—武大本科生课程

第5讲　概率分类(II)

(Lecture 4 Probability classification: Part 2)

武汉大学计算机学院机器学习课程组

Email: snowfly_li@163.com

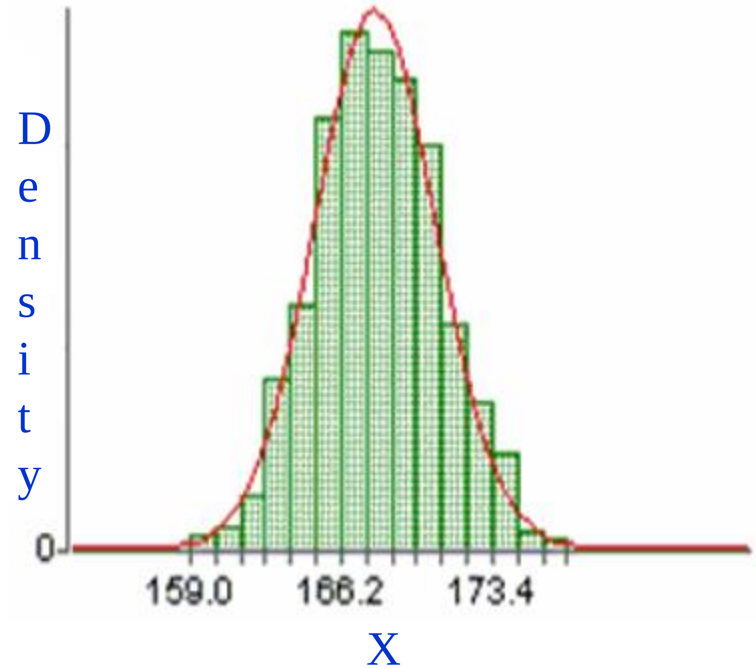
第4章 统计决策理论

(Chapter 4 Statistical decision theory)

内容目录 (以下红色字体为本讲3学时讲授内容)

4.5 正态分布模式的贝叶斯决策

实践中，许多随机数据是由**大量的、独立的、小效果的**多种因素的综合作用形成的，此时正态分布 (高斯分布) 是一种合理的近似。



正态分布概率模型的优点：

* 物理上的合理性。
* 数学上的简单性。

图中为某大学男大学生的身高数据，红线是拟合的密度曲线。可见，其身高应服从正态分布。

4.5.1 相关知识概述

1)二次型

设一向量$\boldsymbol{X}=[x_1,\cdots x_n]^{\mathrm{T}}$，矩阵 $\boldsymbol{A}=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$

则 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{AX}$ 称为二次型。

含义：是一个二次齐次多项式，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{AX}=\sum_{i,j=1}^{n} a_{ij}x_ix_j$

二次型中的矩阵$\boldsymbol{A}$是一个对称矩阵，即$a_{ij}=a_{ji}$。

2)正定二次型

$\forall \boldsymbol{X}\neq \boldsymbol{0}$ (即$\boldsymbol{X}$分量不全为零)，总有$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{AX}>0$，则称此二次型是正定的，而其对应的矩阵$\boldsymbol{A}$称为**正定矩阵**。

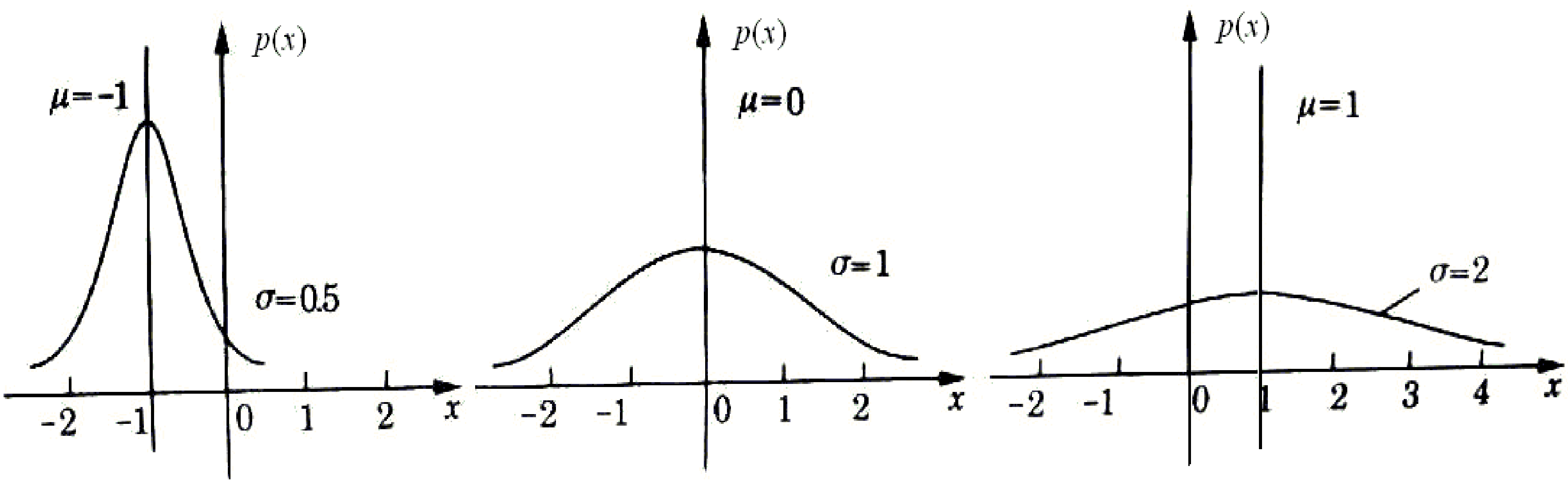
3) 单变量(一维)的正态分布

概率密度函数定义为：

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

曲线如图示：

①μ= -1，σ=0.5；　②μ= 0，σ=1；　③μ= 1，σ=2.



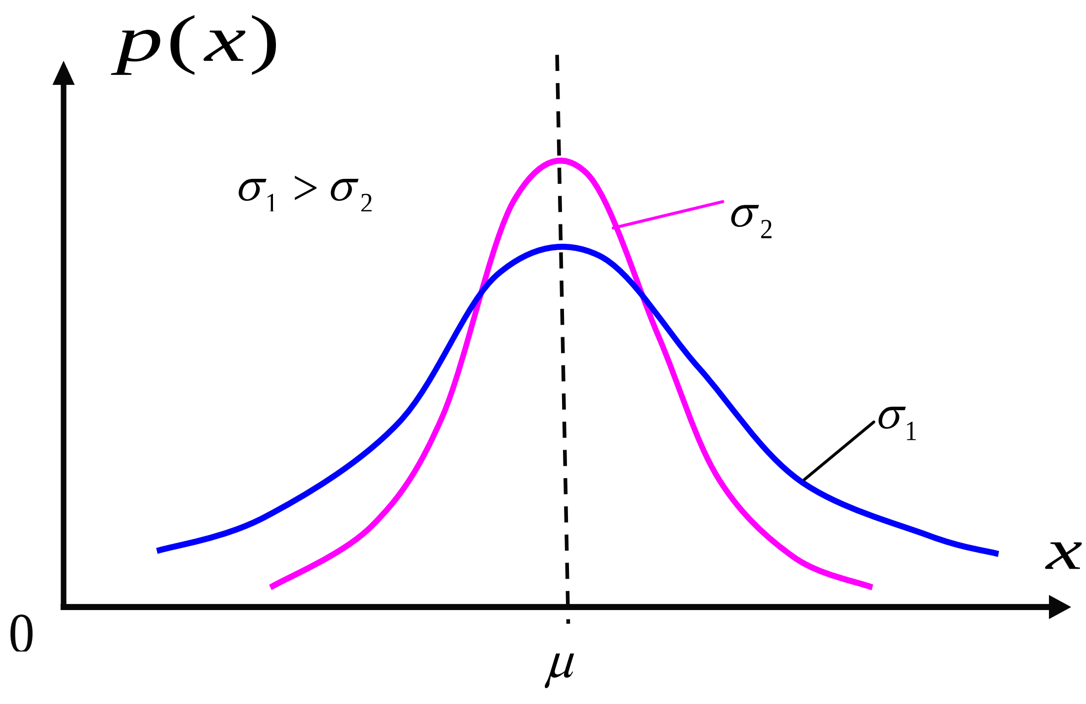
一维正态曲线的性质：

(1) 曲线在 x 轴的上方，与x轴不相交。

(2) 曲线关于直线 $x=\mu$对称。

(3) 当 $x=\mu$时，曲线位于最高点。

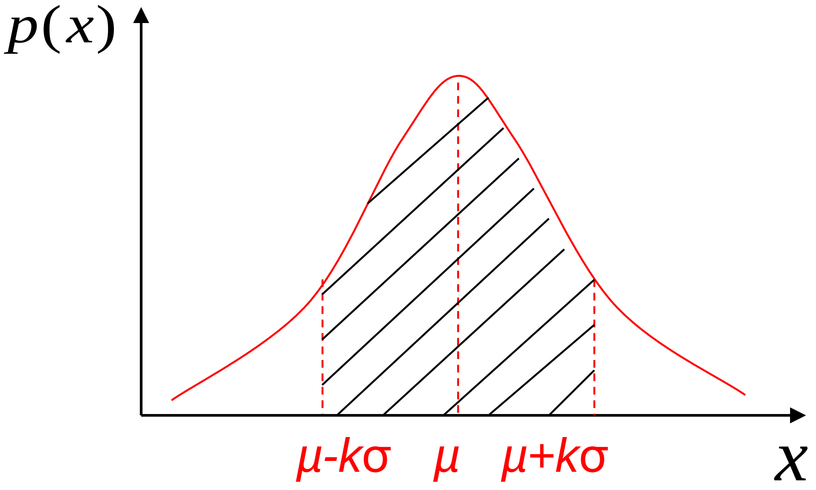
(4) 当$x<\mu$时，曲线上升；当$x>\mu$时，曲线下降。并且当曲线向左、右两边无限延伸时，以x轴为渐近线，向它无限靠近。



(5) μ一定时，曲线的形状由σ确定。σ越大，曲线越“矮胖”，表示总体的分布越分散；σ越小，曲线越“瘦高”，表示总体的分布越集中。

4) 3σ规则

$$P\{\mu-k\sigma\le x\le\mu+k\sigma\}=\begin{cases}0.683, & 当k=1时\\ 0.954, & 当k=2时\\ 0.997, & 当k=3时\end{cases}$$



服从正态分布的随机变量x取$\mu\pm3\sigma$范围内的概率几乎达到1，这就是3 σ原则。

即：绝大部分样本都落在了均值μ附近$\pm3\sigma$的范围内，因此正态密度曲线完全可由均值和方差来确定，常简记为：$p(x)\sim N\left(\mu,\sigma^2\right)$

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

5)多变量(n维)正态分布

概率密度函数定义为：

$$p(\boldsymbol{X})=\frac{1}{(2\pi)^{n/2}|\boldsymbol{C}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right\}$$

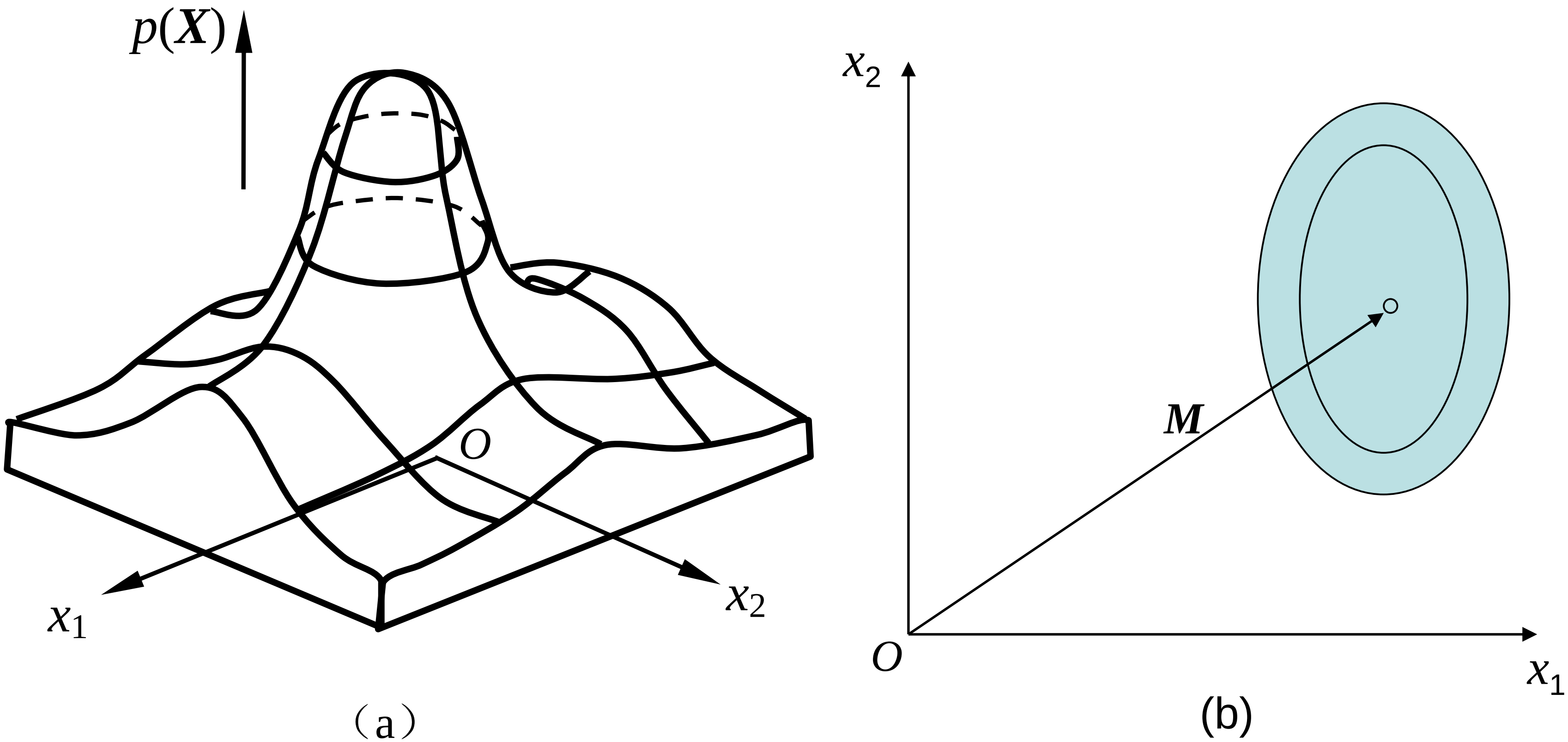
式中：$\boldsymbol{X}=[x_1,\cdots,x_n]^{\mathrm{T}}$；$\boldsymbol{M}=[m_1,\cdots,m_n]^{\mathrm{T}}$；

$$\boldsymbol{C}=\begin{bmatrix}\sigma_{11}^2 & \cdots & \sigma_{1n}^2\\ \vdots & & \vdots\\ \sigma_{n1}^2 & \cdots & \sigma_{nn}^2\end{bmatrix}$$

为协方差矩阵，是对称正定矩阵，独立元素有$n(n+1)/2$个；

$|\boldsymbol{C}|$：协方差矩阵$\boldsymbol{C}$的行列式。

多维正态概率密度函数完全由它的均值向量$\boldsymbol{M}$和协方差矩阵$\boldsymbol{C}$所确定，简记为：$p(\boldsymbol{X})\sim N(\boldsymbol{M},\boldsymbol{C})$

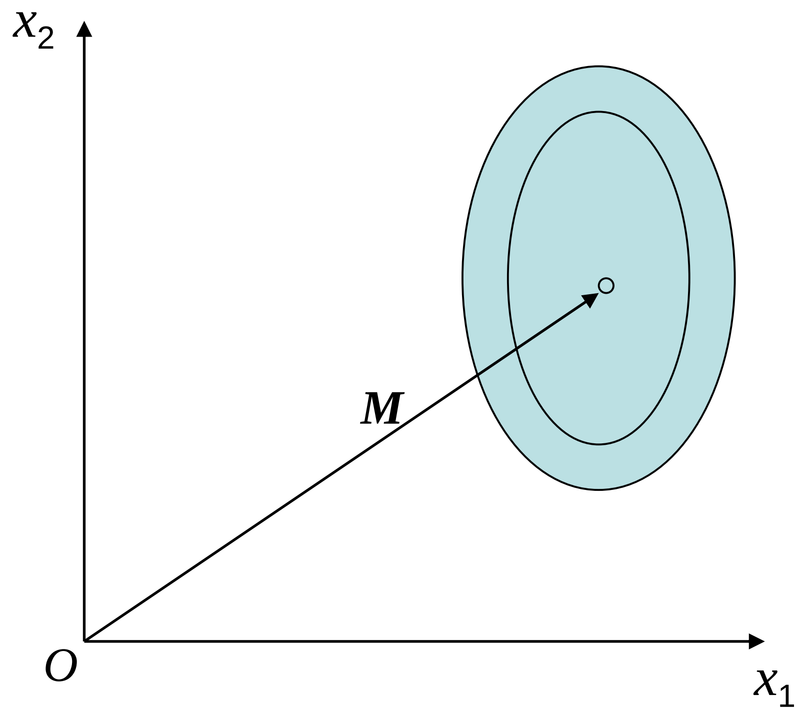


以二维正态密度函数为例：

等高线(等密度线)投影到x_1ox_2面上为椭圆，从原点O到点M 的向量为均值$\boldsymbol{M}$。

椭圆的位置：由均值向量$\boldsymbol{M}$决定；

椭圆的形状：由协方差矩阵$\boldsymbol{C}$决定。



二维正态分布概率密度函数的等密度线(等高线)投影到$x_1 O x_2$面上为椭圆(见图), **M**是均值向量, 决定椭圆的位置.

椭圆的形状由协方差矩阵**C**决定, 椭圆在平行于x_1轴的方向上受x_1的方差σ_{11}^2的影响, 在平行于x_2轴的方向上受x_2的方差σ_{22}^2的影响, 在其他方向上受x_1和x_1的协方差σ_{ij}^2的影响, 这里i, j=1, 2且$i \neq j$. 椭圆的主轴方向由**C**的特征向量决定, 主轴的长度与相应的特征值成正比.

4.5.2 正态分布的最小错误率Bayes决策规则

前面介绍的Bayes决策事先必须求出$p(\boldsymbol{X}|\omega_i)$, $P(\omega_i)$ 。而当类概密/似然$p(\boldsymbol{X}|\omega_i)$呈正态分布时，只需要知道 $\boldsymbol{M}$ 和 $\boldsymbol{C}$ 即可。

1) 多类情况

具有M 种模式类别的多变量正态分布概率密度函数为：

$$p(\boldsymbol{X} \mid \omega_i) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{C}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}\boldsymbol{C}_i^{-1}(\boldsymbol{X}-\boldsymbol{M}_i)\right\}$$

$$i = 1,2,\cdots,M$$

每一类模式的分布密度都完全被其均值向量$\boldsymbol{M}_i$和协方差矩阵$\boldsymbol{C}_i$所规定，其定义为：

$$\boldsymbol{M}_i = E_i[\boldsymbol{X}]$$

$$\boldsymbol{C}_i = E_i[(\boldsymbol{X}-\boldsymbol{M}_i)(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}]$$

协方差矩阵$\boldsymbol{C}_i$：反映样本分布区域的形状；

均值向量$\boldsymbol{M}_i$：表明了区域中心的位置。

$$p(\boldsymbol{X}\mid\omega_i)=\frac{1}{(2\pi)^{n/2}|\boldsymbol{C}_i|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}\boldsymbol{C}_i^{-1}(\boldsymbol{X}-\boldsymbol{M}_i)\right\}$$

最小错误率Bayes决策中，ω_i类的判别函数为 $p(\boldsymbol{X}\mid\omega_i)P(\omega_i)$，对于正态概率密度函数，为方便计算，这里取对数：

$$\ln[p(\boldsymbol{X}\mid\omega_i)P(\omega_i)]=\ln\left[p(\boldsymbol{X}\mid\omega_i)\right]+\ln\left[P(\omega_i)\right]$$

$$=\ln P(\omega_i)-\frac{n}{2}\ln 2\pi-\frac{1}{2}\ln|\boldsymbol{C}_i|-\frac{1}{2}\left\{(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}\boldsymbol{C}_i^{-1}(\boldsymbol{X}-\boldsymbol{M}_i)\right\}$$

对数是单调递增函数，取对数后仍有相对应的分类性能。

去掉与i无关的项，得到多类判别函数：

$$\boxed{d_i(\boldsymbol{X})=\ln P(\omega_i)-\frac{1}{2}\ln|\boldsymbol{C}_i|-\frac{1}{2}\left\{(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}\boldsymbol{C}_i^{-1}(\boldsymbol{X}-\boldsymbol{M}_i)\right\}}\quad\text{(a)}$$

$$i=1,2,\cdots,M$$

—— **正态分布的最小错误率Bayes决策的判别函数**

$$d_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{C}_i\right| - \frac{1}{2}\left\{(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}\boldsymbol{C}_i^{-1}(\boldsymbol{X}-\boldsymbol{M}_i)\right\}$$

决策规则同前：

$$\boxed{\text{若 } d_j(\boldsymbol{X}) > d_i(\boldsymbol{X}),\quad i = 1,2,\cdots,M,\ i \neq j \qquad \text{则 } \boldsymbol{X} \in \omega_j}$$

$d_i(\boldsymbol{X})$表示的决策面是超二次曲面(hyperquadric：可能是超球面hyper-sphere 、超椭球面hyper-ellipsoid、超双曲面hyperboloid 、超抛物面hyper-paraboloid)。当**X**是二维模式时，$d_i(\boldsymbol{X})$表示的决策面为二次曲线(圆、椭圆、双曲线、抛物线)。可见对正态分布模式的Bayes分类器，两类模式之间用一个二次决策面分开，就可以求得最优的分类效果。

2) 两类问题

(1) 当$C_1 \neq C_2$时： $p(\boldsymbol{X} \mid \omega_1) \sim N(M_1, C_1) \quad p(\boldsymbol{X} \mid \omega_2) \sim N(M_2, C_2)$

对应判别函数
$$\begin{cases} d_1(\boldsymbol{X}) = \ln P(\omega_1) - \dfrac{1}{2}\ln|\boldsymbol{C}_1| - \dfrac{1}{2}\{(\boldsymbol{X}-\boldsymbol{M}_1)^{\mathrm{T}}\boldsymbol{C}_1^{-1}(\boldsymbol{X}-\boldsymbol{M}_1)\} \\ d_2(\boldsymbol{X}) = \ln P(\omega_2) - \dfrac{1}{2}\ln|\boldsymbol{C}_2| - \dfrac{1}{2}\{(\boldsymbol{X}-\boldsymbol{M}_2)^{\mathrm{T}}\boldsymbol{C}_2^{-1}(\boldsymbol{X}-\boldsymbol{M}_2)\} \end{cases}$$

决策规则：

$$\text{若} d_1(\boldsymbol{X}) - d_2(\boldsymbol{X}) \begin{cases} > 0, & \text{则 } \boldsymbol{X} \in \omega_1 \\ < 0, & \text{则 } \boldsymbol{X} \in \omega_2 \end{cases} \qquad (a\text{-}1)$$

决策面方程：$d_1(\boldsymbol{X}) - d_2(\boldsymbol{X}) = 0$

决策面是$\boldsymbol{X}$的二次型方程决定的超曲面(超球面/超椭球面/超双曲面/超抛物面)。二维决策界面如右图所示。

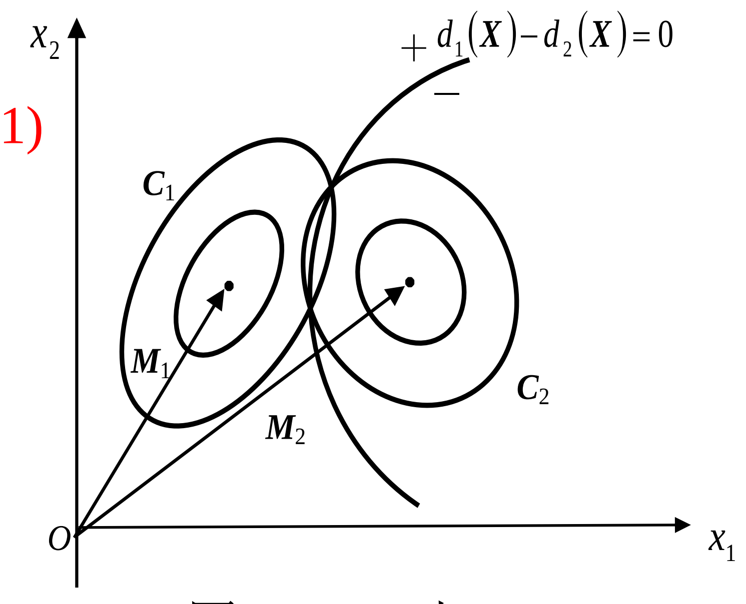


图 $\boldsymbol{C}_1 \neq \boldsymbol{C}_2$时

$$d_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln|\boldsymbol{C}_i| - \frac{1}{2}\left\{(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}\boldsymbol{C}_i^{-1}(\boldsymbol{X}-\boldsymbol{M}_i)\right\} \quad (a)$$

(2) 当$\boldsymbol{C}_1=\boldsymbol{C}_2=\boldsymbol{C}$时：由式$(a)$ 有

$$d_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{1}{2}\left\{(\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{M}_i^{\mathrm{T}})\left(\boldsymbol{C}^{-1}\boldsymbol{X} - \boldsymbol{C}^{-1}\boldsymbol{M}_i\right)\right\}$$

$$= \ln P(\omega_i) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{1}{2}\left\{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X} - \underline{\underline{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_i}} - \underline{\underline{\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}}} + \boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_i\right\}$$

展开相同，合并

$$= \ln P(\omega_i) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{1}{2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X} + \underline{\underline{\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}}} - \frac{1}{2}\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_i \qquad i=1,2$$

由此导出判别界面为：

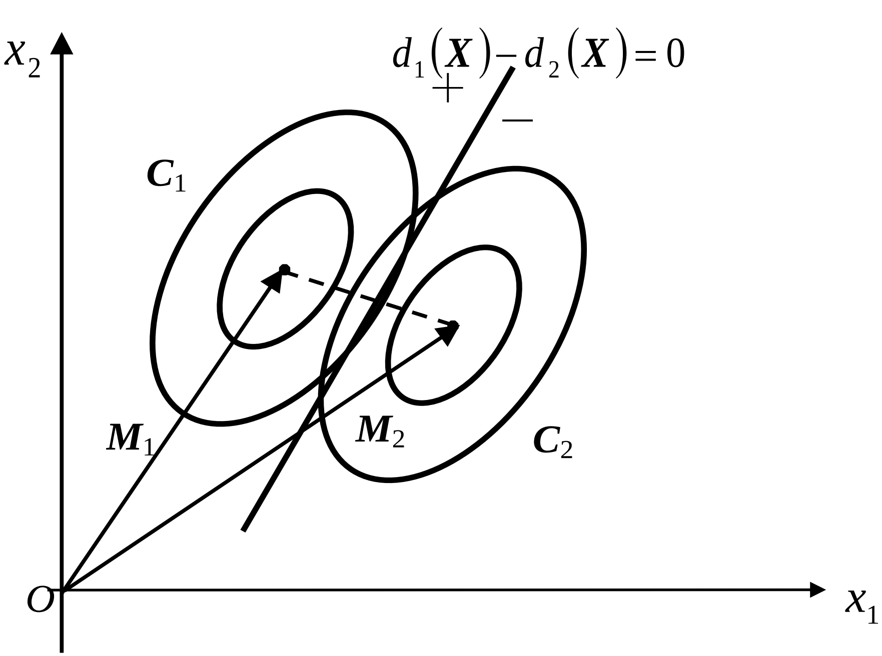
两类相同，抵消

$$d_1(\boldsymbol{X}) - d_2(\boldsymbol{X}) \qquad (a\text{-}2)$$

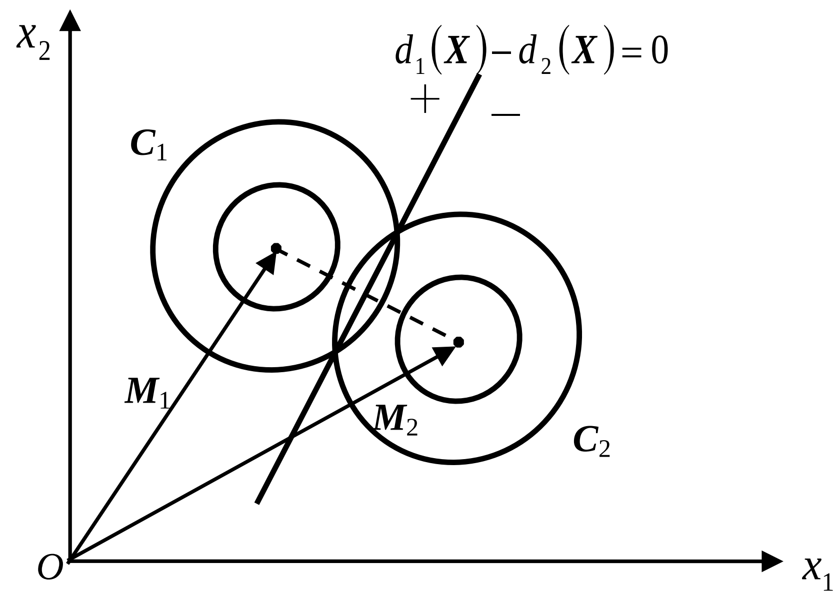
$$= \ln P(\omega_1) - \ln P(\omega_2) + \left(\boldsymbol{M}_1 - \boldsymbol{M}_2\right)^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X} - \frac{1}{2}\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_1 + \frac{1}{2}\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_2 = 0$$

为$\boldsymbol{X}$的线性函数，是一超平面。当为二维时，判别界面为一直线，如图(a)所示。

$$d_1(\boldsymbol{X})-d_2(\boldsymbol{X})$$
$$=\ln P(\omega_1)-\ln P(\omega_2)+\left(\boldsymbol{M}_1-\boldsymbol{M}_2\right)^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}-\frac{1}{2}\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_1+\frac{1}{2}\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_2=0$$



图(a) $\boldsymbol{C}_1=\boldsymbol{C}_2=\boldsymbol{C}$， $P(\omega_1)>P(\omega_2)$



图(b) $\boldsymbol{C}_1=\boldsymbol{C}_2=\boldsymbol{I}$且先验概率相等

(3) 当$\boldsymbol{C}_1=\boldsymbol{C}_2=\boldsymbol{I}$且$P(\omega_1)=P(\omega_2)=\frac{1}{2}$时：

$$d_1(\boldsymbol{X})-d_2(\boldsymbol{X})=\left(\boldsymbol{M}_1-\boldsymbol{M}_2\right)^{\mathrm{T}}\boldsymbol{X}-\frac{1}{2}\left(\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{M}_1-\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{M}_2\right)\quad(a\text{-}3)$$

判别界面如图(b)所示。

➤ 情况(3)讨论：

$$C_i=\sigma^2 I=\begin{bmatrix}\sigma_{11}^2 & \dots & 0\\ \dots & \dots & \dots\\ 0 & \dots & \sigma_{nn}^2\end{bmatrix}$$，只有方差（协方差为零）。

对于二类问题，$C_1=C_2=C=\sigma^2 I,\sigma_{11}^2=\sigma_{11}^2=\sigma^2,C^{-1}=\left(1/\sigma^2\right)I,$

决策面方程：$d_1(\mathbf{x})-d_2(\mathbf{x})=0$。根据$d_1(\boldsymbol{X})-d_2(\boldsymbol{X})$ (a-2)

$$=\ln P(\omega_1)-\ln P(\omega_2)+\left(\boldsymbol{M}_1-\boldsymbol{M}_2\right)^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}-\frac{1}{2}\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_1+\frac{1}{2}\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_2=0$$

决策面方程可等价表示为:

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

其中$\mathbf{w}=\mathbf{M}_1-\mathbf{M}_2$,

$$\mathbf{x}_0=\frac{1}{2}(\mathbf{M}_1+\mathbf{M}_2)-\frac{\sigma^2(\mathbf{M}_1-\mathbf{M}_2)}{\|\mathbf{M}_1-\mathbf{M}_2\|}\ln\frac{P(\omega_1)}{P(\omega_2)}$$

若$\sigma^2=1,P(\omega_1)=P(\omega_2)=0.5$,有:

$$\mathbf{x}_0=\frac{1}{2}(\mathbf{M}_1+\mathbf{M}_2)-\frac{(\mathbf{M}_1-\mathbf{M}_2)}{\|\mathbf{M}_1-\mathbf{M}_2\|}\ln\frac{0.5}{0.5}=\frac{1}{2}(\mathbf{M}_1+\mathbf{M}_2)$$

➢ 情况(3)讨论(Cont.):

二类2维特征情况下$\omega_i=\omega_1\omega_2$

(a)因为$C_i=I,\sigma_i^2=\sigma^2=1$,协方差为零。所以等概率面是一个圆形。

(b) Q W与$(\mathbf{x}-\mathbf{x}_0)$内积为0，因此分界面$H$与$W$垂直

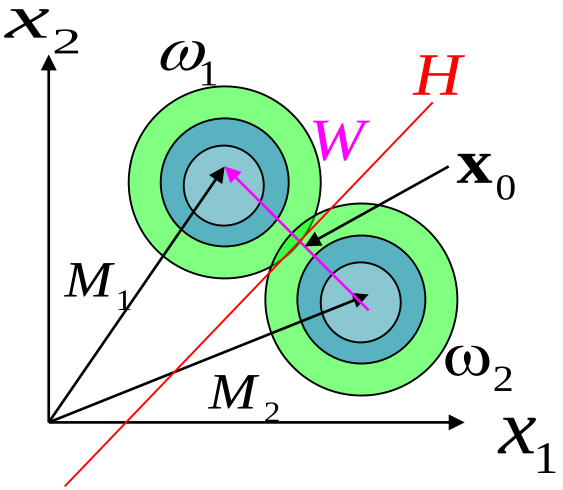
又 Q $W=M_i-M_j=M_1-M_2$，所以W与M_1-M_2同相(同方向)

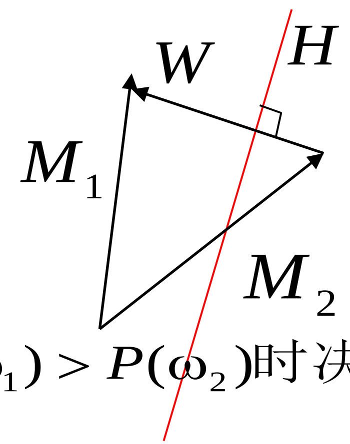
$\therefore$ 决策面H垂直于M的联线。

(c)如果先验概率相等$P(\omega_1)=P(\omega_2)=\dfrac{1}{2}$,$H$通过$\mu$联线的中点。

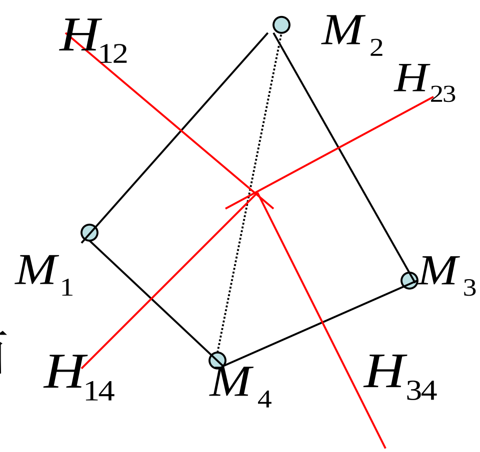
否则就是$P(\omega_1)\neq P(\omega_2)$,$H$离开先验概率大的一类。

(d)对多类情况，用各类的均值联线的垂直线作为分界面。





$P(\omega_1)>P(\omega_2)$时决策面



➢ 情况(1)和情况(2)也可类似讨论(此略)。

➢ 二类问题决策面的各种图形(*: 了解)

决策面方程：$g_i(x) - g_j(x) = 0$

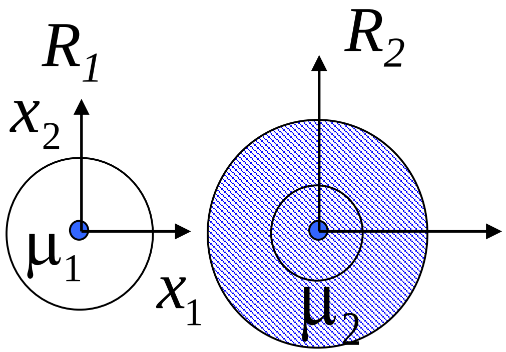
下面看看协方差矩阵不等时，二维特征空间二类问题决策面的各种图形。对于二类问题，有如下条件：

a、二类情况$\omega_1 \omega_2$;

b、$x_1 x_2$为条件独立；

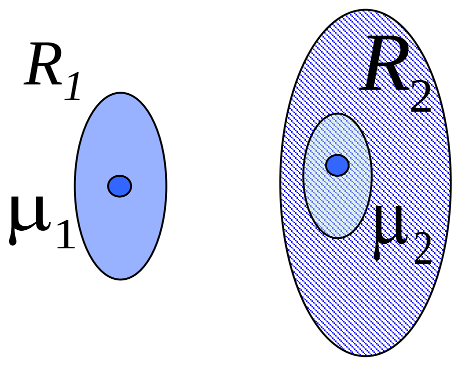
c、先验概率相等。

下面各图展示了二维特征空间两类问题的决策面的各种形式，图中的圆、椭圆表示等概密点轨迹。



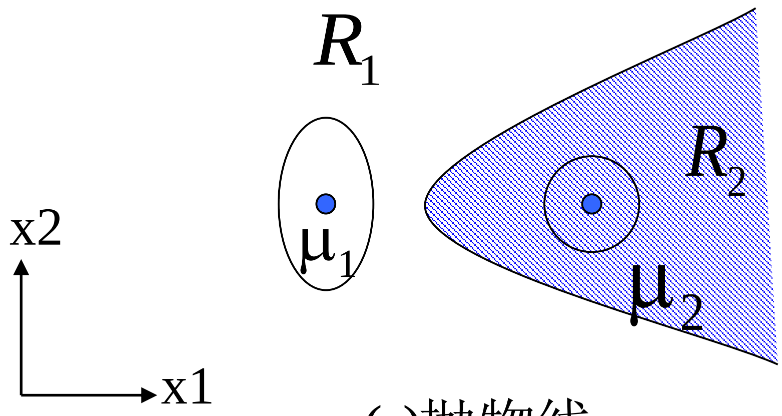
(*a*)圆

(a)图：由于ω2类的方差小，ω2类的模式更集中于μ2，决策面是包围μ2的一个圆



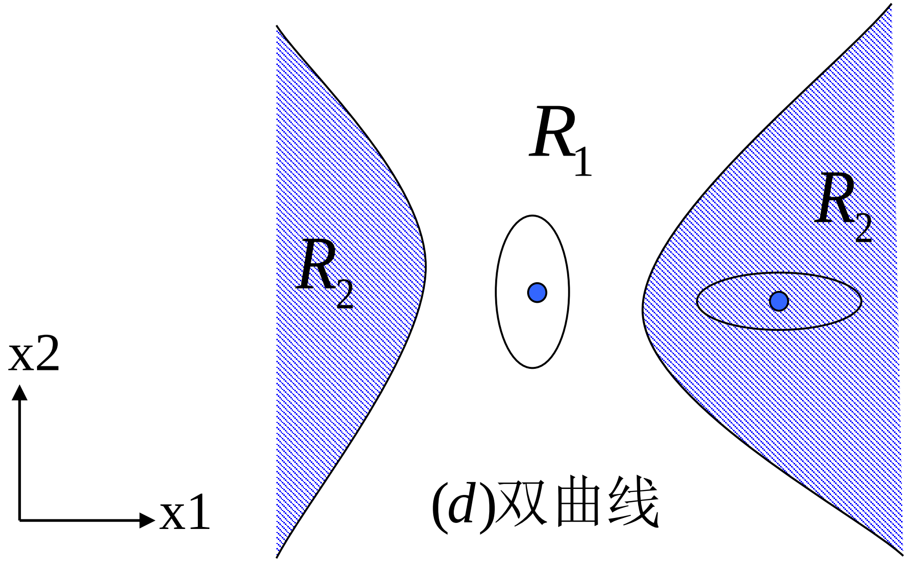
(*b*)椭圆

(b)图：决策面是包围μ2的一个椭圆



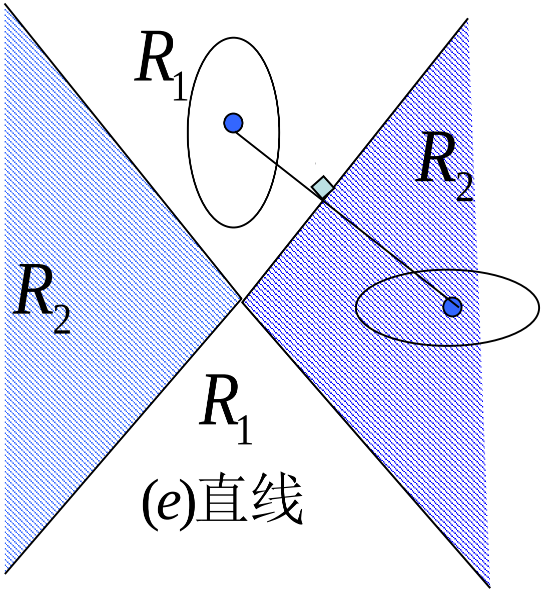
(c)抛物线

(c)图： 类ω1和ω2的x1的方差相同，但ω1的x2方差较ω2的x2方差大，从而x2值较大的模式更可能来自ω1类，因此决策面向右弯，呈抛物线状



(*d*)双曲线

(d)图：由于ω2类的x1的方差大于ω1类的x1的方差，而两类x2的方差情况正相反，因此决策面呈双曲线状



(e)直线

(e)图：由于两类的分布关于一直线对称，因此双曲线退化为相交直线。

4.5.3 正态分布Bayes分类器算法和例题

正态分布Bayes分类算法(假定各类样本服从正态分布)

1. 输入分类数M；特征数n，待分样本数m。
2. 输入训练样本数N和训练集资料矩阵X (N×n)。并计算有关参数($\boldsymbol{\mu}_i$, Σ_i) [(M_i, C_i)] 。
3. 计算矩阵X中各类的后验概率。
4. 若按最小错误率原则分类，则可根据3的结果判定X中各类样本的类别。
5. 若按最小风险原则分类，则输入各值，并计算X中各样本属于各类时的风险并判定各样本类别。

例4.3 有训练集资料矩阵如下表所示，现已知，$N=9$、$N_1=5$、$N_2=4$、$n=2$、$M=2$，试问，$X=(0,0)^T$应属于哪一类？

训练样本号k	1	2	3	4	5	1	2	3	4
特征 x_1	1	1	0	-1	-1	0	1	0	-1
特征 x_2	0	1	1	1	0	-1	-2	-2	-2
类别	ω_1					ω_2			

解1 假定二类协方差 矩阵不等（$\Sigma_1 \neq \Sigma_2$），则均值:

$$\overline{X}_{11}=\frac{1}{5}(1+1+0-1-1)=0, \overline{X}_{12}=\frac{3}{5}; \overline{X}_{21}=\frac{1}{4}(0+1+0-1)=0, \overline{X}_{22}=\frac{1}{4}(-1-2-2-2)=-\frac{7}{4}$$

$$\text{均值向量}: \overline{X_1}=\left(\overline{X_{11}}, \overline{X_{12}}\right)^T=(0,\frac{3}{5})^T, \overline{X_2}=\left(\overline{X_{21}}, \overline{X_{22}}\right)^T=(0,-\frac{7}{4})^T.$$

协方差矩阵：

$$\sum\nolimits_1=\begin{pmatrix}1 & 0\\ 0 & \frac{3}{10}\end{pmatrix}, \sum\nolimits_2=\begin{pmatrix}\frac{2}{3} & 0\\ 0 & \frac{1}{4}\end{pmatrix}, \sum\nolimits_1=\begin{pmatrix}C_{11} & C_{12}\\ C_{21} & C_{22}\end{pmatrix}$$

(要会协方差矩阵Σ的计算）

$$C_{11}=\frac{1}{5-1}\sum_{k=1}^{5}(x_{1k}-\bar{x}_{11})(x_{1k}-\bar{x}_{11})$$

$$=\frac{1}{4}\left[(1-0)^2+(1-0)^2+(0-0)^2+(-1-0)^2+(-1-0)^2\right]=1$$

$$C_{12}=\frac{1}{5-1}\sum_{k=1}^{5}(x_{1k}-\bar{x}_{11})(x_{2k}-\bar{x}_{12})=0$$

$$C_{12}=C_{21}$$

$$C_{22}=\frac{1}{5-1}\sum_{k=1}^{5}(x_{2k}-\bar{x}_{12})(x_{2k}-\bar{x}_{12})=\frac{3}{10}$$

协方差矩阵为$\sum_1=\begin{pmatrix}1 & 0\\ 0 & \frac{3}{10}\end{pmatrix},\sum_2=\begin{pmatrix}\frac{2}{3} & 0\\ 0 & \frac{1}{4}\end{pmatrix}$(计算方法同上）

利用公式(a−1)： $g(x)=g_2(x)-g_1(x)$

$$=\frac{1}{2}(\mathrm{x}-\overline{\mathrm{x}_1})^T\sum\nolimits_1^{-1}(\mathrm{x}-\overline{\mathrm{x}_1})-\frac{1}{2}(\mathrm{x}-\overline{\mathrm{x}_2})^T\sum\nolimits_2^{-1}(\mathrm{x}-\overline{\mathrm{x}_2})$$

$$+\frac{1}{2}\ln\frac{\left|\sum_1\right|}{\left|\sum_2\right|}-\ln\frac{P(\omega_1)}{P(\omega_2)}\begin{matrix}<\\>\end{matrix}0\Rightarrow x\in\begin{matrix}\omega_1\\ \omega_2\end{matrix} \qquad (a-1)$$

$\mathrm{x}=(x_1,x_2)^T$，将$\mathrm{x}=(x_1,x_2)^T=(0,0)^T$代入得：$g(x)=-10.91<0$

所以判$X=(0,0)^T$属于ω_1类。

令$g(x)=0$得分界线方程为：$\frac{1}{2}x_1^2+\frac{2}{3}x_2^2+18x_2+10.91=0$

这是一个非线性椭圆方程：$\frac{x_1^2}{14.81^2}+\frac{(x_2+13.5)^2}{12.88^2}=1$

解2 假定两类协方差矩阵相等$\Sigma=\Sigma_1+\Sigma_2$

$$\Sigma=\Sigma_1+\Sigma_2=\begin{pmatrix}\frac{5}{3} & 0\\ 0 & \frac{11}{20}\end{pmatrix},\Sigma^{-1}=\begin{pmatrix}\frac{3}{5} & 0\\ 0 & \frac{20}{11}\end{pmatrix},$$

所以代入$\mathbf{x}=(0,0)^T$到公式$(a-2)$，得：

$$g(\mathbf{x})=g_2(\mathbf{x})-g_1(\mathbf{x})$$

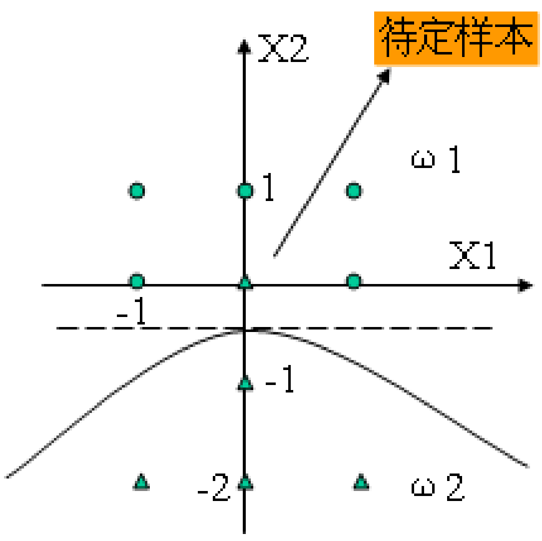
$$=(\overline{\mathrm{x}}_2-\overline{\mathrm{x}}_1)^T\Sigma^{-1}\mathbf{x}+\frac{1}{2}(\overline{\mathrm{x}}_1^T\Sigma^{-1}\overline{\mathrm{x}}_1-\overline{\mathrm{x}}_2^T\Sigma^{-1}\overline{\mathrm{x}}_2)-\ln\frac{P(\omega_1)}{P(\omega_2)}\begin{matrix}<0\\>0\end{matrix}\Rightarrow\mathbf{x}\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}\quad(a\text{-}2)$$

$$=-2.68<0$$

故应把$\mathrm{x}=(0,0)^T$判为ω_1类，

分界线方程为$g(x)=\dfrac{-47}{11}x_2-2.68=0$

从而得$x_2=-0.61$为一直线，如图中虚线所示。



两种解得分界线

例 4.4 设在三维特征空间里，有两类正态分布模式，每类各有4个样本，分别为

$$\omega_1: \quad [1,0,1]^{\mathrm{T}} \quad [1,0,0]^{\mathrm{T}} \quad [0,0,0]^{\mathrm{T}} \quad [1,1,0]^{\mathrm{T}}$$

$$\omega_2: \quad [0,0,1]^{\mathrm{T}} \quad [0,1,1]^{\mathrm{T}} \quad [1,1,1]^{\mathrm{T}} \quad [0,1,0]^{\mathrm{T}}$$

其均值向量和协方差矩阵可用下式估计：

$$\boldsymbol{M}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{X}_{ij}$$

$$\boldsymbol{C}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^{\mathrm{T}} - \boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}}$$

上式中，N_i为类别ω_i中模式的数目，$\boldsymbol{X}_{ij}$代表在第i类中的第j个模式。两类的先验概率 $P(\omega_1)=P(\omega_2)=1/2$ 。试确定两类之间的判别界面。

解：$\boldsymbol{M}_1=\frac{1}{4}\left\{\begin{bmatrix}1\\0\\1\end{bmatrix}+\begin{bmatrix}1\\0\\0\end{bmatrix}+\begin{bmatrix}0\\0\\0\end{bmatrix}+\begin{bmatrix}1\\1\\0\end{bmatrix}\right\}=\frac{1}{4}\begin{bmatrix}3\\1\\1\end{bmatrix}=\frac{1}{4}\begin{bmatrix}3 & 1 & 1\end{bmatrix}^{\mathrm{T}}$

$$\boldsymbol{M}_2=\frac{1}{4}\begin{bmatrix}1 & 3 & 3\end{bmatrix}^{\mathrm{T}}$$

$$\boldsymbol{C}_1=\boldsymbol{C}_2=\frac{1}{16}\begin{bmatrix}3 & 1 & 1\\1 & 3 & -1\\1 & -1 & 3\end{bmatrix}$$ 经计算有 $$\boldsymbol{C}^{-1}=\begin{bmatrix}8 & -4 & -4\\-4 & 8 & 4\\-4 & 4 & 8\end{bmatrix}$$

因协方差矩阵相等，故(a-2)式为其判别式。由于$P(\omega_1)=P(\omega_2)=\frac{1}{2}$

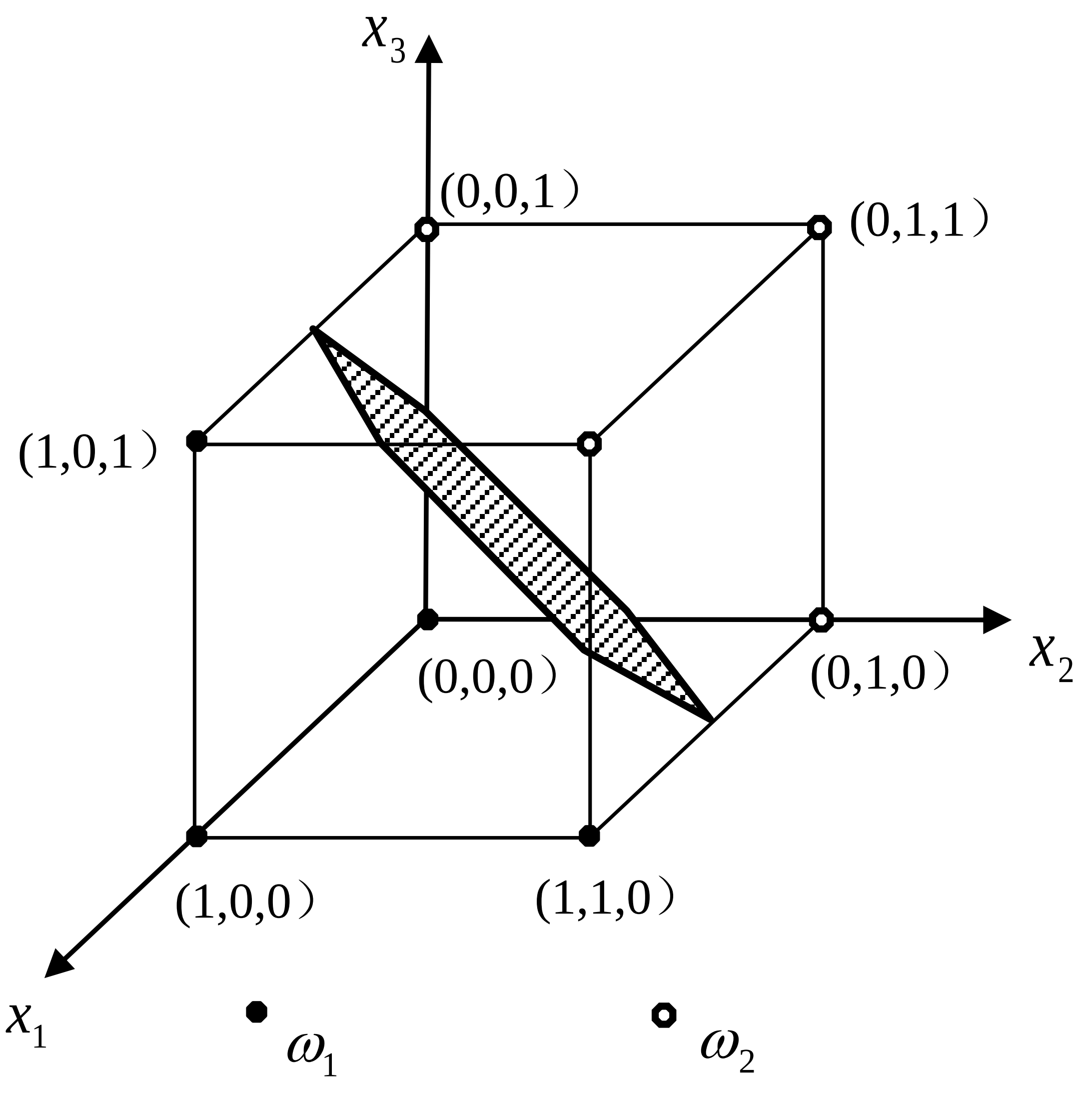
$$d_1(\boldsymbol{X})-d_2(\boldsymbol{X})=(\boldsymbol{M}_1-\boldsymbol{M}_2)^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}-\frac{1}{2}\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_1+\frac{1}{2}\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_2$$

将$\boldsymbol{X}=[x_1,x_2,x_3]^{\mathrm{T}}$ 代入：$d_1(\boldsymbol{X})-d_2(\boldsymbol{X})=8x_1-8x_2-8x_3+4=0$

$d_1(\boldsymbol{X})-d_2(\boldsymbol{X})$ (a-2)

$$=\ln P(\omega_1)-\ln P(\omega_2)+(\boldsymbol{M}_1-\boldsymbol{M}_2)^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{X}-\frac{1}{2}\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_1+\frac{1}{2}\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{M}_2$$

图中画出了**决策面 (判别平面)** 的一部分(见阴影部分)。



例4.5 有训练集资料矩阵如下表所示，现已知， N=9、$N_1=N_2=3$、n=2、M=3，试问，未知样本 $X=(0,0)^T$应属于哪一类？

训练样本号k	1 2 3	1 2 3	1 2 3
特征 x_1	0 1 2	-2 -1 -2	0 1 -1
特征 x_2	1 0 -1	1 0 -1	-1 -2 -2
类别	ω_1	ω_2	ω_3

解1 假定三类协方差不等；

均值向量$\mu_1=(1,0)^T,\mu_2=(-\frac{5}{3},0)^T$，$\mu_3=(0,-\frac{5}{3})^T$

$$协方差矩阵为:\sum_1=\begin{pmatrix}1&0\\0&1\end{pmatrix},\sum_2=\begin{pmatrix}\frac{1}{3}&0\\0&1\end{pmatrix},\sum_3=\begin{pmatrix}1&0\\0&\frac{1}{3}\end{pmatrix}$$

$$所以\sum_1^{-1}=\begin{pmatrix}1&0\\0&1\end{pmatrix},\sum_2^{-1}=\begin{pmatrix}3&0\\0&1\end{pmatrix},\sum_3^{-1}=\begin{pmatrix}1&0\\0&3\end{pmatrix}$$

$$\left|\sum{}_1\right| = 1, \left|\sum{}_2\right| = \frac{1}{3}, \left|\sum{}_3\right| = \frac{1}{3}$$

先验概率$P(\omega_1) = P(\omega_2) = P(\omega_3) = \frac{1}{3}$

代入多类判别函数公式(a)：

$$g_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}\left\{(\boldsymbol{X} - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_i)\right\} \qquad (a)$$

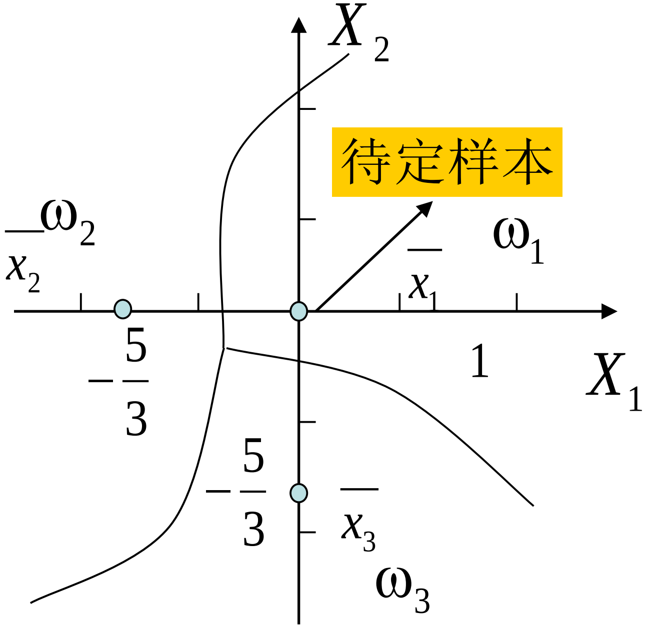
得到：$g_1(x) = -\frac{1}{2}\left(x_1^2 + x_2^2 - 2x_1 + 1\right)$

$$g_2(x) = -\frac{1}{2}\left(3x_1^2 + x_2^2 + 10x_1 + 7.2\right)$$

$$g_3(x) = -\frac{1}{2}\left(x_1^2 + 3x_2^2 + 10x_2 + 7.2\right)$$

将$\mathbf{x} = (0,0)^T$代入得：

$g_1(x) = -0.5, g_2(x) = g_3(x) = -3.6$



故应判样本$X=(0,0)^T$为ω_1类

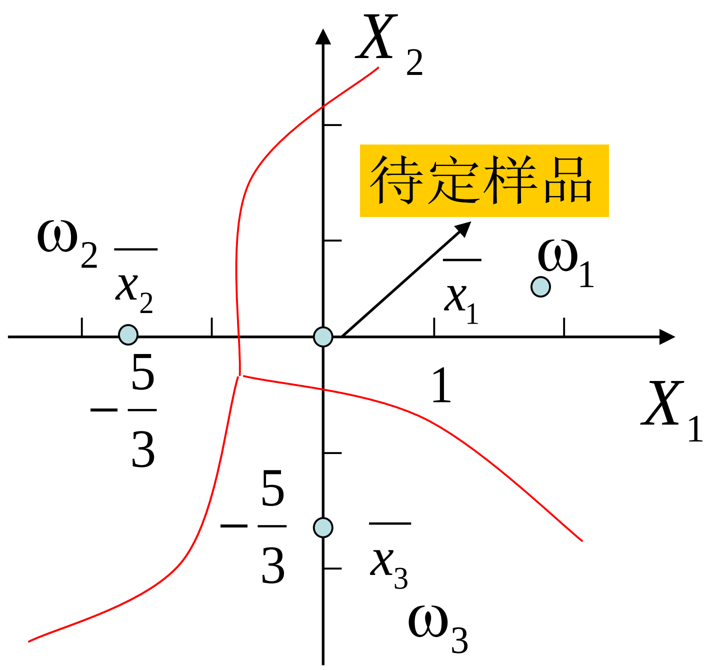
分别令$g_1(x)=g_2(x), g_2(x)=g_3(x), g_3(x)=g_1(x)$

$$g_1(x)-g_2(x)=x_1^2+6x_1+3.1=0$$

$$g_2(x)-g_3(x)=-x_1^2+x_2^2-5x_1+5x_2=0$$

$$g_3(x)-g_1(x)=-x_2^2-2x_1-5x_2-2.6=0$$

- 可得三类分界线，见下图所示：



解2 设三类协方差矩阵相等

$$\sum=\sum_{1}+\sum_{2}+\sum_{3}=\begin{pmatrix}\frac{7}{3} & 0\\ 0 & \frac{7}{3}\end{pmatrix},\sum^{-1}=\begin{pmatrix}\frac{3}{7} & 0\\ 0 & \frac{3}{7}\end{pmatrix}$$

代入多类判别函数公式(a):

$$g_i(\boldsymbol{X})=\ln P(\omega_i)-\frac{1}{2}\ln\left|\Sigma_i\right|-\frac{1}{2}\left\{(\boldsymbol{X}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)\right\} \qquad (a)$$

有:

$$g_1(x)=\frac{3}{7}x_1-\frac{3}{14},g_2(x)=-\frac{5}{7}x_1-\frac{25}{42}$$

$$g_3(x)=-\frac{5}{7}x_2-\frac{25}{42}$$

将$\mathbf{x}=(0,0)^T$ 代入得:

$$g_1(x)=-\frac{3}{14},g_2(x)=g_3(x)=-\frac{25}{42}$$

故应判样本$\mathbf{x}=\left(0,0\right)^T$为$\omega_1$类

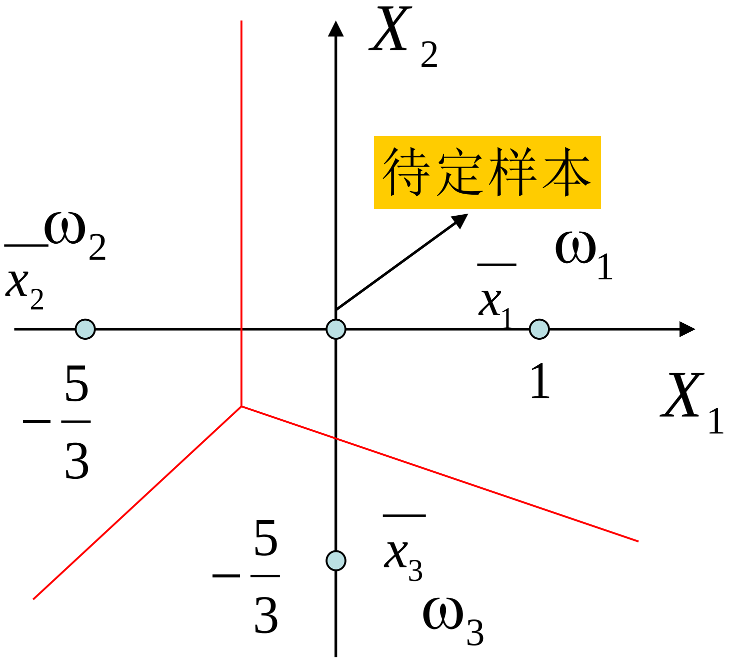
分别令$g_1(x)=g_2(x), g_2(x)=g_3(x), g_3(x)=g_1(x)$

$$g_1(x)-g_2(x)=\frac{8}{7}x_1+\frac{8}{21}$$

$$g_2(x)-g_3(x)=-\frac{5}{7}x_1+\frac{5}{7}x_2$$

$$g_3(x)-g_1(x)=-\frac{3}{7}x_1-\frac{5}{7}x_2-\frac{8}{21}$$

- 可得三类分界线，见右图所示：



4.6 参数密度估计

在前面推导的几中经典统计决策规则中，通常假设先验概率$P(\omega_i)$和类条件概率密度函数$p(X|\omega_i)$是已知的。但是在很多情况中，我们能够利用的只有有限个样本，而$p(X|\omega_i)$和$P(\omega_i)$是未知的，需要根据已有样本进行参数估计，然后将估计值当作真实值来使用。

以下讨论：

已知类别的样本→得到类模式的概率密度 $p(X|\omega_i)$

概率密度的两大类估计方法：

* **参数估计方法：**

已知概率密度函数的形式而函数的有关参数未知，通过估计参数来估计概率密度函数的方法。

两种主要参数估计法：**确定性参数估计方法**把参数看做确定而未知的，典型方法为**最大似然估计**。 **随机参数估计方法**把未知参数当做具有某种分布的随机变量，典型方法为**贝叶斯估计**。

* **非参数估计方法：**

非参数估计就是在概率密度函数的形式未知的条件下，直接利用样本来推断概率密度函数。常用的非参数估计方法有**Parzen窗估计 (KDE, Kernel Density Estimation)** 和k_N -近邻估计。

4.6.1 最大似然估计

设：ω_i类的类条件概率密度函数具有某种确定的函数形式；

$\boldsymbol{\theta}$是该函数的一个未知参数或参数集。

最大似然估计(Maximum Likelihood Estimation, **MLE**)把$\boldsymbol{\theta}$当作确定的未知量进行估计。

1. 似然函数

从ω_i类中独立地抽取N个样本：$X^N = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N\}$

称这N个样本的**联合概率密度**函数 $p(X^N | \boldsymbol{\theta})$ 为**相对于样本集$\boldsymbol{X}^N$的$\boldsymbol{\theta}$的似然函数**。

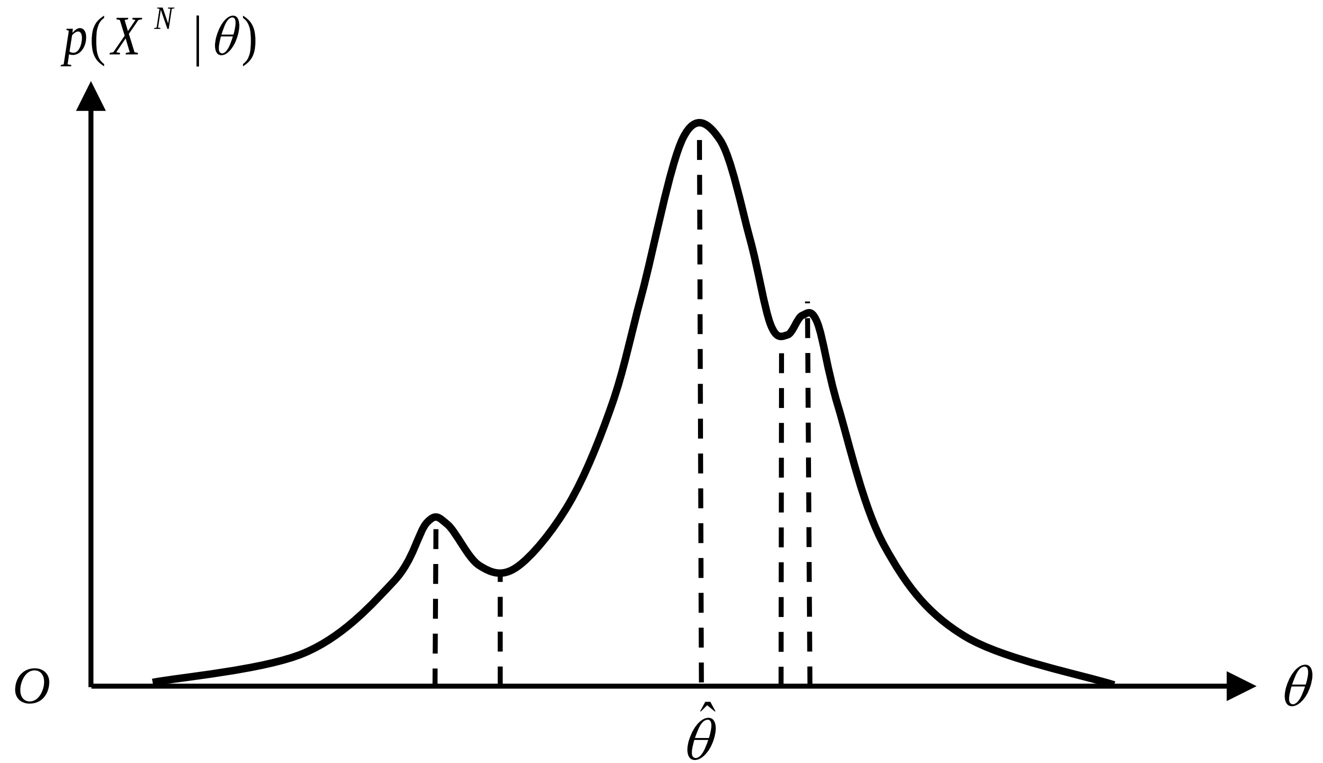
$$L(\boldsymbol{\theta}) = p(X^N | \boldsymbol{\theta}) = p(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N | \boldsymbol{\theta}) = \prod_{k=1}^{N} p(\boldsymbol{X}_k | \boldsymbol{\theta})$$

——在参数$\boldsymbol{\theta}$下观测到的样本集X^N的概率(联合分布)密度

2. 最大似然估计

最大似然估计为样本集构造一个似然函数，通过让似然函数最大化求解出参数$\boldsymbol{\theta}$。其直观解释是，寻求参数的值使得给定的样本集出现的概率(或概率密度函数值)最大。

即求解最优化问题：$\max\limits_{\theta} p(X^N|\boldsymbol{\theta}) = \max\limits_{\theta} \prod_{k=1}^{N} p(\boldsymbol{X}_k|\boldsymbol{\theta})$



$\boldsymbol{\theta}$为一维时的最大似然估计示意图

$\boldsymbol{\theta}$的最大似然估计量$\hat{\boldsymbol{\theta}}$就是使似然函数达到最大值的估计量。

由 $\dfrac{dp(X^N|\boldsymbol{\theta})}{d\boldsymbol{\theta}} = 0$ 求得。

$L(\theta)$的自然对数称为**对数似然函数**，记为$H(\theta)$，即：

$$H(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln p(X^N \mid \boldsymbol{\theta})$$

对数函数是单调递增的，因此，使对数似然函数最大的$\boldsymbol{\theta}$值也必然使似然函数达到最大。

$\boldsymbol{\theta}$的最大似然估计是下面微分方程的解：

$$\frac{\partial H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

设ω_i类模式的概率密度函数有p个未知参数，记为p维向量

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^{\mathrm{T}}$$

此时
$$H(\boldsymbol{\theta}) = \ln p(X^N \mid \boldsymbol{\theta}) = \sum_{k=1}^{N} \ln p(X_k \mid \boldsymbol{\theta})$$

求解最优化问题：$\max\limits_{\theta} \ln p(X^N \mid \boldsymbol{\theta}) = \max\limits_{\theta} \sum_{i=1}^{N} \ln p(\boldsymbol{X}_k \mid \boldsymbol{\theta})$

这是一个不带约束的优化问题，一般情况下可直接求得解析解。也可用梯度法或牛顿法求解。

$$\frac{\partial}{\partial \boldsymbol{\theta}}\left[\sum_{k=1}^{N} \ln p(X_k \mid \boldsymbol{\theta})\right] = 0$$

$$\begin{cases} \displaystyle\sum_{k=1}^{N} \frac{\partial}{\partial \theta_1} \ln p(X_k \mid \boldsymbol{\theta}) = 0 \\ \displaystyle\sum_{k=1}^{N} \frac{\partial}{\partial \theta_2} \ln p(X_k \mid \boldsymbol{\theta}) = 0 \\ \vdots \\ \displaystyle\sum_{k=1}^{N} \frac{\partial}{\partial \theta_p} \ln p(X_k \mid \boldsymbol{\theta}) = 0 \end{cases} \qquad (b)$$

解以上微分方程即可得到$\boldsymbol{\theta}$的最大似然估计值。

3. 正态分布情况举例

例4.6 设ω_i类：正态分布、**一维模式**、概率密度函数为

$$p(\boldsymbol{X} \mid \omega_i) \sim N(\mu, \sigma^2)$$

待估计参数为μ，σ^2。

$p(\boldsymbol{X} \mid \omega_i)$ 可表示为 $p(\boldsymbol{X} \mid \boldsymbol{\theta}) \sim N(\mu, \sigma^2)$ 。

其中， $\boldsymbol{\theta} = [\theta_1, \theta_2]^{\mathrm{T}}$， $\theta_1 = \mu$ ， $\theta_2 = \sigma^2$ 。

若X^N表示从ω_i中独立抽取的N个样本，则θ的似然函数为

$$p(X^N \mid \boldsymbol{\theta}) = \prod_{k=1}^{N} p(\boldsymbol{X}_k \mid \boldsymbol{\theta})$$

其中， $p(\boldsymbol{X}_k \mid \boldsymbol{\theta}) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left[-\dfrac{(\boldsymbol{X}_k - \mu)^2}{2\sigma^2} \right]$

$$\ln p(\boldsymbol{X}_k \mid \boldsymbol{\theta}) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(\boldsymbol{X}_k - \mu)^2}{2\sigma^2}$$

得
$$\begin{cases} \displaystyle\sum_{k=1}^{N} \frac{\partial}{\partial \theta_1} \ln p(\boldsymbol{X}_k \mid \boldsymbol{\theta}) = \sum_{k=1}^{N} \frac{\boldsymbol{X}_k - \theta_1}{\theta_2} = 0 \\ \displaystyle\sum_{k=1}^{N} \frac{\partial}{\partial \theta_2} \ln p(\boldsymbol{X}_k \mid \boldsymbol{\theta}) = \sum_{k=1}^{N} \left[\frac{-1}{2\theta_2} + \frac{(\boldsymbol{X}_k - \theta_1)^2}{2\theta_2^2} \right] = 0 \end{cases}$$

由以上方程组解得均值和方差的估计量为

$$\hat{\mu} = \hat{\theta}_1 = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{X}_k$$

$$\hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{N}\sum_{k=1}^{N} (\boldsymbol{X}_k - \hat{\mu})^2$$

类似地，多元正态分布情况：

$$\hat{\boldsymbol{M}}_i = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{X}_k$$

$$\hat{\boldsymbol{C}}_i = \frac{1}{N}\sum_{k=1}^{N} (\boldsymbol{X}_k - \hat{\boldsymbol{M}}_i)(\boldsymbol{X}_k - \hat{\boldsymbol{M}}_i)^{\mathrm{T}}$$

最大似然估计结果：

- 均值向量的最大似然估计是样本均值；
- 协方差矩阵的最大似然估计是N个矩阵的算术平均。

4.7 最大似然估计在Logistic回归模型训练中的应用

4.7.1 Logistic回归模型

Logistic回归模型又称对数几率回归[01]模型、逻辑回归模型等。

Logistic回归由统计学家David Cos于1958年提出。Logistic回归实质是将数据拟合到一个Logistic函数中，从而预测事件发生的可能性。其因变量可以是二分类或多分类；二分类更为常用，也更加易于解释。

Logistic回归可应用于各个领域，如机器学习、医学领域等。如，在临床医疗中，根据观测的患者多项指标，如性别、年龄、身体质量指数(BMI, Body Mass Index)和血液检测等，预测该患者是否患糖尿病。

本节仅介绍二分类Logistic回归。给定训练样本集$(\mathbf{x}_i, y_i)$, i=1,…,m；其中$\mathbf{x}_i$为n维特征向量，y_i为类别标签，其取值为0或1。Logistic回归研究是的**样本特征向量为x的条件下样本类别为正类(y=1)的概率**，记$\mathbf{x}$的条件下y取1的概率(事件发生)为$p=P(y=1|\mathbf{x})$，y取0的概率(事件不发生)为1-p。

- 事件发生与事件不发生的概率之比，称为“**事件发生的胜率**”（也称“**几率**”/“**优势比**”）odds：

$$\frac{p}{1-p}$$

●事件发生胜率的对数，称为“**事件的对数胜率**”(“**对数几率**”/“**对数优势比**”)：

$$\mathrm{logit}(p) = \ln(\frac{p}{1-p})$$

Logistic回归实际上是**对事件的对数胜率进行线性建模，可表示为**：

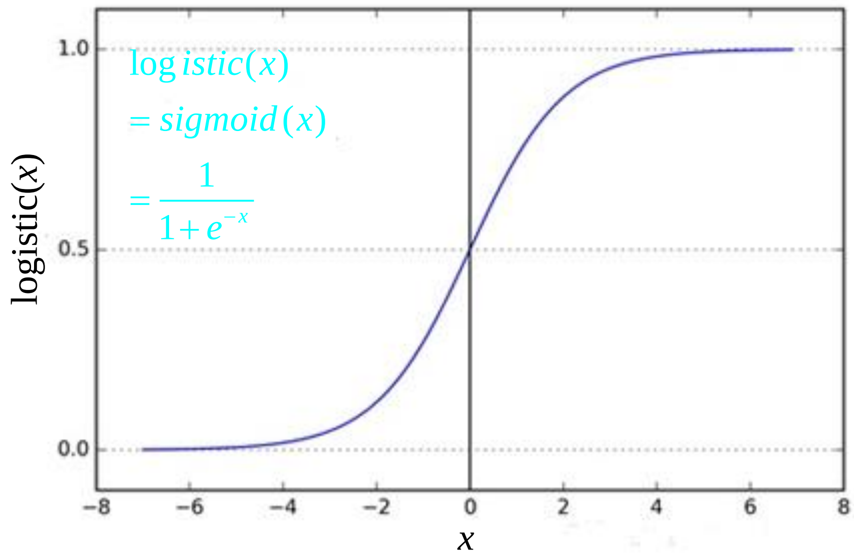
$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})}\right) = \mathbf{w}^T\mathbf{x} \equiv \mathbf{w}\cdot\mathbf{x}$$

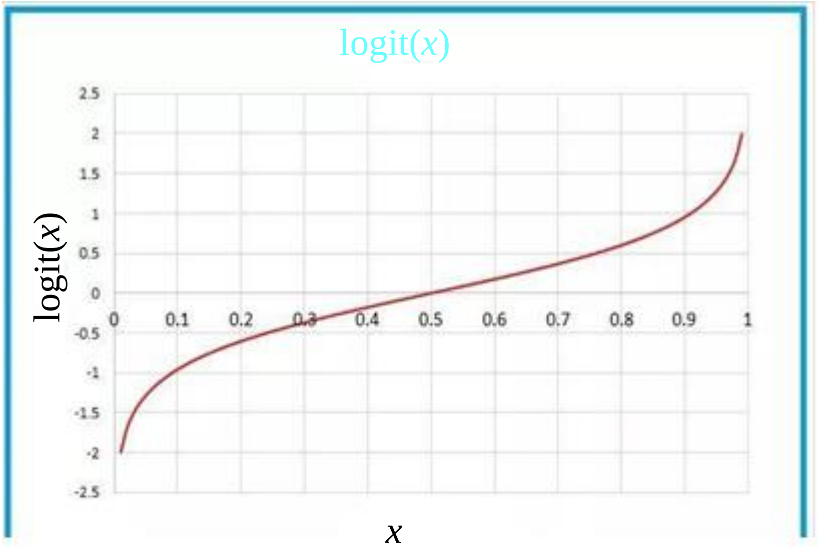
$$= w_0 + w_1x_1 + \cdots + w_nx_n$$

且有：$$p = \frac{e^{(w_0+w_1x_1+\cdots+w_nx_n)}}{1+e^{(w_0+w_1x_1+\cdots+w_nx_n)}}$$

注意：这里的“事件”是指：在样本特征为**x**的条件下，样本类别为正类(y=1)的概率！

Logistic函数与Logit函数：互为反函数





$$\text{logistic}(x) = \frac{1}{1+e^{-x}}$$

$$\text{logit}(x) = \ln\frac{x}{1-x}$$

$$\text{logistic}^{-1}(p) = \text{logit}(p)$$

$$p = \text{logistic}(\text{logit}(p))$$

逻辑回归——既不合逻辑，也不是回归！

进一步观察：由于拟合值$f(\mathbf{x})$大于零的程度越大，则离开类间边界、向正类纵深越远，被认为属于正类的可能性越大。反之，如果 $f(\mathbf{x})$小于零的程度越大，则离开类间边界、向负类纵深越远，被认为属于正类的可能性越小。

这意味着，$f(\mathbf{x})$与$p(y=1|\mathbf{x}) \equiv p(\mathbf{x})$之间不仅存在一一对应关系，而且单调性相同（同增同减），因此虽不适合对$p(\mathbf{x})$直接拟合，但是可以转而对$\text{logit}\,(p(\mathbf{x}))$进行线性拟合：$\text{logit}\,(p(\mathbf{x})) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$。

如果模型已经拟合好，则对一个新的$\mathbf{x}^\mathbf{0} = (x_1^0, x_2^0, \ldots x_n^0)$，先计算

$\text{logit}(p(y^0 = 1|\mathbf{x}^\mathbf{0})) = w_0 + w_1x_1^0 + w_2x_2^0 + \cdots + w_nx_n^0$，然后计算

$\text{logistic}\left(\text{logit}\left(p\left(y^0 = 1|\mathbf{x}^\mathbf{0}\right)\right)\right)$，这实际上就是$p\left(y^0 = 1|\mathbf{x}^\mathbf{0}\right)$！

判别：如果 $p\left(y^0 = 1|\mathbf{x}^\mathbf{0}\right) \geq 0.5$就判$\mathbf{x}^\mathbf{0}$为1（正类），否则判为0（负类）。

4.7.2 Logistic回归模型的求解(最大似然估计)

下面讨论如何根据训练数据, 对逻辑回归模型进行求解。

问题：希望用数据点$(\mathbf{x}, y)$的对数胜率$\mathrm{logit}(p(y|\mathbf{x}))$进行线性回归建模

$$\mathrm{logit}\big(p(y=1|\mathbf{x})\big) = w^T\mathbf{x} \equiv \boldsymbol{w}\cdot\mathbf{x}$$

由前面分析知，训练数据点$(\mathbf{x}, y)$在该模型 $\boldsymbol{w}$ 下属于正类的概率值为

$$p(y=1|\mathbf{x}) = \mathrm{logistic}(\boldsymbol{w}\cdot\mathbf{x}) = \frac{1}{1+\exp(-\boldsymbol{w}\cdot\mathbf{x}\ )} \equiv h(w;\mathbf{x}) \equiv h(\mathbf{x})$$

而属于负类的概率值为 $(1-h(\mathbf{x}))$。由于y 取值于$\{0, 1\}$，$y|\mathbf{x}$的概率分布(也即该数据点关于参数$\boldsymbol{w}$似然值)可以写为一个精致的表达式

$$p(y|\mathbf{x}) = p(y=1|\mathbf{x})^y[1-p(y=1|\mathbf{x})]^{1-y}$$

简记：$\pi(\mathbf{x}) \equiv p(y=1|\mathbf{x})$。因此，$(\mathbf{x}, y)$的似然函数为：

$$p(y|\mathbf{x}) = \pi(\mathbf{x})^y[1-\pi(\mathbf{x})]^{1-y}$$

根据模型拟合的最大似然原则，对于训练样本集$\{(\mathbf{x}_i, y_i)\}_{i=1}^N$，由于样本独立同分布(*i.i.d*, independent, identically distributed), 则训练样本集关于模型***w***的联合似然函数为：

$$L(\mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i) = \prod_{i=1}^{N} [\pi(\mathbf{x}_i)]^{y_i} [1-\pi(\mathbf{x}_i)]^{1-y_i}$$

这样就转化为模型***w***的最大似然参数估计问题。可采用梯度法或牛顿法对下面对数似然函数进行求解：

$$\begin{aligned} H(\mathbf{w}) &= \ln L(\mathbf{w}) \\ &= \sum_{i=1}^{N} [y_i \ln \pi(\mathbf{x}_i) + (1-y_i) \ln(1-\pi(\mathbf{x}_i))] \\ &= \sum_{i=1}^{N} [y_i \ln \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} + \ln(1-\pi(\mathbf{x}_i))] \\ &= \sum_{i=1}^{N} [y_i (\mathbf{w} \cdot \mathbf{x}_i) - \ln(1+\exp(\mathbf{w} \cdot \mathbf{x}_i))] \end{aligned}$$

$$\max_{\mathbf{w}} H(\mathbf{w}) = \max_{\mathbf{w}} \sum_{i=1}^{N} [y_i(\mathbf{w}\cdot\mathbf{x}_i) - \ln(1+\exp(\mathbf{w}\cdot\mathbf{x}_i))]$$

以上对数似然函数求极大值等价于对以下目标函数/准则函数$J(\mathbf{w})$求极小值：

$$J(\mathbf{w}) = -\sum_{i=1}^{N} [y_i(\mathbf{w}\cdot\mathbf{x}_i) - \ln(1+\exp(\mathbf{w}\cdot\mathbf{x}_i))]$$

$$= \sum_{i=1}^{N} [\ln(1+\exp(\mathbf{w}\cdot\mathbf{x}_i)) - y_i(\mathbf{w}\cdot\mathbf{x}_i)]$$

求目标函数$J(\mathbf{w})$对**w**的梯度如下：

$$\nabla J(\mathbf{w}) = \sum_{i=1}^{N} \left(\frac{\mathbf{x}_i}{1+\exp(\mathbf{w}\cdot\mathbf{x}_i)} - y_i\mathbf{x}_i \right) = \sum_{i=1}^{N} (\pi(\mathbf{x}_i) - y_i)\mathbf{x}_i$$

梯度下降法(取负梯度)求解最优解**w***/$\hat{\mathbf{w}}$ 迭代公式：

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \Delta\mathbf{w}$$

$$= \mathbf{w}(k) - \rho\nabla J(\mathbf{w})$$

$$= \mathbf{w}(k) - \rho\sum_{i=1}^{N} (\pi(\mathbf{x}_i) - y_i)\mathbf{x}_i$$

牛顿法求解最优解**w***/$\hat{\mathbf{w}}$ 迭代公式见文献[01]P59公式(3.29)-(3.31)(*, 了解)。

4.7.3 Logistic回归程序示例

例4.7 心脏病科研数据集Logistic回归。

1. 在Kaggle的Datasets页面中Search关键字”heart”就能找到该数据集， 可下载到本地解压的文件名为heart.csv。导入heart数据集，显示数据集前五条记录特征字段：

File Edit View Insert Cell Kernel Widgets Help 可信的 Python 3

运行 代码

In [1]:

```
import numpy as np # 导入NumPy数学工具箱
import pandas as pd # 导入Pandas数据处理工具箱
df_heart = pd.read_csv("d:/download/heart-dataset/heart.csv")  # 读取文件
df_heart.head() # 显示前5行数据
```

Out[1]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

age: 年龄, sex: 性别(1-男,0-女), cp: 胸痛类型, trestbps: 休息时血压, chol: 胆固醇, fbs: 血糖(1-超标,0-未超标), restecg: 心电图, thalach: 最大心率, exang: 运动后心绞痛(1-是,0-否), oldpeak: 运动后ST段低压, slope: 运动高峰期ST段的斜率, ca: 主动脉荧光造影染色数, thal: 缺陷各类, target: 标签字段(0-无心脏病,1-有心脏病)。

2. 用Pandas value_counts方法输出数据集中患心脏病和没有患心脏病的人数：

```
In [2]:  df_heart.target.value_counts() # 输出分类值及各个类别数目

Out[2]: 1    165
        0    138
        Name: target, dtype: int64
```

此步是必要的。因为如果某一类别比例特别低(如300个数据中只有3个人患病)，那么这样的数据集直接采用逻辑回归方法可能是不适宜的。

3. 构建特征集和标签集：

下面代码构建特征张量和标签张量，并输出张量的形状。

```
In [3]:  X = df_heart.drop(['target'], axis = 1) # 构建特征集
         y = df_heart.target.values # 构建标签集
         y = y.reshape(-1,1) # -1是相对索引，等价于len(y)
         print("张量X的形状:", X.shape)
         print("张量y的形状:", y.shape)

         张量X的形状: (303, 13)
         张量y的形状: (303, 1)
```

4. 按照8:2的比例准备训练集和测试集：

```
In [4]:  from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.2) #按照8:2的比例准备训练集和测试集
```

5. 导入Sklearn机器学习库的Logistic回归模型并进行训练和测试：

In [5]:

```
from sklearn.linear_model import LogisticRegression #导入逻辑回归模型
lr = LogisticRegression() # lr代表逻辑回归模型
lr.fit(X_train,y_train) # fit相当于梯度下降
print("Sklearn逻辑回归测试准确率{:.2f}%".format(lr.score(X_test,y_test)*100))
```

```
Sklearn逻辑回归测试准确率85.25%
```

4.8 非参数概率密度估计(*: 选学)

4.8.1 基本方法 (直方图)

根据样本直接估计类概率密度函数的方法。

1. 出发点：基于事实

随机向量$\boldsymbol{X}$落入区域R的概率P为：$P=\int_R p(\boldsymbol{X})d\boldsymbol{X}$ 。

$p(\boldsymbol{X})$：样本$\boldsymbol{X}$的概率密度函数。

上式表明, 概率P是密度函数$p(\boldsymbol{X})$的一种平均形式, 对P的估计就是估计出$p(\boldsymbol{X})$的这个平均值。

设从密度函数为$p(\boldsymbol{X})$的总体中独立抽取的样本$\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_N$。若观察到：$N$个样本中有$k_N$个落入区域$R$中，则可以合理认为

$$\hat{P}\approx k_N/N$$

其中$\hat{P}$：$\boldsymbol{X}$落入区域R中概率P的估计。

类概率密度函数$p(X)$的估计：

设$p(\boldsymbol{X})$连续，区域R足够小且体积为V， $p(\boldsymbol{X})$在R中没有变化，$\boldsymbol{X}$是R中的点。设$\hat{p}(X)$为$\boldsymbol{X}$点概率密度的估计

$$P = \int_R p(\boldsymbol{X})d\boldsymbol{X} = p(\boldsymbol{X})V$$

因此有：$\hat{P} @ \int_R \hat{p}(\boldsymbol{X})dX \approx \hat{p}(\boldsymbol{X})V$,

前面有：$\hat{P} \approx \dfrac{k}{N}$

于是有：$\hat{p}(\boldsymbol{X}) \approx \dfrac{k/N}{V}$ (4-8-1)

2. 存在两个问题

1) 固定V，样本数增多，则k/N以概率1收敛。但只能得到在某一体积V中的平均估计。

2) N固定，V趋于零， $p(\boldsymbol{X}) \approx 0$或发散到无穷大。没有意义。

必须注意V、k、k/N 随N变化的趋势和极限，保持式(4-8-1)合理性。

在式(4-8-1)中，如果固定R，则体积V固定，样本数$N \to \infty$，则$k/N \to P$，此时：

$$p(\mathbf{X}) \to \frac{P}{V} = \frac{\int_{\mathrm{R}} p(\mathbf{X})d\mathbf{X}}{\int_{R} d\mathbf{X}}$$

即上式得到的是概率密度函数$p(\boldsymbol{X})$的空间平均估计值。

要想得到概率密度函数$p(\boldsymbol{X})$, 而不是$p(\boldsymbol{X})$的空间平均估计值, 就需要让R的体积V趋近于0。 若把样本数N固定, 令V趋于0, 以至于R不包含任何样本, 此时, $p(\boldsymbol{X}) \approx 0$, 这种估计是没有意义的; 或者恰有一个或几个样本同$\mathbf{X}$重合, 此时, $p(\boldsymbol{X})$为无穷大, 同样也没有意义。

3. 估计的步骤:

* 构造一个包含$\boldsymbol{X}$的区域序列R_1，R_2，…，R_N，…

* 对R_1采用一个样本估计，对R_2采用两个样本，……

* 假定N时刻的样本数为N, R_N的体积是V_N，落入R_N中的样本数目是k_N，$\hat{p}_N(\boldsymbol{X})$是$p(\boldsymbol{X})$的第$N$次估计，有:

$$\hat{p}_N(\boldsymbol{X}) = \frac{k_N/N}{V_N} \tag{4-8-2}$$

4. 为确保估计合理性应满足的三个条件

1) $\lim\limits_{N\to\infty} V_N = 0$ —— $\hat{p}_N(\boldsymbol{X})$能代表$\boldsymbol{X}$点的密度$p(\boldsymbol{X})$

2) $\lim\limits_{N\to\infty} k_N = \infty$ —— 使式右边能以概率1收敛于$p(\boldsymbol{X})$

3) $\lim\limits_{N\to\infty} k_N/N = 0$ —— 落入R_N中的样本数始终是总数中的极小部分

5. 两种非参数估计法： Parzen窗估计法、 k_N -近邻估计法

满足前面三个条件的区域序列主要有两种选择方法：

(1) **Parzen窗法**。 选定一个中心在$\boldsymbol{X}$处的区域R_N, 其体积为V_N(例如$V_N = 1/\sqrt{N}$, 然后计算落入其中的样本数k_N, 用来估计局部密度$p_N(\boldsymbol{X})$的值。

(2) **k_N -近邻法**。 选定一个k_N值(例如$k_N = \sqrt{N}$), 以$\boldsymbol{X}$为中心构造一个区域R_N, 其体积为V_N, 使R_N恰好包含k_N个样本, 这时的体积V_N用来估计$p_N(\boldsymbol{X})$。

4.8.2 Parzen窗法[02][03]
(核密度估计/KDE, Kernel Density Estimation)

1. Parzen窗估计的基本概念

设区域R_N：d维超立方体，边长：h_N，则

$$V_N = h_N^d$$

定义窗函数$\phi(\mathrm{u})$：

$$\phi(u) = \begin{cases} 1, & \text{当} \ |u_j| \le \dfrac{1}{2};\ j = 1, 2, \cdots, d \\ 0, & \text{其它} \end{cases}$$

以原点为中心的超立方体

其中$\mathbf{u} = (u_1, u_2, \cdots, u_d)$

当$\boldsymbol{X}_i$落入以$\boldsymbol{X}$为中心，体积为V_N的超立方体时：

$$\phi[(\boldsymbol{X} - \boldsymbol{X}_i)/h_N] = 1$$

否则

$$\phi[(\boldsymbol{X} - \boldsymbol{X}_i)/h_N] = 0$$

落入以$\boldsymbol{X}$为中心的超立方体V_N内的样本$\boldsymbol{X}_i$的个数为

$$k_N = \sum_{i=1}^{N} \phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right) \qquad (4\text{-}8\text{-}3)$$

代入 $\hat{p}_N(\boldsymbol{X}) = \dfrac{k_N/N}{V_N}$ 得

$$\hat{p}_N(\boldsymbol{X}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right) \qquad (4\text{-}8\text{-}4)$$

—— Parzen窗法基本公式

实质：

窗函数的作用是平滑，样本$\boldsymbol{X}_i$对$\boldsymbol{X}$处的密度的估计所起的作用，取决于它 ($\boldsymbol{X}_i$) 到$\boldsymbol{X}$的距离。

为使 $\hat{p}_N(X)$成为密度函数，$\phi(u)$应满足的两个条件：

1) $\phi(u) \geq 0$；　　2) $\int \phi(u)du = 1$。

2. 窗函数的选择

一维形式

1) 方窗函数

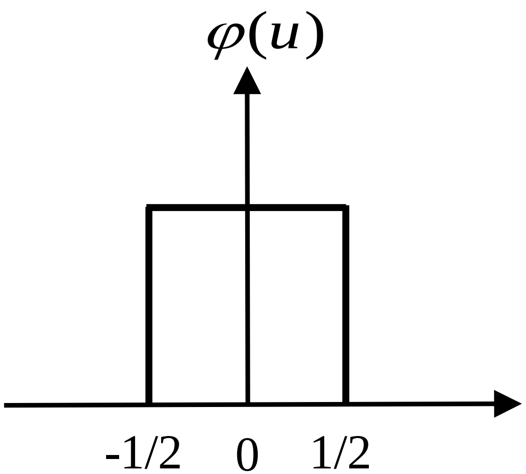
$$\phi(u)=\begin{cases}1, & |u|\le \frac{1}{2}\\ 0, & \text{其它}\end{cases}$$

2) 正态窗函数

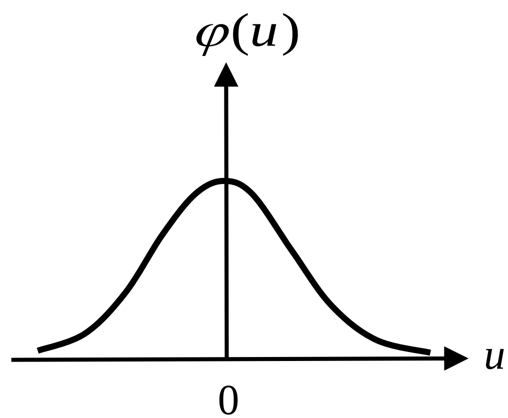
$$\phi(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$$

3) 指数窗函数

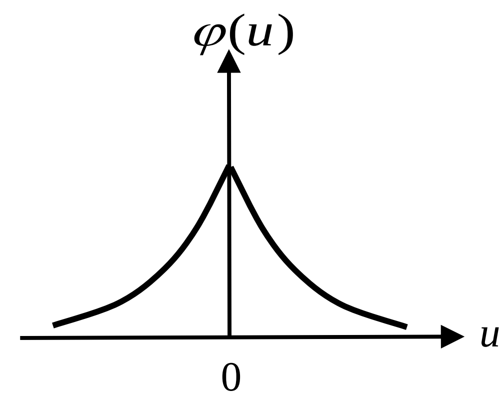
$$\phi(u)=\exp(-|u|)$$



（a）



（b）



（c）

满足条件 $\phi(u)\ge 0$ 和 $\int \phi(u)du=1$ 的都可以作为窗函数。

最终估计效果的好坏与样本情况、窗函数以及窗函数参数的选择有关。

3. 窗宽 h_N 对估计量 $\hat{p}_N(X)$ 的影响

定义 $$\delta_N(\boldsymbol{X}) = \frac{1}{V_N}\phi\left(\frac{\boldsymbol{X}}{h_N}\right)$$

有 $$\hat{p}_N(\boldsymbol{X}) = \frac{1}{N}\sum_{i=1}^{N}\delta_N(\boldsymbol{X} - \boldsymbol{X}_i)$$

h_N 影响 $\delta_N(\boldsymbol{X})$ 的幅度，对 $\hat{p}_N(X)$ 的影响很大。

如何选取根据经验折中考虑。

4. 估计量 $\hat{p}_N(X)$ 的统计性质

满足某些限制条件时，$\hat{p}_N(\boldsymbol{X})$ 渐近无偏和平方误差一致。

限制条件：

1) 总体的密度函数 $p(\boldsymbol{X})$ 在 $\boldsymbol{X}$ 点连续；

2) 窗函数满足以下条件：

$$\left.\begin{aligned} &\phi(u) \geq 0 \\ &\int \phi(u) du = 1 \end{aligned}\right\}$$

保证 $\hat{p}_N(\boldsymbol{X})$ 有密度函数的性质

$$\sup_u \phi(u) < \infty$$

保证 $\phi(u)$有界

$$\lim_{\|u\| \to \infty} \phi(u) \prod_{i=1}^{d} u_i = 0$$

使 $\phi(u)$随u的增加较快趋于零

3) 窗函数受下列条件的约束：

$$\left.\begin{aligned} &\lim_{N \to \infty} V_N = 0 \\ &\lim_{N \to \infty} N V_N = \infty \end{aligned}\right\}$$

使体积随N的增大趋于零时，速度低于N增加的速度

Parzen窗法特点：

* 具有一般性，适用于单峰、多峰形式。

* 要得到较精确的估计必须抽取大量的样本。

（一般非参数估计法的共同问题）

比参数估计法多得多；

样本数目随模式维数一般呈指数规律增长。

(*: 了解)

Parzen窗估计(KDE)法Python案例实践：可参考CSND博文链接：https://yuanyx.blog.csdn.net/article/details/115175706**。**

本章小结：

1.贝叶斯定理(Bayes theorem)

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

$$\boxed{P(\omega_i \mid \mathbf{x}) = P(\mathbf{x} \mid \omega_i)P(\omega_i) / P(\mathbf{x}) = P(\mathbf{x} \mid \omega_i)P(\omega_i) \Big/ \sum_{j=1}^{M} P(\mathbf{x} \mid \omega_j)P(\omega_j)}$$

$P(\omega_i|x)$ 称为后验概率，对模式识别而言可理解为x来自ω_i类的概率，即 x已知的情况下其类别属于ω_i的概率为$P(\omega_i|x)$；$P(x|\omega_i)$称为类条件概率密度(简称似然)； $P(\omega_i)$称为先验概率。

2.最小错误率Bayes决策/最大后验Bayes决策

3.最小风险Bayes决策

$$R(\alpha_i \mid \mathbf{x}) @ E\left[\lambda\left(\alpha_i, \omega_j\right)\right] = \sum_{j=1}^{M} \lambda\left(\alpha_i, \omega_j\right) P\left(\omega_j \mid \mathbf{x}\right)$$

若$R(\alpha_k \mid \mathbf{x}) = \min\limits_{i=1,2,L\ ,M} R(\alpha_i \mid \mathbf{x})$，则判决$\mathbf{x} \in \omega_k$。

4. Naïve Bayes决策

5. 正态分布Bayes决策

6. 最大似然估计

求最大似然函数主要步骤:

(1)写出似然函数或对数似然函数

$$L(\boldsymbol{\theta}) = p(X^N | \boldsymbol{\theta}) \text{或} H(\boldsymbol{\theta}) = \ln p(X^N | \boldsymbol{\theta})$$

(2)对似然函数或对数似然函数求偏导，令 $\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$ 或 $\frac{\partial H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$，

求出$\boldsymbol{\theta}$出最大似然估计(解析法)。

或者梯度下降法等迭代法求出$\boldsymbol{\theta}$最大似然估计。

7. Parzen窗估计/KDE (*: 选学)

除了本课件中介绍的Parzen窗估计(KDE)方法，KDE的Python案例实践参考CSND博文链接：https://yuanyx.blog.csdn.net/article/details/115175706**。**

8. 最大似然估计在Logistic回归模型训练中的应用

$$H(\mathbf{w}) = \ln L(\mathbf{w}) = \sum_{i=1}^{N}[y_i(\mathbf{w}\cdot\mathbf{x}_i) - \ln(1+\exp(\mathbf{w}\cdot\mathbf{x}_i))]$$

$$J(\mathbf{w}) = -H(\mathbf{w}) = \sum_{i=1}^{N}[\ln(1+\exp(\mathbf{w}\cdot\mathbf{x}_i)) - y_i(\mathbf{w}\cdot\mathbf{x}_i)]$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} H(\mathbf{w}) \Leftrightarrow \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$$

目标函数/准则函数$J(\mathbf{w})$对$\mathbf{w}$的梯度：

$$\nabla J(\mathbf{w}) = \sum_{i=1}^{N}\left(\frac{\mathbf{x}_i}{1+\exp(\mathbf{w}\cdot\mathbf{x}_i)} - y_i\mathbf{x}_i\right) = \sum_{i=1}^{N}(\pi(\mathbf{x}_i) - y_i)\mathbf{x}_i$$

用梯度下降法迭代训练求$\mathbf{w}$最优解$\hat{\mathbf{w}}$：

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \rho\nabla J(\mathbf{w})$$

$$= \mathbf{w}(k) - \rho\sum_{i=1}^{N}(\pi(\mathbf{x}_i) - y_i)\mathbf{x}_i$$

课外作业3-4(概率分类):

1. 就分类而言，在模型训练前，对一个样本所来自的总体的类别变量分布的先验知识，与在模型训练后对该类别变量分布的后验知识，主要区别是什么？在什么意义下，两种是一致的？

2. 朴素贝叶斯分类方法的基本假设是什么？试举一例说明该假设的合理性。

3. 假设在某个地区的疾病普查中，异常细胞（ω_1）和正常细胞（ω_2）的先验概率分别为 $P(\omega_1)=0.1$， $P(\omega_2)=0.9$。现有一待识别细胞，其观察值为 X，从类概率密度分布曲线上查得 $p(X\mid\omega_1)=0.4$， $p(X\mid\omega_2)=0.2$ 试对该细胞利用最小错误率贝叶斯决策规则进行分类。

4. 对前一题中两类细胞的分类问题（异常细胞 ω_1，正常细胞 ω_2），除已知的数据外，若损失函数的值分别为 $L_{11}=0$， $L_{21}=6$， $L_{12}=1$， $L_{22}=0$，试用最小风险贝叶斯决策规则对细胞进行分类。

5. 考虑下表中的数据集。

样本序号	A	B	C	类别
1	0	0	1	−
2	1	0	1	+
3	0	1	0	−
4	1	0	0	−
5	1	0	1	+
6	0	0	1	+
7	1	1	0	−
8	0	0	0	−
9	0	1	0	+
10	1	1	1	+

（1）估计以下条件概率

$P(A=1|+)$，$P(B=1|+)$，$P(C=1|+)$，$P(A=1|-)$，$P(B=1|-)$，$P(C=1|-)$。

（2）根据估计的条件概率，使用朴素贝叶斯方法预测样本($A=1$，$B=1$，$C=1$)的类别。

（3）比较$P(A=1)$，$P(B=1)$和 $P(A=1，B=1)$，陈述变量 A、B 之间的统计关系。

（4）比较$P(A=1|+)$，$P(B=1|+)$和 $P(A=1，B=1|+)$，给定类 +， 变量 A、B 条件独立吗？

课外作业3-4(概率分类)参考答案

1. 答:

(1)关于参数（这里是样本所属总体的类别变量的分布参数）的先验知识是指在进行实验前获得的关于参数分布的信息，而关于参数的后验分布则是值在结合了样本中关于总体参数的信息后关于参数分布的信息。在这个意义上，先验信息和后验信息是一致的，都是关于参数的分布知识，区别就在于是否利用了某个样本或样本集。因此，谈先验、后验之分需要针对某个样本集来说。

(2) 昨天的后验就是今天的先验，今天的先验就是明天的后验！

2. 答：

基本假设就是在已知样本的类别时，样本的各个特征（属性）之间是条件独立的。比如，设“一个人是否高收入”为样本的类别（Y=1表示高收入），同时设“一个人是否经常购买高档白酒”为属性X1（X1=1表示经常购买高档白酒），设“一个人是否经常购买高档服装”为属性X2（X2=1表示经常购买高档服装）。（1）在不知道一个人的收入情况下，也就是说在人群总体中进行一般调查时，从计算角度看，属性X1和属性X2肯定存在一定的“相关关系”（也就是说，肯定有部分人员既经常购买高档白酒，也经常购买高档服装），因此不能排除这两个属性在人群总体中存在一定的依赖关系。（2）但是在知道了一个人是高收入的情况下，“是否经常购买高档白酒”与“是否经常购买高档服装”就是互不依赖的，因为此时这两个属性都是真正的原因“高收入”引起的，而不是相互引起的（经常买高品质白酒的人，并一定会经常买高档衣服，反之亦然），因此这两个属性，在已知“一个人是高收入”的条件下，可以认为是条件独立的。

3. 解：

解(1)： $$P(\omega_2 \mid X)=\frac{p(X \mid \omega_2)P(\omega_2)}{\sum_{i=1}^{2} p(X \mid \omega_i)P(\omega_i)}=\frac{0.2\times 0.9}{0.2\times 0.9+0.4\times 0.1}\approx 0.818$$

$$P(\omega_1 \mid X)=\frac{0.4\times 0.1}{0.2\times 0.9+0.4\times 0.1}\approx 0.182$$

$$\because P(\omega_2 \mid X)>P(\omega_1 \mid X)$$

$\therefore X \in \omega_2$（正常）

解(2)： $$p(X \mid \omega_2)P(\omega_2)=0.2\times 0.9=0.18$$

$$p(X \mid \omega_1)P(\omega_1)=0.4\times 0.1=0.04$$

$$\because p(X \mid \omega_2)P(\omega_2)>p(X \mid \omega_1)P(\omega_1)$$

$\therefore X \in \omega_2$ （正常）

4. 解(1)：当 X 被判为 ω_1 类时：

$$d_1(X) = L_{11}p\left(X \mid \omega_1\right)P(\omega_1) + L_{12}p\left(X \mid \omega_2\right)P(\omega_2)$$

$$= 0 \times 0.4 \times 0.1 + 1 \times 0.2 \times 0.9 = 0.18$$

当 X 被判为 ω_2 类时：

$$d_2(X) = L_{21}p\left(X \mid \omega_1\right)P(\omega_1) + L_{22}p\left(X \mid \omega_2\right)P(\omega_2)$$

$$6 \times 0.4 \times 0.9 + 0 \times 0.2 \times 0.1 = 2.16$$

Q $d_1(X) < d_2(X)$， $\therefore X \in \omega_1$ （异常）

解(2)： $l_{12}(X) = \dfrac{p(X \mid \omega_1)}{p(X \mid \omega_2)} = \dfrac{0.4}{0.2} = 2$

$$\theta_{12} = \frac{(L_{12} - L_{22})P(\omega_2)}{(L_{21} - L_{11})P(\omega_1)} = \frac{(1-0) \times 0.9}{(6-0) \times 0.1} = 1.5$$

Q $l_{12}(X) > \theta_{12}$， $\therefore X \in \omega_1$ （异常）

5.解：（1）根据数据集计算的:

$P(A=1|+)=0.6$, $P(B=1|+)=0.4$, $P(C=1|+)=0.8$, $P(A=1|-)=0.4$, $P(B=1|-)=0.4$, and $P(C=1|-)=0.2$

（2）记R:（$A=1,B=1,C=1$）为测试样本。为计算$P(+|R),P(-|R)$，根据贝叶斯公式，需计算$P(+),P(-),P(R|+),\mathrm{P}(R|-)$。根据数据集可以计算:

$P(+)=P(-)=0.5$，而

$$P(R|+)=P(A=1|+)\times P(B=1|+)\times P(C=1|+)=0.192$$
$$P(R|-)=P(A=1|-)\times P(B=1|-)\times P(C=1|-)=0.032$$

于是有：$P(+|R)>P(-|R)$，因此该测试样本应该判为类+。

（3）$P(A=1)=0.5$，$P(B=1)=0.4$，$P(A=1,B=1)=0.2$，因此有

$$P(A=1,B=1)=P(A=1)\times P(B=1)$$

因此，可认为A与B是相互独立的。

（4）根据数据集计算得:

$$P(A=1|+)=0.6$$
$$P(B=1|+)=0.4$$
$$P(A=1,B=1|+)=0.2$$

此时:

$$P(A=1,B=1|+)\neq P(A=1|+)\times P(B=1|+)$$

因此可以认为，在给定了类别+的条件下，变量A和变量B并不统计独立。

- **本章主要参考文献**

[01] 周志华，机器学习-对数几率回归: **57-60**, 清华版，2016

[02] 齐敏等编，模式识别导论-**Parzen窗法: 117-121**，清华版，2009

[03] 李弼程等编, 模式识别原理与应用-**Parzen窗法: 36-42**, 西电版, 2008

[04] 核密度估计及其Python实践，CSND博文：
https://yuanyx.blog.csdn.net/article/details/115175706

[05] 埃塞姆•阿培丁，机器学习导论(第3版)， 范明 译，机工版，2016

[06] 李航，统计学习方法(第2版)，清华版，2019

[07] Andrew Gelman et al. Bayesian Data Analysis(3rd ed), CRC Press, 2020

End of this lecture.

Thanks !