Chapter 2 Bayesian Decision Theory

Bayesian Decision Theory

Pattern Recognition is a decision process in essence

Bayesian decision theory is a **statistical approach** to pattern recognition

Basic Assumptions

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

Key Principle: Bayes Theorem

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

$$P\left(\omega_j|x\right) = \frac{p\left(x|\omega_j\right) \cdot P\left(\omega_j\right)}{p(x)}$$

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

- X: the observed sample / **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** (先验概率) that H holds
  e.g. the probability of catching a salmon
- $P(X|H)$: the **likelihood** (似然度) 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** (后验概率) that H holds given X
  e.g. the probability of X being salmon given its length is 3-inch

State of Nature

- Future events that might occur
- State of nature is unpredictable
- Regarded as a random variable

e.g. let w denote the (discrete) random variable representing the **state of nature (class)** of fish types

> $w = w_1$: sea bass / $w = w_2$: salmon

Prior Probability

Prior probability is the probability distribution which **reflects one’s prior knowledge on the random variable**

- the catch produced as much sea bass as salmon $\rightarrow P(w_1) = P(w_2) = \frac{1}{2}$
- the catch produced more sea bass than salmon $\rightarrow P(w_1) = \frac{2}{3}; P(w_2) = \frac{1}{3}$

Class-conditional probability density function

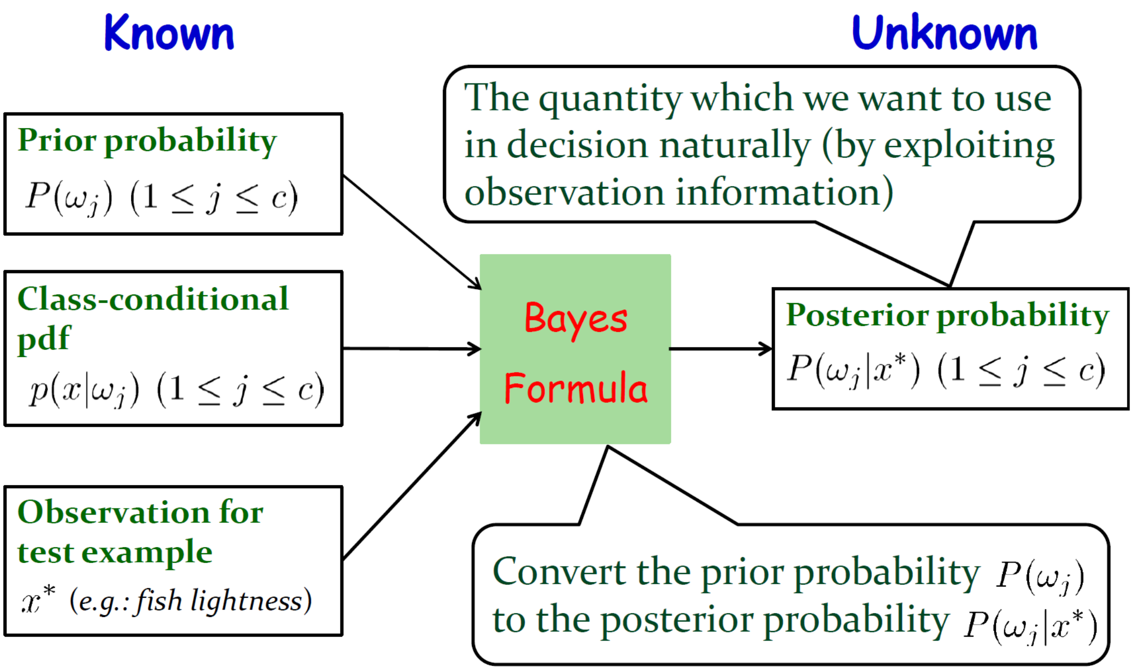
It is a probability density function (pdf) for x given that the state of nature (class) is ω

$$p(x|\omega) \geq 0 \quad \int_{-\infty}^{\infty} p(x|\omega)dx = 1$$

The class-conditional pdf describes the difference in the distribution of observations under different classes

$$p\left(x|\omega_1\right) \text{ should be different to } p\left(x|\omega_2\right)$$

Decision After Observation



Bayes Decision Rule

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

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

Two Special Cases

- Equal prior probability: Depends on the likelihood $P\left(x|\omega_j\right)$

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

- Equal likelihood: Degenerate to naïve decision rule

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

Example: Cancer or not?

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

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

Solution:

Define the notation:

$$\omega_1: \text{ cancer} \quad \omega_2: \text{ no cancer} \quad x \in \{+, -\}$$

Calculate likelihood:

$$\begin{array}{ll} P(\omega_1) = 0.008 & P(\omega_2) = 1 - P(\omega_1) = 0.992 \\ P(+ \mid \omega_1) = 0.98 & P(- \mid \omega_1) = 1 - P(+ \mid \omega_1) = 0.02 \\ P(- \mid \omega_2) = 0.97 & P(+ \mid \omega_2) = 1 - P(- \mid \omega_2) = 0.03 \end{array}$$

Calculate posterior probability:

$$P(\omega_1 \mid +) = \frac{P(\omega_1) P(+ \mid \omega_1)}{P(+)} = \frac{P(\omega_1) P(+ \mid \omega_1)}{P(\omega_1) P(+ \mid \omega_1) + P(\omega_2) P(+ \mid \omega_2)} = 0.2085$$

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

We find out that

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

So he/she have no cancer!

Feasibility of Bayes Formula

Get to know the Prior probability $P(w)$ and likelihood $p(x \mid w)$

- Counting relative frequencies (相对频率)
  e.g. Suppose we have randomly picked 1209 cars in the campus, got prices from their owners, and measured their heights

$$\begin{array}{l} cars\ in\ w_1 = 221 \longrightarrow P(\omega_1) = \frac{221}{1209} = 0.183 \\ cars\ in\ w_2 = 988 \longrightarrow P(\omega_1) = \frac{988}{1209} = 0.817 \end{array}$$

- Conduct density estimation (概率密度估计)
  Discretize the height spectrum (say [0.5m, 2.5m]) into 20 intervals each with length 0.1m, and then count the number of cars falling into each interval for either class

  Suppose $x = 1.05$，and x falls into interval $I_x = [1.0m, 1.1m] For$\omega{1}$, carsinI${x} $is46; For$\omega{2}$, carsinI${x}$ is 59
  So

$$\begin{array}{l} p(x = 1.05 \mid w_1) = \frac{46}{221} = 0.2081 \\ p(x = 1.05 \mid w_2) = \frac{59}{988} = 0.0597 \end{array}$$

Is Bayes Decision Rule Optimal?

Bayes Decision Rule (In case of two classes)

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

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

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

Under Bayes decision rule, we have

$$P(\text{ error } \mid x) = \min\left[P\left(\omega_1 \mid x\right), P\left(\omega_2 \mid x\right)\right]$$

The **average probability of error** over all possible x must be **as small as possible**

The General Case of Bays Decision Rule

- By allowing to **use more than one feature** (d-dimensional Euclidean space)

$$x \in \mathbf{R} \Longrightarrow \mathbf{x} \in \mathbf{R}^d$$

- By allowing more than two states of nature (finite set of c states of nature)

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$$

- By **allowing actions other than merely deciding the state of nature** (finite set of a possible actions)

$$\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$$

  Note that $c \neq a$

Loss function

$\lambda(w_j, \alpha_j)$ [also written as $\lambda(\alpha_i \mid w_j)$] is the loss incurred for taking action α_i when the state of nature is w_j

$$\lambda : \Omega \times \mathcal{A} \rightarrow R$$

A simple loss function

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

Action

Given a particular x, we have to decide which **action** to take.

With the loss of taking each action α_i, to minimize the loss $\lambda(\alpha_i \mid w_j)$

However, the true state of nature is uncertain: **Expected (average) loss**

Expected loss (Conditional risk)

Average by **enumerating** over all possible states of nature

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

where $\lambda\left(\alpha_i \mid \omega_j\right)$ is the incurred loss of taking action α_i in case of true state of nature being w_j, and $P\left(\omega_j \mid \mathbf{x}\right)$ is the probability of w_j being the true state of nature.

e.g.

$$\begin{aligned}
R\left(\alpha_1 \mid \mathbf{x}\right) & =\sum_{j=1}^2 \lambda\left(\alpha_1 \mid \omega_j\right) \cdot P\left(\omega_j \mid \mathbf{x}\right) \\
& =\lambda\left(\alpha_1 \mid \omega_1\right) \cdot P\left(\omega_1 \mid \mathbf{x}\right)+\lambda\left(\alpha_1 \mid \omega_2\right) \cdot P\left(\omega_2 \mid \mathbf{x}\right) \\
& =5 \times 0.01+60 \times 0.99=59.45
\end{aligned}$$

Task

find a mapping form patterns to actions

$$\alpha: \mathbf{R}^d \rightarrow \mathcal{A} \text{ (decision function)}$$

In other words, for every $\mathbf{x}$, the decision function $\alpha(\mathbf{x})$ assumes one of the a actions $\{\alpha_1, \ldots, \alpha_a\}$

Overall risk R

expected loss with decision function $\alpha(\cdot)$

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

where $p(\mathbf{x})$ is the pdf for pattern

For every $\mathbf{x}$, we ensure that the conditional risk $R(\alpha(\mathbf{x}) \mid \mathbf{x})$ is as small as possible, so the risk over all possible $\mathbf{x}$ must be as small as possible.

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

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

Example 1: Two-Category Classification

Classification setting

- $\Omega=\{w_1, w_2\}$: two states of nature
- $\mathcal{A}=\{\alpha_1, \alpha_2\}$: α_1 means decide w_1, α_2 means decide w_2
- $\lambda_{ij}=\lambda(\alpha_i \mid w_j)$: the loss incurred for deciding w_i when the true state of nature is w_j

The conditional risk

$$\begin{aligned}
& R\left(\alpha_1 \mid \mathbf{x}\right)=\lambda_{11} \cdot P\left(\omega_1 \mid \mathbf{x}\right)+\lambda_{12} \cdot P\left(\omega_2 \mid \mathbf{x}\right) \\
& R\left(\alpha_2 \mid \mathbf{x}\right)=\lambda_{21} \cdot P\left(\omega_1 \mid \mathbf{x}\right)+\lambda_{22} \cdot P\left(\omega_2 \mid \mathbf{x}\right)
\end{aligned}$$

If we assume that

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

by definition we get

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

by rearrangement we get

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

by Bayes theorem we get

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

because $\lambda_{21}-\lambda_{11}>0$, which means the loss for being error is ordinarily greater than the loss for being correct, we get

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

where $\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}$ is the **likelihood ratio** and $\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$ is a **constant** independent of $\mathbf{x}$

Example 2: Minimum-Error-Rate Classification

Classification setting

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

Zero-one (symmetrical) loss function

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

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

Proof

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

$1-P\left(\omega_i \mid \mathbf{x}\right)$ is the **error rate**, the probability that action α_i is wrong

To **minimum error rate**, decide ω_i if $P\left(\omega_i \mid \mathbf{x}\right)>P\left(\omega_j \mid \mathbf{x}\right)$ for all $j \neq i$

Minimax Criterion (最小最大化准则)

Generally, we assume that the **prior probabilities** over the states of nature $\Omega=\left\{\omega_1, \omega_2, \ldots, \omega_c\right\}$ **are fixed**.

Nonetheless, in some cases we need to design classifiers which can perform well under **varying prior probabilities**.

e.g. the prior probabilities of catching a sea bass or salmon fish might **vary in different regions**

The **minimax criterion** aims to find the classifier which can **minimize the worst overall risk** for **any value of the priors**

Example of Two-category classification

Suppose the two-category classifier $\alpha(\cdot)$ decides the feature of ω_1 in region $\mathcal{R}_1$ and decides the feature of ω_2 in region $\mathcal{R}_2$. Here, $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathbf{R}^d$ and $\mathcal{R}_1 \cap \mathcal{R}_2 = \emptyset$

$$\begin{aligned} R &= \int R(\alpha(\mathbf{x}) \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{R}_1} R(\alpha_1 \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_2} R(\alpha_2 \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

where

$$\begin{aligned} \int_{\mathcal{R}_1} R(\alpha_1 \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} &= \int_{\mathcal{R}_1} \sum_{j=1}^{2} R(\alpha_1 \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{R}_1} \sum_{j=1}^{2} \lambda_{1j} \cdot P(\omega_j) \cdot p(\mathbf{x} \mid \omega_j) d\mathbf{x} \\ &= \int_{\mathcal{R}_1} \left[\lambda_{11} \cdot P(\omega_1) \cdot p(\mathbf{x} \mid \omega_1) + \lambda_{12} \cdot P(\omega_2) \cdot p(\mathbf{x} \mid \omega_2)\right] d\mathbf{x} \end{aligned}$$

the same

$$\int_{\mathcal{R}_2} R(\alpha_2 \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} = \int_{\mathcal{R}_2} \left[\lambda_{21} \cdot P(\omega_1) \cdot p(\mathbf{x} \mid \omega_1) + \lambda_{22} \cdot P(\omega_2) \cdot p(\mathbf{x} \mid \omega_2)\right] d\mathbf{x}$$

Rewrite the overall risk R as a function of $P(\omega_1)$ via

$$\begin{gathered} P(\omega_1) = 1 - P(\omega_2) \\ \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1) d\mathbf{x} = 1 - \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} \end{gathered}$$

[In the second equation, $\int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1) d\mathbf{x}$ stands for the ratio of the sample $\mathbf{x}$ (its true state of nature is w_1)have been classified as w_1 (view as True Positive); while $\int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x}$ stands for the ratio of the sample $\mathbf{x}$ (its true state of nature is w_1)have been classified as w_2 (view as False Positive);]

we get

$$\begin{aligned} R =& \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} \\ & + P(\omega_1) \left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}\right] \end{aligned}$$

where R_{mm} stands for minimax risk

$$\begin{aligned} R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} \\ &= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} \end{aligned}$$

so for minimax solution we required:

$$\left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}\right] = 0$$

Discriminant Function (判别函数)

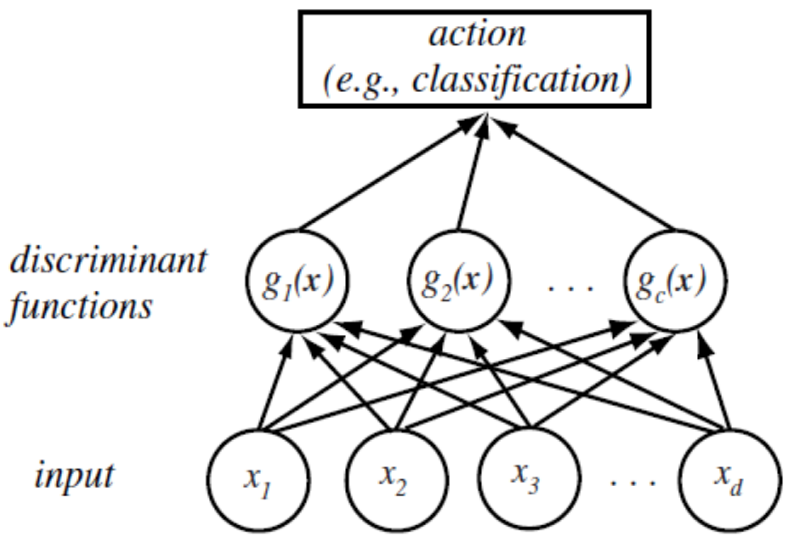
$$g_i : \mathbf{R}^d \to \mathbf{R} \quad (1 \leq i \leq c)$$

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

- Useful way to represent classifiers
- One function per category
- Various **discriminant functions** may leads to **Identical classification results**
  $f(\cdot)$ is a monotonically increasing function (单调递增函数), then

  $$f(g_i(\mathbf{x})) \Longleftrightarrow g_i(\mathbf{x})$$

  which means they are **equivalent in decision**



Type

- Minimum risk

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

- Minimum-error-rate

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

Decision region (决策区域)

From c discriminant functions, get c decision regions

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

Definition

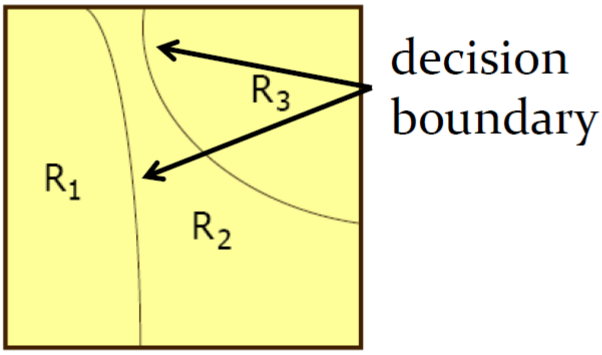
$$\mathcal{R}_i = \{\mathbf{x} \mid \mathbf{x} \in \mathbf{R}^d : g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall j \neq i\}$$

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

Decision boundary (决策边界)

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

On decision boundaries, more than one $g_i(\cdot)$ is maximum



Some Probability

Expected Value μ

$\sim$ stands for has the distribution

- Discrete

$$
\begin{aligned}
& x \in \mathcal{X}=\{x_1, x_2, \ldots, x_c\} \\
& x \sim P(\cdot) \\
& \mathcal{E}[x]=\sum_{x \in \mathcal{X}} x \cdot P(x)=\sum_{i=1}^c x_i \cdot P\left(x_i\right)
\end{aligned}
$$

- Continuous

$$
\begin{aligned}
& x \in \mathbf{R} \\
& x \sim p(\cdot) \\
& \mathcal{E}[x]=\int_{-\infty}^{\infty} x \cdot p(x) d x
\end{aligned}
$$

Variance σ^2

$$
\operatorname{Var}[x]=\mathcal{E}\left[(x-\mathcal{E}[x])^2\right]
$$

- Discrete

$$
\operatorname{Var}[x]=\sum_{i=1}^c\left(x_i-\mu\right)^2 \cdot P\left(x_i\right)
$$

- Continuous

$$
\operatorname{Var}[x]=\int_{-\infty}^{\infty}(x-\mu)^2 \cdot p(x) d x
$$

Vector Random Variables

- **joint** pdf

$$
\mathbf{x} \sim p(\mathbf{x})=p\left(x_1, x_2, \ldots, x_d\right)
$$

- **marginal** pdf

$$
\begin{gathered}
p\left(\mathbf{x}_1\right)=\int p\left(\mathbf{x}_1, \mathbf{x}_2\right) d \mathbf{x}_2 \\
\left(\mathbf{x}_1 \cap \mathbf{x}_2=\emptyset ; \mathbf{x}_1 \cup \mathbf{x}_2=\mathbf{x}\right)
\end{gathered}
$$

Expected vector

$$\mathcal{E}\left[x_i\right]=\int_{-\infty}^{\infty} x_i \cdot p\left(x_i\right) d x_i \quad(1 \leq i \leq d)$$

Covariance Matrix

Symmetric & **Positive semidefinite**

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

$$\begin{aligned} \sigma_{ij}=\sigma_{ji} & =\mathcal{E}\left[\left(x_i-\mu_i\right)\left(x_j-\mu_j\right)\right] \\ & =\int_{-\infty}^{\infty} \int_{-\infty}^{\infty}\left(x_i-\mu_i\right)\left(x_j-\mu_j\right) \cdot p\left(x_i, x_j\right) d x_i d x_j \end{aligned}$$

$$\sigma_{ii}=\operatorname{Var}\left[x_i\right]=\sigma_i^2$$

where $p\left(x_i, x_j\right)$ is **marginal pdf** on a pair of random variables $\left(x_i, x_j\right)$, exported from **joint pdf**.

Gaussian Density in Multivariate Case

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

where

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

and

$$\begin{aligned} & \mathrm{x}=\left(x_1, x_2, \ldots, x_d\right)^t: d \text{-dimensional column vector} \\ & \mu=\left(\mu_1, \mu_2, \ldots, \mu_d\right)^t: d \text{-dimensional mean vector} \end{aligned}$$

$$\begin{aligned} (\mathbf{x}-\boldsymbol{\mu})^t &: 1 \times d \text{ matrix} \\ \boldsymbol{\Sigma}^{-1} &: d \times d \text{ matrix} \\ (\mathbf{x}-\boldsymbol{\mu}) &: d \times 1 \text{ matrix} \\ (\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &: scalar\ (1 \times 1\ matrix) \end{aligned}$$

because Σ^{-1} is positive definite

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

Minimum-error-rate classification

$$g_i(\mathbf{x})=P\left(\omega_i \mid \mathbf{x}\right)$$

which the same as

$$g_i(\mathbf{x})=\ln P\left(w_i \mid \mathbf{x}\right)=\ln p\left(\mathbf{x} \mid w_i\right)+\ln P\left(w_i\right)$$

assume that

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

so we have

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

where $\frac{d}{2}\ln 2\pi$ is a constant that could be ignored

Case 1: $\boldsymbol{\Sigma}_i=\sigma^2\mathbf{I}$

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

so we get

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

rearrange

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

where $\mathbf{x}^t\mathbf{x}$ is the same for all states of nature which could be ignored

Finally we get a **Linear discriminant functions** (线性判别函数)

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

where $\mathbf{w}_i=\frac{1}{\sigma^2}\boldsymbol{\mu}_i$ is weight vector and $w_{i0}=-\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i+\ln P(\omega_i)$ is threshold/bias

Case 2: $\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}$

$(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ is call squared **Mahalanobis distance** (马氏距离)

when $\boldsymbol{\Sigma}=\mathbf{I}$ it reduces to Euclidean distance

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

where $\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}$ is the same for all states of nature which could be ignored

Finally we get a **Linear discriminant functions**

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

where $\mathbf{w}_i=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ is weight vector and $w_{i0}=-\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i+\ln P(\omega_i)$is threshold/bias

Case 3: $\boldsymbol{\Sigma}_{i}=Arbitrary (任意的)$

Using **quadratic discriminant function** (二次判别函数)

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

where

$$\begin{gathered}\mathbf{W}_i=-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\ \text{quadratic matrix}\\ \mathbf{w}_i=\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\ \text{weight vector}\\ w_{i0}=-\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i-\frac{1}{2}\ln|\boldsymbol{\Sigma}_i|+\ln P(\omega_i)\ \text{threshold/bias}\end{gathered}$$