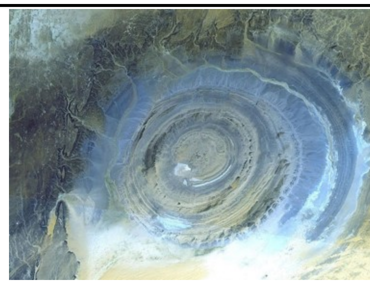


第三讲 贝叶斯决策函数

主讲：王 星
办公电话:86-10-82500167
电子邮箱:wangxingwisdom@126.com

内容大纲

- 引言和基本概念
- 0-1损失函数下连续变量的分类决策
- 正态分布下的线性判别分界面和二次判别分界面
- 朴素Bayes
- 离散变量的Bayes网络

三种主要的分类函数求解方法

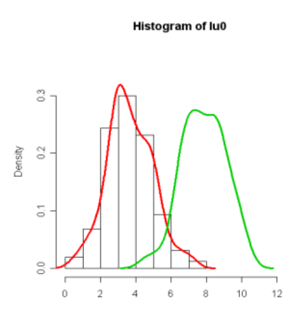
- 1. 经验风险最小化原则
  选择一个分类函数集G,搜索$\hat{g}$使得经验函数最小化.
- 2.回归: 找到函数 $\hat{\eta}(x)$定义

$$g(x)=\begin{cases}1 & \hat{\eta}(x)>1/2\\ 0 & others\end{cases}$$

- 3.密度估计

第一节 基本概念

- The sea bass/salmon example
- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2
- $P(x \mid \omega_1)$ and $P(x \mid \omega_2)$ describe the difference in lightness between populations of sea bass and salmon



二分类问题

- Credit scoring: Inputs are income and savings.
  Output is low-risk vs high-risk
- Input: $\boldsymbol{x} = [x_1,x_2]^T$,Output: C Î {0,1}
- Prediction:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1,x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

or equivalently

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1,x_2) > P(C=0 \mid x_1,x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

Bayes' Rule

- Posterior, likelihood, evidence

posterior *likelihood* *prior*

$$P(\omega_j \mid x) = P(x \mid \omega_j) \, . \, P(\omega_j) \, / \, P(x)$$

evidence

– Where in case of two categories

$$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j) P(\omega_j)$$

– Posterior = (Likelihood. Prior) / Evidence

- Decision given the posterior probabilities

X is an observation for which:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ True state of nature = ω_1
if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ True state of nature = ω_2

Therefore:

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

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

- Minimizing the probability of error

- Decide $\omega 1$ *if* $P(\omega 1 \mid x) > P(\omega 2 \mid x)$;
otherwise decide $\omega 2$

- Therefore:

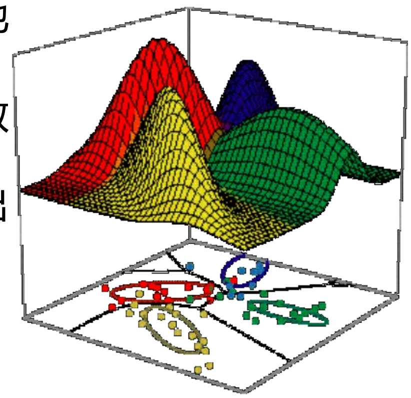
$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$
$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

$$\Updownarrow$$

$$P(error \mid x) = \min\, [P(\omega 1 \mid x), P(\omega 2 \mid x)]$$

推广到一般情况 Generalization

- 观测特征可能会多于一个;
- 判断的类别可能多于两个；
- 可能存在不同于判别的其他决策行为;
- 通过引入更一般的损失函数来替代概率误差:损失函数测量了依照每个行为所做出每次决策的代价.



多类的情况

- 状态集: Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or “categories”)
- 决策集: Let $\{\alpha_1, \alpha_2, \ldots, \alpha_t\}$ be the set of t possible actions
- Let $\ell(\alpha_i, \omega_j)$ be the loss incurred for taking

  action α_i when the state of nature is ω_j
- For any x define conditional risk:

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} l(\alpha_i, \omega_j) P(\omega_j \mid x) \quad for\ i = 1, \ldots, t$$

决 策 论(补充)

决策是对不确定的方案做决定，它的基本特点是通过计算后果来做选择。（包括状态的选择、参数的选择到模型的选择）

可能情况	自己生产 δ_1	转让专利 δ_2
	-80	-40

比如：新药品的生产，有两种方案：自己生产，或转让专利。损失根据不同的销售前景有不同。

可能情况	自己生产 δ_1	转让专利 δ_2
销售好 θ_1	-80	$\boxed{-40}$
销售一般 θ_2	-20	$\boxed{-10}$
销售较差 θ_3	$\boxed{10}$	-5

极小极大原则之下，最好的决策是：转让专利。

决策论的几个基本概念

- **参数集**：参数的所有自然可能的不同的状态 $\Theta=\{\theta\}$;
- **行动集**：所有可能的决策结果 $A=\{a\}$;

定义：损失函数 对于一个行动 a 和参数 θ，评价不同行动或决策优劣的函数 $l(\theta,a)$，一般的损失函数有绝对损失，平方损失，线性损失等。

For each in a set of actions $a \in \mathcal{A}$, if the parameter is θ, a *loss* $L(a,\theta)$ is associated with choosing action a.

The *risk* is the expected loss:

$$R = \int_{\Theta} L(a,\theta)\, dF(\theta)$$

and one chooses the action that minimizes the risk.

统计决策论的基本概念

一个好的决策应该在参数空间上一致地好，
模型选择应该在函数空间上一致的好

最小期望损失原则：

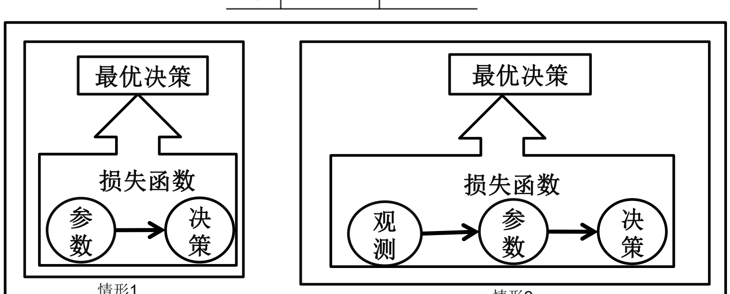
$\pi(\theta_1)=0.2, \pi(\theta_2)=0.5, \pi(\theta_3)=0.3$，计算期望损失

$$R(\delta_1)=-80\times 0.2-20\times 0.5+10\times 0.3=\boxed{-23}$$

$$R(\delta_2)=-40\times 0.2-10\times 0.5-5\times 0.3=-14.5$$

例子（继续）：假如公司做了一项研究，收集了一些信息，比如调查了一些专家，他们认为公司的品牌知名度x的取值和θ参数 有关系，做了一些分析，根据这些分析结果如何决策？ $p(x\,|\,\theta)$

	$x=0$	$x=1$
θ_1	0.3	0.7
θ_2	0.6	0.4
θ_3	0.7	0.3



两种情形下的最优决策

情形 2. 状态不可直接观测，但其可能性可借助其他更易观测的变量推断出来，将 $P(\theta)$ 由分布 $P(\theta|x)$ 代替就产生了条件风险的概念。

【条件风险的定义】：参数状态集设为 $\Theta=\{\theta_1, \theta_2, ...\}$，由观测集 X 决定，$D=\{\delta\}$ 是决策集，损失函数是 $L(\theta, \delta)$，条件风险 $R(\delta|x)$ 定义为

$$R(\delta|x)=\int_{\Theta} l(\theta,\delta)dP(\theta|x)=\int_{(\Theta,X)}\frac{1}{p(x)}l(\theta,\delta)dP(x|\theta)P(\theta)$$

最优决策如下：

$$\delta^*|x=\arg\min_{\delta\in\Delta} R(\delta|x)$$

•[例]: 两类分类问题:

- State: $\{\omega_1, \omega_2\}$,
- Action :

  α_1 : *deciding* ω_1

  α_2 : *deciding* ω_2

- Loss: $\ell_{ij} = \ell(\alpha_i, \omega_j)$:loss incurred for deciding α_i when the true state of nature is ω_j

- Conditional risk:

$$R(\alpha_1|x) = \ell_{11}P(\omega_1|x) + \ell_{12}P(\omega_2|x)$$

$$R(\alpha_2|x) = \ell_{21}P(\omega_1|x) + \ell_{22}P(\omega_2|x)$$

Our decision rule is the following:

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

action α_1: "decide ω_1" is taken

Remember:

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

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

This results in the equivalent rule :

decide ω_1 if:

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

and decide ω_2 otherwise

The preceding rule is equivalent to the following rule:

$$if\ \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \frac{l_{12} - l_{22}}{l_{21} - l_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

结论: 贝叶斯决策规则可以解释成如果似然比超过某个不依赖于观测值x的阈值,那么判断为ω_1 .

课堂案例

Select the optimal decision where:

= $\{\omega_1, \omega_2\}$

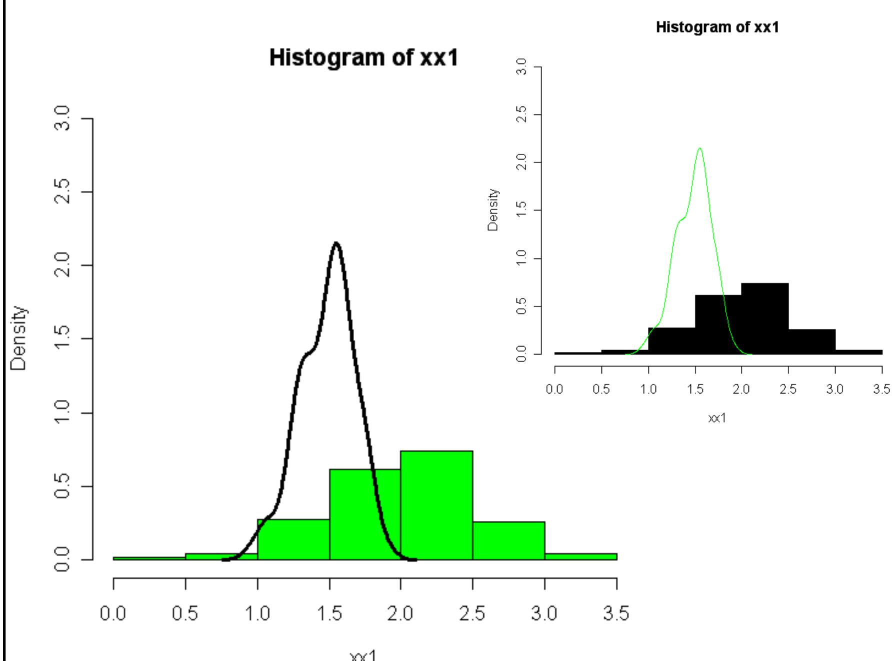
$P(x \mid \omega_1)$ → N(2, 0.5^2) (Normal distribution)

$P(x \mid \omega_2)$ → N(1.5, 0.2^2)

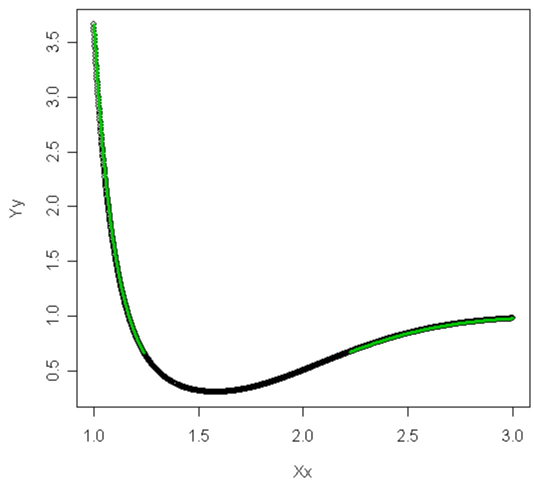
$P(\omega_1) = 2/3$

$P(\omega_2) = 1/3$

$$l = \begin{bmatrix} -1 & 2 \\ 3 & -4 \end{bmatrix} \qquad l_2 = \begin{bmatrix} 0 & 1 \\ 3 & 0 \end{bmatrix}$$



Decision Boundary



20201113课下作业1

在R的library(ISLR)中有一个棒球球员薪资与技术能力的数据Hitters，根据薪水高低标记为{ω_1, ω_2}，高于3/4的标记为ω_1, 低于3/4的标记为ω_2, x选为Hits

根据题意设置如下分布，Select the optimal decision where:

$P(x \mid \omega_1)$ ⟹ $N(\mu_1, \sigma_1^2)$(Normal distribution)

$P(x \mid \omega_2)$ ⟹ $N(\mu_2, \sigma_2^2)$

$P(\omega_1)$,$P(\omega_2)$，随机选出 2/3比例的数据作为训练数据，其余数据作为测试数据，结合如下损失函数和参数的极大似然估计给出决策区域，给出决策边界，讨论错误率。根据实际决策场景自定义损失函数的合理代价，根据自定义损失函数重新讨论决策。

$$l_2 = \begin{bmatrix} 0 & 1 \\ 3 & 0 \end{bmatrix}$$

$$l_2 = \begin{bmatrix} & \\ & \end{bmatrix}$$

20201113课下作业2

在R的library(ISLR)中有一个棒球球员薪资与技术能力的数据Hitters，根据薪水高低标记为$\{\omega_1, \omega_2\}$，高于3/4的标记为ω_1, 低于3/4的标记为ω_2, x选为Hits

根据题意设置如下分布，Select the optimal decision where:

$P(x \mid \omega_1)$ ⇒ $p1(x)$

$P(x \mid \omega_2)$ ⇒ $p2(x)$

$P(\omega_1),P(\omega_2)$，不做正态分布假设，用非参数密度估计各自的密度，选择合适的带宽，结合鲜艳分布，随机选出 2/3比例的数据作为训练数据，其余数据作为测试数据，结合如下损失函数和参数的极大似然估计给出决策区域，给出决策边界，讨论错误率。与1的结果进行比较

$$l_2 = \begin{bmatrix} 0 & 1 \\ 3 & 0 \end{bmatrix}$$

20201113课下作业3

根据上一个作业中薪水高低的标记方式 $\{\omega_1, \omega_2\}$，高于3/4的标记为ω_1, 低于3/4的标记为ω_2, x选为Hits，CAtBat,CHits

根据题意设置 $P(\omega_1),P(\omega_2)$，随机选出 2/3比例的数据作为训练数据，其余数据作为测试数据，给出LDA的判别结果，画出判别图

第二节 0-1损失函数下的分类决策

- Introduction of the zero-one loss function:

$$l(\alpha_i,\omega_j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases} \qquad i,j=1,\ldots,c$$

- Therefore, the conditional risk is:

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

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

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

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

$$Let\ \frac{l_{12}-l_{22}}{l_{21}-l_{11}}.\frac{P(\omega_2)}{P(\omega_1)}>\theta_l \text{ then decide } \omega_1 \text{ if : } \frac{P(x\mid\omega_1)}{P(x\mid\omega_2)}>\theta_l$$

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

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

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

$$if\ \ l=\begin{pmatrix}0 & 2\\ 1 & 0\end{pmatrix} then\ \ \theta_l=\frac{2P(\omega_2)}{P(\omega_1)}=\theta_b$$

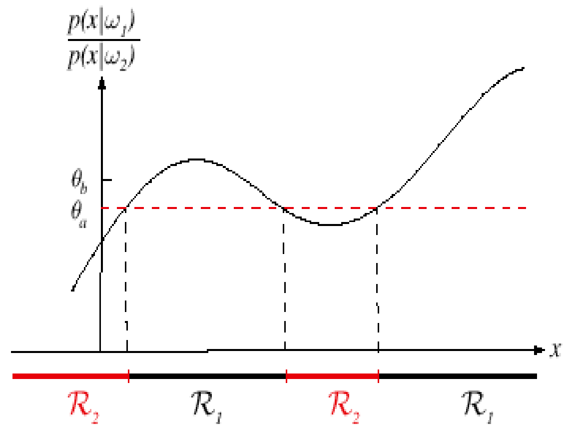
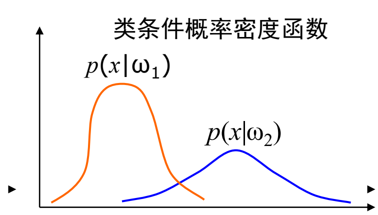
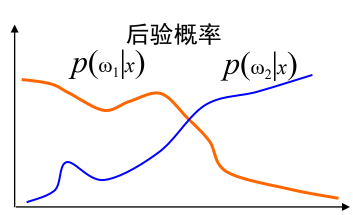


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, 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.

最优分类器

- 最优分类器?
- 已有:类条件概率密度函数
  - This is called the class-conditional probability describing the probability of occurrence of the features on category.
- 欲求:后验概率
  - make a decision that maximize the conditional probability of the object, given certain feature measurements.
  - Also called posterior probability function.





- *Let* $g_i(x) = - R(\alpha_i \mid x)$
  (max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take

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

  (max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$$

判别函数可以替换为任意一个单调增函数f(g(.))

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

(ln: natural logarithm!)

- Feature space divided into c decision regions(判决空间被分成c个判决区域)

$$\text{if } g_i(x) > g_j(x)\ \forall j \neq i \text{ then } x \text{ is in } \mathcal{R}_i$$

($\mathcal{R}_i$ means assign x to ω_i)

二分类问题

  – 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

- Multivariate density
  - Multivariate normal density in d dimensions is:

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

where:

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

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

$\Sigma = d*d$ covariance matrix

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

Discriminant Functions for the Normal Density二次判别函数

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

- The minimum error-rate classification can be achieved by the discriminant function

Quadratic Discriminant Analysis

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

QDA

- In Multinormal case,If $p(x\mid \omega_i)\sim N(\mu_i,\Sigma_i)$

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

$$r_i^2(x) = \frac{1}{2}(x-\mu_i)^t \sum_i^{-1}(x-\mu_i)$$

$$f^*(x) = \begin{cases} 1 & r_1^2(x) < r_0^2(x) + 2\ln\dfrac{\pi_1}{\pi_0} + \ln\dfrac{|\Sigma_0|}{|\Sigma_1|} \\ 0 & others \end{cases}$$

推广到多类

$Y=\{1,\ldots,c\}$,若$p(x|y=j)$是正态的,则$Bayes\ decition\ function$:

$$f^*(x) = \arg\max f_i(x)$$

$$f_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln(\pi_i)$$

用估计表示

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

- 定理: 若 p(x| ω_i)~$N(\mu_i,\Sigma_i)$

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

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{1.i}, \hat{\Sigma}_i = \frac{1}{n_i}\sum_{i=1}^{n_i}(X_{1.i}-\hat{\mu}_i)(X_{1.i}-\hat{\mu}_i)^T$$

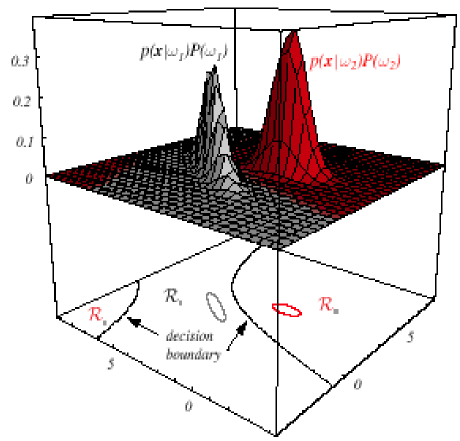


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

情形1. $\Sigma_i = \sigma^2.I$

(I stands for the identity matrix)

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

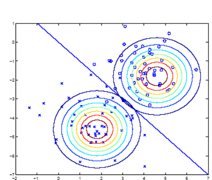
$$\| x-\mu_i \|^2 = (x-\mu_i)^t(x-\mu_i) = x^t x - 2\mu_i^t x + \mu_i^t \mu_i$$

$g_i(x) = w_i^t x + w_{i0}$ (linear discriminant function)

where :

$$w_i = \frac{\mu_i}{\sigma^2};\ w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t \mu_i + \ln P(\omega_i)$$

(w_{i0} is called the threshold for the ***i***th category!)



- A classifier that uses linear discriminant functions is called "a linear machine"
  (使用线性判别函数的分类模型,称为线性机)

- The decision surfaces for a linear machine are pieces of hyperplanes defined by:
  (线性判别机是超平面)

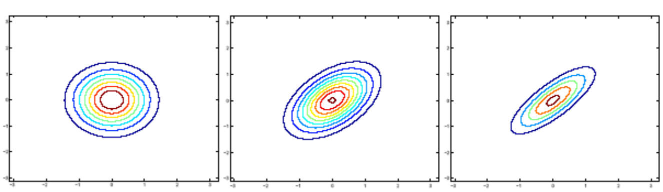
$$g_i(x) = g_j(x)$$

$$w^t(x - x_0) = 0$$

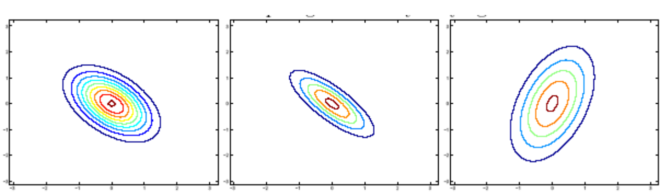
where :

$$w = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\| \mu_i - \mu_j \|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

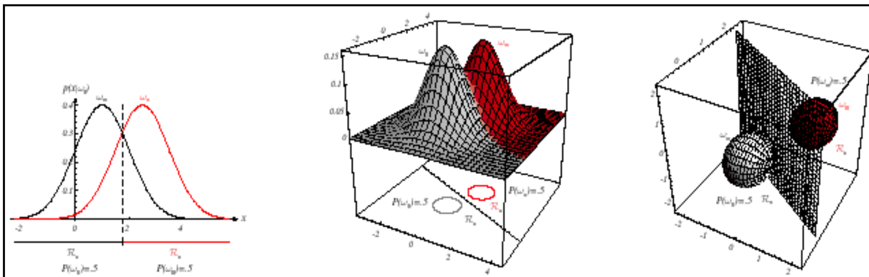


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

– The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

always orthogonal to the line linking the means!

$$\textit{if} \;\; P(\omega_i) = P(\omega_j) \;\; \text{then} \;\; x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

先验概率相等和不等的分界面