

第9章 朴素贝叶斯

贝叶斯



Thomas Bayes

1702 - 1761

贝叶斯流派

- 贝叶斯分析是整个机器学习的基础框架。
- 概率：一件事发生的频率，这叫做客观概率。
- ◆ 贝叶斯：**概率是我们个人的一个主观概念，表明我们对某个事物发生的相信程度。**
- ◆ 拉普拉斯：概率论不过是简化为计算的常识
(Probability theory is nothing but common sense reduced to calculation) 。
- ◆ 这是贝叶斯流派的核心，它解决的是来自外部的信息与我们大脑内信念的交互关系。

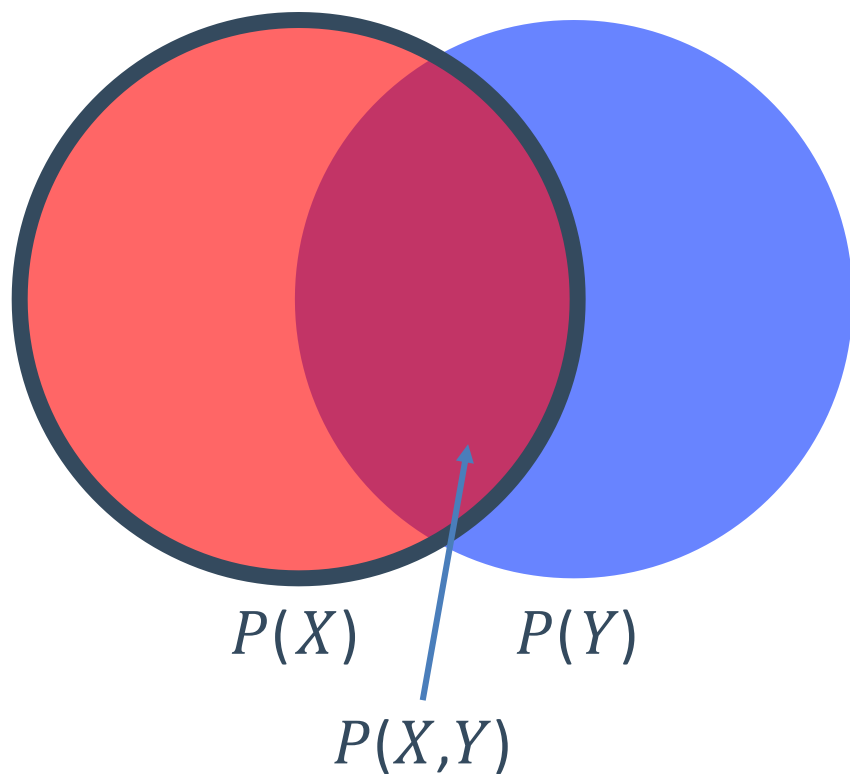
贝叶斯流派

- 两种对于概率的解读区别了频率流派和贝叶斯流派。
- 不理解主观概率就无法理解贝叶斯定律的核心思想。
- 贝叶斯分析的思路对于由证据的积累来推测一个事物发生的概率具有重大作用，它告诉我们当我们要预测一个事物，我们需要的是首先根据已有的经验和知识推断一个先验概率，然后在新证据不断积累的情况下调整这个概率。整个通过积累证据来得到一个事件发生概率的过程我们称为贝叶斯分析。

贝叶斯分类算法

- 贝叶斯分类算法是统计学的一种分类方法，它是一类利用概率统计知识进行分类的算法。
- 在许多场合，朴素贝叶斯(Naïve Bayes, NB)分类算法可以与决策树和神经网络分类算法相媲美，该算法能运用到大型数据库中，而且方法简单、分类准确率高、速度快。
- 由于贝叶斯定理假设一个属性值对给定类的影响独立于其它属性的值，而此假设在实际情况中经常是不成立的，因此其分类准确率可能会下降。
- 衍生出许多降低独立性假设的贝叶斯分类算法，如半朴素贝叶斯、贝叶斯网络等。

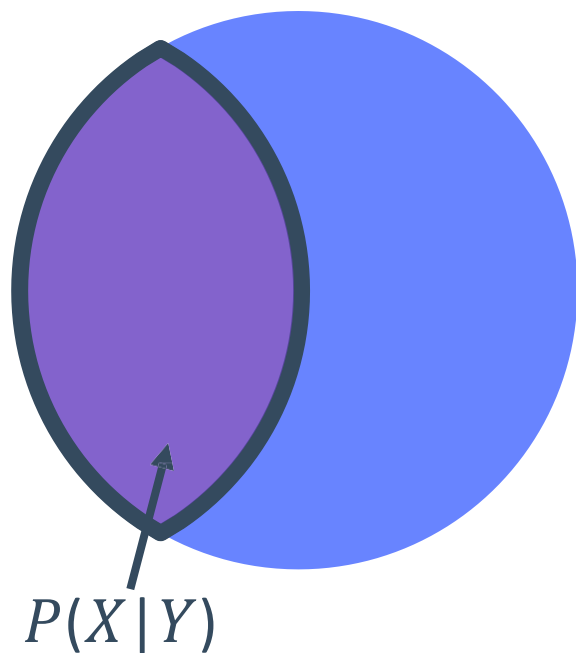
概率基本知识



单个事件概率: $P(X)$ $P(Y)$

联合事件概率: $P(X, Y)$

概率基本知识

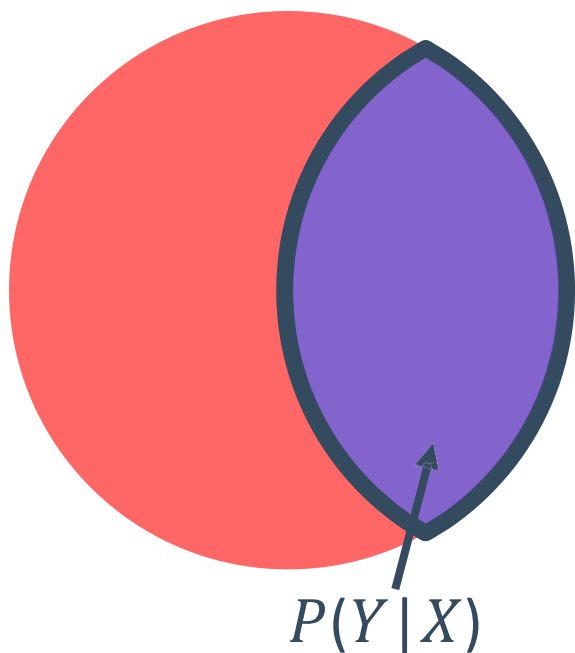


单个事件概率: $P(X), P(Y)$

联合事件概率: $P(X, Y)$

条件概率: $P(X|Y)$

概率基本知识

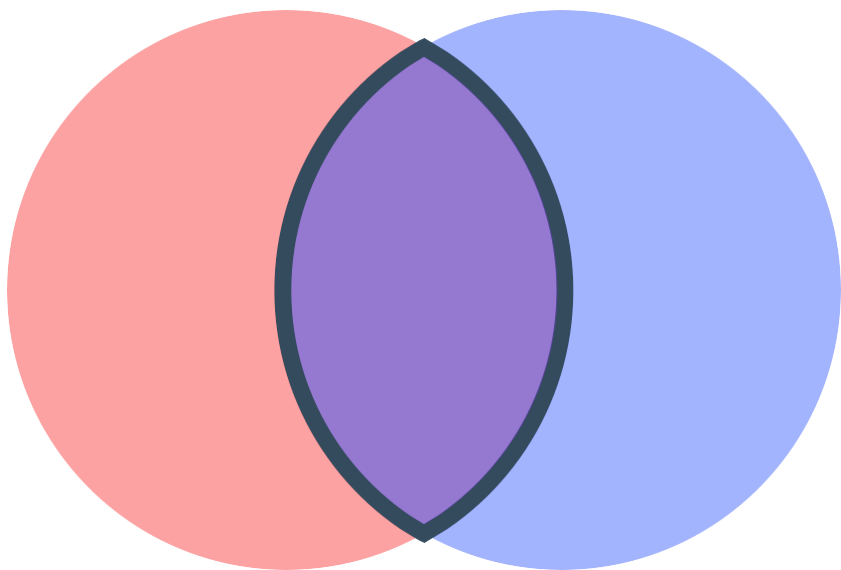


单个事件概率: $P(X), P(Y)$

联合事件概率: $P(X, Y)$

条件概率: $P(X|Y), P(Y|X)$

概率基本知识



单个事件概率: $P(X), P(Y)$

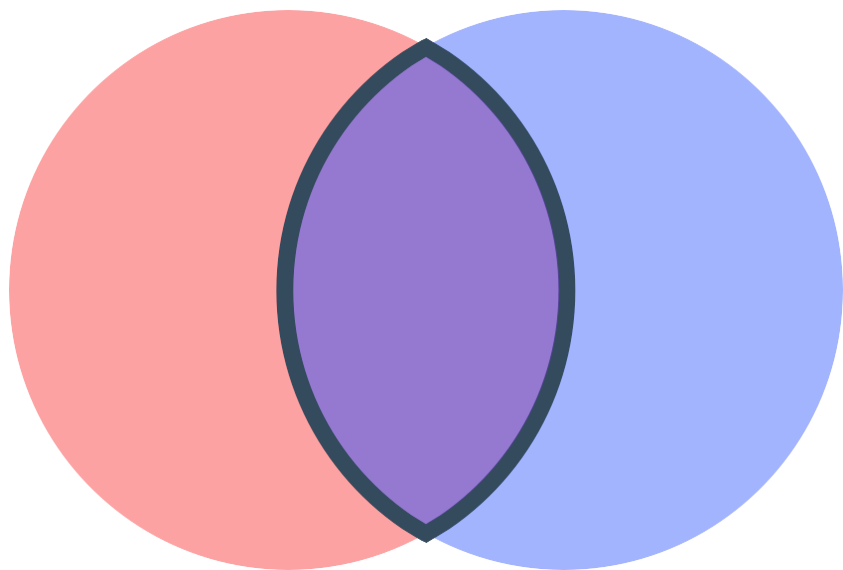
联合事件概率: $P(X, Y)$

条件概率: $P(X|Y), P(Y|X)$

条件和联合概率的关系:

$$\begin{aligned} P(X, Y) &= P(Y|X) * P(X) \\ &= P(X|Y) * P(Y) \end{aligned}$$

贝叶斯公式推导



条件和联合概率的关系：

$$\begin{aligned} P(X,Y) &= P(Y|X)*P(X) \\ &= P(X|Y)*P(Y) \end{aligned}$$



$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

$$P(X) = \sum_Z P(X,Z) = \sum_Z P(X|Z) * P(Z)$$

贝叶斯定理

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

似然估计

类别的先验概率

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

类别的后验概率

变量的先验概率

贝叶斯定理

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

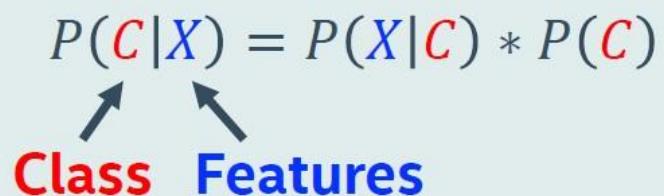
$$posterior = \frac{likelihood * prior}{evidence}$$

贝叶斯定理

例：在肿瘤病人诊断中， X 表示各个指标的检测结果， C 表示诊断的类别，分别为恶性肿瘤和良性肿瘤。 $P(C|X)$ 表示给定当前指标的检测结果，预测恶性肿瘤和良性肿瘤的概率分别是多少？（现有病人样本1000人，100人为恶性肿瘤，900人为良性肿瘤。假设只需检测一项指标，其取值范围为 $[0, 100]$ ，医生根据经验划分成 $[0, 20)$ ， $[20, 50)$ 和 $[50, 100]$ 三个区间。根据历史数据，恶性肿瘤和良性肿瘤对应这三个区间内指标的概率分别为 $(0.1, 0.3, 0.6)$ ， $(0.5, 0.3, 0.2)$ 。）

朴素贝叶斯分类器

给定特征向量(X),
计算其属于每个
类别 (C) 的概率

$$P(\text{Class}|\text{Features}) = P(\text{Features}|\text{Class}) * P(\text{Class})$$


朴素贝叶斯分类器

给定特征向量(X),
计算其属于每个
类别 (C) 的概率

$$P(C|X) = P(X|C) * P(C)$$

很难计算所有特
征的联合概率

$$\begin{aligned} P(C|X) &= P(X_1, X_2, \dots, X_n|C) * P(C) \\ &= P(X_1|X_2, \dots, X_n, C) * P(X_2, \dots, X_n|C) * P(C) \\ &\dots \end{aligned}$$

朴素贝叶斯分类器

给定特征向量(X),
计算其属于每个
类别 (C) 的概率

解决方案: **假设**
给定类别, 所有
特征相互独立

这就是“朴素”
的假设

$$P(C|X) = P(X|C) * P(C)$$

$$P(C|X) = P(X_1|C) * P(X_2|C) * P(X_n|C) * P(C)$$

$$P(C|X) = P(C) \prod_{i=1}^n P(X_i|C)$$

朴素贝叶斯分类器

给定特征向量(X),
计算其属于每个
类别 (C) 的概率

$$P(C|X) = P(X|C) * P(C)$$

按照最大后验概率规则，把X分
入概率最大的类别

$$\underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} P(C_k) \prod_{i=1}^n P(X_i|C_k)$$

对数技巧

很多概率值连乘，
容易造成浮点计
算下界溢出

取对数把乘法转
化成加法

$$\operatorname{argmax}_{k \in \{1, \dots, K\}} P(\mathbf{C}_k) \prod_{i=1}^n P(X_i | \mathbf{C}_k)$$

$$\log(P(\mathbf{C}_k)) \sum_{i=1}^n \log(P(X_i | \mathbf{C}_k))$$

朴素贝叶斯分类器

例：现有一个数据集，给出了不同天气条件与是否出去打网球的映射关系。预测在给定的天气下是否出去玩。

天气	打网球
晴天	不打
阴天	打
下雨	打
晴天	打
晴天	打
阴天	打
下雨	不打
下雨	不打
晴天	打
下雨	打
晴天	不打
阴天	打
阴天	打
下雨	不打

频率表

天气	不打	打
晴天	2	3
阴天	0	4
下雨	3	2
总计	5	9

似然表

天气	不打	打	$P(X)$	
晴天	2	3		
阴天	0	4		
下雨	3	2		
总计	5	9	--	
$P(C)$			--	
$P(X C)$			--	

案例：预测打网球

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
D15	Overcast	Mild	Normal	Weak	?

案例：预测打网球

$$P(\text{Play}=\text{Yes}) = 9/14$$

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

$$P(\text{Play}=\text{No}) = 5/14$$

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

使用训练数据构建概率查找表

拉普拉斯平滑技术

问题：有些类别里没有的特征将会导致这些条件概率值为"0"

解决方案：在这些条件概率的分子和分母上各加1

$$P(C|X) = \underbrace{P(X_1|C)}_0 * P(X_2|C) * P(C)$$

$$P(X_1|C) = \frac{1}{Count(C) + n}$$

$$P(X_2|C) = \frac{Count(X_2 \& C) + 1}{Count(C) + m}$$

朴素贝叶斯模型的类型

朴素贝叶斯模型	数据类型
贝努利模型	二值(0/1, T/F...)
多项式模型	离散值 (如, 计数)
高斯模型	连续数值型

不同朴素贝叶斯模型的区别，主要在于它们对概率分布 $P(X_i|C)$ 所做的不同假设。

结合不同特征类型

问题

模型特征包含不同的数据类型（连续的和类别的）

解决

方案 1： 将连续特征离散化成类别变量，然后应用多项式模型

方案 2： 用高斯模型拟合连续特征，用多项式模型拟合分类变量，然后再结合成一个“元模型”

文本分类

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
A very close game	?

朴素贝叶斯算法小结

主要优点：

- 朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。
- 对小规模的数据表现很好，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。
- 运行速度快，能处理多分类任务， 可以进行是实时预测。
- 对缺失数据不太敏感，算法也比较简单，容易实现，很容易进行模型训练，适合各种规模的数据集，常用于文本分类。

朴素贝叶斯算法小结

主要缺点：

- 理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型给定输出类别的情况下，假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。
- 需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。
- 由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。
- 对输入数据的表达形式很敏感。