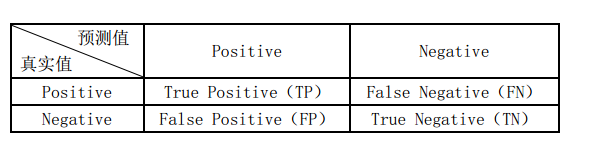
# 概念

目录

[概念 1](#_Toc38565349)

[1. 各种精确度问题 1](#_Toc38565350)

## 各种精确度问题



TP:实际正样本 预测正样本

FP:实际负样本 预测正样本

FN:实际正样本 预测负样本

TN:实际负样本 预测负样本

#### Accuracy(准确率)

预测对的样本占总样本的比重

*Accuracy*=*TP*+*TN*/(*TP*+*TN*+*FP*+*FN*)

#### Precision(精确率)

预测为正的样本中有多少是真正的正样本

*Precision*=*TP*/(*TP*+*FP*)

#### 召回率(Recall)

针对原本样本，样本中的正例有多少被预测准确的

*Recall*=*TP*/(*TP*+*FN*)

#### F1-score 是基于召回率和精确率计算的

*F*1*score*=2∗*Precision*∗*Recall*/(*Precision*+*Recall*)

#### 词袋模型

将文档表示成特征矢量，基本思想是对于一个文本忽略其词序语法句法 ，仅仅将其看作是词汇的集合，而文本中的每个词汇是独立的。

将每篇文章看作一个袋子，然后根据袋子里词汇出现的频率进行分类

#### CountVectorizer

函数只考虑每个单词出现的频率；然后构成一个特征矩阵，每一行表示一个训练文本的词频统计结果

#### count\_vector.get\_feature\_names()

构建词汇表

#### 1.朴素贝叶斯

#### 1.1贝叶斯原理

**如果我们事先不知道袋子里面黑球和白球的比例，而是通过我们摸出来的球的颜色，能判断出袋子里面黑白球的比例么？**

这是因为，贝叶斯原理与其他统计学推断方法截然不同，它是建立在主观判断的基础上：在我们不了解所有客观事实的情况下，同样可以先估计一个值，然后根据实际结果不断进行修正。

先验概率

通过经验来判断事情发生的概率

后验概率

后验概率就是发生结果之后，推测原因的概率，比如说某人查出来了患有“贝叶死”，那么患病的原因可能是 A、B 或 C。\*\*患有“贝叶死”是因为原因 A 的概率就是后验概率

朴素贝叶斯是一种非常简单的用于不同相互独立特征下的分类 假设每种输入变量是独立的

朴素贝叶斯分类常用于文本分类，尤其是对于英文等语言来说，分类效果很好。

朴素贝叶斯分类算法是学习效率和分类效果较好的分类器之一。直观的文本分类算法，也是最简单的贝叶斯分类器，具有很好的可解释性，朴素贝叶斯算法特点是假设所有特征的出现相互独立互不影响，每一特征同等重要。

虽然"所有特征彼此独立"这个假设，在现实中不太可能成立，但是它可以大大简化计算，而且有研究表明对分类结果的准确性影响不大。

公式

假设某个体有n项特征（Feature），分别为F1、F2、...、Fn。现有m个类别（Category），分别为C1、C2、...、Cm。贝叶斯分类器就是计算出概率最大的那个分类，也就是求下面这个算式的最大值

由贝叶斯定理 　P(A|B) = P(B|A) P(A) / P(B)

P(C|F1F2...Fn)  
　　= P(F1F2...Fn|C)P(C) / P(F1F2...Fn)

由于 P(F1F2...Fn) 对于所有的类别都是相同的，可以省略，问题就变成了求

P(F1F2...Fn|C)P(C)

更进一步贝叶斯，假设所有的特征是独立的

P(F1F2...Fn|C)P(C)  
　　= P(F1|C)P(F2|C) ... P(Fn|C)P(C)

MultinomialNB假设特征的先验概率为多项式分布：

[拉普拉斯平滑处理 Laplace Smoothing](https://www.cnblogs.com/bqtang/p/3693827.html)

#### 逻辑回归

一种用于解决二分类（0 or 1）问题的机器学习方法，用于估计某种事物的可能性。

注意，这里用的是“可能性”，而非数学上的“概率”，logisitc回归的结果并非数学定义中的概率值，不可以直接当做概率值来用。该结果往往用于和其他特征值加权求和，而非直接相乘。

和线性回归的区别是使用了sigmoid函数

线性回归预测输出的是（-∞,+∞）

而逻辑回归输出的是{0,1}

调参

* + - penalty： L1/L2
    - Dual:目标函数是原始形式还是对偶形式
    - Tol：优化算法的停止条件
    - C= 正则化系数 系数越小 越强
    - fit\_intercept 常数项
    - solver ：

liblinear 坐标轴下降

lbfgs 拟牛顿法

newton-cg 牛顿法的一种

sag 随机平均梯度下降

#### 决策树

#### 如何选择特征作为节点：获得信息增益最高的特征就是最好的选择

#### class\_weight： 指定样本各类别的的权重，主要是为了防止训练集某些类别的样本过多导致训练的决策树过于偏向这些类别。这里可以自己指定各个样本的权重，如果使用“balanced”，则算法会自己计算权重，样本量少的类别所对应的样本权重会高。

**max\_depth：**  int or None, optional (default=None) 一般来说，数据少或者特征少的时候可以不管这个值。如果模型样本量多，特征也多的情况下，推荐限制这个最大深度，具体的取值取决于数据的分布。常用的可以取值10-100之间。常用来解决过拟合

#### 随机森林

随机森林是一种灵活的、便于使用的机器学习算法，即使没有超参数调整，大多数情况下也会带来好的结果。

#### LDA

是一种文档主题生成模型，它是一种主题模型，它包含文章，主题和词三个部分，它可以将文档集中每篇文档的主题以概率分布的形式给出

从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。

LDA就是要干的事就是根据给定的一篇文档，推测其主题分布。它是一个无监督学习，类似于聚类。