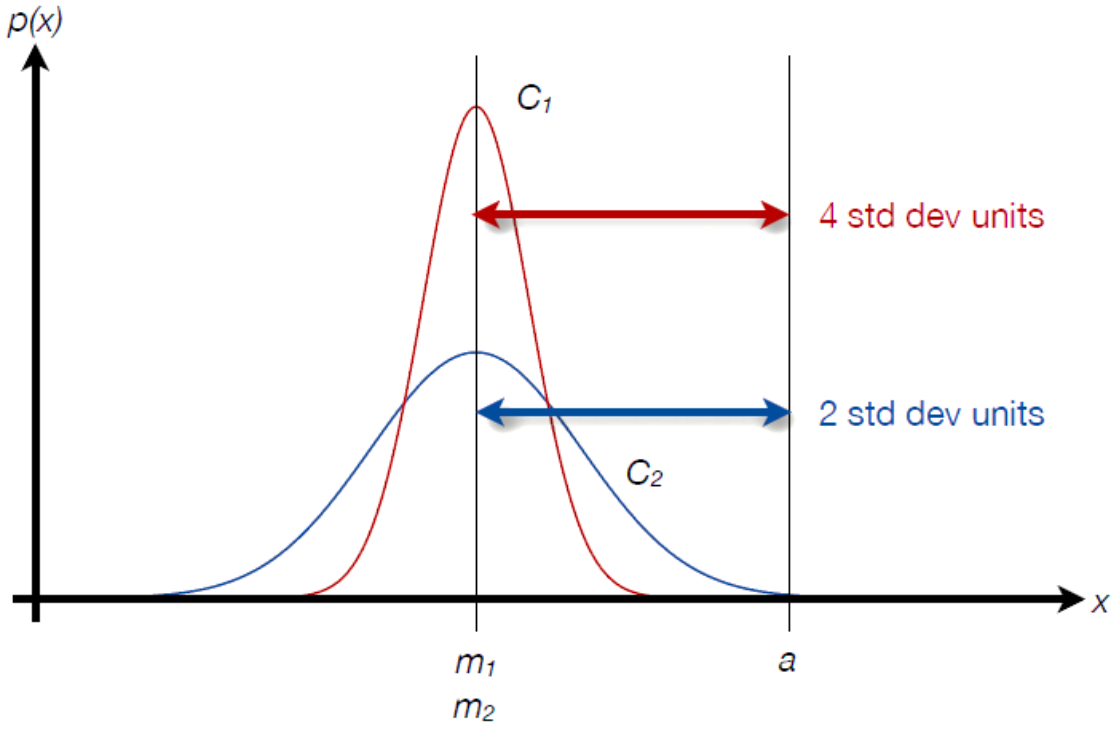
机器学习

于元隆、朱丹红

福州大学计算机与大数据学院
Email: yu.yuanlong@fzu.edu.cn

MICD分类器的问题

- MICD分类器会选择方差较大的类。



基于距离的决策：

- 仅考虑每个类别各自观测到的训练样本的分布情况。例如，均值（MED分类器）和协方差（MICD分类器）。
- 没有考虑类的分布等先验知识。例如，类别之间样本数量的比例，类别之间的相互关系。

概率：通常用来表达事物处于每种取值状态的可能性。

- 推理可分为确定性（Certainty）推理和概率推理。所谓确定性推理是指类似如下的推理过程：

如条件B存在，就一定会有结果A。现在已知条件B存在，可以得出结论是结果A一定也存在。

“如果考试作弊，该科成绩就一定是0分。”这就是一条确定性推理。

- 而概率推理（Probabilistic Reasoning）是不确定性推理，它的推理形式可以表示为：

如条件B存在，则结果A发生的概率为P(A|B)。P(A|B)也称为结果A发生的条件概率（Conditional Probability）。

“如果考前未复习，该科成绩有50%的可能性不及格。”这就是一条概率推理。

- 通常情况下，条件概率从大量实践中得来，它是一种经验数据的总结，但对于我们判别事物和预测未来没有太大的直接作用。

- 我们更关注的是如果我们发现了某个结果（或者某种现象），那么造成这种结果的原因有多大可能存在呢？这就是**逆概率推理**的含义。

- 即：已知条件B存在，则结果A存在的概率为P(A|B)。如果发现结果A出现了，求条件B存在的概率P(B|A)？

  例如：已知某位人物如果是罪犯，他留下某些线索的概率；那么如果发现了一些线索，他是罪犯的概率是多少？

  再如：已知患有某种疾病，会出现某种症状的概率；那么如果医生发现某位患者出现了某种症状，他患有该种疾病的概率是多少？

解决逆概率推理问题：贝叶斯公式

贝叶斯公式是由托马斯·贝叶斯于1763年提出的，它的数学表述为：

设试验 E 的样本空间为 S，A 为 E 的事件，B_1，B_2，…，B_c 为 S 的一个划分，且 $P(A)>0$，$P(B_i)>0\ (i=1,2,\cdots\cdots,c)$，则

$$P(B_i|A)\models\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{c}P(A|B_j)P(B_j)}=\frac{P(A|B_i)P(B_i)}{P(A)}$$

$$P(Bi|A)$$

称为**后验概率（Posterior Probability）**，表示事件A（结果A）出现后，各不相容的条件Bi存在的概率，它是在结果出现后才能计算得到的，因此称为“后验”。

$$P(A|B_j)$$

称为**类条件概率（Class-conditional Probability）**，表示在各条件Bi存在时，结果事件A发生的概率。也称**观测似然**。

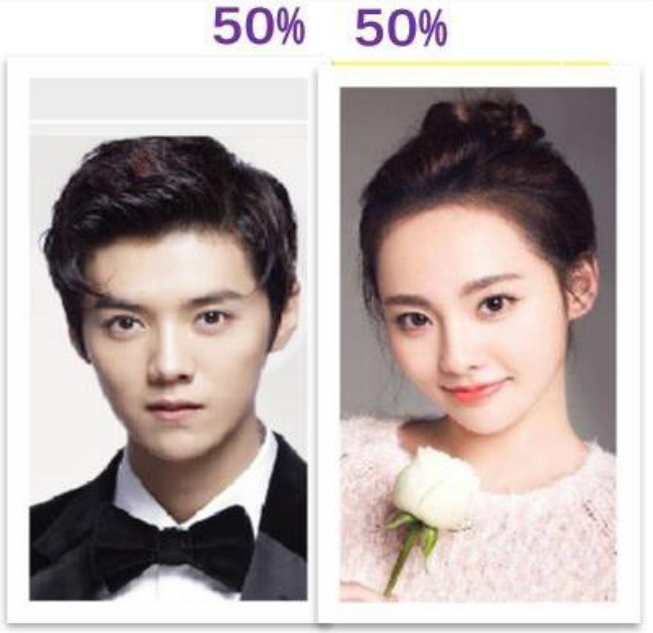
$$P(B_j)$$

称为**先验概率（Priori Probability）**，表示各不相容的条件Bi出现的概率，它与结果A是否出现无关，仅表示根据先验知识或主观推断，认为总体上各条件出现的可能性有什么差别。

$$P(A)=\Sigma P(A|B_j)P(B_j)$$

由先验概率和类条件概率计算得到，它表达了结果A在各种条件下出现的总体概率，称为结果A的**全概率（Total Probability）**。

- 因此，当我们把**贝叶斯公式**应用于不确定统计分类时，我们就得到了根据样本的**特征值**来进行类别划分的一种不确定分类器。
- 它可以计算出该样本属于每一个类别的概率是多少。



需要计算：后验概率P(男性|长发)和P(女性|长发)。

假设**类条件概率**（**观测似然**）
即男性留长发的概率P(长发|男性)=5%，
女性留长发的概率P(长发|女性)= 70%。

- 如果我们在一个特殊的环境中，男性女性的先验概率并不相同，例如在福州大学，男性大约占总人群的75%，女性大约占25%。
- 此时我们如果遇到一个背影是长发的人，他是男性或女性的后验概率为：

P(长发|男性) =0.05，P(长发|女性) =0.7，设 P(男性) =75%，P(女性) =25%，则：

P(男性|长发)=(0.05*0.75)/(0.05*0.75+0.7*0.25)=3/17=0.176。

- 可以看到：在各类别先验概率不均等时，后验概率也会发生很大的变化。

- 贝叶斯分类的特点：
  - 1、**先验概率必须是已知的。**在没有获得任何信息的时候，如果要进行分类判别，只能依据各类存在的先验概率，将样本划分到先验概率大的一类中，风险会比较小。
  - 2、**以新获得的信息对先验概率进行修正。**获得了更多关于样本特征的信息后，即得到观测概率后，可以依照贝叶斯公式对先验概率进行修正，得到后验概率。
  - 3、**分类决策存在错误率。**由于贝叶斯分类是在样本取得某特征值时对它属于各类的概率进行推测，并无法获知样本真实的类别归属情况，所以分类决策一定存在错误率，即使错误率很低。

- 某地发生了一起交通事故肇事逃逸事件，现场有一位目击者，他非常肯定地说，他看见肇事车的车标是右侧的车标，而不是左侧的车标。如果这个目击者的可信度达到99%，就是说只有1%的可能性他会在两个车标中认错。

- 请问：肇事车的车标是右侧车标的可能性有多大?

- 根据我们的直觉，我们会相信这个可信度非常高的目击者，会认为接受他的证词错误率会比较小。

  P(认成右标|实为左标) =0.01，

  P(认成右标|实为右标) =0.99，

  设P(左标) =P(右标) =50%，

  则P(实为右标|认成右标) =(0.99*0.5)/(0.99*0.5+0.01*0.5)=0.99

- 结果显示，我们的直觉是正确的，肇事车的车标确实是右侧车标的概率达到了99%。
- 我们接纳目击者的证词，认定肇事车是右侧的车型错误率很低。

市场占有率	
50%	50%

市场占有率	
50%	50%
99%	1%

P(实为右标|认成右标)=?

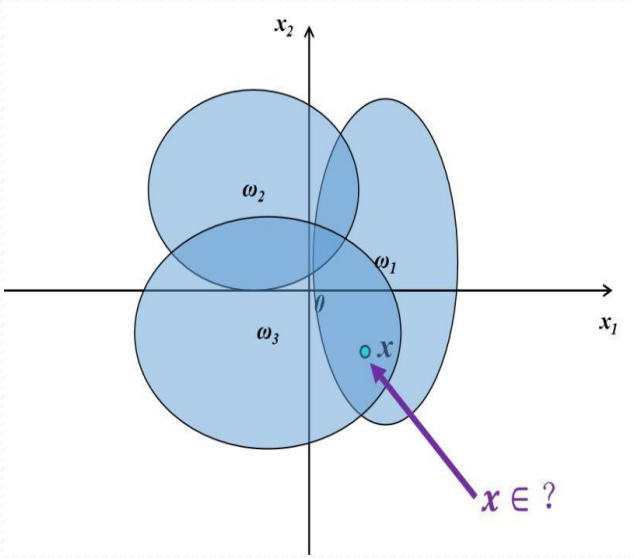
- 对于不确定的统计分类，已知的是每个类别的样本取得不同特征向量的概率（也就是该类样本的统计分布），现在需要实现的是如何依据某个待识别样本的特征向量，计算出该样本属于每一个类的概率？

1、先验概率：每一类的样本的整体出现概率。

2、类条件概率：把每个类中样本取得某个具体特征向量值的概率。

3、后验概率：把待计算的样本取得某一个具体特征向量值时属于每一类的概率。

使用贝叶斯公式可以解决分类问题。



例：2019年以来，人感染某类流感的病例开始出现，并造成了一定的社会恐慌。根据目前数据统计，该病的总体发病率大约为**1000万分之一**，对照普通流感的发病率可高达**30%**。研究发现，**易感人群中99%**的人感染某类流感病例曾出现过发热、咳嗽等急性呼吸道感染症状，而同样的易感人群中**80%**的普通流感患者也出现过同样症状。

（1）现有一位患者属于易感人群，并出现了发热、咳嗽等急性呼吸道感染症状，请问是否应当按照人某类流感疑似病例对待？

计算该患者实为某类流感感染的后验概率：

P(H7N9) =0.0000001/(0.0000001+0.3)

P(流感) =0.3/(0.0000001+0.3)

P(症状|某类流感) =0.99　　P(症状|流感) =0.8

P(某类流感|症状)=

(0.0000001/(0.0000001+0.3)*0.99)/(0.0000001/(0.0000001+0.3)*0.99+0.3/(0.0000001+0.3)*0.8)=0.000000412

- 这个后验概率是非常低的，应当还是将他按普通流感对待。

- 但是，某类流感致死率高，又有强烈的传染性，万一我们误诊了，后果会比较严重。当然，如果我们是把普通流感患者误诊为某类流感患者，他可能就是受到些惊吓，或者被隔离后生活不便，没有太大的社会风险。

- 如果我们是把一名某类流感患者误诊为普通流感患者，而没有采取合理有效的措施对他进行隔离和治疗的话，可能就会给患者本人和整个社会带来**巨大的损失**。

- 这个例子提示我们，当我们使用贝叶斯分类器时，仅仅考虑识别错误率低是不够的，还**应当把我们所采取的分类决策所带来的后果考虑进去**，这就是“最小风险贝叶斯分类器”的由来。

机器学习

于元隆、朱丹红

福州大学计算机与大数据学院
Email: yu.yuanlong@fzu.edu.cn

贝叶斯分类器

概率的观点

- 随机性：每个样本是一次随机采样，样本个体具有随机性。
- 例如，鲈鱼和三文鱼的分类：即使是属于同一类的鱼，长度和亮度等特征在不同的个体样本上也是有变化的。
- **机器学习**所要做的是：反复观测采样，找出数据蕴含的概率分布规律。
- 推理决策：根据学习出来的概率分布规律来做决定。

- 概率：通常用来表达事物处于每种取值状态的可能性。

  每维**特征**构成一个随机变量，其概率分布由两个元素组成：

  - 1）该特征的取值空间（连续或者离散）。
  - 2）在该特征维度上，样本处于各个取值状态的可能性。

- 从概率的观点看，给定一个测试模式x，决策其属于哪个类别需要依赖于如下条件概率：$P(C|x)$

- 输入模式$\boldsymbol{x}$：随机变量（单维特征）或向量（高维特征）。
- 类别输出$\boldsymbol{C}$：随机变量，取值是所有类别标签$\{Ci\}$。
- 针对每个类别Ci，该条件概率可以写作：$P(Ci|\boldsymbol{x})$
- 该条件概率也称作后验概率(posterior probability)，表达给定模式$\boldsymbol{x}$属于类Ci可能性。

- 理想情况下，对于给定的一个测试样本，我们希望选择它属于拥有最高概率的那个类。
- 这里的概率是指经过观测以后，测试样本x属于$C1$ 或者$C2$ 类的后验概率。

$$P(C_i|\underline{x}) \underset{C_j}{\overset{C_i}{\gtrless}} P(C_j|\underline{x})$$

例如：一个2类问题，C_1诊断为患有某病，C_2诊断为无病，

则：$P(C_1)$表示某地区的人患有此病的概率。

$P(C_2)$表示该地区人无此病的概率。

通过统计资料得到

若用某种方法检测是否患有某病，假设 X 表示“试验反应呈阳性”。则：

值低 / 高 ？
√

$P(X|C_2)$ 表示无病的人群做该试验时反应呈阳性(显示有病)的概率。

$P(C_2|X)$ 表示试验呈阳性的人中，实际没有病的人的概率。

值低 / 高 ？
√

$P(X|C_1)$ 表示患病人群做该试验时反应呈阳性的概率。

$P(C_1|X)$ 表示试验呈阳性的人中，实际确实有病的人的概率。

后验概率计算

- 如何得到后验概率：使用贝叶斯规则
  - 先验概率、观测似然、后验概率

$$P(C_i|\underline{x}) = \frac{p(\underline{x}|C_i)P(C_i)}{p(\underline{x})}$$

Where $p(\underline{x}|C_i)$ is the class conditional probability density (p.d.f.), which needs to be estimated from the available samples or otherwise assumed.

$P(C_i)$ is the 'a priori' (before measurement) probability of class C_i.

$$p(\underline{x}) = \sum_j p(\underline{x}|C_j)P(C_j)$$

例 假定在细胞识别中，病变细胞的先验概率和正常细胞的先验概率分别为：

$$P(\omega_1)=0.05,\ P(\omega_2)=0.95$$

现有一待识别细胞，其观察值为X，观测似然取值：

$$p(\boldsymbol{X}\mid\omega_1)=0.5 \qquad p(\boldsymbol{X}\mid\omega_2)=0.2$$

试对细胞X进行分类。

解：[方法1] 通过后验概率计算。

$$P(\omega_1 \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \omega_1)P(\omega_1)}{\sum_{i=1}^{2} p(\boldsymbol{X} \mid \omega_i)P(\omega_i)}$$

$$= \frac{0.5 \times 0.05}{0.05 \times 0.5 + 0.95 \times 0.2} \approx 0.16$$

$$P(\omega_2 \mid \boldsymbol{X}) = \frac{0.2 \times 0.95}{0.05 \times 0.5 + 0.95 \times 0.2} \approx 0.884$$

$$\because P(\omega_2 \mid \boldsymbol{X}) > P(\omega_1 \mid \boldsymbol{X}) \qquad \therefore \boldsymbol{X} \in \omega_2$$

[方法2]：利用先验概率和类概率密度计算。

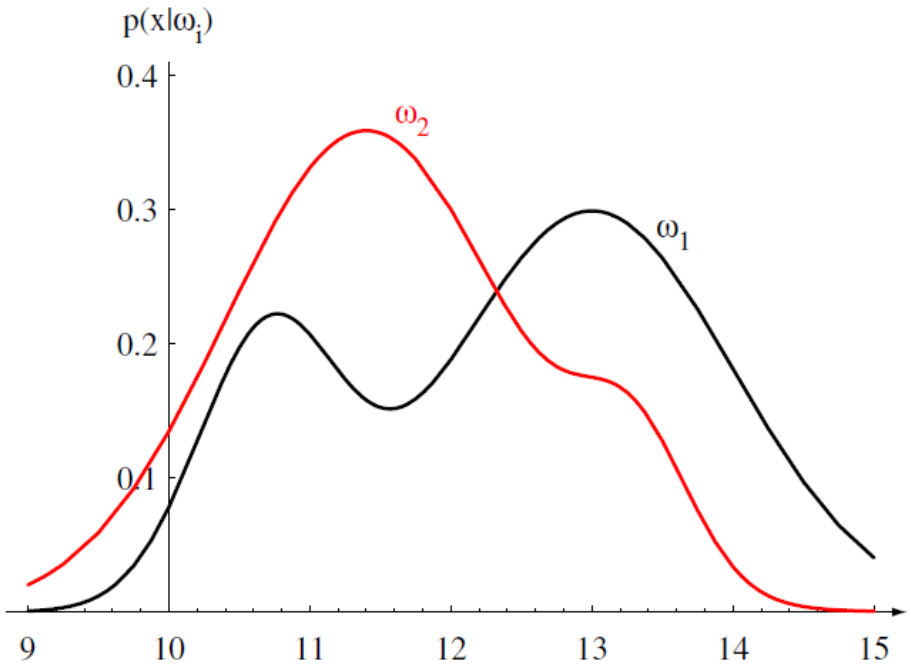
$$p(\boldsymbol{X}\mid\omega_1)P(\omega_1)=0.5\times0.05=0.025$$

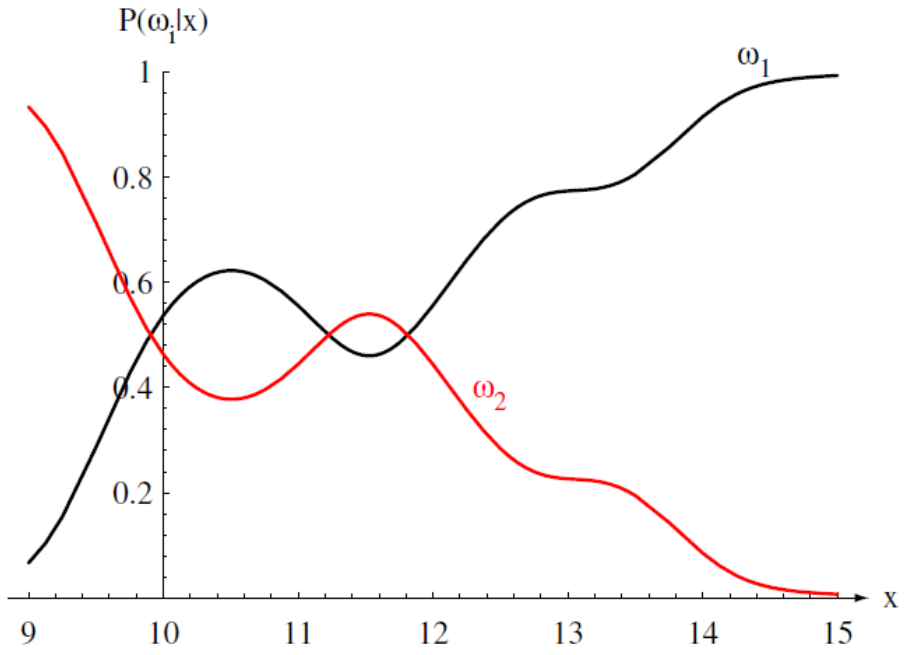
$$p(\boldsymbol{X}\mid\omega_2)P(\omega_2)=0.2\times0.95=0.19$$

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

$$\therefore \boldsymbol{X}\in\omega_2$$，是正常细胞。

- 如何得到后验概率：使用贝叶斯规则
  - 先验概率：P(C1)=2/3,P(C2)=1/3
  - 观测似然、后验概率



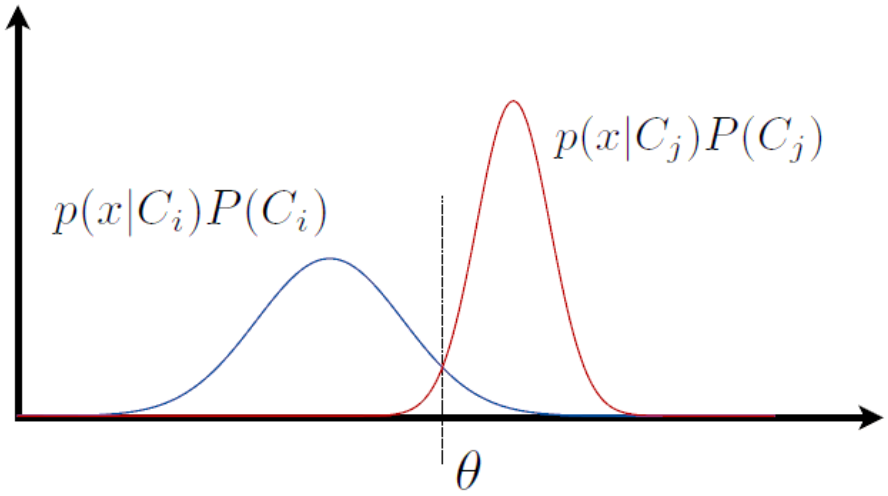


最大后验概率（MAP）分类器

- Maximum a posterior (MAP)分类器：
- 将测试样本分给后验概率最大的那个类。

$$P(C_i|\underline{x}) \underset{C_j}{\overset{C_i}{\gtrless}} P(C_j|\underline{x})$$

$$p(\underline{x}|C_i)P(C_i) \underset{C_j}{\overset{C_i}{\gtrless}} p(\underline{x}|C_j)P(C_j)$$

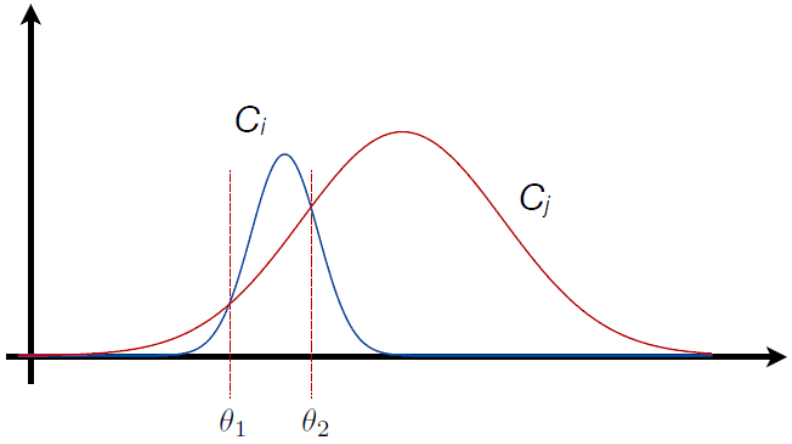


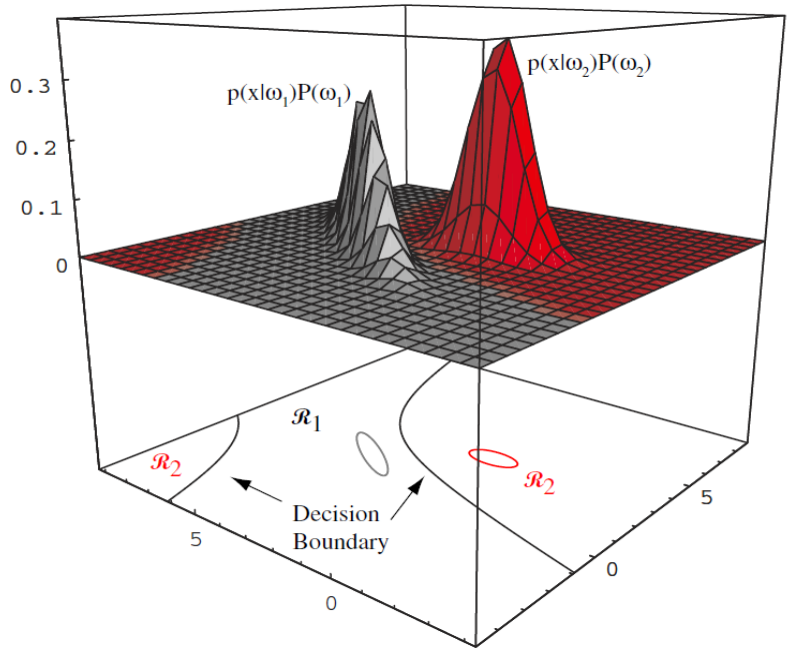
MAP分类器的决策边界

- MAP分类器的决策边界方程：

$$p(\underline{x}|C_i)P(C_i) = p(\underline{x}|C_j)P(C_j)$$

  - 在一维空间，通常有两条分类边界。
  - 在n维空间，分类边界非常复杂。
  - 实际的分类边界取决于先验概率和观测似然所用的形式。

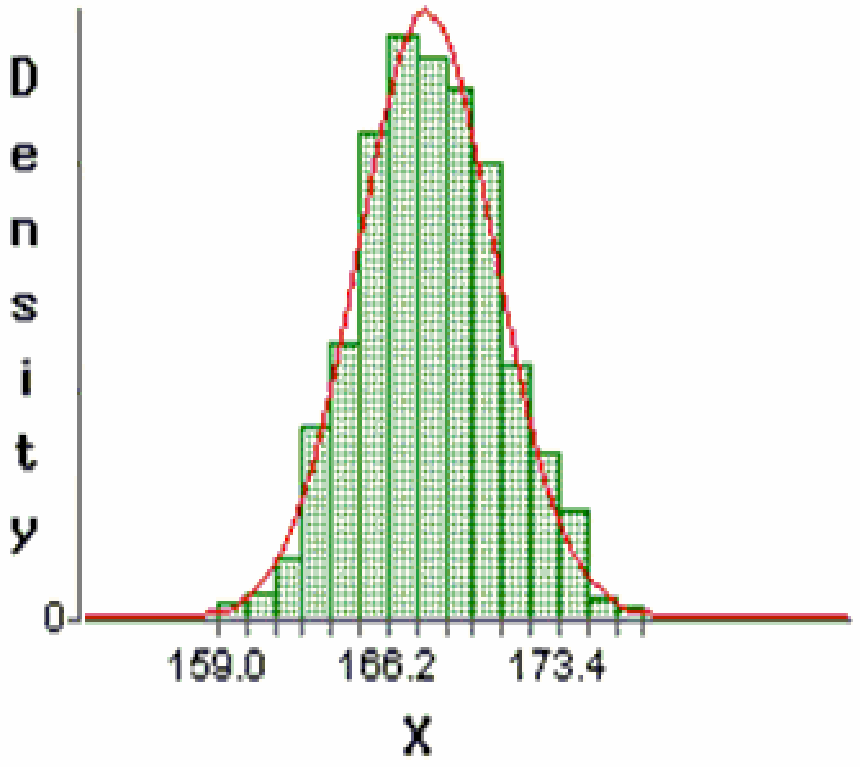




观测似然：单变量高斯分布

许多实际的数据集：
均值附近分布较多样本；
距均值点越远，样本分布越少。
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。

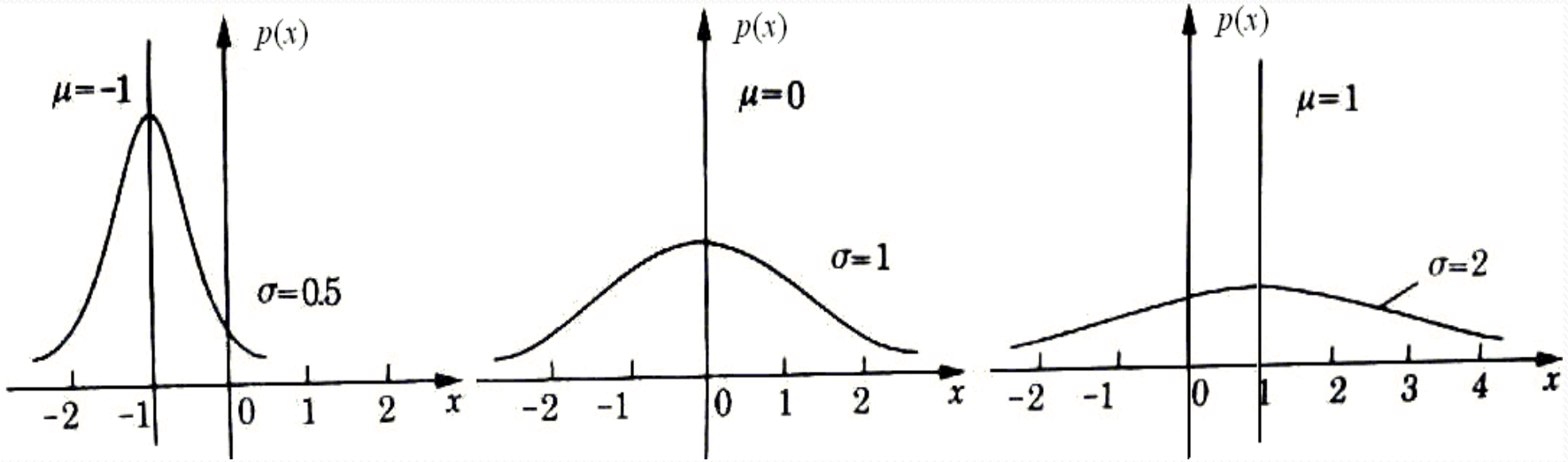
单变量（一维）的正态分布

概率密度函数定义为：

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

曲线如图示：

①μ= -1，σ=0.5；　②μ= 0，σ=1；　③μ= 1，σ=2 .



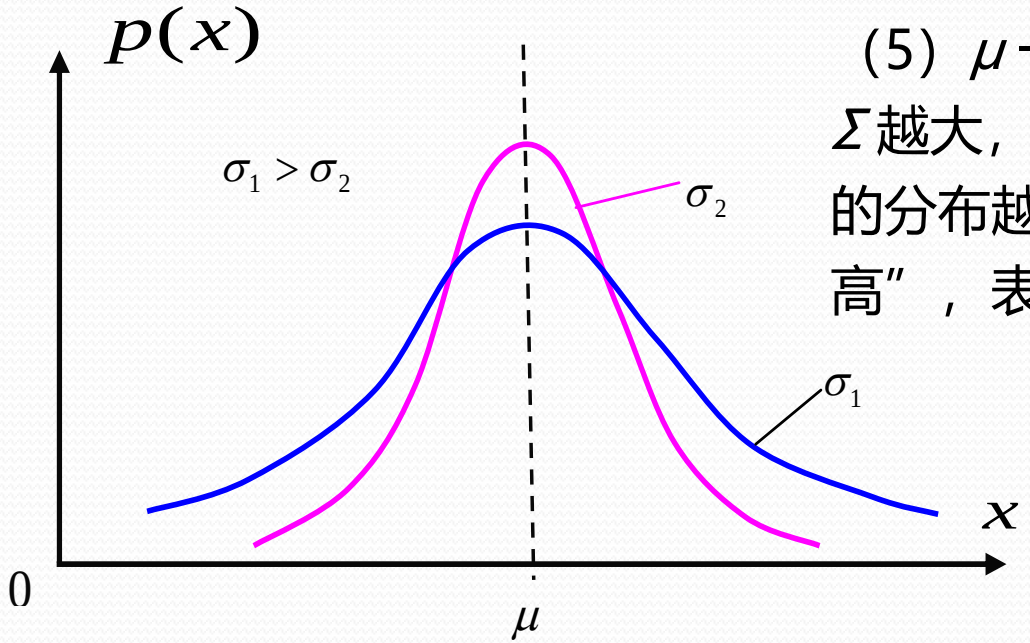
一维正态曲线的**性质**：

（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。

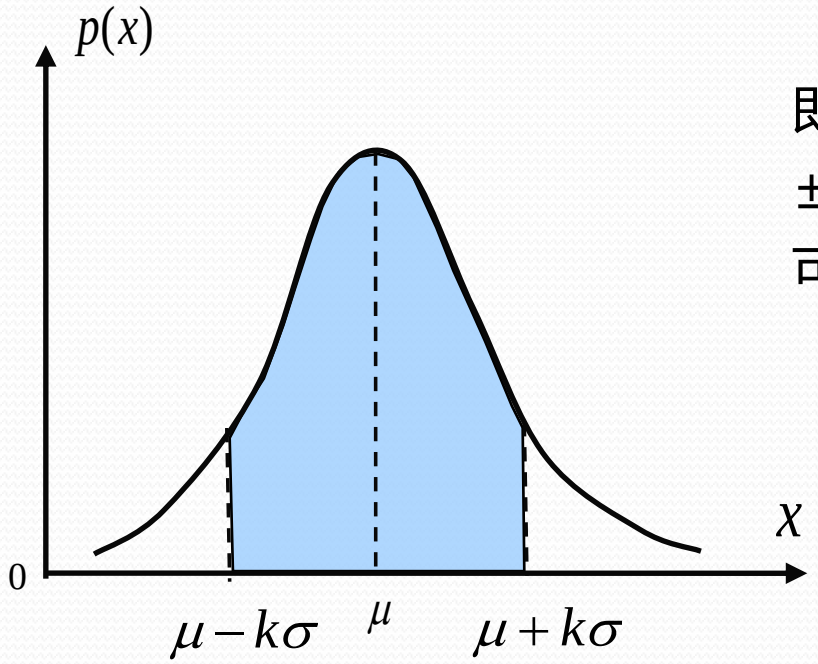
（4）当$x<\mu$时，曲线上升；当$x>\mu$时，曲线下降.并且当曲线向左、右两边无限延伸时，以x轴为渐近线，向它无限靠近。



（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。

3σ规则

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



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：

$$p(x)\sim N(\mu,\sigma^2)$$

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

多变量（n维）正态随机向量

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

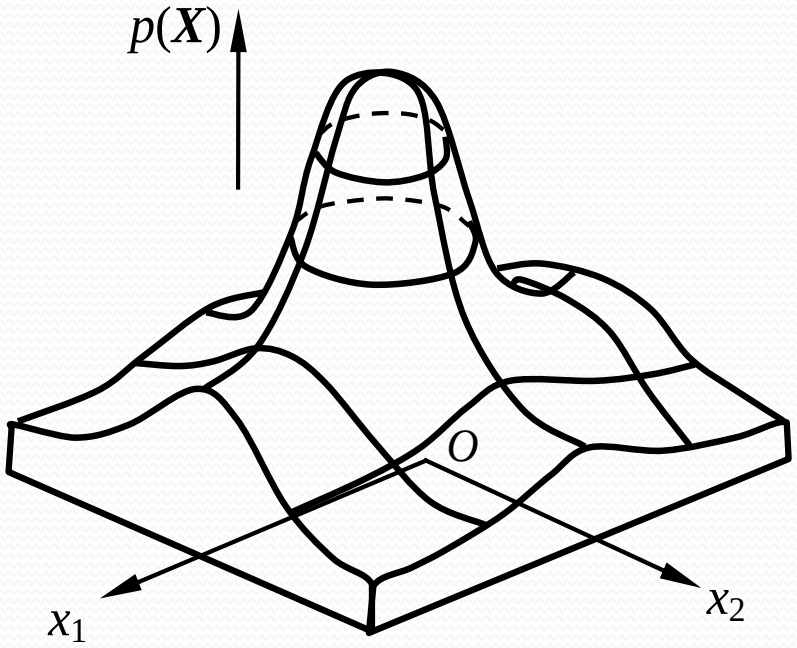
$$\boldsymbol{C}=\begin{bmatrix}\sigma_{11}^2 & \cdots & \sigma_{1n}^2\\ \vdots & & \vdots\\ \sigma_{n1}^2 & \cdots & \sigma_{nn}^2\end{bmatrix}$$

为协方差矩阵，是对称正定矩阵。

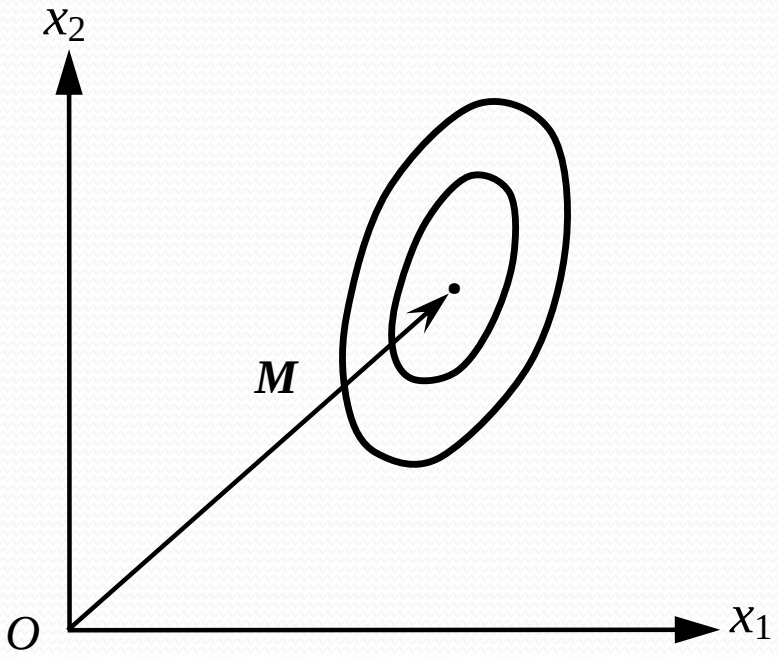
$|\boldsymbol{C}|$：协方差矩阵$\boldsymbol{C}$的行列式。

多维正态密度函数完全由它的均值 $\boldsymbol{M}$ 和协方差矩阵 $\boldsymbol{C}$ 所确定，简记为：

$$p(\boldsymbol{X})\sim N(\boldsymbol{M},\boldsymbol{C})$$



（a）



（b）

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。

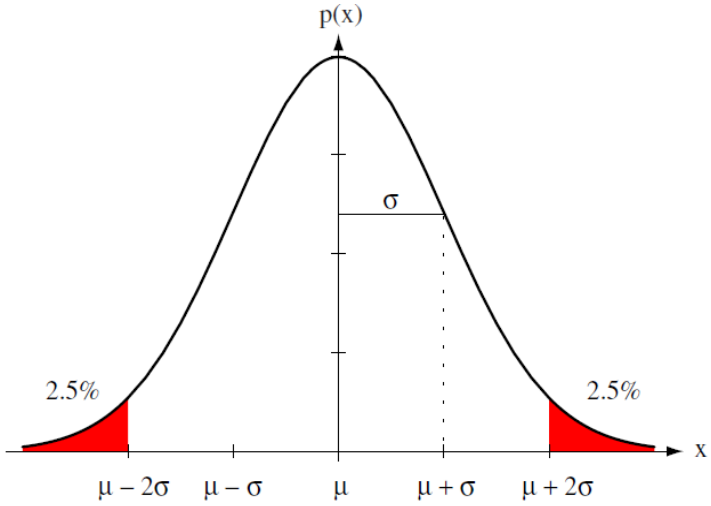
- 虽然它的MAP分类决策规则很简单，但是由于**观测似然**分布的形式不确定，所以，我们很难找到一个形式简单的分类决策边界。

- 如果各类样本特征向量取值的类条件概率(观测似然)满足正态分布，决策规则会不会更简化?

观测似然：单变量高斯分布

- 假设观测似然是单变量高斯分布：

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2} \quad i = 1, 2,$$



- 带入MAP分类器，两边都取log：

$$p(\underline{x}|C_i)P(C_i) \underset{C_j}{\overset{C_i}{\gtrless}} p(\underline{x}|C_j)P(C_j)$$

$$log\left(\frac{p(\underline{x}|C_i)}{p(\underline{x}|C_j)}\right) \underset{C_j}{\overset{C_i}{\gtrless}} log\left(\frac{P(C_j)}{P(C_i)}\right)$$

$$\left(\frac{x-\mu_j}{\sigma_j}\right)^2 - \left(\frac{x-\mu_i}{\sigma_i}\right)^2 \underset{C_j}{\overset{C_i}{\gtrless}} 2log\left(\frac{P(C_j)\sigma_i}{P(C_i)\sigma_j}\right)$$

- 练习题：

设一维两类模式满足正态分布，它们的均值和方差分别为，$m_1=0$，$s_1=2$，$m_2=2$，$s_2=2$，$p(x) \sim N(m,s)$，$P(\omega_1)= P(\omega_2)$。试算出判决边界点，并绘出它们的概率密度函数曲线（示意图）；试确定样本-3，-2，1，3，5各属哪一类。

高斯观测似然时的决策边界

- 为了得到决策边界，决策方程两边设置相等，从而解得x：

$$\left(\frac{1}{\sigma_j^2}-\frac{1}{\sigma_i^2}\right)x^2-2\left(\frac{\mu_j}{\sigma_j^2}-\frac{\mu_i}{\sigma_i^2}\right)x+\frac{\mu_j^2}{\sigma_j^2}-\frac{\mu_i^2}{\sigma_i^2}-2log\left(\frac{P(C_j)\sigma_i}{P(C_i)\sigma_j}\right)=0$$

- 如果$\boldsymbol{\sigma_i=\sigma_j}$，只有一条分类边界：

$$2\left(\frac{\mu_i-\mu_j}{\sigma^2}\right)x+\frac{\mu_j^2-\mu_i^2}{\sigma^2}-2log\left(\frac{P(C_j)}{P(C_i)}\right)=0$$

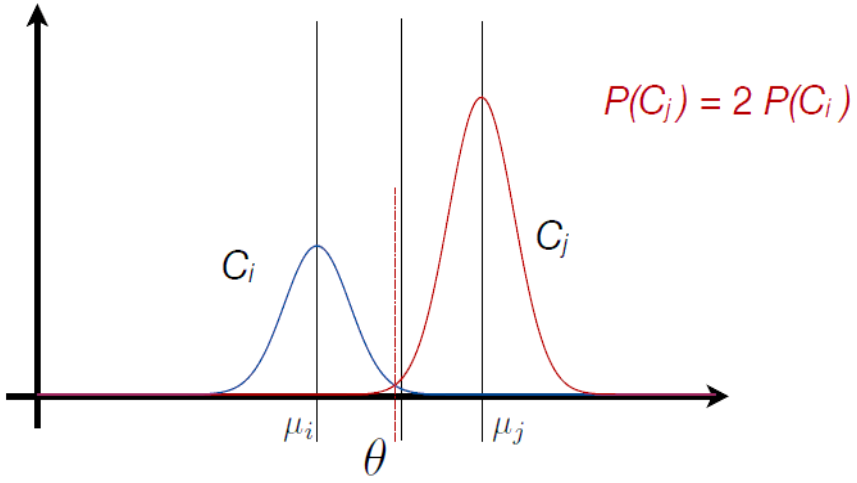
- 从而解出x，得到该条分类边界θ的方程表达：

$$\theta=\frac{\mu_i+\mu_j}{2}+\frac{\sigma^2}{\mu_i-\mu_j}log\left(\frac{P(C_j)}{P(C_i)}\right)$$

- 假设$\mu_j > \mu_i$，如果$P(C_j) > P(C_i)$，则第二项为负值，且得到分类边界：

$$\theta = \frac{\mu_i + \mu_j}{2} + \frac{\sigma^2}{\mu_i - \mu_j} log\left(\frac{P(C_j)}{P(C_i)}\right)$$

$$\theta = \frac{\mu_i + \mu_j}{2} - \delta$$



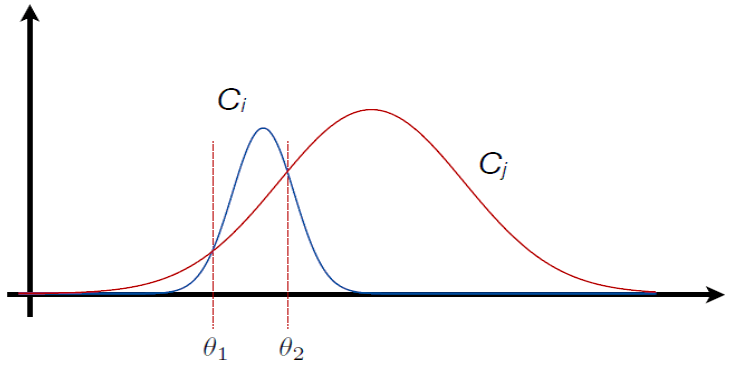
- 比较MED分类边界，在方差相同的情况下，MAP分类边界偏向于先验可能性较小的类。
- 假设 $\mu_j < \mu_i$，也可以得到相同结论。

$$\theta_{MED} = \theta_{MICD} = \frac{\mu_i + \mu_j}{2}$$

- 通常情况下， $\boldsymbol{\sigma_i \neq \sigma_j}$ ，此时分类边界有两个解：

$$\left(\frac{1}{\sigma_j^2}-\frac{1}{\sigma_i^2}\right)x^2-2\left(\frac{\mu_j}{\sigma_j^2}-\frac{\mu_i}{\sigma_i^2}\right)x+\frac{\mu_j^2}{\sigma_j^2}-\frac{\mu_i^2}{\sigma_i^2}-2log\left(\frac{P(C_j)\sigma_i}{P(C_i)\sigma_j}\right)=0$$

- θ_1和θ_2是假设 μ_i, μ_j , σ_i, σ_j, $P(C_i)$, $P(C_j)$ 的函数。



- 和MICD分类器决策边界比较：

$$\boldsymbol{x}^T(\Sigma_1^{-1}-\Sigma_2^{-1})\boldsymbol{x}+2(\boldsymbol{\mu}_2^T\Sigma_2^{-1}-\boldsymbol{\mu}_1^T\Sigma_1^{-1})\boldsymbol{x}+\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2=0$$

- 当$\sigma_i > \sigma_j$且先验概率相等时，$\delta>0$，MAP分类器倾向选择σ_j类，即方差较小（紧致）的类。

高斯观测似然时的分类器比较

- 比较MAP和MICD分类器：

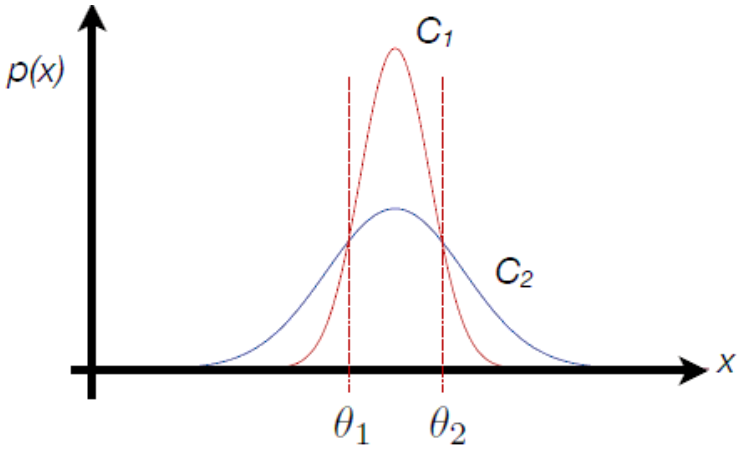
MAP:

$$\left(\frac{x-\mu_j}{\sigma_j}\right)^2 - \left(\frac{x-\mu_i}{\sigma_i}\right)^2 \underset{C_j}{\overset{C_i}{\gtrless}} 2log\left(\frac{P(C_j)\sigma_i}{P(C_i)\sigma_j}\right)$$

MICD:

$$\left(\frac{x-\mu_j}{\sigma_j}\right)^2 - \left(\frac{x-\mu_i}{\sigma_i}\right)^2 \underset{C_j}{\overset{C_i}{\gtrless}} 0$$

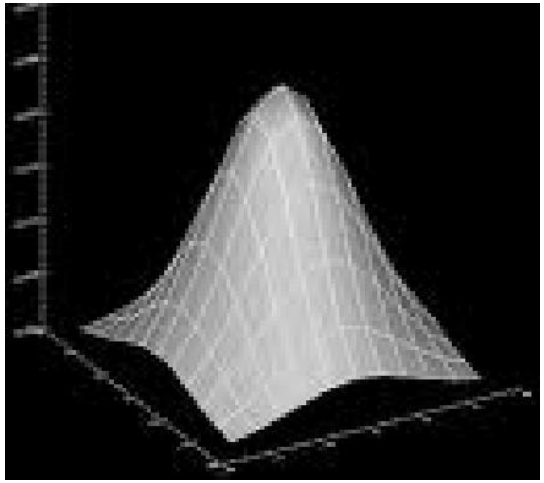
- MAP分类器决策时，偏向于先验较大可能性的类、分布较为紧致的类。

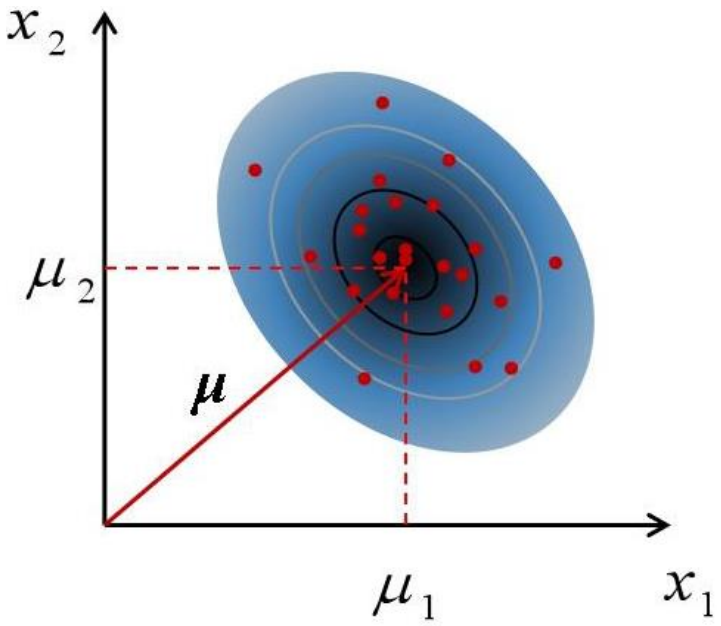


观测似然：多变量高斯分布

- 假设观测似然是多变量高斯分布，即特征是多维度。

$$p(\underline{x}|C_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} e^{\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)} \quad i = 1, 2, ..., N$$





观测似然：多变量高斯分布

- 假设观测似然是多变量高斯分布，即特征是多维度。

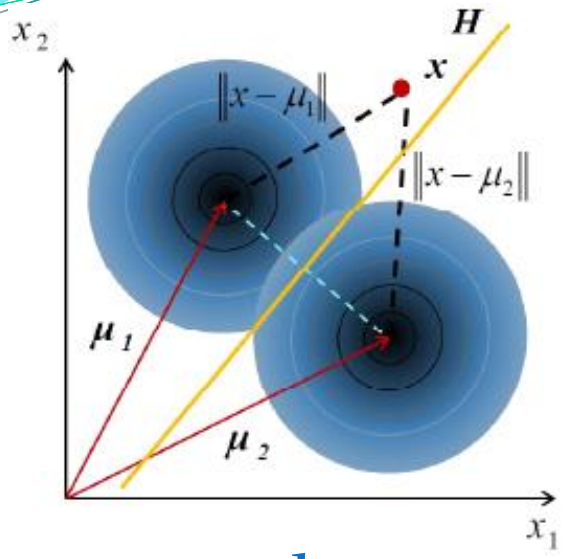
$$p(\underline{x}|C_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} e^{\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)} \quad i = 1, 2, ..., N$$

- 代入MAP分类器，两边都取log：

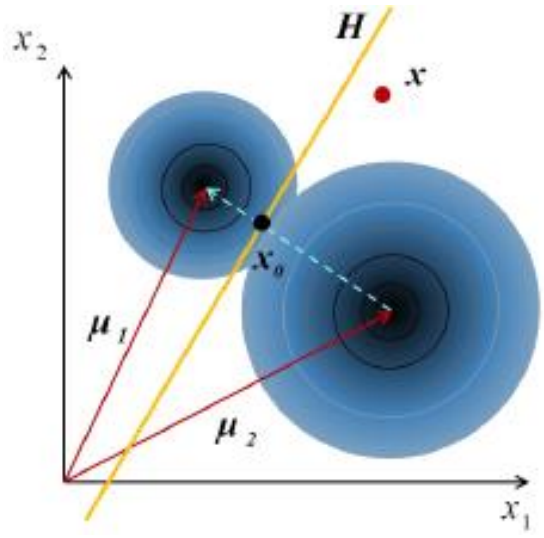
$$(\underline{x}-\underline{\mu}_j)^T\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j) - (\underline{x}-\underline{\mu}_i)^T\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) \underset{C_j}{\overset{C_i}{\gtrless}} 2log\left(\frac{P(C_j)|\Sigma_i|^{1/2}}{P(C_i)|\Sigma_j|^{1/2}}\right)$$

- 决策边界是一个超二次型，但始终是偏移MICD决策边界如下距离：

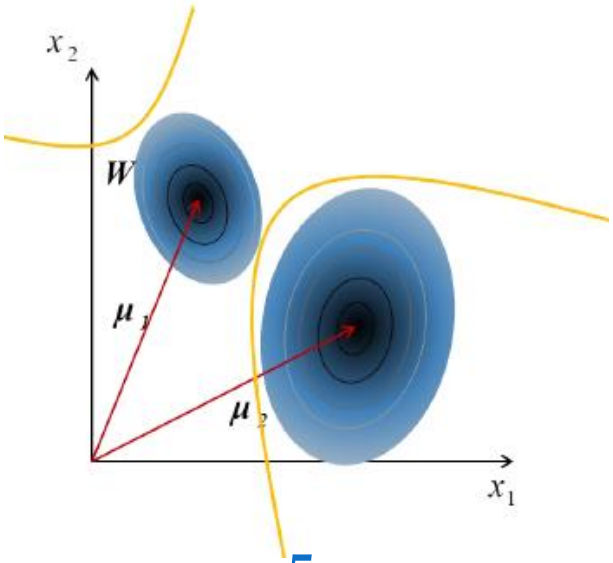
$$\underbrace{2log\left(\frac{P(C_j)}{P(C_i)}\right)}_{\text{A priori term - favours the class which is more likely}} + \underbrace{log\left(\frac{|\Sigma_i|}{|\Sigma_j|}\right)}_{\text{Favours the class whose covariant matrix has the smallest determinant (equivalent to smallest volume of the unit std dev contour)}}$$


x_2
H
x
$\|x-\mu_1\|$
$\|x-\mu_2\|$
μ_1
μ_2
x_1

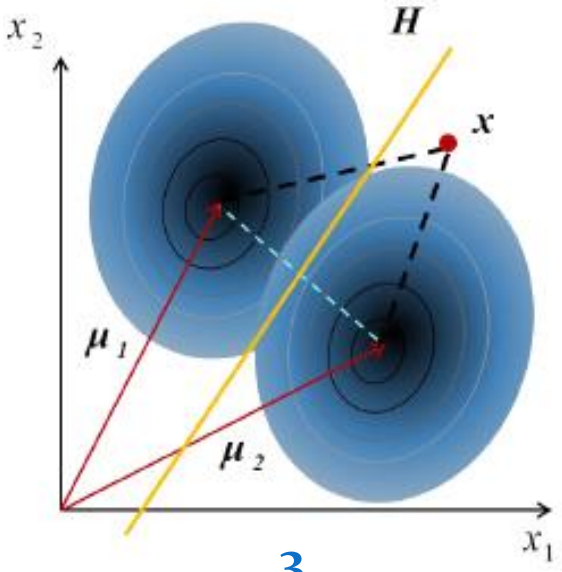
1


x_2
H
x
x_0
μ_1
μ_2
x_1

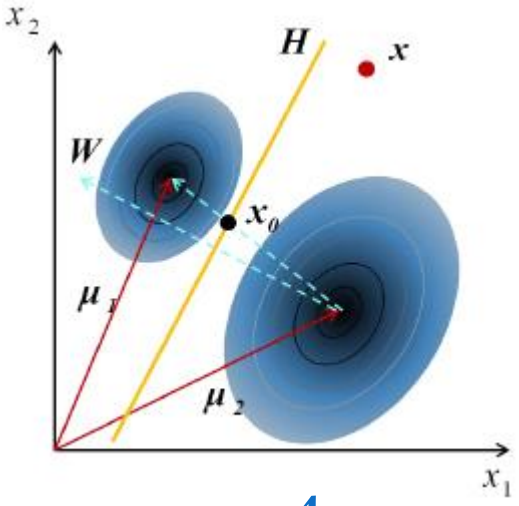
2


x_2
W
μ_1
μ_2
x_1

5


x_2
H
x
μ_1
μ_2
x_1

3


x_2
H
x
W
x_0
μ_1
μ_2
x_1

4

例： 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：

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

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

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：$P(\omega_1) = P(\omega_2) = 1/2$

试确定两类之间的决策边界。

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

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

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

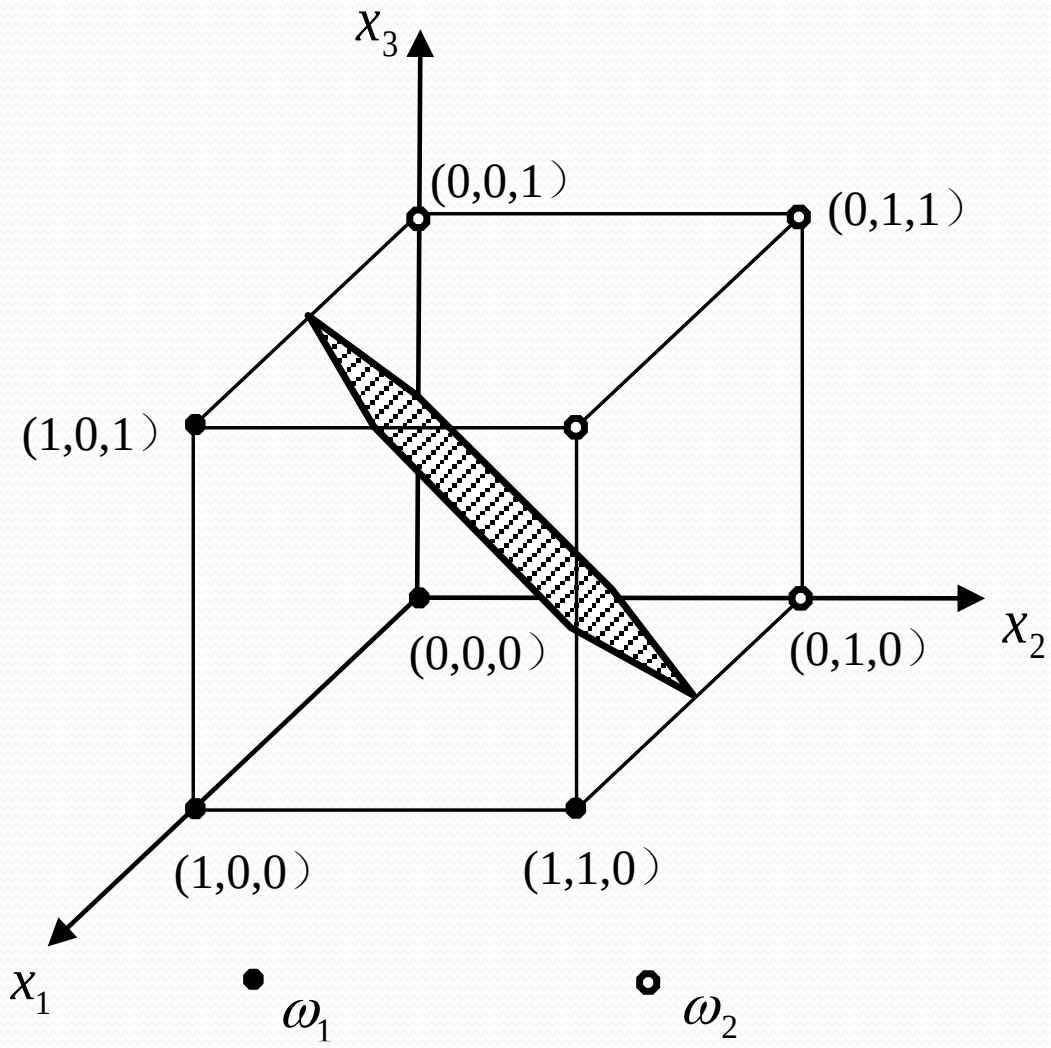
$$\boldsymbol{C}_1=\boldsymbol{C}_2=\frac{1}{16}\begin{bmatrix}3&1&1\\1&3&-1\\1&-1&3\end{bmatrix}$$ 经计算有：$$\boldsymbol{C}^{-1}=\begin{bmatrix}8&-4&-4\\-4&8&4\\-4&4&8\end{bmatrix}$$

协方差矩阵相等。由于 $P(\omega_1)=P(\omega_2)=\frac{1}{2}$

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

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

图中画出判别平面的一部分。



最大似然（ML）分类器

- 如果没有先验概率，或者不考虑先验概率，则MAP分类器变为最大似然分类器，即Maximum Likelihood(ML)分类器。

$$x \in C_i, \quad \text{if } p(x|C_i) > p(x|C_j),\ \forall j \neq i$$

- 如果观测似然是高斯分布，则ML分类器为：

$$(\underline{x}-\underline{\mu}_j)^T\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j) - (\underline{x}-\underline{\mu}_i)^T\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) \underset{C_j}{\overset{C_i}{\gtrless}} 2log\left(\frac{|\Sigma_i|^{1/2}}{|\Sigma_j|^{1/2}}\right)$$

损失函数

- 通常情况下，基于分类的决策机制是与决策损失（loss）相连的。同时，针对各个类，决策损失也是不一样的。
- 例如，测试样本本来属于类2，但是分类器把它错误的分类为属于类1，则该项决策导致的损失有可能会大于相反的情况导致的损失。
- 尤其是当一个类非常稀少、但错误决策又会产生很大代价的情况，例如，银行卡盗刷行为。

$$p(\underline{x}|C_1)P(C_1) \underset{2}{\overset{1}{\gtrless}} p(\underline{x}|C_2)P(C_2)$$

- 假设C2类非常稀少，会导致$P(C1) \gg P(C2)$。对于分类器而言，如果把测试样本都判断为属于C1类，可以获得最低的整体错误率。但是，这是正确的分类决策吗?

条件风险

- 如何解决该问题？
  - 提高特征选择的性能，针对某些x的取值空间，使得观测似然$P(x|C_2) \gg P(x|C_1)$，以此对冲先验概率的巨大差异。但是，在具体实施时，如何选择特征才能获得上述结果是较为困难的。
  - 另一个方法就是给错误分类分配一个损失系数。
- 将分类的动作表示为：$\{\alpha_1, \alpha_2, \ldots, \alpha_{NC}\}$
- 真值是C_j，但是决策动作是α_2的损失表达为：$\lambda(\alpha_i|C_j)$。
- 给定一个测试模式x，可以定义采取某个分类动作α_i的条件风险为：

$$R(\alpha_i|\underline{x}) = \sum_{j=1}^{N_c} \lambda(\alpha_i|C_j) P(C_j|\underline{x})$$

Nc is the number of classes

The loss associated with action i when the true state is Cj

- 针对所有决策动作和候选类别，可以用一个矩阵来表示对应的损失值。

  以信用卡盗刷为例，损失矩阵可以表示如下：

	决策为盗刷	决策为正常
真值为盗刷	0	1000
真值为正常	1	0

- 损失λij具体数值可以手动设计、也可以通过机器学习训练。

风险贝叶斯分类决策

- 决策策略：**最小化整体风险。**针对一个测试样本 x，选择条件风险最低的类。

$$R(\alpha_i|\underline{x}) \underset{\alpha_i}{\overset{\alpha_j}{\gtrless}} R(\alpha_j|\underline{x})$$

Shorthand notation:

$$\lambda_{ij} \equiv \lambda(\alpha_i|C_j) \equiv \text{cost of action i given class j}$$

风险贝叶斯分类决策：两类的情况

- 如果假设错误分类到C1和错误分类到C2的损失是相同的，则为MAP分类。

Assume: $\lambda_{11} = \lambda_{22} = 0$ (no loss for correct choice)

$\lambda_{12} = \lambda_{21} = 1$ (same loss for incorrect choice)

$$R(\alpha_1|\underline{x}) \underset{\alpha_1}{\overset{\alpha_2}{\gtrless}} R(\alpha_2|\underline{x})$$

$$\mathbf{0}\,\lambda_{11}P(C_1|\underline{x}) + \mathbf{1}\,\lambda_{12}P(C_2|\underline{x}) \underset{\alpha_1}{\overset{\alpha_2}{\gtrless}} \mathbf{1}\,\lambda_{21}P(C_1|\underline{x}) + \mathbf{0}\,\lambda_{22}P(C_2|\underline{x})$$

$$P(C_2|\underline{x}) \underset{\alpha_1}{\overset{\alpha_2}{\gtrless}} P(C_1|\underline{x})$$

This is the MAP classifier.

- 如果假设错误分类到C1和错误分类到C2的损失是不同的：
- 如果$\lambda_{12} > \lambda_{21}$，意味着错误分类到C2的损失大于错误分类到C1的损失。
- 结论：针对该情况，通过使用风险贝叶斯决策，将有更多可能性选择C2类。

Assume: $\lambda_{11} = \lambda_{22} = 0$ (no loss for correct choice)

$\lambda_{12} \neq \lambda_{21}$ (different losses for incorrect choices)

$$\lambda_{11}P(C_1|\underline{x}) + \lambda_{12}P(C_2|\underline{x}) \underset{\alpha_1}{\overset{\alpha_2}{\gtrless}} \lambda_{21}P(C_1|\underline{x}) + \lambda_{22}P(C_2|\underline{x})$$

Use Bayes' theorem: $$\lambda_{12}p(\underline{x}|C_2)P(C_2) \underset{\alpha_1}{\overset{\alpha_2}{\gtrless}} \lambda_{21}p(\underline{x}|C_1)P(C_1)$$

风险贝叶斯分类决策：例子1

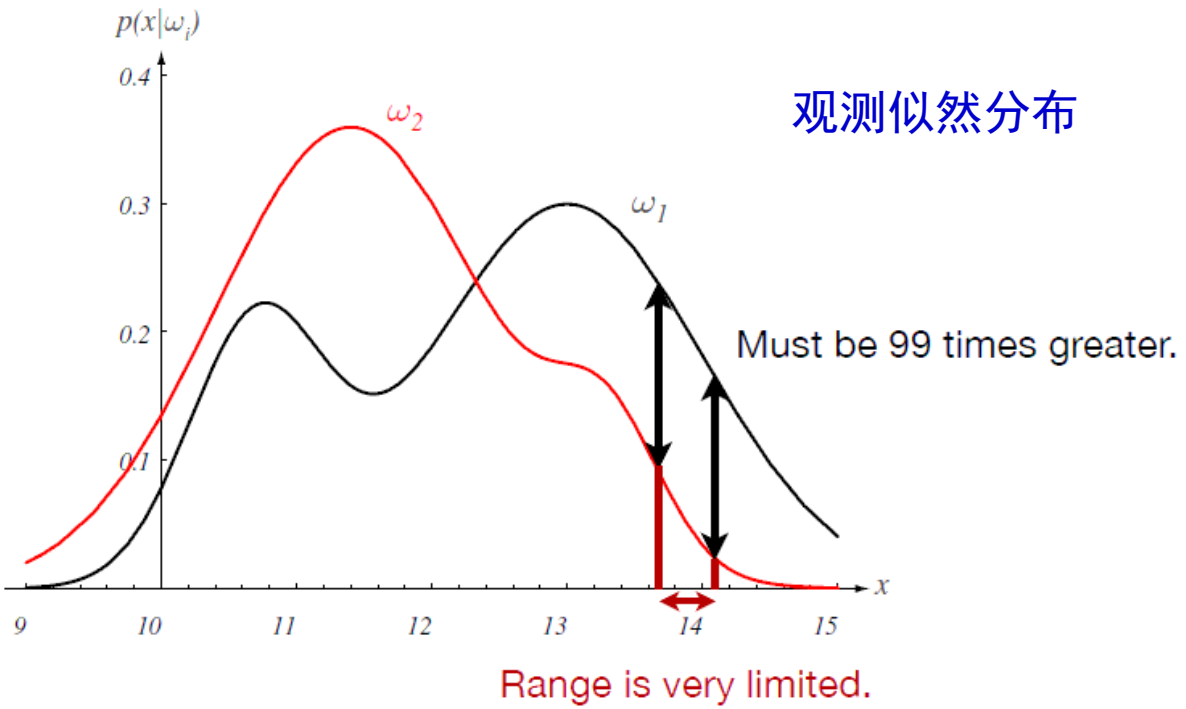
- 信用卡盗刷行为分类：根据用户的信用卡使用特征，实时监测每一笔交易是合法还是非法盗刷。
- 特征：交易地点、交易类型、速度等。

Assuming that the amount of fraudulent activity is about 1% of the total credit card activity:

C_1 = Fraud $P(C_1) = 0.01$

C_2 = No fraud $P(C_2) = 0.99$

- 如果两种错误分类的损失都是一样的话，相当于是MAP分类器。
- 如果要判断属于盗刷类（C1）类，该类观测似然要至少比C2类高99倍，从而导致特征只有在很小的范围才能达到目标。



风险贝叶斯分类决策：损失的设置

- 所以，不同的错误决策对应的损失系数不应该一样。
- 如果把盗刷错误判断为合法，则会遭受财产损失，损失程度会很重。
- 如果过多的把合法刷卡错误判断为盗刷，也会一定程度影响客户对银行的反感，但是损失程度没有那么重。
- 因此，两种错误分类决策对应的损失系数可以设置为：

$$\lambda_{12} = 0.01 \qquad \lambda_{21} = 0.5$$

A missed fraud is 50 times more expensive than accidentally freezing a card due to legitimate use.

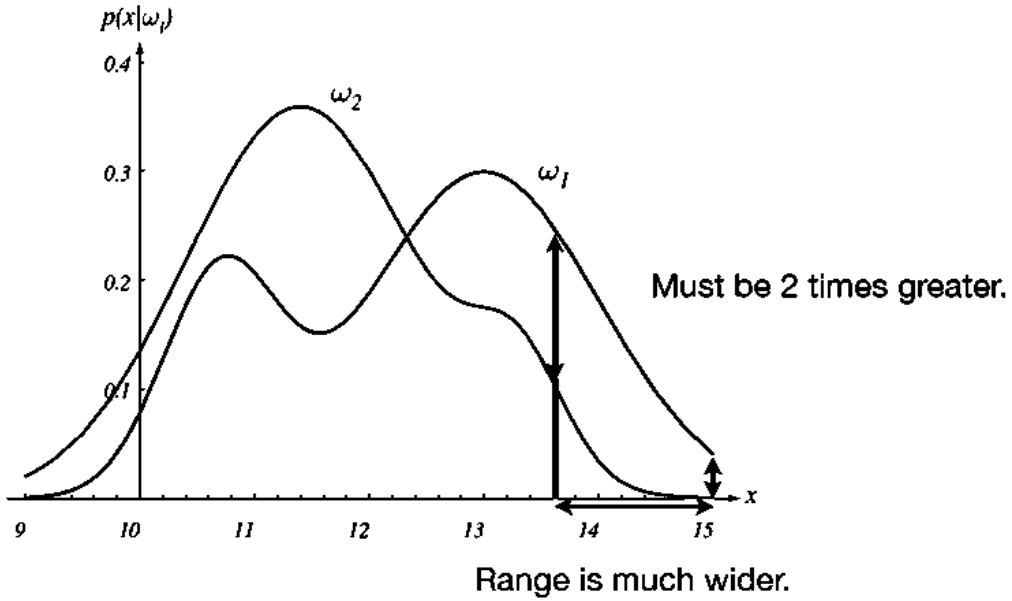
风险贝叶斯分类决策：例子2

- 信用卡盗刷行为分类：使用不同程度的损失系数。

$$p(\underline{x}|C_1) \underset{\alpha_2}{\overset{\alpha_1}{\gtrless}} 99 \frac{\lambda_{12}}{\lambda_{21}} p(\underline{x}|C_2)$$

$$p(\underline{x}|C_1) \underset{\alpha_2}{\overset{\alpha_1}{\gtrless}} 99 \frac{1}{50} p(\underline{x}|C_2)$$

$$p(\underline{x}|C_1) \underset{\alpha_2}{\overset{\alpha_1}{\gtrless}} 2\, p(\underline{x}|C_2)$$



- 使用不同程度的损失系数，分类边界还是产生了较大变化。

例 在细胞识别中，病变细胞和正常细胞的先验概率分别为：

$$P(\omega_1)=0.05,\quad P(\omega_2)=0.95$$

现有一待识别细胞，观察值为$\boldsymbol{X}$，观测似然取值：

$$p(\boldsymbol{X}\mid\omega_1)=0.5,\quad p(\boldsymbol{X}\mid\omega_2)=0.2$$

损失函数分别为$L_{11}=0$，$L_{21}=10$，$L_{22}=0$，$L_{12}=1$。按最小风险贝叶斯决策分类。

对样本 X:

当X被判为ω_1类时:

$$r_1(\boldsymbol{X}) = L_{11}\, p(\boldsymbol{X} \mid \omega_1) P(\omega_1) + L_{12}\, p(\boldsymbol{X} \mid \omega_2) P(\omega_2)$$

当X被判为ω_2类时:

$$r_2(\boldsymbol{X}) = L_{21}\, p(\boldsymbol{X} \mid \omega_1) P(\omega_1) + L_{22}\, p(\boldsymbol{X} \mid \omega_2) P(\omega_2)$$

决策规则:

若 $r_1(\boldsymbol{X}) < r_2(\boldsymbol{X})$　　则 $\boldsymbol{X} \in \omega_1$

若 $r_1(\boldsymbol{X}) > r_2(\boldsymbol{X})$　　则 $\boldsymbol{X} \in \omega_2$

即：

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

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

$$\frac{p(\boldsymbol{X}\mid\omega_1)}{p(\boldsymbol{X}\mid\omega_2)}>\frac{(L_{12}-L_{22})P(\omega_2)}{(L_{21}-L_{11})P(\omega_1)}$$

则 $\boldsymbol{X}\in\omega_1$

则 $\boldsymbol{X}\in\omega_2$

$$\frac{p(\boldsymbol{X}\mid\omega_1)}{p(\boldsymbol{X}\mid\omega_2)} > \frac{(L_{21}-L_{22})P(\omega_2)}{(L_{12}-L_{11})P(\omega_1)}$$

令：$l_{12}(\boldsymbol{X}) = \dfrac{p(\boldsymbol{X}\mid\omega_1)}{p(\boldsymbol{X}\mid\omega_2)}$，称**似然比**；

$\theta_{12} = \dfrac{(L_{21}-L_{22})P(\omega_2)}{(L_{12}-L_{11})P(\omega_1)}$，为阈值。

判别步骤：

① 定义损失函数L_{ij}。

② 计算θ_{12}。

③ 计算$l_{12}(\boldsymbol{X})$。

④ 若 $l_{12}(\boldsymbol{X}) > \theta_{12}$，则 $\boldsymbol{X} \in \omega_1$

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

若 $l_{12}(\boldsymbol{X}) = \theta_{12}$，任意判决

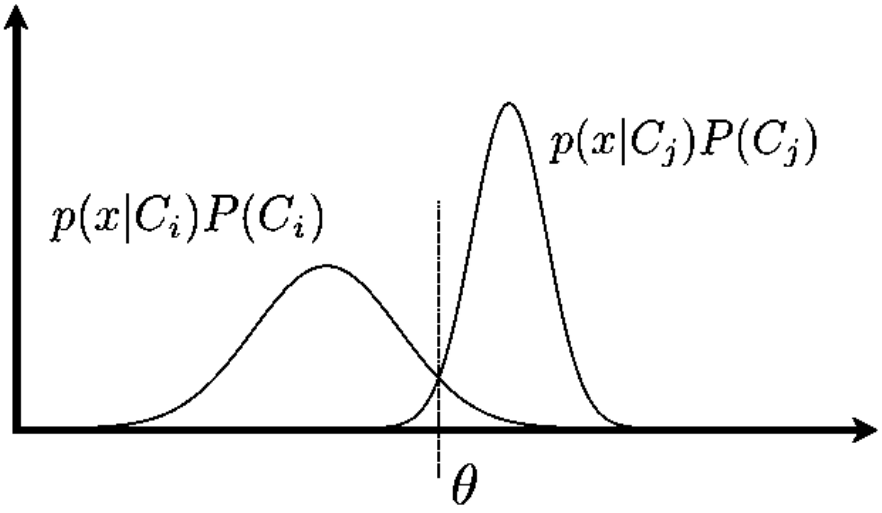
最大后验概率（MAP）分类器

- 最大后验概率(MAP)分类器：将测试样本分给后验概率最大的那个类。

$$P(C_i|\mathbf{x}) \gtrless_{C_j}^{C_i} P(C_j|\mathbf{x})$$

$$\frac{P(\mathbf{x}|C_i)P(C_i)}{P(\mathbf{x})} \gtrless_{C_j}^{C_i} \frac{P(\mathbf{x}|C_j)P(C_j)}{P(\mathbf{x})}$$

where $P(\mathbf{x}) = \sum_k P(\mathbf{x}|C_k)P(C_k)$



Note that $P(\mathbf{x}|C_i)P(C_i) = P(\mathbf{x}, C_i)$

MAP分类器：平均概率误差最小

- 给定一个样本x，MAP分类决策产生的误差可以用概率误差表达。
- 概率误差等于选错类所对应的后验概率。

$$P(error|\mathbf{x}) = \begin{cases} P(C_2|\mathbf{x}) & \text{if we decide } \mathbf{x} \in R_1 \\ P(C_1|\mathbf{x}) & \text{if we decide } \mathbf{x} \in R_2 \end{cases}$$

- 给定所有样本（N为样本个数），分类决策产生的平均概率误差为：样本的概率误差的均值。

$$P(error) = \frac{1}{N}\sum_{\mathbf{x}\in R_1 \cup R_2} P(error, \mathbf{x}) = \frac{1}{N}\sum_{\mathbf{x}\in R_1 \cup R_2} P(error|\mathbf{x})P(\mathbf{x})$$

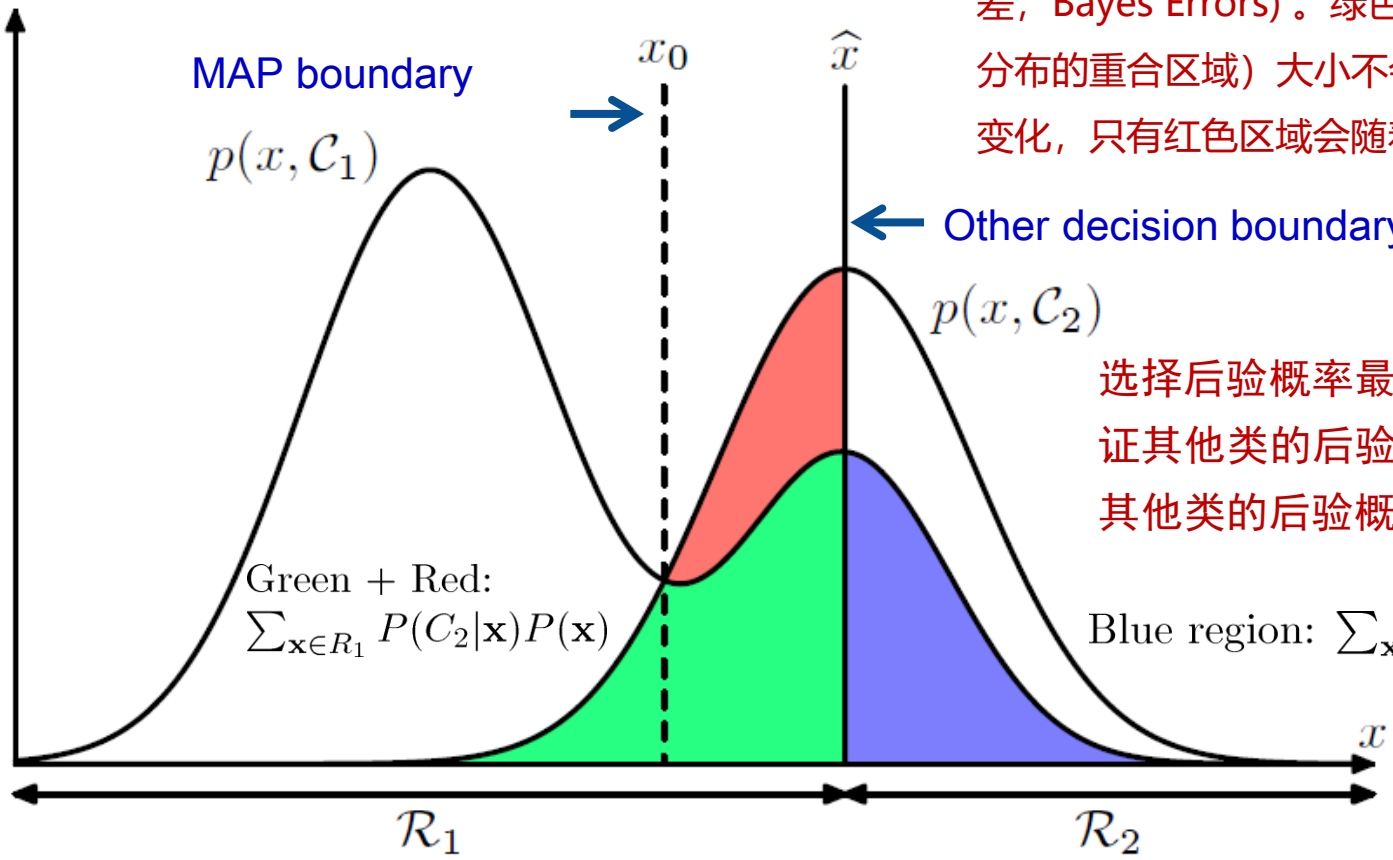
$$= \frac{1}{N}\left[\sum_{\mathbf{x}\in R_1} P(C_2|\mathbf{x})P(\mathbf{x}) + \sum_{\mathbf{x}\in R_2} P(C_1|\mathbf{x})P(\mathbf{x})\right]$$

- 给定所有样本，MAP分类器选择后验概率最大的类，等于最小化平均概率误差：因为针对每个x，MAP分类器都选择一个概率误差最小的，则对所有样本而言，平均概率误差也是最小的。

MAP分类器：分类误差最小化

- 概率误差最小化：就是分类误差最小化。

目标：最小化红色、绿色和蓝色区域总和(贝叶斯误差，Bayes Errors)。绿色和蓝色区域（两个类概率分布的重合区域）大小不会随着决策边界的移动而变化，只有红色区域会随着决策边界的移动而变化。



选择后验概率最大的类，虽然不能保证其他类的后验概率为0，但能够保证其他类的后验概率是小的。

多类情况错误率

设共有M类，当判决 $\boldsymbol{X} \in \omega_i$ 时：

$$\text{错误率}=\sum_{\substack{j=1\\ j\neq i}}^{M}\int_{R_i} P(\omega_j \mid \boldsymbol{X})p(\boldsymbol{X})d\boldsymbol{X} = \sum_{\substack{j=1\\ j\neq i}}^{M}\int_{R_i} p(\boldsymbol{X} \mid \omega_j)P(\omega_j)d\boldsymbol{X}$$

当X判为任何一类时，都存在这样一个可能的错误，故

总错误率为：

$$P(e)=\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M}\int_{R_i} p(\boldsymbol{X} \mid \omega_j)P(\omega_j)d\boldsymbol{X}$$

正确分类概率

$$P(c)=\sum_{i=1}^{M}\int_{R_i} p(\boldsymbol{X} \mid \omega_i)P(\omega_i)d\boldsymbol{X}$$

则：$P(e)=1-P(c)$

把样本归入到后验概率最大的类别中，那么显然分类正确率也是最大的，所以错误率也是最小的。

贝叶斯分类器

- 贝叶斯分类器：将测试样本分给风险最小的那个类。决策就是选取的类别。
- 针对每个样本x，决策属于C_i类的动作对应的风险评估计算如下：

$$R(\alpha_i|\mathbf{x}) = \sum_k \lambda_{ik} P(C_k|\mathbf{x})$$

$\lambda(\alpha_i|C_k)$：真值是C_k，但是决策动作是α_i的损失。

风险评估的含义：对于属于 C_i类 这个决策，机器并不知道正确类是什么，所以需要通过计算该决策与所有其他可能的正确决策选项相比的损失之和，作为评估该决策风险的依据。

- 因此，针对一个样本x，贝叶斯决策选择风险最小的类，其决策公式为：

$$R(\alpha_i|\mathbf{x}) \gtrless_{\alpha_i}^{\alpha_j} R(\alpha_j|\mathbf{x})$$

- Loss matrix：表征机器决策与其他可能正确决策选项之间的损失关系。

	cancer	normal
cancer	0	1000
normal	1	0

贝叶斯分类器：期望损失最小化

- 给定所有样本（N为样本个数），贝叶斯分类决策的**期望损失**(expected loss)为：所有样本决策到每一个类的风险损失的平均值。

$$R(\{\mathbf{x}\}) = \sum_i R(\alpha_i|\{\mathbf{x}\}) = \frac{1}{N}\sum_i \sum_k \lambda_{ik} \sum_{\mathbf{x}\in R_i} P(C_k|\mathbf{x})$$

x is decided to be Ci

- 给定所有样本，贝叶斯分类决策目标函数是：**最小化期望损失**。

- 如何实现期望损失最小化？通过针对每个样本选择风险最小所属的类，使得决策判断属于类Ci的所有样本,贝叶斯决策保证平均风险最小。

$$\min R(\{\mathbf{x}\}) = \min \sum_i R(\alpha_i|\{\mathbf{x}\}) \Rightarrow \sum_i \sum_{\mathbf{x}\in R_i} [\min \sum_k \lambda_{ik} P(C_k|\mathbf{x})]$$

$$\Rightarrow \min \sum_k \lambda_{ik} P(C_k|\mathbf{x}) = \min R(\alpha_i|\mathbf{x})$$

拒绝选项

- 以MAP分类器为例，针对每个测试样本，分类器通过比较所有类的后验概率，选择最大的后验概率所属的类。
- 但是，有可能最大后验概率的绝对值比较小，即属于该类的决策有很大的不确定性。为了避免出现错误决策，分类器可以选择决绝。
- 如何拒绝？引入阈值θ：

$$\text{Reject } \mathbf{x} \in C_i, \text{ if } P(C_i|\mathbf{x}) \leq \theta$$

$$\text{when } P(C_i|\mathbf{x}) > P(C_j|\mathbf{x}), \forall i \neq j$$

当θ =1，所有样本的任何决策都会被拒绝。

当$\theta < 1/K$，所有样本的决策都不会被拒绝，K是类别的个数。

- 拒绝选项也可用到贝叶斯分类器：当最小的风险仍然大于阈值时，选择拒绝。