# Naive Bayes

## Bayes公式

> 1. 先验概率：事件发生前的预判概率(可以基于历史数据/背景/主观观点)
> 2. 后验概率：事件发生后求的反向条件概率

### Bayes公式

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

$P(A)$是先验概率，一般是人主观给出的；$P(B)$是先验概率，往往通过全概率公式求出；$P(B|A)$是条件概率，一般是通过历史数据统计得到；$P(A|B)$是后验概率，通常是求解目标。

## Naive Bayes算法核心思想

Naive Bayes算法定义：设$X\{a_1,a_2,\cdots,a_n\}$是待分类项，每个$a_i$是$x$的一个特征属性，且**特征属性之间相互独立**。设$C\{y_1,y_2,\cdots,y_n\}$是一个类别集合，$P(y_k|x)=\max\{P(y_1|X),P(y_2|X),\cdots,P(y_n|X)\}$，则$X\in y_k$。

Naive Bayes是生成方法，通过考虑特征概率来预测分类：对给出的待分类样本$X\{a_1,a_2,\cdots,a_n\}$求解此样本出现的条件下各个类别$y_i$出现的概率，哪个$P(y_i|X)$最大就认为属于哪个类别。

### Naive Bayes的计算过程

1. 找到训练样本集(已知分类的待分类项集合)，统计得到在各类别下各个特征属性的条件概率估计：$P(a_1,y_1),\cdots,P(a_n|y_1)\cdots\cdots P(a_1|y_n),\cdots,P(a_n|y_n)$
2. 根据Bayes公式：$P(y_i|X)=\frac{P(X|y_i)P(y_i)}{P(X)}$。$P(X|y_i)=\prod_{k=1}^{n}P(a_k|y_i)$，$P(a_k|y_i)$是指在类别$y_i$中，特征元素$a_j$出现的概率，可求解为$P(a_k|y_i)=\frac{|\text{训练样本为}y_i\text{时},a_j\text{出现的次数}|}{|y_i\text{训练样本数}|}$；$P(y_i)$是指在训练样本中$y_i$出现的概率，可近似求解为$P(y_i)=\frac{|y_i|}{D}$

## 多项式/伯努利/高斯Naive Bayes

e.g. 文本分类：单词是对应的特征$a_j$，类别标签是$y$。每个单词的频次可能大于1，如果直接以单词频次参与统计计算，则为**多项式朴素贝叶斯**；如果以是否出现(0和1)参与统计计算，则为**伯努利朴素贝叶斯**。

## 多项式Naive Bayes

设某文档$d=(t_1,t_2,\cdots,t_k)$，$t_k$是该文档中出现过的单词，允许重复。

先验概率$P(c)=\frac{\text{类}c\text{下单词总数}}{\text{整个训练样本的单词总数}}$

类条件概率$P(t_k|c)=\frac{\text{类}c\text{下单词}t_k\text{在各文档中出现过的次数之和}+1}{\text{类}c\text{下单词总数}+|V|}$，$V$是训练样本单词表，$|V|$表示训练样本包含多少单词。

## 伯努利Naive Bayes

对应上面的多项式Naive Bayes：

先验概率$P(c)=\frac{\text{类}c\text{下文档总数}}{\text{整个训练样本的文档总数}}$

类条件概率$P(t_k|c)=\frac{\text{类}c\text{下包含单词}t_k\text{的文档数}+1}{\text{类}c\text{下文档总数}+2}$

## 高斯Naive Bayes

特征$x_i$是连续变量，假设在$y_i$的条件下，$x$服从高斯分布。则有

$P(x_i|y_k)=\frac{1}{\sqrt{2\pi\sigma_{y_{k,i}}^2}}e^{-\frac{(x_i-\mu_{y_{k,i}})^2}{2\sigma_{y_{k,i}}^2}}$，其中$\mu_{y_{k,i}}$表示类别为$y_k$的样本中，第$i$维特征的均值；$\sigma_{y_{k,i}}^2$表示类别为$y_k$的样本中，第$i$维特征的方差。

# 平滑处理

原因：可能出现零概率问题

方法：Laplace平滑，e.g. 文本分类$P(x_i|y)=\frac{N_{yi}+\alpha}{N_y+n\alpha}$，其中$N_{yi}$是类$y$的所有样本中特征$x_i$的特征值之和；$N_y$表示类$y$的所有样本中全部特征的特征值之和；$\alpha\in[0,1]$表示平滑值；$n$表示特征总数。