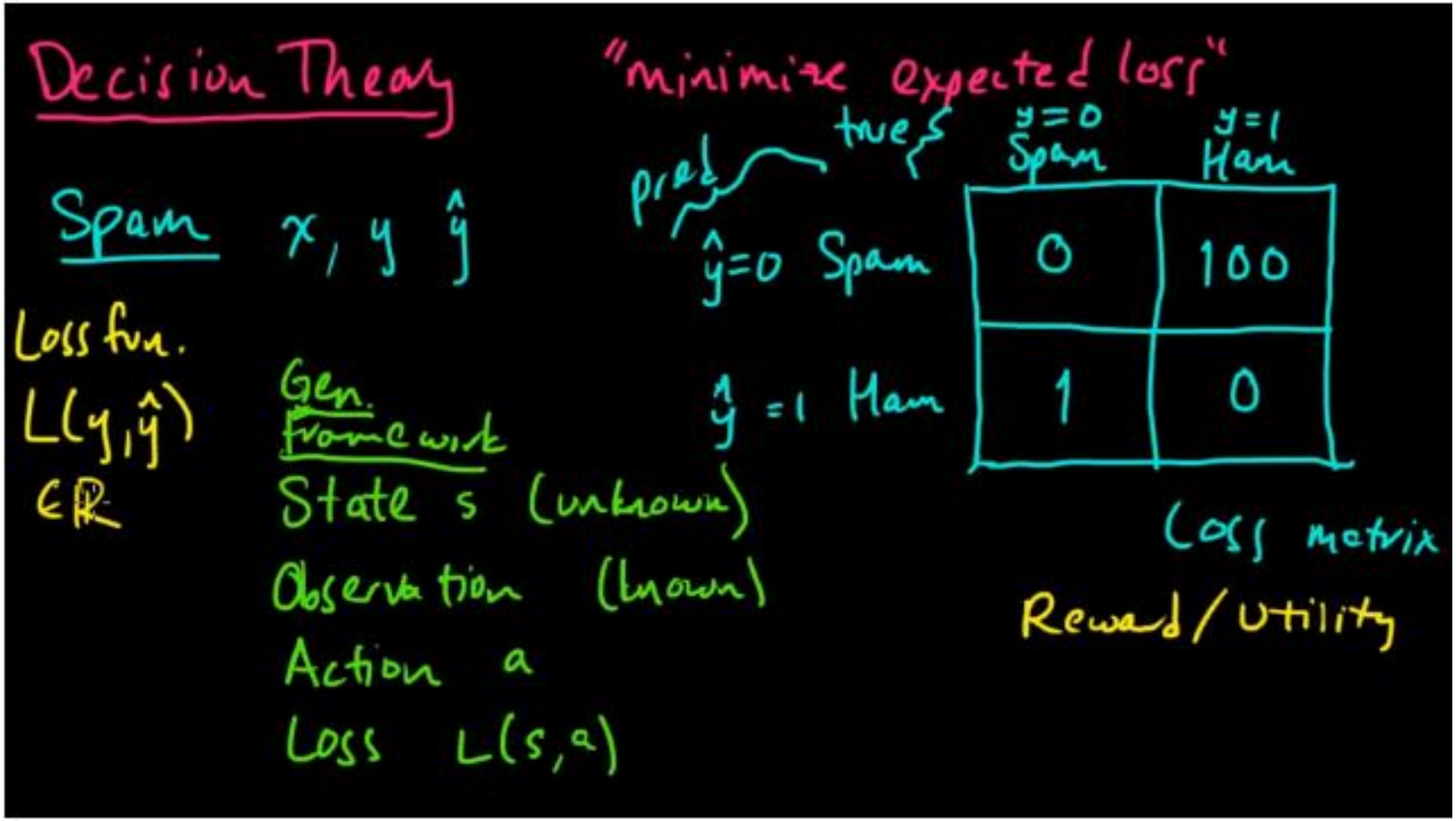
Bayesian Decision Theory

Slides are adapted from Jason Corso, George Bebis and Sargur Srihari

based on the content from Duda, Hart & Stork

Motivation



Decision Theory
"minimize expected loss"
Spam
x, y ŷ
Loss fun.
L(y, ŷ)
ER
Gen. Framework
State s (unknown)
Observation (known)
Action a
Loss L(s, a)
pred
true
y=0 Spam
y=1 Ham
ŷ=0 Spam
ŷ=1 Ham
0
100
1
0
Loss matrix
Reward/Utility

Reverend Thomas Bayes

1702-1761

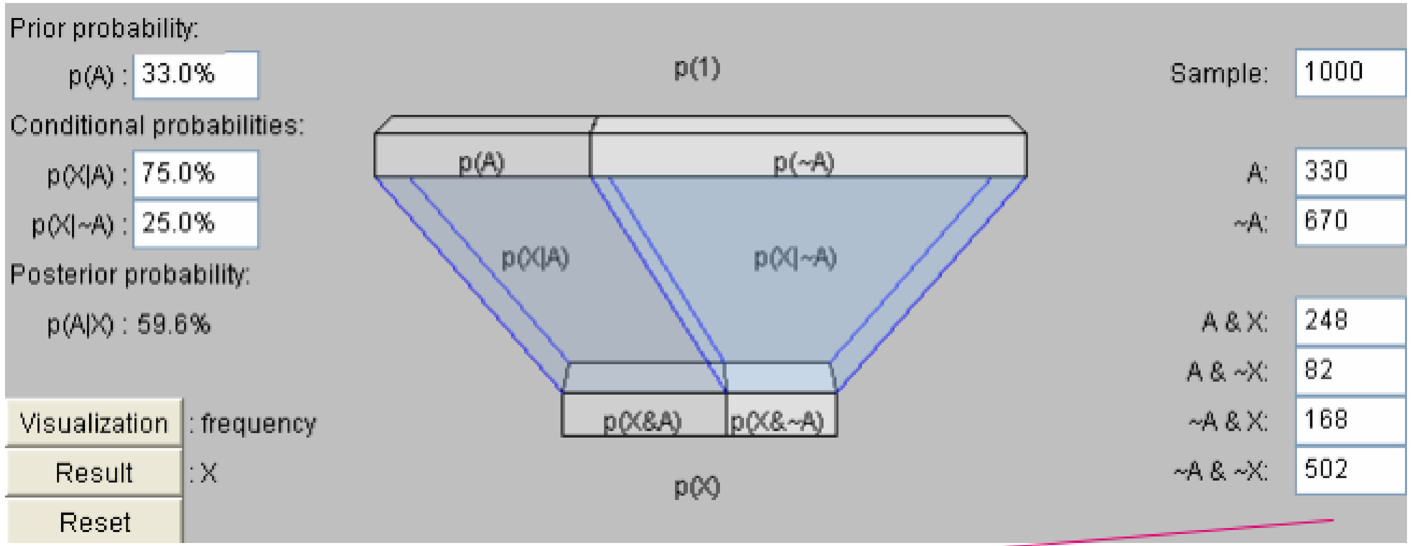
Bayes set out his theory of probability in *Essay towards solving a problem in the doctrine of chances* published in the *Philosophical Transactions of the Royal Society of London* in 1764. The paper was sent to the Royal Society by Richard Price, a friend of Bayes', who wrote:-

I now send you an essay which I have found among the papers of our deceased friend Mr Bayes, and which, in my opinion, has great merit... In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.

Bayes Rule

Two Classes(A, ~A) , Single Binary-Valued Feature (X,~X)

Known



Data

By Conditional Probability Rule,

$$p(X / A) = \frac{p(X \,\&\, A)}{p(A)}$$

$$= \frac{.248}{.330} = 0.7515$$

$$p(X / \sim A) = \frac{p(X \,\&\, \sim A)}{p(\sim A)}$$

$$= \frac{.168}{.670} = 0.2507$$

By Bayes Rule, $P(A / X) = \dfrac{P(X / A)P(A)}{P(X)}$

$$= \frac{P(X / A)P(A)}{P(X \,\&\, A) + P(X \,\&\, \sim A)}$$

$$= \frac{P(X / A)P(A)}{P(X / A)P(A) + P(X / \sim A)P(\sim A)}$$

$$= \frac{0.75 \times 0.33}{0.75 \times 0.33 + 0.25 \times 0.67}$$

$$= \frac{.2475}{.2475 + .1675} = \frac{.2475}{.415} = 0.596$$

Bayesian Decision Theory

- Fundamental statistical approach to statistical pattern classification
- Quantifies trade-offs between classification using probabilities and costs of decisions
- Assumes all relevant probabilities are known

Bayesian Decision Theory

It is the decision making when all underlying probability distributions are known.
It is optimal given the distributions are known.

For two classes ω_1 and ω_2 ,

Prior probabilities for an unknown new observation:

$P(\omega_1)$: the new observation belongs to class 1
$P(\omega_2)$: the new observation belongs to class 2
$P(\omega_1) + P(\omega_2) = 1$

It reflects our prior knowledge. It is our decision rule when no feature on the new object is available:
Classify as class 1 if $P(\omega_1) > P(\omega_2)$

Bayesian Decision Theory

- Design classifiers to make decisions subject to minimizing an expected ”risk”.
  - The simplest risk is the classification error (i.e., assuming that misclassification costs are equal).
  - When misclassification costs are not equal, the risk can include the cost associated with different misclassifications.

Example

- Recall our example from the first lecture on classifying two fish as salmon or sea bass.
- And recall our agreement that any given fish is either a salmon or a sea bass; DHS call this the **state of nature** of the fish.
- Let's define a (probabilistic) variable ω that describes the state of nature.

$$\omega = \omega_1 \quad \text{for sea bass} \qquad (1)$$
$$\omega = \omega_2 \quad \text{for salmon} \qquad (2)$$

- Let's assume this two class case.

Salmon

Sea Bass

Terminology - consider the sea bass/salmon example

- 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)

Prior Probability

- The *a priori* or **prior** probability reflects our knowledge of how likely we expect a certain state of nature before we can actually observe said state of nature.
- In the fish example, it is the probability that we will see either a salmon or a sea bass next on the conveyor belt.
- Note: The prior may vary depending on the situation.
  - If we get equal numbers of salmon and sea bass in a catch, then the priors are equal, or **uniform**.
  - Depending on the season, we may get more salmon than sea bass, for example.
- We write $P(\omega = \omega_1)$ or just $P(\omega_1)$ for the prior the next is a sea bass.
- The priors must exhibit exclusivity and exhaustivity. For c states of nature, or classes:

$$1 = \sum_{i=1}^{c} P(\omega_i) \tag{3}$$

Prior Probability

- The catch of salmon and sea bass is equiprobable
  - $P(\omega_1) = P(\omega_2)$ (uniform priors)
  - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

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?
- $\boxed{\text{Decide } \omega_1 \text{ if } P(\omega_1) > P(\omega_2)\text{; otherwise decide } \omega_2.}$
- What can we say about this decision rule?
  - Seems reasonable, but it will **always** choose the same fish. Favours the most likely class.
  - 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.)

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

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

Features and Feature spaces

- A **feature** is an observable variable.
- A **feature space** is a set from which we can sample or observe values.
- Examples of features:
  - Length
  - Width
  - Lightness
  - Location of Dorsal Fin
- For simplicity, let's assume that our features are all continuous values.
- Denote a scalar feature as x and a vector feature as $\mathbf{x}$. For a d-dimensional feature space, $\mathbf{x} \in \mathbb{R}^d$.

Class Conditional probability density $p(x/\omega_j)$ (*likelihood*) :

- The class-conditional probability density function is the probability density function for x, our feature, given that the state of nature is ω
- e.g., how frequently we will measure a pattern with feature value *x* given that the pattern belongs to class ω_j
- $P(x \mid \omega_1)$ and $P(x \mid \omega_2)$ describe

the difference in lightness

between populations of

sea-bass and salmon

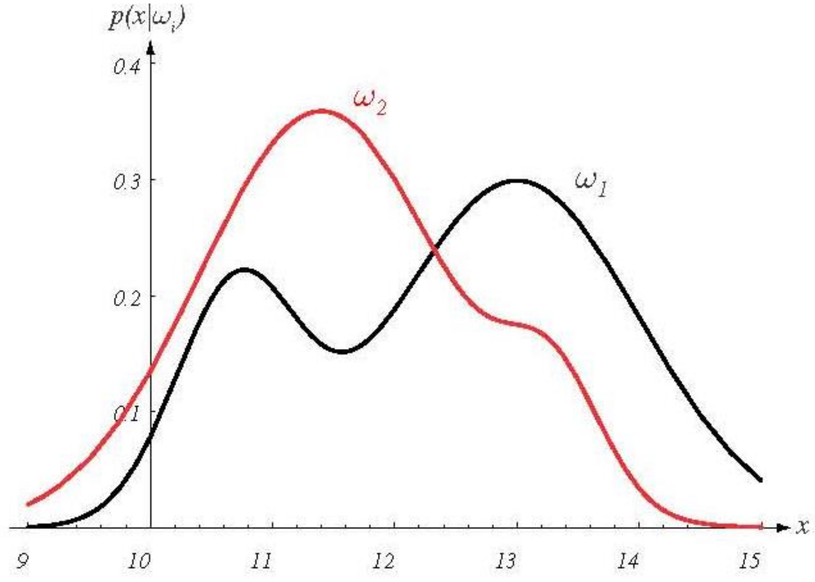


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

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:

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

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

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

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

Decision Rule Using Conditional Probabilities

- Using Bayes' rule:

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

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

likelihood ratio threshold

Decision Rule Using Conditional Probabilities (cont'd)

$p(x/\omega_j)$ $\qquad P(\omega_1) = \frac{2}{3} \qquad P(\omega_2) = \frac{1}{3}$ $\qquad P(\omega_j/x)$

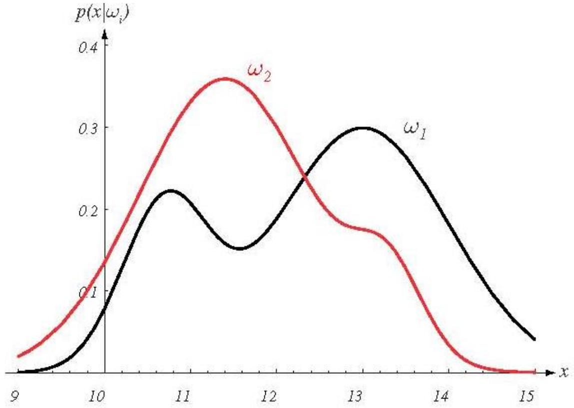


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

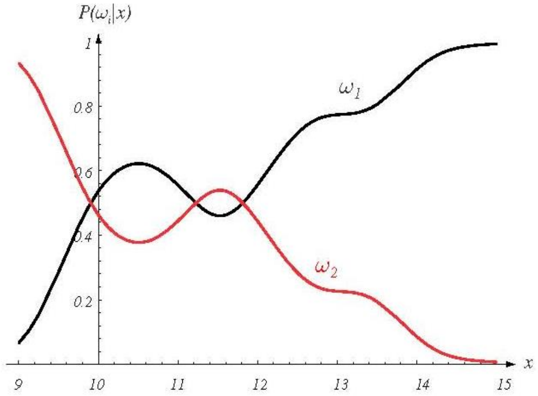


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

Probability of error

- x is an observation for which:
- if P(ω1 | x) > P(ω2 | x) True state of nature = ω1
- if P(ω1 | x) < P(ω2 | x) True state of nature = ω2

- For a given observation x, we would be inclined to let the posterior govern our decision:

$$\omega^* = \arg\max_i P(\omega_i|\mathbf{x}) \quad (\text{}$$

- What is our **probability of error**?
- For the two class situation, we have

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

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.

- Probabilities can only come from experiments.

(2) **Bayesian** (subjective) approach.

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

Example (objective approach)

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

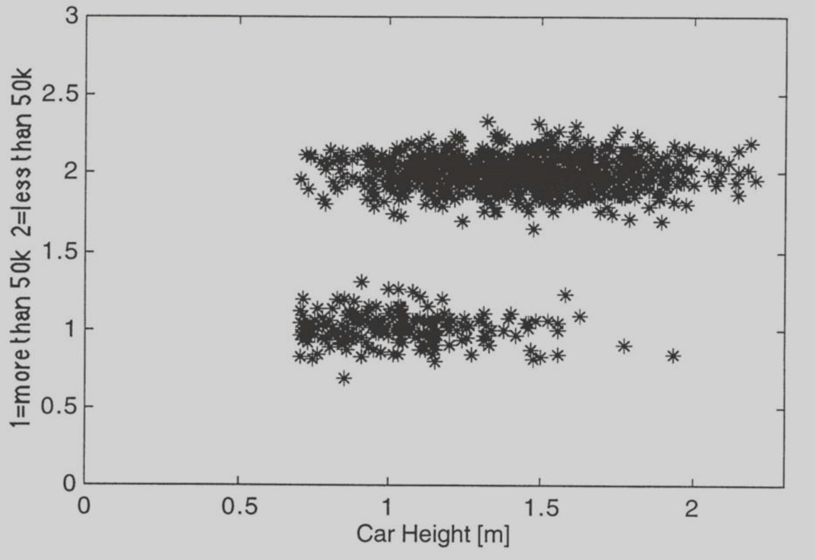
- Use the Bayes' rule to compute the posterior probabilities:

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

- 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: $\#C_1=221$ $\#C_2=988$



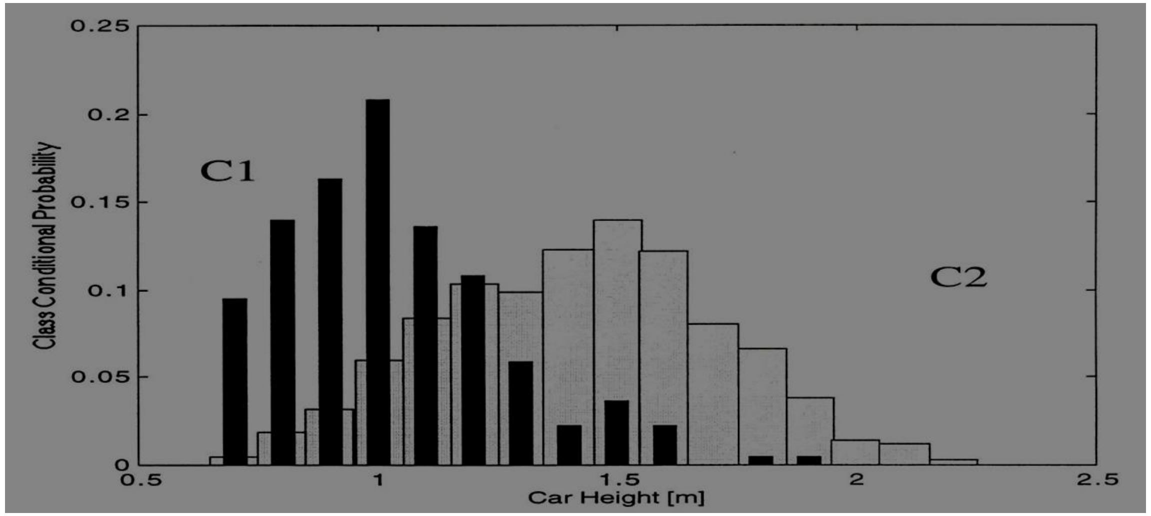
$$P(C_1) = \frac{221}{1209} = 0.183$$

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

Example (cont'd)

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

$p(x/C_i)$



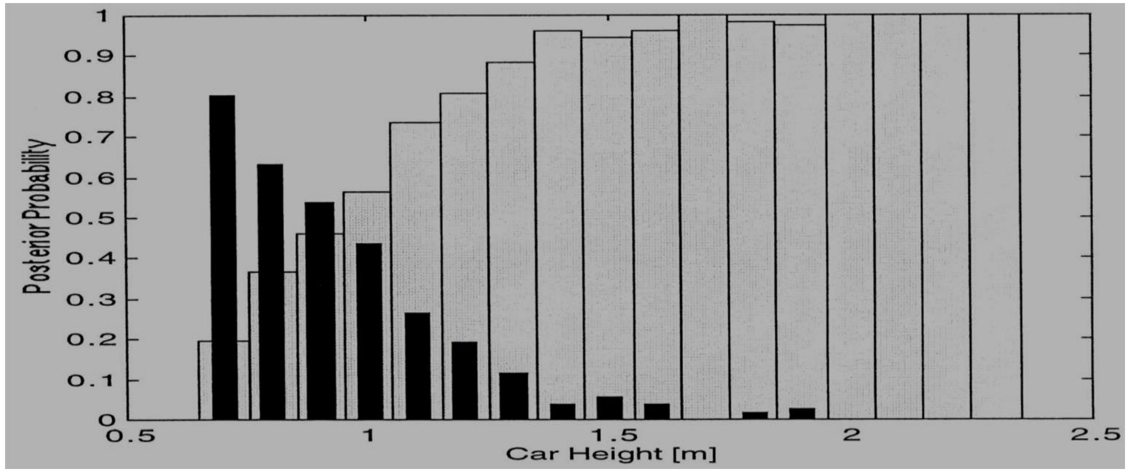
Example (cont'd)

- Calculate the posterior probability for each bin:

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

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

$P(C_i/x)$



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) - Refusing to make a decision in close or bad cases!.
- Employ a more general error function (i.e., expected "risk") by associating a "cost" (based on a "loss" function) with different errors.
- Note that, the loss function states how costly each action taken is

Loss Function

- A **loss function** states exactly how costly each action is.
- As earlier, we have c classes $\{\omega_1, \dots, \omega_c\}$.
- We also have a possible actions $\{\alpha_1, \dots, \alpha_a\}$.
- The loss function $\lambda(\alpha_i|\omega_j)$ is the loss incurred for taking action α_i when the class is ω_j.
- The **Zero-One Loss Function** is a particularly common one:

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, 2, \dots, c \qquad ($$

It assigns no loss to a correct decision and uniform unit loss to an incorrect decision.

If action α_i is taken and the true state of nature is ω_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* which is the *error rate*

Expected loss

- We can consider the loss that would be incurred from taking each possible action in our set.
- The **expected loss** or conditional risk is by definition

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

- The **zero-one conditional risk** is

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

- Hence, for an observation x, we can minimize the expected loss by selecting the action that minimizes the conditional risk.
- (Teaser) You guessed it: this is what Bayes Decision Rule does!

Overall Risk

R = Sum of all $\underbrace{R(\alpha_i \mid x)}_{\text{Conditional risk}}$ *for i = 1,…,a*

Minimizing R $\Longleftrightarrow$ Minimizing $R(\alpha_i \mid x)$ *for* $i = 1, \ldots, a$

Expected Loss with action i

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

Select the action α_i for which $R(\alpha_i \mid x)$ is minimum

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

Bayes risk = best performance that can be achieved

Overall Risk

- Suppose *$\alpha(\mathbf{x})$* is a general decision rule that determines which action $\alpha_{1,}$ $\alpha_{2, \dots,}$ α_l to take for every **x**.

- The overall risk is defined as:

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

Clearly, we want the rule α(·) that minimizes R(α(x)|x) for all x.

- 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 $R(\alpha_i / \mathbf{x})$
- The resulting minimum R^* is called *Bayes risk* and is the best (i.e., optimum) performance that can be achieved:

Example: Two-category classification

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

loss incurred for deciding ω_i when the true state of nature is ω_j

- The conditional risks are:

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

$$\Downarrow$$

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

Example: Two-category classification

- 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

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

or (i.e., using likelihood ratio)

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

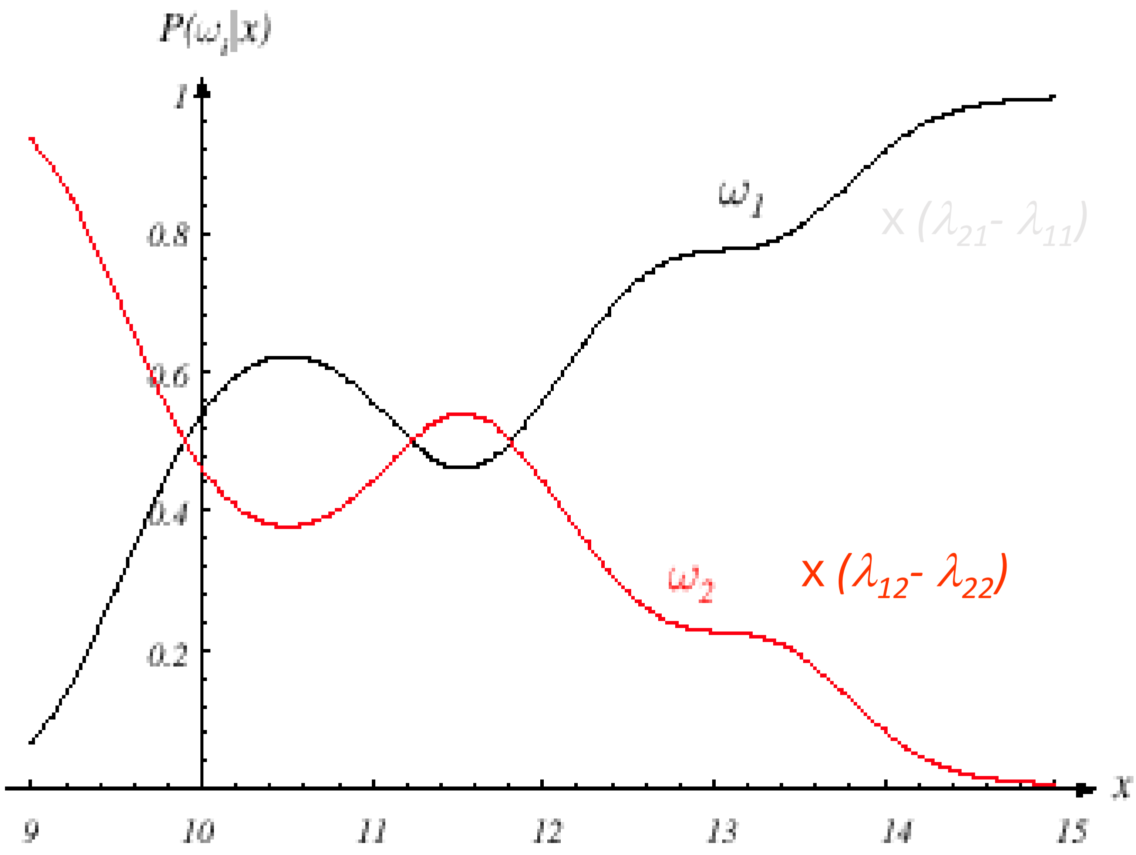


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

Two-Category Decision Theory: Chopping Machine

α_1 = *chop*
α_2 = *DO NOT chop*
ω_1 = *NO hand in machine*
ω_2 = *hand in machine*

$\lambda_{11} = \lambda(\alpha_1 \mid \omega_1) = \$ \quad 0.00$
$\lambda_{12} = \lambda(\alpha_1 \mid \omega_2) = \$ \quad 100.00$
$\lambda_{21} = \lambda(\alpha_2 \mid \omega_1) = \$ \quad 0.01$
$\lambda_{22} = \lambda(\alpha_1 \mid \omega_1) = \$ \quad 0.01$

Therefore our rule becomes

$$(\lambda_{21} - \lambda_{11})\, P(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22})\, P(x \mid \omega_2)\, P(\omega_2)$$

$$0.01 \;\; P(x \mid \omega_1)\, P(\omega_1) > 99.99\, P(x \mid \omega_2)\, P(\omega_2)$$

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

decide ω_1 if:

$$(\lambda_{21} - \lambda_{11})\, P(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22})\, P(x \mid \omega_2)\, P(\omega_2)$$

and decide ω_2 otherwise

Special Case:
Zero-One Loss Function

- Assign the same loss to all errors:

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

- All errors are equally costly.
- The conditional risk corresponding to this loss function:

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

$$= \sum_{j \neq i} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

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

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 turns out to be the average probability error!

Loss function

Let $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ denote the loss for deciding class i when the true class is j

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x})$$
$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

In minimizing the risk, we decide class one if

$$R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$$

Rearrange it, we have

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

Loss function

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

Example:

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

$$then \ \ \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

$$\lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$$

$$then \ \ \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$

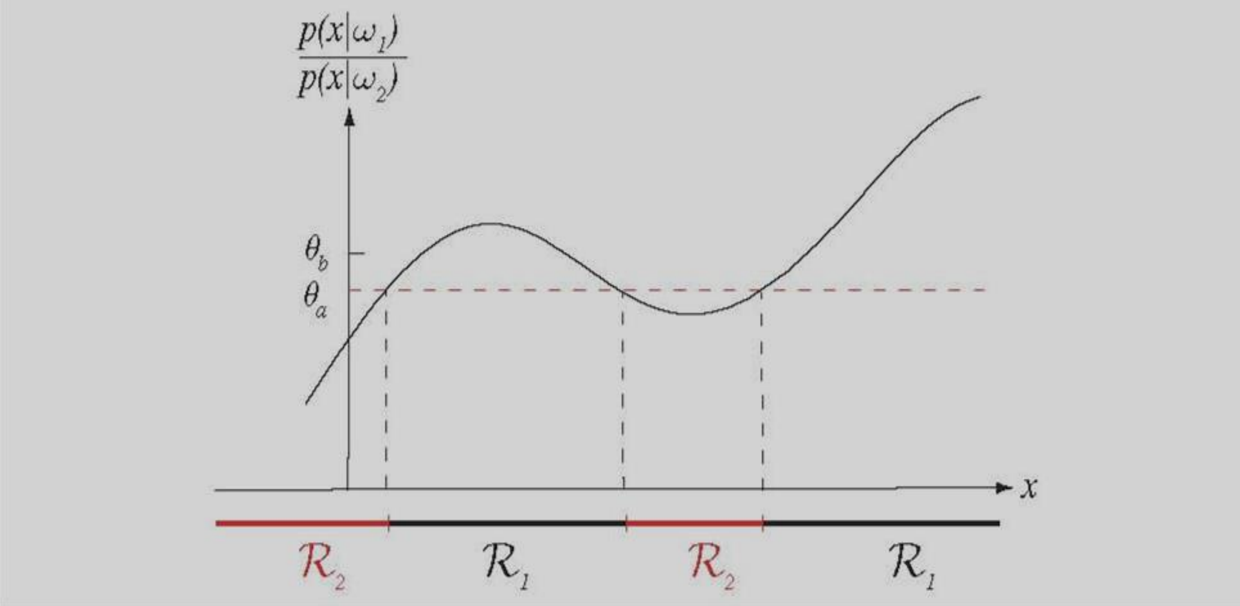
Example

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)

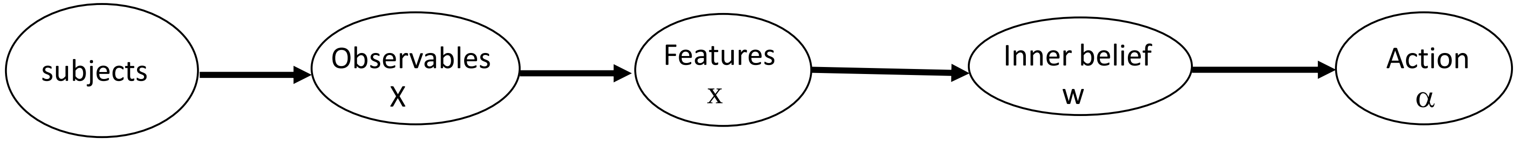
$$\theta_a = P(\omega_2)/P(\omega_1)$$

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

assume: $\lambda_{12} > \lambda_{21}$

Diagram of pattern classification

Procedure of pattern recognition and decision making



X--- all the observables using existing sensors and instruments
x --- is a set of features selected from components of X, or linear/non-linear functions of X.
w --- is our inner belief/perception about the subject class.
α --- is the action that we take for x.

We denote the three spaces by

$$x \in \Omega^{\mathrm{d}}, \quad w \in \Omega^{\mathrm{C}}, \quad \alpha \in \Omega^{\alpha}$$

$$x = (x_1, x_2, \ldots, x_{\mathrm{d}}) \quad \text{is a vector}$$

$$w \;\; \text{is the index of class,} \;\; \Omega^{\mathrm{C}} = \{w_1, w_2, \ldots, w_{\mathrm{k}}\}$$

Examples

Ex 1: Fish classification

$X = I$ is the image of fish,

x =(brightness, length, fin#, ...)

w is our belief what the fish type is

Ω^c={"sea bass", "salmon", "trout", …}

α is a decision for the fish type,

in this case $\Omega^c = \Omega^\alpha$

Ω^α ={"sea bass", "salmon", "trout", …}

Ex 2: Medical diagnosis

X= all the available medical tests, imaging scans that a doctor can order for a patient

x =(blood pressure, glucose level, cough, x-ray….)

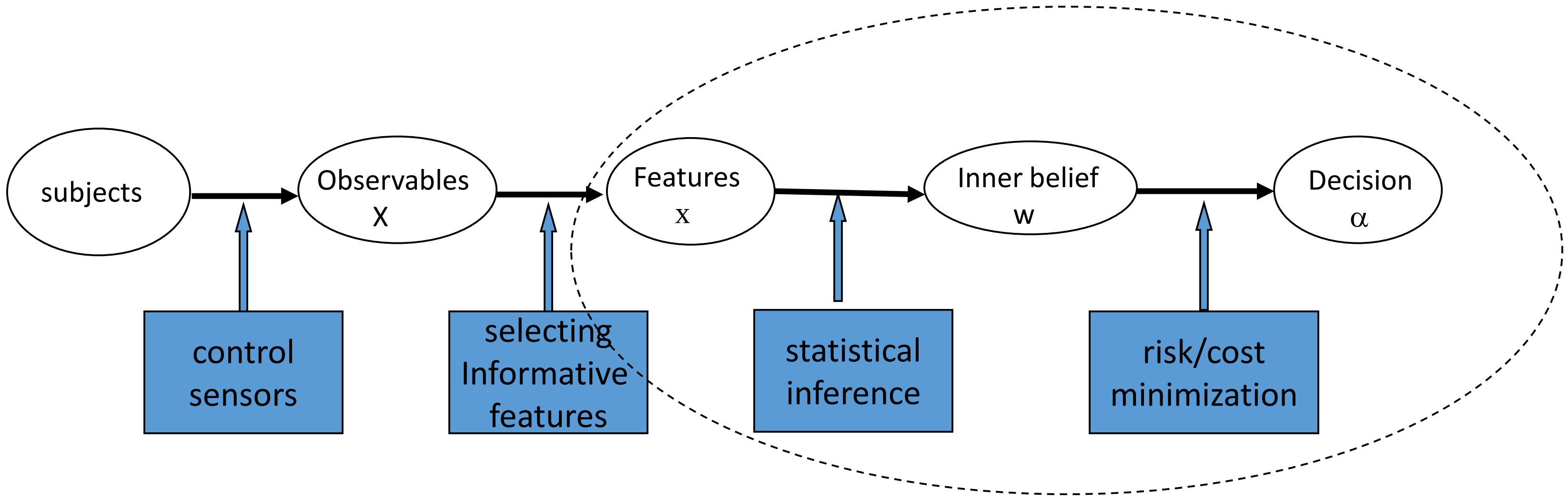
w is an illness type

Ω^c={"Flu", "cold", "TB", "pneumonia", "lung cancer"…}

α is a decision for treatment,

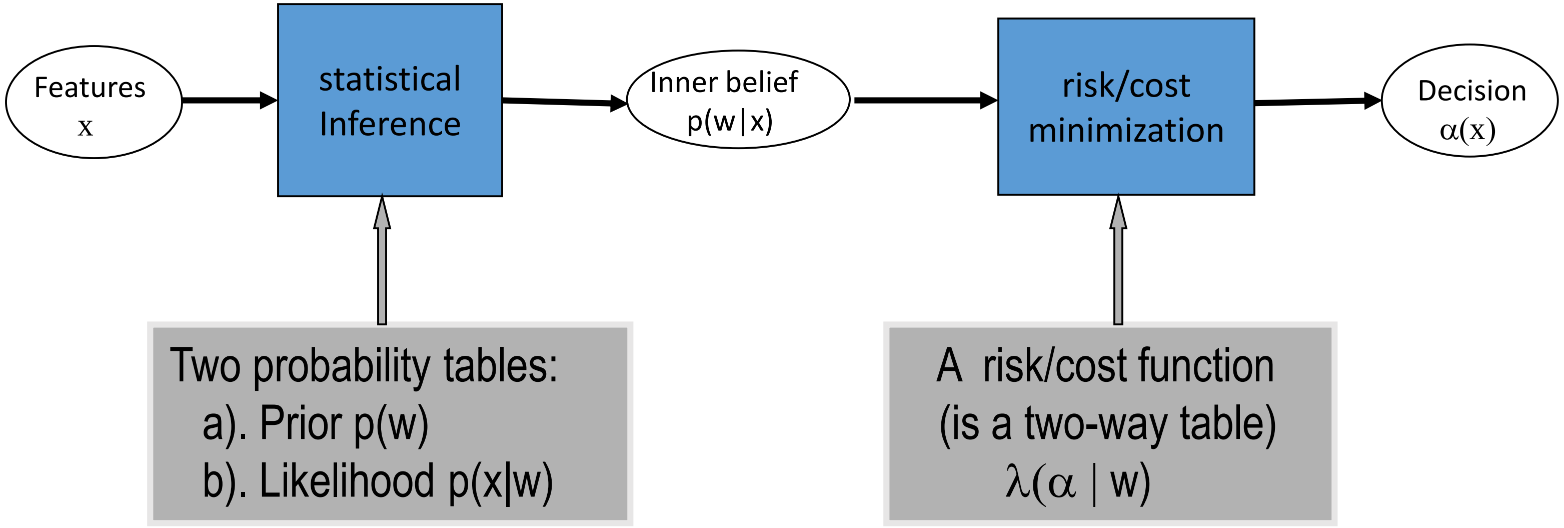
Ω^α ={"Tylenol", "Hospitalize", …}

Tasks



In Bayesian decision theory, we are concerned with the last three steps in the big ellipse assuming that the observables are given and features are selected.

Bayesian Decision Theory



The belief on the class w is computed by the Bayes rule

$$p(w \mid x) = \frac{p(x \mid w)\, p(w)}{p(x)}$$

The risk is computed by

$$R(\alpha_{\mathrm{i}} \mid x) = \sum_{\mathrm{j}=1}^{k} \lambda(\alpha_{\mathrm{i}} \mid \mathrm{w}_{\mathrm{j}})\mathrm{p}(\mathrm{w}_{\mathrm{j}} \mid x)$$

Decision Rule

A decision rule is a mapping function from feature space to the set of actions

$$\alpha(x):\ \Omega^{\mathrm{d}} \rightarrow \Omega^{\alpha}$$

we will show that randomized decisions won't be optimal.

A decision is made to minimize the average cost / risk,

$$R = \int R(\alpha(x) \mid x)\, p(x)\ \mathrm{dx}$$

It is minimized when our decision is made to minimize the cost / risk for each instance x.

$$\alpha(x) = \arg\min_{\Omega^{\alpha}}\ R(\alpha \mid x) \ = \arg\min_{\Omega^{\alpha}} \sum_{j=1}^{k} \lambda(\alpha \mid w_j)\, p(w_j \mid x)$$

Bayesian error

In a special case, like fish classification, the action is classification, we assume a 0/1 error.

$$\lambda(\alpha_i | w_j) = 0 \quad if \quad \alpha_i = w_j$$
$$\lambda(\alpha_i | w_j) = 1 \quad if \quad \alpha_i \neq w_j$$

The risk for classifying x to class α_i is,

$$R(\alpha_i | x) = \sum_{w_j \neq \alpha_i} p(w_j | x) = 1 - p(\alpha_i | x)$$

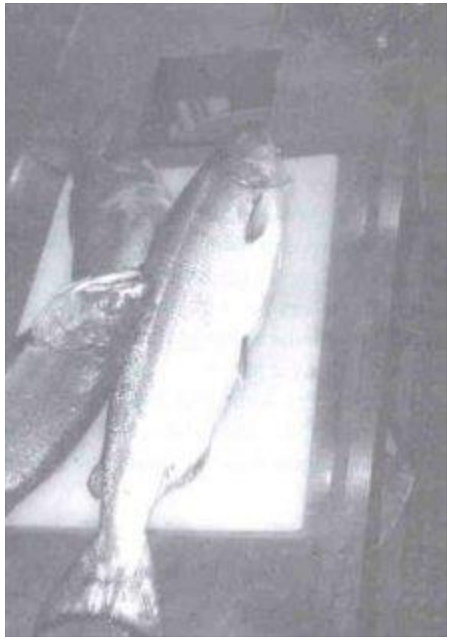
The optimal decision is to choose the class that has maximum posterior probability

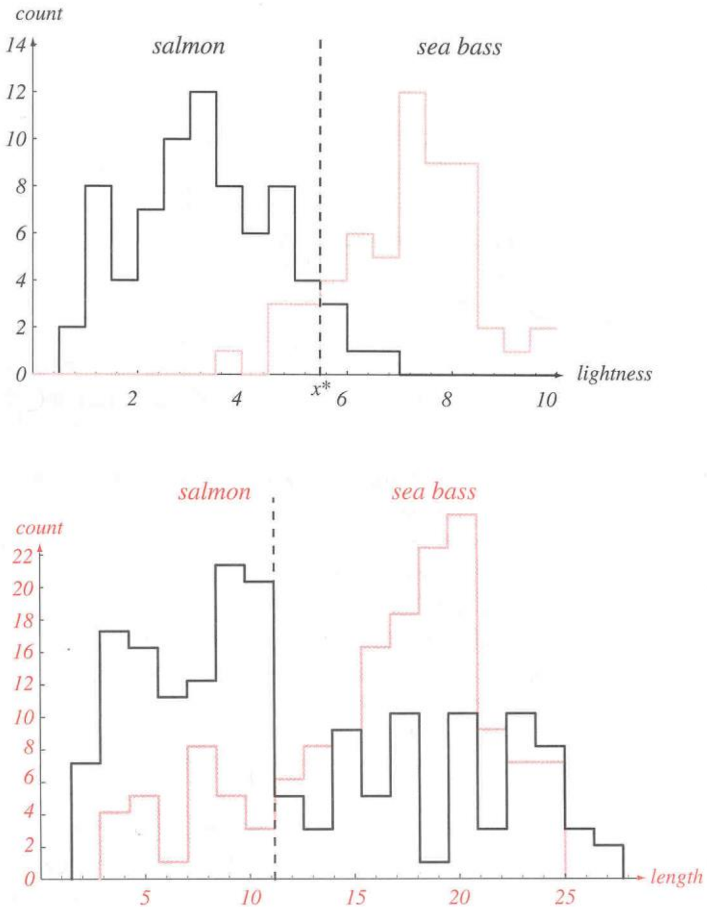
$$\boxed{\alpha(x) = \arg\min_{\Omega^\alpha} (1 - p(\alpha | x)) = \arg\max_{\Omega^\alpha} p(\alpha | x)}$$

The total risk for a decision rule, in this case, is called the Bayesian error

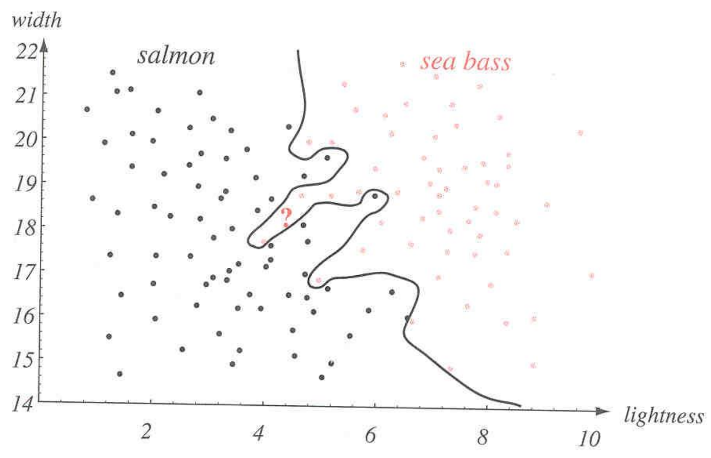
$$R = p(error) = \int p(error | x) p(x) dx = \int (1 - p(\alpha(x) | x)) p(x) dx$$

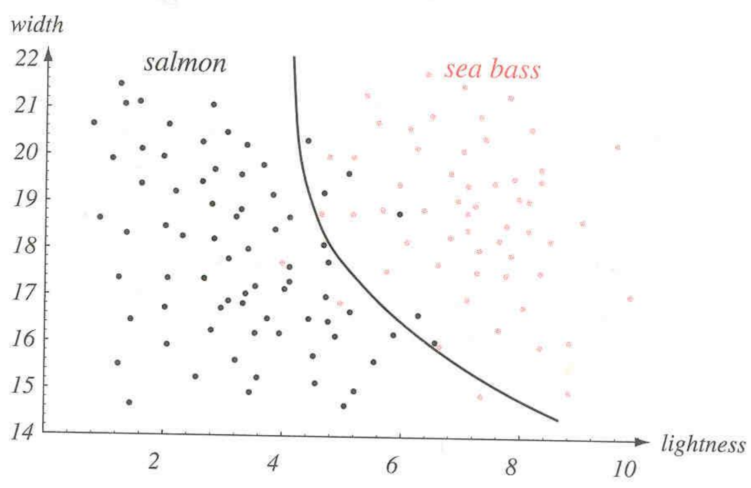
An example of fish classification





Decision/classification Boundaries





Discriminant Functions

- **Discriminant Functions** are a useful way of representing pattern classifiers.
- Let's say $g_i(\mathbf{x})$ is a discriminant function for the ith class.
- This classifier will assign a class ω_i to the feature vector $\mathbf{x}$ if

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

or, equivalently

$$i^* = \arg\max_i g_i(x) \ , \quad \text{decide} \quad \omega_{i^*} \ .$$

$$\alpha(x) = \arg\max \{g_1(x), g_2(x), \ldots\ldots, \ g_k(x)\}$$

Decision surface defined by

$g_i(x) = g_j(x)$

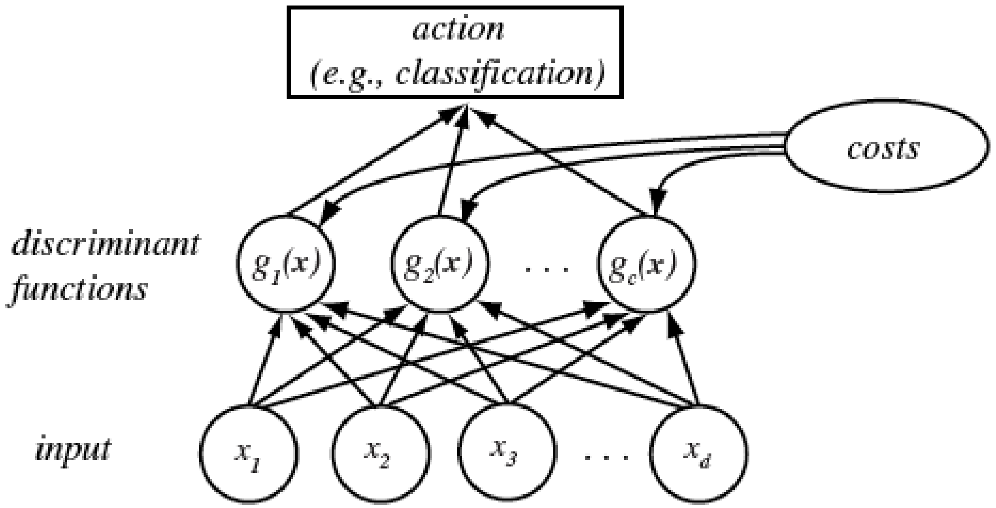


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction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

Bayes Discriminants

- General case with risks

$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x}) \tag{27}$$

$$= -\sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \tag{28}$$

- Can we prove that this is correct?
- **Yes!** The minimum conditional risk corresponds to the maximum discriminant.

- In the case of zero-one loss function, the Bayes Discriminant can be further simplified:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \ . \tag{29}$$

Uniqueness of discriminants

- Is the choice of discriminant functions unique?
- No!
- Multiply by some positive constant.
- Shift them by some additive constant.
- For monotonically increasing function $f(\cdot)$, we can replace each $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$ without affecting our classification accuracy.
  - These can help for ease of understanding or computability.
  - The following all yield the same exact classification results for minimum-error-rate classification.

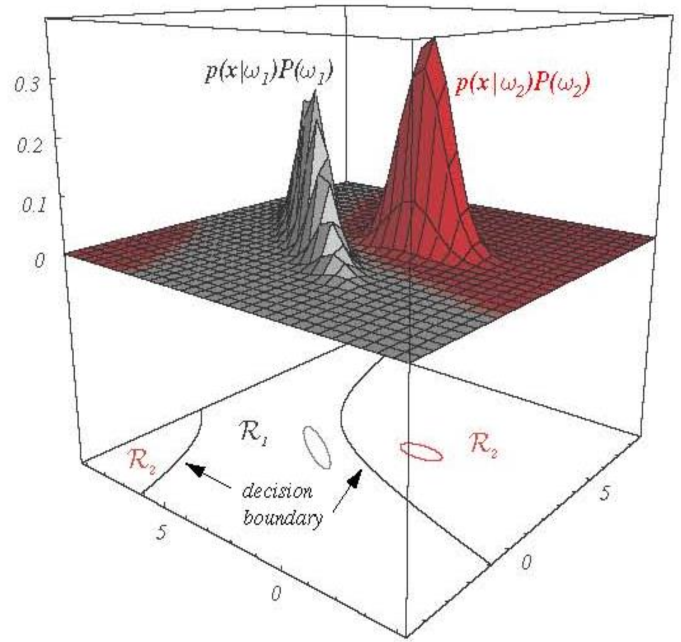
$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_j p(\mathbf{x}|\omega_j)P(\omega_j)} \tag{30}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i) \tag{31}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \tag{32}$$

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:

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

Case of two categories

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

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

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

- Examples:

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

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

Discriminant function for discrete features

Discrete features: $x = [x_1, x_2, \ldots, x_d]^t$, $x_i \in \{0,1\}$

$$p_i = P(x_i = 1 \mid \omega_1)$$

$$q_i = P(x_i = 1 \mid \omega_2)$$

The likelihood will be:

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

$$P(\mathbf{x}|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}$$

Discriminant function for discrete features

The discriminant function:

$$g_1(x) = \log\left[p(\mathbf{x}|\omega_1)p(\omega_1)\right]$$

$$g_2(x) = \log\left[p(\mathbf{x}|\omega_2)p(\omega_2)\right]$$

The likelihood ratio:

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} = \prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

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

$$= \sum_{i=1}^{d}\left[x_i \ln\frac{p_i}{q_i} + (1-x_i)\ln\frac{1-p_i}{1-q_i}\right] + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

Discriminant function for discrete features

So the decision surface is again a hyperplane.

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

$$w_i = \ln \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \qquad i = 1, \dots, d$$

$$w_0 = \sum_{i=1}^{d} \ln \frac{1 - p_i}{1 - q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

The Univariate Normal Density

- Easy to work with analytically
- A lot of processes are asymptotically Gaussian
- Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

- Continuous univariate normal, or **Gaussian**, density:

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

- The **mean** is the expected value of x is

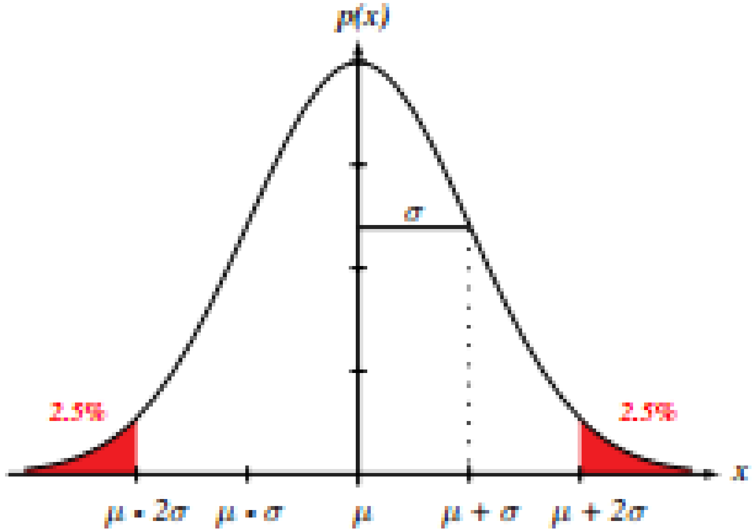
$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} x p(x) dx .$$

- The **variance** is the expected squared deviation

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

Univariate Normal Density

- Samples from the normal density tend to cluster around the mean and be spread-out based on the variance.



- The normal density is completely specified by the mean and the variance. These two are its **sufficient statistics**.
- We thus abbreviate the equation for the normal density as

$$p(x) \sim N(\mu, \sigma^2) \tag{43}$$

Entropy

- **Entropy** is the uncertainty in the random samples from a distribution.

$$H(p(x)) = -\int p(x) \ln p(x) dx \tag{44}$$

- The normal density has the maximum entropy for all distributions have a given mean and variance.
- What is the entropy of the uniform distribution?
- The uniform distribution has maximum entropy (on a given interval).

Multivariate density: N(μ , Σ)

- Multivariate normal density in d dimensions:

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

where:

$x = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose of a vector)

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d*d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse of Σ, respectively

- The covariance matrix is always symmetric and positive semidefinite; we assume Σ is positive definite so the determinant of Σ is strictly positive
- Multivariate normal density is completely specified by [d + d(d+1)/2] parameters
- If variables x_1 *and* x_2 are *statistically independent* then the *covariance* of x_1 *and* x_2 *is zero.*

Multivariate Normal Density

- The multivariate Gaussian in d dimensions is written as

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

- Again, we abbreviate this as $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- The sufficient statistics in d-dimensions:

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

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

The Covariance Matrix

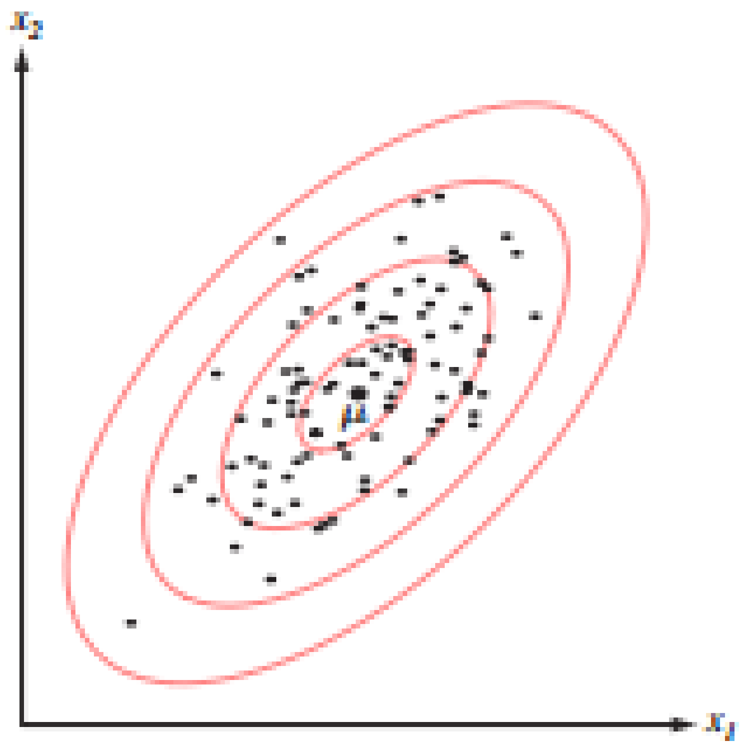
$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements σ_{ii} are the variances of the respective coordinate x_i.
- The off-diagonal elements σ_{ij} are the covariances of x_i and x_j.
- What does a $\sigma_{ij} = 0$ imply?
- That coordinates x_i and x_j are statistically independent.
- What does $\boldsymbol{\Sigma}$ reduce to if all off-diagonals are 0?
- The product of the d univariate densities.

Mahalanobis Distance

- The shape of the density is determined by the covariance $\mathbf{\Sigma}$.
- Specifically, the eigenvectors of $\mathbf{\Sigma}$ give the principal axes of the hyperellipsoids and the eigenvalues determine the lengths of these axes.
- The loci of points of constant density are hyperellipsoids with constant **Mahalonobis distance**:

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \qquad (48)$$



Reminder of some results for random vectors

Let Σ be a *kxk* square symmetrix matrix, then it has *k* pairs of eigenvalues and eigenvectors. A can be decomposed as:

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \ldots\ldots + \lambda_k e_k e_k' = P \Lambda P'$$

Positive-definite matrix:

$$x' \Sigma x > 0, \forall x \neq 0$$

$$\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_k > 0$$

$$Note: x' \Sigma x = \lambda_1 (x' e_1)^2 + \ldots\ldots + \lambda_k (x' e_k)^2$$

Normal density

Whitening transform:

P : eigen vector matrix

Λ : diagonal eigen value matrix

$$A_w = P\Lambda^{-\frac{1}{2}}$$

$$A_w^t \Sigma A_w$$

$$= \Lambda^{-\frac{1}{2}} P^t \Sigma P \Lambda^{-\frac{1}{2}}$$

$$= \Lambda^{-\frac{1}{2}} P^t P \Lambda P^t P \Lambda^{-\frac{1}{2}}$$

$$= I$$

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \ldots\ldots + \lambda_k e_k e_k' = P \Lambda P'$$

Linear Combinations of Normals

- Linear combinations of jointly normally distributed random variables, independent or not, are normally distributed.
- For $p(\mathbf{x}) \sim N((\mu), \mathbf{\Sigma})$ and $\mathbf{A}$, a d-by-k matrix, define $\mathbf{y} = \mathbf{A}^\top \mathbf{x}$. Then:

$$p(\mathbf{y}) \sim N(\mathbf{A}^\top \boldsymbol{\mu}, \mathbf{A}^\top \mathbf{\Sigma} \mathbf{A}) \quad (49)$$

- With the covariance matrix, we can calculate the dispersion of the data in any direction or in any subspace.

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \mathbf{\Sigma}) \qquad \mathbf{y} = \mathbf{A}^t \mathbf{x}$$

$$p(\mathbf{y}) \sim N(\mathbf{A}^t \boldsymbol{\mu}, \mathbf{A}^t \mathbf{\Sigma} \mathbf{A})$$

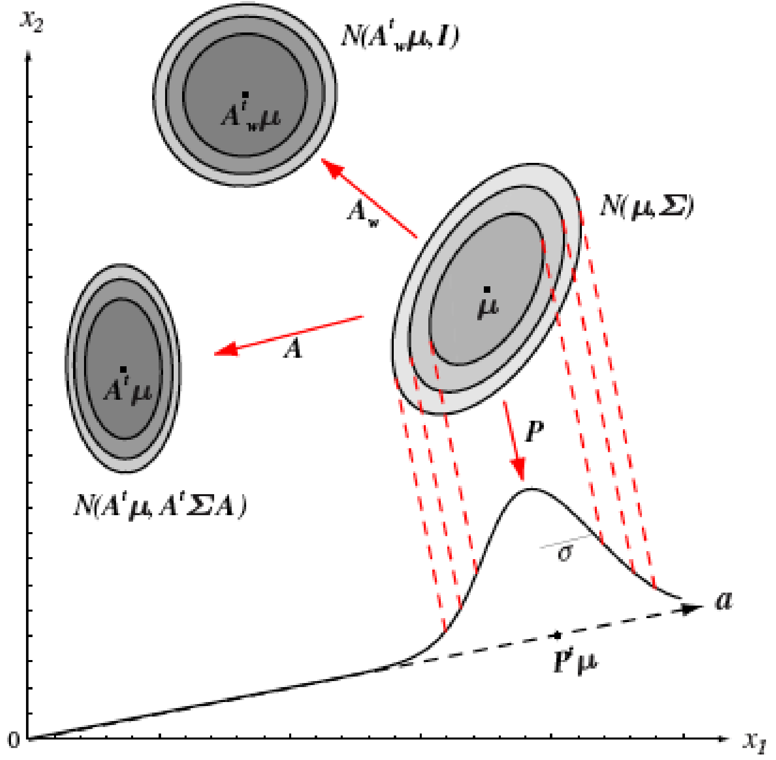


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, $\mathbf{A}$, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\mathbf{\Sigma}\mathbf{A})$. Another linear transformation—a projection $\mathbf{P}$ onto a line defined by vector $\mathbf{a}$—leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original x_1x_2-space. A whitening transform, $\mathbf{A}_w$, leads to a circularly symmetric Gaussian, here shown displaced. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

General Discriminant Function for Multivariate Gaussian Density

Recall the minimum error rate discriminant

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

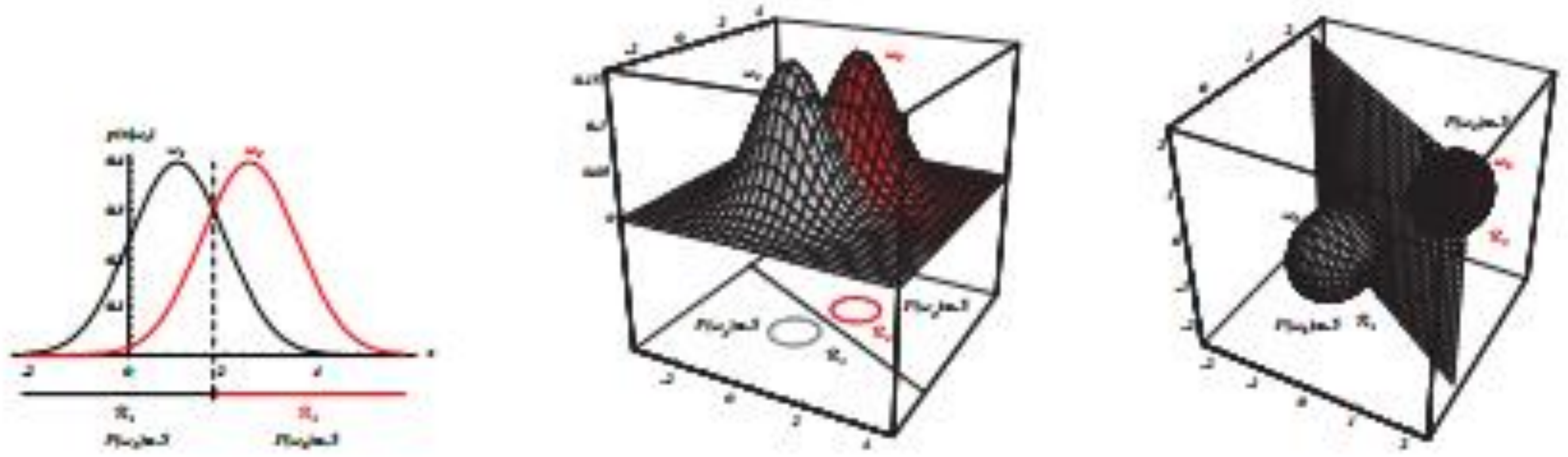
$$N(\mu,\Sigma) \quad = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\mathbf{x}-\mu)^t\Sigma^{-1}(\mathbf{x}-\mu)]$$

- If we assume normal densities, i.e., if $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, then the general discriminant is of the form

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

Simple Case: Statistically Independent Features with Same Variance

- What do the decision boundaries look like if we assume $\Sigma_i = \sigma^2 \mathbf{I}$?
- They are hyperplanes.



A classifier that uses linear discriminant functions is called “a linear machine”
The decision surfaces for a linear machine are pieces of hyperplanes defined by:

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

Multivariate Gaussian Density: Case I

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t\Sigma_i^{-1}(\mathbf{x}-\mu_\mathbf{i}) \boxed{-\frac{d}{2}\,ln2\pi - \frac{1}{2}\ln|\Sigma_i|} + \ln P(\omega_i)$$

- **$\Sigma_i = \sigma^2$ I** (diagonal matrix)
  - Features are statistically independent
  - Each feature has the same variance

- Think of this discriminant as a combination of two things
  1. The distance of the sample to the mean vector (for each i).
  2. A normalization by the variance and offset by the prior.

- If we disregard $\frac{d}{2}\,ln2\pi$ and $\frac{1}{2}\ln|\Sigma_i|$ (constants):

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

where $\|\mathbf{x}-\mu_i\|^2 = (\mathbf{x}-\mu_\mathbf{i})^t(\mathbf{x}-\mu_\mathbf{i})$

- Expanding the above expression:

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

Case1 : $\Sigma_i = \sigma^2 I$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\boxed{\mathbf{x}^\top\mathbf{x}} - 2\boldsymbol{\mu}_i^\top\mathbf{x} + \boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \ . \qquad (52)$$

- The quadratic term $\mathbf{x}^\top\mathbf{x}$ is the same for all i and can thus be ignored.
- This yields the equivalent **linear discriminant functions**

$$g_i(\mathbf{x}) = \mathbf{w}_i^\top\mathbf{x} + w_{i0} \qquad (53)$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \qquad (54)$$

$$w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i + \ln P(\omega_i) \qquad (55)$$

- w_{i0} is called the **bias**.

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

- The decision surfaces for a linear discriminant classifiers are hyperplanes defined by the linear equations $g_i(\mathbf{x}) = g_j(\mathbf{x})$.
- The equation can be written as

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \tag{56}$$

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \tag{57}$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{58}$$

- These equations define a hyperplane through point x_0 with a normal vector $\mathbf{w}$.

i.e. With equal prior, x_0 is the middle point between the two means.
The decision surface is a hyperplane, perpendicular to the line between the means.

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

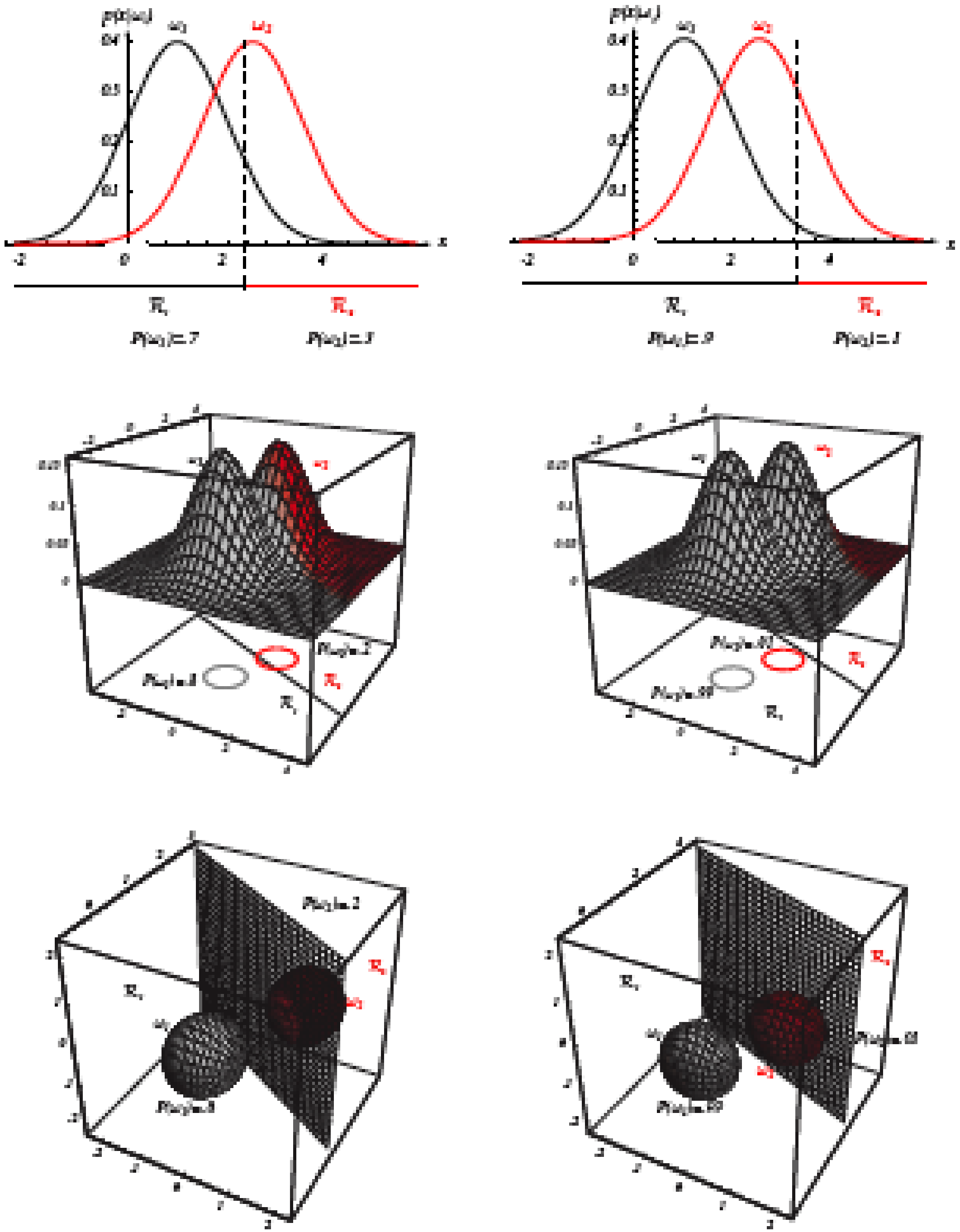
$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

$$\text{where } \mathbf{w} = \mu_i - \mu_j, \text{ and } \mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$$

- Properties of decision boundary:
  - It passes through $\mathbf{x_0}$
  - It is orthogonal to the line linking the means.
  - If σ is very small, the position of the boundary is insensitive to $P(\omega_i)$ and $P(\omega_j)$

When $P(\omega_i)$ are equal, then the discriminant becomes:

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



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

With unequal prior probabilities, the decision boundary shifts to the less likely mean.

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

$$x_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

Multivariate Gaussian Density: Case 2

Recall $$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t\Sigma_i^{-1}(\mathbf{x}-\mu_\mathbf{i}) \boxed{-\frac{d}{2}\,ln2\pi - \frac{1}{2}\ln|\Sigma_i|} + \ln P(\omega_i)$$

- **$\Sigma_i = \Sigma$**

- The clusters have hyperellipsoidal shape and same size (centered at μ).

- If we disregard $\frac{d}{2}\,ln2\pi$ and $\frac{1}{2}\ln|\Sigma_i|$ (constants):

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

- Expanding the above expression and disregarding the quadratic term:

$$\boxed{\begin{array}{c} g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0} \\ (\textit{linear discriminant}) \end{array}}$$

where $\mathbf{w}_i = \Sigma^{-1}\mu_i$, and $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + \ln P(\omega_i)$

Case 2 : $\Sigma_i = \Sigma$

- Decision boundary is determined by hyperplanes; setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$:

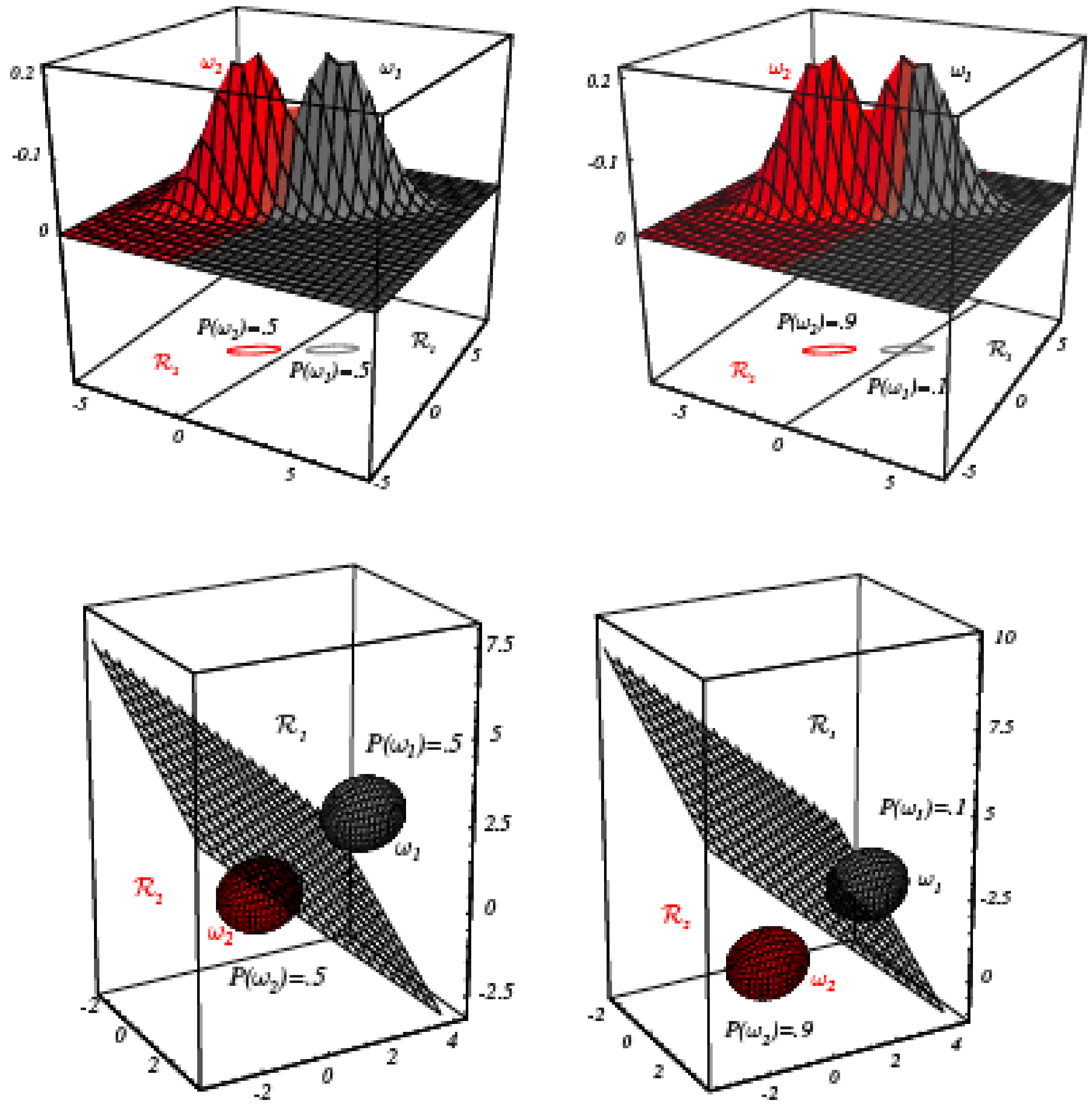
$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

where $\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$ and $\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$

Case 2 : $\Sigma_i = \Sigma$

Properties of hyperplane (decision boundary) for equal but asymmetric Gaussian distributions

- It passes through $\mathbf{x}_0$
- It is not orthogonal to the line between the means.
- If *$P(\omega_i)$ and $P(\omega_j)$ are not equal,* then $\mathbf{x}_0$ shifts away from the most likely category.



Case 2 : $\Sigma_i = \Sigma$

- Mahalanobis distance classifier
  - When $P(\omega_i)$ are equal, then the discriminant becomes:

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

$$\Downarrow$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_\mathbf{i})^t \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{i})$$

Multivariate Gaussian Density: Case 3

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

- Σ_i= arbitrary

- The clusters have different shapes and sizes (centered at μ).

- If we disregard $\frac{d}{2} ln 2\pi$ (constant):

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

(quadratic discriminant)

where $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$, $\mathbf{w}_i = \Sigma_i^{-1}\mu_i$, and $w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i - \frac{1}{2} ln|\Sigma_i| + \ln P(\omega_i)$

- Decision boundary is determined by superquadrics; setting hyperquadrics;

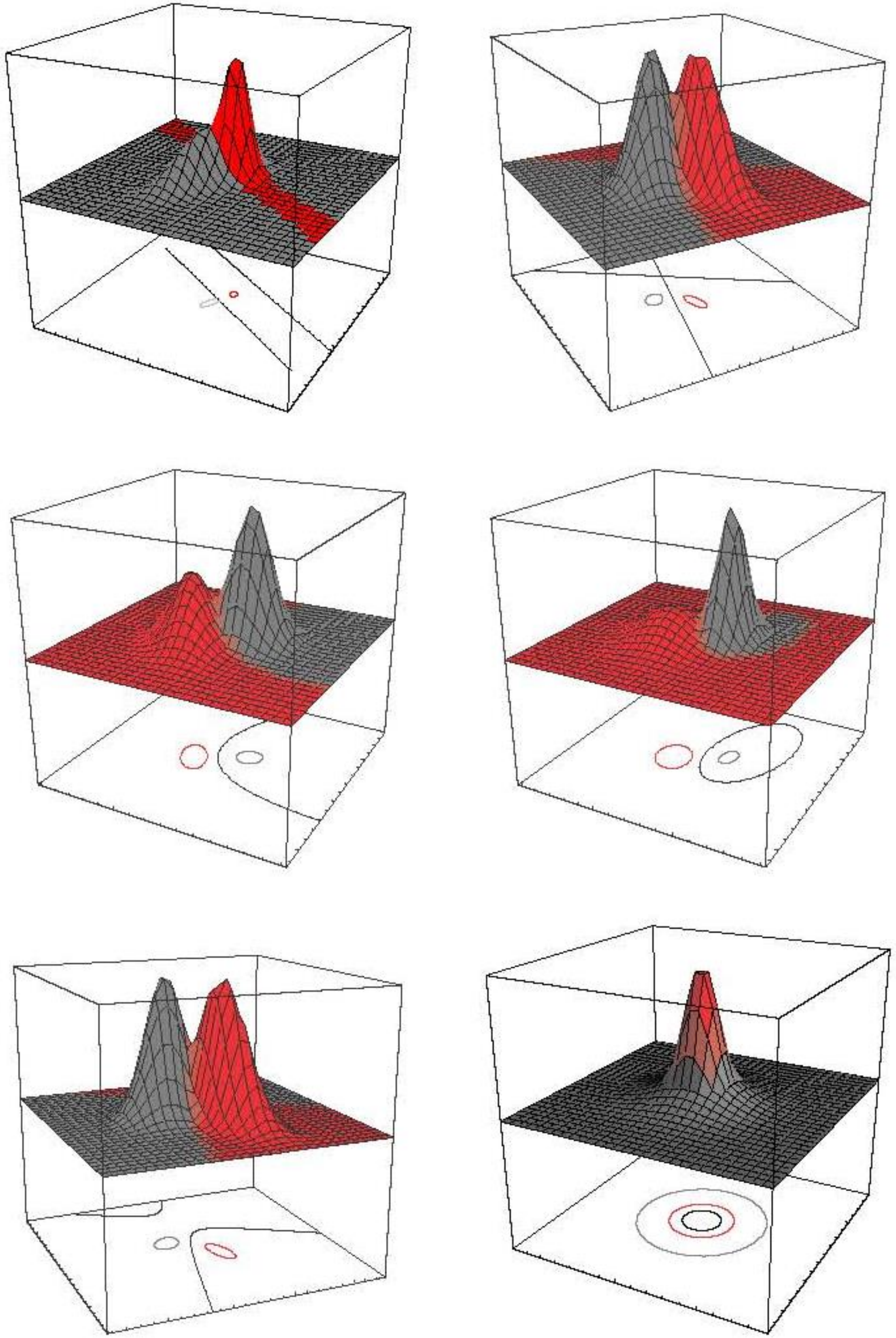
e.g., hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids etc.

Case 3: Σ_i= arbitrary

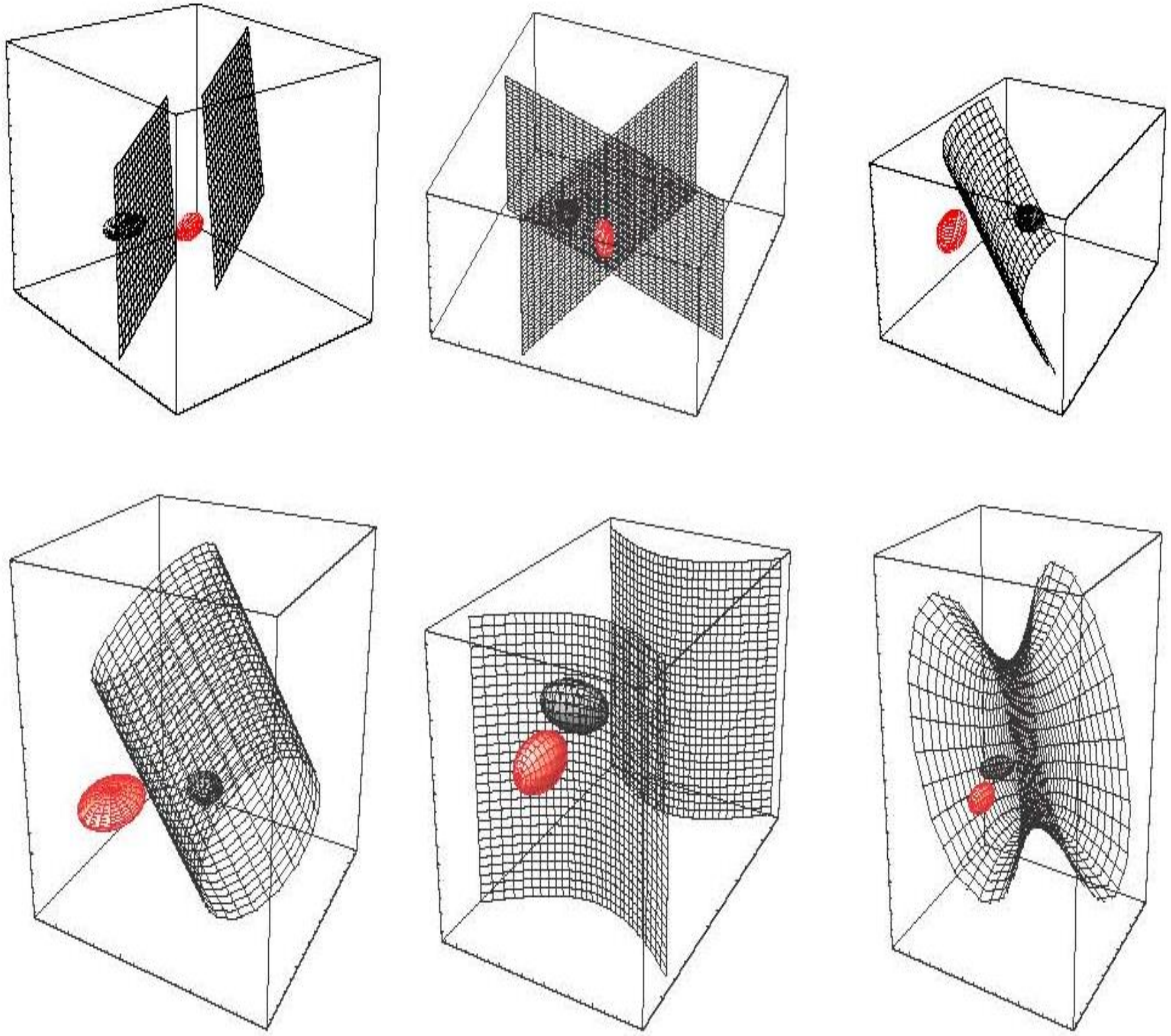
non-linear

decision

boundaries

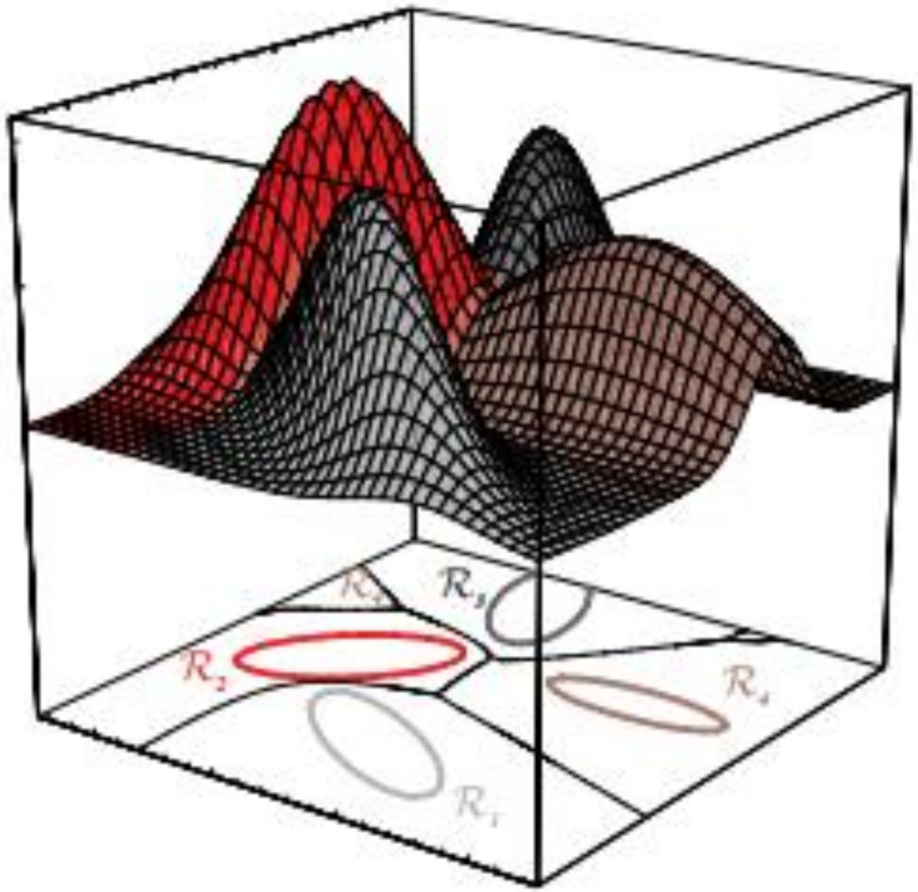


Case 3: Σ_i= arbitrary



Case 3: Σ_i= arbitrary

General Case for Multiple Categories



Quite A Complicated Decision Surface!

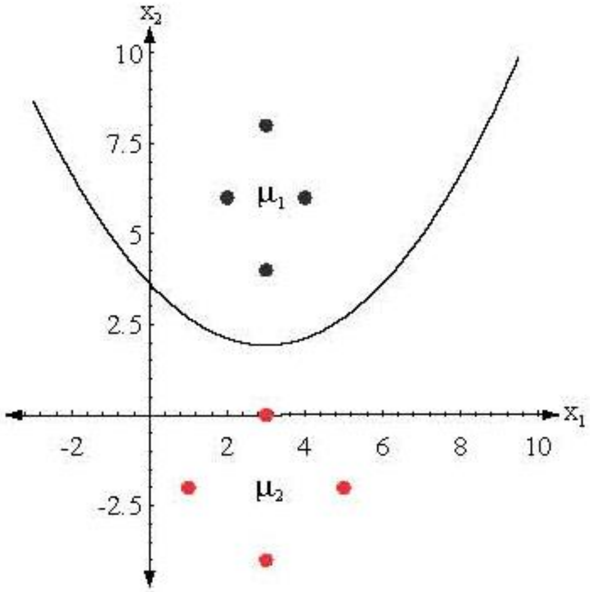
Example - Case 3

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

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

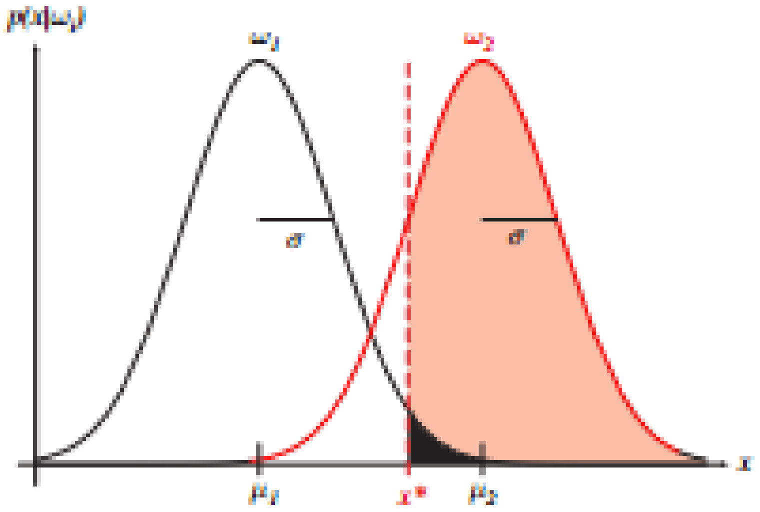
$P(\omega_1)=P(\omega_2)$

boundary does not pass through midpoint of μ_1,μ_2



Signal Detection Theory

- A fundamental way of analyzing a classifier.
- Consider the following experimental setup:



- Suppose we are interested in detecting a single pulse.
- We can read an internal signal x.
- The signal is distributed about mean μ_2 when an external signal is present and around mean μ_1 when no external signal is present.
- Assume the distributions have the same variances, $p(x|\omega_i) \sim N(\mu_i, \sigma^2)$.

Signal Detection Theory

- The detector uses x^* to decide if the external signal is present.
- **Discriminability** characterizes how difficult it will be to decide if the external signal is present without knowing x^*.

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma} \tag{63}$$

- Even if we do not know μ_1, μ_2, σ, or x^*, we can find d' by using a **receiver operating characteristic** or ROC curve, as long as we know the state of nature for some experiments

Receiver Operating Characteristics

- A **Hit** is the probability that the internal signal is above x^* given that the external signal is present

$$P(x > x^*|x \in \omega_2) \tag{64}$$

- A **Correct Rejection** is the probability that the internal signal is below x^* given that the external signal is not present.

$$P(x < x^*|x \in \omega_1) \tag{65}$$

- A **False Alarm** is the probability that the internal signal is above x^* despite there being no external signal present.

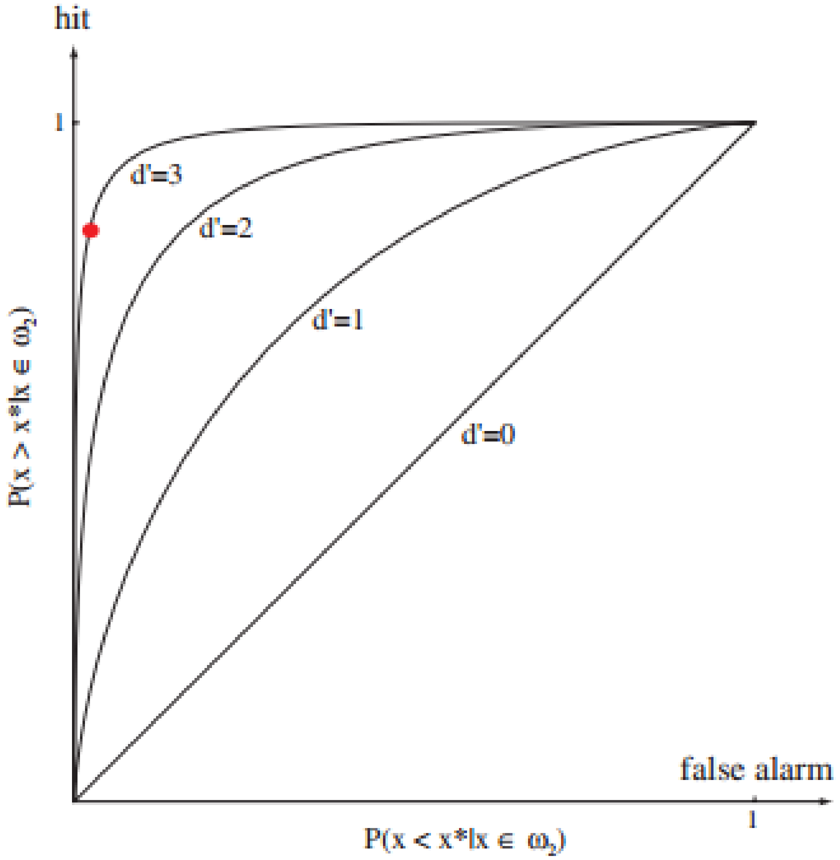
$$P(x > x^*|x \in \omega_1) \tag{66}$$

- A **Miss** is the probability that the internal signal is below x^* given that the external signal is present.

$$P(x < x^*|x \in \omega_2) \tag{67}$$

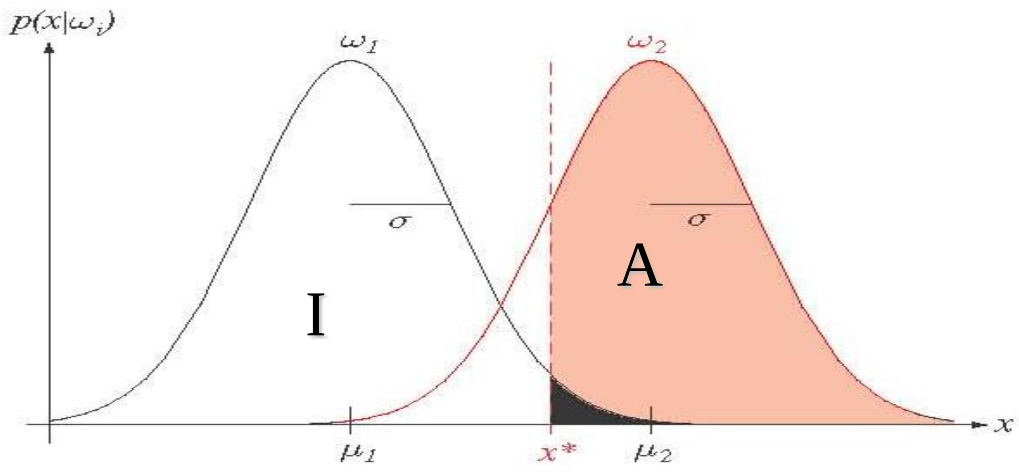
ROC

- We can experimentally determine the rates, in particular the Hit-Rate and the False-Alarm-Rate.
- Basic idea is to assume our densities are fixed (reasonable) but vary our threshold x^*, which will thus change the rates.
- The receiver operating characteristic plots the hit rate against the false alarm rate.
- What shape curve do we want?



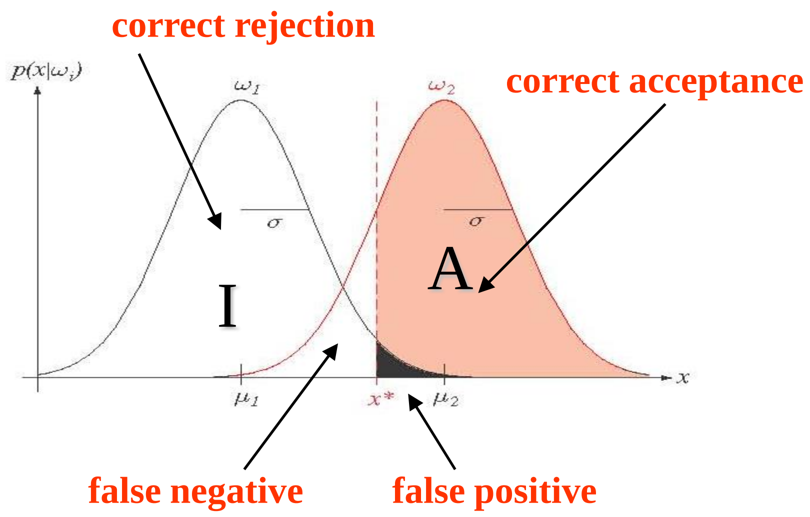
Example: Person Authentication

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



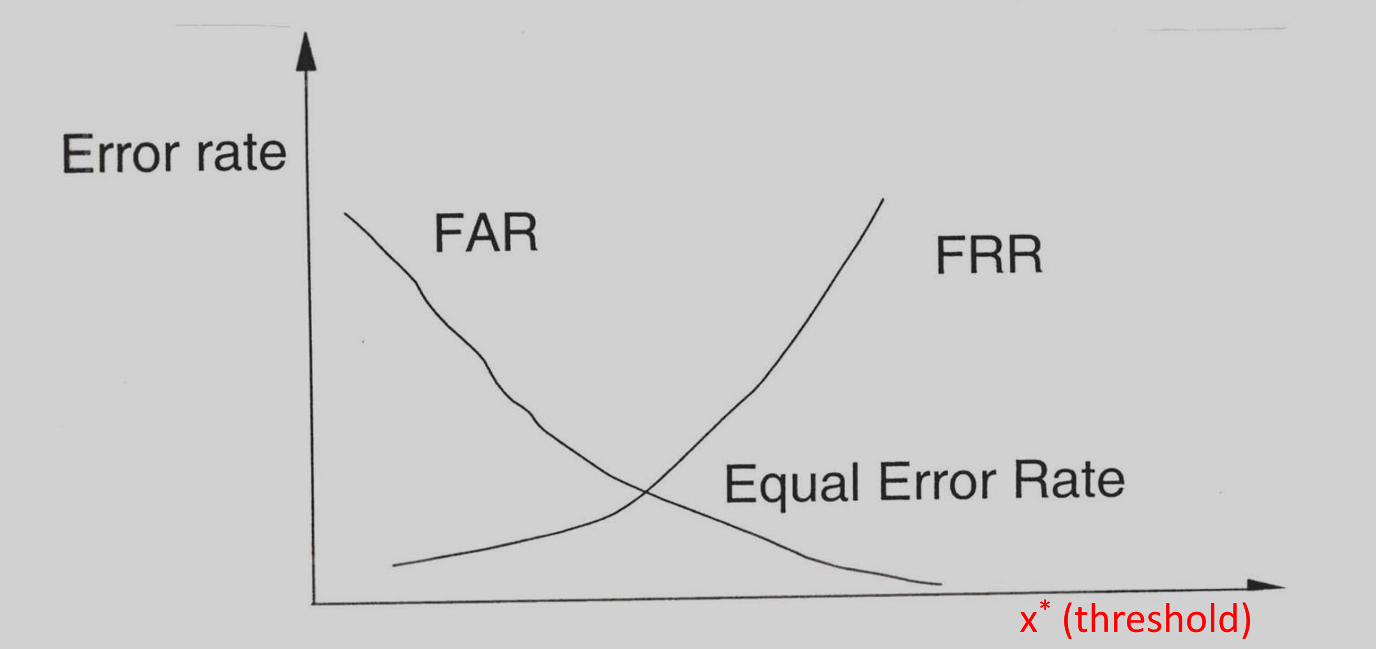
Example: Person Authentication

- 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



Error vs Threshold

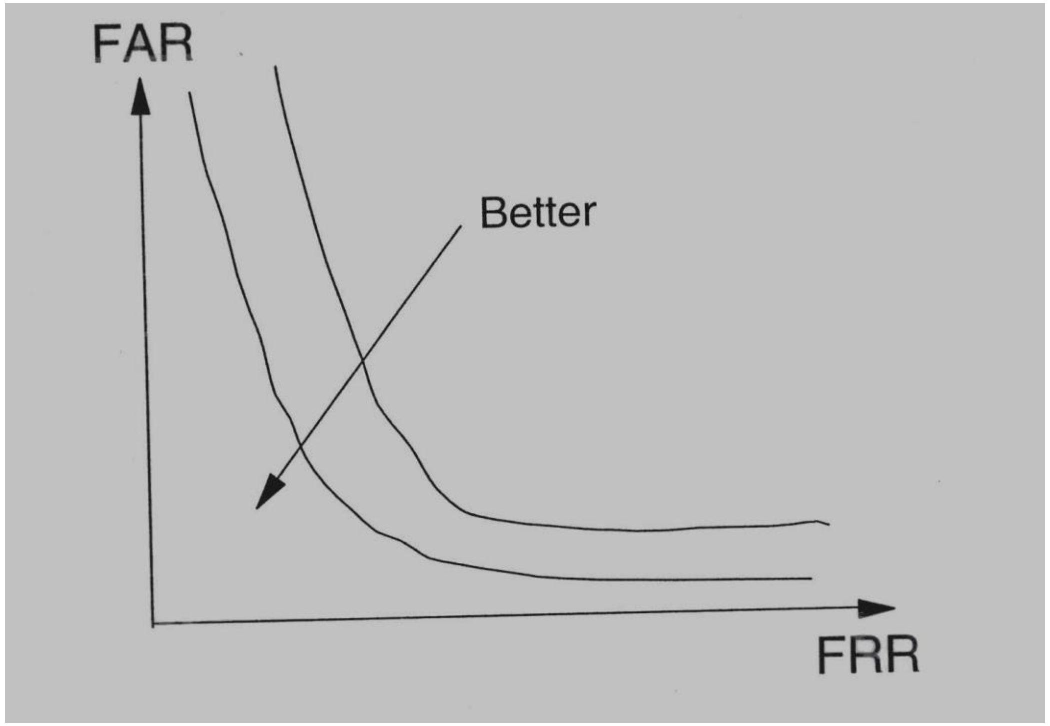
ROC Curve



FAR: False Accept Rate (False Positive)
FRR: False Reject Rate (False Negative)

False Negatives vs Positives

ROC Curve



FAR: False Accept Rate (False Positive)
FRR: False Reject Rate (False Negative)