模式识别阅读笔记

为了不浪费时间，这个笔记我就不写公式了。只提供给一些概念性的知识供回忆。

I 概论

模式识别。什么是模式，模式(pattern),是一种较为抽象的东西。举个例子，比喻猫狗，你为什么能判断猫是猫，狗是狗，一般人们很难说清，这是种难以形容的东西。又比如你怎么判断一个东西是照片还是一副画，是什么类型的画，这些都难以表达，也很难用什么数学函数表达式直接表示出来。这些东西就叫做模式，对于人来说，识别不同的模式是一种简单的事请，但对于计算机来说并不简单，但计算机擅长计算这种人并不擅长的东西。所以我们的目的不是让计算机去模仿人脑，而是通过数学的方式让计算机去理解人的识别。

模式识别有分为**有监督学习**和**无监督学习**:

- 有监督学习比较著名的就是分类问题，也是当今模式识别领域运用最多的。监督指的是有样本也有标签。例如图像猫狗识别，能够给你很多张图片并告诉你猫还是狗，当模型训练出来后就可以做分类问题了。
- 无监督学习更多的都是聚类问题，但聚类完之后得对结果进行解释。

在模式识别的分类问题中中通常有以下几个步骤:

- 预处理
- 特征提取
- 输入分类器进行训练
- 将训练好的模型用于预测

II 统计决策方法

这里我们以贝叶斯的方法来进行，**贝叶斯决策理论称作统计决策理论。**

2.1 贝叶斯公式

学过概率论的应该都对贝叶斯公式有深刻的印象，这是一种从果到因推断出从因到果的过程

$$P\left(A_i \mid B\right)=\frac{P\left(B \mid A_i\right) P\left(A_i\right)}{\sum_j P\left(B \mid A_j\right) P\left(A_j\right)}$$

在模式识别中，贝叶斯公式主要是下面的形式

$$P\left(\omega_i \mid \boldsymbol{x}\right)=\frac{P\left(\boldsymbol{x} \mid \omega_i\right) P\left(\omega_i\right)}{\sum_j P\left(\boldsymbol{x} \mid \omega_j\right) P\left(\omega_j\right)}$$

其中$\boldsymbol{x}$代表输入，是一个多维特征向量,ω_i代表第几个分类。$P\left(\omega_j\right)$叫做**先验概率**，$P\left(\omega_i \mid \boldsymbol{x}\right)$叫做**后验概率**。上面这个显然是一种条件概率，等式左边其实就是可以看出当输入是$\boldsymbol{x}$时，样本属于ω_i概率。通过这个概率，我们可以引出以下几种决策的方式。

2.2 最小错误率贝叶斯决策

任何一个样本都可能会出错，对于某个样本的**错误率**为$p(e|x)$，显然对于一个二分类问题，样本x出错的概率为$P(w_2|x)$（若决策$x \in \omega_1$）或者$P(w_1|x)$（若决策$x \in \omega_2$）。对于多分类问题错误率定义为

$$P(e)=\int P(e|x)p(x)dx$$

所谓最小错误率决策，就是使上述的错误率最小化，也就是使后验概率最大的决策。也就是我们最常见的，选择使$P\left(\omega_i \mid \boldsymbol{x}\right)$最大的$\omega_i$为最佳决策。

但这个规则在二分类问题的情况下还可以整理一下

$$\text{若 } l(x)=\frac{p\left(\boldsymbol{x} \mid \omega_1\right)}{p\left(\boldsymbol{x} \mid \omega_2\right)} \gtrless \lambda=\frac{P\left(\omega_2\right)}{P\left(\omega_1\right)}, \quad \text{则 } \boldsymbol{x} \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$$

上面这个公式仔细想一下就行，$p\left(\boldsymbol{x} \mid \omega_i\right)$反映了某一类中观察到特征值$x$的相对可能性，所以被称为似然度,主要λ叫做**似然比(likelihood ratio)**。

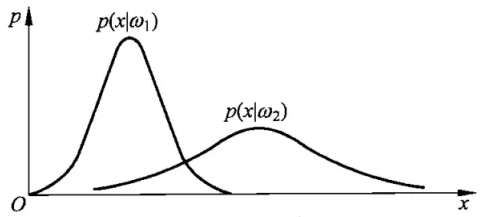


图 2-1　类条件概率密度

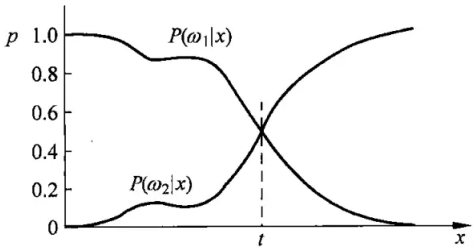


图 2-2　后验概率

图2.2中的虚线就被叫做**决策面**或者分类面。

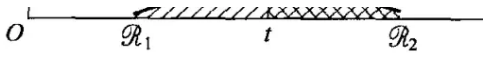


图 2-3　错误率

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

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

试对该细胞 x 进行分类。

解：利用贝叶斯公式，分别计算出 ω_1 及 ω_2 的后验概率

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

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

根据贝叶斯决策规则式(2-8)，因为

$$P(\omega_1|\boldsymbol{x})=0.818>P(\omega_2|\boldsymbol{x})=0.182$$

2.3最小风险贝叶斯决策

同上面不一样的是，我们同样使用贝叶斯公式算出了后验概率$P(\omega_i|x)$，但此时我们不是单纯的选择最大那个了。在某一些，情况下，我们想要宁可错杀不可放过。比如检测疾病，或者故障维修的时候。所以我们需要评估假如选错了会带来的风险，并将其量化。

对于输入x，假如由k种决策，包含把它分到c类中的一个，或者说判断其不属于任何一类。定义一个决策空间：

$$\mathscr{A}=\{\alpha_1,\alpha_2,\cdots,\alpha_k\}$$

同时定义**风险函数**,设对于实际状态为 ω_j的向量 x 采取决策 α_i所带来的损失为

$$\lambda\left(\alpha_i,\omega_j\right), \quad i=1,\cdots,k, \quad j=1,\cdots,c$$

对于某个样本 $\boldsymbol{x}$, 它属于各个状态的后验概率是 $P(\omega_j \mid \boldsymbol{x}), j=1, \cdots, c$, 对它采取决策 α_i, $i=1, \cdots, k$ 的期望损失是

$$R(\alpha_i \mid \boldsymbol{x})=E[\lambda(\alpha_i, \omega_j) \mid \boldsymbol{x}]=\sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j \mid \boldsymbol{x}), \quad i=1, \cdots, k$$

我们的目的就是要使期望损失最小,使用以下决策

$$\text{若 } \lambda_{11} P(\omega_1 \mid \boldsymbol{x})+\lambda_{12} P(\omega_2 \mid \boldsymbol{x}) \lessgtr \lambda_{21} P(\omega_1 \mid \boldsymbol{x})+\lambda_{22} P(\omega_2 \mid \boldsymbol{x}), \text{则 } \boldsymbol{x} \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$$

其中, $\lambda_{12}=\lambda(\alpha_1, \omega_2)$ 是把属于第 2 类的样本分为第 1 类时的损失, $\lambda_{21}=\lambda(\alpha_2, \omega_1)$ 是把属于第 1 类的样本分为第 2 类时的损失, $\lambda_{11}=\lambda(\alpha_1, \omega_1)$、$\lambda_{22}=\lambda(\alpha_2, \omega_2)$ 是决策正确 (把第 1 类决策为 第 1 类和把第 2 类决策为第 2 类) 时的损失。通常, $\lambda_{11}=\lambda_{12}=0$; 不失一般性, 我们可以假设 $\lambda_{11}<\lambda_{21}$,$\lambda_{22}<\lambda_{12}$ 。

实际上很好理解，对于多分类来说，使用不同决策的风险函数与对应的后验概率相乘，计算当做某一种决策所承担的风险，然后选择风险最小的一种就行。

例 2.2 在例 2.1 给出条件的基础上,利用表 2-2 的决策表,按最小风险贝叶斯决策进行分类。

表 2-2 例 2.2 的决策表

决策	状 态	
	ω_1	ω_2
α_1	0	6
α_2	1	0

解：已知条件为

$$\begin{aligned}
&P(\omega_1)=0.9, && P(\omega_2)=0.1 \\
&P(\boldsymbol{x} \mid \omega_1)=0.2, && P(\boldsymbol{x} \mid \omega_2)=0.4 \\
&\lambda_{11}=0, && \lambda_{12}=6 \\
&\lambda_{21}=0, && \lambda_{22}=0
\end{aligned}$$

根据例 2.1 的计算结果可知后验概率为

$$P(\omega_1 \mid \boldsymbol{x})=0.818, \quad P(\omega_2 \mid \boldsymbol{x})=0.182$$

再按式 (2-26) 计算出条件风险

$$\begin{aligned}
&R(\alpha_1 \mid \boldsymbol{x})=\sum_{j=1}^{2} \lambda_{1j} P(\omega_j \mid \boldsymbol{x})=\lambda_{12} P(\omega_2 \mid \boldsymbol{x})=1.092 \\
&R(\alpha_2 \mid \boldsymbol{x})=\lambda_{21} P(\omega_1 \mid \boldsymbol{x})=0.818 \\
&R(\alpha_1 \mid \boldsymbol{x})>R(\alpha_2 \mid \boldsymbol{x})
\end{aligned}$$

即决策为 ω_2 的条件风险小于决策为 ω_1 的条件风险, 因此我们采取决策行动 α_2, 即判断待识 别的细胞 x 为 ω_2 类一异常细胞。

2.4 两类错误率、Neyman-Pearson决策与ROC

下面讨论一些概念性的问题。在医学领域，人们经常会用阳性阴性代表两类，或者说正样本和负样本。正常的决策可能会有以下几种情况：

决策	状态	状态
	阳性	阴性
阳性	真阳性(TP)	假阳性(FP)
阴性	假阴性(FN)	真阴性(TN)

上面需要牢记，同时定义**灵敏度**Sn为实际是阳性的样本中被检测为阳性的，**特异度**Sp为实际上是阴性的样本中被检测为阴性的。注意灵敏度越高特异度就越低。

$$\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$
$$\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

同时把所有实际上是阴性样本中被为阳性检测(假阳性)称为**第一类错误**(误报或虚报)，把所有阳性样本中被检测为阴性(假阴性)称为**第二类错误**(漏报)。显然对于病人来说，漏报是比误报还严重的，有些时候我们希望将第二类错误降低到一定程度时使第一类错误更低，但我觉得应该不会考所以我暂时不讲。

现在来看**ROC曲线**,几乎所有机器学习得评价中都会看见这个曲线。

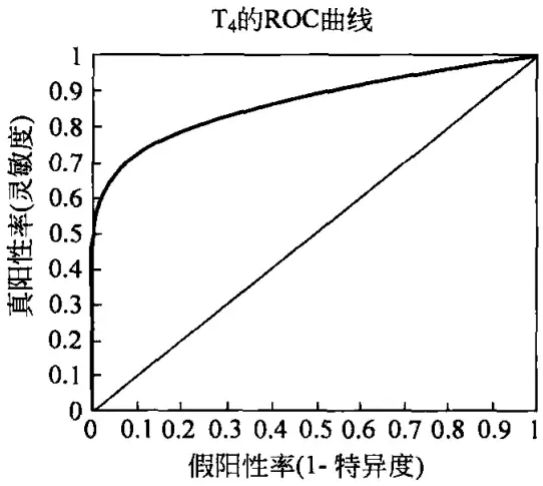


图 2-5　ROC 曲线

这个显然当ROC曲线要在对角线之上才是正常的，否则真阳性率比假阳性率还低就没有意义的。这个曲线是通过改变决策阈值来实现的。比如有(0,0)点这种全识别成阴性这种特殊情况。显然有，当这个ROC曲线越往上凸是越好的，或者说其曲线下的相对面积(AUC)越大越好。AUC就成了展现决策方法性能的判别。

2.5 正态分布的统计决策

下面这个部分我认为虽然很复杂但是很有意思。首先需要明确一个概念，在这一小节我们假设对于某一种分类ω_i假设其x是**服从正态分布**的。正态分布我们都知道有均值和方差，比如对于人脸识别来说，一张标准的大众脸就是均值，而其他人的不同长相就是方差。

单变量的正态分布很简单

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

有期望和标准差

$$\mu = E\{x\} = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x$$
$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \mathrm{d}x$$

那么多变量会稍稍复杂亿点点

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

式中: $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^{\mathrm{T}}$ 是 d 维列向量; $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_d]^{\mathrm{T}}$ 是 d 维均值向量; Σ 是 $d \times d$ 维协方程矩阵, Σ^{-1} 是 $\boldsymbol{\Sigma}$ 的逆矩阵, $|\boldsymbol{\Sigma}|$ 是 $\boldsymbol{\Sigma}$ 的行列式。不难证明,协方差矩阵总是对称非负定阵, 且可表示为

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1d}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 & \cdots & \sigma_{2d}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1d}^2 & \sigma_{2d}^2 & \cdots & \sigma_{dd}^2 \end{bmatrix}$$

接下来引出一个新的概念：等密度点轨迹的超椭球面。

正态分布抽取的样本大部分落在一个区域中，区域中心由μ决定，大小由$\boldsymbol{\Sigma}$决定。可以想象，其等密度的轨迹应该是一个超椭球面(可以把它想象成三维空间)，且由下面这个方程确定

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \text{常数}$$

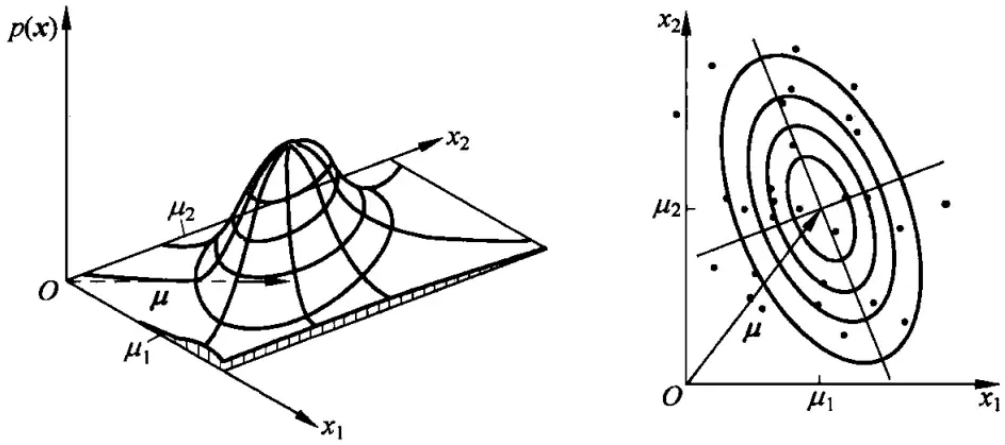


图 2-7　正态分布的等密度点的轨迹为超椭球面

并且将其到中心μ的距离定义为γ，这个距离被称作Mahalanobis距离(马氏距离)。

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \gamma^2$$

另外再复习一下，相互独立一定不相关，不相关不一定相互独立。

那么在这种正态分布的假设下，之前最小错误率贝叶斯决策和决策面的有关公式，可以写出一个判别函数为

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

决策面方程为

$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

化简为

$$
\begin{aligned}
&-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-(\boldsymbol{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]\\
&-\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_i|}{|\boldsymbol{\Sigma}_j|}+\ln\frac{P(\omega_i)}{P(\omega_j)}=0
\end{aligned}
$$

这个公式看起来很复杂，当考虑到实际问题的一些情况的时候，是可以化简的。

1. 第一种情况$\boldsymbol{\Sigma}_i=\sigma^2 I, i=1,2,\cdots,c$

在这种情况下，每类协方差矩阵都相等，且各个特征间都是相互独立的。而且方差还相等。虽然这种情况很理想，但对于某些相关性不强的情况我认为也可以近似一下。这种情况最为简单，相当于该类样本集中一个个**超球体**之内。

$$
\boldsymbol{\Sigma}_i=\begin{bmatrix}\sigma^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma^2\end{bmatrix}
$$

把上面的公式化简一下，忽略一些和类别无关的变量，最终就有结果

$$
g_i(\boldsymbol{x})=\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}_i)+\ln P(\omega_i)
$$

$$
(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}_i)=\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2=\sum_{i=1}^{d}(x_j-\mu_{ij})^2,\quad i=1,\cdots,c
$$

上面看起来我们已经转化为一个欧氏距离的平方了，特别是当**先验概率$P(w_j)$都相等时**，可以忽略末尾项，此时结果就仅仅和输入与分类器几类的均值向量的欧氏距离有关，x就属于$\min\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2$的类。

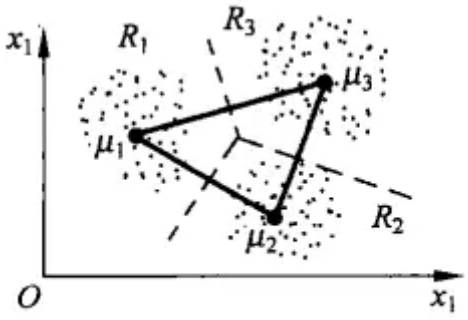
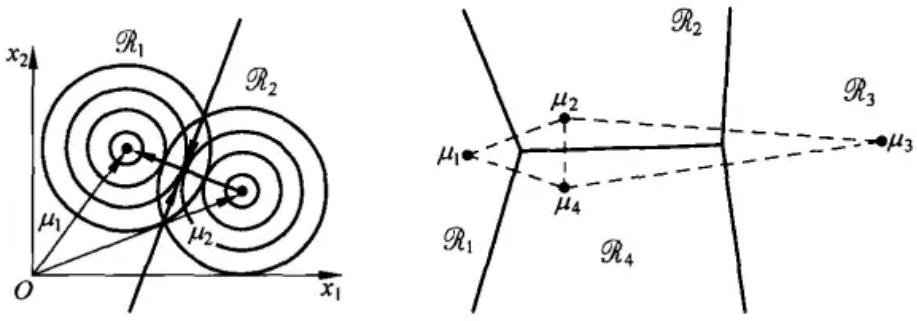


图 2-8　最小距离分类器



(a) 两类情况　　(b) 多类情况

图 2-9　正态分布且 $P(\omega_i)=P(\omega_j)$，$\boldsymbol{\Sigma}_i=\sigma^2\boldsymbol{I}$ 时的决策面

上面决策面是个超平面，显然通过两个中心的连线中点且与连线正交。上面的一些决策面实际上可以用方程表示出来：

$$\boldsymbol{w}^T(\boldsymbol{x}-\boldsymbol{x}_0)=0$$

$$\boldsymbol{w}=\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$$

$$\boldsymbol{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

当先验概率不相等，决策面与先验概率相等时决策面平行，只是向先验概率小的方向偏移。

2. 第二种情况：$\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}$

各类的协方差都相等，稍微复杂了一些，从几何上看，该样本集中于一个个超椭球体之内，且每一个椭球体都相等

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

在先验概率相等的情况下，忽略末尾项，则最终结果只和x到每类的均值点μ_i的马氏距离的平方γ^2相关。这实际上是一个线性判别函数，决策面是一个超平面，这里超平面的方程就不详细给出了，具体看书。

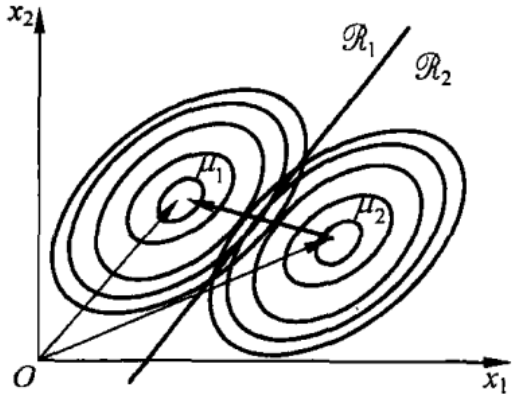


图 2-10　正态分布且 $P(\omega_i)=P(\omega_j)$，$\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}_j$ 时的决策面

3.第三种情况：各类协方差不相等

太复杂了，看看就好，我觉得不会考

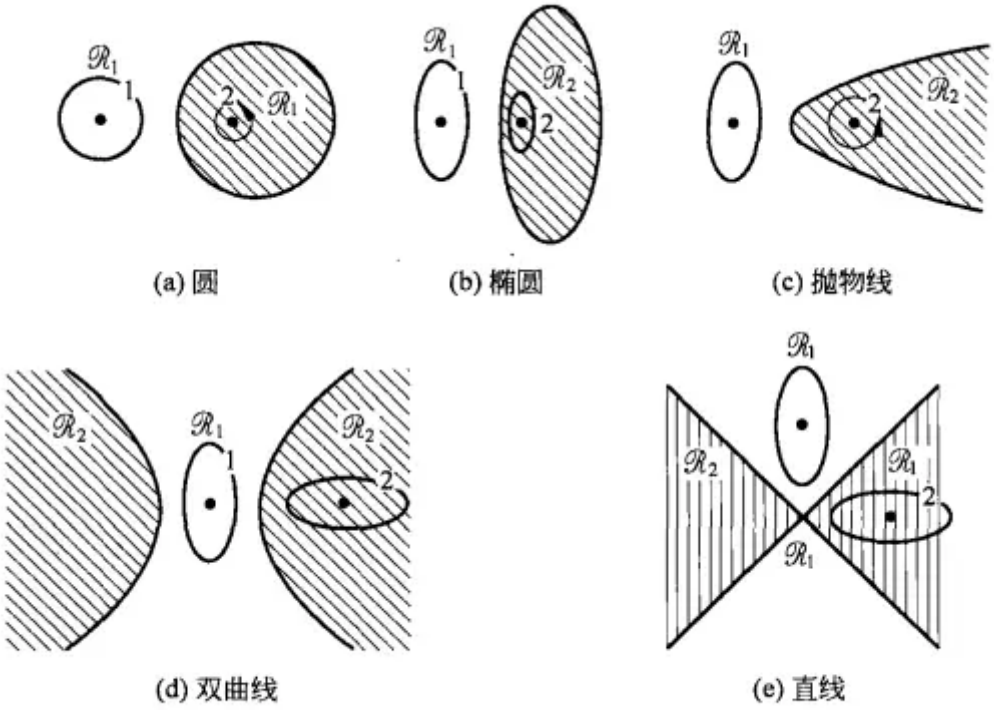


图 2-11　正态分布下的几种决策面形式

2.6 错误率计算

事实上大部分时候错误率都是很不好算的，但现在我们既然规定了正态分布，那似乎就可以开始计算错误率了。回顾一下最小错误率贝叶斯决策规则的负对数似然比。

$$h(\boldsymbol{x})=-\ln l(\boldsymbol{x})=-\ln p\left(\boldsymbol{x} \mid \omega_1\right)+\ln p\left(\boldsymbol{x} \mid \omega_2\right) \lessgtr \ln\left[\frac{P\left(\omega_1\right)}{P\left(\omega_2\right)}\right] \rightarrow x \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$$

$h(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的函数, $\boldsymbol{x}$ 是随机向量, 因此 $h(\boldsymbol{x})$ 是随机变量。我们记它的分布密度函数为 $p\left(h \mid \omega_1\right)$ 。 由于它是一维密度函数, 易于积分, 所以用它计算错误率有时较为方便。有

$$\begin{aligned}
&P_1(e)=\int_{\mathscr{S}_2} p\left(x \mid \omega_1\right) \mathrm{d} \boldsymbol{x}=\int_t^{\infty} p\left(h \mid \omega_1\right) \mathrm{d} h \\
&P_2(e)=\int_{\mathscr{S}_1} p\left(x \mid \omega_2\right) \mathrm{d} x=\int_{-\infty}^t p\left(h \mid \omega_2\right) \mathrm{d} h \\
&t=\ln \left[P\left(\omega_1\right) \mid P\left(\omega_2\right)\right]
\end{aligned}$$

这个还是很方便的

$$
\begin{aligned}
P_1(e) &= \int_t^{\infty} p\left(h \mid \omega_1\right) \mathrm{d}h \\
&= \int_t^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{h+\eta}{\sigma}\right)^2\right\} \mathrm{d}h \\
&= \int_t^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{h+\eta}{\sigma}\right)^2\right\} \mathrm{d}\left(\frac{h+\eta}{\sigma}\right) \\
&= \int_{\frac{t+n}{a}}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\xi^2\right) \mathrm{d}\xi \\
P_2(e) &= \int_{-\infty}^{t} p\left(h \mid \omega_2\right) \mathrm{d}h \\
&= \int_{-\infty}^{t} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{h-\eta}{\sigma}\right)^2\right\} \mathrm{d}\left(\frac{h-\eta}{\sigma}\right) \\
&= \int_{\infty}^{\frac{t}{\sigma}} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\xi^2\right) \mathrm{d}\xi
\end{aligned}
$$

其中

$$
\eta = \frac{1}{2}\left[\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right]
$$

$$
t = \ln\left[\frac{P\left(\omega_1\right)}{P\left(\omega_2\right)}\right], \quad \sigma = \sqrt{2\eta}
$$

$h(\boldsymbol{x})$ 的概率密度函数如图 2-12 所示。阴影部分相当于最小错误率贝叶斯决策的错误率

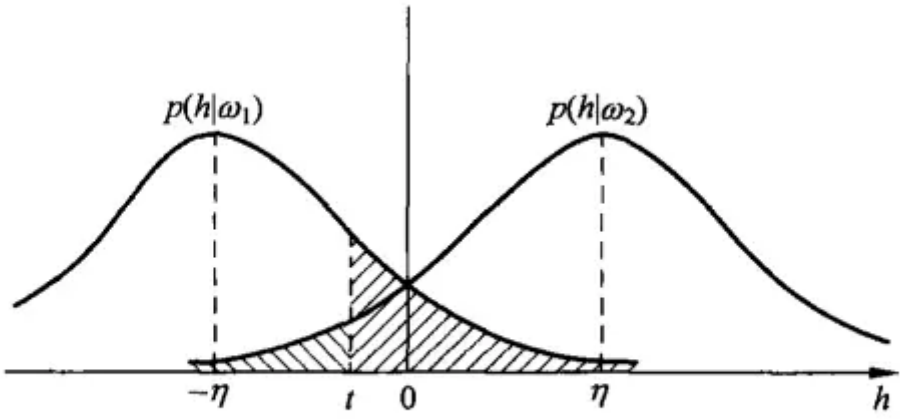


图 2-12　负对数似然比 $h(\boldsymbol{x})$ 的概率密度函数

2.7 离散概率模型下的统计决策举例

前面是实际上讲的都是连续情况下的，都是一些概率密度函数，其实离散情况在我看来更加简单，并且同样有一个很简单方法，马尔可夫模型。下面看一个案例

我们知道，生物的基因组是由 A 、 T 、 G 、 C四种核苷酸组成的序列。人的一套基因组单 链是由约 30 亿个 A 、 T 、 G 、 C 组成的一个超长的序列，可以把它看成是一本由四个字母写成 的天书。这样一个超长的字符串或其中一个子串,并不是由四个字母随机组成的，而是遵循 着很多特殊的规律。比如，由于一些生物化学机制的作用，基因组同一条链上出现相连的 C 和 G 的概率要比随机情况小很多。把相连的 C 和 G 叫做一个 CpG 双核苒酸。人们已经观 察到，CpG 在基因组上出现的平均频率比根据 C 、 G 各自出现的频率估计的组合出现频率小 很多,但是， 这些有限的 CpG在基因组上分布的位置不是均匀的， 而是倾向于集中在相对较 短的一些片段上,这种 CpG相对富集的区域被称作 CpG岛,就像大海上的小岛一样。CpG岛在基因组上有重要的功能， 研究CpG岛的识别是非常有意义的。

首先介绍一下马尔可夫模型，这是一种以状态转移为核心的模型，你也可以向有限状态机一样画一个状态转移图。对于**一阶马尔可夫模型**，核心思想就是，**下一个的状态仅和上一个状态有关**，比如在ATGC中，我们以DNA轴为主线，通常实际上是以时间顺序的，但是在这里我们把DNA序列看作时间序列，就有$x_i = \{A, T, G, C\}$

$$P\left(x_i \mid x_{i-1}, x_{i-2}, \cdots, x_1\right) = P\left(x_i \mid x_{i-1}\right)$$

并定义转移概率有

$$a_{st} = P\left(x_i = t \mid x_{i-1} = s\right)$$

即从一个状态转移到另一个状态的概率，对一个长度为 L 的序列, 我们观察到这个序列的概率是

$$P(x) \overset{\text{def}}{=} P\left(x_i, x_2, \cdots, x_L\right) = P\left(x_1\right) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$

在DNA案例中，我们有转移概率矩阵

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.282

图 2-15　CpG 岛与非 CpG 岛状态转移矩阵的例子

那么现在的问题就是，给你一个DNA链判断其是否是CpG岛。因为CpG岛和非CpG岛的转移概率矩阵是不同的，有一个非常直观的思想就是，用分别用CpG岛和非CpG岛的转移矩阵计算 $P\left(x_1\right) \prod_{i=2}^{L} a_{x_{i-1} x_i}$。这就代表走DNA能够走出这一条路线的状态的可能性。看是CpG岛的转移矩阵算出的可能性大和非CpG岛的转移矩阵算出的可能性大，哪一个大就是哪个。

用书里的话来说，这与前面贝叶斯决策的似然比很相似。在前面讲述连续变量的贝叶斯决策时, 用类条件概率密度来描述各类样本的特征分布。对于一个特定样本 x,根据类条件概率密度计算似然 比 $l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)}$ 并与一定的阈值做比较来进行判别。当采用最小错误率准则且两类先验 概率相等时, 阈值是 1 , 即如果似然比大于 1 则判别为第一类, 小于 1 则判别为第二类。

在 CpG 岛的识别中, 把 CpG 岛一类记作“+”, CpG 岛情况下的马尔可夫转移概率记作 $a^{+}_{x_{i-1} x_i}$;把非 CpG 岛一类记作“-”, 非 CpG 岛情况下的马尔可夫转移概率记 $a^{-}_{x_{i-1} x_i}$。为 了考虑到长序列处理方便, 可以采用下面的对数似然比 (log likelihood ratio) 来进行判别

$$S(x) = \log \frac{P(x \mid +)}{P(x \mid -)} = \log \frac{\prod_{i=1}^{L} a^{+}_{x_{i-1} x_i}}{\prod_{i=1}^{L} a^{-}_{x_{i-1} x_i}} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1} x_i}}{a_{x_{i-1} x_i}}$$

这个又叫做对数几率比。

但实际上，上面那个转移概率矩阵是需要我们自己估计的，或者可以称为训练。只需要找到一些CpG岛和非CpG岛的片段，统计AGCT后面出现AGCT的次数就可以估计了。

除了马尔可夫还有**隐马尔可夫模型**

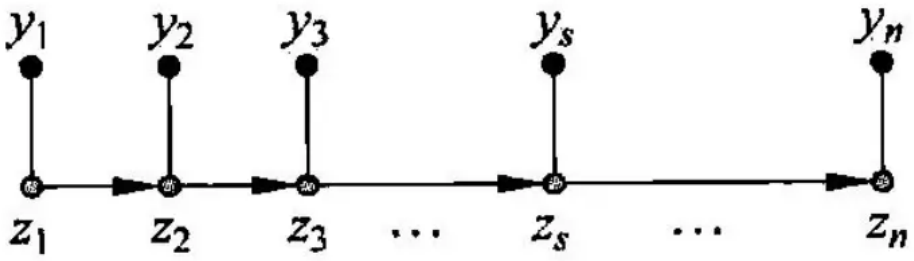


图 2-18　隐马尔可夫模型示意图

这里书上提到了Viterbi算法和Gibbs采样法可以求解。