Xi'an Jiaotong-Liverpool University

Pattern Recognition

Lecture 03(b). Bayesian Decision Theory

Dr. Shanshan.ZHAO & Dr. Yuxuan.ZHAO

School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University

Academic Year 2023-2024

Table of Contents

1. Bayes' Rule
2. Minimizing the misclassification rate

Bayesian Decision theory

Design classifiers to recommend decisions that minimize some total expected "risk".

- fundamental statistical approach to the problem of pattern classification
- ideal case, optimal classifier
- compare with all other classifiers

Fish example

- Reconsider the problem: Classify two fish as salmon or seabass.

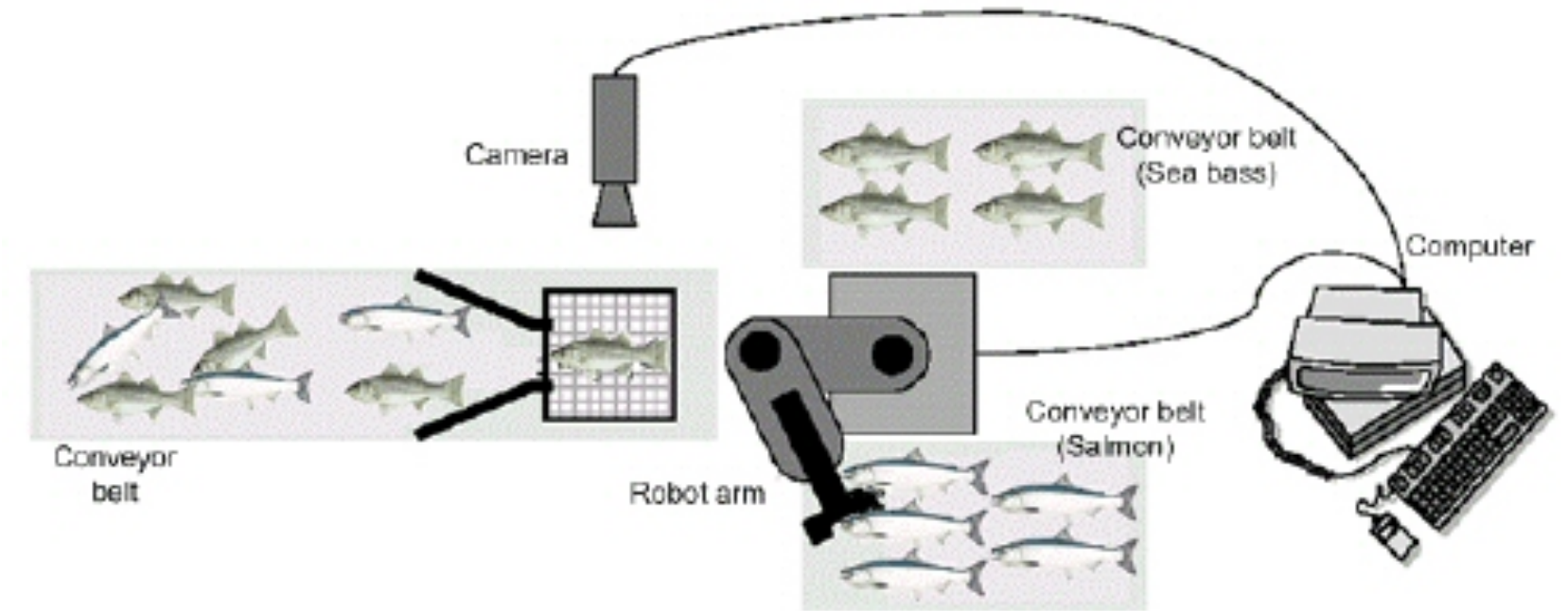


Figure 1: The fish packing system.

Fish example

- Assume any given fish is either a salmon or a sea bass.
- Let's define a variable ω that describes the *state of nature*

$$\omega = \omega_1 \quad \text{for sea bass}$$
$$\omega = \omega_2 \quad \text{for salmon}$$

(a) (b)

Figure 2: The objects to be classified: a. salmon; b. sea bass

1 Bayes' Rule

2 Minimizing the misclassification rate

Prior Probability

If sea bass is produced as much as salmon, we would say that the next fish is equally like to be sea bass or salmon.
More generally, we assume there is a **prior probability**.

- The a priori or prior probability reflects our knowledge of how likely we expect a certain category before we can actually observe.
- The priors must exhibit exclusivity and exhaustivity. For c states of nature, or classes:
  $\sum_{i=1}^{c} P(\omega_i) = 1$
  In the fish example, $P(\omega_1) + P(\omega_2) = 1$

Decision Rule with Only Priors

A **decision rule** prescribes what action to take based on observed input.

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

- Favours the most likely class
- This rule will be making the same decision all times -i.e., optimum if no other information is available

What can we say about this decision rule?

Decision Rule with Only Priors

A **decision rule** prescribes what action to take based on observed input.
Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2

- Favours the most likely class
- This rule will be making the same decision all times -i.e., optimum if no other information is available

What can we say about this decision rule?
Seems reasonable, but it will always make the same choice. It doesn't make sense if we were to judge many fish.

Features and Feature Spaces

Mostly, we are not asked to make decisions with such little information. We might use some measurements or features to improve our classifier.

- 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, etc.
- For simplicity, we assume our features are all continuous values.
- Denote a scalar feature as x and a vector features as x. For a d-dimensional feature space, $x \in \mathscr{R}^d$.

Conditional probability density

- The Conditional probability density $p(x|\omega_i)$is also called *likelihood*. It shows the probability density of feature x, given that it belongs to class ω_i.
- Example

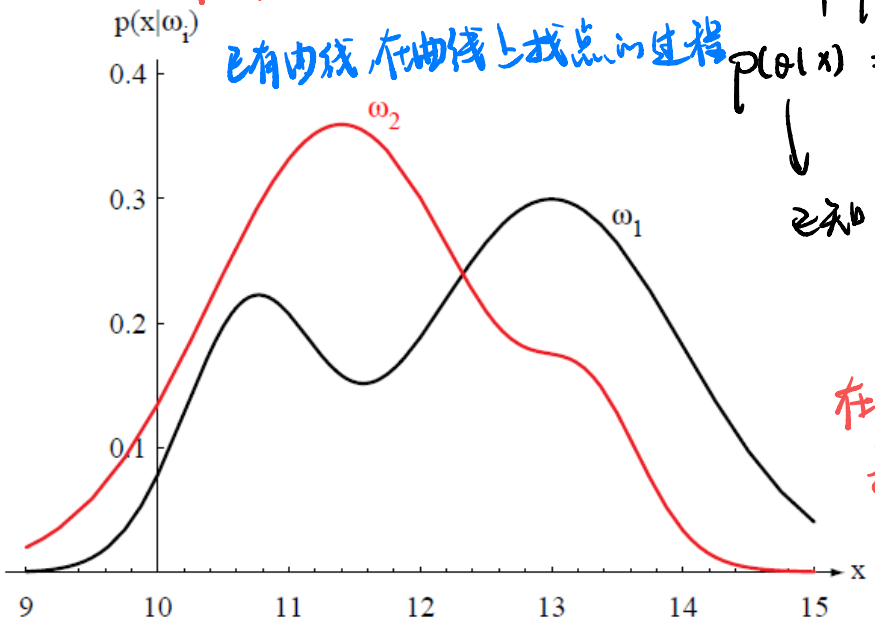


Figure 3: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x

Posterior Probability

- Now that we know the prior distribution as well as the conditional density, how does this affect our decision rule?
- Posterior probability is the probability of a class given our observations: $P(\omega|x)$.
- Bayes Formula:

同时发生的概率

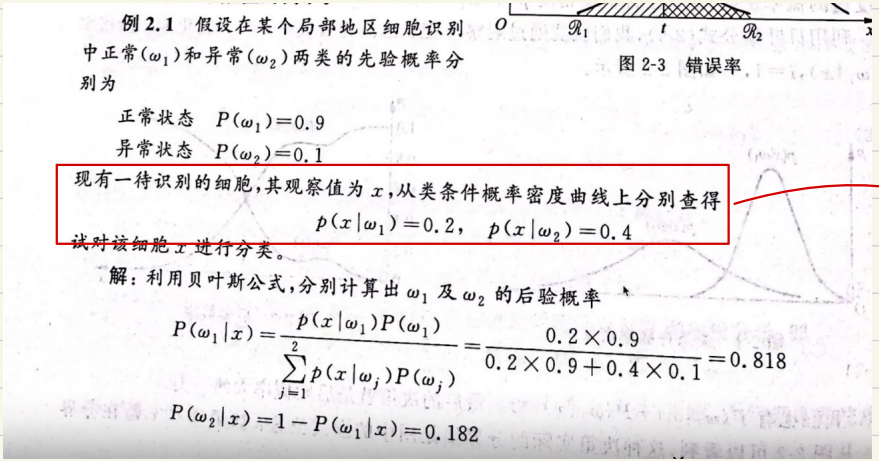
$$P(\omega, x) = P(\omega|x)p(x) = p(x|\omega)P(\omega) \quad (1)$$

$$P(\omega|x) = \frac{p(x|\omega)P(\omega)}{p(x)} \quad (2)$$

$$= \frac{p(x|\omega)P(\omega)}{\sum_i p(x|\omega_i)P(\omega_i)} \quad (3)$$

- Notice that the likelihood and the prior govern the posterior. The $p(x)$ evidence term is a scale-factor to normalize the density.

- P(ω|x) 表示在给定数据 x 的情况下，类别 ω 的后验概率。这是我们想要计算的结果，即在观测到数据 x 后，类别 ω 是多么可能。
- P(ω) 表示先验概率，即在观测任何数据之前，类别 ω 的概率分布。这是我们对类别 ω 的先验信念或先验估计。
- p(x|ω) 表示似然性，即在已知类别 ω 的情况下，观测到数据 x 的概率。这个部分描述了给定类别下数据 x 的分布。



例 2.1 假设在某个局部地区细胞识别中正常(ω_1)和异常(ω_2)两类的先验概率分别为

正常状态 $P(\omega_1)=0.9$

异常状态 $P(\omega_2)=0.1$

现有一待识别的细胞，其观察值为 x，从类条件概率密度曲线上分别查得

$$p(x|\omega_1)=0.2,\quad p(x|\omega_2)=0.4$$

试对该细胞 x 进行分类。

图 2-3 错误率

查得？

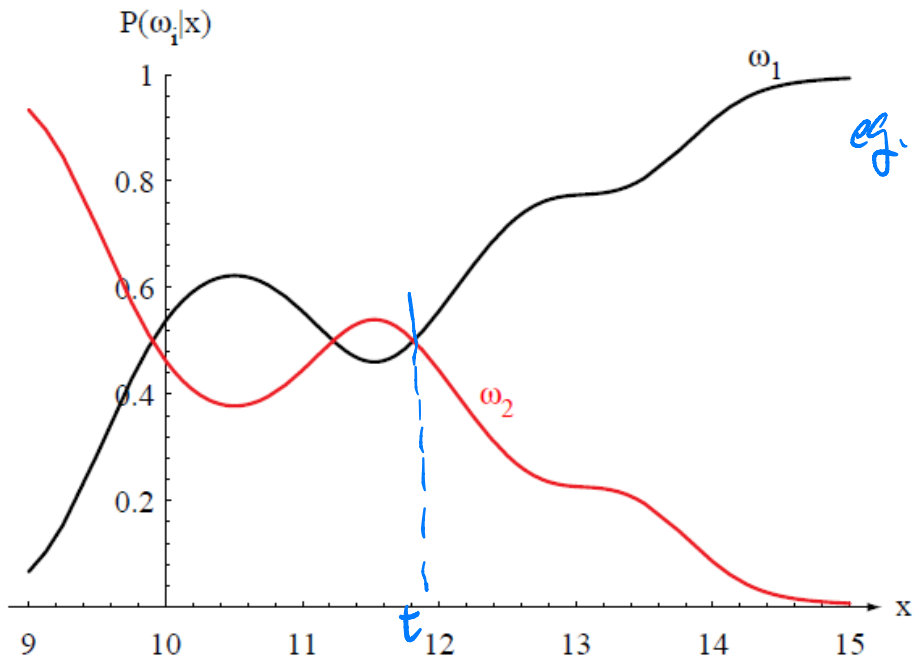
解：利用贝叶斯公式，分别计算出 ω_1 及 ω_2 的后验概率

$$P(\omega_1|x)=\frac{p(x|\omega_1)P(\omega_1)}{\sum_{j=1}^{2}p(x|\omega_j)P(\omega_j)}=\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}=0.818$$

$$P(\omega_2|x)=1-P(\omega_1|x)=0.182$$

Posterior Probability

For the case of $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$, the posterior is



后验概率
eg. 当x>t，
决策成ω1，因为P(ω1|x)更高

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

$$\omega = arg \max_i P(\omega_i|x) \tag{4}$$

Decision Rule Using Posteriors

Recap: Using Bayes' rule, the posterior probability of category ω_j given measurement x is given by: Probabilities *Bayes rule*.

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

The *Bayes classification rule* can be stated as

- Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; or
- Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$;

  Decision making relies on both the priors and the likelihoods and Bayesian Decision Rule combines them to achieve the minimum probability of error.

Error Probability

For the two class situation, we have

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1 \end{cases} \quad (5)$$

We can minimize the probability of error by :

$$\text{Decide} \quad \omega_1 \quad \text{if} \quad P(\omega_1|x) > P(\omega_2|x); \quad \text{otherwise decide} \quad \omega_2 \quad (6)$$

$$P(error|x) = \min[P(\omega_1|x), P(\omega_2|x)] \quad (7)$$

And , this minimizes the average probability of error too:

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)dx \quad (8)$$

Because the integral will be minimized when we can ensure each $P(error|x)$is as small as possible.

多类别决策过程中，要把特征空间分割成 $\mathcal{R}_1$，$\mathcal{R}_2$，…，$\mathcal{R}_c$ 个区域，可能错分的情况很多，平均错误概率 $P(e)$ 将由 $c(c-1)$ 项组成，即

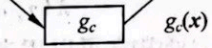


图 2-4　多类判别决策

$$
\begin{aligned}
P(e) = {} & [P(\boldsymbol{x} \in \mathcal{R}_2 \mid \omega_1) + P(\boldsymbol{x} \in \mathcal{R}_3 \mid \omega_1) + \cdots + P(\boldsymbol{x} \in \mathcal{R}_c \mid \omega_1)]P(\omega_1) \\
& + [P(\boldsymbol{x} \in \mathcal{R}_1 \mid \omega_2) + P(\boldsymbol{x} \in \mathcal{R}_3 \mid \omega_2) + \cdots + P(\boldsymbol{x} \in \mathcal{R}_c \mid \omega_2)]P(\omega_2) \\
& + \cdots \\
& + [P(\boldsymbol{x} \in \mathcal{R}_1 \mid \omega_c) + P(\boldsymbol{x} \in \mathcal{R}_2 \mid \omega_c) + \cdots + P(\boldsymbol{x} \in \mathcal{R}_{c-1} \mid \omega_c)]P(\omega_c)
\end{aligned}
$$

c 行

每行 $c-1$ 项

$$
= \sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} [P(\boldsymbol{x} \in \mathcal{R}_j \mid \omega_i)]P(\omega_i)
$$

该式计算量比较大，可以通过计算平均正确率 $P(c)$ 来计算错误率

$$
P(c) = \sum_{j=1}^{c} P(\boldsymbol{x} \in \mathcal{R}_j \mid \omega_j)P(\omega_j) = \sum_{j=1}^{c} \int_{\mathcal{R}_j} p(\boldsymbol{x} \mid \omega_j)P(\omega_j)\mathrm{d}\boldsymbol{x} \tag{2-18}
$$

$$
P(e) = 1 - P(c) = 1 - \sum_{j=1}^{c} P(\omega_j) \int_{\mathcal{R}_j} p(\boldsymbol{x} \mid \omega_j)\mathrm{d}\boldsymbol{x} \tag{2-19}
$$

1 Bayes' Rule

2 Minimizing the misclassification rate

Minimizing the misclassification rate

贝叶斯分类器在最小化分类错误率上是最优的

Goal: To make as a few misclassifications as possible.
A mistake occurs when an input vector belongs to class ω_1 is assigned to class ω_2 or vice versa.

同时发生 $P(A\cap B)=P(B|A)P(A)$

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

不太明白！

Question: Is it true that the probability of misclassification is minimised by assigning each point to the class with the maximum posterior probability? True

使错误率最小的分类决策就是使后验概率最大的决策？ 互补关系？

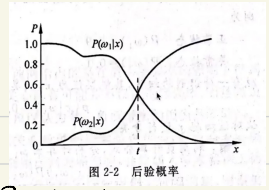


图 2-2 后验概率

错误率：

决策边界把 x 轴分割成 2 个区域，分别称为第一类和第二类的决策区域 R_1 和 R_2。R_1 为 $(-\infty, t)$，R_2 为 $(t, +\infty)$。

样本在 R_1 中，但属于第二类的概率 / 样本在 R_2 中，但属于第一类的概率 => 出现错误的概率

$$P(error) = \int_{-\infty}^{t} P(x|\omega_2)P(\omega_2)dx + \int_{t}^{+\infty} P(x|\omega_1)P(\omega_1)dx$$

也可以写成：

$$\begin{aligned} P(error) &= P(x \in R_1, \omega_2) + P(x \in R_2, \omega_1) \\ &= P(x \in R_1 | \omega_2) \cdot P(\omega_2) + P(x \in R_2 | \omega_1) \cdot P(\omega_1) \\ &= P(\omega_2) \cdot \int_{R_1} P(x|\omega_2)dx + P(\omega_1) \int_{R_2} P(x|\omega_1)dx \end{aligned}$$

$$= P(\omega_2) \cdot P_2(e) + P(\omega_1) P_1(e)$$

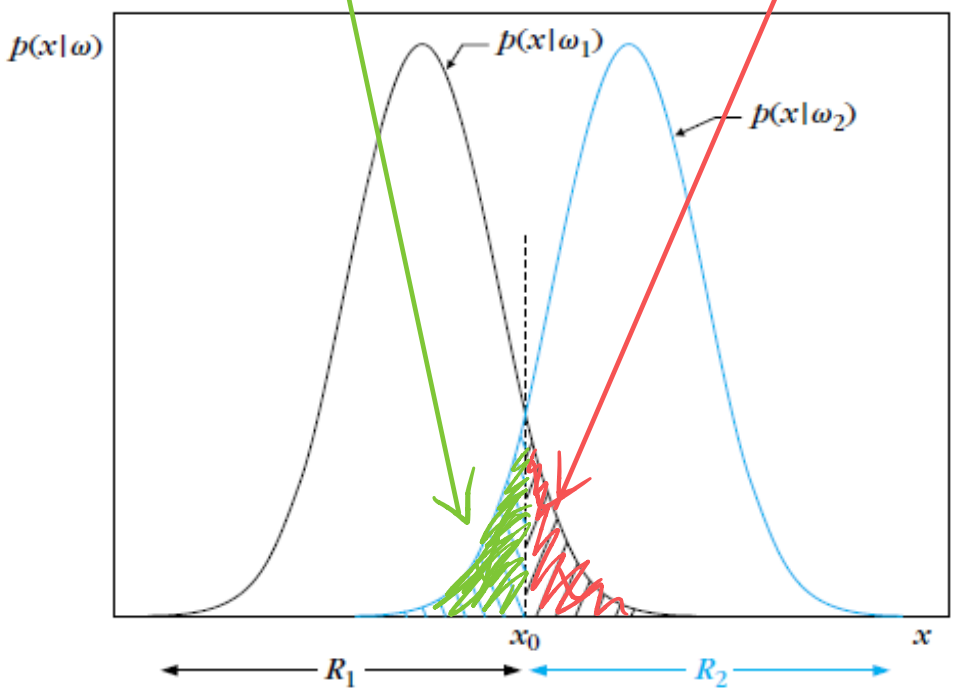
其中 $P_1(e) = \int_{R_2} P(x|\omega_1)dx$ => 把第一类样本决策为第二类的错误率

$P_2(e) = \int_{R_1} P(x|\omega_2)dx$ => 把第二类样本决策为第一类的错误率

Minimizing the misclassification rate

Example 1. The two regions R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes. $P(\omega_1) = P(\omega_2) = \frac{1}{2}$

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)dx \qquad (9)$$



Minimizing the misclassification rate

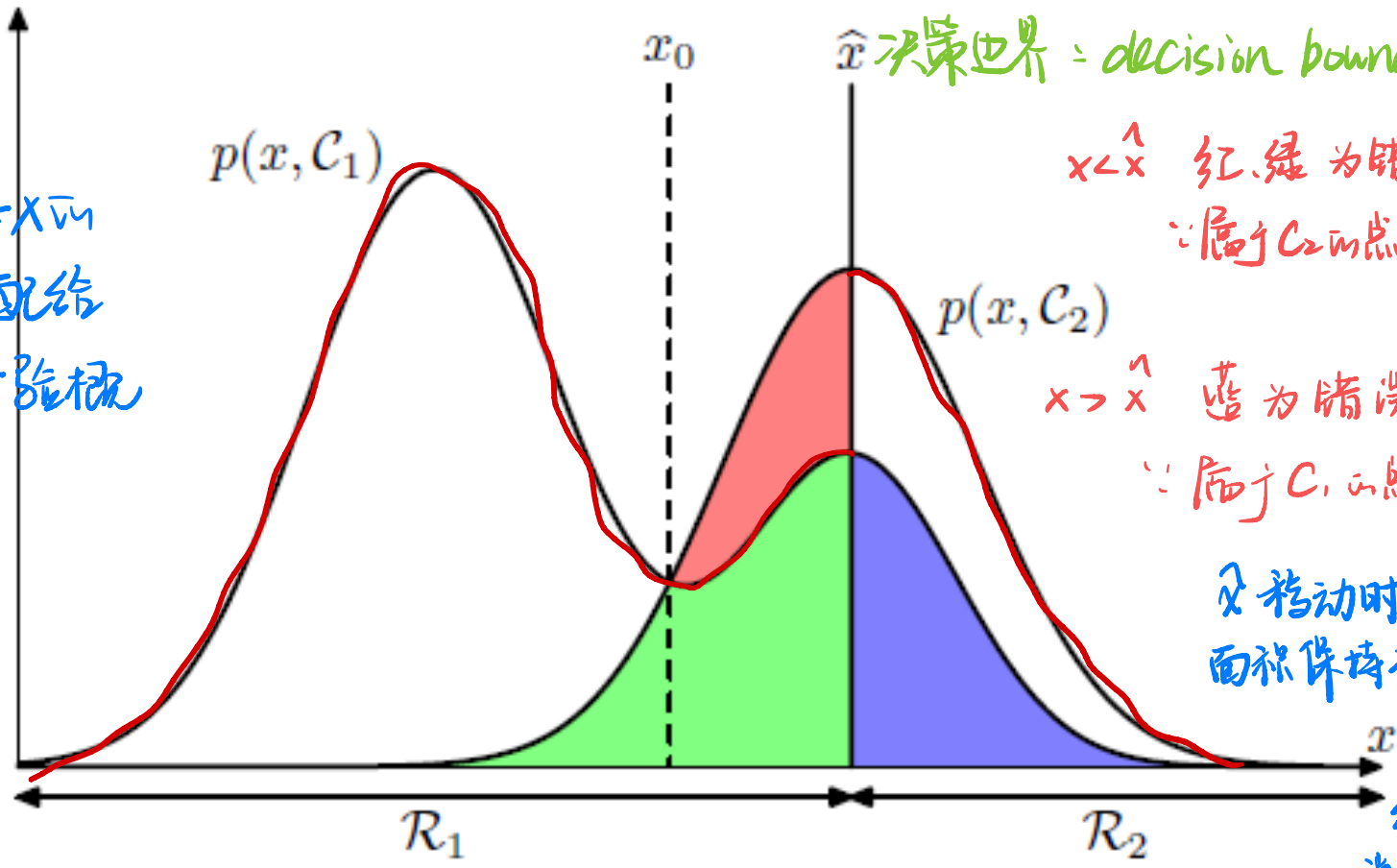


Figure 5: Schematic illustration of the joint probabilities $p(x, C_k)$ for each of two classes plotted against x, together with the decision

Minimizing the misclassification rate

It is possible to extend this justification for a decision rule based on **maximum posterior probability**.
Therefore, we consider the probability for a pattern being correctly classified $P(correct)$.
Exercises:

- 1. $P(correct) = ?$
- 2. Prove that the maximum posterior probability decision rule is equivalent to minimising the probability of misclassification.

ion

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

$$
\begin{aligned}
P(\text{error}) &= P(x \in \mathcal{R}_1, \omega_2) + P(x \in \mathcal{R}_2, \omega_1) \\
&= \int_{\mathcal{R}_1} p(x, \omega_2) dx + \int_{\mathcal{R}_2} p(x, \omega_1) dx \\
&= \int_{\mathcal{R}_1} p(x|\omega_2) P(\omega_2) dx + \int_{\mathcal{R}_2} p(x|\omega_1) P(\omega_1) dx \\
&= P(\omega_2) \int_{\mathcal{R}_1} p(x|\omega_2) dx + P(\omega_1) \int_{\mathcal{R}_2} p(x|\omega_1) dx \\
&= P(\omega_2)[1 - \int_{\mathcal{R}_2} p(x|\omega_2) dx] + P(\omega_1)[1 - \int_{\mathcal{R}_1} p(x|\omega_1) dx] \\
&= P(\omega_2) + P(\omega_1) - [\int_{\mathcal{R}_2} p(x|\omega_2) P(\omega_2) dx + \int_{\mathcal{R}_1} p(x|\omega_1) P(\omega_1) dx] \\
&= 1 - P(\text{correct})
\end{aligned}
$$

Thank You !

$\mathcal{Q}$ & $\mathcal{A}$