贝叶斯分类器

1 判别模型vs生成模型

判别模型和生成模型

监督学习方法又分

生成方法（Generative approach）和**判别方法**（Discriminative approach）

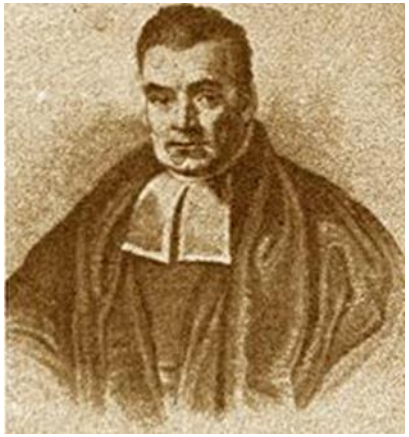
所学到的模型分别称为

生成模型（Generative Model）和**判别模型**（Discriminative Model)。

判别模型（Discriminative Model)	**生成模型（Generative Model)**
由数据直接学习决策函数Y=f(X)或者条件概率分布P(Y|X)作为预测的模型，即判别模型。基本思想是有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。 即：直接估计$P(Y\|X)$	由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布$P(Y\|X)$。具体来说，利用训练数据学习$P(X\|Y)$和$P(Y)$的估计，得到联合概率分布：$P(X,Y)=P(Y)P(X\|Y)$，再利用它进行分类。 即：估计$P(X\|Y)$ 然后推导$P(Y\|X)$
线性回归、逻辑回归、感知机、决策树、支持向量机……	朴素贝叶斯、HMM、深度信念网络(DBN)……

贝叶斯

- 贝叶斯(约1701-1761) Thomas Bayes，英国数学家。贝叶斯在数学方面主要研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计决策理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。他死后，理查德·普莱斯(Richard Price)于1763年将他的著作《机会问题的解法》(An essay towards solving a problem in the doctrine of chances)寄给了英国皇家学会，对于现代概率论和数理统计产生了重要的影响



贝叶斯

- 贝叶斯决策就是在不完全情报下，对部分未知的状态用主观概率估计，然后用贝叶斯公式对发生概率进行修正，最后再利用期望值和修正概率做出最优决策。
- 贝叶斯决策方法是统计模型决策中的一个基本方法，其基本思想是：

  1.已知类条件概率密度参数表达式和先验概率

  2.利用贝叶斯公式转换成后验概率

  3.根据后验概率大小进行决策分类

2 贝叶斯概率基础

- **先验概率**：根据历史资料或主观判断所确定的各种事件发生的概率。用$P(Y)$来代表在没有训练数据前假设Y拥有的初始概率。
- 先验概率可分为两类：
  - **客观先验概率：是指利用过去的历史资料计算得到的概率(如：在自然语言处理中，从语料库中统计词语的出现频率：客观先验概率)**；
  - **主观先验概率**：**是指在无历史资料或历史资料不全的时候，只能凭借人们的主观经验来判断取得的概率。**

2 贝叶斯概率基础

- **后验概率**：是指利用**贝叶斯公式，结合调查等方式**获取了新的附加信息，对先验概率修正后得到的更符合实际的概率。即根据已经发生的事件来分析得到的概率。以 $P(Y/X)$ 代表假设 X 成立的情下观察到 Y 数据的概率，因为它反映了在看到训练数据 X 后 Y 成立的置信度。
- **联合概率**：联合概率是指在多元的概率分布中多个随机变量分别满足各自条件的概率。X 与 Y 的联合概率表示为 $P(X,Y)$

2 贝叶斯概率基础

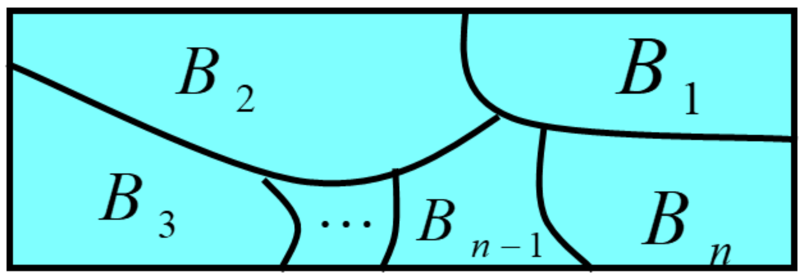
- **条件概率**：是指当**条件事件**发生后，该事件发生的概率。
- 条件概率公式:

$$P(X|Y) = P(X,Y)/P(Y)$$

$$P(Y|X) = P(X,Y)/P(X)$$

2 贝叶斯概率基础

- 全概率公式
- 设$B_1, B_2, \ldots, B_n$是两两互斥的事件，且$P(B_i)>0$，$i=1, 2, \ldots, n$, $B_1+B_2+\ldots+B_n=\Omega$。



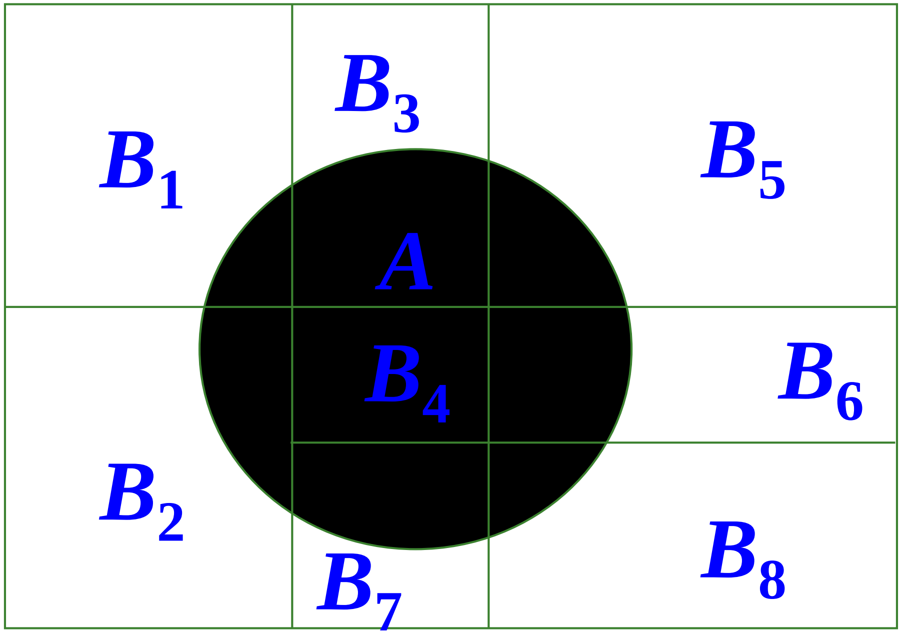
- 另有一事件$A = AB_1 + AB_2 + \ldots + AB_n$

$$P(A)=\sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

2 贝叶斯概率基础

- 全概率公式可看成是"由原因推结果"，即：每个原因对结果的发生有一定"作用"，结果发生的可能性与各种原因的"作用"大小有关。
- 全概率公式表达了它们之间的关系。

B_i是原因
A是结果



2 贝叶斯概率基础

- 贝叶斯公式(后验概率公式)
- 设先验概率为$P(B_i)$，调查所获的新附加信息为$P(A|B_i)$ ($i=1, 2, \dots, n$)，则贝叶斯公式计算的后验概率为：

$$P(B_i \mid A) = P(B_i)P(A|\ B_i) \Big/ \sum_{i=1}^{n} P(B_i)P(A|\ B_i)$$

  - 该公式于1763年由贝叶斯(Bayes)导出。
  - 该公式是在观察到事件A已发生的条件下，寻找导致A发生的每个原因的概率。

2 贝叶斯概率基础

- 贝叶斯公式(后验概率公式)
- 对于分类问题，假设有*N*种可能的类别标记，即 $y=\{c_1,c_2,\dots,c_N\}$根据贝叶斯公式，给定*x*是类别*c*的概率

类标记 c 相对于样本 $\mathbf{x}$ 的“类条件概率”（class-conditional probability），或称“似然”。

$$P(c\mid\mathbf{x})=\frac{\boxed{P(c)}\,\boxed{P(\mathbf{x}\mid c)}}{\boxed{P(\mathbf{x})}}$$

先验概率
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“证据”（evidence）因子，与类标记无关

2 贝叶斯概率基础

- 例：某电子设备厂所用的元件由三家元件厂提供，根据以往记录，这三个厂家的次品率分别为0.02，0.01和0.03，提供元件的份额分别为0.15，0.8和0.05，设这三家的产品在仓库是均匀混合的，且无区别的标志。
  - 问题1：在仓库中，随机抽取一个元件，求它是次品的概率；
  - 问题2：在仓库中，随机抽取一个元件，若已知它是次品，则该次品来自三家供货商的概率分别是多少？

2 贝叶斯概率基础

- 【解】设A表示“取到的元件是次品”，B_i表示“取到的元件是由第i个厂家生产的”，则

  $P(B_1)=0.15$，$P(B_2)=0.8$，$P(B_3)=0.05$

- 问题1：由**全概率公式**可得，

$$P(A) = P(B_1)*P(A|B_1) + P(B_2)*P(A|B_2)$$
$$+ P(B_3)*P(A|B_3)$$
$$= 0.15*0.02+0.8*0.01+0.05*0.03$$
$$= 0.0125$$

2 贝叶斯概率基础

- 【解】设A表示“取到的元件是次品”，B_i表示“取到的元件是由第i个厂家生产的”，则

  $P(B_1)=0.15$，$P(B_2)=0.8$，$P(B_3)=0.05$

- 问题2：由**贝叶斯公式**可得，

$$P(B_1|A) = P(B_1)*P(A|B_1)/P(A)$$
$$= 0.15*0.02/0.0125 = 0.24$$
$$P(B_2|A) = P(B_2)*P(A|B_2)/P(A)$$
$$= 0.8*0.01/0.0125 = \mathbf{0.64}$$
$$P(B_3|A) = P(B_3)*P(A|B_3)/ P(A)$$
$$= 0.05*0.03/0.0125 = 0.12$$

贝叶斯分类

- 数据挖掘中以贝叶斯定理为基础的分类方法，代表性技术有朴素贝叶斯分类和贝叶斯信念网络两种。
- 朴素贝叶斯分类假定一个属性值对给定类的影响独立于其他属性的值，即在属性间不存在依赖关系，也因此称为“朴素的”。
- 贝叶斯信念网络也可以用于分类，它是图形模型。它优于朴素贝叶斯，它能够处理属性子集间有依赖关系的分类。
- 采用监督式的学习方式

3 朴素贝叶斯分类

- 贝叶斯决策（Bayesian decision theory）是在概率框架下实施决策的基本方法。
  - 对于分类问题，在所有相关概率都已知的理想情形下，贝叶斯决策考虑如何基于这些概率和误判损失来选择最优的类别标记
- 假设有*N*种可能的类别标记，即 $y=\{c_1,c_2,\ldots,c_N\}$ ，λ_{ij} 是将一个真实标记为c_j的样本误分类为c_i所产生的损失。基于后验概率 $P\{c_i \mid \mathbf{x}\}$ 可获得将样本 X分类为 c_i 所产生的期望损失(expected loss)，即在样本上的“条件风险”(conditional risk)

$$R(c_i \mid \mathbf{x})=\sum_{j=1}^{N}\lambda_{ij}P(c_j \mid \mathbf{x})$$

3 朴素贝叶斯分类

- 贝叶斯决策的任务是寻找一个判定准则 $h: X \mapsto Y$ 以最小化总体风险

$$R(h) = \mathbf{E}_x \left[R(h(\mathbf{x}) \mid \mathbf{x}) \right]$$

- 显然，对每个样本 X ,若*h*能最小化条件风险 $R(h(\mathbf{x}) \mid \mathbf{x})$ 则总体风险 $R(h)$ 也将被最小化。
- 这就产生了贝叶斯判定准则(Bayes decision rule)：为最小化总体风险，只需在每个样本上选择那个能使条件风险 $R(c \mid \mathbf{x})$ 最小的类别标记，即

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}}\, R(c \mid x)$$

此时，被称为贝叶斯最优分类器(Bayes optimal classifier)，与之对应的总体风险 $R(h^*)$ 称为贝叶斯风险 (Bayes risk)

3 朴素贝叶斯分类

- 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij} 可写为

$$\lambda_{i,j} \begin{cases} 0, & \text{if } \ i = j; \\ 1, & \text{otherwise}, \end{cases}$$

- 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x})$$

- 于是，最小化分类错误率的贝叶斯最优分类器为

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} \, P(c \mid x)$$

- 即对每个样本 $\mathbf{x}$，选择能使后验概率 $P(c \mid \mathbf{x})$ 最大的类别标记。

3 朴素贝叶斯分类

- 不难看出，使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c \mid \mathbf{x})$。
- 然而，在现实中通常难以直接获得。机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率$P(c \mid \mathbf{x})$。
- 主要有两种策略：
  - 判别式模型：直接建模 $P(c \mid \mathbf{x})$, 如决策树，BP神经网络，支持向量机
  - 生成模式：先对联合概率分布 $P(\mathbf{x}, c)$建模，然后由条件概率公式生成

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})}$$

3 朴素贝叶斯分类

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

类标记 c 相对于样本 $\mathbf{x}$ 的“类条件概率”（class-conditional probability），或称“似然”。

$$P(c \mid \mathbf{x}) = \frac{\boxed{P(c)}\boxed{P(\mathbf{x} \mid c)}}{\boxed{P(\mathbf{x})}}$$

先验概率
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“证据”（evidence）因子，与类标记无关

3 朴素贝叶斯分类

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$可写成

类标记 c 相对于样本 $\mathbf{x}$ 的“类条件概率”（class-conditional probability），或称“似然”。

$$P(c \mid \mathbf{x}) = \frac{\boxed{P(c)}\boxed{P(\mathbf{x} \mid c)}}{\boxed{P(\mathbf{x})}}$$

先验概率
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“证据”（evidence）因子，与类标记无关

- 主要困难：类条件概率 $P(\mathbf{x} \mid c)$是所有属性上的联合概率难以从有限的训练样本估计获得
- 估计类条件概率的常用策略：先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布参数估计。

3 朴素贝叶斯分类

- 概率模型的训练过程就是参数估计过程，统计学界的两个学派提供了不同的方案
- 频率主义学派 (frequentist)认为参数虽然未知，但却存在客观值，因此可通过优化似然函数等准则来确定参数值
- 贝叶斯学派 (Bayesian)认为参数是未观察到的随机变量、其本身也可由分布，因此可假定参数服从一个先验分布，然后基于观测到的数据计算参数的后验分布。
- 贝叶斯学派很古老，但是从诞生到一百年前一直不是主流。主流是频率学派。频率学派的权威皮尔逊和费歇尔都对贝叶斯学派不屑一顾，但是贝叶斯学派硬是凭借在现代特定领域的出色应用表现为自己赢得了半壁江山。

贝叶斯学派

- 贝叶斯学派的思想可以概括为 先验概率+数据=后验概率。也就是说我们在实际问题中需要得到的后验概率，可以通过先验概率和数据一起综合得到。数据大家好理解，被频率学派攻击的是先验概率，一般来说先验概率就是我们对于数据所在领域的历史经验，但是这个经验常常难以量化或者模型化，于是贝叶斯学派大胆的假设先验分布的模型，比如正态分布，beta分布等。这个假设一般没有特定的依据，因此一直被频率学派认为很荒谬。虽然难以从严密的数学逻辑里推出贝叶斯学派的逻辑，但是在很多实际应用中，贝叶斯理论很好用，比如垃圾邮件分类，文本分类。

3 朴素贝叶斯分类

- 估计类条件概率 $P(\mathbf{x} \mid c)$ 的常用策略：先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布参数估计。

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})}$$

- 可以使用极大似然估计法来进行参数估计。需注意的是，这种参数化的方法虽能使类条件概率估计变得相对简单，但估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布。
- 朴素贝叶斯分类器采用了“属性条件独立性假设” (attribute conditional independence assumption)：每个属性独立地对分类结果发生影响。

3 朴素贝叶斯分类

- 朴素贝叶斯分类器采用了“属性条件独立性假设”：每个属性独立地对分类结果发生影响。基于属性条件独立性假设，

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

其中 d 为属性数目，x_i 为 x在第i个属性上的取值

- 由于对所有类别来说$P(x)$相同，因此贝叶斯判定准则有

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}}\, P(c \mid x)$$

$$\Updownarrow$$

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}}\, P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

3 朴素贝叶斯分类

- 如何训练？：朴素贝叶斯分类器的训练器的训练过程就是基于训练集*D*估计类先验概率$P(c)$并为每个属性估计条件概率$P(x_i \mid c)$
  - 令 D_c 表示训练集*D*中第*c*类样本组合的集合，若有充足的独立同分布样本，则可容易地估计出类先验概率
$$P(c) = \frac{|D_c|}{D}$$
  - 对离散属性而言，令 D_{c,x_i} 表示 D_c 中在第*i*个属性上取值为 x_i 的样本组成的集合，则条件概率 $P(x_i \mid c)$ 可估计为
$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{D}$$

3 朴素贝叶斯分类

- 如何训练？：朴素贝叶斯分类器的训练器的训练过程就是基于训练集*D*估计类先验概率$P(c)$并为每个属性估计条件概率$P(x_i \mid c)$
  - 对连续属性而言可考虑概率密度函数，通常假定该属性服从高斯分布 $p(x_i \mid c) \sim N(\mu_{c,i}, \sigma^2_{c,i})$ ，其中 $\mu_{c,i}$ 和 $\sigma^2_{c,i}$分别是第*c*类样本在第 *i*个属性上取值的均值和方差，则有

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}{}^2}\right)$$

3 朴素贝叶斯分类

- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题。
- 为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“拉普拉斯修正”（Laplacian correction）
- 令*N*表示训练集*D*中可能的类别数，N_i 表示第*i*个属性可能的取值数，则上式分别修正为

$$\hat{P}(c)=\frac{|D_c|+1}{|D|+N} \qquad \hat{P}(x_i \mid c)=\frac{|D_{c,x_i}|+1}{|D|+N_i}$$

  - 加入拉普拉斯平滑之后，避免了出现概率为0 的情况，又保证了每个值都在0到1 的范围内，又保证了最终和为1 的概率性质

3 朴素贝叶斯分类

假设我们正在构建一个分类器，该分类器说明文本是否与运动(Sports)有关。我们的训练数据有5句话：

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

我们想要计算句子“A very close game”是 Sports 的概率以及它不是 Sports 的概率。

即$P(\text{Sports} \mid \text{a very close game})$这个句子的类别是Sports的概率

3 朴素贝叶斯分类

特征：单词的频率

已知贝叶斯定理$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$，则：

$$P(\text{ Sports } | \text{ a very close game })$$

$$=\frac{P(\text{ a very close game } | \text{ Sports }) \times P(\text{ Sports })}{P(\text{ a very close game })}$$

由于我们只是试图找出哪个类别有更大的概率，可以舍弃除数，只是比较

$P(\text{ a very close game } | \text{ Sports }) \times P(\text{ Sports })$ 和

$P(\text{ a very close game } | \text{ Not Sports }) \times P(\text{ Not Sports })$

3 朴素贝叶斯分类

我们假设一个句子中的每个单词都与其他单词无关。

$$P(\text{ a very close game }) = P(a) \times P(\text{ very }) \times P(\text{ close }) \times P(\text{ game })$$

$$P(\ a\ very\ close\ game|Sports\ ) = P(a|\ Sports\ ) \times P(\ very\ |\ Sports) \times P(\ close\ |\ Sports\ ) \times P(\ game\ |\ Sports\ )$$

3 朴素贝叶斯分类

计算每个类别的先验概率：

对于训练集中的给定句子，

$P(\text{Sports })$的概率为⅗。

$P(\text{Not Sports })$是⅖。

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

然后，在计算$P(game|Sports)$就是“game”有多少次出现在Sports的样本，然后除以sports为标签的文本的单词总数（3+3+5=11）。

因此，$P(game|Sports)=\frac{2}{11}$。

“close”不会出现在任何sports样本中！那就是说$P(close|Sports)=0$。

3 朴素贝叶斯分类

通过使用一种称为**拉普拉斯平滑**的方法：我们为每个计数加1，因此它永远不会为零。为了平衡这一点，我们将可能单词的数量添加到除数中，因此计算结果永远不会大于1。

14个单词

在这里的情况下，可能单词是['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match']。

由于可能的单词数是14，因此应用平滑处理可以得到

$$P(\text{game} \mid \text{sports}) = \frac{2+1}{11+14}$$

3 朴素贝叶斯分类

通过使用一种称为**拉普拉斯平滑**的方法：我们为每个计数加1，因此它永远不会为零。为了平衡这一点，我们将可能单词的数量添加到除数中，因此计算结果永远不会大于1。

14个单词

在这里的情况下，可能单词是['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match']。

由于可能的单词数是14，因此应用平滑处理可以得到

$$P(\text{ game} \mid \text{sports }) = \frac{2+1}{11+14}$$

3 朴素贝叶斯分类

Word	P (word | Sports)	P (word | Not Sports)
a	$(2+1)\div(11+14)$	$(1+1)\div(9+14)$
very	$(1+1)\div(11+14)$	$(0+1)\div(9+14)$
close	$(0+1)\div(11+14)$	$(1+1)\div(9+14)$
game	$(2+1)\div(11+14)$	$(0+1)\div(9+14)$

$$P(a|\text{ Sports }) \times P(\text{ very }|\text{ Sports }) \times P(\text{ close }|\text{ Sports }) \times P(\text{ game }|\text{ Sports }) \times P(\text{ Sports })$$
$$= 2.76 \times 10^{-5} = 0.0000276$$

$$P(a|\text{ Not Sports }) \times P(\text{ very }|\text{ Not Sports }) \times P(\text{ close }|\text{ Not Sports })$$
$$\times P(\text{ game }|\text{ Not Sports }) \times P(\text{ Not Sports})$$
$$= 0.572 \times 10^{-5} = 0.00000572$$

由于0.0000276大于0.00000572，我们的分类器预测“A very close game”是Sport类。

3 朴素贝叶斯算法流程

- **输入**：训练集为 m 个样本 n 个维度 $T=(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)$,共有 K 个特征输出类别，分别为 $y\in\{c_1,c_2,\ldots,c_K\}$。
- **输出**：实例 $x(test)$ 的分类。
- **算法流程如下：**

1.首先计算 Y 的 K 个先验概率

$$P(Y=c_k)$$

2.然后计算条件概率分布：

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\ldots,X^{(n)}=x^{(n)}|Y=c_k)$$

由于上式的参数是指数级别，无法计算。所以根据特征条件独立假设，可以化简为下式。

$$P(X=x|Y=c_k)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

3.根据贝叶斯原理，计算后验概率：

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum\limits_k P(X=x|Y=c_k)P(Y=c_k)}$$

其中$\sum\limits_k P(X=x|Y=c_k)P(Y=c_k)\Leftrightarrow P(X=x)$

带入 $P(X=x|Y=c_k)=\prod\limits_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$ 得到

$$P(Y=c_k|X=x)=\frac{\prod\limits_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)P(Y=c_k)}{\sum\limits_k\prod\limits_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)P(Y=c_k)}$$

由于所有c_k的$P(X=x)$都是相同的，这样我们可以把输出结果化简成，上式再变为如下：

$$P(Y=c_k|X=x)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)P(Y=c_k)$$

4.计算 $X(test)$ 的类别

$$y_{(test)}=arg\ \max_{c_k}\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}_{(test)}|Y=c_k)P(Y=c_k)$$

3 朴素贝叶斯分类总结

- 优点：
  - 计算速度最快的演算法，有稳定的分类效率；
  - 规则清楚易懂，对缺失数据不太敏感；
  - 对小规模的数据表现很好，能个处理多分类任务，适合增量式训练；
- 缺点：
  - 如果样本数据分布不能很好的代表样本空间分布，那先验概率容易测不准；
  - 假设变量间为独立互不影响（实际中很难满足，是个非常严格的条件），因此使用时需要谨慎分析变量间的相关性。