



Machine learning

Bayesian Decision Theory

Mohammad-Reza A. Dehaqani

dehaqani@ut.ac.ir

Bayesian Decision Theory



- Bayesian Decision Theory is a fundamental statistical approach that quantifies the tradeoffs 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 Revisited



- **State of nature** is a random variable.
 - Define w as the **type** of fish we observe (state of nature, class) where
 - $w = w_1$ for sea bass,
 - $w = w_2$ for salmon.
- $P(w_1)$ is the a priori probability that the next fish is a sea bass.
- $P(w_2)$ 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 $P(w_1)$ and $P(w_2)$?
 - Set $P(w_1) = P(w_2)$ 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
 - $P(w_1) + P(w_2) = 1$
- (**exclusivity** and **exhaustivity**).

Approaches for building a classifier.



- In classification, the goal is to find a mapping from inputs X to outputs w given a labeled set of input-output pairs (training set)
 - Binary classification (two classes labels)
 - Multi-class classification.
- Approaches for building a classifier.
 - **Generative approach:** This approach creates a joint model of feature vectors and classes ($p(x,w)$) and then drive $p(w_j|k)$
 - **Discriminative approach:** This approach creates a model of the form of $p(w_j|k)$ directly.



Generative Classifiers:

- **Definition:** Model the joint probability $P(X, Y)$, i.e., the distribution of features $\textcolor{blue}{X}$ and labels $\textcolor{blue}{Y}$.
- **Examples:** Naive Bayes, Hidden Markov Models (HMM).
- **Interpretation:**

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Generative model first learns the class-conditional distribution $P(X|Y)$ and the prior $\textcolor{blue}{P(Y)}$, then applies Bayes' rule to predict the class.

Discriminative Classifiers:

- **Definition:** Directly model the conditional probability $P(Y|X)$ without modeling $\textcolor{blue}{P(X)}$
- **Examples:** Logistic Regression, SVM, Neural Networks.
- **Interpretation:**

$P(Y|X)$ directly predicted from the data.

- **Discriminative model** focuses on finding decision boundaries that directly classify $\textcolor{blue}{X}$ into class $\textcolor{blue}{Y}$.



Making a Decision

- How can we make a decision with only the prior information? (**Decision Rule**)

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

- What is the **probability of error** for this decision?

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

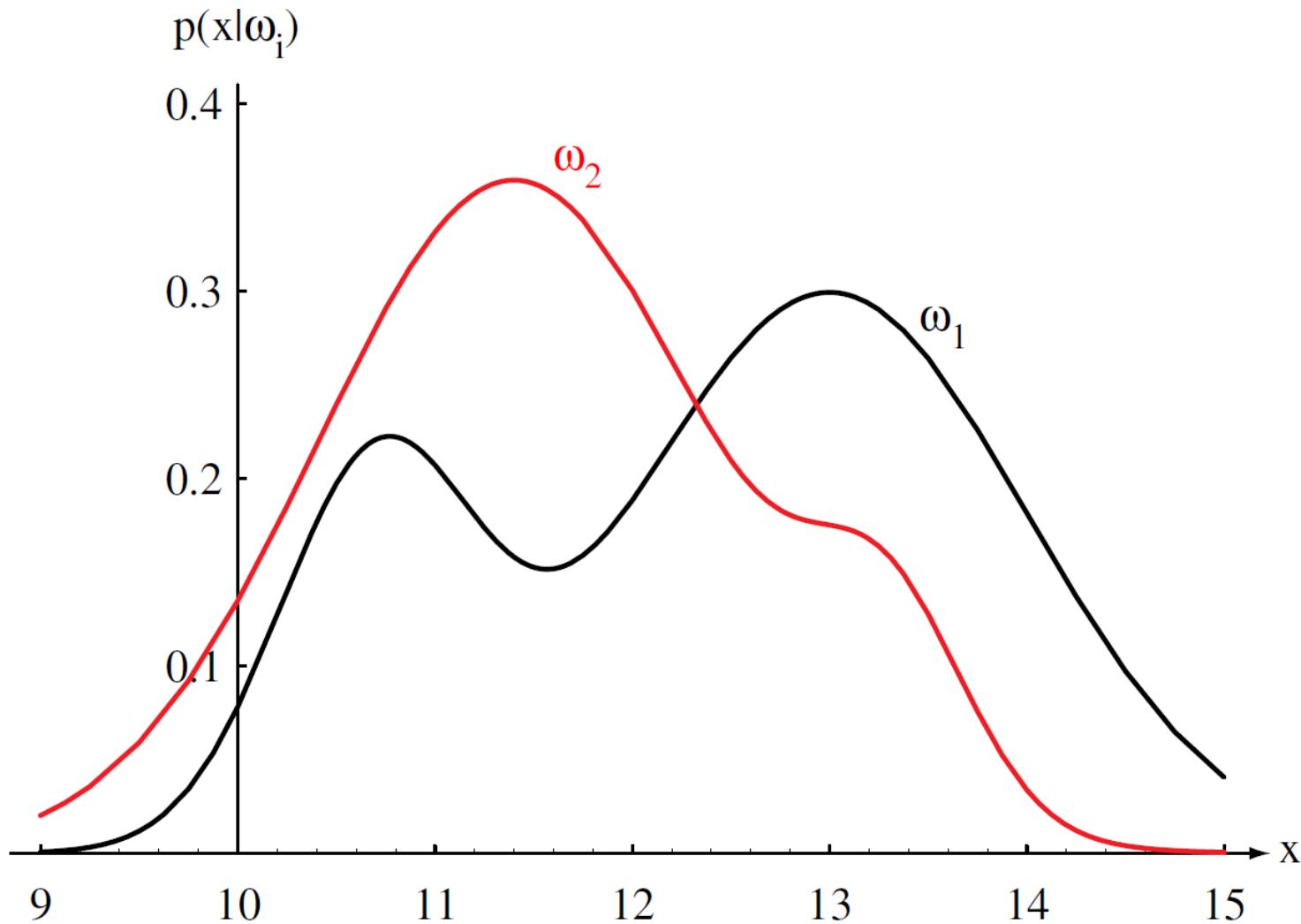
Class-Conditional Probabilities



- Let's try to **improve** the decision using the **lightness measurement** x . (importing **data** to decision model)
 - Let x be a **continuous random** variable.
- Define $p(x|w_j)$ as the **class-conditional probability** density (probability of x given that the **state of nature** is w_j for $j = 1, 2$).
- $p(x|w_1)$ and $p(x|w_2)$ describe the **difference** in lightness between populations of sea bass and salmon.



Hypothetical class-conditional probability density functions for two classes.





Posterior Probabilities

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

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

- Where $p(x) = \sum_{j=1}^2 p(x|\omega_j)P(\omega_j)$.



Bayes' formula

- Bayes' formula can be expressed informally in English by saying that

$$posterior = \frac{likelihood \times prior}{evidence}.$$

- $p(x|w_j)$ is called the **likelihood** and $p(x)$ is called the **evidence**.



Making a Decision

- How can we make a decision **after observing the value of x ?**

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

- Rewriting the rule gives

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

- Note that, at every x , $P(w_1|x) + P(w_2|x) = 1$.



Evidence

- The *evidence*, $p(x)$ is **unimportant** as far as **making a decision is concerned**
- It is basically just a **scale factor** that states how frequently we will actually measure a pattern with feature value x ;
- By eliminating this scale factor, we obtain the following completely equivalent decision rule:

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



Where do Probabilities come from?

- There are two competitive answers:

Relative frequency (objective**) approach.**

- Probabilities can only come **from experiments**.

Bayesian (subjective**) approach.**

- Probabilities may reflect **degree of belief** and can be based on opinion.



Probability of Error (misclassification)

- What is the **probability of error** for this decision?

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

- What is the **average probability of error**?

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

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

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



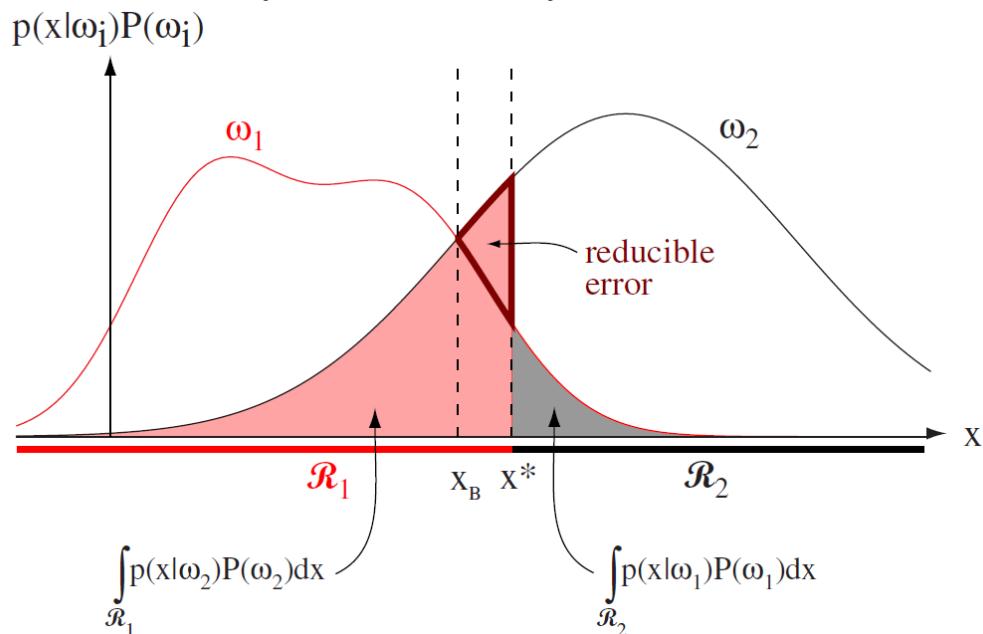
Bayesian classifier

The Bayesian classifier is optimal with respect to minimizing the classification probability

Decision region:

Let $R_1(R_2)$ be the region in the feature space in which we decide in favor of $w_1(w_2)$.

$$P(\text{error}) = P(x \in \mathcal{R}_1, \omega_2) + P(x \in \mathcal{R}_2, \omega_1)$$



$$= \int_{\mathcal{R}_1} P(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R}_2} P(x|\omega_1)P(\omega_1)dx$$



Proof of optimality

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

Since $\mathcal{R}_1 \cup \mathcal{R}_2$ covers all the feature space ,we have

$$P(\omega_1) = \int_{\mathcal{R}_1} P(\omega_1|x)p(x)dx + \int_{\mathcal{R}_2} P(\omega_1|x)p(x) dx$$

The probability of error equals to

$$P(\text{error}) = \int_{\mathcal{R}_1} P(\omega_2|x)p(x)dx + \underbrace{\int_{\mathcal{R}_2} P(\omega_1|x)p(x))dx}_{+ \int_{\mathcal{R}_1} P(\omega_1|x)p(x)dx - \int_{\mathcal{R}_1} P(\omega_1|x)p(x)dx}$$

$$P(\text{error}) = P(\omega_1) - \int_{\mathcal{R}_1} [P(\omega_1|x) - P(\omega_2|x)]p(x)dx$$



Proof of optimality

- The probability of error equals to

$$P(error) = P(\omega_1) - \int_{\mathcal{R}_1} [P(\omega_1|x) - P(\omega_2|x)]p(x)dx$$

- The probability of error is minimized if R_1 is the region of the space in which

$$[P(\omega_1|x) - P(\omega_2|x)] > 0 \rightarrow x \in \omega_1$$

- Then R_2 becomes the region where the reverse is true, i.e. is the region of the space in which

$$[P(\omega_1|x) - P(\omega_2|x)] < 0 \rightarrow x \in \omega_2$$

x can belong to R1 or R2 (not both of them)



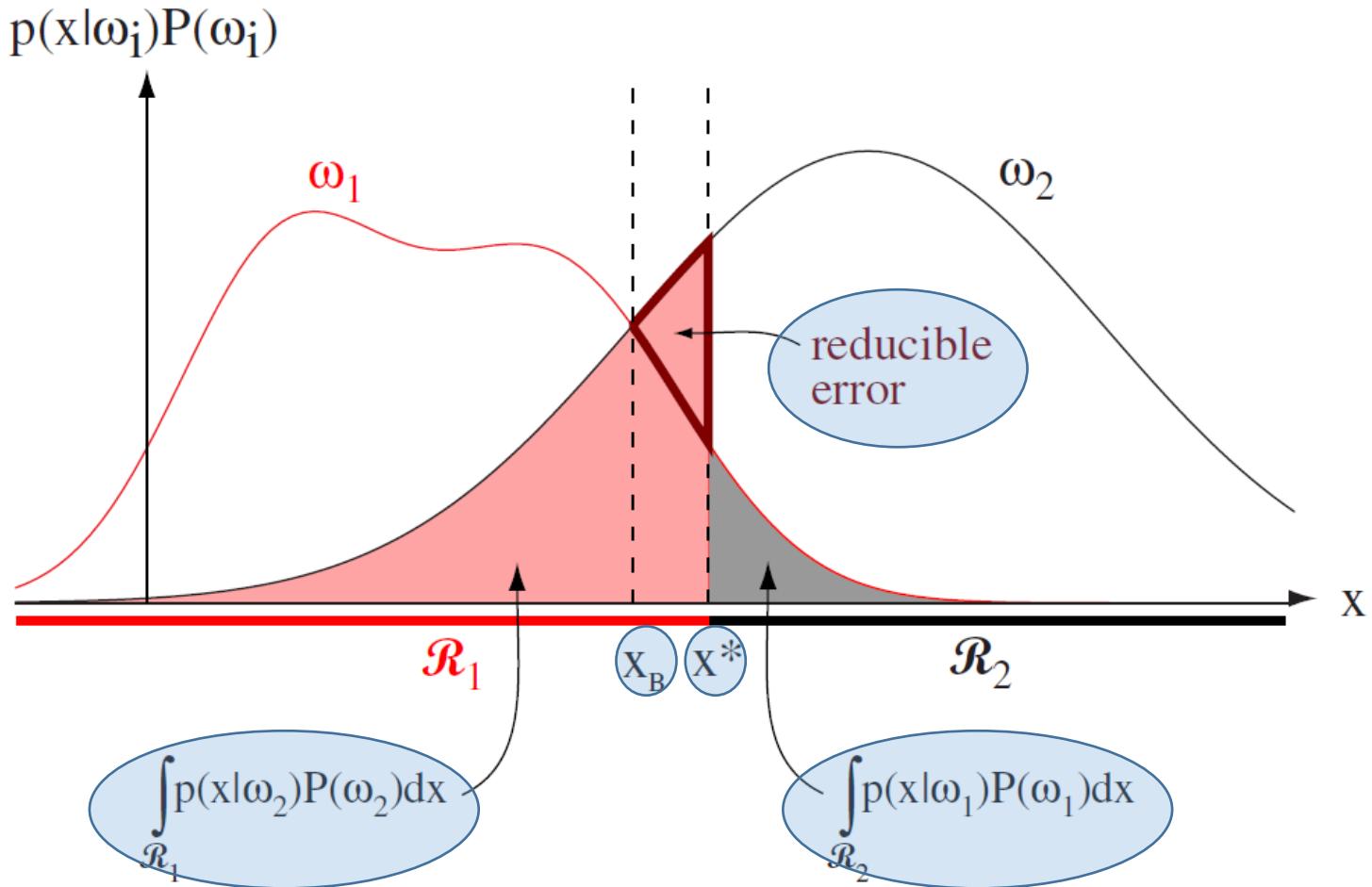
Error Probabilities and Integrals

- For the two-category case

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



Optimal classifier



Components of the probability of error for equal priors and the **non-optimal decision point x^*** .

The optimal point x^B **minimizes** the total shaded area and gives the Bayes error rate.



Error Probabilities and Integrals

- For the **multicategory** case

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



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 (more than two classes)



- $p(\mathbf{x}|\omega_j)$ is the class-conditional probability density function.
- $P(\omega_j)$ is the prior probability that nature is in state ω_j
- The posterior probability can be computed as

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

- Where

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



A More General Theory

- Employ a more general **error function** (i.e., expected “**risk**”) by associating a “**cost**” (based on a “**loss**” function) with different errors
- Let $\{w_1, \dots, w_c\}$ be the finite set of c states of nature (**classes, categories**).
- Let $\{a_1, a_2, \dots, a_a\}$ be the finite set of a **possible actions**.
- Let $I(a_i | w_j)$ be the **loss** incurred for taking action a_i when the state of nature is w_j .
- Let x be the d -component vector-valued random variable called the **feature vector** .

Relation of $I(w_i | w_j)$ and confusion matrix



- **Loss Function $I(w_i | w_j)$** : This is a cost associated with classifying an instance from **true class w_j** as **predicted class w_i** .
- The loss function quantifies how “bad” it is to misclassify.
- It can be defined using the **confusion matrix**. The matrix entries n_{ij} represent counts of true class j predicted as class i . Each misclassification comes with a cost $I(w_i | w_j)$.
- The loss function $I(w_i | w_j)$ maps directly to the off-diagonal entries in the confusion matrix, since these entries represent misclassifications.
- A Bayesian minimum-risk classifier would aim to minimize the misclassifications reflected in the confusion matrix by selecting the class that minimizes the expected loss for each input x



Conditional Risk

- Suppose we observe \mathbf{x} and take **action** a_i .
 - If the **true state** of nature is w_j , we incur the **loss** $l(a_i | w_j)$.
- The **expected loss** with taking action a_i is

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

- which is also called the **conditional risk**



Minimum-Risk Classification

- The general **decision rule** $a(\mathbf{x})$ tells us which action to take for observation \mathbf{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: choose the action a_i for which $R(a_i|\mathbf{x})$ is minimum.

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

- $R(a_i|\mathbf{x})$ is the **conditional risk associated** with action a_i
- The resulting minimum overall risk is called the **Bayes risk** and is the best performance that can be achieved.



Two-Category Classification

- Define
 - a_1 : deciding w_1 ,
 - a_2 : deciding w_2 ,
 - $I_{ij} = I(a_i | w_j)$
- **Conditional risks** can be written as:

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

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

- The fundamental rule is to decide ω_1 if $R(\alpha_1 | \mathbf{x}) < R(\alpha_2 | \mathbf{x})$



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 w_1 if

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)$$

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}.$$

- Comparing **the likelihood ratio** to a threshold that is **independent** of the observation \mathbf{x}

Minimum-Error-Rate Classification



- In classification problems, actions are decisions on classes (a_i is deciding w_i).
 - If action a_i is taken and the true state of nature is w_j , then the **decision is correct** if $i = j$ and in **error** if $i \neq j$.
- If errors are to be avoided, it is natural to seek a **decision rule** that **minimizes the probability of error**, i.e., the ***error rate***



zero-one loss function

- Define the zero-one loss function

$$\lambda(\alpha_i | \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \dots, c.$$

(all errors are **equally** costly).

- **Conditional risk** becomes

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

Minimum-Error-Rate Classification



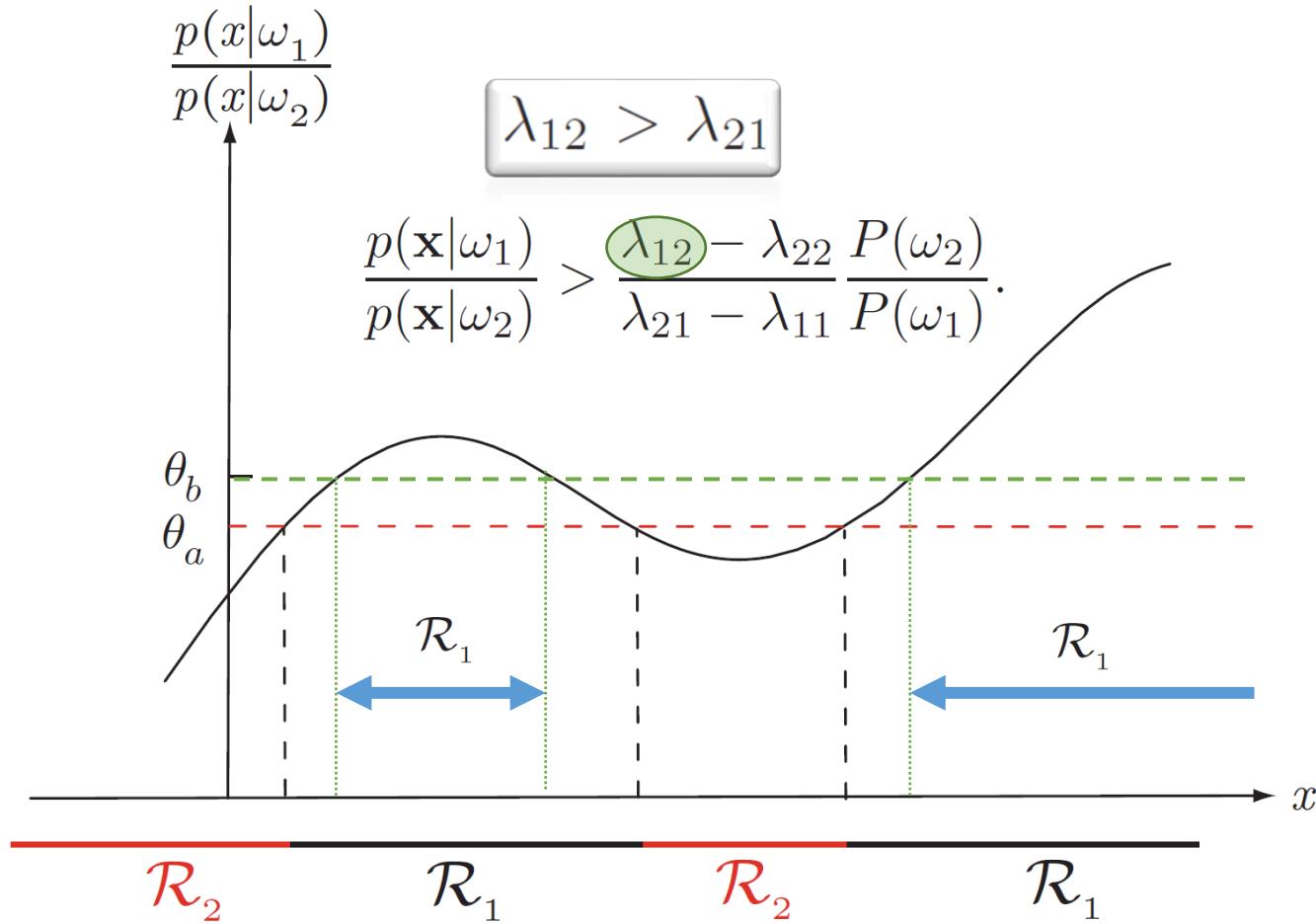
- **Minimizing** the risk requires **maximizing** $P(\omega_i | \mathbf{x})$ and results in the **minimum-error decision** rule

Decide ω_i if $P(\omega_i | \mathbf{x}) > P(\omega_j | \mathbf{x})$ for all $j \neq i$.

- Decide base on **posterior**
- The resulting error is called the **Bayes error** and is the best performance that can be achieved.



Minimum-Error-Rate Classification



The likelihood ratio $p(\mathbf{x}|\omega_1)/p(\mathbf{x}|\omega_2)$. The threshold θ_a is computed using the priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$, and a zero-one loss function. If we [penalize mistakes in classifying ω_2 patterns as ω_1] more than the converse, we should increase the threshold to θ_b .

Neyman-Pearson Criterion



- In some problems, we may wish to minimize the overall risk **subject to a constraint**
- For example, we might wish to minimize the total risk subject to the constraint

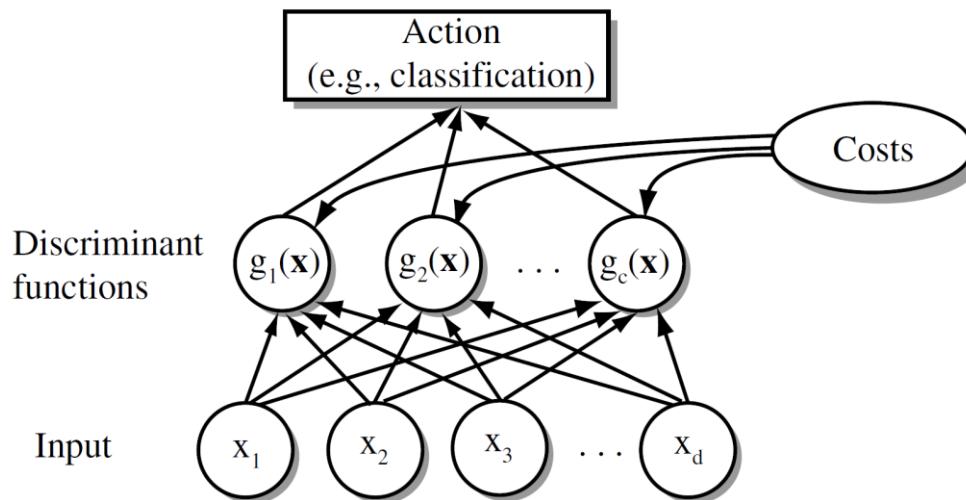
$$\int R(\alpha_i | \mathbf{x}) d\mathbf{x} < \text{constant} \text{ for some particular } i.$$

- For instance, in our fish example, there might be some **government regulation** that we **must not misclassify more than 1% of salmon as sea bass**

Discriminant Functions

- A useful way of **representing classifiers** is through discriminant functions $g_i(\mathbf{x})$, $i = 1, \dots, c$, where the classifier **assigns** a feature vector \mathbf{x} to class w_i if

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





Discriminant Functions

- For the classifier that minimizes **conditional risk**

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

- For the classifier that minimizes error

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



Discriminant Functions

- These functions **divide the feature space into c decision regions** ($\mathcal{R}_1, \dots, \mathcal{R}_c$), separated by **decision boundaries**.
- Note that the results do not change even if we replace every $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$ where $f(\cdot)$ is a **monotonically increasing** function (e.g., logarithm).
- This may lead to significant analytical and computational **simplifications**.



e. g. for minimum-error-rate classification

- Any of the following choices gives identical classification results

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

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

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

Confusion Matrix for Multi-Class Classifiers



- **Confusion Matrix Overview:**

- The confusion matrix is a $C \times C$ table used to evaluate the performance of a classifier.
- **Columns represent the true classes, and rows represent the predicted classes.**

Confusion Matrix (Example for 3 Classes):

	True Class 1	True Class 2	True Class 3
Pred Class 1	n_{11}	n_{12}	n_{13}
Pred Class 2	n_{21}	n_{22}	n_{23}
Pred Class 3	n_{31}	n_{32}	n_{33}

- n_{ij} is the number of times a true class j was predicted as class i .
- Diagonal elements ($n_{11}, n_{22}, n_{33}, \dots$) represent correct predictions.

Metrics Derived from the Confusion Matrix; Key Metrics for Classifier Evaluation



- **Accuracy:**

$$Accuracy = \frac{Correct\ Predictions}{Total\ Instances} = \frac{\sum_{i=1}^C n_{ii}}{\sum_{i=1}^C \sum_{j=1}^C n_{ij}}$$

- Overall, how many instances were correctly classified.

- **Precision** (for Class iii):

$$Precision_i = \frac{n_{ii}}{\sum_{j=1}^C n_{ij}} = \frac{True\ Positive}{Predicted\ Positive\ for\ Class\ i}$$

- How many of the predicted instances for class iii were actually correct.

- **Recall** (for Class iii):

$$Recall_i = \frac{n_{ii}}{\sum_{j=1}^C n_{ji}} = \frac{True\ Positive}{Actual\ Positive\ for\ Class\ i}$$

- How many instances of class iii were correctly predicted.

- **F1-Score** (for Class iii):

$$F1_i = 2 \cdot \frac{Precision_i \cdot Recall_i}{Precision_i + Recall_i}$$



Recall, precision, and f1 score

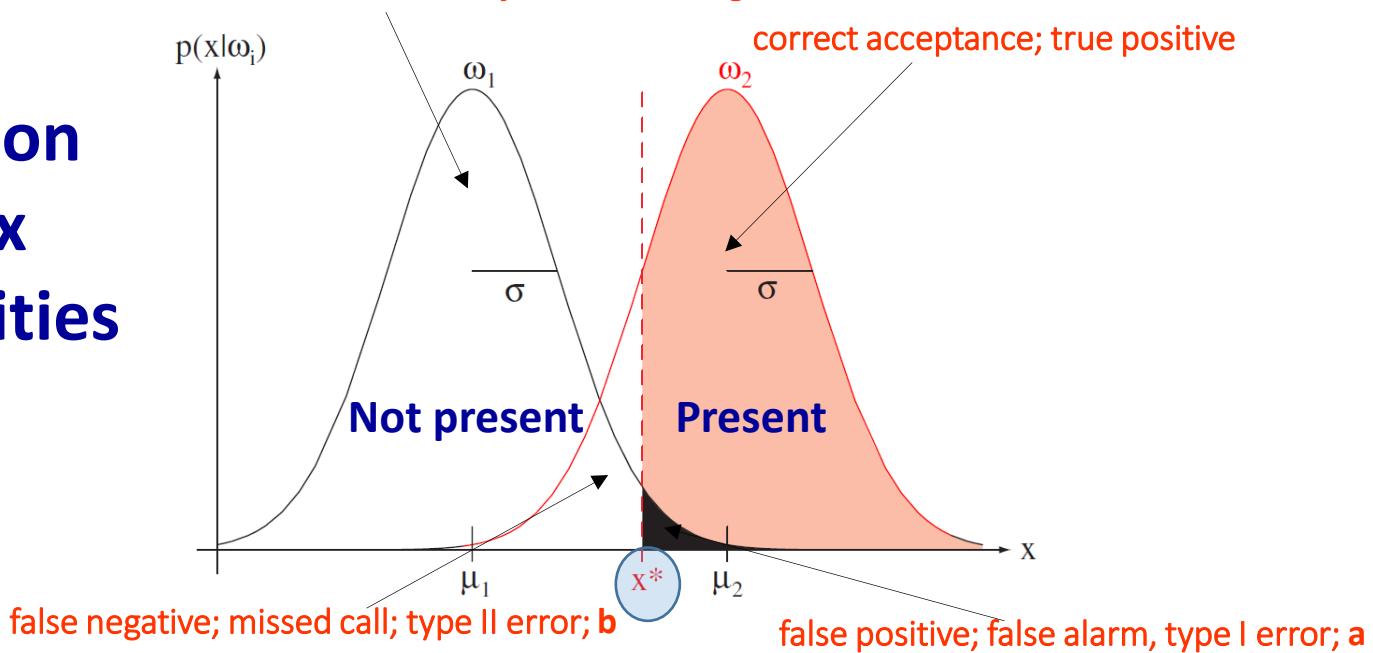
- **Precision for class i** measures how many of the instances predicted as class iii are actually from class i.
- Precision answers the question, "Of all the instances predicted as class i, how many were correct?"
- **Recall for class i** measures how many of the instances that actually belong to class i were correctly predicted as class i.
- The **F1 score** is a robust metric for evaluating classifiers, particularly in scenarios with **imbalanced** datasets or when the **cost of false positives and false negatives varies** (Trade-off Between Precision and Recall).
- The **harmonic mean** used in the F1 score tends to be more sensitive to small values than the arithmetic mean. This means that if either precision or recall is low, the F1 score will also be low, emphasizing the importance of improving the worse-performing metric
- In multi-class classification, there are three primary ways to report the F1 score: **macro-averaged**, **micro-averaged**, and **weighted-averaged** F1 scores. The choice of which to use depends on the balance between classes and the specific focus of the evaluation.



Signal Detection Theory

- There is an **internal signal** (such as a voltage) x , whose value has mean μ_2 when the external signal (pulse) is **present**, and mean μ_1 when it is **not present**.
- Because of noise the actual value is a **random variable**

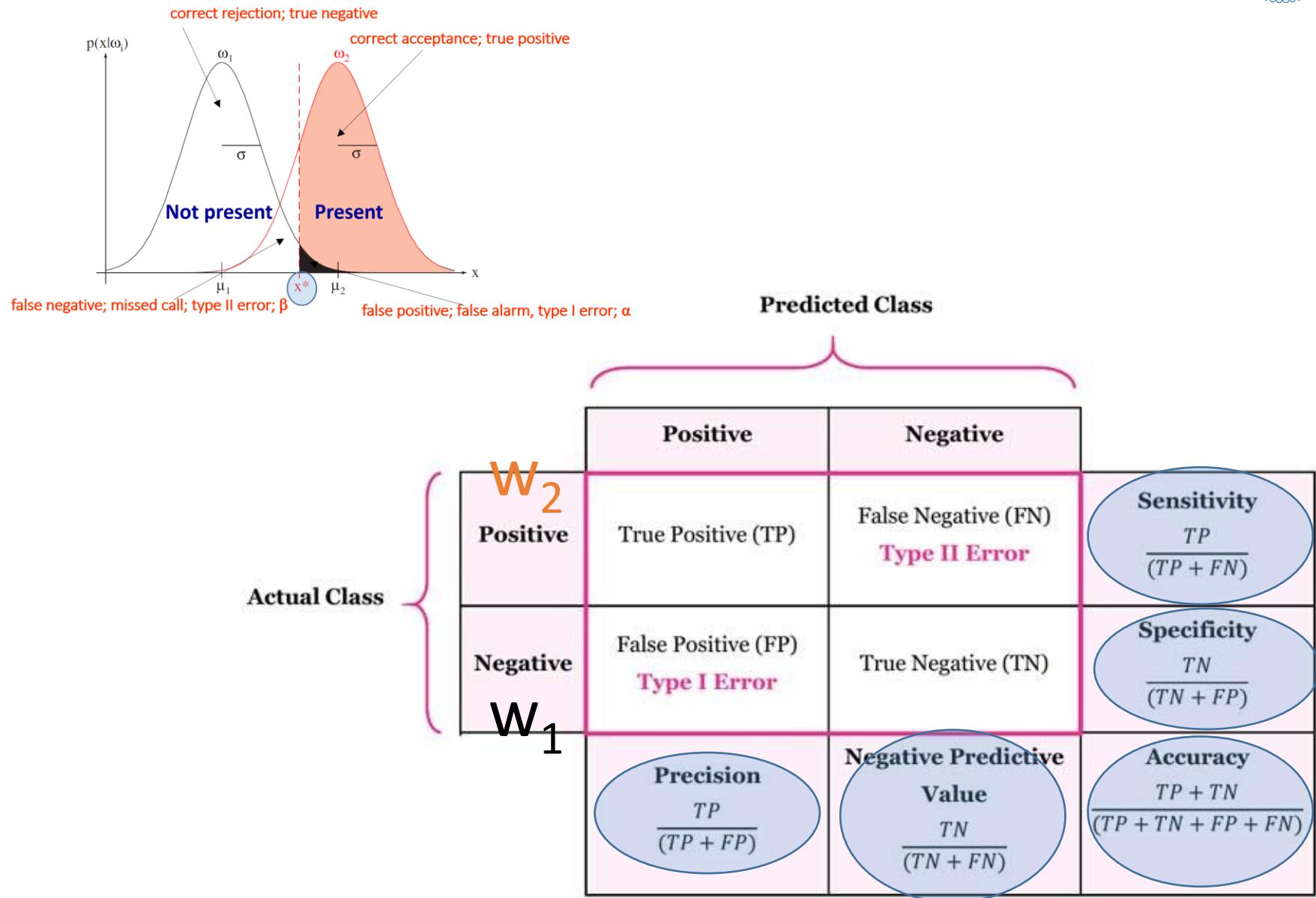
**Confusion
matrix
probabilities**



- The detector (classifier) employs a **threshold** value x^* for determining whether the external pulse is present,



2 classes confusion matrix





Discriminability

- A measure for discriminating whether the pulse is present or not
- Independent of the **choice of x^*** and the **decision strategy**
- Describes the **inherent** and unchangeable properties due to **noise** and the **strength** of ability the external signal

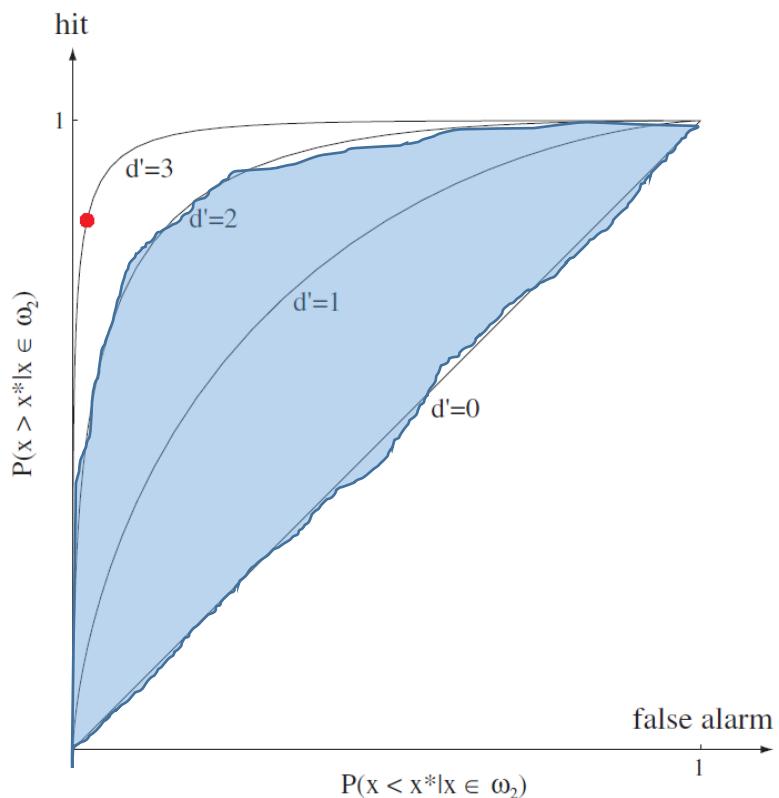
$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}.$$





Operating Characteristics

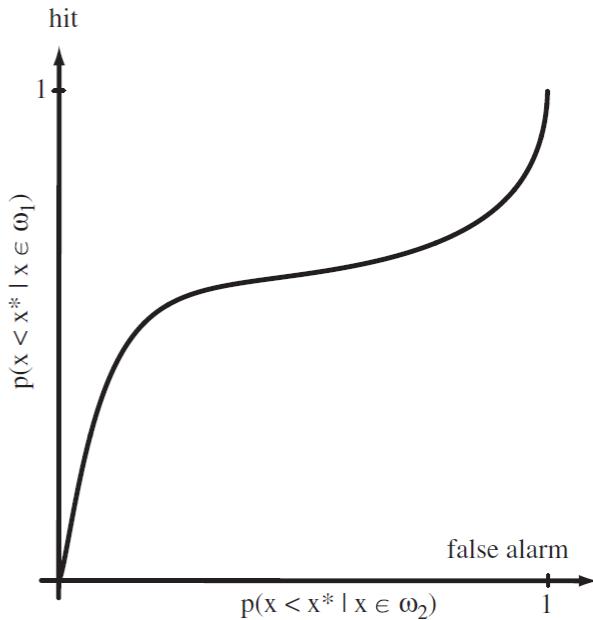
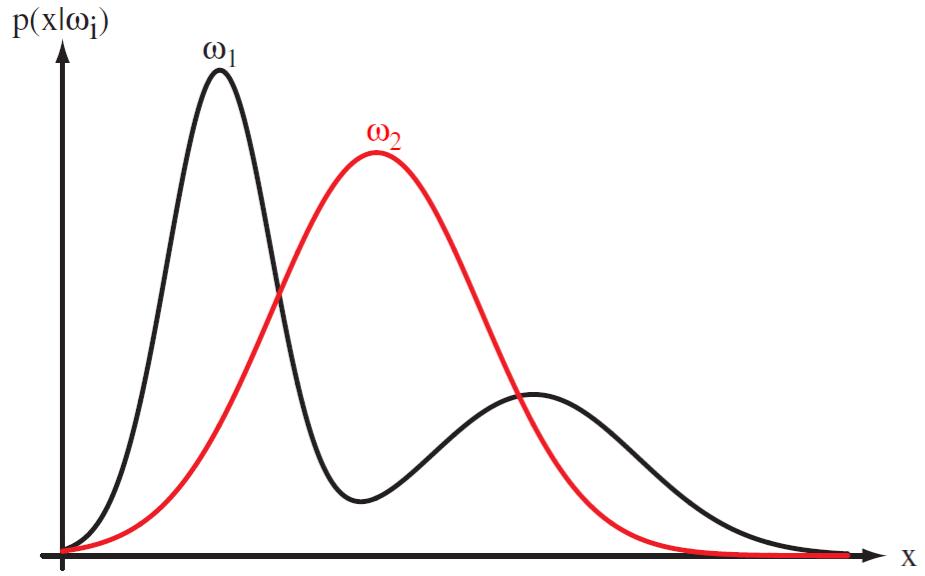
- We do not know μ_1, μ_2, σ nor x^* , but we know the state of nature and the decision of the system. We want to find d'
- Hit and false alarm can be computed by large number of trials. (assume x^* is fixed)
- If the densities are fixed but the threshold x^* is changed, then our hit and false alarm rates will also change.
- Thus we see that for a given discriminability d , our point will move along a smooth curve — a *receiver operating characteristic* or **ROC curve**



The area under a receiver operating characteristic (ROC) curve, abbreviated as AUC, is a single scalar value that measures the overall performance of a binary classifier
Non-parametric measure



Example: general operating characteristic curve



operating characteristic curves are generally **not**
symmetric,



Applications of ROC Curves

- **(ROC)** curve is a graphical representation used to evaluate the performance of a **binary classifier** across various threshold settings
- **Model Comparison:**
 - ROC curves are particularly useful for comparing the performance of multiple binary classifiers. The classifier with the highest AUC is typically considered the best.
- **Threshold Selection:**
 - ROC curves help in choosing an optimal threshold based on the desired trade-off between sensitivity and specificity, especially in contexts where the costs of false positives and false negatives are not equal.
- **Imbalanced Datasets:**
 - ROC curves are effective in evaluating classifiers trained on imbalanced datasets, where traditional metrics (like accuracy) may be misleading.

The Two-Category Case; *dichotomizer*



- It is more common to define a **single discriminant function**

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

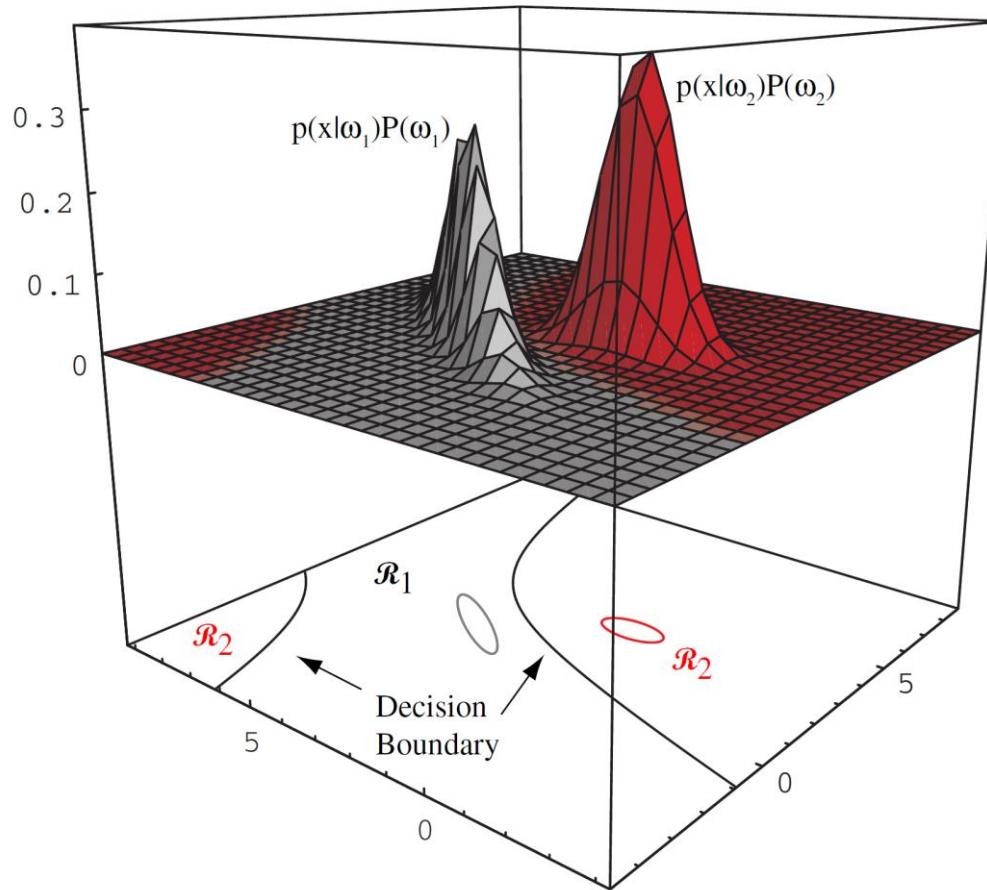
Decide ω_1 if $g(\mathbf{x}) > 0$; otherwise decide ω_2

- the minimum-error-rate discriminant function can be written:

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

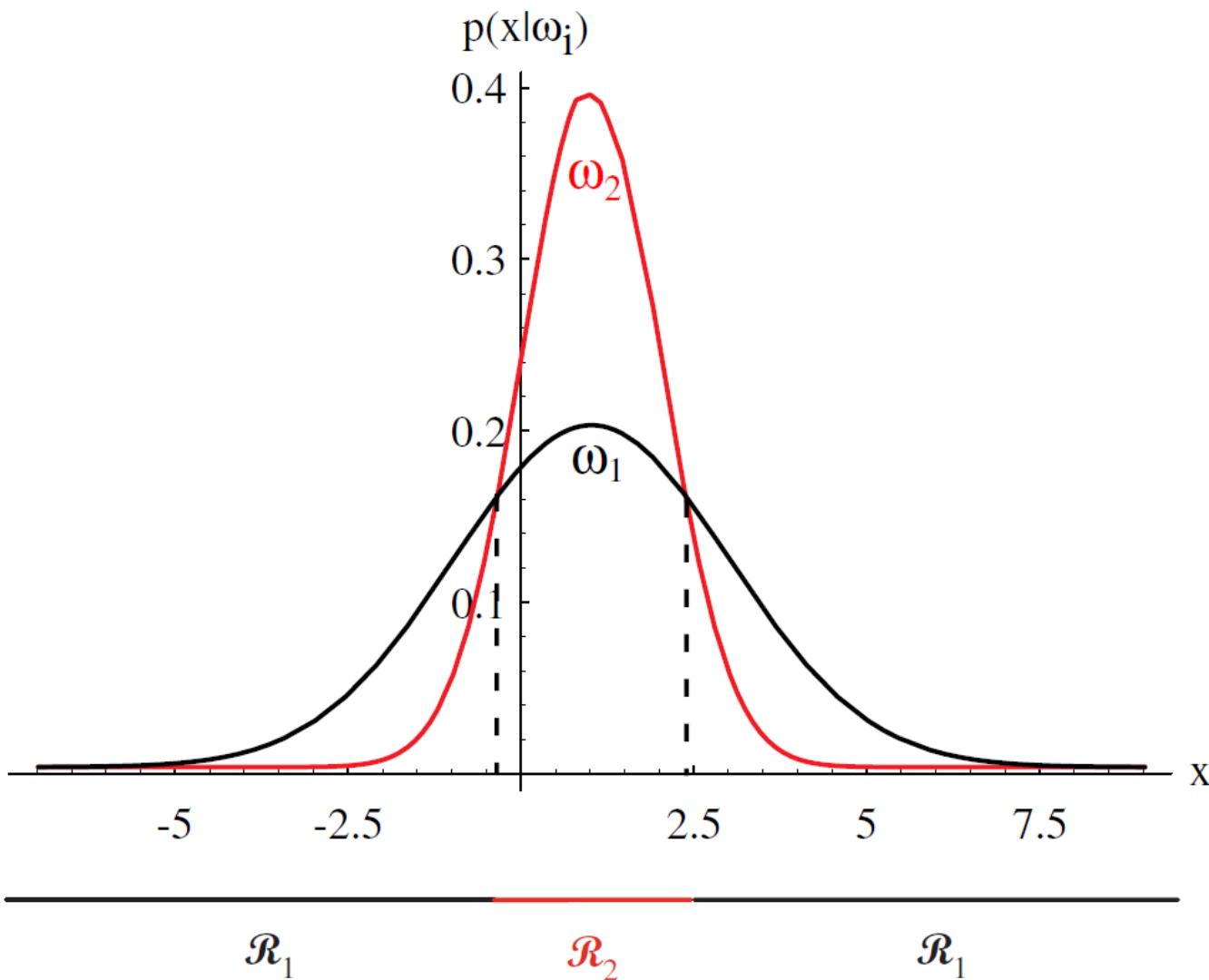
$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}.$$

Two-dimensional two-category classifier



the probability densities are Gaussian (with $1/e$ ellipses shown), the **decision boundary consists of two hyperbolas**, and thus the decision region R_2 is not simply connected.

Non-simply connected decision regions



The Gaussian Density



- Gaussian can be considered as a model where the feature vectors for a given class are **continuous-valued**.
- Some properties 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 transformations** of a Gaussian are also Gaussian.
 - **Uncorrelatedness** implies **independence**.



Univariate Gaussian

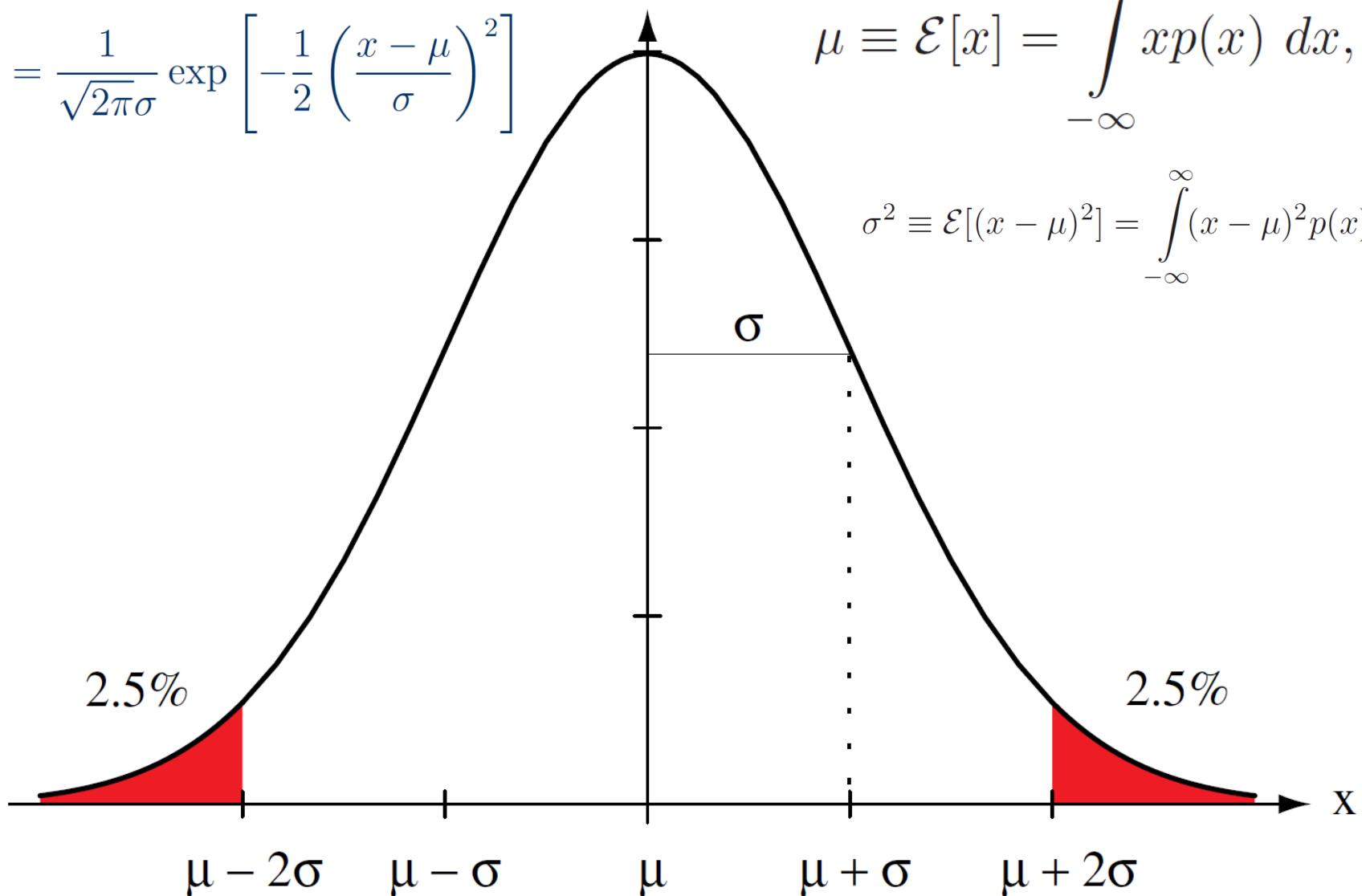
$$p(x) = N(\mu, \sigma^2)$$

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

$$p(x)$$

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x) \, dx,$$

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, dx.$$





entropy

- The entropy is a non-negative quantity that describes the **fundamental uncertainty** in the values of points selected randomly from a distribution

$$H(p(x)) = - \int p(x) \ln p(x) dx,$$

- It measured in ***nats*** (**The natural unit of information**).
- If a \log_2 at is used instead, the unit is the ***bit***.



Multivariate Gaussian

- General multivariate normal density in d dimensions

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

- Where

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

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

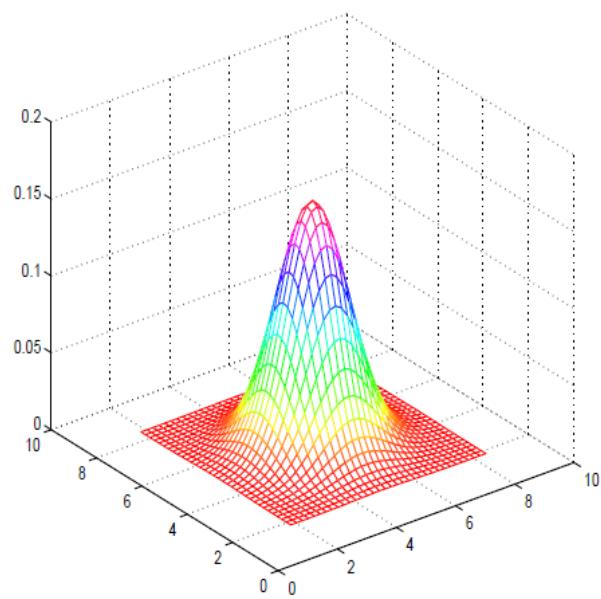
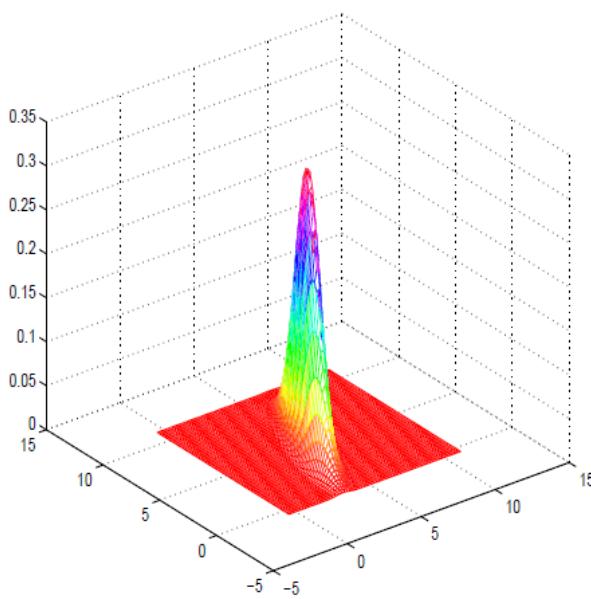
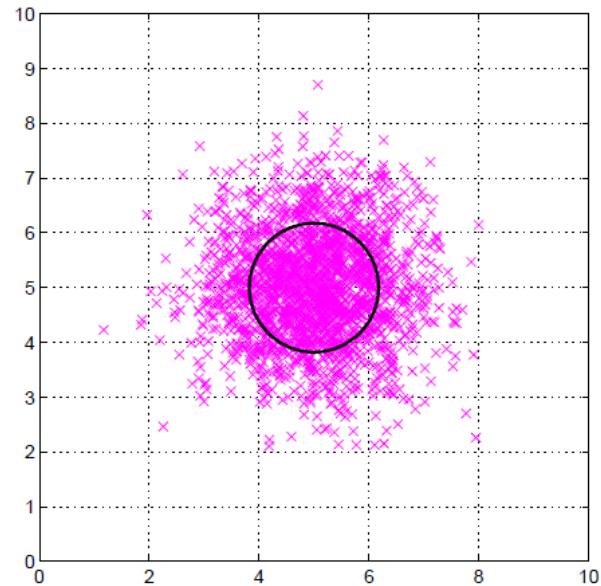
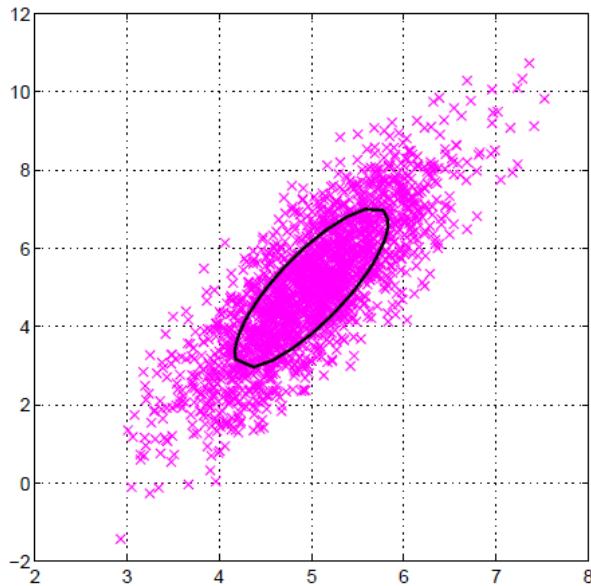
$$\mu_i = \mathcal{E}[x_i] \quad \sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)].$$

- Σ is **positive definite**, so that the determinant of Σ is strictly **positive**

Example

$$\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 0.5 \end{pmatrix}$$

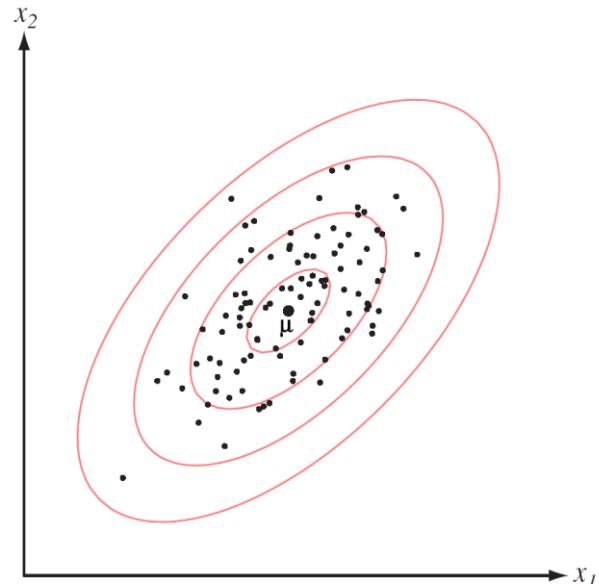
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$





Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ .

- The loci of **points of constant density** are the **ellipses** for which $(\mathbf{x} - \mathbf{m})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{m})$ is constant, where the **eigenvectors** of \mathbf{S} determine the **direction** and the corresponding **eigenvalues** determine the **length** of the principal axes.
- The quantity $r^2 = (\mathbf{x}-\mathbf{m})^T \mathbf{S}^{-1} (\mathbf{x}-\mathbf{m})$ is called the squared **Mahalanobis** distance from \mathbf{x} to μ .





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:

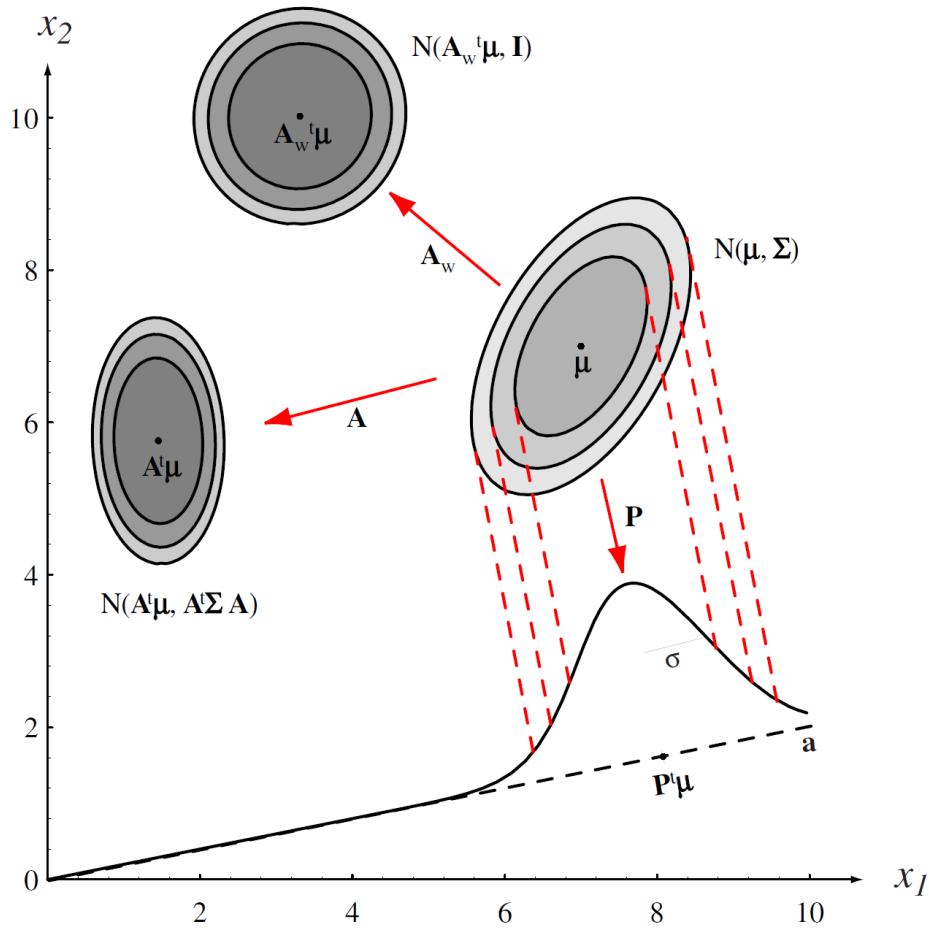
$$\mathbf{A}_w = \Phi \Lambda^{-1/2}$$

- Where
 - \mathbf{F} is the matrix whose **columns** are the **orthonormal eigenvectors** of \mathbf{S}
 - \mathbf{L} is the **diagonal** matrix of the corresponding eigenvalues,
 - Gives a **covariance matrix** equal to the identity matrix \mathbf{I} .



Linear Transformations

- \mathbf{A} is a d -by- k matrix and $\mathbf{y} = \mathbf{A}^t \mathbf{x}$ is a k -component vector,
- If $k = 1$ and \mathbf{A} is a unit-length vector \mathbf{a} , $y = \mathbf{a}^t \mathbf{x}$ is a scalar that represents the **projection of x onto a line in the direction of a** ;
- In general then, **knowledge of the covariance matrix** allows us to calculate the dispersion of the data in **any direction**, or in any **subspace**.





Discriminant Functions for the Gaussian Density

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

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

- For $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

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



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

- Features are **statistically independent**, and each feature has the **same variance**, σ^2
- Samples fall in equal-size **hyperspherical clusters**
- $|\Sigma_i| = \sigma^{2d}$ and $\Sigma_i^{-1} = (1/\sigma^2)\mathbf{I}$
- $|\Sigma_i|$ and the $(d/2) \ln 2\pi$ term are **independent of i**

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

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

where $\|\cdot\|$ is the *Euclidean norm*, that is,

$$\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t (\mathbf{x} - \boldsymbol{\mu}_i).$$



linear discriminant functions

- Expansion of the quadratic form $(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})$ yields

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} [\mathbf{x}^T \mathbf{x} - 2\boldsymbol{\mu}_i^T \mathbf{x} + \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i] + \ln P(\omega_i),$$

- It seems quadratic, but term $\mathbf{x}^T \mathbf{x}$ is the same for all i , making it an **ignorable additive constant**
- Thus, we obtain the **linear discriminant functions**

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

- Where

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

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

We call w_{i0} the **threshold** or **bias** in the i^{th} category



Decision Boundaries

- Decision boundaries are the hyperplanes $g_i(x) = g_j(x)$, and can be written as

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

- Where

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

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

- Hyperplane separating R_i and R_j passes through the point \mathbf{x}_0 and is **orthogonal** to the **vector \mathbf{w}** .



Minimum-distance classifier

- Special case when $P(w_i)$ are the same for c classes the $\ln P(w_i)$ term becomes unimportant additive **constant** that can be ignored and it is **minimum-distance classifier**

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

- The minimum-distance classifier that uses the decision rule

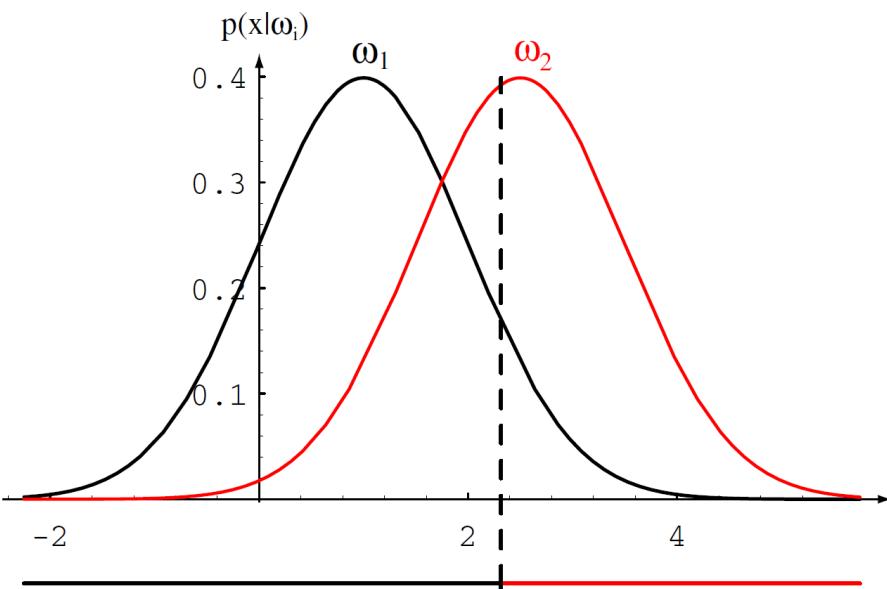
assign \mathbf{x} to w_{i^*} where $i^* = \arg \min_{i=1,\dots,c} \|\mathbf{x} - \boldsymbol{\mu}_i\|$.



Shift of decision boundary by changing priors

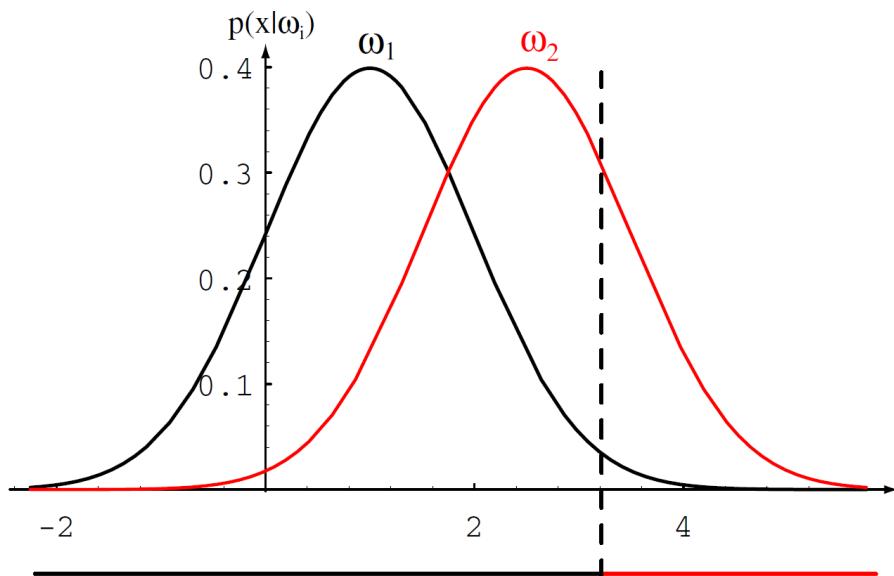
- If $P(\omega_i) \neq P(\omega_j)$, then x_0 shifts away from the most likely category.
- If σ is very small, the position of the boundary is insensitive to $P(\omega_i)$ and $P(\omega_j)$

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



$$\begin{aligned} \mathcal{R}_1 \\ P(\omega_1) = .7 \end{aligned}$$

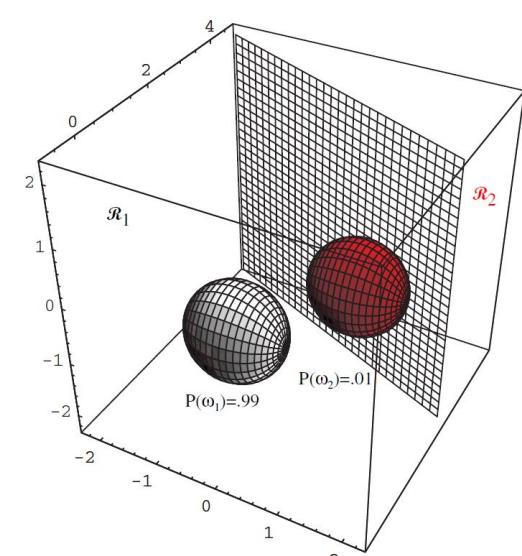
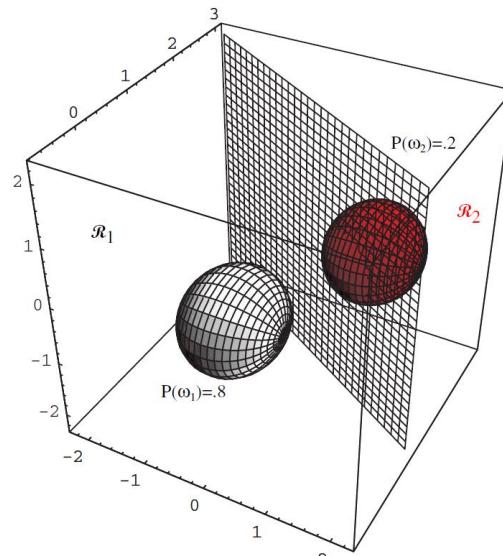
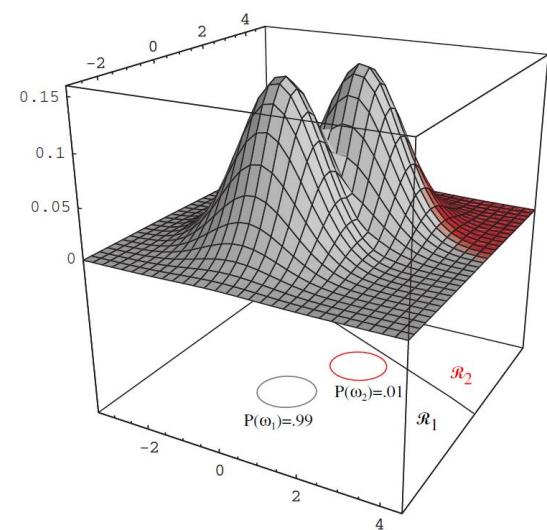
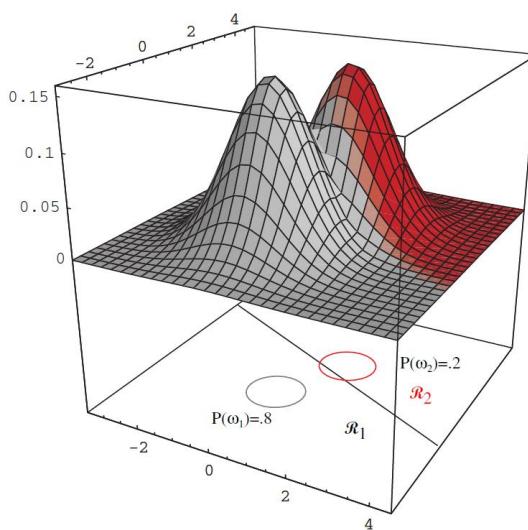
$$\begin{aligned} \mathcal{R}_2 \\ P(\omega_2) = .3 \end{aligned}$$



$$\begin{aligned} \mathcal{R}_1 \\ P(\omega_1) = .9 \end{aligned}$$

$$\begin{aligned} \mathcal{R}_2 \\ P(\omega_2) = .1 \end{aligned}$$

$$\Sigma_i = \sigma^2 \mathbf{I}$$



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 2: $\Sigma_i = \Sigma$

- When the **covariance** matrices **for all of the classes** are **identical** but otherwise arbitrary
- Samples fall in **hyperellipsoidal** clusters of **equal size and shape**
- $|\Sigma_i|$ and the $(d/2) \ln 2\pi$ term are **independent of i**

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + \ln p(w_i)$$

- To classify a feature vector x , measure the squared **Mahalanobis distance from x** to each of the c mean vectors, and assign x to the category of the **nearest mean**



Expanding the Quadratic Term

$$(X - \mu_i)^T \Sigma^{-1} (X - \mu_i) = X^T \Sigma^{-1} X - 2X^T \Sigma^{-1} \mu_i + \mu_i^T \Sigma^{-1} \mu_i$$

This expression contains:

- A **quadratic term**: $X^T \Sigma^{-1} X$
- A **linear term**: $-2X^T \Sigma^{-1} \mu_i$
- A **constant term**: $\mu_i^T \Sigma^{-1} \mu_i$

However, because all classes share the same covariance matrix Σ , the quadratic term $X^T \Sigma^{-1} X$ is the **same** for all classes. It does not affect the decision-making process and can be omitted from the discriminant function.

Thus, the simplified discriminant function becomes:

$$g_i(x) = X^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln p(w_i)$$

This is now a **linear function** of x , meaning it depends on x only through the term $X^T \Sigma^{-1} \mu_i$, which is a dot product between x and some vector (related to μ_i)



Discriminant functions

- In the expansion of **quadratic** form $(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{m}_i)$, the **quadratic** term $\mathbf{x}^t \Sigma^{-1} \mathbf{x}$ is **independent of i** ; so

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0},$$

- Where

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

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i).$$

Decision Boundaries Between Classes: Hyperplanes



$$g_i(x) = g_j(x)$$

This gives:

$$X^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln p(w_i) = X^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + \ln p(w_j)$$

Simplifying the above equation:

$$X^T \Sigma^{-1} (\mu_i - \mu_j) = \frac{1}{2} (\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j) + \ln \frac{p(w_i)}{p(w_j)}$$

This is an equation of a **hyperplane** in the feature space, where:

- $\Sigma^{-1} (\mu_i - \mu_j)$ is the **normal vector** to the hyperplane,
- The right-hand side is the **offset** of the hyperplane.



Decision boundaries

- Since the discriminants are **linear**, the resulting decision boundaries are again **hyperplanes**

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

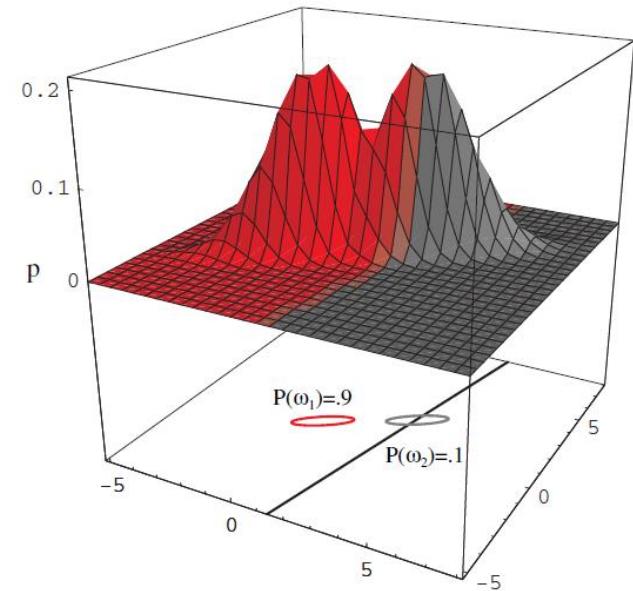
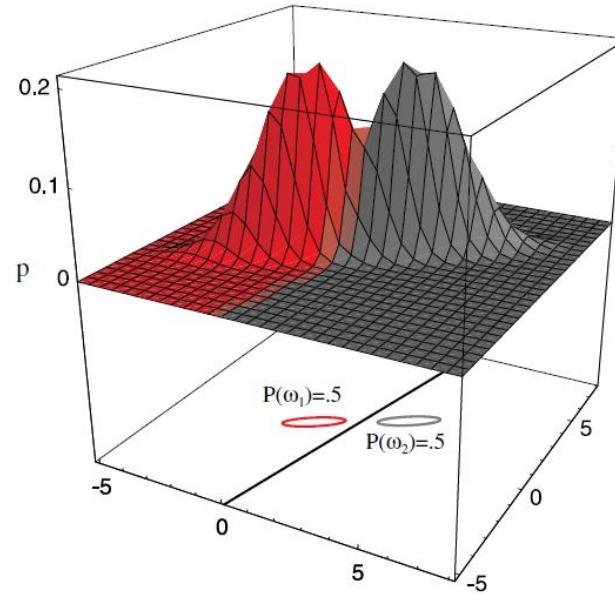
- Where

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

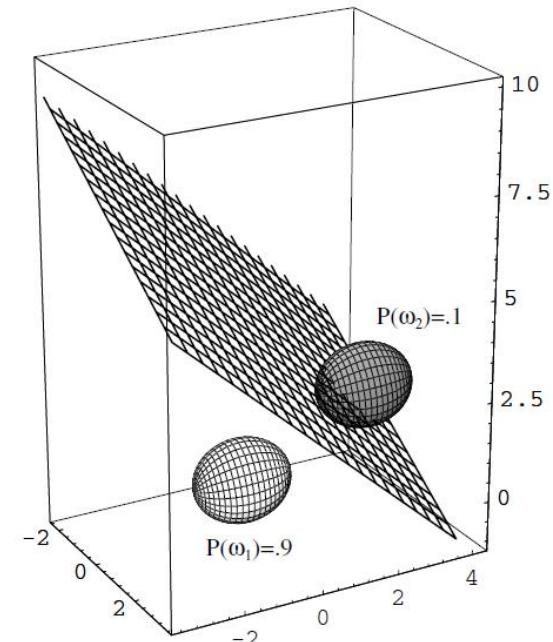
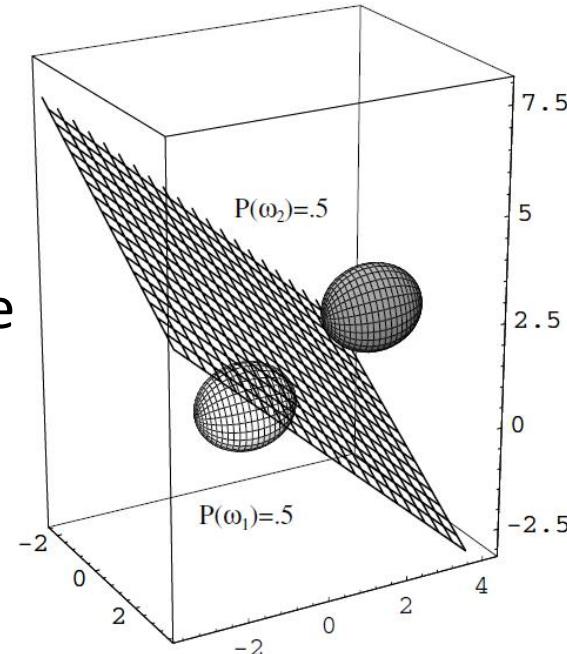
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln [P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

- Hyperplane passes through \mathbf{x}_0 but is **not necessarily orthogonal** to the line between the means. Since $\mathbf{w}=\Sigma^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$ is generally not in the direction of $\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$

Probability densities
with equal but
asymmetric
Gaussian
distributions.



The decision
hyperplanes **are not**
necessarily
perpendicular to the
line connecting the
means.





Case 3: $\Sigma_i = \text{arbitrary}$

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln p(w_i)$$

- Only $(d/2) \ln 2\pi$ term will dropped

- Discriminant functions are:

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0},$$

- Where

$$\mathbf{W}_i = -\frac{1}{2} \Sigma_i^{-1},$$

$$\mathbf{w}_i = \Sigma_i^{-1} \mu_i$$

$$w_{i0} = -\frac{1}{2} \mu_i^t \Sigma_i^{-1} \mu_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i).$$

- Decision boundaries are **hyperquadrics**

Decision Boundaries: Quadratic Surfaces



$$g_i(x) = g_j(x)$$

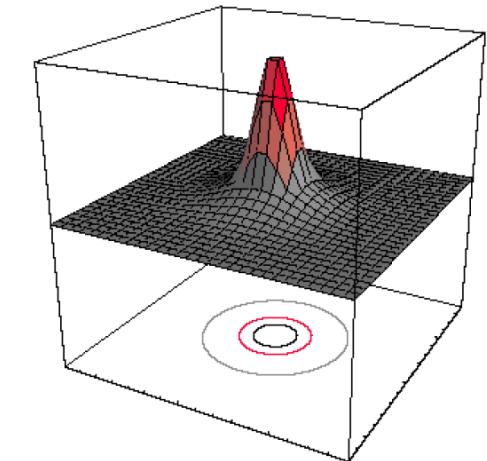
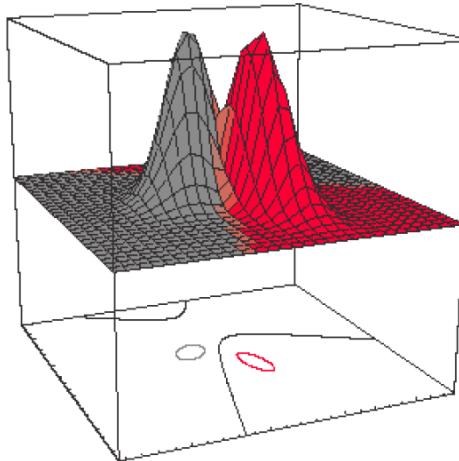
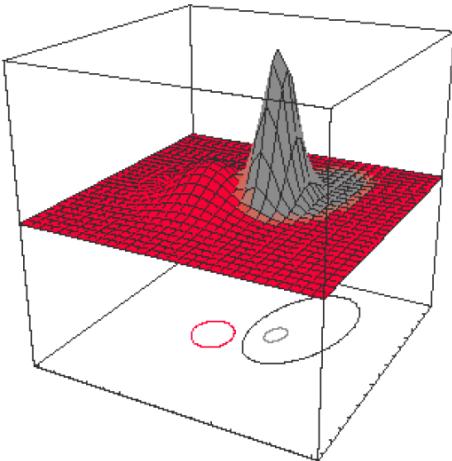
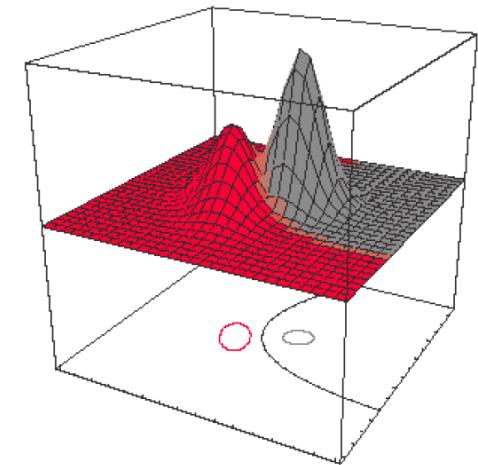
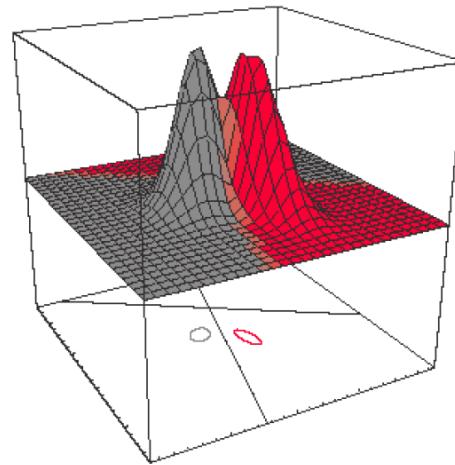
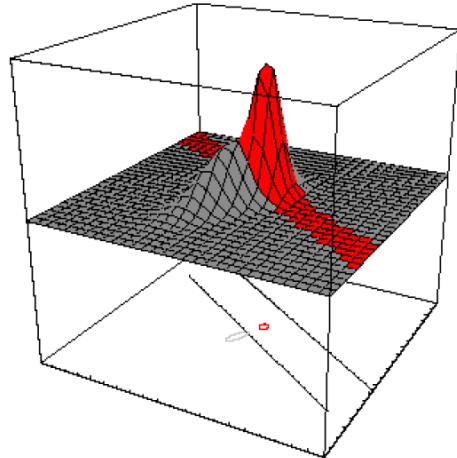
This gives:

$$-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln p(w_i) = -\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) - \frac{1}{2} \ln |\Sigma_j| + \ln p(w_j)$$

Simplifying the above equation:

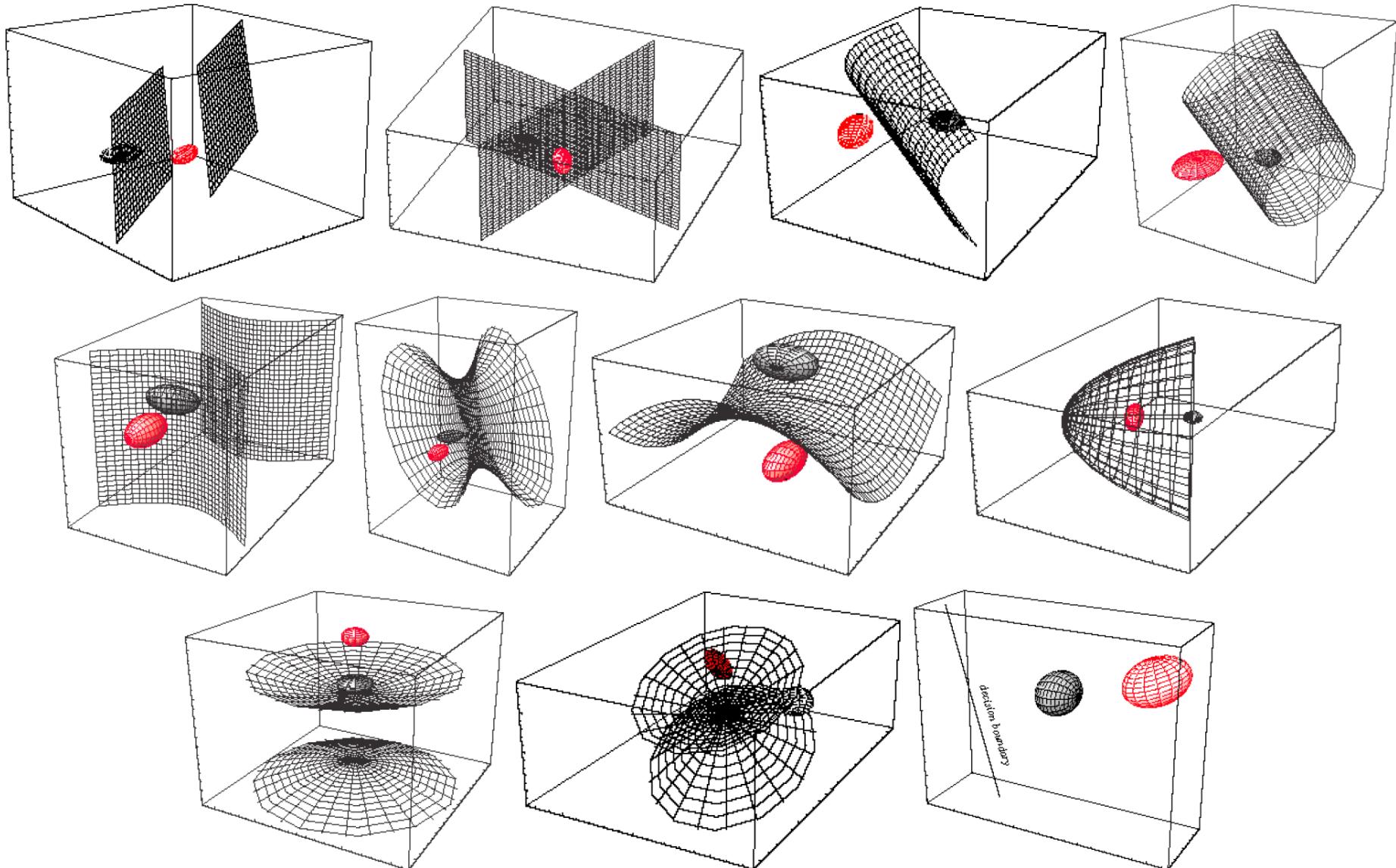
$$\frac{1}{2} \left[(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) - (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right] = \ln \frac{p(w_i)}{p(w_j)} + \frac{1}{2} \ln \frac{|\Sigma_i|}{|\Sigma_j|}$$

This equation represents a **quadratic form** in \mathbf{x} . It is a quadratic equation in the elements of \mathbf{x} , which defines a **quadratic surface** as the decision boundary between the two classes.



Arbitrary Gaussian distributions lead to Bayes decision
boundaries that are **general hyperquadrics**.

Case 3: $\Sigma_i = \text{arbitrary}$

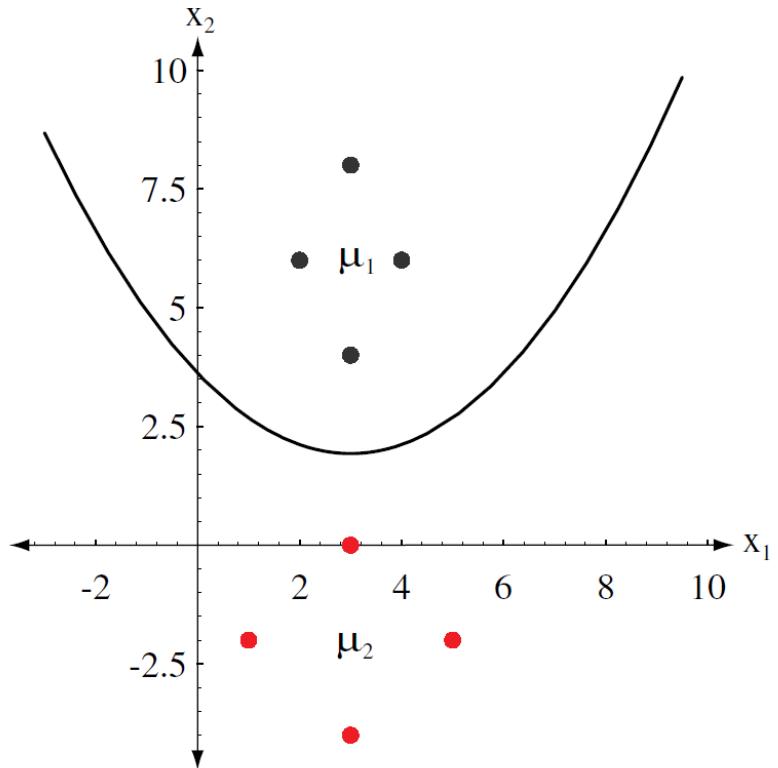




Decision regions for two-dimensional Gaussian data

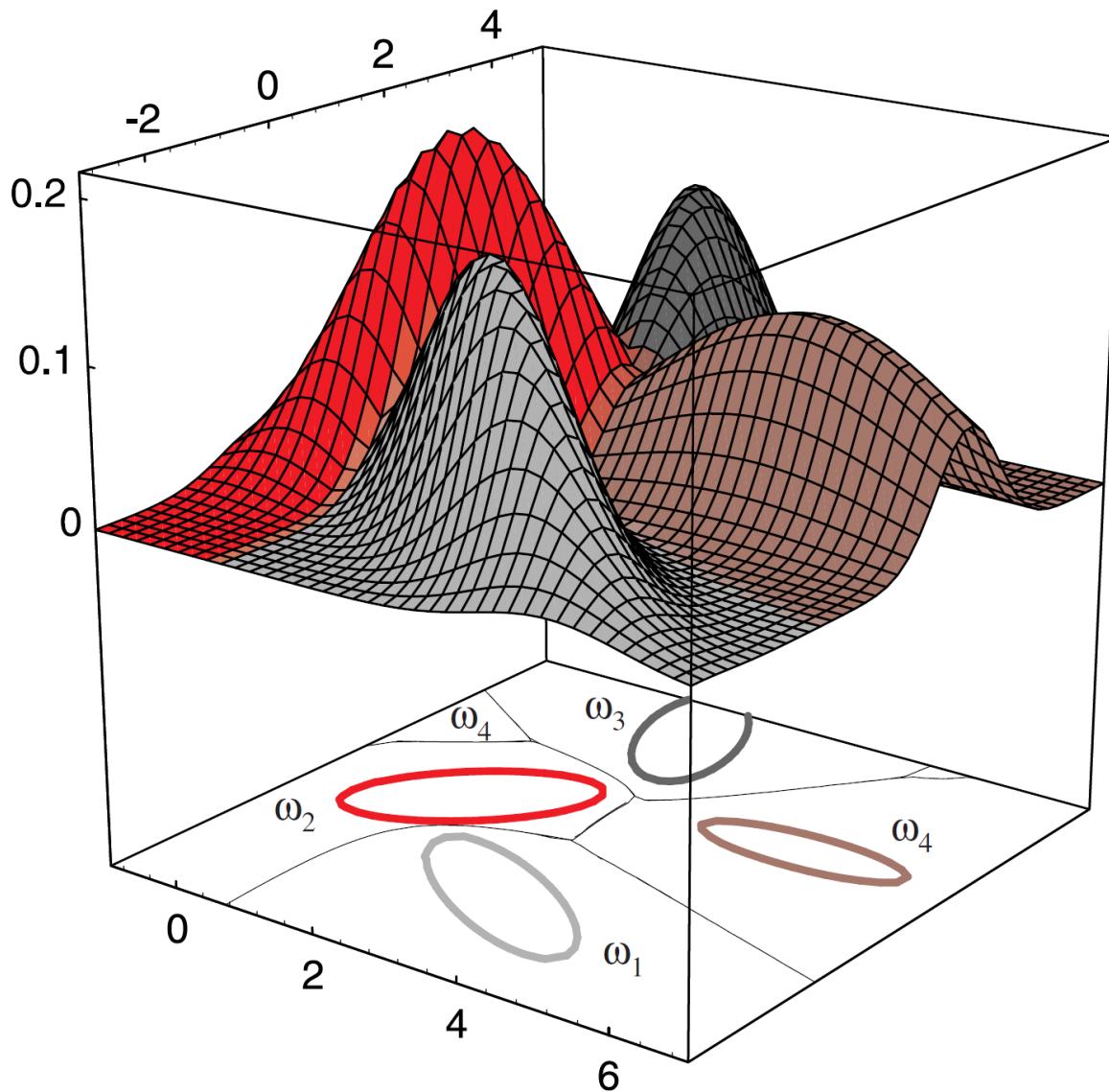
$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5,$$



decision boundary:

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



The **decision regions** for four normal distributions. Even with such a **low number of categories**, the shapes of the boundary regions can be rather complex.