BAYES CLASSIFIER

Dr. Umarani Jayaraman
Assistant Professor

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
DESIGN AND MANUFACTURING,
KANCHEEPURAM

Chapter 2

Bayesian Decision Theory

Decision Theory

Decision

Make choice under uncertainty

Pattern Recognition

Pattern → Category

Given a test sample, its category is uncertain and a decision has to be made

In essence, PR is a decision process

Bayesian Decision Theory

Bayesian decision theory is a statistical approach to pattern recognition

The fundamentals of most PR algorithms are rooted from Bayesian decision theory

Basic Assumptions

- The decision problem is posed (formalized) in **probabilistic** terms
- All the relevant probability values are known

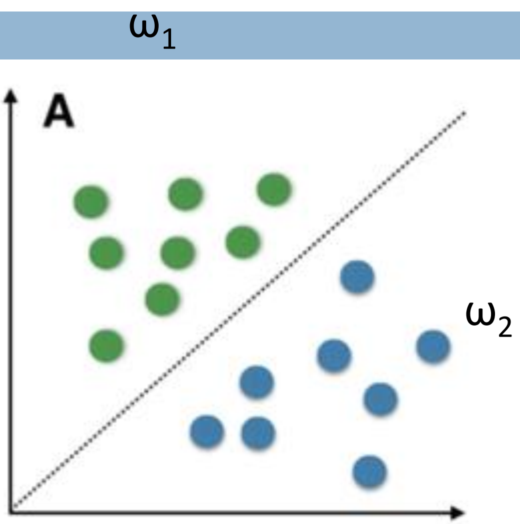
Key Principle

Bayes Theorem

Linear separable classes

Let ω_1 and ω_2

are two classes



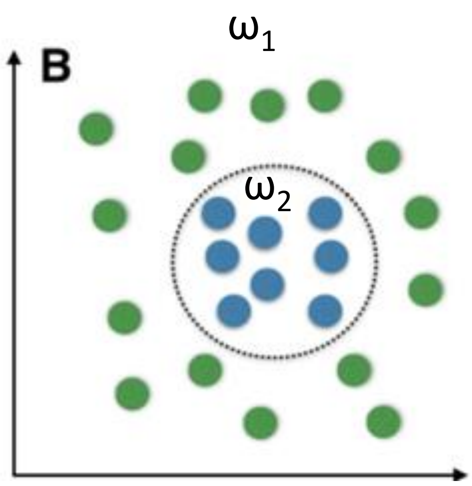
- In this case, the two classes can be separated by linear boundary, this is also known as linearly separable classes

- This is supervised learning

Non linear separable classes

Among these non-linear separable classes, the most common are

1. **Quadratic classifier**
2. **Cubic classifier**

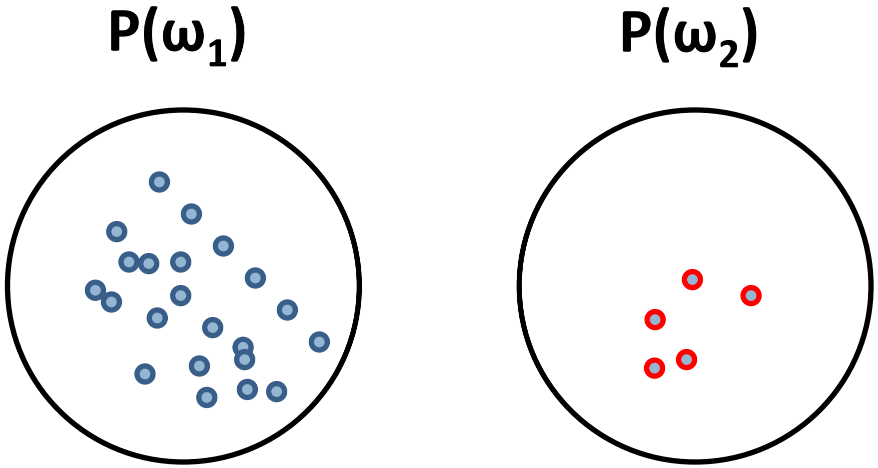


Two class problem

- **Consider any manufacturing company which produces goods for example steel plant**
- **Quality control department should take the decision in which of the two classes it should go**
- **ω_1 -> accept**
- **ω_2 -> reject**

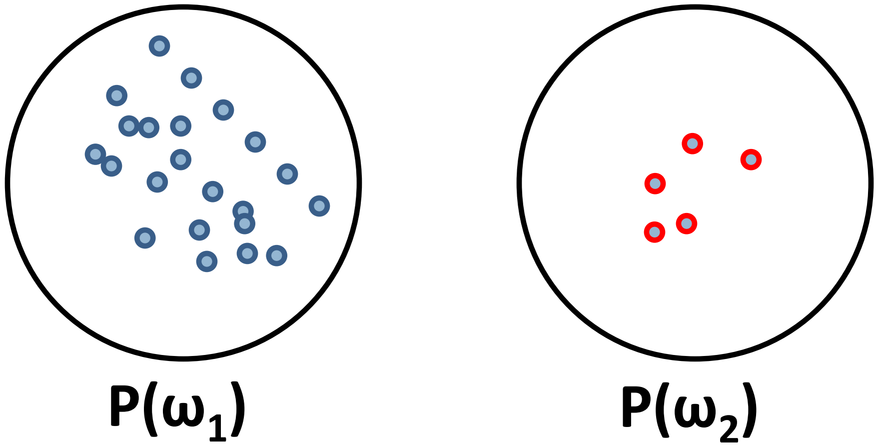
Two class problem

- ω_1 -> accept
- ω_2 -> reject
- We may take the previous history (i.e) how many objects are accepted and how many are rejected by the quality control department



Two class problem

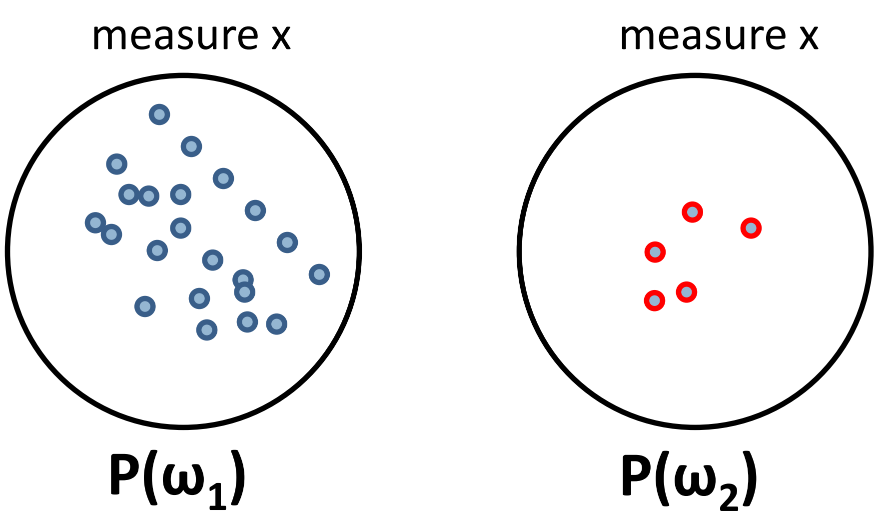
- $P(\omega_1) > P(\omega_2) \Rightarrow \omega_1$
- $P(\omega_1) < P(\omega_2) \Rightarrow \omega_2$
- **But, this is not really logical because the object is always accepted or always be rejected based on a priori probability (i.e) $p(\omega_1)$ and $p(\omega_2)$**



Solution: Incorporate observations into decision!

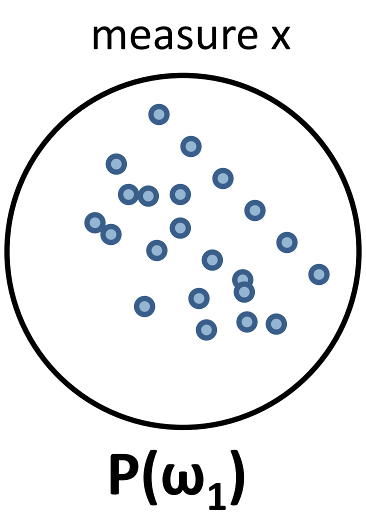
Two class problem

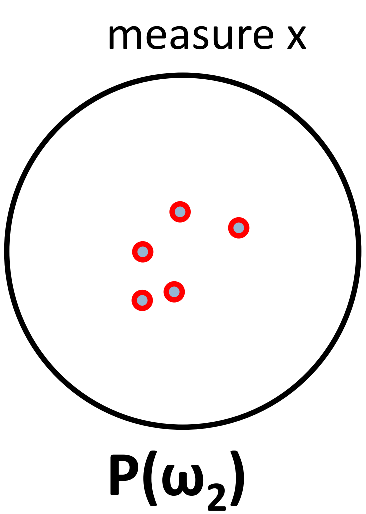
- Let the observation be x
- We can find out $P(x/\omega 1)$ and $P(x/\omega 2)$
- It is nothing but probability density function x taking the objects from class $\omega 1$ and $\omega 2$ respectively (class conditional PDF)



Two class problem

- Now the decision rule may be
- $P(\omega1/x) > P(\omega2/x) \Rightarrow \omega1$, in favor of class $\omega1$
- $P(\omega1/x) < P(\omega2/x) \Rightarrow \omega2$, in favor of class $\omega2$
- A more logical will be if this $P(\omega1/x)$ and $P(\omega2/x)$ can be combined with **a priori probability** $P(\omega1)$ and $P(\omega2)$





Decision After Observation

Known

Unknown

Prior probability

$P(\omega_j) \ (1 \leq j \leq c)$

Class-conditiona l pdf

$p(x|\omega_j) \ (1 \leq j \leq c)$

Observation for test example

x^* (*e.g.: fish lightness*)

Bayes Formula

Posterior probability

$P(\omega_j|x^*) \ (1 \leq j \leq c)$

The quantity which we want to use in decision naturally (by exploiting observation information)

Convert the prior probability $P(\omega_j)$ to the posterior probability $P(\omega_j|x^*)$

Bayes Formula Revisited

From the preliminary probability theory,

Joint probability density function (Joint PDF)

$$p(\omega, x)$$

Law of total probability

$$p(\omega, x) = P(\omega|x) \cdot p(x)$$
$$p(\omega, x) = P(\omega) \cdot p(x|\omega)$$

$$P(\omega|x) \cdot p(x) = P(\omega) \cdot p(x|\omega)$$

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

Bayes Formula Revisited (Cont.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad (1 \leq j \leq c) \quad \text{(Bayes Formula)}$$

Bayes Decision Rule

$$\boxed{\text{if } P(\omega_j|x) > P(\omega_i|x),\ \forall i \neq j \implies \text{Decide } \omega_j}$$

- $P(\omega_j)$ and $p(x|\omega_j)$ are assumed to be known
- $p(x)$ is irrelevant for Bayesian decision (serving as a normalization factor, not related to any state of nature)

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

Bayes Formula Revisited (Cont.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)}$$ **= P(x/ω1). P(ω1) > P(x/ω2). P(ω2) => ω1**

Special Case I: Equal prior probability

$$P(\omega_1) = P(\omega_2) = \cdots = P(\omega_c) = \frac{1}{c}$$

→ Depends on the likelihood $P(x|\omega_j)$

Special Case II: Equal likelihood

$$p(x|\omega_1) = p(x|\omega_2) = \cdots = p(x|\omega_c)$$

→ Depends on a priori probability $P(\omega_j)$

Special Case III: otherwise, prior probability and likelihood function together in Bayesian decision process

Bayes Theorem

Bayes theorem $P(H|X) = \dfrac{P(H)P(X|H)}{P(X)}$

X: the observed sample (also called **evidence**; *e.g.: the length of a fish*)

H: the hypothesis (*e.g. the fish belongs to the "salmon" category*)

P(*H*): the **prior probability** that *H* holds (*e.g. the probability of catching a salmon*)

P(*X*|*H*): the **likelihood** of observing *X* given that *H* holds (e.g. *the probability of observing a 3-inch length fish which is salmon*)

P(*X*): the **evidence probability** that *X* is observed (*e.g. the probability of observing a fish with 3-inch length*)

P(*H*|*X*): the **posterior probability** that *H* holds given *X* (e.g. *the probability of X being salmon given its length is 3-inch*)

Thomas Bayes (1702-1761)

Bayes Classifier

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad (1 \le j \le c) \quad \text{(Bayes Formula)}$$

$$\boxed{\text{if } P(\omega_j|x) > P(\omega_i|x),\ \forall i \neq j \implies \text{Decide } \omega_j}$$

Example 1

- Two boxes B1 and B2 contain 100 and 200 light bulbs respectively. The first box (B1) has 15 defective bulbs and the second has 5 defective bulbs
  - a) Suppose a box is selected at random and one bulb is picked out. What is the probability it is defective?
  - b) Suppose the bulb we tested was defective what is the probability it came from box 1?

Example 1- cont.

	Defective	Not defective
B1	15	85
B2	5	195

Example 1- cont.

- Since the box is selected at random they are equally likely
- P(B1)= P(B2)= ½ = 0.5
- P(D/B1)= 15/100= 0.15
- P(D/B2)=5/100=0.25
- P(D)= P(D/B1). P(B1) + P(D/B2). P(B2)
- P(D)=(0.15x0.5)+(0.025x0.5)=0.0875
- Thus there is about 9% probability a bulb is defective.

Example 1- cont.

- $P(B1/D) = P(D/B1).\ P(B1)\ /\ P(D)$
- $P(B1/D) = 0.15x0.5\ /\ 0.0875 = 0.8571$
- $0.8571 > 0.5$
- Recall box 1 has three times more defective bulbs compared to box 2

Example 2

- **Problem statement**
- A new medical test is used to detect whether a patient has a certain cancer or not, whose test result is either + (*positive*) or - (*negative*)
- For patient with this cancer, the probability of returning *positive* test result is 0.98
- For patient without this cancer, the probability of returning *negative*
- test result is 0.97
- The probability for any person to have this cancer is 0.008
- **Question**
- If *positive* test result is returned for some person, does he/she have this kind of canceror not?

Example 2- cont.

	Positive (+)	Negative (-)
Class ω_1 Cancer	0.98	
Class ω_2 No Cancer		0.97

Question

If *positive* test result is returned for some person,does he/she have this kind of cancer or not?

Idea:

$P(\omega_i /+) = ? \Rightarrow P(\omega_1 /+) = ?\ ,\ P(\omega_2 /+) = ?$

If $P(\omega_1 /+) > P(\omega_2 /+) \Rightarrow \omega_1$

If $P(\omega_1 /+) < P(\omega_2 /+) \Rightarrow \omega_2$

Example 2 (Cont.)

ω_1 : cancer $\qquad \omega_2$: no cancer $\qquad x \in \{+, -\}$

$P(\omega_1) = 0.008 \qquad P(\omega_2) = 1 - P(\omega_1) = 0.992$

$P(+ \mid \omega_1) = 0.98 \qquad P(- \mid \omega_1) = 1 - P(+ \mid \omega_1) = 0.02$

$P(- \mid \omega_2) = 0.97 \qquad P(+ \mid \omega_2) = 1 - P(- \mid \omega_2) = 0.03$

$$P(\omega_1 \mid +) = \frac{P(\omega_1)P(+ \mid \omega_1)}{P(+)} = \frac{P(\omega_1)P(+ \mid \omega_1)}{P(\omega_1)P(+ \mid \omega_1) + P(\omega_2)P(+ \mid \omega_2)}$$

$$= \frac{0.008 \times 0.98}{0.008 \times 0.98 + 0.992 \times 0.03} = 0.2085$$

$$P(\omega_2 \mid +) = 1 - P(\omega_1|+) = 0.7915$$

$P(\omega_2 \mid +) > P(\omega_1|+)$

No cancer!

Error in Bayes Classifier

Probability of Error

Bayes Risk Classifier

Bayes Minimum Error Rate Classifier

Bayes Formula Revisited (Cont.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad (1 \leq j \leq c)$$ (Bayes Formula)

Bayes Decision Rule

$$\boxed{\text{if } P(\omega_j|x) > P(\omega_i|x),\ \forall i \neq j \implies \text{Decide } \omega_j}$$

- $P(\omega_j)$ and $p(x|\omega_j)$ are assumed to be known
- $p(x)$ is irrelevant for Bayesian decision (serving as a normalization factor, not related to any state of nature)

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

Bayes Formula Revisited (Cont.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)}$$ **= P(x/ω1). P(ω1) > P(x/ω2). P(ω2) => ω1**

Special Case I: Equal prior probability

$$P(\omega_1) = P(\omega_2) = \cdots = P(\omega_c) = \frac{1}{c}$$

Depends on the likelihood $P(x|\omega_j)$

Special Case II: Equal likelihood

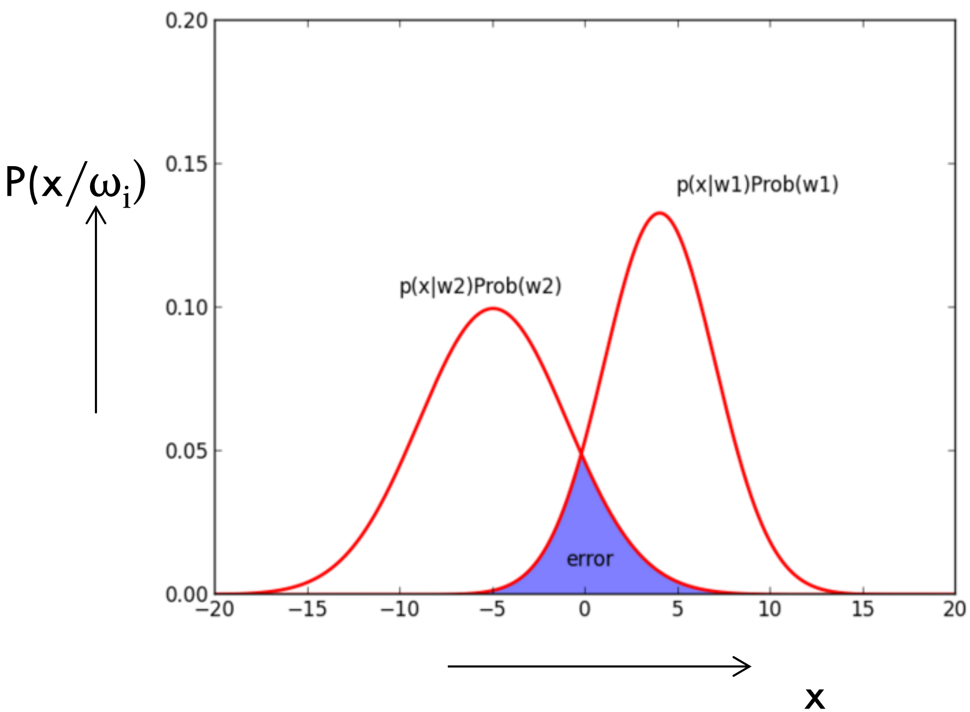
$$p(x|\omega_1) = p(x|\omega_2) = \cdots = p(x|\omega_c)$$

Depends on a priori probability $P(\omega_j)$

Special Case III: otherwise, prior probability and likelihood function together in Bayesian decision process

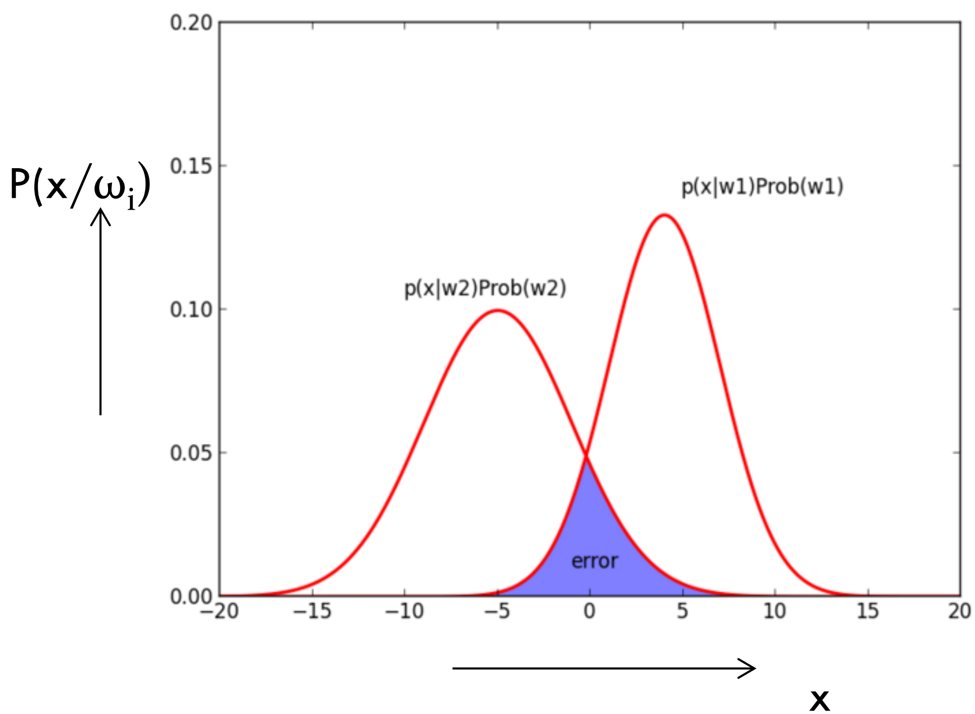
Error: Probability of error

- $P(x/\omega 1).\ P(\omega 1) > P(x/\omega 2).\ P(\omega 2) \Rightarrow \omega 1$



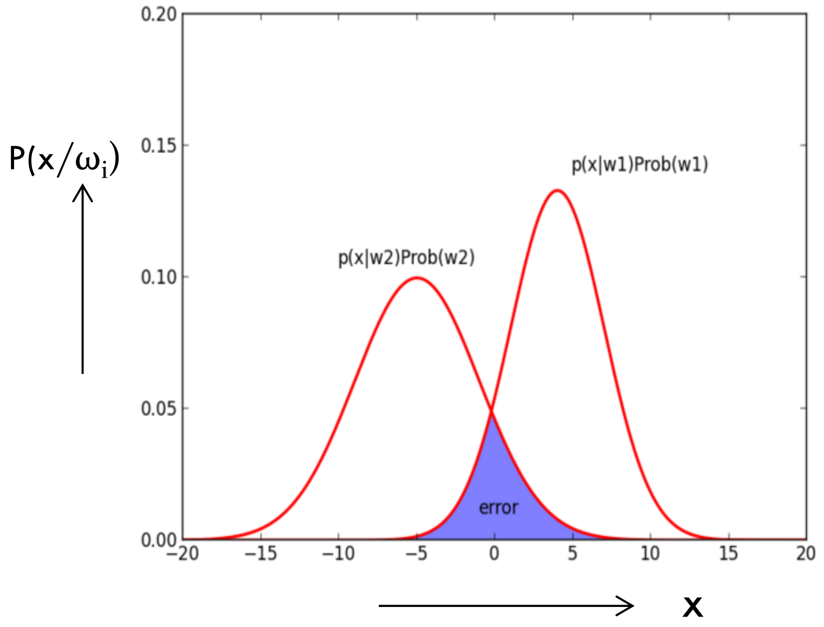
Error: Probability of error

- $P(x/\omega_1).\ P(\omega_1) > P(x/\omega_2).\ P(\omega_2) \Rightarrow \omega_1$
- There is a finite probability of error
- If we decide in favor of $\omega_1 \Rightarrow P(\omega_2/x)$
- If we decide in favor of $\omega_2 \Rightarrow P(\omega_1/x)$



Error: Probability of error

- $P(x/\omega_1).\ P(\omega_1) > P(x/\omega_2).\ P(\omega_2) \Rightarrow \omega_1$
- There is a finite probability of error
- If we decide in favor of $\omega_1 \Rightarrow P(\omega_2/x)$
- If we decide in favor of $\omega_2 \Rightarrow P(\omega_1/x)$
- Given, the situation like this, the total error
- $P(error \mid x) = \min[P(\omega_1 \mid x), P(\omega_2 \mid x)]$



Error: Probability of error

Bayes Decision Rule (In case of two classes)

$$\text{if } P(\omega_1|x) > P(\omega_2|x), \text{ Decide } \omega_1; \text{ Otherwise } \omega_2$$

Whenever we observe a particular *x*, the probability of error is:

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

Under Bayes decision rule, we have

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

For every *x*, we ensure that *P*(*error* | *x*) is as small as possible ⇒ The average probability of error over all possible *x* must be as small as possible

Is Bayes Decision Rule Optimal?

- For every x, Bayes classifier ensures that P(error/x) is as small as possible
- The average probability of error over all possible *x* must also be as small as possible
- The Bayes rule minimizes the expected error rate
- Minimizing the expected error rate is a pretty reasonable goal
- However, it is not always the best thing to do.

Is Bayes Decision Rule Optimal?

- Consider the situation
- You are designing a pedestrian detection algorithm for an autonomous navigation system
- Your algorithm must decide whether there is a pedestrian crossing the street
- There could be two possible types of error
  - False positive: There is no pedestrian, but the system thinks there is a pedestrian
  - Miss (false negative): There is a pedestrian, but the system thinks there is not

Is Bayes Decision Rule Optimal?

- In this situation, should we give equal weight to these two types of error?
- Solution: To deal with these kind of problem instead of minimizing the error rate, we minimize something called the risk
- First, we define the loss matrix L, which quantifies the cost of making each type of error

Bayes Risk

- Element λ_{ij} of the loss matrix specifies the cost of taking action α_i when the true state of nature is ω_j
- Typically, we set $\lambda_{ii} = 0$ for all i
- Thus a typical loss matrix for m=2, would have the form
- L= $\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$

Bayes Decision Rule –The General Case

- By allowing to use more than one feature

  $x \in \mathbf{R} \implies \mathbf{x} \in \mathbf{R}^d$ (*d*-dimensional Euclidean space)

- By allowing more than two states of nature

  $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ (finite set of *c* states of nature)

- By allowing actions other than merely deciding the state of nature

  $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ (finite set of *a* possibleactions)

$\text{Note : } c \neq a$

Bayes Decision Rule –The General Case (Cont.)

- By introducing a loss function more general than the probability of error

$$\lambda(\alpha_i \mid \omega_j)$$

$\lambda(\alpha_i \mid \omega_j)$ => the loss incurred for taking action α_i when the state of nature is ω_j

For ease of reference, usually written as:

λ_{ij}

A simple loss function

Action / Class	$\alpha_1 =$ "Recipe A"	$\alpha_2 =$ "Recipe B"	$\alpha_3 =$ "No Recipe"
$\omega_1 =$ "cancer"	**5**	**50**	**10,000**
$\omega_2 =$ "no cancer"	**60**	**3**	**0**

Bayes Decision Rule –The General Case (Cont.)

- The problem
  - Given a particular **x**, we have to decide which action to take

$\alpha_i \; (1 \leq i \leq a)$

- We need to know the *loss* of taking each action

true state of nature is ω_j

the action being taken is α_i

incur the loss $\lambda(\alpha_i \mid \omega_j)$

However, the true state of nature is uncertain

Expected (average) loss

Bayes Decision Rule –The General Case

Average by enumerating over all possible states of nature!

Expected loss

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

The incurred loss of taking action α_i in case of true state of nature being ω_j

The probability of ω_j being the true state of nature, given the feature vector x

The expected loss is also named as ***(conditional) risk or risk function***

Bayes Decision Rule –The General Case

- Now, we have to choose that action α_i for which the risk is minimum.
- It is also called as Bayes risk and it is the best performance that can be achieved.

Bayes Decision Rule –The General Case (Cont.)

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} \quad (\text{overall risk})$$

For every **x**, we ensure that the conditional risk $R(\alpha(\mathbf{x}) \mid \mathbf{x})$ is as small as possible

The overall risk over all possible **x** must be as small as possible

Bayes decision rule (*General case*)

$$\alpha(\mathbf{x}) = \arg\min_{\alpha_i \in \mathcal{A}} R(\alpha_i \mid \mathbf{x})$$
$$= \arg\min_{\alpha_i \in \mathcal{A}} \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x})$$

- The resulting overall risk is called the ***Bayes risk*** (denoted as R^*)
- The best performance achievable given $p(\mathbf{x})$ and loss function

Two-Category Classification

Special case

$$\begin{matrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \end{matrix}$$

- $\Omega = \{\omega_1, \omega_2\}$ (two states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2\}$ $(\alpha_1 = \text{decide } \omega_1;\ \alpha_2 = \text{decide } \omega_2)$

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$: the loss incurred for deciding ω_i when the true state of nature is ω_j

The conditional risk:

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

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

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

yes → decide ω_1

no → decide ω_2

Two-Category Classification (Cont.)

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

by definition

$$\lambda_{11} \cdot P(\omega_1 \mid \mathbf{x}) + \lambda_{12} \cdot P(\omega_2 \mid \mathbf{x}) < \lambda_{21} \cdot P(\omega_1 \mid \mathbf{x}) + \lambda_{22} \cdot P(\omega_2 \mid \mathbf{x})$$

by re-arrangement

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

by Bayes theorem

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

$\lambda_{21} - \lambda_{11} > 0$

the loss for being error is ordinarily greater than the loss for being correct

$$\underbrace{\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}}_{\text{likelihood ratio}} > \underbrace{\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}}_{\text{constant } \theta \text{ independent of } \mathbf{x}}$$

Bayes Minimum Risk- Numerical Example 1

Suppose we have:

Class \ Action	$\alpha_1 =$ "Recipe A"	$\alpha_2 =$ "Recipe B"	$\alpha_3 =$ "No Recipe"
$\omega_1 =$ "cancer"	**5**	**50**	**10,000**
$\omega_2 =$ "no cancer"	**60**	**3**	**0**

For a particular $\mathbf{x}$:

$$P(\omega_1 \mid \mathbf{x}) = 0.01$$

$$P(\omega_2 \mid \mathbf{x}) = 0.99$$

Bayes Minimum Risk- Numerical Example 1

calculate the risk involved for various action given in the table

Action / Class	$\alpha_1 =$ "Recipe A"	$\alpha_2 =$ "Recipe B"	$\alpha_3 =$ "No Recipe"
$\omega_1 =$ "cancer"	**5**	**50**	**10,000**
$\omega_2 =$ "no cancer"	**60**	**3**	**0**

For a particular $\mathbf{x}$:

$$P(\omega_1 \mid \mathbf{x}) = 0.01$$

$$P(\omega_2 \mid \mathbf{x}) = 0.99$$

$$R(\alpha_1 \mid \mathbf{x}) = \sum_{j=1}^{2} \lambda(\alpha_1 \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x})$$

$$= \lambda(\alpha_1 \mid \omega_1) \cdot P(\omega_1 \mid \mathbf{x}) + \lambda(\alpha_1 \mid \omega_2) \cdot P(\omega_2 \mid \mathbf{x})$$

$$= 5 \times 0.01 + 60 \times 0.99 = 59.45$$

Similarly, we can get: $R(\alpha_2 \mid \mathbf{x}) = 3.47 \quad R(\alpha_3 \mid \mathbf{x}) = 100$

Bayes Minimum Risk- Numerical Example 2

Spam Filtering: Suppose we have

Class \ Action	$\alpha_1 =$ Keep the mail	$\alpha_2 =$ Delete as Spam
ω1=normal mail	**0**	**3**
ω2=spam mail	**1**	**0**

For a particular **x**:

P(x/ω1)=0.35

P(x/ω2)=0.65

P(ω1)=0.4

P(ω2)=0.6

Bayes Minimum Risk- Numerical Example 2

Spam Filtering: Suppose we have

Action / Class	$\alpha_1 =$ Keep the mail	$\alpha_2 =$ Delete as Spam
ω1=normal mail	**0**	**3**
ω2=spam mail	**1**	**0**

For a particular **x**:

P(x/ω1)=0.35

P(x/ω2)=0.65

P(ω1)=0.4

P(ω2)=0.6

$R(\alpha_1 \mid \mathbf{x})$= 0.736

$R(\alpha_2 \mid \mathbf{x})$= 0.792

Since $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$, we decide take action 1 and decide class 1. Keep the mail

Minimum-Error-Rate Classification

Classification setting

- $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ (c possible states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_c\}$ ($\alpha_i = \text{decide } \omega_i,\ 1 \leq i \leq c$)

Zero-one (symmetrical) loss function

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

- Assign no loss (i.e. 0) to a correct decision by taking action
- Assign a unit loss (i.e. 1) to any incorrect decision (**equal cost**)

Minimum-Error-Rate Classification (Cont.)

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

$$= \sum_{j \neq i} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \; + \; \lambda(\alpha_i \mid \omega_i) \cdot P(\omega_i \mid \mathbf{x})$$

$$= \sum_{j \neq i} P(\omega_j \mid \mathbf{x})$$

$$= 1 - P(\omega_i \mid \mathbf{x})$$

error rate

$P(\omega_i/x)$ the probability that action α_i (decide ω_i) is correct

Minimum error rate

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

Discriminant Functions

Discriminant function for Minimum Risk,

Minimum Error Rate classifier (Bayes Classifier)

Discriminant Function-Multi category case

Classification

Pattern ➔ Category

actions ⟷ **decide categories**

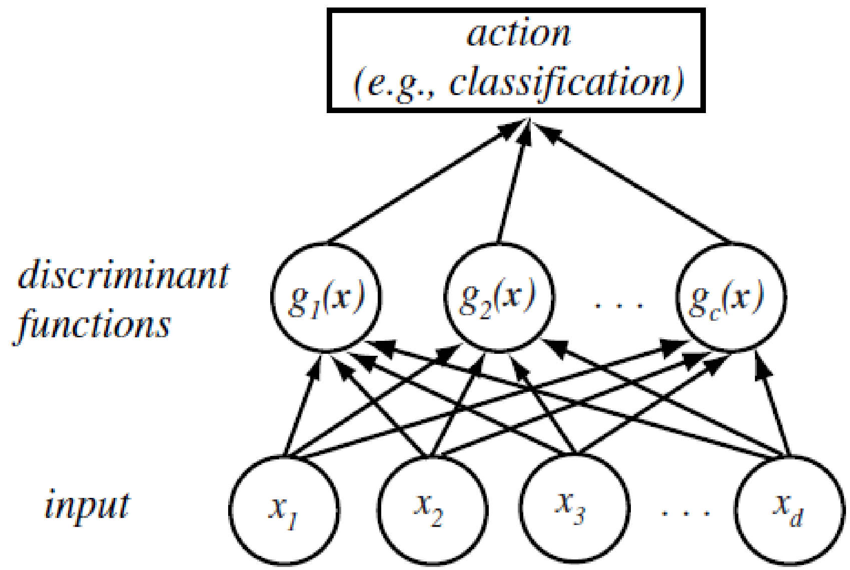
Discriminant functions

$$g_i : \mathbf{R}^d \to \mathbf{R} \qquad (1 \le i \le c)$$

- *Useful way to represent classifiers*
- *One function per category*

Decide ω_i

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

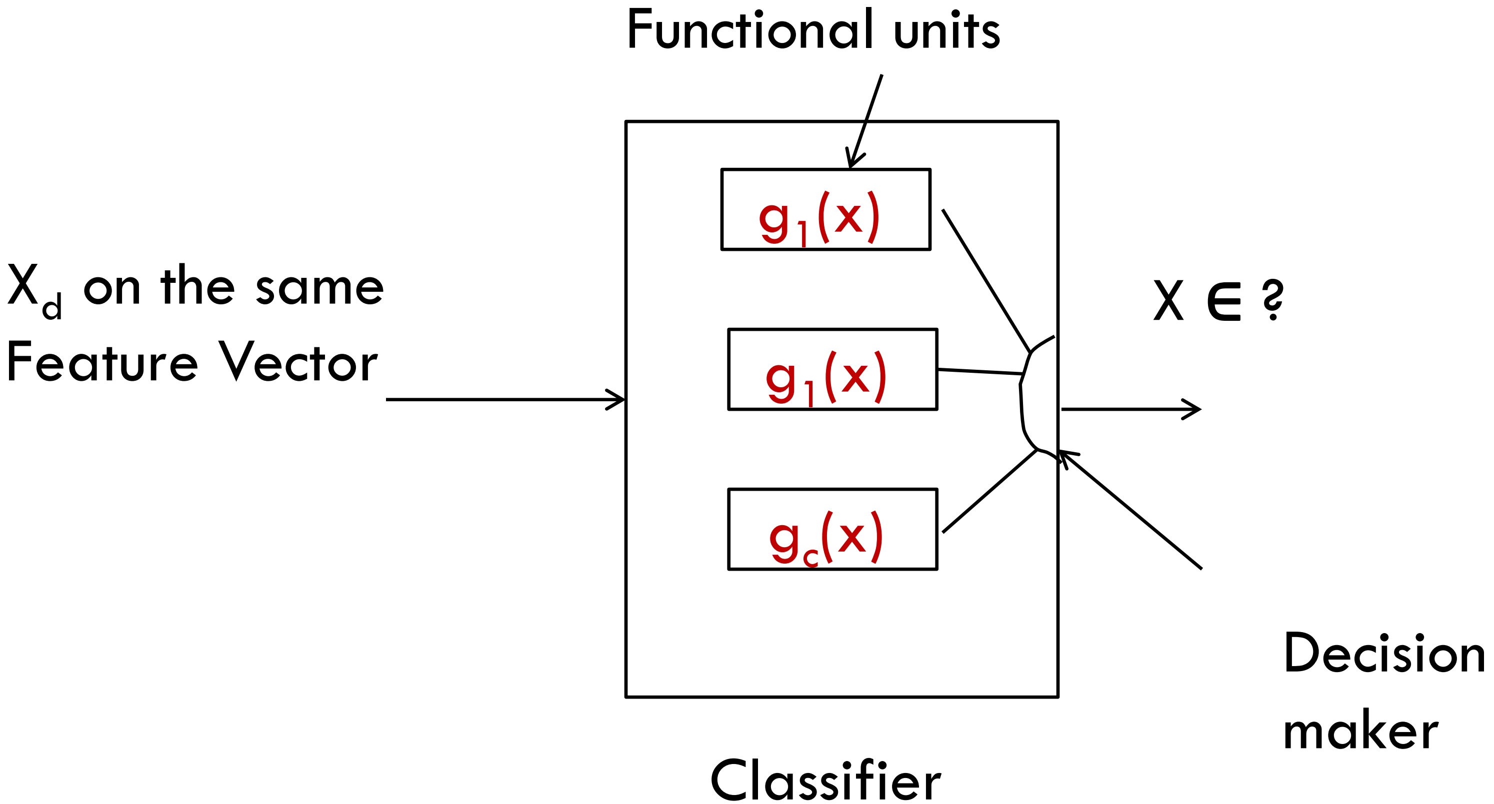


Discriminant Function

- The classifier is viewed as a network or machine that computes c discriminant function
- Select the category corresponding to the maximum discriminant

Discriminant Function

- A network representation of a classifier is shown below



Discriminant Function

- The nature of discriminant classes
- ω_1, ω_2,...., ω_c => c number of classes
- $g_i(x) > g_j(x)$ for all $i \neq j$ => $x \in \omega_i$

Discriminant function under different conditions

Minimum risk: $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x}) \quad (1 \leq i \leq c)$

Minimum-error-rate: $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \quad (1 \leq i \leq c)$

- Hence, the feature vector x can assign to the class which has maximum $g_i(x)$.
- But the choice of discriminant function $g_i(x)$ is not unique, more generally, if we replace every $g_i(x)$ by $f(g_i(x))$, where f(.) is a monotonically increasing function, the resulting classification is unchanged.

Discriminant Function (Cont.)

This observation can lead to significant analytical and computational simplifications

Various discriminant functions $\Rightarrow$ **Identical classification results**

$f(\cdot)$ is a *monotonically increasing function*

$$f(g_i(\mathbf{x})) \iff g_i(\mathbf{x})$$ (*i.e. equivalent in decision*)

e.g.:

$f(x) = k \cdot x \;(k > 0)$ $\Rightarrow$ $f(g_i(\mathbf{x})) = k \cdot g_i(\mathbf{x}) \;(1 \leq i \leq c)$

$f(x) = \ln x$ $\Rightarrow$ $f(g_i(\mathbf{x})) = \ln g_i(\mathbf{x}) \;(1 \leq i \leq c)$

Discriminant Function (Cont.)

Decision region

c discriminant functions ⇒ *c* decision regions

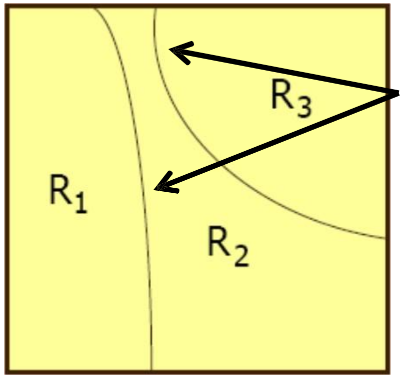
$g_i(\cdot) \ (1 \leq i \leq c)$ ⇒ $\mathcal{R}_i \subset \mathbf{R}^d \ (1 \leq i \leq c)$

$$\mathcal{R}_i = \{\mathbf{x} \mid \mathbf{x} \in \mathbf{R}^d : g_i(\mathbf{x}) > g_j(\mathbf{x}) \ \ \forall j \neq i\}$$

where $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset \ (i \neq j)$ and $\bigcup_{i=1}^{c} \mathcal{R}_i = \mathbf{R}^d$

Decision boundary

surface in feature space where ties occur among several largest discriminant functions



decision boundary

Dichotomizer: The two category case

- Dichotomizer: A Classifier that places a pattern in one of any two categories has a special name called a dichotomizer
- $g_1(x)= g_2(x)$
- $g(x)= g_1(x)- g_2(x)$
- Thus a dichotomizer can be viewed as a machine that computes a single discriminant function g(x), and classifies x according to algebraic sign of the result

Discriminant Function (Cont.)

- **Minimum error rate classifier**
- $g_i(x) = P(\omega_i / x)$
- $g_i(x) = P(x/\omega_i).\ P(\omega_i) / P(x)$
- $g_i(x) = P(x/\omega_i).\ P(\omega_i)$
- $f(g_i(x)) = \ln P(x/\omega_i) + \ln P(\omega_i)$

Discriminant Function (Cont.)

- Minimum error rate classifier: $g_i(x) = P(\omega_i/x)$
- **two category case**
- $g_1(x) = P(\omega_i/x)$; $g_2(x) = P(\omega_i/x)$
- $g(x) \equiv g_1(x) - g_2(x) = 0$
- $g(x) = P(\omega_1/x) - P(\omega_2/x) = 0$
- $g(x) = P(x/\omega_1) . P(\omega_1) - P(x/\omega_2) . P(\omega_2)$
- $g(x) = \ln (P(x/\omega_1) . P(\omega_1)) - \ln(P(x/\omega_2) .P(\omega_2))$
- $g(x) = \ln P(x/\omega_1) + \ln P(\omega_1) - \ln P(x/\omega_2) - \ln P(\omega_2)$
- $g(x) = \ln \frac{P(X/\omega_1)}{P(X/\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$

Discriminant function for Bayes Classifier

- $g_1(x)= P(\omega_i /x)$
- $g_i(x)= \ln P(\omega_i /x)$
- $g(x)= \ln P(x/\omega_i) + \ln P(\omega_i)$
- $P(x/\omega_i)$= class conditional PDF
- $P(\omega_i)$= PDF (a priori probability)
- Note: The structure of baye's classifier is determined by the conditional densities $P(x/\omega_i)$ as well as prior probabilities $P(\omega_i)$

Discriminant function for Bayes Classifier

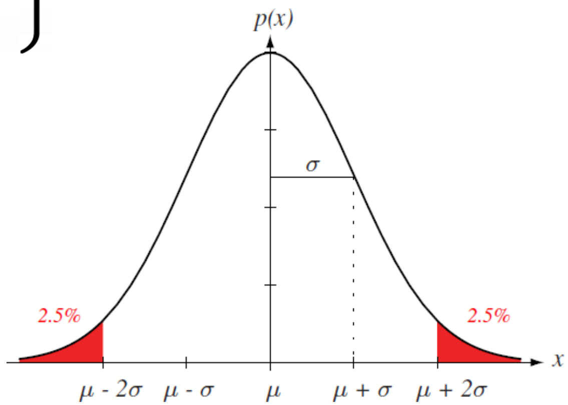
- We can have various types of probability density like i) normal density ii) poison iii) laplacian iv) exponential and so on
- Out of these, the most common Probability Density Function (PDF) which is in use is Normal/ Gaussian density function.

The normal density

The normal density

- Univariate density: for a single variable, the normal /Gaussian density is

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



- P(x) = probability density function
- μ = mean = E(x)
- σ = standard deviation = E $[(x-\mu)^2]$
- This particular PDF is specified by two parameters μ, σ
- In short ≈ $N(\mu, \sigma^2)$

The normal density

- Multivariate Probability Density function
- Here X be a feature vector, $X = [x_1, x_2, \ldots, x_d]^T$

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

- $\mu = E[X]$; x is d-dimensional vector
  - $\mu = [\mu_1, \mu_2, \ldots, \mu_d]^T$
- Σ = covariance matrix, dxd matrix
  - $\Sigma = E[(X-\mu)(X-\mu)^t]$
- In short ≈ $N(\mu, \Sigma)$

The normal density

- What is expected value of individual component?
- <u>Expected value of the i^{th} component</u>
- $\mu_i = E[x_i]$
- $i \neq j;\ \sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$
- $i = j;\ \sigma_{ii} = E[(x_i - \mu_i)(x_i - \mu_i)] = E[(x_i - \mu_i)^2]$

The normal density

- Bivariate Probability Density function (two variables)
- Here X is of the form $[x_{1,} x_2]^T$
- Number of dimension d=2
- Assume x_1 and x_2 are statistically independent, and hence σ_{12} and σ_{21} are 0
- Mean $\mu= [\mu_1, \mu_2]^T$
- $\Sigma=\begin{bmatrix} \sigma_1{}^2 & 0 \\ 0 & \sigma_2{}^2 \end{bmatrix}$

The normal density- bivariate case

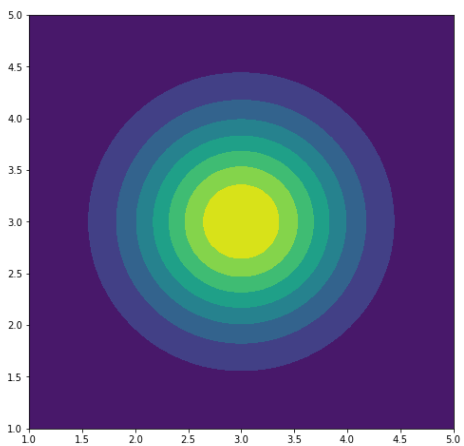
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\,\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

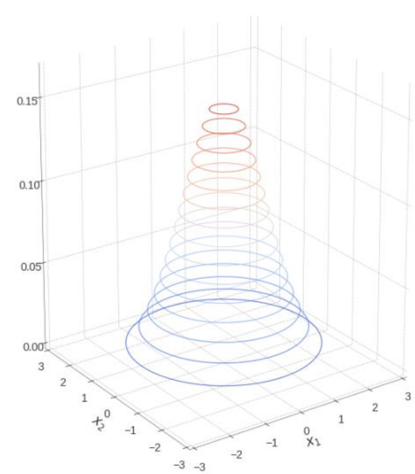
- The above multivariate normal density can be simplified to

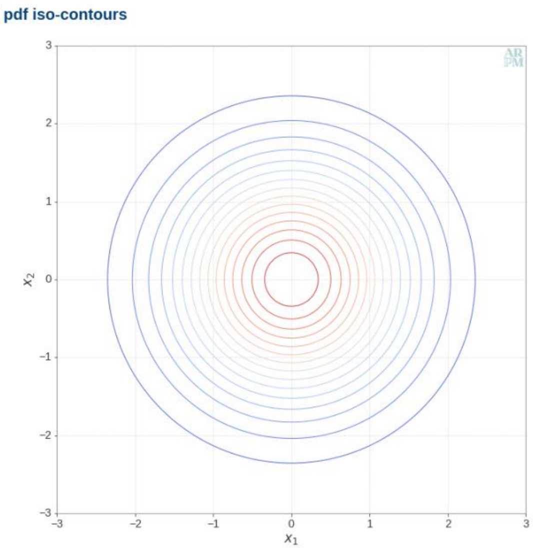
$$P(X) = \frac{1}{2\pi|\Sigma|^{1/2}} exp[\frac{-1}{2}\{(\frac{x_1-\mu_1}{\sigma_1})^2 + \frac{x_2-\mu_2}{\sigma_2})^2\}]$$

The normal density- two dimensional feature space

- **Case 1:**
  - $\sigma_{ij}=0;\ i \neq j$
  - $\sigma_1^2 = \sigma_2^2$; $\sigma_{12} = \sigma_{21} = 0$; $\quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = 5\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- x_1 and x_2 are statistically independent
- Trace the loci of points of constant density for all value of x for which P(x) is constant.
- Those loci of points forms circle

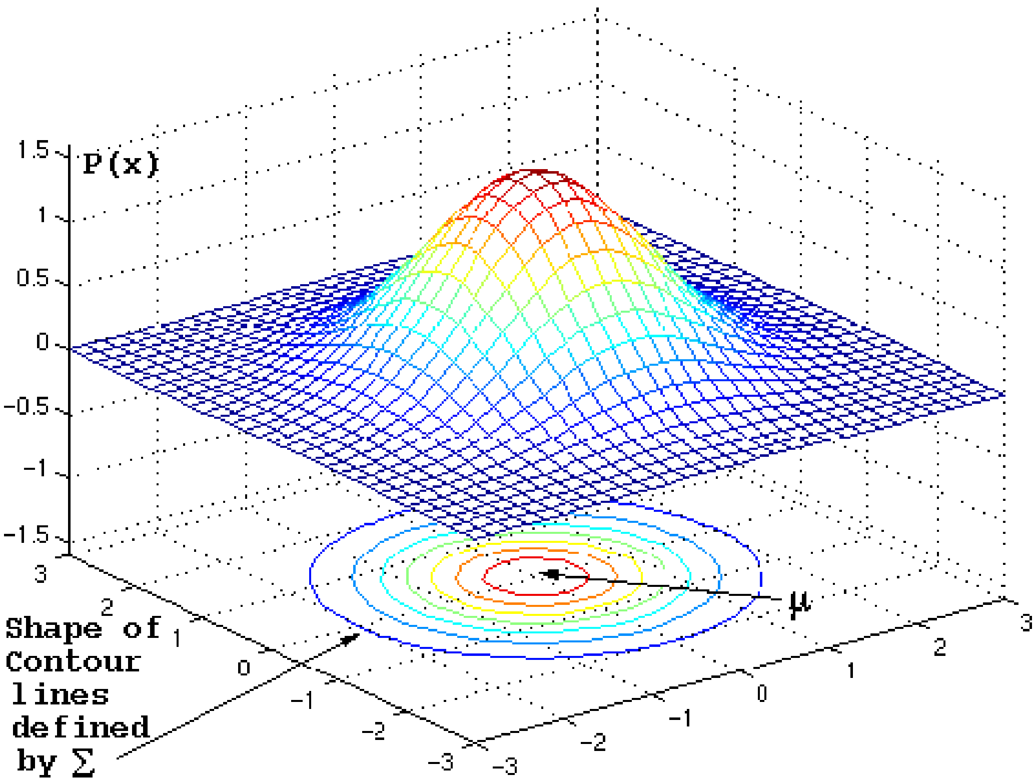






The normal density- two dimensional feature space

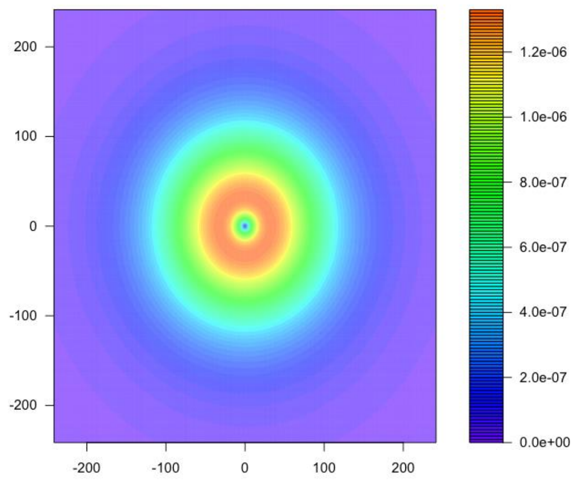
- **Case 1:**



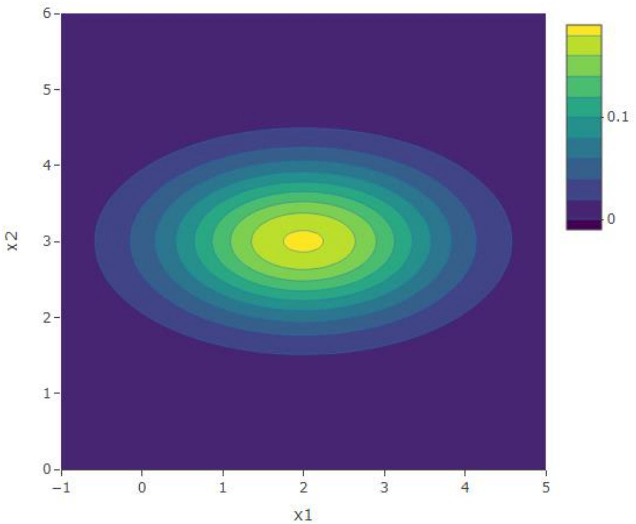
The normal density- two dimensional feature space

- What happens if the variants are different?
- **Case 2:**
  - $\sigma_{ij}=0; i\neq j$
  - $\sigma_1{}^2 \neq \sigma_2{}^2; \sigma_{12}= \sigma_{21} = 0$; $\Sigma=\begin{bmatrix}\sigma_1{}^2 & 0\\ 0 & \sigma_2{}^2\end{bmatrix} = \begin{bmatrix}15 & 0\\ 0 & 5\end{bmatrix}$; $\sigma_1{}^2 > \sigma_2{}^2$
- x_1 and x_2 are not statistically independent. The loci of points forms ellipse

$\sigma_1{}^2 < \sigma_2{}^2$

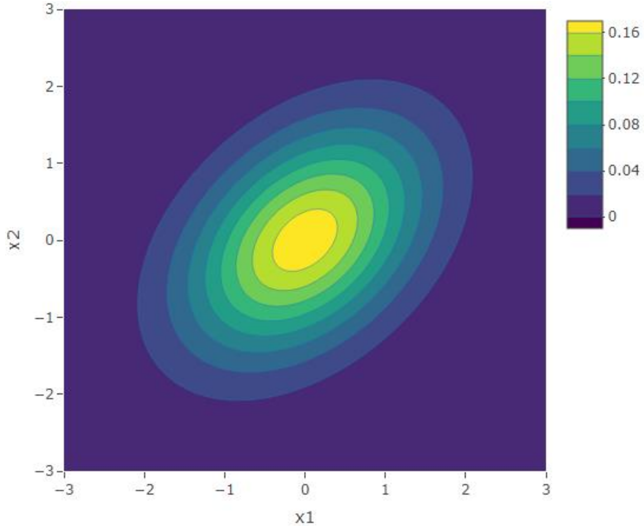


$\sigma_1{}^2 > \sigma_2{}^2$



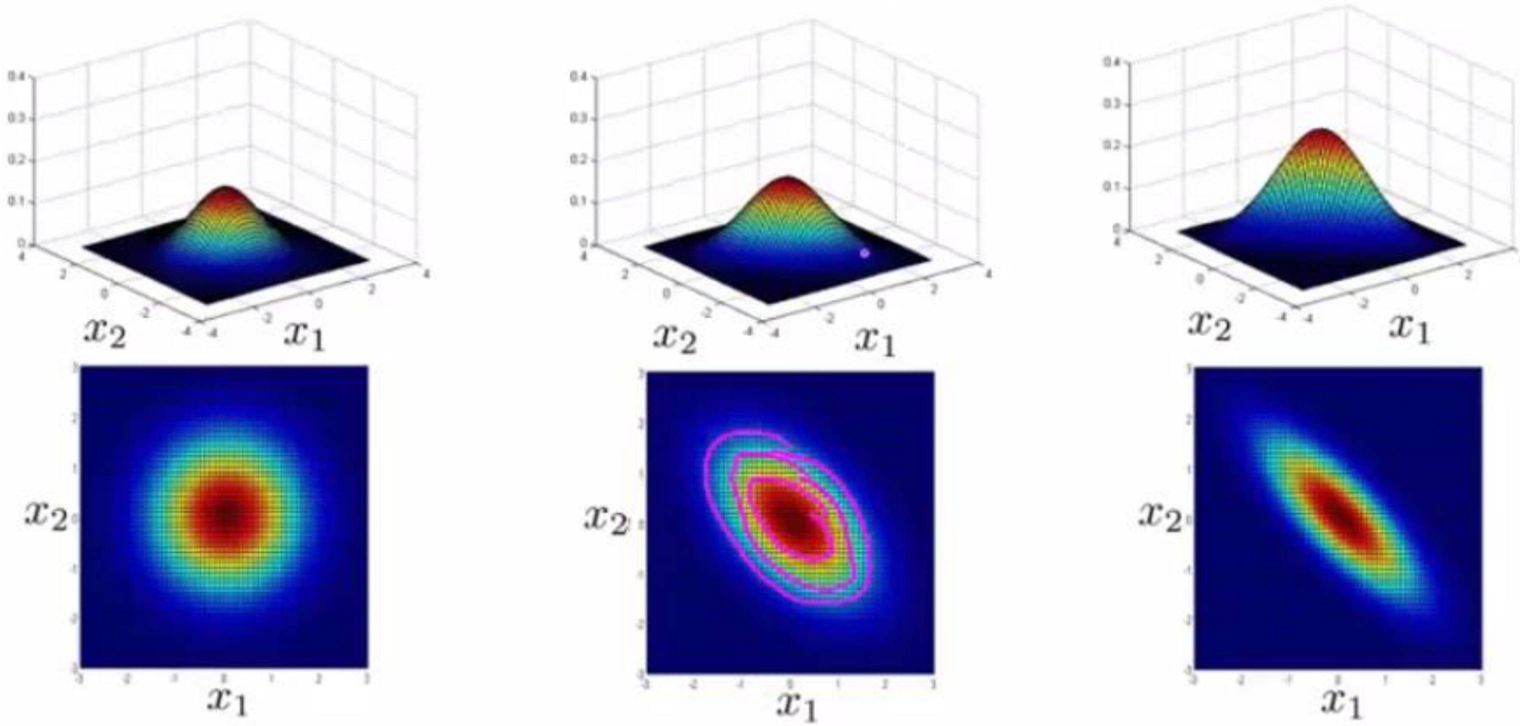
The normal density- two dimensional feature space

- What happens if the variants are different?
- **Case 3:**
  - $\sigma_{ij} \neq 0$; $i \neq j$
  - $\sigma_1^2 \neq \sigma_2^2$; $\sigma_{12} = \sigma_{21} = 0$
- x_1 and x_2 are not statistically independent
- The direction of point distribution is determined by eigenvector of $\sum$
- $\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.3 \\ 0.6 & 0.7 \end{bmatrix}$



Multivariate Gaussian (Normal) examples

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$



Andrew N

Gaussian Density–Multivariate Case

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

Gaussian Density – Multivariate Case

$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\boxed{\mu_i = \mathcal{E}[x_i] \qquad \sigma_{ij} = \sigma_{ji} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]}$$

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

$\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$: *d*-dimensional *column vector*

$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^t$: *d*-dimensional *mean vector*

$$\boldsymbol{\Sigma} = [\sigma_{ij}]_{1 \le i,j \le d} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix}$$

$d \times d$ *covariance matrix*

$|\boldsymbol{\Sigma}|$: determinant

$\boldsymbol{\Sigma}^{-1}$: inverse

Gaussian Density – Multivariate Case

Properties of covariance matrix

Properties of Σ

- symmetric
- Positive semidefinite

$$\boldsymbol{\Sigma} = [\sigma_{ij}]_{1 \leq i,j \leq d} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix}$$

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

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) \cdot \underline{\underline{p(x_i, x_j)}}\, dx_i dx_j$$

$$\sigma_{ii} = \mathrm{Var}[x_i] = \sigma_i^2$$

marginal pdf on a pair of random variables (x_i, x_j)

Gaussian Density– Multivariate Case (Cont.)

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

$(\mathbf{x}-\boldsymbol{\mu})^t : 1 \times d$ matrix

$\boldsymbol{\Sigma}^{-1} : d \times d$ matrix

$(\mathbf{x}-\boldsymbol{\mu}) : d \times 1$ matrix

$\Rightarrow$

$(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

scalar $(1 \times 1$ matrix$)$

$\boldsymbol{\Sigma}$: positive definite $\Rightarrow$ $\boldsymbol{\Sigma}^{-1}$: positive definite

$\Downarrow$

$(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \geq 0$

$\Leftarrow$

$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq 0$

Discriminant Functions for Gaussian Density for Bayes Classifier

Bayes classification: (Minimum error rate classification)

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \quad (1 \le i \le c)$$

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

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

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

Constant, could be ignored

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

Case I: $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$

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

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

Covariance matrix: σ^2 times the identity matrix $\mathbf{I}$

$$\boldsymbol{\Sigma}_i = \sigma^2 \cdot \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{pmatrix} \Rightarrow$$

$$|\boldsymbol{\Sigma}_i| = \sigma^{2d}$$

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

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

$|| \cdot || :$ *Euclidean norm*

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

Case I: $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$ (Cont.)

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

Squared Euclidean distance

the same for all *states of nature*, could be ignored

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

Linear discriminant functions

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

Linear machine or linear equation

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

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

Case II: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

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

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

Covariance matrix: *identical* for all classes

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

$(\mathbf{x} - \boldsymbol{\mu}_i)^t \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}_i)$: squared *Mahalanobis distance*

$\boldsymbol{\Sigma} = \mathbf{I}$ ⇒ reduces to *Euclidean distance*

P. C. Mahalanobis (1893-1972)

Case II: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ (Cont.)

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

the same for all *states of nature*, could be ignored

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

Linear discriminant functions

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$ **Linear machine or linear equation**

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

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

Case III: $\boldsymbol{\Sigma}_i = \textbf{arbitrary}$

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

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

Quadratic discriminant functions

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

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$ *quadratic matrix*

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

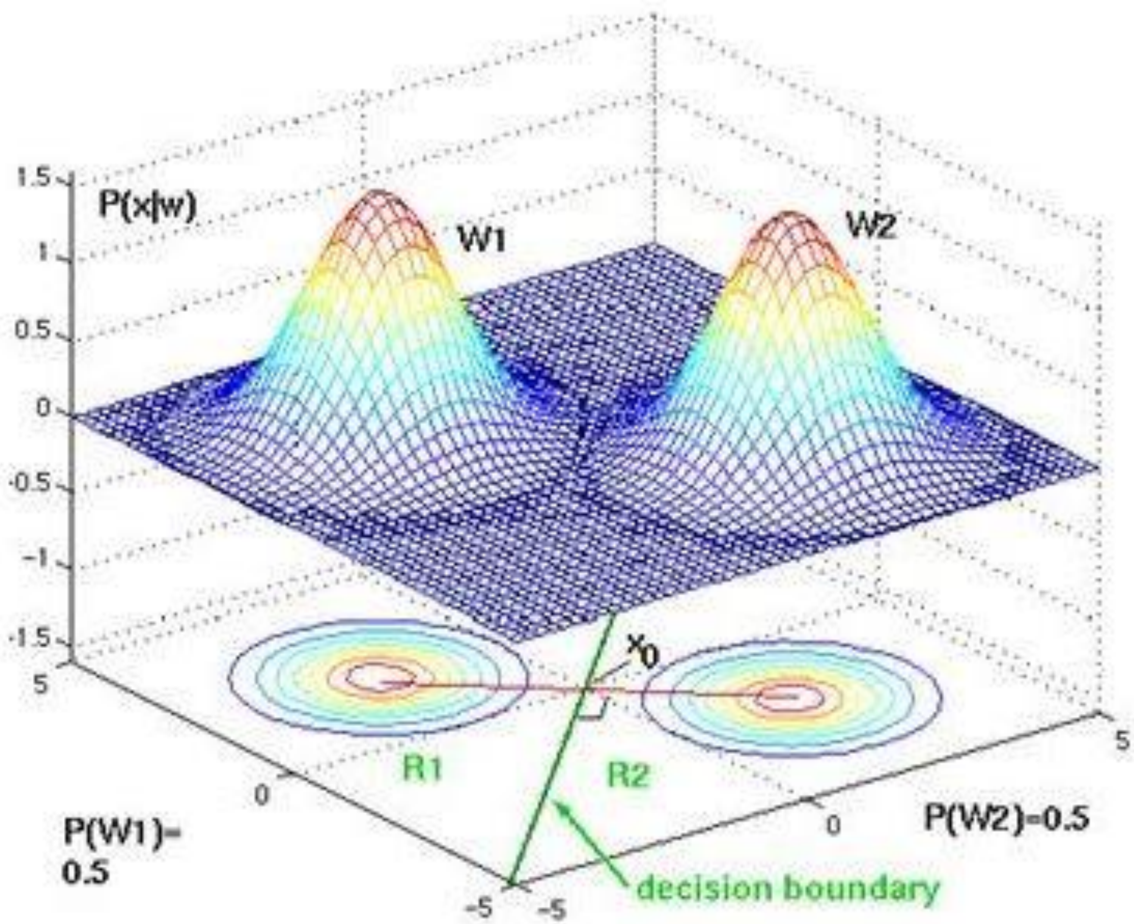
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$ *threshold/bias*

Bayes Decision Boundary for Two Classes

Case 1 & 2 : Linear Separable Cases

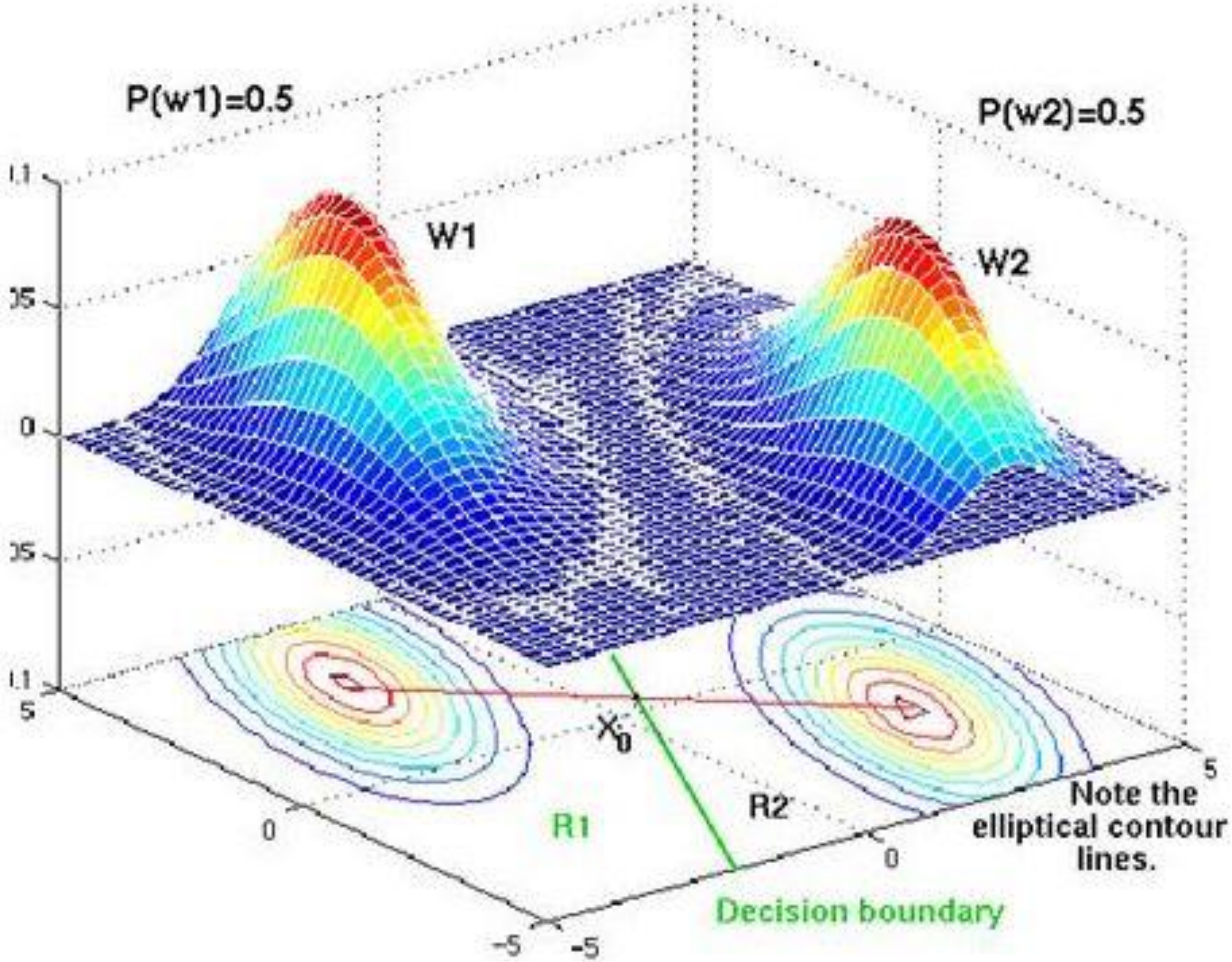
Case 3: Nonlinearly Separable Cases

Case 1

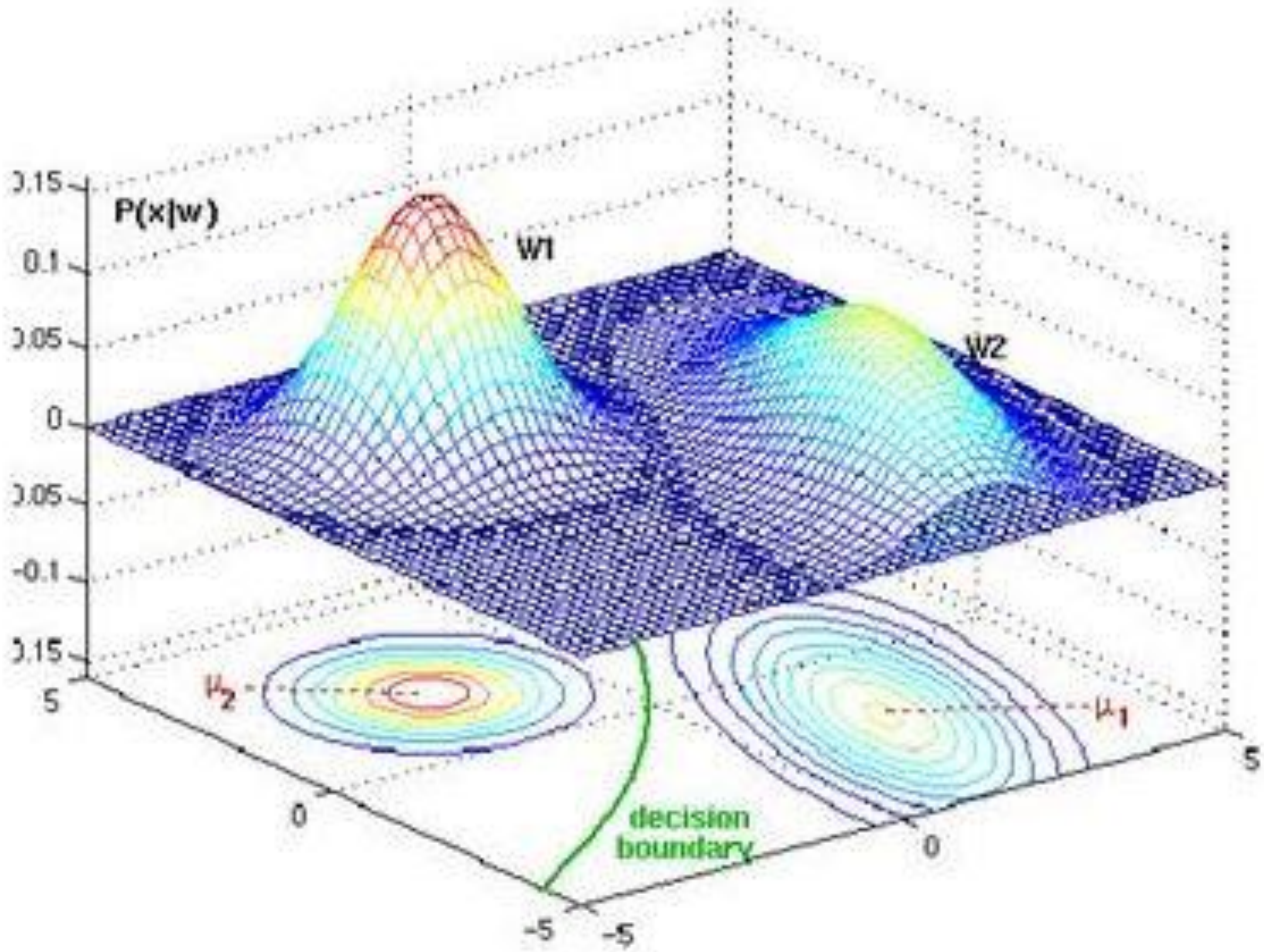


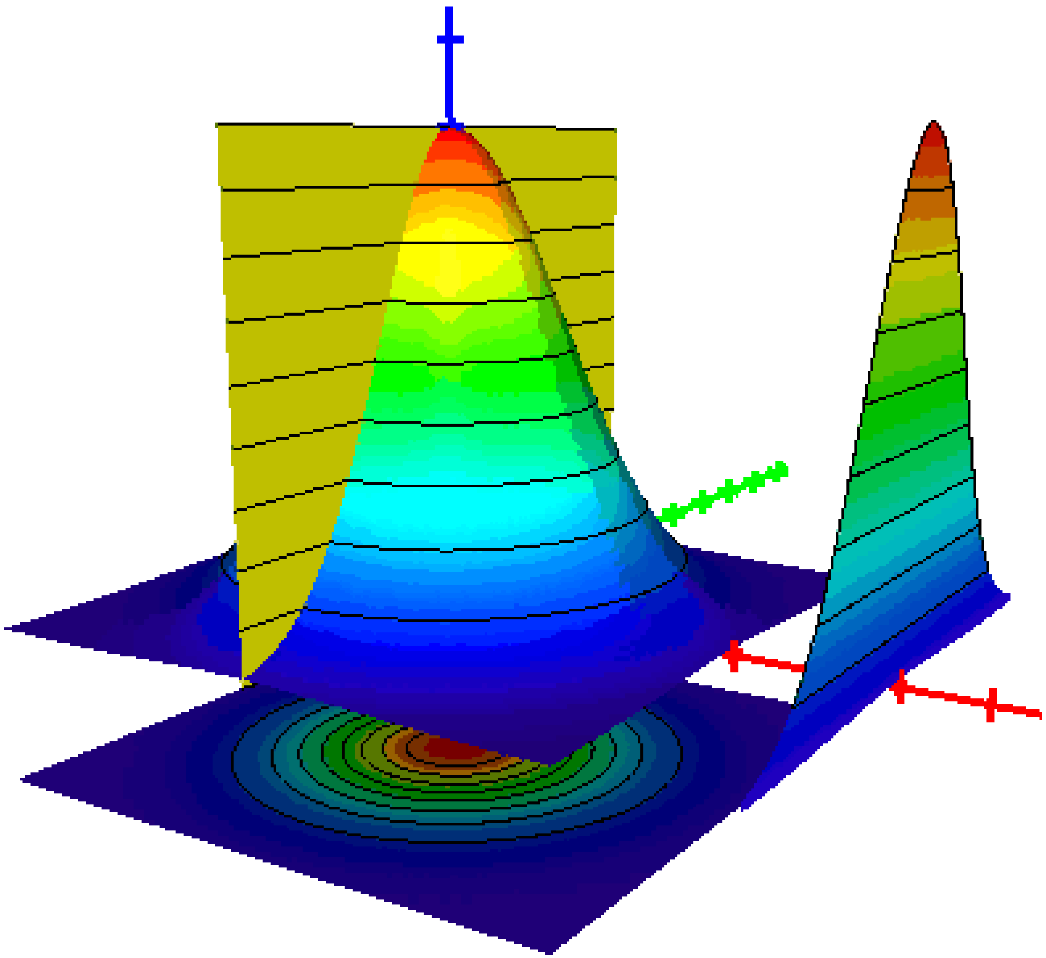
P(x|w)
W1
W2
x0
R1
R2
P(W1)=
0.5
P(W2)=0.5
decision boundary
1.5
1
0.5
0
-0.5
-1
-1.5
5
0
-5
-5
0
5

Case 2

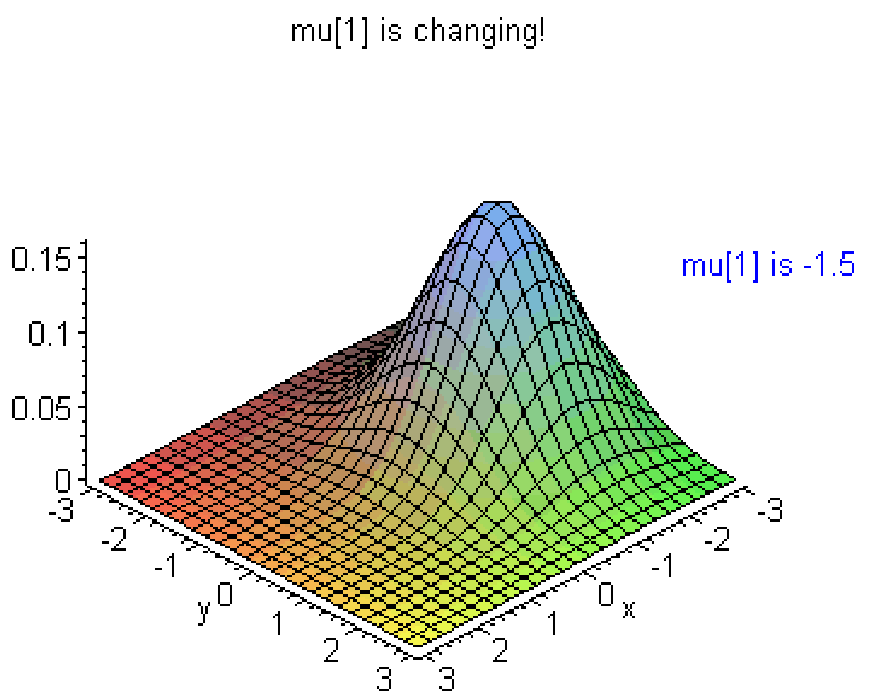

P(w1)=0.5
P(w2)=0.5
W1
W2
X0
R1
R2
Note the elliptical contour lines.
Decision boundary

Case 3

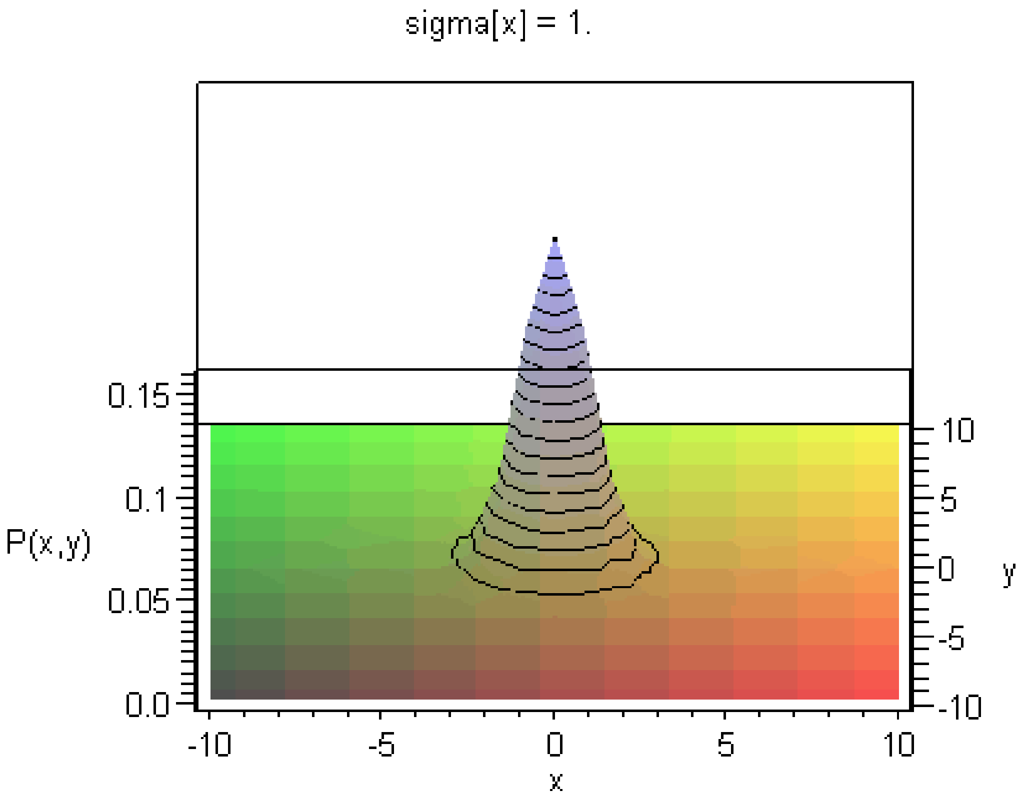




Normal density- mean is changing



Normal density- sigma is changing



Summary

- Bayesian Decision Theory
  - PR: essentially a decision process
  - Basic concepts
    - States of nature
    - Probability distribution, probability density function (pdf)
    - Class-conditional pdf
    - Joint pdf, marginal distribution, law of total probability
  - Bayes theorem
    - Prior + likelihood + observation ➔ Posterior probability
  - Bayes decision rule
    - Decide the state of nature with maximum posterior

Summary (Cont.)

- Feasibility of Bayes decision rule
  - Prior probability + likelihood
  - Solution I: counting relative frequencies
  - Solution II: conduct density estimation (chapters 3,4)
- Bayes decision rule: The general scenario
  - Allowing more than one feature
  - Allowing more than two states of nature
  - Allowing actions than merely deciding state of nature
  - Loss function: $\lambda : \ \Omega \times \mathcal{A} \rightarrow \mathbf{R}$

Summary (Cont.)

- Expected loss (*conditional risk*)

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

Average by enumerating over all possible states of nature

- General Bayes decision rule
  - Decide the action with minimum expected loss
- Minimum-error-rate classification
  - Actions ←→ Decide states of nature
  - Zero-one loss function
    - Assign *no loss/unit loss* for *correct/incorrect* decisions

Summary (Cont.)

- Discriminant functions
  - General way to represent classifiers
  - One function per category/class
  - Induce *decision regions* and *decision boundaries*
- Gaussian/Normal density

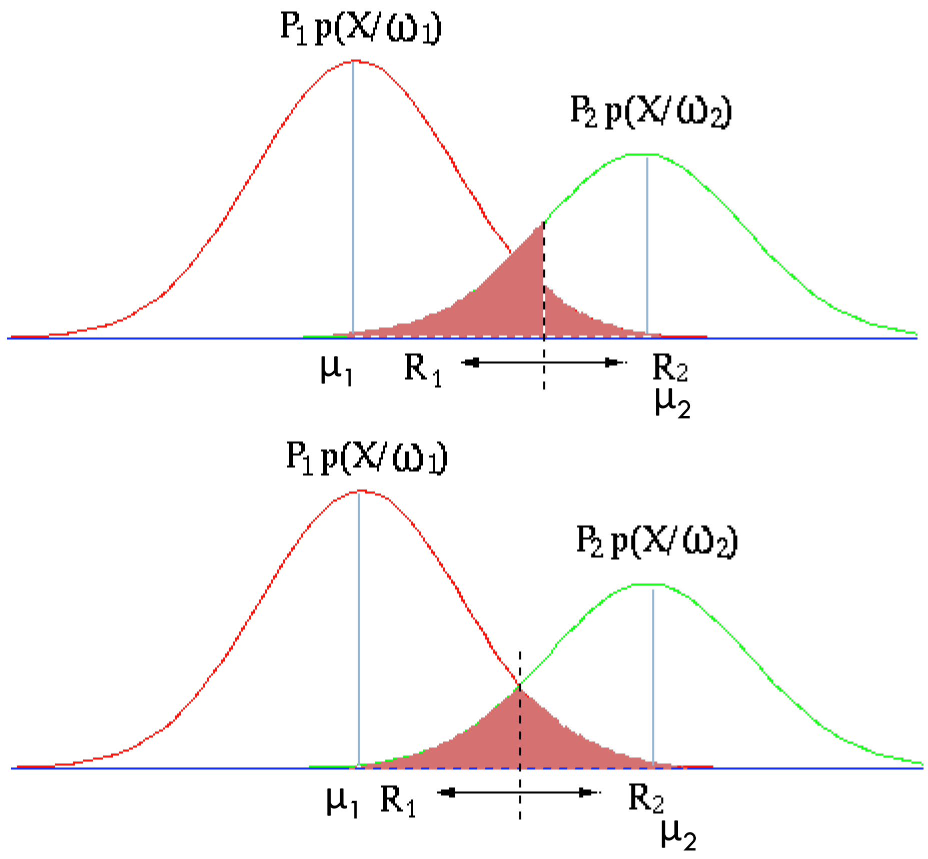
$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \quad p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- Discriminant functions for Gaussian pdf

$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}, \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$: linear discriminant function

$\boldsymbol{\Sigma}_i = \textbf{arbitrary}$: quadratic discriminant function

Thank you



P1 p(X/ω1)
P2 p(X/ω2)
μ1
R1
R2
μ2
P1 p(X/ω1)
P2 p(X/ω2)
μ1
R1
R2
μ2