朴素贝叶斯分类器

石川
量化交易 话题的优秀回答者

1 引言

有监督分类是量化投资中常见的情景之一。比如，我们希望根据上市公司财报中的各种指标特征，区分出优秀的和差劲的股票，这就是一个分类问题。在机器学习中，有监督分类的算法有很多，比如 SVM、ANN 以及基于决策树的 AdaBoost 和随机森林等。这其中自然也少不了今天的主角**朴素贝叶斯分类器（Naïve Bayes classifiers）**。它代表着**一类**应用**贝叶斯定理**的分类器的总称。朴素（naive）在这里有着特殊的含义、代表着一个非常强的假设（下文会解释）。

朴素贝叶斯分类器虽然简单，但是用处非常广泛（尤其是在文本分类方面）。在 IEEE 协会于 2006 年列出的十大数据挖掘算法中，朴素贝叶斯分类器赫然在列（Wu et al. 2008）。捎带一提，另外九个算法是 C4.5、k-Means、SVM、Apriori、EM、PageRank、AdaBoost、kNN 和 CART（那时候深度学习还没有什么发展）。

朴素贝叶斯分类器以贝叶斯定理为基础。下面首先回顾一下贝叶斯定理（熟悉的朋友可以跳过第 2 节）。之后会阐释“朴素”的意义并介绍朴素贝叶斯分类器。文章的最后使用一个例子说明如何使用朴素贝叶斯分类器选股。

贝叶斯定理的推导始于条件概率。**条件概率可以定义为：在事件 B 发生的前提下，事件 A 发生的概率。**数学上用 $P(A|B)$ 来表示该条件概率。条件概率 $P(A|B)$ 的数学定义为：

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

这个公式的白话解释为：**“当 B 发生前提下 A 发生的概率”等于“A 和 B 同时发生的概率”除以“B 发生的概率”。**

生活中条件概率屡见不鲜。比如“在没有赶上 8 点这趟地铁的前提下，上班迟到的概率是多少？”应用条件概率的定义可知“在没有赶上 8 点这趟地铁的前提下，上班迟到的**条件**概率”等于“没赶上 8 点这趟地铁且上班迟到的概率”除以“没赶上 8 点这趟地铁的概率”。将上式左右两边同时乘以 $P(B)$ 得到：

$$P(B)P(A|B) = P(A \cap B)$$

类似的，我们也可以求出 P(B|A)，即在 A 发生的前提下，B 发生的概率是多少。在上面例子中，这对应着“在上班迟到的前提下，没有赶上 8 点这趟地铁的概率是多少”？（上班迟到的原因可能很多，比如没赶上这趟地铁是一个，又比如在公司楼下的咖啡馆里耽搁了 10 分钟也是一个，或者因为早上发烧先去医院了等等。）根据定义：

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

同样，两边同时乘以 $P(A)$，并且由 $P(A \cap B) = P(B \cap A)$，得到：

$$P(A)P(B|A) = P(A \cap B)$$

由此可知 $P(B)P(A|B) = P(A)P(B|A)$。这个结果也可以写作如下形式，即大名鼎鼎的**贝叶斯定理（Bayes rule）**：

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

3 何为“朴素”？

下面我们将贝叶斯定理应用于有监督的分类场景。令 $\mathbf{X}$ 代表一个 n 维特征向量（它代表着一些特征，即 features），这些特征用来描述一个对象；C 代表该对象所属的类别。分类的目的

具体的，假设类别的个数为 K （即 C 的取值有 K 个），那么对于每一个可能的取值（记为 c_k ， $k=1,2,\cdots,K$ ），我们需要根据给定的特征 $\mathbf{X}$ 计算出概率 $P(C=c_k|\mathbf{X})$ 。然后，**只要从所有的 $P(c_k|\mathbf{X})$ 中挑出取值最大的概率对应的 c_k 作为最有可能的分类即可。**

利用贝叶斯定理， $P(C=c_k|\mathbf{X})$ 可以写作：

$$P(C=c_k|\mathbf{X})=\frac{P(\mathbf{X}|C=c_k)P(C=c_k)}{P(\mathbf{X})}$$

由于对有所的 $P(C=c_k|\mathbf{X})$ 来说，上式右侧的分母都相同（和 C 的取值无关），因此我们只需要根据训练集数据来估计所有的 $P(\mathbf{X}|C=c_k)$ 以及所有的 $P(C=c_k)$ 即可。下面来看看为了实现这个目标，需要多大的样本空间。

考虑最简单的情况。假设 n 维向量 $\mathbf{X}$ 中的每一个特征以及类别 C 都是二元的（binary）。因此，特征向量 $\mathbf{X}=\{X_1,X_2,\ldots,X_n\}$ 所有可能的取值为 2^n 个（因为每个 X_i 的取值有 2 个，而一共有 n 个 X_i ）。此外， C 的取值也是 2 个。因此，仅从 $P(C=c_k|\mathbf{X})$ 来说，需要估计的参数就高达 $2\times(2^n-1)$ 个。而这仅仅是从特征空间所有取值组合可能性出发的最低要求。事实上，为了得到准确的参数估计，对于每一个 n 维特征的组合，我们都需要多个观测值来计算 $P(C=c_k|\mathbf{X})$ 的概率。这进一步增加了对样本空间大小的要求。举例来说，如果特征空间的维度 $n=30$ ，那么我们需要估计超过 30 亿个参数！

在现实的应用场景中， $n=30$ 是否常见？非常常见。比如上市公司的特征就可以轻松超过 30 个。而在现实的应用场景中，我们拥有超过 30 亿个样本来估计 30 亿个参数是否常见？痴人说梦。因此，**想利用有限的样本数据估计出所有的 $P(\mathbf{X}|C=c_k)$ 和 $P(C=c_k)$ 是不切实际的。**

为什么有这么多参数需要估计呢？这是因为在求解 $P(\mathbf{X}|C=c_k)$ 时，我们考虑的是 $\mathbf{X}=\{x_1,x_2,\ldots,x_n\}$ 在 $C=c_k$ 这个条件下的**条件联合分布**，这大大增加了待估计的参数的个数。为了解决这个问题， “朴素”闪亮登场。

朴素贝叶斯在求解 $P(\mathbf{X}|C=c_k)$ 时做了一个非常强的假设 —— **条件独立性（conditional independence）。它的意思是在给定的类别 $C=c_k$ 下，不同维度特征的取值之间是相互独立的。**比如令 X_1 和 X_2 代表 n 维里面的两个维度，则 $P(X_1=x_1|C=c_k)$ 的概率与 X_2 的取值无关，即：

$$P(X_1=x_1|C=c_k)=P(X_1=x_1|X_2=x_2,C=c_k)$$

就没什么关系了。当然，打雷和下雨通常在非条件下是相关的，我们仅仅假设在闪电发生的条件下，它们满足条件独立。

上述例子强调了**在朴素贝叶斯中，我们仅仅假设特征之间满足条件独立性，而非一般的独立性。**在条件独立性假设下，反复利用条件概率的定义，$P(\mathbf{X}=(x_1,x_2,\ldots,x_n)|C=c_k)$ 可以写成：

$$\begin{aligned}
P(\mathbf{X}|C=c_k) &= P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|C=c_k) \\
&= P(X_1=x_1|C=c_k)\times P(X_2=x_2|X_1=x_1,C=c_k)\times \\
&\quad P(X_3=x_3|X_1=x_1,X_2=x_2,C=c_k)\times\cdots\times \\
&\quad P(X_n=x_n|X_1=x_1,X_2=x_2,\cdots,X_{n-1}=x_{n-1},C=c_k) \\
&= P(X_1=x_1|C=c_k)\times P(X_2=x_2|C=c_k)\times \\
&\quad P(X_3=x_3|C=c_k)\times\cdots\times \\
&\quad P(X_n=x_n|C=c_k) \\
&= \prod_{i=1}^{n} P(X_i=x_i|C=c_k)
\end{aligned}$$

在前面提及的特征和类别均为 binary 的情况下，这将待估计的参数从 $2\times(2^n-1)$ 个直接减少到 $2n$ 个。**这大大简化了对样本空间的要求以及求解的计算量，使得朴素贝叶斯算法非常简单。条件独立性的假设便是“朴素”一词的来源。**因此，朴素贝叶斯通常也被称为**简单贝叶斯（simple Bayes）**或**独立贝叶斯（independence Bayes）**。

4 朴素贝叶斯分类器

通过上一节对“朴素”含义的说明，朴素贝叶斯分类器的大致轮廓已经比较清晰了。本节就来正式说明其数学表达式。对于特征向量 $\mathbf{X}$ 和类别 C，利用贝叶斯定理和条件独立性的假设，写出每个 $C=c_k$ 的条件概率：

$$\begin{aligned}
P(C=c_k|X_1,X_2,\cdots,X_n) &= \frac{P(C=c_k)P(X_1,X_2,\cdots,X_n|C=c_k)}{P(X_1,X_2,\cdots,X_n)} \\
&= \frac{P(C=c_k)\prod_{i=1}^{n}P(X_i|C=c_k)}{P(X_1,X_2,\cdots,X_n)}
\end{aligned}$$

每个 c_k 求出 $P(C=c_k)\times\prod_i P(X_i=x_i|C=c_k)$ ，并比较这些值中最大的，就可以确定这个新样本的分类：

$$C\leftarrow \arg\max_{c_k} P(C=c_k)\prod_{i=1}^{n} P(X_i|C=c_k)$$

以上就是朴素贝叶斯分类器的数学表达式。

在实际的应用中，根据特征变量是离散的还是了连续的，在使用训练集数据估计 $P(X_i=x_i|C=c_k)$ 时，又有不同的处理方法。在离散的情况下，只需要 counting（计数），即：

$$\hat{P}(X_i=x_i|C=c_k)=\frac{\#\{X_i=x_i\bigwedge C=c_k\}}{\#\{C=c_k\}}$$

其中 $\#\{X_i=x_i\wedge C=c_k\}$ 表示训练集中 $X_i=x_i$ 和 $C=c_k$ 共同发生的次数；$\#\{C=c_k\}$ 表示训练集中 $C=c_k$ 发生的次数。这个估计方法称作**最大似然估计（maximum likelihood estimate）**。

在一些情况下，由于样本数据极度匮乏，很有可能出现某个特征的取值和某个类别的取值在训练集中从未同时出现过，即 $\#\{X_i=x_i\wedge C=c_k\}=0$ ，这会造成对 $P(X_i=x_i|C=c_k)$ 的估计等于零。 $P(X_i=x_i|C=c_k)=0$ 会导致对应的 $P(C=c_k)\times\prod_i P(X_i=x_i|C=c_k)=0$ ，即让我们误以为这个样本属于某个类别 c_k 的概率为 0。这是不合理的，**不能因为一个事件没有观察到就认为该事件不会发生。**

解决这个问题的办法是给每个特征和类别的组合加上给定个数的**虚假样本（“hallucinated” examples）**。假设特征 X_i 的取值有 J 个，并假设为每个 x_i 对应的 $\#\{X_i=x_i\wedge C=c_k\}$ 增加 s 个虚假样本，这样得到对 $P(X_i=x_i|C=c_k)$ 的估计称为**平滑估计（smoothed estimate）**：

$$\hat{P}(X_i=x_i|C=c_k)=\frac{\#\{X_i=x_i\bigwedge C=c_k\}+s}{\#\{C=c_k\}+sJ}$$

特别的，当 $s=1$ 时，上述平滑称为**拉普拉斯平滑（Laplace smoothing）**。类似的，对于 $P(C=c_k)$ 的估计也可以采用平滑的方式：

其中， t 为对每个类增加的虚假样本数， K 是类别个数， $\#\{C\}$ 表示训练集的样本数。

当特征是连续变量时，情况稍微复杂一些。在使用训练集求解 $P(X_i = x_i|C = c_k)$ 时，**需要假设该条件概率分布的形式。**一种常见的假设是认为对于给定的 c_k ， $P(X_i = x_i|C = c_k)$ 满足正态分布，而正态分布的均值和标准差需要从训练集学习得到。这样的模型称为**高斯朴素贝叶斯分类器（Gaussian Naïve Bayes classifier）**。

5 一个例子

下面我们用朴素贝叶斯分类来选股看看。假设描述上市公司的特征有 7 个维度：市盈率、市净率、净资产收益率、总资产周转率变动率、预期盈利增长修正、20 日涨幅、以及市值。为了简化讨论，令每一个特征的取值都是 binary 的，即分为高或者低；进一步令类别也是 binary 的，即好公司（买入后的一段时间内股价上涨）或者差公司（买入后的一段时间内股价下跌）。假设训练集中共有 20 个公司，它们的特征和类别如下表所示。

	市盈率	市净率	净资产收益率	总资产周转率变动率	预期盈利增长修正	20 日涨幅	市值	公司标签（好 = 1/差 = 0）
上市公司 1	低	低	高	高	高	低	低	1
上市公司 2	低	高	高	低	高	高	高	1
上市公司 3	高	高	低	低	低	高	高	0
上市公司 4	低	低	高	高	高	高	高	1
上市公司 5	低	高	高	高	高	高	低	1
上市公司 6	高	低	低	低	低	低	高	0
上市公司 7	高	高	低	高	高	高	高	0
上市公司 8	低	低	高	高	低	低	高	1
上市公司 9	高	低	高	低	低	低	高	0
上市公司 10	低	高	高	低	低	高	低	0
上市公司 11	低	高	低	低	低	高	高	0
上市公司 12	低	低	高	高	高	高	低	1
上市公司 13	高	低	高	高	低	低	低	1
上市公司 14	高	高	低	低	低	低	高	0
上市公司 15	高	高	低	低	高	低	高	0
上市公司 16	低	低	高	低	低	高	高	0
上市公司 17	高	低	高	高	高	低	高	0
上市公司 18	高	低	低	低	高	低	高	0
上市公司 19	低	低	高	高	低	高	低	1
上市公司 20	高	低	高	低	高	低	低	1

使用这个训练集来估计所有的 $P(X_i = x_i|C = c_k)$ 和 $P(C = c_k)$ 的取值。通过计数（counting）以及拉普拉斯平滑就可以求出这些参数的估计量（见下表）。

概率	45.5 %	55.5 %

P(X_i=x_i | C=c_k)

	P(特征=高 | 公司=好)	P(特征=低 | 公司=好)	P(特征=高 | 公司=差)	P(特征=低 | 公司=差)
市盈率	27.3 %	72.7 %	69.2 %	30.8 %
市净率	27.3 %	72.7 %	53.8 %	46.2 %
净资产收益率	90.9 %	9.1 %	38.5 %	61.5 %
总资产周转率变动率	72.7 %	27.3 %	23.1 %	76.9 %
预期盈利增长修正	63.6 %	36.4 %	38.5 %	61.5 %
20 日涨幅	54.5 %	45.5 %	46.2 %	53.8 %
市值	36.4 %	63.6 %	84.6 %	15.4 %

使用这些估计量就可以对任意给定的新公司分类。比如对于某上市公司，它的特征分别为：市盈率低、市净率高、净资产收益率高、总资产周转率变动率高、预期盈利增长修正低、20 日涨幅高、市值低。使用朴素贝叶斯，对好公司和差公司这两类，分别计算 $P(C=c_k)\times\prod_i P(X_i=x_i|C=c_k)$ 的取值：

$$
\begin{aligned}
\text{好公司：}\quad & P(C=c_k)\times\textstyle\prod_i P(X_i=x_i|C=c_i)\\
= \quad & 0.455\times 0.727\times 0.273\times 0.909\times 0.727\times 0.364\times 0.545\times 0.636\\
= \quad & 0.00753
\end{aligned}
$$

$$
\begin{aligned}
\text{差公司：}\quad & P(C=c_k)\times\textstyle\prod_i P(X_i=x_i|C=c_i)\\
= \quad & 0.555\times 0.308\times 0.538\times 0.385\times 0.231\times 0.615\times 0.462\times 0.154\\
= \quad & 0.00036
\end{aligned}
$$

由于 0.00753 > 0.00036，因此朴素贝叶斯分类对该公司的分类结果是好公司。

6 结语 —— 朴素贝叶斯为什么靠谱？

由于条件独立性这个强假设的存在，朴素贝叶斯分类器十分简单。但是，**它仍然有非常不错的**
果。原因何在呢？人们在使用分类器之前，首先做的第一步（也是最重要的一步）往往是**特征**
（feature selection），这个过程的目的就是为了**排除特征之间的共线性、选择相对较为独立**
特征。其次，当我们假设特征之间相互独立时，这事实上就暗含了正则化的过程；而**不考虑变**

实际中往往能够取得非常优秀的结果。Hand and Yu (2001) 通过大量实际的数据表明了这一点。

最后，我们以 Wu et al. (2008) 中对朴素贝叶斯分类器的高度概括作为全文的收尾：

> *The naive Bayes model is tremendously appealing because of its simplicity, elegance, and robustness. It is one of the oldest formal classification algorithms, and yet even in its simplest form it is often surprisingly effective. It is widely used in areas such as text classification and spam filtering. A large number of modifications have been introduced, by the statistical, data mining, machine learning, and pattern recognition communities, in an attempt to make it more flexible, but one has to recognize that such modifications are necessarily complications, which detract from its basic simplicity.*
> *译：朴素贝叶斯模型因其简单、优雅和鲁棒性而极具吸引力。它是最古老的形式化分类算法之一，但即使是最简单的形式，它也常常令人惊讶地有效。它被广泛应用于文本分类和垃圾邮件过滤等领域。统计、数据挖掘、机器学习和模式识别领域的专家们通过对它进行了大量的修改，试图使其更加灵活。但人们必须认识到，这种修改必然是复杂的，这削弱了它的基本简单性。*

参考文献

- Hand, D. J. and K. Yu (2001). Idiot’ s Bayes – not so stupid after all? *International Statistical Review*, Vol. 69(3), 385 – 398.
- Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, Vol. 14(1), 1 – 37.

免责声明：文章内容不可视为投资意见。市场有风险，入市需谨慎。

编辑于 2019-07-02

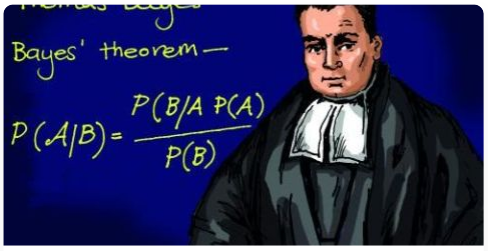
机器学习 贝叶斯统计 贝叶斯理论

知乎 首发于 川流不息

川流不息
北京量信投资管理有限公司是一家在中国基金业协会备案登记的专业私募基金管理人...

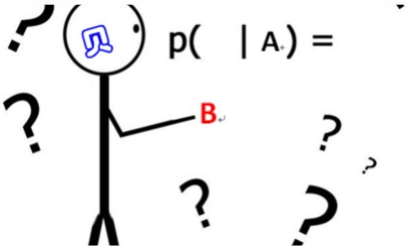
关注专栏

推荐阅读



贝叶斯网络，看完这篇我终于理解了(附代码)！

mantch



在故事中学“朴素贝叶斯（NB）分类算法”（03）

说白了

金融中的机

Coursera近
器学习的专
Learning a
Learning in
Specializati
和强化学习

Pan Y...

4 条评论

切换为时间排序

写下你的评论...

卷毛艺术家 1 年前

楼主你好，请教和个问题。设训练样本X属于{0,1}。假如的预测样本向量X=(1,0,1,1,...,1)_T，它的第二个元素X_2=0，如何计算#{X_2=0 ∧ C=c_k}？谢谢。

赞

石川 (作者) 回复 卷毛艺术家 1 年前

对于 c_k = 0，找训练集样本中所有 X_2 = 0 & C = 0 的样本点，然后统计一共多少个；对 c_k = 1 做同样的事儿不就得到 #{X_2=0 ∧ C=c_k} 了吗？还是我没有理解你的问题？

赞 回复 踩 举报

卷毛艺术家 回复 石川 (作者) 1 年前

理解了。谢谢楼主。

赞

知乎 首发于 川流不息

贝叶斯假设所有feature相互独立，但是实际应用并不能达到，请问如何优化？谢谢！

赞