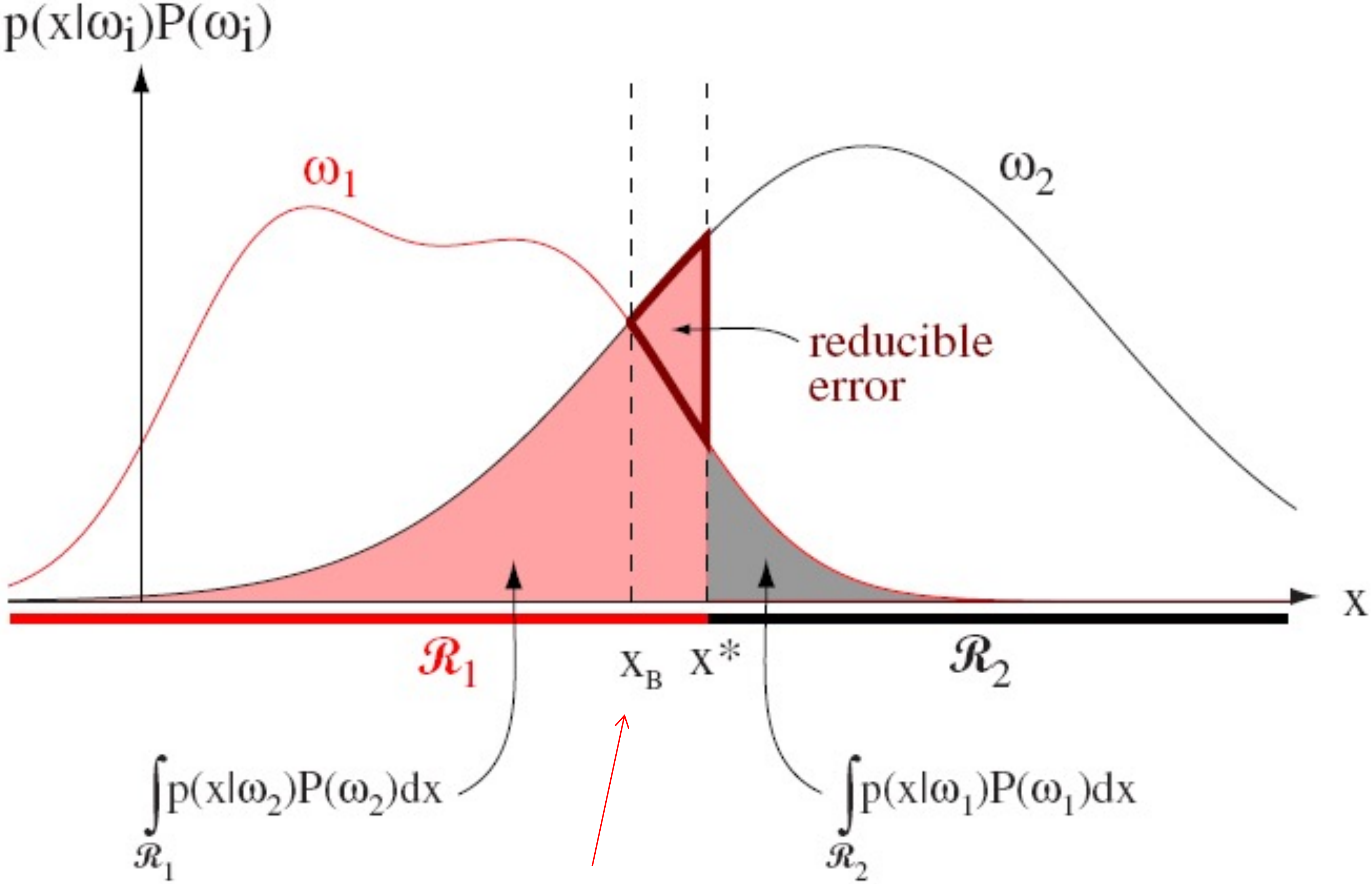
2-class problem

There are two types of errors



$$
\begin{aligned}
P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\
&= P(\mathbf{x} \in \mathcal{R}_2 | \omega_1) P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1 | \omega_2) P(\omega_2) \\
&= \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) P(\omega_1)\ d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) P(\omega_2)\ d\mathbf{x}.
\end{aligned}
$$

Error Probabilities and Integrals



Bayes optimal decision boundary in 1-D case

Figure 2.17: Components of the probability of error for equal priors and (non-optimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 68. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible — this is the Bayes decision and gives the Bayes error rate.

Error Probabilities and Integrals

- Multi-class problem
  - Simpler to computer the prob. of being correct (more ways to be wrong than to be right)

$$\begin{aligned} P(correct) &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, \omega_i) \\ &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i | \omega_i) P(\omega_i) \\ &= \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|\omega_i) P(\omega_i) \, d\mathbf{x}. \end{aligned}$$