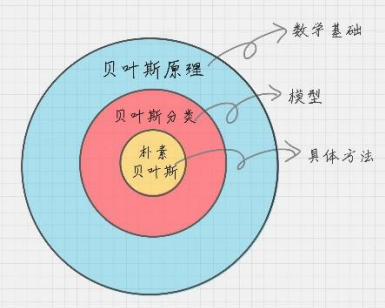
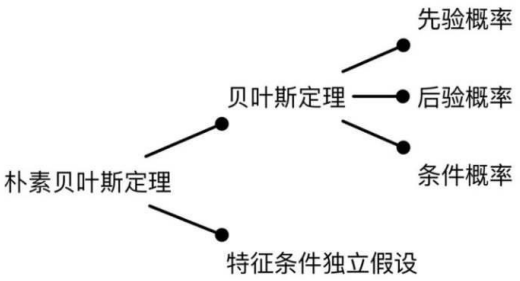
朴素贝叶斯

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法 。

最为广泛的两种分类模型是决策树模型(Decision Tree Model)和朴素贝叶斯模型（Naive Bayesian Model，NBM）。和决策树模型相比，朴素贝叶斯分类器(Naive Bayes Classifier 或 NBC)发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC 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 NBC 模型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，这给 NBC 模型的正确分类带来了一定影响。





1.贝叶斯定理

1 条件概率

设 A，B 是两个事件，且 $P(A)>0$，称

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

为在事件 A 发生的条件下事件 B 发生的条件概率。

2 样本空间的划分

设 S 为试验 E 的样本空间，$B_1,\ B_2,\ \cdots,\ B_n$ 为 E 的一组事件。若

(i) $B_iB_j=\emptyset,\ i\neq j,\ i,j=1,2,\cdots,n;$

(ii) $\mathrm{B}_1\cup B_2\cup\cdots\cup B_n=S,$

则称 $B_1,\ B_2,\ \cdots,\ B_n$ 为样本空间 S 的一个划分。

3 全概率公式

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

$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\cdots+P(A|B_n)P(B_n)$$

称为全概率公式。

2. 贝叶斯公式及应用

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

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n}P(A|B_j)P(B_j)},\quad i=1,2,\cdots,n$$

称为贝叶斯（Bayes）公式。

先验概率与后验概率

举一个例子来理解贝叶斯公式的应用。

对以往数据分析结果表明，当机器调整得良好时，产品的合格率为 98%，而当机器发生某种故障时，其合格率为 55%。每天早上机器开动时，机器调整良好的概率为 95%。试求已知某日早上第一件产品是合格品时，机器调整良好的概率是多少？

设 A 为事件“产品合格”，B 为事件“机器调整良好”。已知$P(A|B)=0.98$，$P(A|\overline{B})=0.55$，$P(B)=0.95$，$P(\overline{B})=0.05$，所需求的概率为$P(B|A)$。用贝叶斯公式来计算

$$\begin{aligned}P(B|A)&=\frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\overline{B})P(\overline{B})}\\&=\frac{0.98\times0.95}{0.98\times0.95+0.55\times0.05}=0.97\end{aligned}$$

这就是说，当生产出第一件产品是合格品时，此时机器调整良好的概率为 0.97。这里，概率 0.95 是由以往的数据分析得到的，叫做先验概率。而在得到信息（即生产出的第一件产品是合格品）之后再重新加以修正的概率（即 0.97）叫做后验概率。有了后验概率

3.朴素贝叶斯分类

求 X 样本属于 C 类别的概率，即当观察到 X 样本出现时，其所属的类别为 C 的概率：

P(C|X)=P(C)P(X|C)/P(X)

P(C)P(X|C)=P(C,X)=P(C,x1,x2,...,xn)

=P(x1,x2,...,xn,C)

=P(x1|x2,...,xn,C)P(x2,x3,...,xn,C)

=P(x1|x2,...,xn,C)P(x2|x3,...,xn,C)P(x3,x4,...,xn,C)

=P(x1|x2,...,xn,C)P(x2|x3,...,xn,C)P(x3|x4,...,xn,C)...P(C)

朴素：条件独立假设，即样本各个特征之间并无关联，不构成条件约束。

=P(x1|C)P(x2|C)P(x3|C)...P(C)

X 样本属于 C 类别的概率，正比于 C 类别出现的的概率乘以 C 类别条件下 X 样本中每一个特征值出现的概率之乘积。

```
import sklearn.naive_bayes as nb
model = nb.GaussianNB()
```