—武大本科生课程

第4讲　概率分类(I)

(Lecture 4 Probability classification: Part 1)

武汉大学计算机学院机器学习课程组

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)

4.0 一些概念回顾/归纳

1.机器学习中的预测性建模方法

(1) 代数方法：判别式建模方法，借助训练数据对观测量和预测量的函数关系进行直接建模，基于函数关系对变量取值进行数值预测。利用矢量空间的直观概念，使用代数方程方法，对模式进行分类

如：**KNN**，感知机，判别分析，决策树，随机森林，支持向量机、逻辑回归，神经网络

(2) 概率方法：生成式建模方法，借助训练数据对同类数据的生成机制(概率分布)进行估计，基于概率关系对变量取值进行概率预测。把模式视为随机变量的抽样，利用统计决策理论(贝叶斯统计)成熟的判决准则与方法，对模式样本进行分类

如：**贝叶斯分类器**、贝叶斯网络(概率图模型)、高斯混合模型、隐马尔可夫模型、受限玻尔兹曼机、生成对抗网络，变分自动编码器

2. 线性可分概念与线性分类算法

一个分类问题是否属于**线性可分**，取决于是否有可能找到一个点、直线、平面或超平面来分离开两个相邻的类别。

如果每个类别样本的分布范围本身是全连通的单一凸集，且互不重叠，则这两个类别一定是线性可分的，见下图所示。

线性分类算法主要有**线性判别函数**、**Fisher法 (*：Ch3 选学)** 、单层感知器、逻辑回归等。

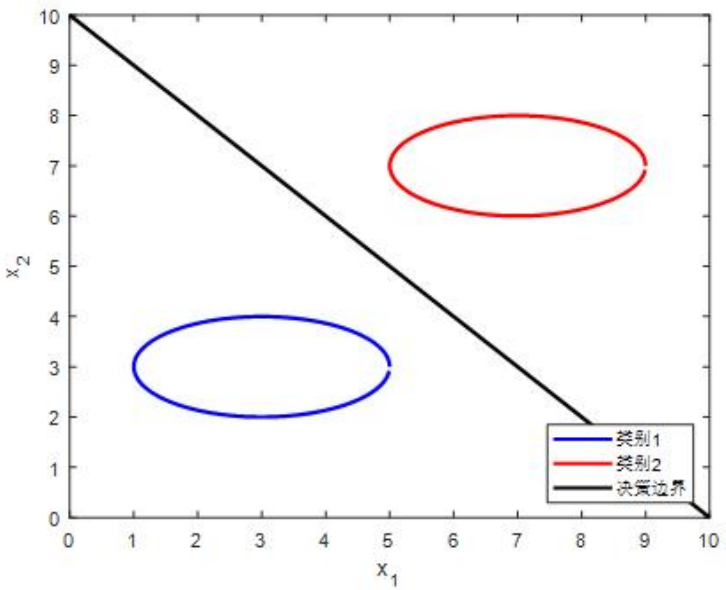


图　线性可分情况

Fisher LDA Python实现两例：(1) 蠓的二分类；(2) 鸢尾花三分类。 见CSDN博文
https://yuanyx.blog.csdn.net/article/details/114813129

3. 回归与分类

线性回归：基于线性模型的回归，相应的线性模型称为线性回归模型。一般而言，解释变量和响应变量都是连续值。

对于任意给定的样本$\mathbf{x}$，线性回归模型表示为：

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m$$

其中$\boldsymbol{w} = (w_1, w_2, \ldots, w_m)^T$为参数向量（权向量）。

线性回归与线性分类问题非常相似，只不过线性回归的响应变量一般是定量型，而线性分类涉及的响应变量主要是定性型，如：有病 vs 无病；穷人 vs 富人。为了适合计算，把这些定性变量用{0, 1}或(−1, +1}代替。

3. 回归与分类

使用符号函数作为判决函数：简单观察可知，前面讨论线性判别分类时，实际上是在线性回归函数$f(\boldsymbol{x})$（也即$d(\boldsymbol{X})/\mathrm{g}(\boldsymbol{X})$）上，复合了一个符号型判决函数：

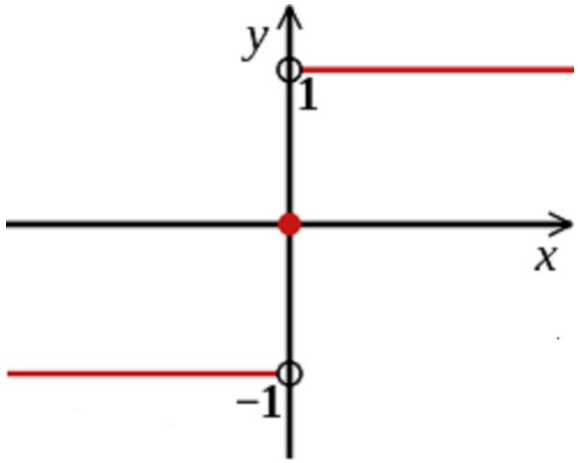
$$\mathrm{y}=\mathbf{sign}(f(\boldsymbol{x}))=\mathbf{sign}(w_1x_1+w_2x_2+\cdots+w_mx_m)$$

这个判决函数非常硬——只看是否大于零，不管与零有多远！

我们如何利用成熟的线性回归方法进行更细致的分类呢？

我们能不能更多地利用判决函数值的大小信息、而不仅仅是利用其正负信息呢？

$$\mathrm{sign(x)}=\begin{cases}1, & \mathrm{x}>0\\ 0, & \mathrm{x}=0\\ -1, & \mathrm{x}<0\end{cases}$$



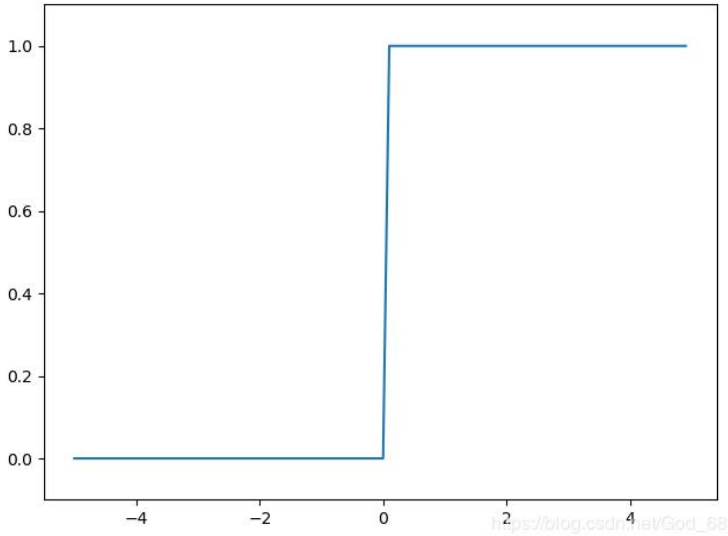
3. 回归与分类

使用跃阶函数作为判决函数：当类标记从{1，−1}改为{1, 0}时，相当于把判决函数改为阶跃函数（step function）

$$\mathbf{step}(x) = \frac{1}{2}(\mathbf{sign}(x) + 1)$$

当$f(x)$大于零时将样本x划分为正类；小于零时将x划分为负类；等于零时将x随机划分。

$$\mathbf{step}\big(f(\boldsymbol{x})\big) = \begin{cases} 0 & f(\boldsymbol{x}) < 0; \\ 0.5 & f(\boldsymbol{x}) = 0; \\ 1 & f(\boldsymbol{x}) > 0; \end{cases}$$



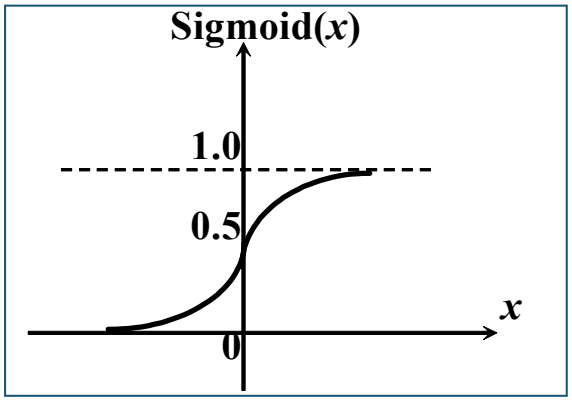
3. 回归与分类

使用Sigmoid作为分类判决函数：具有良好数学性质的判别函数，克服了跃阶函数的不连续性，实现对连续值的离散化

$$g(f(\boldsymbol{x})) = \text{Sigmoid}(f(\boldsymbol{x})) = \frac{1}{1+e^{-w^T X}}$$

令$H(\boldsymbol{x}) = \text{Sigmoid}(f(\boldsymbol{x}))$，$H(\boldsymbol{x})$值域为（0, 1），可将$H(\boldsymbol{x})$看成是一个关于$\boldsymbol{x}$的概率分布，即：

- 若$H(\boldsymbol{x})$的值越接近1，则$\boldsymbol{x}$属于正类的可能性就越大
- 若$H(\boldsymbol{x})$的值越接近0，则$\boldsymbol{x}$属于正类的可能性就越小



4. 两类研究对象

获取样本的观察值时，有两种情况：

- **确定性事件**：事物间有确定的因果关系，即在一定条件下，他必然会发生或必然不发生。
- **随机性事件**：事物间没有确定的因果关系，观察到的特征的值的出现具有统计特性，是一个随机向量。

对于随机模式向量只能利用模式集的统计特性来分类，以使分类器发生错误的概率最小，这就是本章要介绍的统计决策理论。

4.1 Bayes决策的引入

前面我们学习了线性判别分析及分类器设计方法(如线性分类器中Fisher法等)。

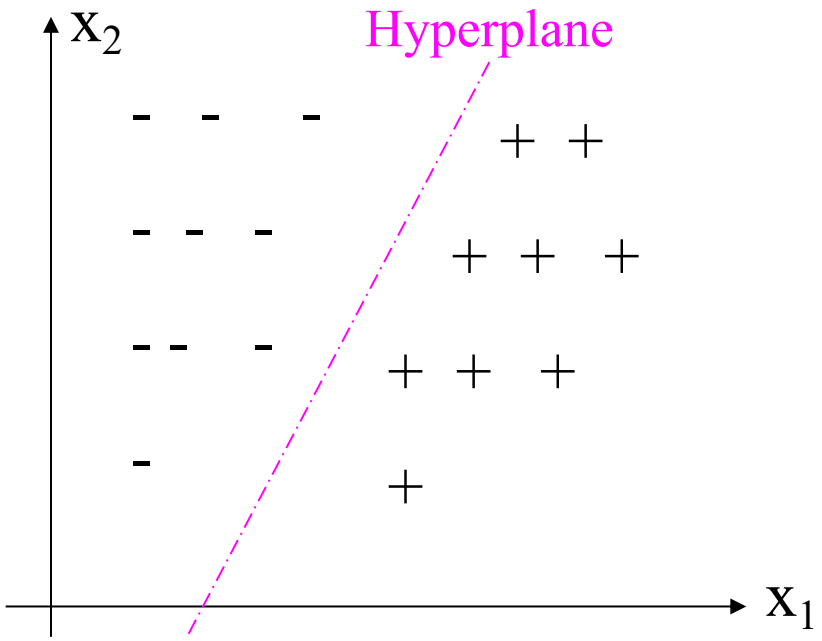
现在的问题是，当分类器设计完成后，对待测样本进行分类，一定能分类正确吗？若有错分情况发生，是在哪种情况下出现的？错分类的可能性有多大？这些都是机器学习与模式识别所涉及的重要问题。

4.1.1 Bayes决策要解决的问题

这里以某制药厂生产的药品检验识别为例，说明Bayes决策要解决的问题。

+:正常药品

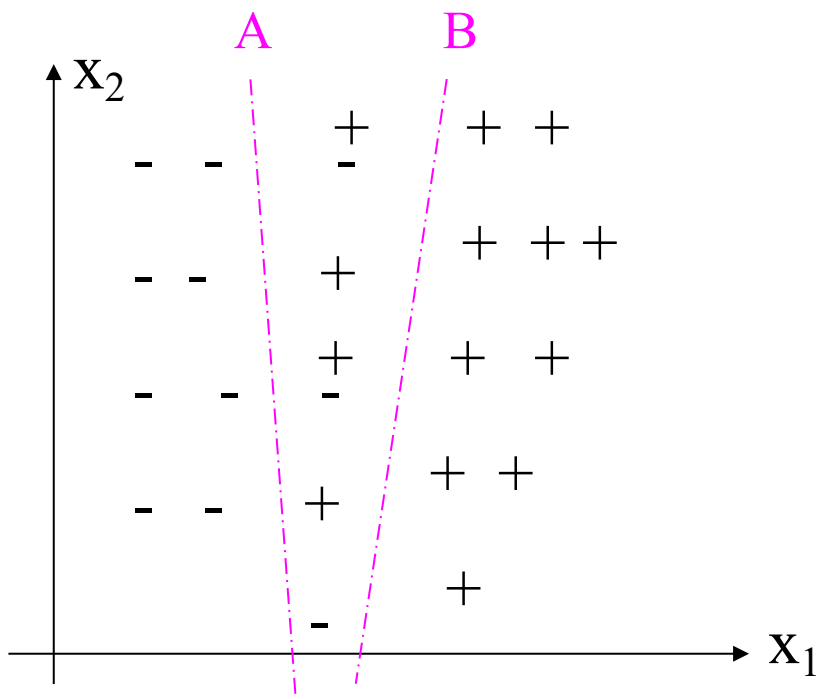
-: 异常药品



线性可分示意图

对于左图的线性分类器，可用一直线作为决策面 (即分界面)。若待识别的药品模式向量**x**被划分到直线的右侧，则其为正常药品，若被划分到直线的左侧，则其为异常药品，可见对其作出决策是很容易的，也不会出现什么差错。

问题：可能出现模棱两可的情况，见下图。此时，任何决策都存在判错的可能性。



线性不可分示意图

在直线A、B之间，属于不同类的样本在特征空间中相互穿插，很难用简单的分界线将它们分开。

如果以错分类最小为原则分类，则图中A可能是最佳分界线，它使错分类的样本数量最小。但是，若将一个“-”样本错分成“+”类，所造成的损失要比将“+”分成“-”类严重，这是由于将异常药品误判正常药品，则会让病人失去治疗时机而遭受大的损失；把正常药品误判为异常药品会给企业带来一点损失，则偏向使对“-”类样本的错分类进一步减少，可能使总的损失(风险)为最小，那么B直线就可能比A直线更适合作为分界线。

可见，分类器参数的选择或者学习过程得到的结果取决于设计者选择什么样的准则函数(目标函数)。不同的准则函数的最优解对应于不同的学习结果，得到性能不同的分类器。

错误分类难以避免，这种可能性可用$P(\omega_i|\ \mathrm{x})$表示，如何做出合理的判决就是Bayes决策所要讨论的问题。其中，最具代表性的是基于最小错误的Bayes决策与基于最小风险的Bayes决策。

(1)基于最小错误的Bayes决策

它指出机器自动识别出现错分类的条件，错分类的可能性如何计算，如何实现使错分类出现的可能性最小。

(2)基于最小风险的Bayes决策

错误分类有不同情况，从前面的图可看出，两种错误造成的损失不一样，不同的错误分类造成的损失会不相同，前一种错误更可怕，因此就要考虑减少因错误分类造成的危害损失。为此，引入一种“风险”与“损失”的概念，希望做到使风险最小，减少危害大的错分类情况。

4.1.2 Bayes公式

托马斯·贝叶斯
(Thomas Bayes）(1702–1761)

英国数学家，做过神父，1742年成为英国皇家学会会员。 Bayes在数学方面主要是研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了Bayes统计理论，对于统计决策函数、统计推断、统计估算等做出了杰出贡献。

4.1.2 Bayes公式

若已知总体共有M类模式，以及各类在这n维特征空间的统计分布，具体说就是已知各类别ω_i，i=1,2,……,M的先验概率$P(\omega_i)$及类条件概率密度函数$P(\mathbf{x}|\omega_i)$。对于待测样本x，Bayes公式可以计算该样本分属各类别的概率$P(\omega_i|\mathbf{x})$，即后验概率。看x属于哪个类的可能性最大，就把x归于可能性最大的那个类，后验概率作为识别对象归属的依据。Bayes公式为：

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

类别的状态是一个随机变量，而某种状态出现的概率是可以估计的。Bayes公式体现了先验概率、类条件概率密度函数和后验概率三者关系的式子。

4.1.2 Bayes公式

Bayes公式可用较直观的非正式英文表示为：

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

类概密$P(x|\omega_i)$也称为ω_i关于x的似然函数，或简称为“类似然”(*class* likelihood)，表明在其他条件都相等的条件下，使得$P(x|\omega_i)$较大ω_i的更有可能是真实的类别。注意到后验(posterior)概率主要是由似然函数和先验概率的乘积所决定的，证据(evidence)因子$P(x)$可以看成是一个标量因子(scalar factor)，是看到观察样本x的边缘概率(marginal probability)，$P(x)=\sum P(x, \omega_i)$（Sum rule, 加和规则），以保证各类别的后验概率总和为1从而满足概率条件。

机器学习中这三个概率的解释：

设随机样本的特征向量为$\boldsymbol{X}$，对应的类别变量用ω表示：

(1) ω_i的先验概率$P(\omega_i)$：根据以前知识和经验得出的、一般情况下**样本**属于ω_i类的概率——属**事先猜测**，故称**先验概率**。

(2) ω_i的后验概率$P(\omega_i|\boldsymbol{X})$：指获得样本$\boldsymbol{X}$后计算出的、当前$\boldsymbol{X}$可能属于$\omega_i$类的概率。虽然仍然是概率意义上的**事后猜测**，但这是在得到样本$\boldsymbol{X}$后的概率，故称为**后验概率**。

(3) $\boldsymbol{X}$的类条件概率$P(\boldsymbol{X}|\omega_i)$：对$\omega_i$类样本，其特征(数)值$\boldsymbol{X}$的概率。例：对一批得病患者进行化验，结果为阳性的概率为95%，ω_1代表“患病”，则“患病者$(\boldsymbol{X}\in\omega_1)$的化验结果$\boldsymbol{X}$为阳性”这一事件的概率可表示为

$$P(\boldsymbol{X}=\text{阳}\,|\,\omega_1)=0.95$$

1.先验概率$P(\omega_i)$-Prior probability

先验概率$P(\omega_i)$针对M个事件出现的可能性而言，不考虑其他任何条件。如，由统计资料表明总药品数为N，其中正常药品数为N1，异常药品数为N2，则：

$$P(\omega_1)=N_1/N$$

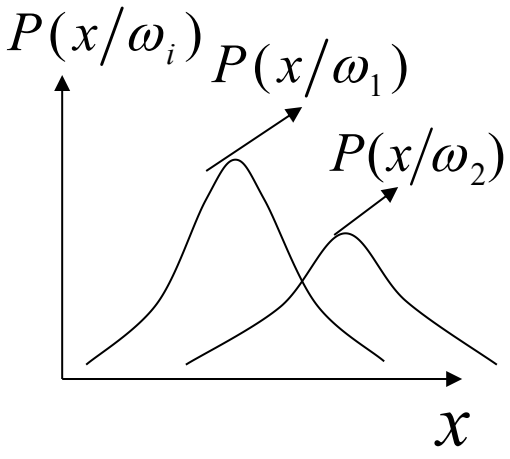
$$P(\omega_2)=N_2/N$$

称$P(\omega 1)$及$P(\omega 2)$为先验概率。显然，在一般情况下，正常药品占药品比例大，即$P(\omega 1)>P(\omega 2)$。仅按先验概率来决策，就会把所有药品都划归为正常药品，并没有达到将正常药品和异常药品区别分开的目的。这表明由先验概率所提供的信息太少。

2. 类条件概率密度函数$P(x|\omega i)$

类条件概率密度函数(Class-conditional probability density function)$P(x|\omega_i)$是指在ωi类条件下x的概率密度，即ωi类样本x的概率分布密度，简称为类概密/似然。

设只用一个特征进行分类，即n=1(特征数目)，并已知这两类的类条件概率密度函数分布，见右图。类概密$P(x|\omega_1)$是正常药品的属性(此处n=1，故为特征数值)分布，类概密$P(x|\omega_2)$是异常药品的属性分布。



类条件概率密度分布

在工程问题中，统计数据往往满足正态分布规律。若采用正态密度函数作为类概密的函数形式，则函数内的参数，如期望、方差是未知的。那么问题就变成如何利用大量样本对这些参数进行估计，只要估计出这些参数，类概密$P(x|\omega i)$就确定了。

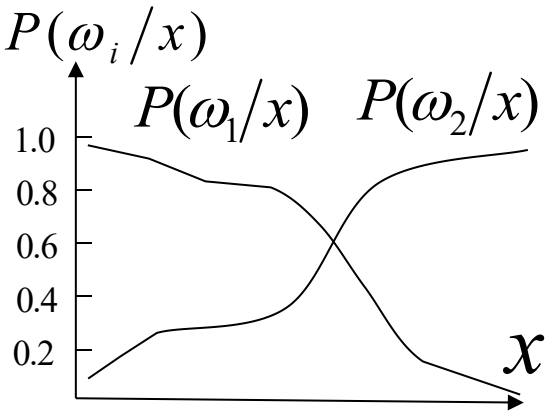
3. 后验概率$P(\omega_i|\ x)$-Posterior probability

条件概率$P(\omega i|\ x)$表示x出现的条件下ωi类出现的概率，称其为类别ωi的后验概率，对模式识别而言可理解为x来自ωi 类的概率。即若所观察到的某一样本的特征向量为x ，而在类中又有不止一类可能呈现这一x特征数值，它属于各类的概率就是$P(\omega i|\ x)$。

4. $P(\omega_1|x)$、 $P(\omega_2|x)$与$P(x|\omega 1)$、 $P(x|\omega 2)$的区别

(1) P(ω1| x)和P(ω2| x)是指在同一条件x下，比较ω1与ω1出现的概率，若P(ω1| x) > P(ω2| x)，则有结论：在x条件下，事件ω1出现的可能性大，见右图。对于两类的情况，则有P(ω1| x) + P(ω2| x)=1。

(2) P(x |ω1)和P(x |ω2)都是指各自条件下出现x的可能性，两者间没有联系，比较两者没有意义。 P(x |ω1)和P(x |ω2)是在不同条件下讨论问题，即使只有两类ω1与ω2， P(x |ω1)+P(x |ω2)≠1。不能仅因为P(x |ω1)>P(x |ω2)，就认为x是ω1类的可能性就大。只有考虑先验概率这一因素，才能决定x条件下，判别为ω1类或ω2类的可能性哪个大。



后验概率分布

思考题：

一口袋里有3只红球、2只白球，采用不放回方式摸取，求：

(1) 第一次摸到红球（记作A）的概率；

(2) 第二次摸到红球（记作B）的概率；

(3) 已知第二次摸到了红球，求第一次摸到的是红球的概率。

思考题参考解答：

1. P(A)=3/5，这就是先验概率；

2. P(B)=P(A)P(B|A)+P(/A)P(B|/A)=3/5×2/4+2/5×3/4=3/5；

3. P(A|B)=P(A)P(B|A)/P(B)=1/2，这就是后验概率。

当根据经验及有关资料推测出主观概率后，对其是否准确没有充分把握时，可采用概率论中的贝叶斯公式进行修正，修正前的概率称为先验概率{P(A)}，修正后的概率称为后验概率{P(A|B}，利用后验概率再进行风险分析。

4.2 分类器的描述方法

4.2.1 基本假设

给定模式空间S，由m个互不相交的模式类集合$\omega 1$，$\omega 2$, …, ωm组成:

(1) 假定类ωi的先验概率为$P(\omega i)$;

(2) 样本(或模式) **x**由特征向量来表示, 同样记为**x**, 假设为d维, 即$\mathbf{x}=(x_1, x_2, \ldots, x_d)$;

(3) 特征向量**x**的取值范围构成特征空间, 记为$\mathbf{R}^d$;

(4) 特征向量**x**的类条件概率密度函数为$p(\mathbf{x}|\omega i)$, 表示当样本**x**∈ωi时, 特征向量**x**的概率密度函数。

(5) 特征向量**x**的后验概率为$P(\omega i|\mathbf{x})$, 表示在特征向量**x**出现的条件下, 样本**x**来自类ωi的概率, 即类ωi出现的概率。

模式识别就是根据特征向量**x**的取值, 依据某个判决准则把样本**x**划分到$\omega 1$，$\omega 2$, …, ωm中的一个。

4.2.2　模式分类器的描述

模式分类器的描述方法有多种, 这里仅介绍以下三种描述方法, 它们之间是统一的。

1. 映射描述法

由于我们获取的有关观察对象的数据总量是有限的，因此，可用一个d+1维向量表示，即：

$$(x_1, x_2, \ldots, x_d; \alpha)$$

其中：$(x1, x2, \ldots, xd)$为特征向量，是特征空间$\mathbf{R}^d$中的一个点；α取值于集合$\{1, 2, \ldots, m\}$，表示模式的真实类别号，是未知的量，m为类别数。模式分类的实质在于实现特征空间$\mathbf{R}^d$到类别号空间$\{1, 2, \ldots, m\}$的一个映射，即

$$\mathbf{R}^d \rightarrow \{1, 2, \ldots, m\}$$

给定一个映射f，就给出了一种模式识别方法，不同的映射对应不同的分类方法，这就是模式识别问题的**映射描述法**。

2. 划分描述法

由于每个特征向量是R^d空间的一个点，且$R^d \rightarrow \{1, 2, \ldots, m\}$是一个多对一的映射，通过映射，本质上实现了对空间$\mathbf{R}^d$的一种划分，即把$R^d$**划分成个m不相重叠的区域**，每一个区域对应一个类别。区域R_i对应第i类ωi。

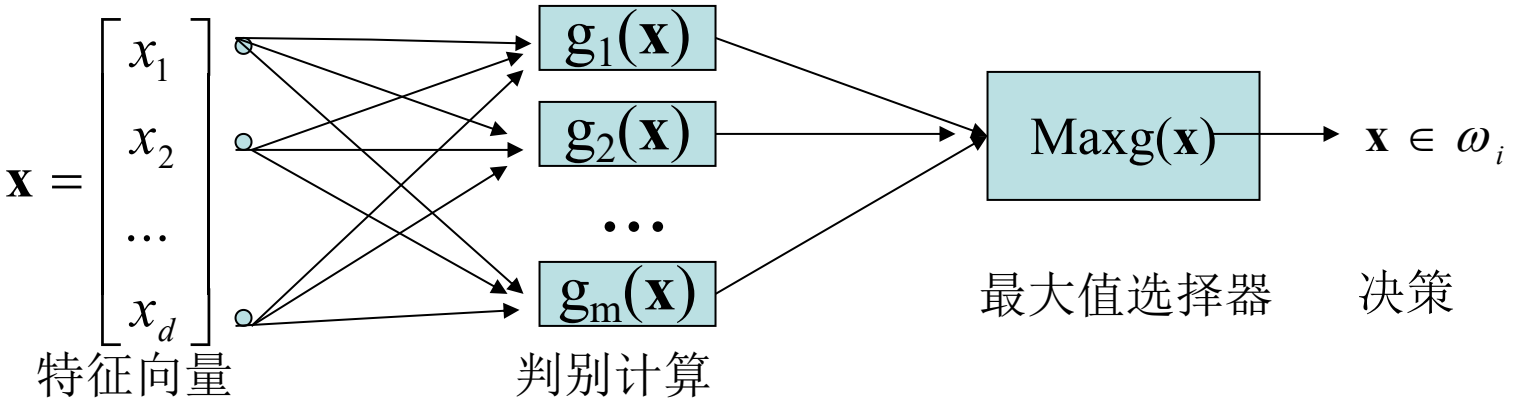
3. 判别函数法

把分类问题对应为R^d空间上的多元函数，通常称为判别函数(或称判决函数)$g_i(\mathbf{x})$， i=1，2，…，m。

将样本$\mathbf{x}$判属有极大(或极小)函数值的那一类。

模式分类实际上是将特征空间划分为不同的决策区域, 相邻决策区域被决策面所分割, 这些决策面是特征空间中的**超曲面**(*hypersurface*), 其决策面方程满足相邻两个决策域的判别函数相等, 即 $g_i(\mathbf{x})=g_j(\mathbf{x})$

分类器可被看作是一个计算m类个判别函数并选取最大(或最小)判决值对应的类别的网络或者机器，见下图。



4.3 最小错误率Bayes决策

(Minimum-error-rate classification)

对于**两类分类问题**，**最小错误率Bayes分类**(**最大后验概率分类**)的基本思想是：对于模式$\mathbf{x}$，如果属于模式类$\omega 1$的概率大于模式类$\omega 2$的概率，则决策模式$\mathbf{x}$属于模式类$\omega 1$；反之，决策模式$\mathbf{x}$属于模式类$\omega 2$。用数学语言描述为：

$$\text{若}P(\omega_1|\mathbf{x}) \underset{<}{\overset{>}{}} P(\omega_2|\mathbf{x}),\text{则}\mathbf{x}\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

--后验概率形式(基本形式1)

由*Bayes*定理

$$P(\omega_i|\mathbf{x})=\frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})},$$

同时考虑到$P(\mathbf{x})>0$,则上面决策可改写为:

$$\text{若}P(\mathbf{x}|\omega_1)P(\omega_1) \underset{<}{\overset{>}{}} P(\mathbf{x}|\omega_2)P(\omega_2),\text{则}\mathbf{x}\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

--类概密形式(基本形式2)

推广到m类，假设待识别的模式为$\mathbf{x}$, 模式(样本)分为m类，各类的先验概率和各类的类概密均已知，就有m个判别函数，由Bayes公式可知：

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

在得到一个观察模式$\mathbf{x}$后，在模式$\mathbf{x}$的条件下，看哪个类的概率最大，就应该把$\mathbf{x}$归于概率最大的那个类。由此，可得到最大后验概率判决准则的几种等价形式：

(1) 若$p(\mathbf{x} \mid \omega_j)P(\omega_j) = \max\limits_{i \in \{1,2,\mathrm{L}\ ,m\}} p(\mathbf{x} \mid \omega_i)P(\omega_i)$, 则$\mathbf{x} \in \omega_j$;

(2) 若$L(\mathbf{x}) = \dfrac{p(\mathbf{x} \mid \omega_j)}{p(\mathbf{x} \mid \omega_i)} > \dfrac{P(\omega_i)}{P(\omega_j)}, i = 1,2,\mathrm{L}\ ,m, i \neq j$, 则$\mathbf{x} \in \omega_j$;

(3) 若$\ln L(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_j) - \ln p(\mathbf{x} \mid \omega_i) > \ln \dfrac{P(\omega_i)}{P(\omega_j)}, i = 1,2,\cdots,m, i \neq j$

则$\mathbf{x} \in \omega_j$;

其中， $L(\boldsymbol{x})$称为似然比(likelihood ratio)， $\ln L(\boldsymbol{x})$称为对数似然比, $P(\omega_i)/P(\omega_j)$称为似然比阈值。

例4.1：假设在某个局部地区的细胞识别中, 第一类表示正常, 第二类表示异常, 两类的先验概率分别为: 正常 $P(\omega 1)=0.9$, 异常$P(\omega 2)=0.1$。 现有一个待识别样本细胞, 其观察值为$\mathbf{x}$, 从类条件概率密度函数曲线$p(\mathbf{x}|\omega \mathbf{i})$上可查得: $p(\mathbf{x}|\omega 1)=0.2$, $p(\mathbf{x}|\omega 2)=0.4$, 试判断该细胞是否正常。

解：该细胞属于正常细胞还是异常细胞，先计算后验概率：计算

$$p(\mathbf{x}|\omega 1)P(\omega 1)=0.2\times 0.9=0.18$$

$$p(\mathbf{x}|\omega 2)P(\omega 2)=0.4\times 0.1=0.04$$

$$p(\mathbf{x}|\omega 1)P(\omega 1)>p(\mathbf{x}|\omega 2)P(\omega 2)$$

根据Bayes判决准则将该细胞判为第一类ω1, 即为正常细胞。

(*：了解)

最大后验概率判决准则使决策的错误率最小？

最大后验概率判决准则的一个优良性质就是使平均错误概率达到最小。因此，最大后验概率判决准则又称为最小错误概率判决准则。

这里以二分类情况为例进行分析。此时，$m=2$，任意一个判决准则对应于特征空间$\mathbf{R}^d$的一个划分：$R=R1 \cup R2$，$R1 \cap R2=\Phi$。为直观起见，假设$\mathbf{x}$只有一个特征，n=1。错误分类有两种情况：①若$\mathbf{x}$原属于$\omega 1$类，却落入$R2$，称为第1类错误；②若$\mathbf{x}$原属于$\omega 2$类，却落入$R1$，称为第2类错误。

第1类错误概率$P1(e)$为：

$$P1(e) = P(\mathbf{x} \in R_2 \mid \omega_1) = \int_{R_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x}$$

第2类错误概率$P2(e)$为：

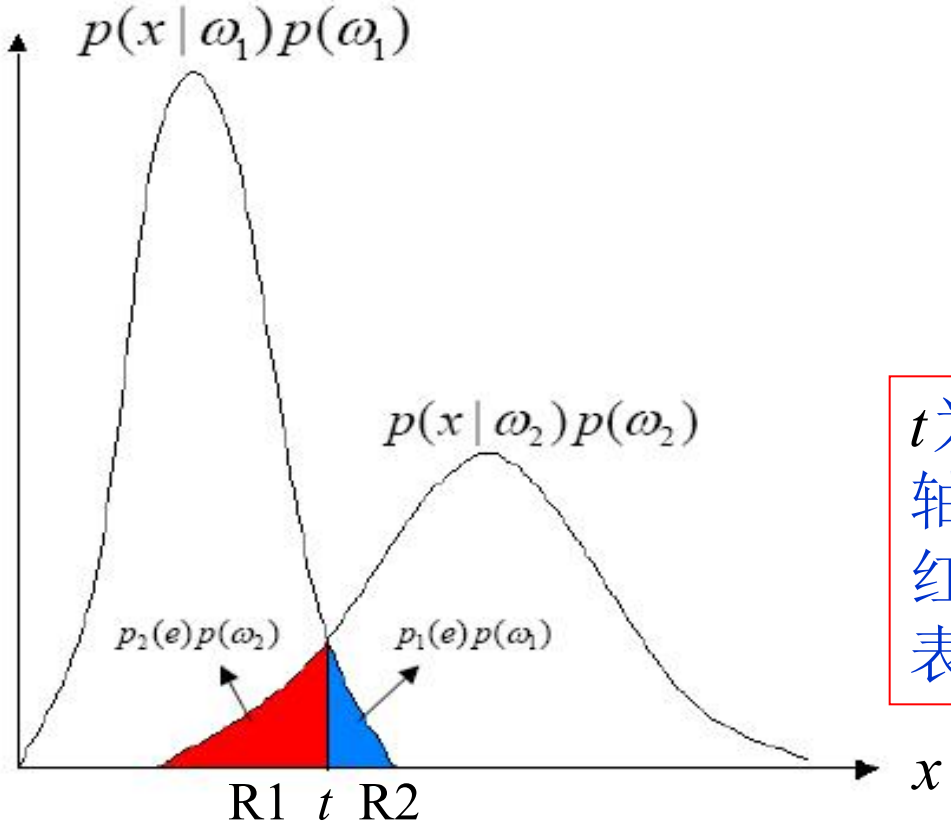
$$P2(e) = P(\mathbf{x} \in R_1 \mid \omega_2) = \int_{R_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}$$

因此，平均错误概率P(e)为：

$$P(e) = P_1(e)P(\omega_1) + P_2(e)P(\omega_2)$$

$$= P(\mathbf{x} \in R_2 \mid \omega_1)P(\omega_1) + P(\mathbf{x} \in R_1 \mid \omega_2)P(\omega_2)$$

$$= P(\omega_1)\int_{R_2} p(\mathbf{x} \mid \omega_1)d\mathbf{x} + P(\omega_2)\int_{R_1} p(\mathbf{x} \mid \omega_2)d\mathbf{x} \qquad (1)$$



平均错误概率计算示意图

t为两类的分界点，将x轴分成两个区域R1和R2。红色和蓝色区域的面积表示平均错误率。

其中：$P_1(e)=\int_{R_2} p(\mathbf{x}\mid\omega_1)d\mathbf{x}, P_2(e)=\int_{R_1} p(\mathbf{x}\mid\omega_2)d\mathbf{x}$

考虑到

$$\int_{R_2} p(\mathbf{x}\mid\omega_1)d\mathbf{x}=1-\int_{R_1} p(\mathbf{x}\mid\omega_1)d\mathbf{x}$$

$$P(e)=P(\omega_1)\int_{R_2} p(\mathbf{x}\mid\omega_1)d\mathbf{x}+P(\omega_2)\int_{R_1} p(\mathbf{x}\mid\omega_2)d\mathbf{x}$$

$$=P(\omega_1)\{1-\int_{R_1} p(\mathbf{x}\mid\omega_1)d\mathbf{x}\}+P(\omega_2)\int_{R_1} p(\mathbf{x}\mid\omega_2)d\mathbf{x}$$

$$=P(\omega_1)+\int_{R_1}\{P(\omega_2)p(\mathbf{x}\mid\omega_2)-P(\omega_1)p(\mathbf{x}\mid\omega_1)\}d\mathbf{x}$$

若要使$P(e)$达到最小，则$\boldsymbol{x}\in\ \omega_1$的决策区域$R1$必须满足：

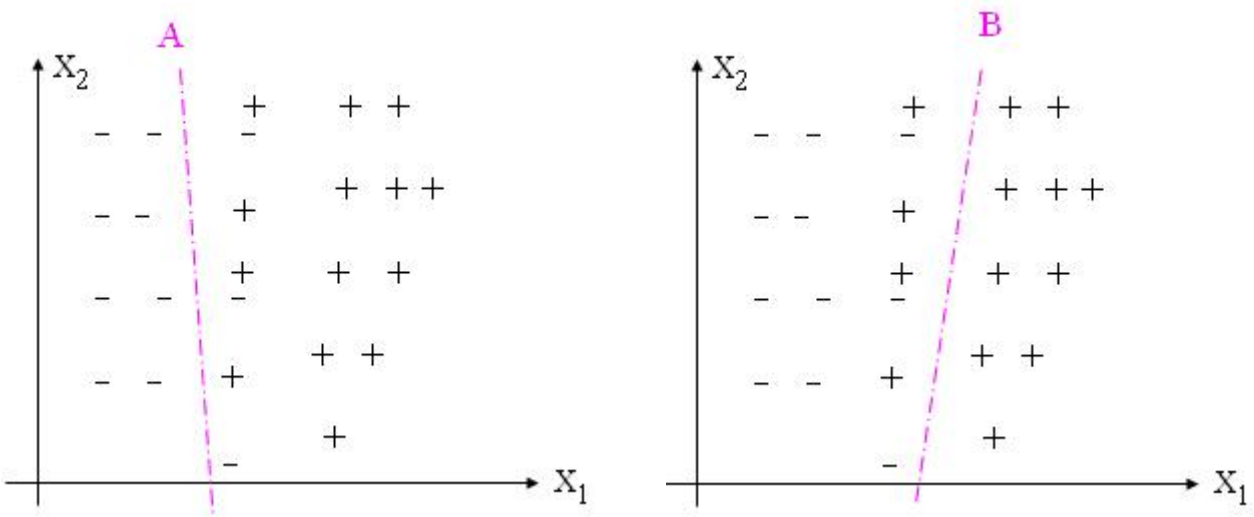
$$R_1=\{\mathbf{x}\,|\,P(\omega_2)p(\mathbf{x}\,|\,\omega_2)-P(\omega_1)p(\mathbf{x}\,|\,\omega_1)<0\}$$

即：$$R_1=\left\{\mathbf{x}\left|\frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)}>\frac{P(\omega_2)}{P(\omega_1)}\right.\right\}$$

上式与最大后验概率判决准则中$x\in\omega_1$的决策区域是一致的，也就是说，最大后验概率判决使平均错误概率达到最小。

4.4 最小风险Bayes决策

上节探讨了使错误概率最小的Bayes决策规则。然而，当接触到实际问题时，可能发现使错误率最小并非是普遍适用的最佳选择。见下图。



(a)基于最小错误　　(b)基于最小风险(扩大错误率，减小损失)

图　基于最小错误分类和基于最小风险分类的比较

图中，直线B的划分把正常药品误判为异常药品，这样扩大了总错误率，会给企业带来一定的损失；直线A的划分将异常药品误判为正常药品，虽然使错误分类最小，但会使病人因失去正确的治疗而遭受极大的损失。可见使错误率最小并不一定是最佳选择。

实际应用时，从根据不同性质的错误会引起不同程度的损失考虑出发，宁可扩大一些总的错误率，但也要使总的损失减少。这时图中的直线B的划分最为实用。这会引进一个与损失有关联的概念-风险(risk)。在做决策时，要考虑所承担的风险。基于最小风险的Bayes决策规则正是为了体现这一点而产生的。

- 若要判断某颗药品是正常(ω_1)还是异常(ω_2)，在判断中可能出现如下情况：
  - 第1种，判对(正常药品→正常药品)λ_{11}；
  - 第2种，判错(正常药品→异常药品)λ_{21}；
  - 第3种，判对(异常药品→异常药品)λ_{22}；
  - 第4种，判错(异常药品→正常药品)λ_{12}。
- 在判断时，除了能做出“是”ω_i类或“不是”ω_i类的动作以外，还可以做出“拒识”的动作。

对于两类分类问题，最小风险Bayes决策的基本思想是：对于模式x，如果将其决策为模式类ω1的风险大于决策为模式类ω2的风险，则决策模式x属于类ω2；反之，决策模式x属于模式类ω1。

为了更好地研究最小风险Bayes分类器，下面先说明几个相关概念。

几个概念

- 设模式(样本)**x**来自类ωi, 可能被判为$\omega 1, \omega 2, \ldots, \omega M$ 中的任何一种, 若允许拒绝判决, 可将拒绝类看成是独立的一类, 记为第$m+1$类, 即ω_{m+1}。
- 行动(action, 或决策)αi：表示把模式**x**判决为ωi类的一次动作 (决策)。

  不同的动作对应于特征空间的不同决策区域Rj, $j \in \{1, 2, \ldots, M\}$。 若$x \in Rj$, 则判决$\mathbf{x} \in \omega j (j=1, 2, \ldots, M)$。这里未考虑拒识情况。
- 损失函数$\lambda_{ii}=\lambda(\alpha_i, \omega_i)$表示模式**x**本来属于$\omega_i$类而错判为$\omega_i$所受损失。因为这是正确判决，故损失最小。
- 损失函数$\lambda_{ij}=\lambda(\alpha_i, \omega_j)$表示模式**x**本来属于$\omega_j$类错判为$\omega_i$所受损失。因为这是错误判决，故损失最大。
- 风险R(期望损失)：对未知模式**x**采取一个判决行动$\alpha(\mathbf{x})$所付出的代价 (损失)。

几个概念(Cont.)

- 条件风险(也叫条件期望损失： 模式向量为x时采取决策为α_i时的风险)：

$$R(\alpha_i \mid \mathbf{x}) = E\left[\lambda(\alpha_i, \omega_j)\right] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j \mid \mathbf{x}), i = 1, 2, ..., c(c \le M).$$

- 在整个特征空间中定义期望风险(expected risk)R(*: 了解)：

$$R = \int_{R^d} R(\alpha(\mathbf{x}) \mid \mathbf{x}) P(\mathbf{x}) d\mathbf{x}$$

$$= \sum_{j}^{M} \int_{R_j} R(\alpha_j \mid \mathbf{x}) P(\mathbf{x}) d\mathbf{x}, \text{(平均风险)}$$

P(x)为样本向量x在R^d空间中的概率密度

- 条件风险(conditional risk)$R(a_i|\mathrm{x})$只反映对某x取值的决策行动α_i所带来的风险。
- 期望风险R则反映在整个特征空间不同的x取值a(x) {决策可看成是随机向量x的函数，记为a(x)}的决策行动所带来的平均风险。(*: 了解)

几个概念(Cont.)

- 在实际应用时, 可以将损失函数λ_{ij}写成如下矩阵形式：

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} & \mathrm{L} & \lambda_{1M} \\ \lambda_{21} & \lambda_{22} & \mathrm{L} & \lambda_{2M} \\ & \mathrm{M} & & \\ \lambda_{c1} & \lambda_{c2M} & \mathrm{L} & \lambda_{cM} \end{bmatrix} (c \le M)$$

称之为损失矩阵(Loss matrix)。

基于最小风险的**Bayes**决策：

决策带来的损失的平均值—风险最小。

最小风险Bayes决策规则：

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

损失函数根据实际问题和经验确定. 若损失函数取:

$$\lambda(\alpha_i, \omega_j)=\begin{cases}0 & (i=j)\\ 1 & (i\neq j)\end{cases} \qquad (i, j=1,2,\ldots, M)$$

则称这种损失函数为0−1损失函数. (zero-one loss function) 此时决策α_j 的条件风险为

$$R(\alpha_j \mid \mathbf{x})=\sum_{i=1}^{M}\lambda\left(\alpha_j,\omega_i\right)P\left(\omega_i \mid \mathbf{x}\right)=\sum_{i\neq j}P(\omega_i \mid \mathbf{x})=1-P(\omega_j \mid \mathbf{x}) \quad (1)$$

从上式可以看出, $R(\alpha_j \mid \mathbf{x})$最小实际上对应于$P(\omega_j \mid \mathbf{x})$最大.

因此, 当取0-1损失函数时, 最小风险贝叶斯判决准则等价于最大后验概率判决准则[公式(1)]。即最大后验概率判决准则是最小风险Bayes判决准则的特例。

对于两分类问题, 条件风险为:

$$R(\alpha_1 \mid \mathbf{x}) = \lambda(\alpha_1, \omega_1) P(\omega_1 \mid \mathbf{x}) + \lambda(\alpha_1, \omega_2) P(\omega_2 \mid \mathbf{x}) \quad (2)$$

$$R(\alpha_2 \mid \mathbf{x}) = \lambda(\alpha_2, \omega_1) P(\omega_1 \mid \mathbf{x}) + \lambda(\alpha_2, \omega_2) P(\omega_2 \mid \mathbf{x}) \quad (3)$$

按最小风险的*Bayes*分类器, 有:

$$[\lambda(\alpha_2, \omega_1) - \lambda(\alpha_1, \omega_1)] P(\omega_1 \mid \mathbf{x})$$

$$> [\lambda(\alpha_1, \omega_2) - \lambda(\alpha_2, \omega_2)] P(\omega_2 \mid \mathbf{x}) \Rightarrow \mathbf{x} \in \omega_1 \quad (4a)$$

$$[\lambda(\alpha_2, \omega_1) - \lambda(\alpha_1, \omega_1)] P(\omega_1 \mid \mathbf{x})$$

$$< [\lambda(\alpha_1, \omega_2) - \lambda(\alpha_2, \omega_2)] P(\omega_2 \mid \mathbf{x}) \Rightarrow \mathbf{x} \in \omega_2 \quad (4b)$$

根据Bayes公式, 有:

$$[\lambda(\alpha_2,\omega_1)-\lambda(\alpha_1,\omega_1)]P(\mathbf{x}\mid\omega_1)P(\omega_1)$$

$$>[\lambda(\alpha_1,\omega_2)-\lambda(\alpha_2,\omega_2)]P(\mathbf{x}\mid\omega_2)P(\omega_2)\Rightarrow\mathbf{x}\in\omega_1 \qquad (5a)$$

$$[\lambda(\alpha_2,\omega_1)-\lambda(\alpha_1,\omega_1)]P(\mathbf{x}\mid\omega_1)P(\omega_1)$$

$$>[\lambda(\alpha_1,\omega_2)-\lambda(\alpha_2,\omega_2)]P(\mathbf{x}\mid\omega_2)P(\omega_2)\Rightarrow\mathbf{x}\in\omega_2 \qquad (5b)$$

$$L(\mathbf{x})=\frac{P(\mathbf{x}\mid\omega_1)}{P(\mathbf{x}\mid\omega_2)}\underset{<}{>}\frac{P(\omega_2)}{P(\omega_1)}\frac{\lambda(\alpha_1,\omega_2)-\lambda(\alpha_2,\omega_2)}{\lambda(\alpha_2,\omega_1)-\lambda(\alpha_1,\omega_1)}\Rightarrow\mathbf{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases} \qquad (6)$$

例 4.2： 在例4.1的基础上，增加条件$\lambda_{11}=0$，$\lambda_{12}=6$，$\lambda_{21}=1$，$\lambda_{22}=0$，试判断该细胞是否正常。

解： 若按最小风险的Bayes判决进行判断，先计算后验概率：

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

$$P(\omega_2/\mathbf{x})=1-P(\omega_1/\mathbf{x})=0.182$$

条件风险：$R(\alpha_1\mid\mathbf{x})=\sum_{i=1}^{2}\lambda_{1i}P(\omega_i\mid\mathbf{x})=1.092$

$$R(\alpha_2\mid\mathbf{x})=\sum_{i=1}^{2}\lambda_{2i}P(\omega_i\mid\mathbf{x})=0.818$$

Q $R(\alpha_1\mid x)>R(\alpha_2\mid x)\therefore x\in$异常细胞(第2类)，因此决策$\omega_1$类风险大。因$\lambda_{12}$=6较大，决策损失起决定作用。

小结：

1.贝叶斯定理(Bayes theorem)

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

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

$P(\omega_i|\mathrm{x})$ 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$P(\omega_i|\mathrm{x})$；$P(\mathrm{x}|\omega_i)$称为类条件概率密度(又称似然函数)； $P(\omega_i)$称为先验概率。

2.最小错误率Bayes分类/最大后验Bayes分类

3.最小风险Bayes分类

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

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

4.4 朴素贝叶斯分类器

1. 贝叶斯定理与最大后验概率Bayes决策(回顾)

假设有一个已标定的数据集$\{\mathbf{x}^{(i)}, y^{(i)}\}$,其中$y^{(i)} \in \{\omega_1, \omega_2, ..., \omega_M\}$,即数据集共有$M$个类别;$\mathbf{x}^{(i)} = (x_1, x_2, ..., x_n)^T$,即模式向量共有$n$个输入特征. 针对一个新的模式样本$\mathbf{x}$, 我们要预测$y$的值, 即对$\mathbf{x}$进行分类, 这是一典型的机器学习分类问题.

对于所求解的问题, 使用概率统计语言可表述为: 当观察到输入模式样本是$\mathbf{x}$时,其所属类别$y = \omega_k$的概率,其中$\omega_k \in \{\omega_1, \omega_2, ..., \omega_M\}$,根据*Bayes' Theorem*,有:

$$p(\omega_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_k) p(\omega_k)}{p(\mathbf{x})}$$

对于一个确定的数据集,ω_k, $p(\mathbf{x})$都是固定值, 因此:

$$p(\omega_k \mid \mathbf{x}) \propto p(\mathbf{x} \mid \omega_k) p(\omega_k)$$

其中,$\propto$ 表示成正比之意.

若采用最小错误率Bayes决策, 对于待识别的模式$\mathbf{x}$,

$$y = f(\mathbf{x}) = \arg\max_{\omega_j} \{p(\mathbf{x} \mid \omega_j) p(\omega_j)\} = \omega_k .$$

2. 朴素贝叶斯分类器(Naïve Bayes Classifier)

根据$\mathbf{x}$为n维特征向量和乘法规则(product rule),有:

$$p(\mathbf{x}\,|\,\omega_k)p(\omega_k) = p(\omega_k,\mathbf{x}) = p(\omega_k,x_1,\ldots,x_n)$$

根据条件概率的定义以及链式法则:

$$p(\omega_k,x_1,\ldots,x_n) = p(x_1,\ldots,x_n,\omega_k)$$

$$p(x_1,\ldots,x_n,\omega_k) = p(x_1\,|\,x_2\ldots,x_n,\omega_k)p(x_2\ldots,x_n,\omega_k)$$

$$= p(x_1\,|\,x_2\ldots,x_n,\omega_k)p(x_2\,|\,x_3,\ldots,x_n,\omega_k)\ldots p(x_n\,|\,\omega_k)p(\omega_k)$$

朴素贝叶斯决策中最前面的定语朴素"Naive"是指条件独立假设,即事件之间没有关联关系. 也就是输入特征要满足条件独立(conditionally independent)假设,即当$i \neq j$时,模式向量$\mathbf{x}$中,x_i和x_j是不相关的,通俗地说就是x_i是否发生和x_j没关系.

根据特征条件独立假设：$p(x_i \mid x_{i+1},...,x_n,\omega_k) = p(x_i \mid \omega_k)$

有：$p(x_1,...,x_n,\omega_k) = p(x_1 \mid \omega_k)p(x_2 \mid \omega_k)...p(x_n \mid \omega_k)p(\omega_k)$

$$=p(\omega_k)\prod_{i}^{n} p(x_i \mid \omega_k)$$

$p(\omega_k \mid \mathbf{x}) \propto p(\mathbf{x} \mid \omega_k)p(\omega_k)$

$= p(\omega_k, \mathbf{x}) = p(\omega_k, x_1,...,x_n)$

$$=p(\omega_k)\prod_{i}^{n} p(x_i \mid \omega_k)$$

Naive Bayes决策(Naive Bayes分类器)：对于待识别的模式$\mathbf{x}$,

$$y = f(\mathbf{x}) = \arg\max_{\omega_k}\{p(\omega_k)\prod_{i}^{n} p(x_i \mid \omega_k)\}.$$

其中,$\prod$ 是连乘符号,$p(x_i \mid \omega_k)$表示当类别为ω_k时特征x_i出现的概率,该值可由数据集统计出来.

3. 朴素贝叶斯分类器主要优缺点

优点：(1)对数据预测便捷、高效，特别对于多元分类任务；(2)当特征相互独立假设成立时，其预测能力优于逻辑回归等其他算法，适合增量式训练，尤其是数据量超出内存时，仍可成批地进行增量训练。

缺点：(1)朴素贝叶斯分类器假设条件在实际中有时很难成立，在特征个数比较多或者特征间相关性较大时，分类效果不好；(2)需要知道先验概率，且先验概率很多时候取决于假设，分类决策存在错误率；(3)对输入数据的表达形式很敏感。

4. 朴素贝叶斯分类器主要应用场景

(1)实时预测：朴素贝叶斯算法简捷，因此，它能用于实时预测。
(2)多分类预测：适用于目标变量为多类别的任务，能预测多类目标变量的概率。
(3)文本分类/垃圾邮件过滤/情感分析：主要用于文本分类的朴素贝叶斯分类器(由于多类问题和独立规则更好的结果)，与其他算法相比具有更高的成功率。 因此，它广泛用于垃圾邮件过滤(识别垃圾邮件)和情感分析(在社交媒体分析中，识别积极和消极的客户情绪)
(3)推荐系统：用朴素贝叶斯分类器和协作过滤共同构建推荐系统，该系统采用机器学习和数据挖掘技术来过滤看不见的信息并预测用户是否会喜欢给定的资源。一个简单的例子就是淘宝上的商品推荐。

5. 朴素贝叶斯分类器Python编程

在Python的Sklearn库中有朴素贝叶斯算法程序包，它包括三种朴素贝叶斯分类器：

(1)高斯GaussianNB：用于任意连续数据的分类，它假定特征服从正态分布。

(2)多项式MultinomialNB：用于计数数据(即每个特征代表某个对象的计数，比如一个单词在句子中的出现次数)的分类。主要用于文本数据分类。

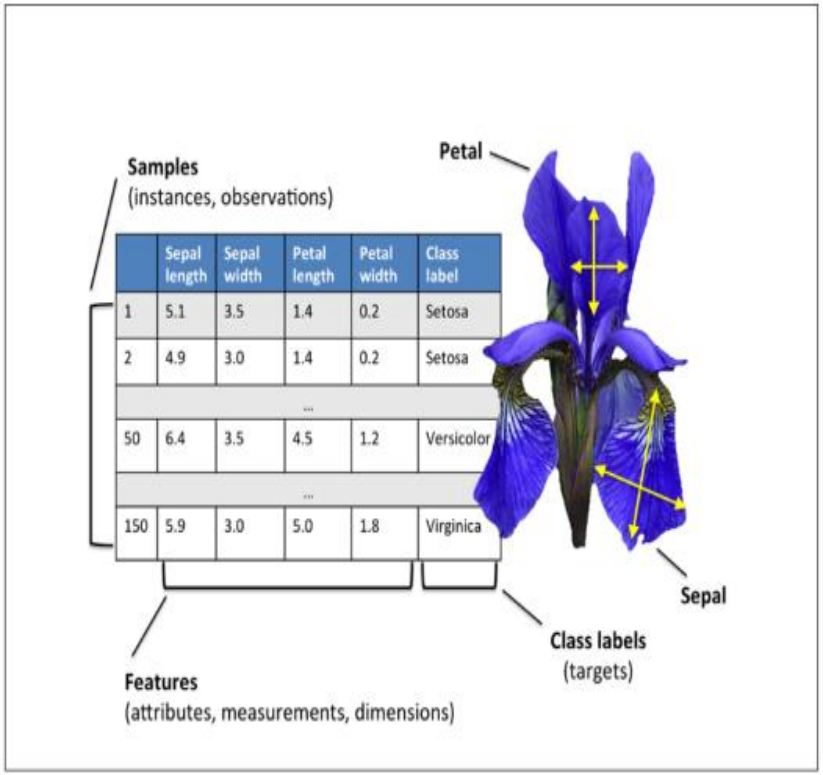
(3)伯努利BernoulliNB：用于二分类数据的0-1分类。主要用于文本数据分类。如在文本分类中的“1、0”分别是“词语出现在文档中”和“词语文档不出现在文档中”。

GaussianNB主要用于高维数据，而另外两种朴素贝叶斯模型则广泛用于稀疏计数数据，比如文本。 MultinomialNB的性能通常优于BernoulliNB，特别是在包含很多非零特征的数据集(即大型文档)上。

例：使用Sklearn库自带的Iris数据集，用朴素贝叶斯分类器对数据集进行分类。

Iris数据集介绍：

Iris(鸢尾花)数据集是常用于机器学习的一种多分类实验数据集，它是由著名统计学家Fisher于1936年收集整理。Iris数据集包含150个样本记录 (见右图)，分为3类，每类50个样本数据，每个样本数据包含4个特征/属性。可通过萼片(Sepal)长度、萼片宽度、花瓣(Petal)长度、花瓣宽度4个特征预测鸢尾花属于(Setosa山鸢尾，Versicolour杂色鸢尾，Virginica维吉尼亚鸢尾)三种鸢尾花卉中的哪一类。 Iris数据集中各样本的四个特征长度和宽度均以cm为单位。



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

In [2]:
```
from sklearn.datasets import load_iris   #导入iris数据集
iris = load_iris()   #加载iris数据集
print(iris.data)
```

```
[[5.1 3.5 1.4 0.2]
 [4.9 3.  1.4 0.2]
 [4.7 3.2 1.3 0.2]
 [4.6 3.1 1.5 0.2]
 [5.  3.6 1.4 0.2]
 [5.4 3.9 1.7 0.4]
 [4.6 3.4 1.4 0.3]
 [5.  3.4 1.5 0.2]
 [4.4 2.9 1.4 0.2]
 [4.9 3.1 1.5 0.1]
 [5.4 3.7 1.5 0.2]
 [4.8 3.4 1.6 0.2]
 [4.8 3.  1.4 0.1]
 [4.3 3.  1.1 0.1]
 [5.8 4.  1.2 0.2]
 [5.7 4.4 1.5 0.4]
 [5.4 3.9 1.3 0.4]
 [5.1 3.5 1.4 0.3]
 [5.7 3.8 1.7 0.3]
 [5.1 3.8 1.5 0.2]
```

In [3]:
```
print(iris.target)            #输出真实标签
print(len(iris.target))       #150个样本，每个样本4个特征
print(iris.data.shape)
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]
150
(150, 4)
```

以下是鸢尾花三分类问题，采用朴素贝叶斯分类器的Python程序运行界面截图：

In [1]:

```
#Naive Bayes classifier
#鸢尾花的Naive Bayes三分类Python程序
#Filename: Bayes.ipynb
import numpy as np
from sklearn import naive_bayes
from sklearn import datasets
np.random.seed(1000)
iris=datasets.load_iris()
iris_x=iris.data
iris_y=iris.target
indices=np.random.permutation(len(iris_x))
#随机选取数据集中的140个样本作为训练集
iris_x_train=iris_x[indices[:-10]]
iris_y_train=iris_y[indices[:-10]]
#数据集剩下的10个样本作为测试集
iris_x_test=iris_x[indices[-10:]]
iris_y_test=iris_y[indices[-10:]]
#定义Naive Bayes分类器对象naive_bayes_clf
naive_bayes_clf = naive_bayes.GaussianNB()     #GaussianNB(MultinomialNB/BernoulliNB)
#调用该对象的训练方法
naive_bayes_clf.fit(iris_x_train,iris_y_train)
#调用该对象的测试方法
iris_y_pred=naive_bayes_clf.predict(iris_x_test)
print('测试数据集的正确标签为:',iris_y_test)
print('测试数据集的预测标签为:',iris_y_pred)
from sklearn.metrics import accuracy_score
testing_acc=accuracy_score(iris_y_test,iris_y_pred)*100
print('Bayes分类器测试准确率: {:.2f}%'.format(testing_acc))
```

```
测试数据集的正确标签为: [2 0 0 1 2 0 1 1 1 1]
测试数据集的预测标签为: [2 0 0 1 2 0 1 1 1 1]
Bayes分类器测试准确率: 100.00%
```

以下是鸢尾花三分类问题朴素贝叶斯分类器的Python程序清单：

```
#Naive Bayes classifier
#鸢尾花的Naive Bayes三分类Python程序
#Filename: Naive_Bayes.ipynb
import numpy as np
from sklearn import naive_bayes
from sklearn import datasets
np.random.seed(1000)
iris=datasets.load_iris()
iris_x=iris.data
iris_y=iris.target
indices=np.random.permutation(len(iris_x))
#随机选取数据集中的140个样本作为训练集
iris_x_train=iris_x[indices[:-10]]
iris_y_train=iris_y[indices[:-10]]
#数据集剩下的10个样本作为测试集
iris_x_test=iris_x[indices[-10:]]
iris_y_test=iris_y[indices[-10:]]
#定义Naive Bayes分类器对象naive_bayes_clf
naive_bayes_clf = naive_bayes.GaussianNB()    #GaussianNB(MultinomialNB/BernoulliNB)
#调用该对象的训练方法
naive_bayes_clf.fit(iris_x_train,iris_y_train)
#调用该对象的测试方法
iris_y_pred=naive_bayes_clf.predict(iris_x_test)
print('测试数据集的正确标签为:',iris_y_test)
print('测试数据集的预测标签为:',iris_y_pred)
from sklearn.metrics import accuracy_score
testing_acc=accuracy_score(iris_y_test,iris_y_pred)*100
print('朴素Bayes分类器测试准确率: {:.2f}%'.format(testing_acc))
```

例：朴素贝叶斯分类器Python编程。

训练样本集：

x_i	$(-3,7)^T$,	$(1,5)^T$,	$(1,2)^T$,	$(-2,0)^T$,	$(2,3)^T$,	$(-4,0)^T$,	$(-1,1)^T$,	$(1,1)^T$,	$(-2,2)^T$,	$(2,7)^T$,	$(-4,1)^T$,	$(-2,7)^T$
y_i	1,	1,	2,	2,	2,	1,	1,	2,	1,	1,	1,	2

试利用以上训练样本集编制Gaussian Naïve Bayes分类器Python程序，并对下面两个待识别模式进行预测分类：

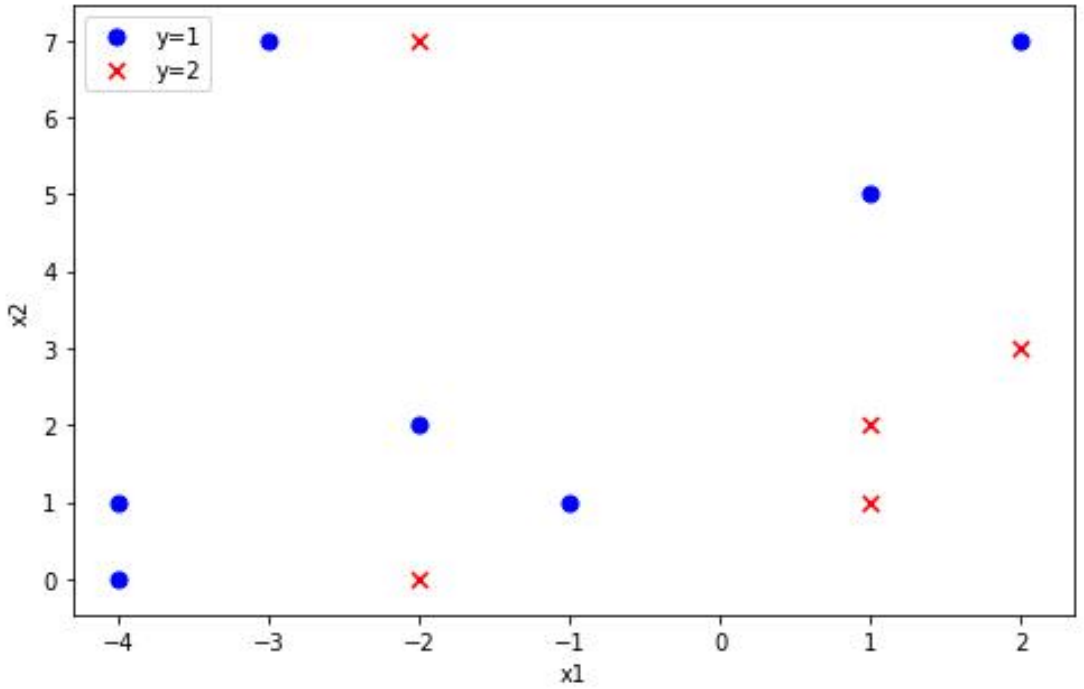
$(1,2)^T$, $(3,7)^T$

Python程序清单：

```
#Filename: GaussianNB_example.ipynb
import matplotlib.pyplot as plt
#Import Library of Gaussian Naive Bayes model
from sklearn.naive_bayes import GaussianNB
import numpy as np
#Assumed you have, X (predictor) and y (target) for training data set and X_test(predictor) of test_dataset
#设二维特征向量为x，对应类别标签为y
X=np.array([[-3,7],[1,5], [1,2], [-2,0], [2,3], [-4,0], [-1,1], [1,1], [-2,2],[2,7], [-4,1], [-2,7]])
y=np.array([1,1, 2, 2, 2, 1, 1, 2, 1, 1, 1, 2])
X_test=np.array([[1,2],[3,7]])
id1=np.where(y==1)
id2=np.where(y==2)
fig,ax =plt.subplots(figsize=(8,5))
ax.scatter(X[id1,0],X[id1,1],s=50, c='b', marker='o', label='y=1')
ax.scatter(X[id2,0],X[id2,1],s=50, c='r', marker='x', label='y=2')
ax.legend()
ax.set_xlabel('x1')
ax.set_ylabel('x2')
plt.show()
```

```
#Create a Gaussian Naive Bayes Classifier(采用高斯朴素贝叶斯分类器)
model=GaussianNB()
#Train the model using the training sets
model.fit(X,y)
#Predict Output
predicted= model.predict(X_test)
print(predicted)
```

输出结果：
[2, 1]



结论：采用Naïve Bayes Classifier，第1个待识别模式归为第2类，第2个待识别模式归为第1类。

```
# LogisticRegression Predict (法2：采用逻辑回归预测)
from sklearn.linear_model.logistic import LogisticRegression
classifier=LogisticRegression()
classifier.fit(X,y)
#Equation coefficient and Intercept
print('Coefficient: n', classifier.coef_)
print('Intercept: n', classifier.intercept_)
predictions=classifier.predict(X_test)
print(predictions)
```

输出结果：

```
Coefficient: n [[ 0.37237264 -
0.14496394]]
Intercept: n [0.27308522]
[2, 2]
```

结论：采用逻辑回归预测，第1个待识别模式归为第2类，第2个待识别模式归为第2类。

本例两种方法结果比较：由于模式样本是自己构造的，无法判断测试集预测结果的准确性。但是，我们能将模式向量 $[1,2]^T$、$[3,7]^T$放到上面**x**的分类散点图坐标系中进行观察，可以看出用朴素贝叶斯分类器预测相对准确一些。

思考题/练习题：

(1) Naïve Bayes Categorization Problem

Probability	positive	negative
P(C)	0.5	0.5
P(small | C)	0.4	0.4
P(medium | C)	0.1	0.2
P(large | C)	0.5	0.4
P(red | C)	0.9	0.3
P(blue | C)	0.05	0.3
P(green | C)	0.05	0.4
P(square | C)	0.05	0.4
P(triangle | C)	0.05	0.3
P(circle | C)	0.9	0.3

The pattern/instance to be recognized:
$\mathbf{x}$=(medium, red, circle)^T

Probability	positive	negative
P(C)	0.5	0.5
P(medium | C)	0.1	0.2
P(red | C)	0.9	0.3
P(circle | C)	0.9	0.3

The pattern/instance to be recognized:
x=(medium, red, circle)^T

Solution:

P(positive | **x**) = P(positive)*P(medium | positive)*P(red | positive)*P(circle | positive) / P(**x**)
0.5 * 0.1 * 0.9 * 0.9
= 0.0405 / P(**x**) = 0.0405 / 0.0495 = 0.8181

P(negative | **x**) = P(negative)*P(medium | negative)*P(red | negative)*P(circle | negative) / P(**x**)
0.5 * 0.2 * 0.3 * 0.3
= 0.009 / P(**x**) = 0.009 / 0.0495 = 0.1818

$\Rightarrow$ **x**ϵpositive

P(positive | **x**) + P(negative | **x**) = 0.0405 / P(**x**) + 0.009 / P(**x**) = 1

P(**x**) = 0.0405 + 0.009 = 0.0495

(2) Naïve Bayes Diagnosis Problem

C = {allergy, cold, well}

e_1 = sneeze; e_2 = cough; e_3 = fever

The pattern to be recognized: **x** = (sneeze, cough, ¬fever)^T

Probability	**Well**	**Cold**	**Allergy**
P(c_i)	0.9	0.05	0.05
P(sneeze|c_i)	0.1	0.9	0.9
P(cough|c_i)	0.1	0.8	0.7
P(fever|c_i)	0.01	0.7	0.4

Probability	Well	Cold	Allergy
$P(c_i)$	0.9	0.05	0.05
$P(\text{sneeze} \mid c_i)$	0.1	0.9	0.9
$P(\text{cough} \mid c_i)$	0.1	0.8	0.7
$P(\text{fever} \mid c_i)$	0.01	0.7	0.4

The pattern to be recognized:

$\mathbf{x}=(\text{sneeze, cough, } \neg\text{fever})^T$

Solution:

$P(\text{well} \mid \mathbf{x}) = (0.9)(0.1)(0.1)(0.99)/P(\mathbf{x})=0.0089/P(\mathbf{x})$

$P(\text{cold} \mid \mathbf{x}) = (0.05)(0.9)(0.8)(0.3)/P(\mathbf{x})=0.01/P(\mathbf{x})$

$P(\text{allergy} \mid \mathbf{x}) = (0.05)(0.9)(0.7)(0.6)/P(\mathbf{x})=0.019/P(\mathbf{x})$

Most probable category: allergy

$P(\mathbf{x}) = 0.0089 + 0.01 + 0.019 = 0.0379$

$P(\text{well} \mid \mathbf{x}) = 0.23$

$P(\text{cold} \mid \mathbf{x}) = 0.26$

$P(\text{allergy} \mid \mathbf{x}) = 0.50$

End of this lecture.

Thanks!