

Pattern Classification

All materials in these slides were taken from

Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000

with the permission of the authors and the publisher

Chapter 2

Bayesian Decision Theory (Sections 2.3-2.5)

- Minimum-Error-Rate Classification
- Classifiers, Discriminant Functions and Decision Surfaces
- The Normal Density

- Definitions of actions
- Action α_i : assign the test sample to the i -th class
- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

2.3 Minimum-Error-Rate Classification

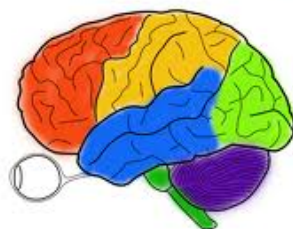
- Actions are decisions on classes

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 $i \neq j$



Decision Making



- The loss function for above case is the zero-one loss function:

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

Therefore, the conditional risk is:

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

Minimum-Error
Bayesian Decision !

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

- Minimize the risk requires to maximize $P(\omega_i | x)$
(since $R(\alpha_i | x) = 1 - P(\omega_i | x)$)

- For Minimum error rate

- Decide ω_i if $P(\omega_i | x) > P(\omega_j | x) \quad \forall j \neq i$

Minimum-Error and Maximum income in expectation may be not optimal

© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

ABC咨询公司



“我们这次推荐的策略比较冒险，要是您能在离开前就把钱付了我们会万分感谢的。”



How to minimize the risk ?



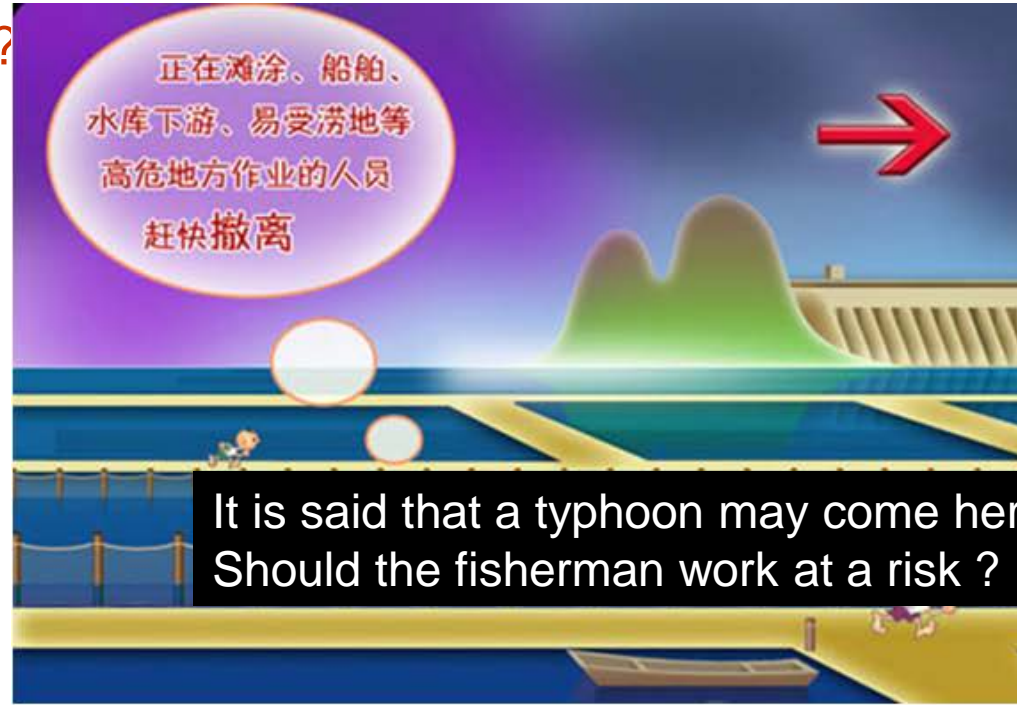
What should we do ?

How to define the risk (loss function) ?

How to obtain the minimum-risk
and make the decision safe enough ?



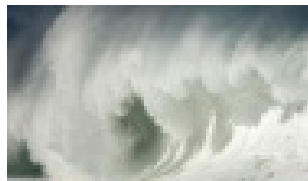
A earthquake may occur at a probability of 50%.
Should we issue the earthquake forecast ?



It is said that a typhoon may come here.
Should the fisherman work at a risk ?



It is said that a seaquake may come at a probability of



Minimum-risk decision

The conditional risk

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



$\lambda(\alpha_i | \omega_j)$ can be defined in accordance with the real applications.

$\lambda(\alpha_i | \omega_j) \equiv \lambda_{ij}$:The loss (cost) in the case where the class label is j but action α_i is adopted (the test sample is classified into the i -th class).

For two-class problem



$$R(\alpha_1 | x) = \lambda(\alpha_1 | \omega_1)P(\omega_1 | x) + \lambda(\alpha_1 | \omega_2)P(\omega_2 | x)$$

$$R(\alpha_2 | x) = \lambda(\alpha_2 | \omega_1)P(\omega_1 | x) + \lambda(\alpha_2 | \omega_2)P(\omega_2 | x)$$

If $R(\alpha_1 | x) < R(\alpha_2 | x)$

Then the test sample is classified into the first class ω_1

Minimum-risk decision

$$\begin{aligned} R(\alpha_1 | x) < R(\alpha_2 | x) &\longrightarrow \lambda_{21} \frac{P(x | \omega_1)P(\omega_1)}{P(x)} + \lambda_{22} \frac{P(x | \omega_2)P(\omega_2)}{P(x)} > \\ &\quad \lambda_{11} \frac{P(x | \omega_1)P(\omega_1)}{P(x)} + \lambda_{12} \frac{P(x | \omega_2)P(\omega_2)}{P(x)} \\ &\quad \downarrow \\ \frac{P(x | \omega_1)}{P(x | \omega_2)} &> \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} \end{aligned}$$

- Regions of decision and zero-one loss function, therefore:

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

- If λ is the zero-one loss function which means:

$$\lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

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

$$\text{if } \lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \text{ then } \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$

A high λ_{ij} means a high loss (cost).

λ_{ii} may be or be not zero.



An example: cancer diagnosis (+: 49%; - : 51%. How to reduce the risk ?)

An example: earthquake forecast



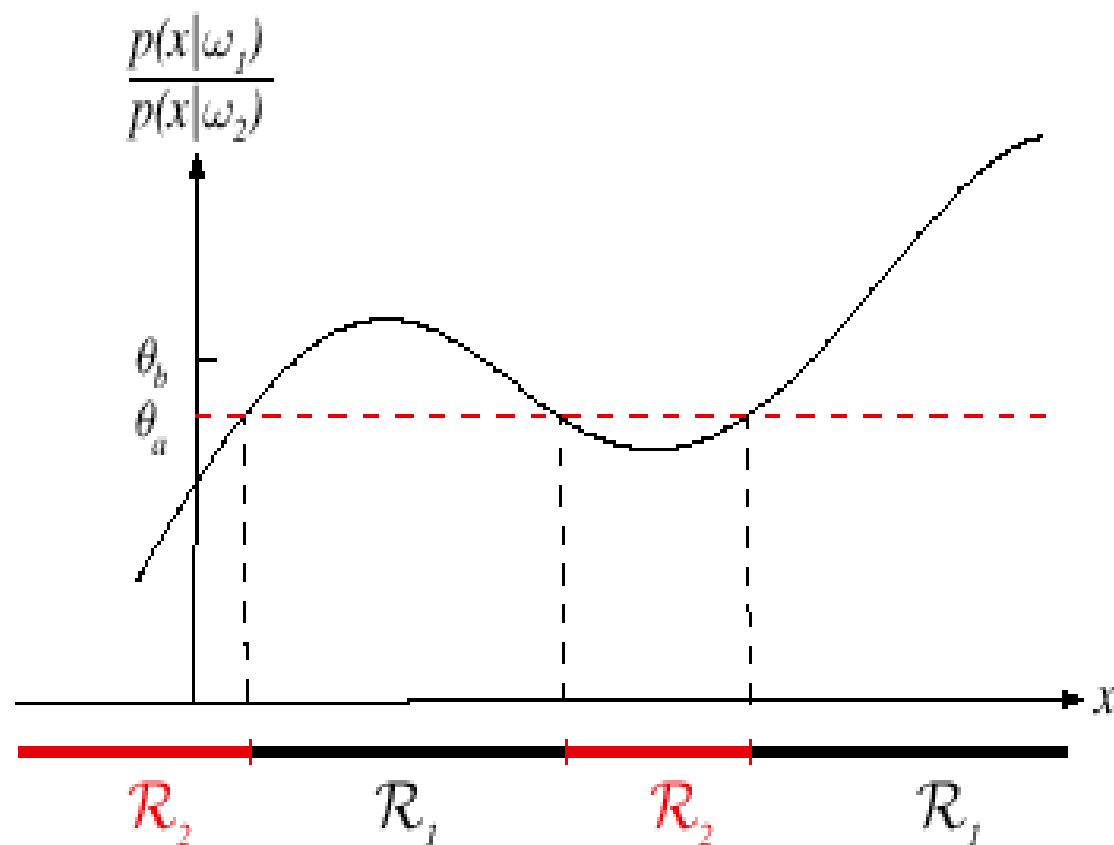


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a . If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b , and hence \mathcal{R}_1 becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

■ Example of Minimum-Error-Rate Classification

- The **error-rate** is the sole criterion of classification



例题:

地震预报是比较困难的一个课题,可以根据地震与生物异常反应之间的联系来进行研究。根据历史记录的统计,地震前一周内出现生物异常反应的概率为 50%,而一周内没有发生地震但也出现了生物异常反应的概率为 10%。假设某一个地区属于地震高发区,发生地震的概率为 20%。问:

如果某日观察到明显的生物异常反应现象,是否应当预报一周内将发生地震?

解:

把地震是否发生设成两个类别:发生地震为 ω_1 , 不发生地震为 ω_2 ;

则两个类别出现的先验概率 $P_1=0.2$, $P_2=1-0.2=0.8$;

设地震前一周是否出现生物异常反应这一事件设为 x , 当 $x=1$ 时表示出现了, $x=0$ 时表示没出现;

则根据历史记录统计可得, : $p(x=1|\omega_1)=0.5$, $p(x=1|\omega_2)=0.1$

设地震前一周是否出现生物异常反应这一事件设为 x ，当 $x=1$ 时表示出现了， $x=0$ 时表示没出现；

则根据历史记录统计可得，： $p(x=1|\omega_1)=0.5$ ， $p(x=1|\omega_2)=0.1$

所以，某日观察到明显的生物异常反应现象，此时可以得到将发生地震的概率为：

$$\begin{aligned} p(\omega_1|x=1) &= (P_1 \times p(x=1|\omega_1)) / (P_1 \times p(x=1|\omega_1) + P_2 \times p(x=1|\omega_2)) \\ &= (0.2 \times 0.5) / (0.2 \times 0.5 + 0.8 \times 0.1) = 5/9 \end{aligned}$$

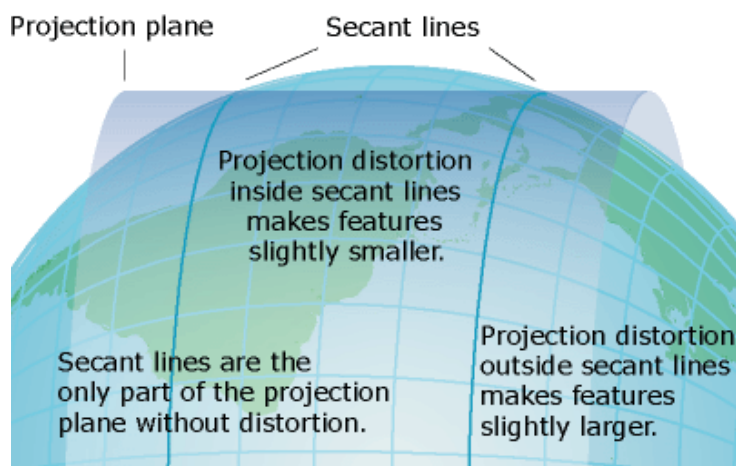
而不发生地震的概率为：

$$\begin{aligned} p(\omega_2|x=1) &= (P_2 \times p(x=1|\omega_2)) / (P_1 \times p(x=1|\omega_1) + P_2 \times p(x=1|\omega_2)) \\ &= (0.8 \times 0.1) / (0.2 \times 0.5 + 0.8 \times 0.1) = 4/9 \end{aligned}$$

因为 $p(\omega_1|x=1) > p(\omega_2|x=1)$ ，所以在观察到明显的生物异常反应现象时，发生地震的概率更高，所以应当预报一周内将发生地震。

Disadvantage of Minimum-Error-Rate Classification

Minimum-Error might be not optimal. The cost is lower than the tunnel



Minimum-risk is better !!

Minimum-risk

例题：

对于上例中的地震预报问题，假设预报一周内发生地震，可以预先组织抗震救灾，由此带来的防灾成本会有 2500 万元，而当地震确实发生时，由于地震造成的直接损失会有 1000 万元；假设不预报将发生地震而地震又发生了，造成的损失会达到 5000 万元。请问在观察到明显的生物异常反应后，是否应当预报一周内将发生地震？

解：

设决策 1 为发布地震预报，决策 2 为不发布地震预报，则



Minimum-risk

发生了地震，而提前发布了地震预报，此时的损失为 $\lambda_{11}=2500+1000=3500$ 万元；

发生了地震，而没有提前发布地震预报，此时的损失为 $\lambda_{21}=5000$ 万元；

没有发生地震，而提前发布了地震预报，此时的损失为 $\lambda_{12}=2500$ 万元；

没有发生地震，而没有提前发布地震预报，此时的损失为 $\lambda_{22}=0$ 元；

则在观察到明显的生物异常反应现象时，发布地震预报的条件风险为：

$R(\text{发布地震预报}|x=1) = \lambda_{11} \times p(\omega_1|x=1) + \lambda_{12} \times p(\omega_2|x=1) = 3500 \times 5/9 + 2500 \times 4/9 = 3056$ 万元；

而不发布地震预报带来的综合损失为：

$R(\text{不发布地震预报}|x=1) = \lambda_{21} \times p(\omega_1|x=1) + \lambda_{22} \times p(\omega_2|x=1) = 5000 \times 5/9 = 2778$ 万元；

因为 $R(\text{发布地震预报}|x=1) > R(\text{不发布地震预报}|x=1)$

所以，发布地震预报风险更大，不应该发布地震预报。



2.4 Classifiers, Discriminant Functions and Decision Surfaces

■ The multi-category case

- Set of discriminant functions $g_i(x)$, $i = 1, \dots, c$
- The classifier assigns a feature vector x to class ω_i
if: $g_i(x) > g_j(x) \quad \forall j \neq i$

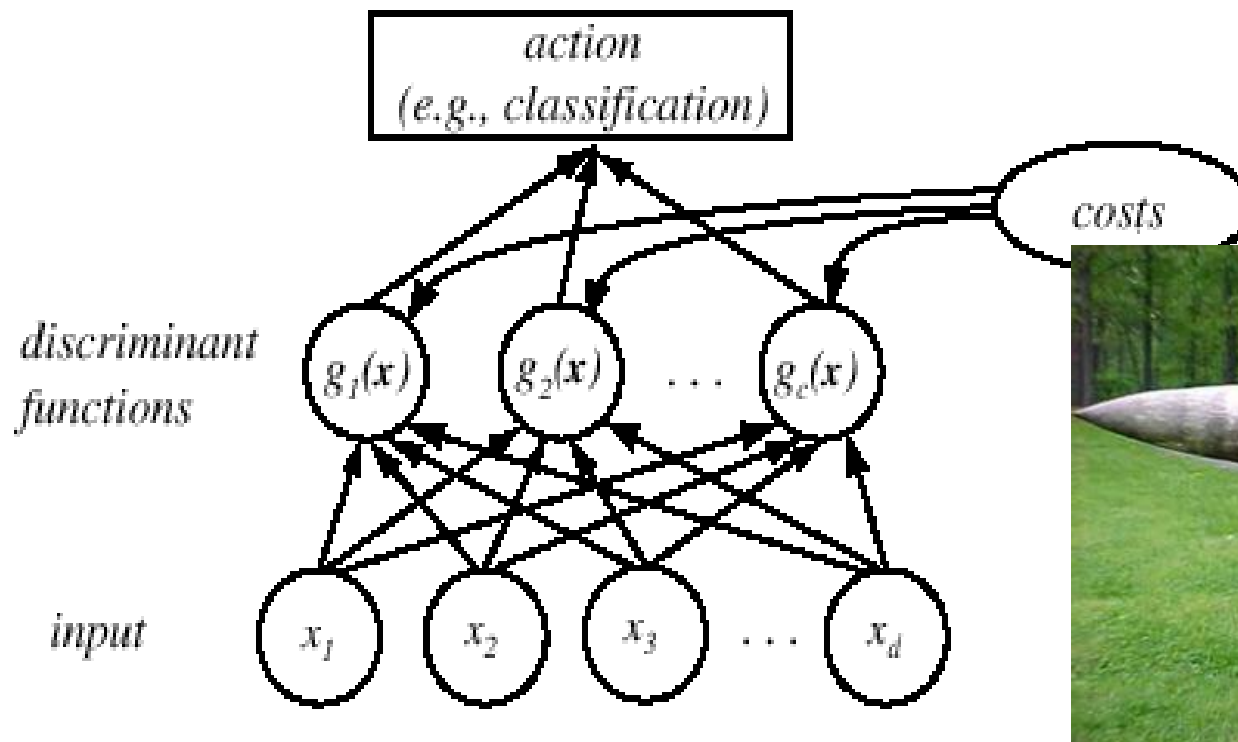


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction 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.

- For the minimum *risk* case

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

(max. discriminant corresponds to min. risk!)

- For the minimum error rate case, we take

$$g_i(x) = P(\omega_i | x)$$

(max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv p(x | \omega_i) P(\omega_i)$$

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

(ln: natural logarithm!)



- Feature space divided into c decision regions

if $g_i(x) > g_j(x) \forall j \neq i$ then x is in R_i

(R_i means to assign x to ω_i)

- The two-category case

- A classifier is a “dichotomizer” that has two discriminant functions g_1 and g_2

Let $g(x) \equiv g_1(x) - g_2(x)$

Decide ω_1 if $g(x) > 0$; Otherwise decide ω_2



- The computation of $g(x)$

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

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



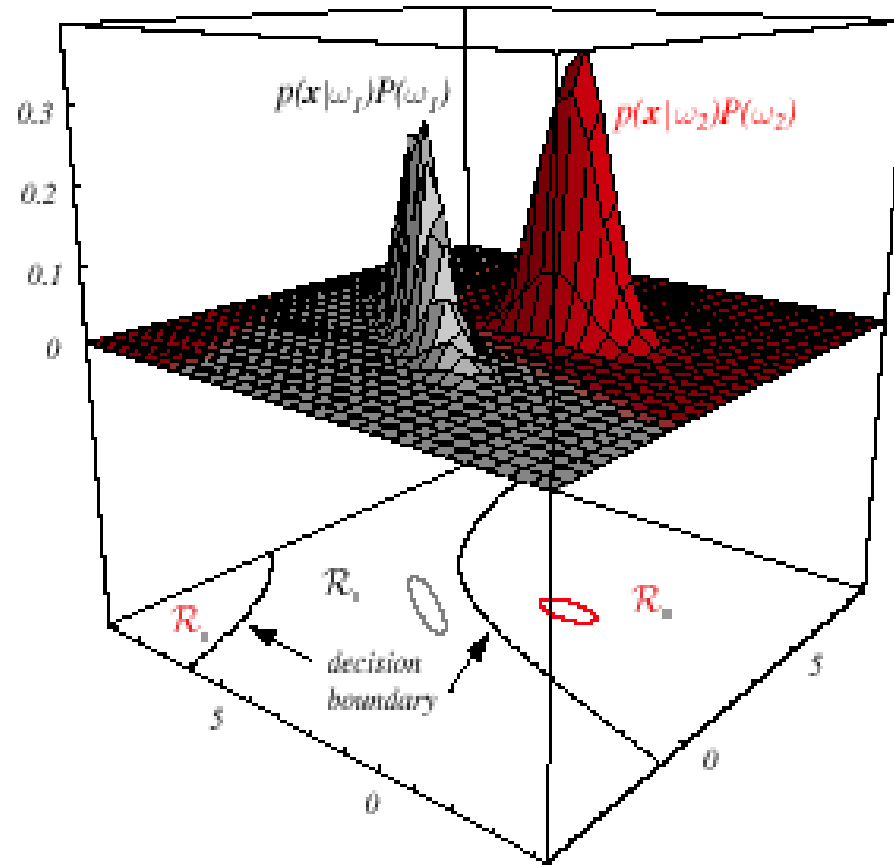
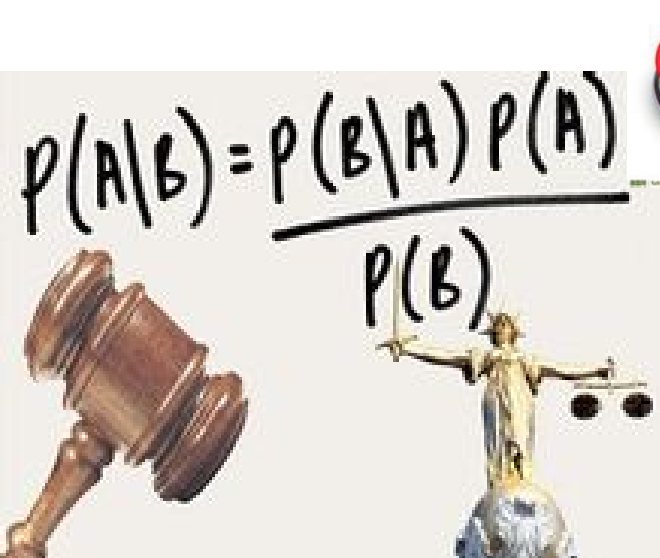


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region \mathcal{R}_2 is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

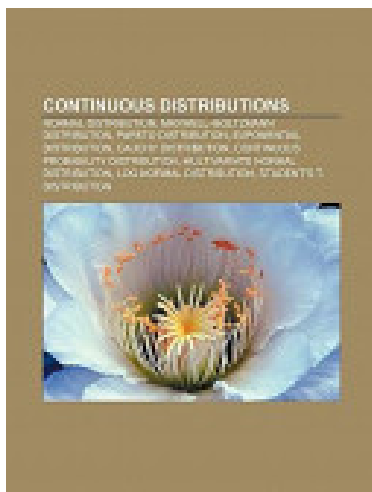
Difficulties of Bayesian decision

- It's not easy to obtain the probability (prior and conditional probability)





- To assume that the conditional probability satisfies the normal distributions is a basic way.
- The real-world verifies that the normal distributions is really consistent with the true distribution



2.5 The Normal Density

■ Univariate density

- Continuous density
- A lot of processes are asymptotically Gaussian
- Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

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

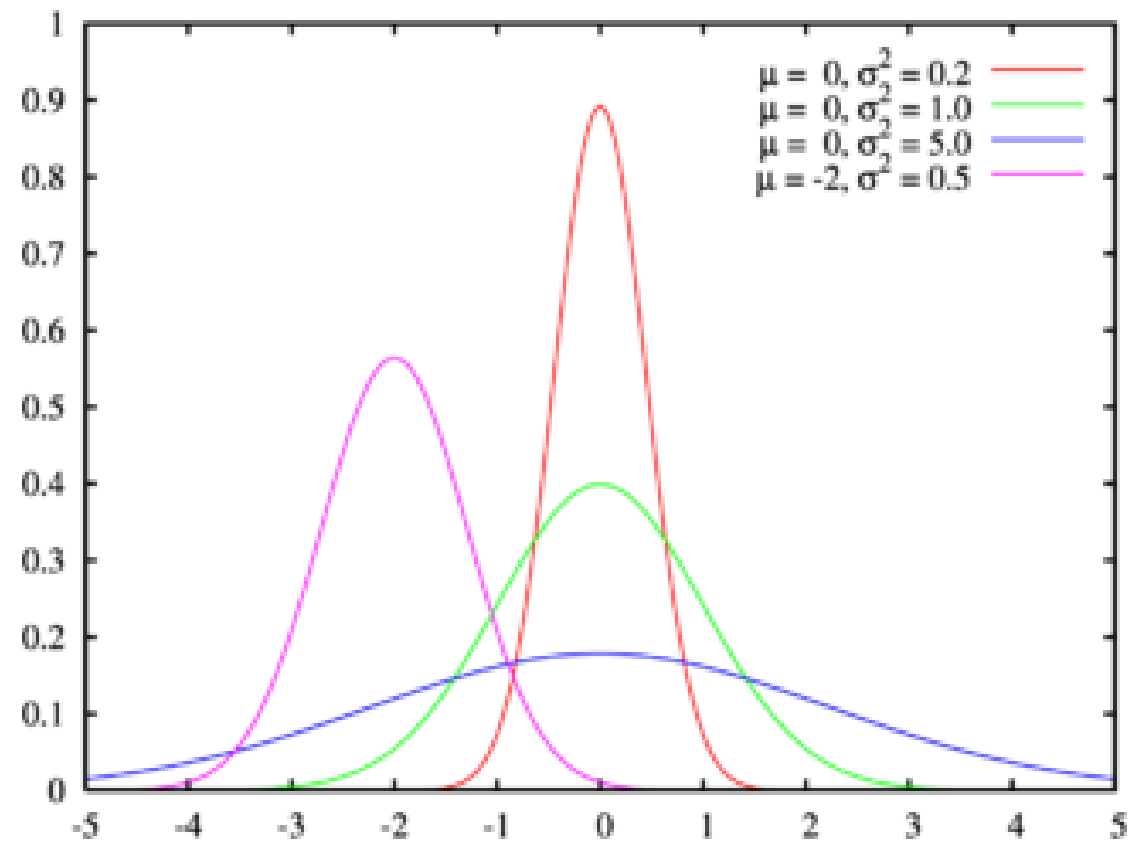
Where:

μ = mean (or expected value) of x

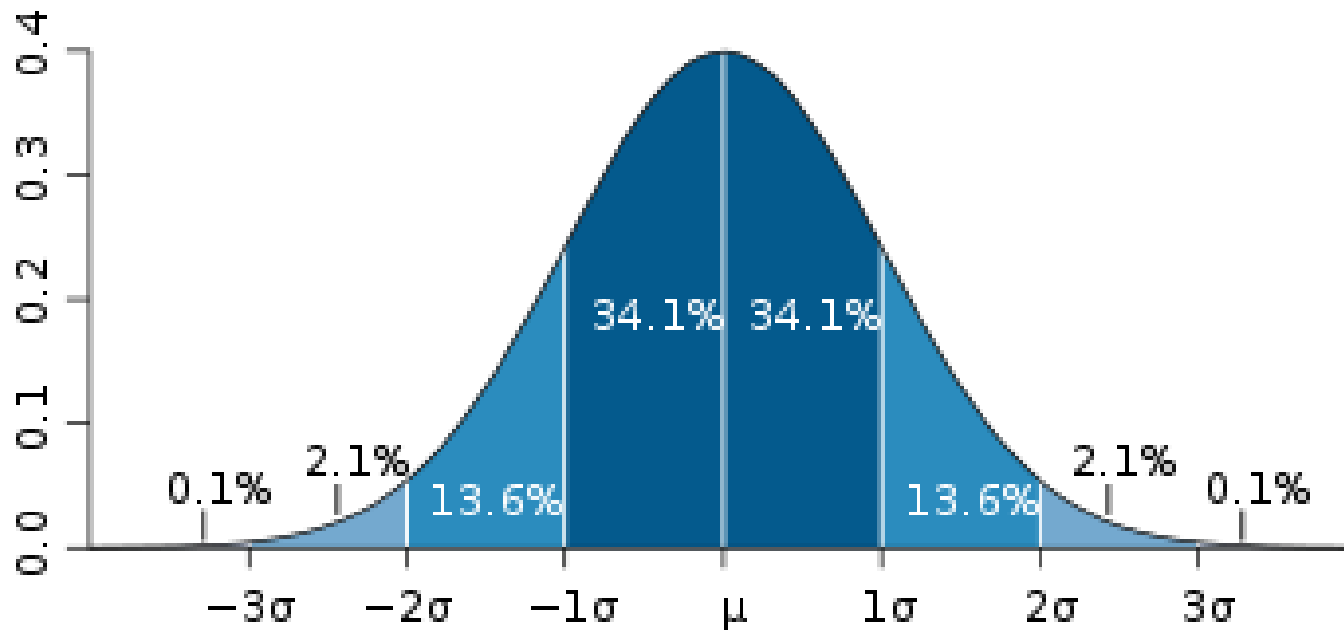
σ^2 = expected squared deviation or variance



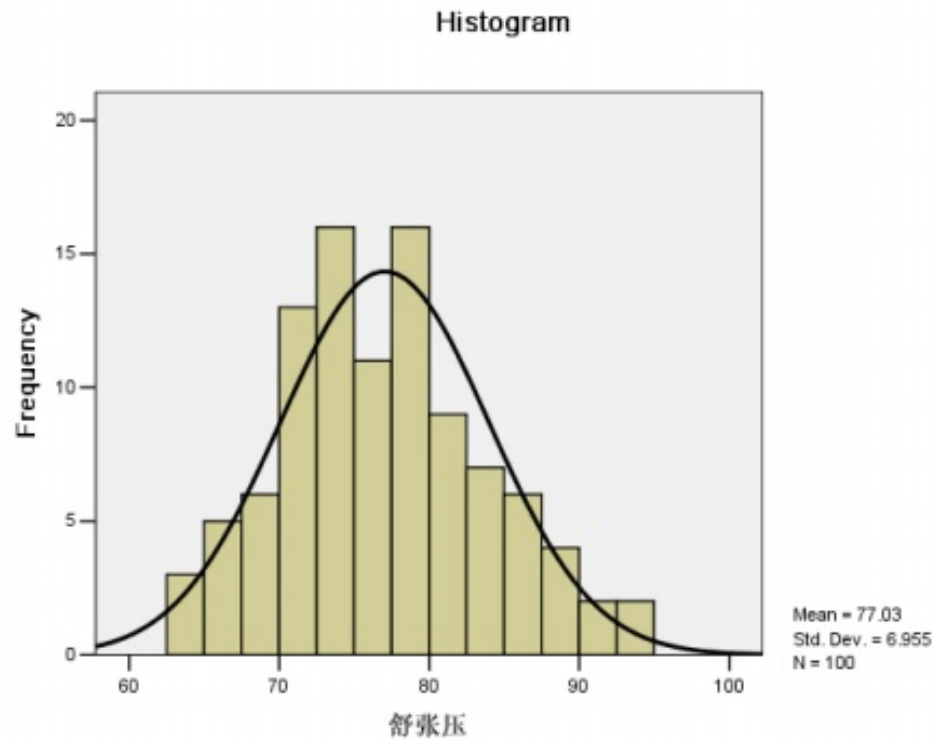
Examples of normal distributions



The 68-95-99.7 rule



Examples in the real world



Multivariate density

□ Multivariate normal density in d dimensions is:

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

where:


$\mathbf{x} = (x_1, x_2, \dots, x_d)^t$ (t stands for the transpose vector form)

$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_d)^t$ mean vector

$\Sigma = d \times d$ covariance matrix

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




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

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

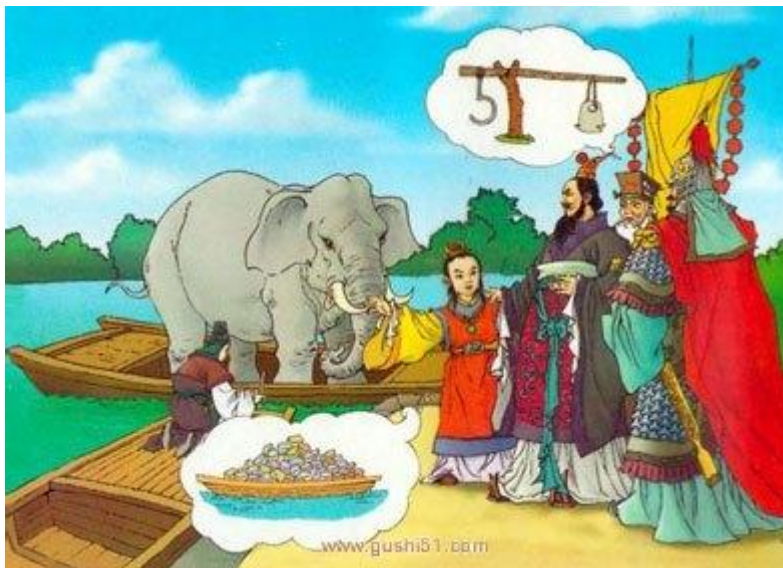
$$\mu_i = \mathcal{E}[x_i]$$

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

- Mahalanobis distance from x to μ

$$r^2 = (x - \mu)^t \Sigma^{-1} (x - \mu)$$

Mahalanobis distance is widely used in pattern recognition

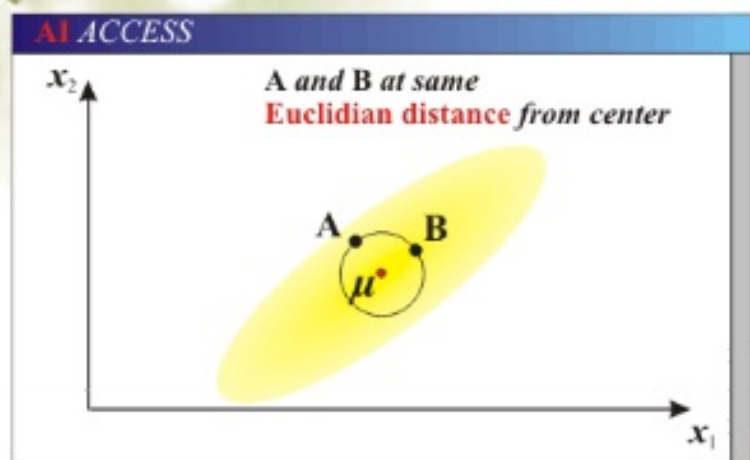


❖ 我们熟悉的欧氏距离虽然很有用，但在解决多元数据的分析问题时，就显示出了它的不足之处。一是它没有考虑到总体的变异对“距离”远近的影响，显然一个变异程度大的总体可能与更多样品近些，即使它们的欧几里得距离不一定最近；另外，欧几里得距离受变量的量纲影响，这对多元数据的处理是不利的。

马氏距离优点

它不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关；由标准化数据和中心化数据(即原始数据与均值之差)计算出的二点之间的马氏距离相同。马氏距离还可以排除变量之间的相关性的干扰。

欧氏距离与马氏距离的区别与联系



❖ 欧式距离

❖ 马氏距离

