- **贝叶斯判别**

根据概率判别规则，有：

若 $P(\omega_1|\boldsymbol{x})>P(\omega_2|\boldsymbol{x})$，则 $\boldsymbol{x}\in\omega_1$

若 $P(\omega_1|\boldsymbol{x})<P(\omega_2|\boldsymbol{x})$，则 $\boldsymbol{x}\in\omega_2$

由贝叶斯定理，后验概率 $P(\omega_i|\boldsymbol{x})$可由类别ω$_i$的先验概率 $P(\omega_i)$和 $\boldsymbol{x}$ 的条件概率密度 $p(\boldsymbol{x}|\omega_i)$来计算，即：

$$P(\omega_i|\boldsymbol{x})=\frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})}=\frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{\sum_{i=1}^{2}p(\boldsymbol{x}|\omega_i)P(\omega_i)}$$

这里 $p(\boldsymbol{x}|\omega_i)$也称为似然函数。将该式代入上述判别式，有：

若 $p(\boldsymbol{x}|\omega_1)P(\omega_1)>p(\boldsymbol{x}|\omega_2)P(\omega_2)$，

则 $\boldsymbol{x}\in\omega_1$

若 $p(\boldsymbol{x}|\omega_1)P(\omega_1)<p(\boldsymbol{x}|\omega_2)P(\omega_2)$，

则 $\mathrm{x}\in\omega_2$

或

若 $l_{12}(\boldsymbol{x})=\dfrac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)}>\dfrac{P(\omega_2)}{P(\omega_1)}$， 则 $\boldsymbol{x}\in\omega_1$

若 $l_{12}(\boldsymbol{x})=\dfrac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)}<\dfrac{P(\omega_2)}{P(\omega_1)}$， 则 $\boldsymbol{x}\in\omega_2$

其中，l_{12} 称为似然比，$P(\omega_2)/P(\omega_1)=\theta_{21}$ 称为似然比的判决阈值，此判别称为贝叶斯判别。

贝叶斯定理：

条件概率

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

全概公式：

$$P(A)=P(A\cap B)+P(A\cap B^C)=P(A|B)\times P(B)+P(A|B^c)\times P(B^c)$$

- **贝叶斯判别计算实例**

已知：$P(\omega_1)=0.2$，$P(\omega_2)=0.8$，

$P\,(x=\text{异常}|\omega_1)=0.6$，$P\,(x=\text{正常}|\omega_1)=0.4$，

$P\,(x=\text{异常}|\omega_2)=0.1$，$P\,(x=\text{正常}|\omega_2)=0.9$

利用贝叶斯公式，有：

$$\begin{aligned}P(\omega_1\,|\,x=\text{异常})&=\frac{P(x=\text{异常}\,|\,\omega_1)P(\omega_1)}{P(x=\text{异常})}\\&=\frac{P(x=\text{异常}\,|\,\omega_1)P(\omega_1)}{P(x=\text{异常}\,|\,\omega_1)P(\omega_1)+P(x=\text{异常}\,|\,\omega_2)P(\omega_2)}\\&=\frac{0.6\times0.2}{0.6\times0.2+0.8\times0.1}=0.6\end{aligned}$$

似然比：$l_{12}=\dfrac{P(\mathrm{x}=\text{异常}|\omega_1)}{P(\mathrm{x}=\text{异常}|\omega_2)}=\dfrac{0.6}{0.1}=6$

判决阈值：$\theta_{21}=\dfrac{\mathrm{P}(\omega_2)}{\mathrm{P}(\omega_1)}=\dfrac{0.8}{0.2}=4$

$l_{12}>\theta_{21}$，所以判断为第一类，即地震

- **最小平均条件风险表达式**

按贝叶斯公式，最小平均条件风险可写成：

$$r_j(\boldsymbol{x})=\frac{1}{p(\boldsymbol{x})}\sum_{i=1}^{M}L_{ij}\,p(\boldsymbol{x}\,|\,\omega_i)P(\omega_i)$$

因 1/p(x)为公共项，可舍去，因此可简化为：

$$r_j(\boldsymbol{x})=\sum_{i=1}^{M}L_{ij}\,p(\boldsymbol{x}\mid\omega_i)P(\omega_i)$$

这也是贝叶斯分类器，只是它的判别方法不是按错误概率最小作为标准，而是按平均条件风险作为标准。

- **两类（*M*=2）情况的贝叶斯最小风险判别**

选 M=2，即全部的模式样本只有ω$_1$和ω$_2$两类，要求分类器将模式样本分到ω$_1$和ω$_2$两类中，则平均风险可写成：

当分类器将 $\boldsymbol{x}$ 判别为ω$_1$时：

$$r_1(\boldsymbol{x})=L_{11}p(\boldsymbol{x}|\omega_1)P(\omega_1)+L_{21}p(\boldsymbol{x}|\omega_1)P(\omega_2)$$

当分类器将 x 判别为ω$_2$时：

$$r_2(\boldsymbol{x})=L_{12}p(\boldsymbol{x}\mid\omega_1)P(\omega_1)+L_{22}p(\boldsymbol{x}\mid\omega_2)P(\omega_2)$$

若 $r_1(x)<r_2(x)$，则 x 被判定为属于ω$_1$，此时：

$$L_{11}p(\boldsymbol{x}\mid\omega_1)P(\omega_1)+L_{21}p(\boldsymbol{x}\mid\omega_2)P(\omega_2)<$$
$$L_{12}p(\boldsymbol{x}\mid\omega_1)P(\omega_1)+L_{22}p(\boldsymbol{x}\mid\omega_2)P(\omega_2)$$

即

$$(L_{21}-L_{22})p(\boldsymbol{x}\mid\omega_2)P(\omega_2)<(L_{12}-L_{11})p(\boldsymbol{x}\mid\omega_1)P(\omega_1)$$

通常取 $L_{ij}>L_{ii}$，有：

当$\dfrac{p(\boldsymbol{x}\mid\omega_1)}{p(\boldsymbol{x}\mid\omega_2)}>\dfrac{P(\omega_2)}{P(\omega_1)}\cdot\dfrac{L_{21}-L_{22}}{L_{12}-L_{11}}$时， $\boldsymbol{x}\in\omega_1$

该式左边为似然比：$l_{12}=\dfrac{p(\boldsymbol{x}\mid\omega_1)}{p(\boldsymbol{x}\mid\omega_2)}$

右边为阈值：$\theta_{21}=\dfrac{P(\omega_2)}{P(\omega_1)}\cdot\dfrac{L_{21}-L_{22}}{L_{12}-L_{11}}$

故得两类模式的贝叶斯判别条件为：

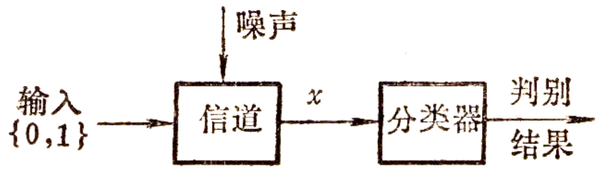
（1）若 $l_{12}(\boldsymbol{x}) > \theta_{21}$，则 $\boldsymbol{x} \in \omega_1$

（2）若 $l_{12}(\boldsymbol{x}) < \theta_{21}$，则 $\boldsymbol{x} \in \omega_2$

（3）若 $l_{12}(\boldsymbol{x}) = \theta_{21}$，则可做任意判别。

通常，当判别正确时，不失分，可选常数 $L_{11}= L_{22} = 0$；判别错误时，可选 $L_{12} = L_{21} = 1$，此时 $\theta_{21} = \dfrac{P(\omega_2)}{P(\omega_1)}$。

- **两类（*M*=2）情况的贝叶斯最小风险判别实例**



如图所示为一信号通过一受噪声干扰的信道。

信道输入信号为 0 或 1，噪声为高斯型，其均值 $\mu=0$，方差为 σ^2。

信道输出为 x，试求最优的判别规则，以区分 x 是 0 还是 1。

设送 0 为 ω_1 类，送 1 为 ω_2 类，从观察值 x 的基础上判别它是 0 还是 1。直观上可以看出，若 x<0.5 应判为 0，x>0.5 应判为 1。用贝叶斯判别条件分析：设信号送 0 的先验概率为 $P(0)$，送 1 的先验概率为 $P(1)$，$\boldsymbol{L}$ 的取值为：

$$L = \begin{array}{c c} & \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{array}{c} \omega_1 \\ \omega_2 \end{array} & \begin{pmatrix} 0 & L_{12} \\ L_{21} & 0 \end{pmatrix} \end{array}$$

这里 a_1 和 a_2 分别对应于输入状态为 0 和 1 时的正确判别，L_{12} 对应于实际上是 ω_1 类但被判成 ω_2 类(a_2)时的代价，L_{21} 对应于实际上是 ω_2 类但被判成 ω_1 类(a_1)时的代价。正确判别时 $\boldsymbol{L}$ 取 0。

当输入信号为 0 时，受噪声为正态分布 $N(0,\ \sigma^2)$的干扰，其幅值大小的概率密度为：

$$p(x\mid\omega_1)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}$$

当输入信号为 1 时：$p(x\mid\omega_2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-1)^2}{2\sigma^2}}$

则似然比为：$l_{12}=\frac{p(x\mid\omega_1)}{p(x\mid\omega_2)}=e^{\frac{1-2x}{2\sigma^2}}$，

若 $l_{12}>\theta_{21}$，即 $e^{\frac{1-2x}{2\sigma^2}}>\theta_{21}\Rightarrow x<\frac{1}{2}-\sigma^2\ln\theta_{21}$，（$\theta_{21}=\frac{P(\omega_2)}{P(\omega_1)}\cdot\frac{L_{21}-L_{22}}{L_{12}-L_{11}}$）则 $x\in\omega_1$，此时信号应是 0，即

$$x<\frac{1}{2}-\sigma^2\ln\left(\frac{L_{21}}{L_{12}}\cdot\frac{P(1)}{P(0)}\right)$$

若取 $L_{21}=L_{12}=1$，$P(1)=P(0)$,则 $x<1/2$ 判为 0。

若无噪声干扰，即 $\sigma^2=0$，则 $x<1/2$ 判为 0。

- **多类（*M* 类）情况的贝叶斯最小风险判别**

对于 M 类情况，若 $r_i(\boldsymbol{x})<r_j(\boldsymbol{x}),\ j=1,2,\ldots,M,\ j\neq i$，则 $x\in\omega_i$。

$\boldsymbol{L}$ 可如下取值（仍按判对失分为 0，判错失分为 1 记）：

$$L_{ij}=\begin{cases}0 & when\ i=j\\ 1 & when\ i\neq j\end{cases}$$

则条件平均风险可写成：

$$
\begin{aligned}
r_j(\boldsymbol{x}) &= \sum_{i=1}^{M} L_{ij}\, p(\boldsymbol{x}\,|\,\omega_i)P(\omega_i) \\
&= L_{1j}\, p(\boldsymbol{x}\,|\,\omega_1)P(\omega_1) + \cdots + L_{jj}\, p(\boldsymbol{x}\,|\,\omega_j)P(\omega_j) + \cdots + L_{Mj}\, p(\boldsymbol{x}\,|\,\omega_M)P(\omega_M) \\
&= \sum_{i=1}^{M} p(\boldsymbol{x}\,|\,\omega_i)P(\omega_i) - p(\boldsymbol{x}\,|\,\omega_j)P(\omega_j) \\
&= p(\boldsymbol{x}) - p(\boldsymbol{x}\,|\,\omega_j)P(\omega_j)
\end{aligned}
$$

由 $r_i(\boldsymbol{x}) < r_j(\boldsymbol{x})$，有当 $p(\boldsymbol{x}\,|\,\omega_i)P(\omega_i) > p(\boldsymbol{x}\,|\,\omega_j)P(\omega_j)$ 时，$\boldsymbol{x} \in \omega_i$，对应于判别函数为：取 $d_i(\boldsymbol{x}) = p(\boldsymbol{x}\,|\,\omega_i)P(\omega_i), i = 1, 2, \ldots, M$，则对于全部 $j \neq i$ 的值，若 $d_i(\boldsymbol{x}) > d_j(\boldsymbol{x})$，则 $\boldsymbol{x} \in \omega_i$。

- **M 种模式类别的多变量正态类密度函数**

具有 M 种模式类别的多变量正态类密度函数为：

$$
p(\boldsymbol{x}\,|\,\omega_i) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{C}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^T \boldsymbol{C}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i)\right\}, i = 1, 2, \cdots, M
$$

其中，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：

$$\boldsymbol{m}_i = E_i\{\boldsymbol{x}\}$$

$$\boldsymbol{C}_i = E_i\{(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T\}$$

$\boldsymbol{m}_i = E_i(\boldsymbol{x})$ 表示对类别属于 ω_i 的模型的数学期望。

在上述公式中，$\boldsymbol{x}$ 是 n 为列向量，$|\boldsymbol{C}_i|$ 为矩阵 $\boldsymbol{C}_i$ 的行列式，协方差矩阵 $\boldsymbol{C}_i$ 是对称的正定矩阵，其对角线上的元素 C_{kk} 是模式向量第 k 个元素的方差，非对角线上的元素 C_{jk} 是 $\boldsymbol{x}$ 的第 j 个分量 x_j 和第 k 个分量 x_k 的协方差。当 x_j 和 x_k 统计独立时，$C_{jk}=0$。当协方差矩阵的全部非对角线上的元素都为零时，多变量正态类密度函数可简化为 n 个单

变量正态类密度函数的乘积。

已知类别 ω_i 的**判别函数**可写成如下形式：

$$d_i(\boldsymbol{x}) = p(\boldsymbol{x} \mid \omega_i)P(\omega_i), i = 1, 2, \cdots, M$$

对于正态密度函数，可取自然对数的形式以方便计算（因为自然对数是单调递增的，取对数后不影响相应的分类性能），则有：

$$d_i(\boldsymbol{x}) = \ln p(\boldsymbol{x} \mid \omega_i) + \ln P(\omega_i), i = 1, 2, \cdots, M$$

代入正态类密度函数，有：

$$d_i(\boldsymbol{x}) = \ln P(\omega_i) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\boldsymbol{C}_i\right|$$
$$-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)\boldsymbol{C}_i^{-1}(\boldsymbol{x} - \boldsymbol{m}_i), i = 1, 2, \cdots, M$$

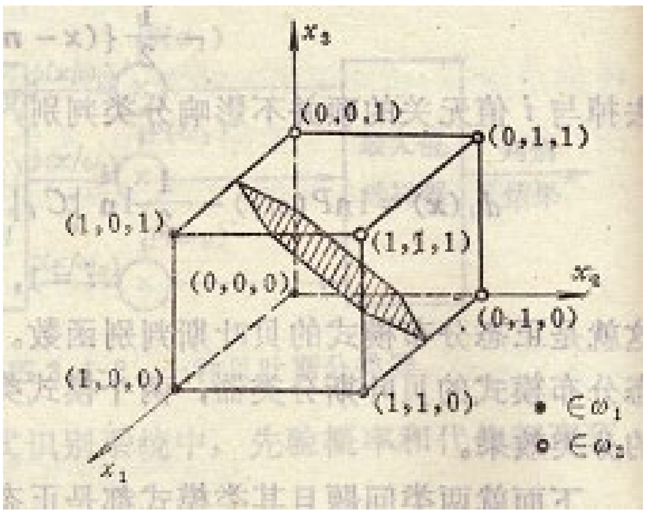
去掉与 i 无关的项（并不影响分类结果），有：

$$d_i(\boldsymbol{x}) = \ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{C}_i\right| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)\boldsymbol{C}_i^{-1}(\boldsymbol{x} - \boldsymbol{m}_i), i = 1, 2, \cdots, M$$

即为正态分布模式的贝叶斯判别函数。

- **两类问题且其类模式都是正态分布的实例**

$P(\omega_1)=P(\omega_2)=1/2$，求其判别界面。



模式的均值向量 $\boldsymbol{m}_i$ 和协方差矩阵 $\boldsymbol{C}_i$ 可用下式估计：

$$m_i = \frac{1}{N_i}\sum_{x^j \in \omega_i} x^j \quad i = 1,2$$

$$C_i = \frac{1}{N_i}\sum_{x^j \in \omega_i} (x^j - m_i)(x^j - m_i)^T \quad i = 1,2$$

其中 N_i 为类别 ω_i 中模式的数目。由上式可求出：

$$m_1 = \frac{1}{4}(3\ 1\ 1)^T$$

$$m_2 = \frac{1}{4}(1\ 3\ 3)^T$$

$$C_1 = C_2 = C = \frac{1}{16}\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{pmatrix}, \quad C^{-1} = 4\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}$$

设 $P(\omega_1)=P(\omega_2)=1/2$，因 $C_1=C_2$，

根据两类问题且其类模式都是正态分布（协方差矩阵相同）时的判别界面方程：

$$d_1(x) - d_2(x) = \ln P(\omega_1) - \ln P(\omega_2) + (m_1 - m_2)^T C^{-1} x - \frac{1}{2} m_1^T C^{-1} m_1 + \frac{1}{2} m_2^T C^{-1} m_2 = 0$$

则判别界面为：

$$\begin{aligned} d_1(x) - d_2(x) &= (m_1 - m_2)^T C^{-1} x - \frac{1}{2} m_1^T C^{-1} m_1 + \frac{1}{2} m_2^T C^{-1} m_2 \\ &= 8x_1 - 8x_2 - 8x_3 + 4 = 0 \end{aligned}$$

- **均值和协方差矩阵的估计量定义**

设模式的类概率密度函数为 $p(x)$，则其均值向量定义为：

$$m = E(x) = \int_x x p(x) dx$$

其中，样本 x 和均值向量 m 为 n 维向量，即 $x = (x_1, x_2, \ldots, x_n)^T$，$m = (m_1, m_2, \ldots, m_n)^T$。

若以样本的平均值作为均值向量的近似值，则均值估计量 $\hat{m}$ 为：

$$\hat{\boldsymbol{m}} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}^j$$

其中 N 为样本的数目, $\boldsymbol{x}^j$ 表示第 j 个样本。

协方差矩阵为：

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

其每个元素 c_{lk} 定义为：

$$\begin{aligned} c_{lk} &= E\{(x_l - m_l)(x_k - m_k)\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_l - m_l)(x_k - m_k)p(x_l, x_k)dx_l dx_k \end{aligned}$$

其中，x_l、x_k 和 m_l、m_k 分别为 $\boldsymbol{x}$ 和 $\boldsymbol{m}$ 的第 l 和 k 个分量。

协方差矩阵写成向量形式为：

$\boldsymbol{C} = E\{(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^T\} = E\{\boldsymbol{x}\boldsymbol{x}^T\} - \boldsymbol{m}\boldsymbol{m}^T$ **（注：用乘法分配率展开** $E\{\boldsymbol{x}\boldsymbol{m}^T\} = E\{\boldsymbol{x}\}\boldsymbol{m}^T = \boldsymbol{m}\boldsymbol{m}^T, E\{\boldsymbol{m}\boldsymbol{x}^T\} = \boldsymbol{m}E\{\boldsymbol{x}^T\} = \boldsymbol{m}\boldsymbol{m}^T$ **）**

协方差矩阵的估计量（当 N>>1 时）为：

$$\hat{\boldsymbol{C}} \approx \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}^j - \hat{\boldsymbol{m}})(\boldsymbol{x}^j - \hat{\boldsymbol{m}})^T$$

这里，样本模式总体为$\{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^k, \ldots, \boldsymbol{x}^N\}$。因为计算估计量时没有真实的均值向量 $\boldsymbol{m}$ 可用，只能用均值向量的估计量 $\hat{\boldsymbol{m}}$ 来代替，会存在偏差。

- **均值和协方差矩阵估计量的迭代运算形式**

假设已经计算了 N 个样本的均值估计量，若再加上一个样本，其新的估计量 $\hat{\boldsymbol{m}}(N+1)$ 为：

$$\hat{\boldsymbol{m}}(N+1)=\frac{1}{N+1}\sum_{j=1}^{N+1}\boldsymbol{x}^{j}=\frac{1}{N+1}\left[\sum_{j=1}^{N}\boldsymbol{x}^{j}+\boldsymbol{x}^{N+1}\right]=\frac{1}{N+1}[N\hat{\boldsymbol{m}}(N)+\boldsymbol{x}^{N+1}]$$

其中 $\hat{\boldsymbol{m}}(N)$ 为从 N 个样本计算得到的估计量。迭代的第一步应取 $\hat{\boldsymbol{m}}(1)=\boldsymbol{x}^{1}$。

协方差矩阵估计量的迭代运算与上述相似。取 $\hat{\boldsymbol{C}}(N)$ 表示 N 个样本时的估计量为：

$$\hat{\boldsymbol{C}}(N)=\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}^{j}(\boldsymbol{x}^{j})^{T}-\hat{\boldsymbol{m}}(N)\hat{\boldsymbol{m}}^{T}(N)$$

加入一个样本，则：

$$\begin{aligned}\hat{\boldsymbol{C}}(N+1)&=\frac{1}{N+1}\sum_{j=1}^{N+1}\boldsymbol{x}^{j}(\boldsymbol{x}^{j})^{T}-\hat{\boldsymbol{m}}(N+1)\hat{\boldsymbol{m}}^{T}(N+1)\\&=\frac{1}{N+1}\left[\sum_{j=1}^{N}\boldsymbol{x}^{j}(\boldsymbol{x}^{j})^{T}+\boldsymbol{x}^{N+1}(\boldsymbol{x}^{N+1})^{T}\right]-\hat{\boldsymbol{m}}(N+1)\hat{\boldsymbol{m}}^{T}(N+1)\\&=\frac{1}{N+1}[N\hat{\boldsymbol{C}}(N)+N\hat{\boldsymbol{m}}(N)\hat{\boldsymbol{m}}^{T}(N)+\boldsymbol{x}^{N+1}(\boldsymbol{x}^{N+1})^{T}]-\\&\quad\frac{1}{(N+1)^{2}}[N\hat{\boldsymbol{m}}(N)+\boldsymbol{x}^{N+1}][N\hat{\boldsymbol{m}}(N)+\boldsymbol{x}^{N+1}]^{T}\end{aligned}$$

（将 $\hat{\boldsymbol{m}}(N+1)=\frac{1}{N+1}(N\hat{\boldsymbol{m}}(N)+\boldsymbol{x}^{N+1}),\sum_{j=1}^{N}\boldsymbol{x}^{j}(\boldsymbol{x}^{j})^{T}=N\hat{\boldsymbol{C}}(N)+N\hat{\boldsymbol{m}}(N)\hat{\boldsymbol{m}}^{T}(N)$ 代入上式第二步，可得最后的式子）

其中，$\hat{\boldsymbol{C}}(1)=\boldsymbol{x}^{1}(\boldsymbol{x}^{1})^{T}-\hat{\boldsymbol{m}}(1)\hat{\boldsymbol{m}}^{T}(1)$ 且 $\hat{\boldsymbol{m}}(1)=\boldsymbol{x}^{1}$，因此 $\hat{\boldsymbol{C}}(1)=\boldsymbol{0}$ 为零矩阵。

● 一般概念

设$\{\boldsymbol{x}^1, \boldsymbol{x}^2,\ldots, \boldsymbol{x}^N\}$为 N 个用于估计一未知参数 $\boldsymbol{\theta}$ 的密度函数的样本，$\boldsymbol{x}^i$ 被一个接着一个逐次地给出。于是用贝叶斯定理，可以得到在给定了 $\boldsymbol{x}^1, \boldsymbol{x}^2,\ldots, \boldsymbol{x}^N$ 之后，$\boldsymbol{\theta}$ 的后验概率密度的迭代表示式为：

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^N)=\frac{p(\boldsymbol{x}^N\,|\,\boldsymbol{\theta},\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})\,p(\theta\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})}{p(\boldsymbol{x}^N\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})}$$

#注#

$$\begin{aligned}p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^N)&=\frac{p(\boldsymbol{x}^N\,|\,\boldsymbol{\theta},\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})\,p(\boldsymbol{\theta},\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})}{p(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^N)}\\&=\frac{p(\boldsymbol{x}^N\,|\,\boldsymbol{\theta},\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})\,p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})\,p(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})}{p(\boldsymbol{x}^N\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})\,p(\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})}\\&=\frac{p(\boldsymbol{x}^N\,|\,\boldsymbol{\theta},\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})\,p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})}{p(\boldsymbol{x}^N\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})}\end{aligned}$$

其中，对于 $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^N)$ 而言，$p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})$ 是它的先验概率，当加入新的样本 $\boldsymbol{x}_N$ 后，得到经过修正的新的概率密度 $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^N)$。如此一步步向前推，则 $p(\boldsymbol{\theta})$ 应为最初始的先验概率密度，当读入第一个样本 $\boldsymbol{x}^1$ 时，经过贝叶斯定理计算，可得到后验概率密度 $p(\theta\,|\,\boldsymbol{x}^1)$。以此为新的一步，将 $p(\theta\,|\,\boldsymbol{x}^1)$作为第二步计算的先验概率密度，读入样本 $\boldsymbol{x}_2$，又得到第二步的后验概率密度 $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\boldsymbol{x}^2)$，依此可以算出最后的后验概率密度 $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^N)$，从而得到最终的结果。

这里，需要先知道最初始的概率密度函数 $p(\boldsymbol{\theta})$。至于全概率 $p(\boldsymbol{x}^N\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})$ 则可通过下式算出：

$$p(\boldsymbol{x}^N\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})=\int_{\boldsymbol{x}}p(\boldsymbol{x}^N\,|\,\boldsymbol{\theta},\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})p(\boldsymbol{\theta}\,|\,\boldsymbol{x}^1,\cdots,\boldsymbol{x}^{N-1})d\boldsymbol{\theta}$$

该值与未知量 $\boldsymbol{\theta}$ 无关，可认为是一定值。

● 单变量正态密度函数的均值学习

设一个模式样本集，其类概率密度函数是单变量正态分布 $N(\theta, \sigma^2)$，均值 θ 待求，即

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

给出 N 个训练样本 $\{x^1, x^2,\ldots, x^N\}$，用贝叶斯学习计算其均值估计量。

设最初的先验概率密度 $p(\theta)$ 为 $N(\theta_0,\sigma_0^2)$，这里 θ_0 是凭先验知识对未知量 θ 的“最好”推测，σ_0^2 表示上述推测的不确定性度量。这里可以假定 $p(\theta)$ 是正态的，因为均值的估计量是样本的线性函数，因样本 x 是正态分布的，因此 $p(\theta)$ 取为正态分布是合理的，这样计算起来可比较简单。

初始条件已知，即 $p(\theta)$ 为 $N(\theta_0,\sigma_0^2)$，$p(x^1|\theta)$ 为 $N(\theta, \sigma^2)$，由贝叶斯公式 $p(\theta|x^1)=a\,p(x^1|\theta)\,p(\theta)$，可得：

$$p(\theta|x^1)=a\cdot\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x^1-\theta}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right]$$

其中 a 是一定值。由贝叶斯法则有：

$$p(\theta|x^1,\cdots,x^N)=\frac{p(x^1,\cdots,x^N|\theta)p(\theta)}{\int_\varphi p(x^1,\cdots,x^N|\theta)p(\theta)d\theta}$$

这里 Φ 表示整个模式空间。由于每一次迭代是从样本子集中逐个抽取一个变量，所以 N 次运算是独立地抽取 N 个变量，因此上式可写成：

$$p(\theta|x^1,\cdots,x^N)=a\left\{\prod_{k=1}^{N}p(x^k|\theta)\right\}p(\theta)$$

代入 $p(x^k|\theta)$ 和 $p(\theta)$ 的值，得：

$$p(\theta\mid x^1,\cdots,x^N)=a\left\{\prod_{k=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x^k-\theta}{\sigma}\right)^2\right]\right\}\cdot\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right]$$

$$=a'\exp\left[-\frac{1}{2}\left\{\sum_{k=1}^{N}\left(\frac{x^k-\theta}{\sigma}\right)^2\right\}+\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right]$$

$$=a''\exp\left[-\frac{1}{2}\left\{\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\theta^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N}x^k+\frac{\theta_0}{\sigma_0^2}\right)\theta\right\}\right]$$

上式每一步中与 θ 无关的项都并入常数项 a' 和 a''，这样 p($\theta\mid x^l$, …, x^N)是 θ 平方函数的指数集合，仍是一正态密度函数。将它写成 $N(\theta_N,\sigma_N^2)$ 的形式，即：

$$p(\theta\mid x_1,\cdots,x_N)=\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\theta-\theta_N}{\sigma_N}\right)^2\right]$$

$$=a'''\exp\left[-\frac{1}{2}\left(\frac{\theta^2}{\sigma_N^2}-2\frac{\theta_N\theta}{\sigma_N^2}\right)\right]$$

将上述两式相比较，得：

$$\frac{1}{\sigma_N^2}=\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}$$

$$\frac{\theta_N}{\sigma_N^2}=\frac{1}{\sigma^2}\sum_{k=1}^{N}x^k+\frac{\theta_0}{\sigma_0^2}=\frac{N}{\sigma^2}\hat{m}_N+\frac{\theta_0}{\sigma_0^2}$$

解出 θ_N 和 σ_N，得：

$$\theta_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\hat{m}_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\theta_0$$

$$\sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

即根据对训练样本集 $\{x^i\}_{i=1,2,\ldots,N}$ 的观察，求得均值 θ 的后验概率密度 p($\theta\mid x^i$)为 $N(\theta_N,\sigma_N^2)$，其中 θ_N 是经过 N 个样本观察之后对均值的最好估计，它是先验信息（即 θ_0，σ_0^2 和 σ^2）与训练样本所给信息（即 N

和$\hat{m}_N$）适当结合的结果，是用N个训练样本对均值的先验估计 θ_0 的补充；σ_N^2是对这个估计的不确定性的度量，因σ_N^2随N的增加而减小，因此当$N\to\infty$时，σ_N^2趋于零。由于 θ_N是$\hat{m}_N$和θ_0的线性组合，两者的系数都非负且其和为1，因此只要$\sigma_0\neq 0$，当$N\to\infty$时，θ_N趋于样本均值的估计量$\hat{m}_N$。

图中所示为一正态密度的均值学习过程，每增加一次对样本的预测，都可减小对θ估计的不确定性，所以$p(\theta|x^1,...,x^N)$变得越来越峰形突起，且其均值与估计量$\hat{m}_N$之间的偏差的绝对值亦越来越小。