**【量化课堂】朴素贝叶斯入门**

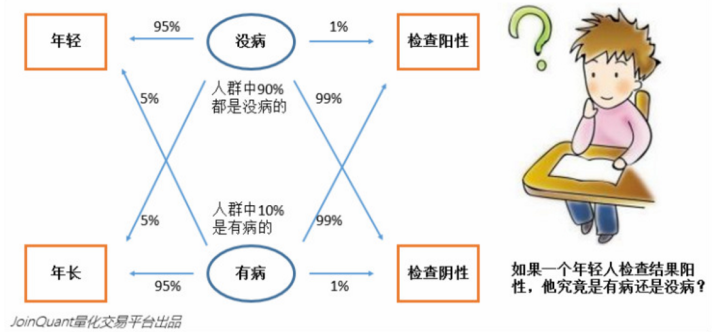
**导语：**在现实生活中，我们经常要利用观测现象（特征数据）推测现象背后的原因。例如我们看到草地湿了，需要判断是不是下雨导致的；今天的交易量大涨，需要判断是有新资金入场、还是存量资金雄起了一把；去医院体检，检查结果为阳性，是因为真的得病了，还是因为医院的误诊。朴素贝叶斯算法可以利用历史数据的分布，给你一个最有可能的结果，使你犯错误的概率最小化。

本文由JoinQuant量化课堂推出 。难度标签为进阶上，理解深度标签：level-0

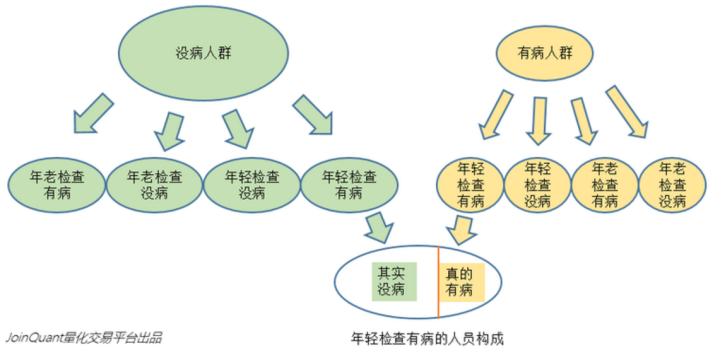
作者：yongpeng.r

编辑：宏观经济算命师

**1. 一个例子说清楚的事情绝不用定义：**

先说一句特别复杂的朴素贝叶斯介绍：朴素贝叶斯法是基于贝叶斯定理，特征条件独立假设和后验概率最大化的分类方法。不知道你们看完了什么感受，反正我是一脸懵逼。下面我用一个小例子让大家明白这到底是怎么回事。  
举个例子：某一种病，年轻人得病的概率远远小于年长的。如果一个年轻人检查为阳性，那么他就能直接被确诊吗？要知道，检查为阳性，可能会是误诊的哦。这个时候，朴素贝叶斯就登场了。我们不妨把各种可能性列出来，画一个图：  
  
上图中，箭头附近的数字表示各种情况的概率。例如，没病然而检查为阳性（说明误诊了）的概率是1%，没病而且检查为阴性的概率是99%。如果一个年轻人去医院体检，体检结果是有病，那么这个人到底是有病还是没病呢？或者说这个人真实得病的概率有多大呢？

有的人可能会说，既然有病的人会有99%的被确诊，至少得病的概率比没病的概率要高吧。其实，一个年轻人检查出有病，真正得病的概率比没病的概率还要低！

现在我们来分析下：假设人群中有20000人，按照第一个图中的患病的概率，18000人是没有病的，2000人是有病的；在18000个没有病的人中，年轻人的概率为95%，检查为阳性的概率为1%，那么“年轻”并且“检查为阳性”并且“没病”的人数一共有18000⋅95%⋅1%=17118000⋅95%⋅1%=17118000⋅95%⋅1%=171人；在2000个有病的人中，年轻人的概率是5%，检查为阳性的概率是99%。那么年轻人得病并且被检查出来的人数为2000⋅5%⋅99%=992000⋅5%⋅99%=992000⋅5%⋅99%=99人。如果我们现在只知道这个人是年轻人，而且检查结果是阳性：那么他有可能是本身真的得病并且被检查出来的人，也有可能是误诊了的人。前者的概率就是99/(99+171)=36.7%99/(99+171)=36.7%99/(99+171)=36.7%。后者为171/(99+171)=63.3%171/(99+171)=63.3%171/(99+171)=63.3%。很显然，我们有更大的可能性相信他没有得病（所以现实生活中，医生会让你多复诊几次）。  
  
在上面的整个分析过程中，我们就分别用到了特征条件独立假设，贝叶斯定理，条件概率最大化这几个知识点。下面我们进行详细说明。

**2. 特征的条件独立假设**

我们的目的是通过“目前已知的数据”判断未知的结果，那么这个“目前已知的数据”就被称为特征。在上面判断有没有得病的例子中，特征就是这个人“是否年轻”以及“检查结果是否为阳性”。

这里我们要做一个重要的假设：上述两个特征之间是独立的。在判断这个人有没有病的时候，我们认为这个人“是否年轻”和“检查是否为阳性”之间没有联系。因此，随机抽取一个检查者，他“年轻”并且“检查结果为阳性”的概率就等于“年轻”的概率乘以“检查结果为阳性”的概率。

上面这个假设就是条件独立假设。如果变量不满足独立性，则不可以将两者的概率相乘，比如天空有云的概率是0.5，下雨的概率是0.33，但下雨和“天空中有云”不是独立的，就不能得到“即有云又下雨”的概率为0.5⋅0.330.5⋅0.330.5⋅0.33这个结论。

**3.贝叶斯定理**

贝叶斯定理主要描述在给定特征数据的情况下，判定属于某个类别的概率。在下面的公式中，样本的数据用X=xX=xX=x表示，样本的类别属于某个类别用Y=ckY=ckY=ck表示。

P(Y=ck|X=x)=(P(X=x|Y=ck)P(Y=ck))/(∑kP(X=x|Y=ck)P(Y=ck))

P(Y=ck|X=x)=(P(X=x|Y=ck)P(Y=ck))/(∑kP(X=x|Y=ck)P(Y=ck))P(Y=ck|X=x)=(P(X=x|Y=ck)P(Y=ck))/(∑kP(X=x|Y=ck)P(Y=ck))

在章节1给出的例子中，注意看我们计算真正得病的公式：

年轻人检查为阳性而真的得病的概率=A/B=99/(99+171)

年轻人检查为阳性而真的得病的概率=A/B=99/(99+171)年轻人检查为阳性而真的得病的概率=A/B=99/(99+171)

其中，A为真实得病的人中“检查为阳性”并且“年轻”的数目，B为人群中所有“检查为阳性”并且“年轻”的人的数目，其实这个公式就是贝叶斯定理的公式。现在感觉贝叶斯定理是不是简单了很多。

**4. 后验概率最大化**

我们知道一个“年轻”人“检查为阳性”，现在需要你告诉他有病还是没病。那么很显然，我们只要用上文公式做计算，看看这个“年轻”并且“检查结果为阳性”的人到底是得病的概率更高，还是没得病的概率更高。这个就是后验概率最大化的直观解释。在上面的例子中，我们就可以告诉他，检查结果为阳性也不意味着你就得病了，但是为了安全起见，需要后续跟进复查。

**5. 朴素贝叶斯的具体使用-sklearn**

上面通过了一个小例子介绍了一下朴素贝叶斯算法的算法原理，但是如何在实际的代码中使用朴素贝叶斯算法帮助我们完成分类呢？下面介绍下python环境中的朴素贝叶斯算法是如何使用的。

特征是通过收盘价数据计算的SMA，WMA，MOM指标，训练样本的特征是从2007-1-4到2016-6-2中截止前一天的SMA，WMA，MOM指标，训练样本的类别标签是2007-1-4日到2016-6-2中每一天的涨跌情况，涨了就是True，跌了就是False，测试样本是2016-6-3日的三个指标以及涨跌情况。我们可以判定之后判断结果是正确还是错误，如果通过朴素贝叶斯判断的结果和当天的涨跌情况相符，则输出True，如果判断结果和当天的涨跌情况不符，则输出False。（和SVM那一篇中例子的作用是一样滴，只是为了展示如何使用，不对预测的准确性做担保啊）

本文由JoinQuant量化课堂推出，版权归JoinQuant所有，商业转载请联系我们获得授权，非商业转载请注明出处。