# NBC

Naive Bayes Classifier——朴素贝叶斯分类器

1. 基础算法：
2. 设，每个为的一个特征属性，为类别集合
3. 用Bayes公式计算

假设各特征属性条件独立：

1. 将未知样本归为2)中计算出得到条件概率最大的那一类

注：由于Bayes公式中分母相同，只要比较分子大小即可

1. 的估计：
2. 特征属性为离散值时

只要将训练样本中各个划分在每个类别中出现的频率用于估计即可

1. 特征属性为连续值时

通常假定其值服从正态分布，即

1. 优点：
2. 当特征属性充分多时，对个别属性抗干扰性强【详见5.例子1】
3. 对缺失数据不敏感，在这方面优于决策树模型
4. 算法稳定，对不同数据特点分类性能差别不大，健壮性比较好
5. 分类过程中时间、空间开销小
6. 理论上有最小的错误率（因为是严谨的全等号概率公式推导）
7. 缺点
8. 当某个类别下某个特征划分没有出现时，这样，这将导致分类器质量大大降低

**改进（Laplace平滑，离散型）**：对每类别下所有划分的计数+1，这样如果训练样本

数量充分大时对结果不会产生影响

1. 各特征属性条件独立的假设一般不成立，对正确分类带来影响

**改进1：贝叶斯网络**

**改进2：半朴素贝叶斯**

将特征相关的属性分成一组，假设不同组中的属性是相互独立的，同一组 中的属性是相互关联的

**改进3：TAN**(tree augmented naive Bayes)

允许每个属性结点最多可以依赖一个非类结点

1. 例子1（离散变量）：检测SNS(Social Network Site)中不真实账号

设一个账号特征属性有日志数量/注册天数，好友数量/注册天数，是否使用真实头像，取值为别为：，，，类别集合为

①统计训练样本中真实账号、不真实账号概率：

，

②统计每个类别条件下各个特征属性划分的概率：

，，

，，

，，

，，

，，，

③用朴素贝叶斯分类器鉴别一个日志数量/注册天数=0.1，好友数量/注册天数=0.2，使用非真实头像账号的真实性：

④故判定这个账号为真实账号，尽管这个账号没有使用真实头像

1. 例子2（连续变量）：根据人类身体特征判断性别

设一个人特征属性有身高（英尺），体重（磅），脚掌（英寸），类别集合为，训练样本如下表所示：

性别　　身高（英尺）　体重（磅）　　脚掌（英寸）

男 　　　6 　　　　　　 180　　　　　12

男 　　　5.92　　　　　 190　　　　　11

男 　　　5.58　　　　　 170　　　　　12

男 　　　5.92　　　　　 165　　　　　10

女 　　　5 　　　　　　 100　　　　　6

女 　　　5.5 　　　　　 150　　　　　8

女 　　　5.42　　　　　 130　　　　　7

　　女 　　　5.75　　　　　 150　　　　　9

①统计训练样本中男女概率：

②统计每个类别条件下各个特征属性划分的概率：

假设正态分布：

③用朴素贝叶斯分类器鉴别一个身高6英尺，体重130磅，脚掌8英寸的人是男是女：

1.5789\*5.9867e-06\*\*0.001311\*0.5=6.2e-09

0.2235\*0.01679\*0.2867\*0.5=5.4e-04

④故判定这个人为女

1. 当特征维度参差不齐时，也可以用类似的思想做：

例子3：垃圾邮件鉴别

假设邮件中只有1,2,3三个词，前两封只有两个词，后两封有3个次词，类别集合为 ，训练样本如下表所示：

第一个词 第二个词 第三个词 分类

1 2 - 1(垃圾邮件)

2 1 - 0(非垃圾邮件)

1 3 2 0(非垃圾邮件)

3 3 3 1(垃圾邮件)

①统计训练样本中垃圾邮件概率：

②统计每个类别条件下各个特征属性划分的概率，不用Laplace平滑：

如果用Laplace平滑，公式应变为：

其中表示邮件总数，表示第封邮件的长度，表示词典中词的总个数，

为示性函数，当大括号中事件成立时为1，否则为0

分子表示垃圾邮件中词k出现的次数总和，

分母表示垃圾邮件词数总和，

所以整体表示垃圾邮件中词k的出现概率

③鉴别一封的邮件是否垃圾邮件：

④故判定这封邮件为垃圾邮件