第 6 章 朴素贝叶斯算法

目录

第 6 章 朴素贝叶斯

6.1 朴素贝叶斯算法

在前面几章内容中，笔者分别介绍了一种回归模型和两种分类模型以及模型的改善与泛化。在接下来的一章中，笔者将开始介绍下一个新的分类模型——朴素贝叶斯（Naive Bayes, NB）。那么什么又是朴素贝叶斯呢？从名字也可以看出，朴素贝叶斯算法与贝叶斯公式有着莫大的关联，说得简单点朴素贝叶斯就是由贝叶斯公式加“朴素”这一条件所构成。

在看贝叶斯算法的相关内容时，相信各位读者一定会被突如其来的数学概念搞得头昏脑胀。比如先验概率(Prior Probability)、后验概率(Posteriori Probability,) 极大似然估计(Maximum Likelihood Estimation)，极大后验概率估计(Maximum A Posteriori Estimation)，等。所以下面，笔者将先简单的介绍一下这几个概念，让读者先对这部分内容有一个感性的认识，然后再继续介绍后面的内容。

6.1.1 概念介绍

1) 先验概率

所谓先验概率指的就是根据历史经验得出来的概率。例如可以通过西瓜的颜色、敲的声音来判断其是否成熟。因为你已经有了通过颜色和声音来判断的“经验”，不管这个经验是你自己学会的还是别人告诉你的。又如在某 2 分类数据集中，其中正样本有 4 个，负样本有 6 个，那么通过这个数据集能够学习到的先验知识便是任取一个样本，其为正样本的可能性为 40%，为负样本的可能性为 60%。最后举个例子，假如办公室失窃了，理论上每个人都可能是小偷。但可以根据对每个人的了解分析得出一个可能性，比如张三偷窃的可能性为 20%，李四偷窃的可能性为 30%，王五偷窃的可能性为 50%，而这就被称之为先验概率，它是通过历史经验得来的。

2) 后验概率

所谓后验概率指的就是通过贝叶斯公式推断得到的结果。例如上述例子中，不可能因为负样本出现的可能性为 60%就判定任意取出的样本为负样本；也不能因为王五偷窃的可能性最大就判定每次办公室失窃都是他所为。先验知识只能帮助我们先取得一个大致的判断，而事实情况需要根据先验概率和条件概率来进行计算。

3) 极大后验概率估计

一言以蔽之，极大后验概率指的是在所有后验概率中选择其中最大的一个。例如上述例子中，根据先验概率和条件概率便可以计算出每个样本属于正样本还是负样本的后验概率。最后在判断该样本属于何种类别时，挑选后验概率最大的类别即可。

4) 极大似然估计

所谓极大似然估计（即最大似然估计）指的是用来估计使得当前已知结果最有可能发生的模型参数值（可以参见 3.4.3 节中的介绍）。例如上述例子中，已知的当前结果为正样本有 4 个，负样本有 6 个。那么什么样的模型参数能够使得这一结果最可能发生呢？此时只需要最大化式(6-1)即可。

$$\binom{10}{4} p^4(1-p)^6 \tag{6-1}$$

其中 p 为属于正样本的概率。

6.1.2 理解朴素贝叶斯

由贝叶斯公式可知

$$P(B|A)=\frac{P(AB)}{P(A)} \tag{6-2}$$

假设 B 为最终的分类标签， A 为一系列的特征属性，那么在使用朴素贝叶斯进行样本分类的时候，实际计算的应该就是每个样本在当前的特征取值为 A 的情况下，它属于类别 B 的概率。因此，进一步当计算出特征值 A 属于每个类别的概率后，再挑选概率值最大时所对应的类别即可作为该样本的分类。但是，在实际情况中对于 A,B 之间的联合概率分布 $P(AB)$ 是不知道的，说得直白点就是我们并不知道数据集的生成规则。但是，好在可通过先验概率分布 $P(A)$ 乘以条件概率分布 $P(B|A)$ 来得到联合分布，即将公式(6-2)转换为

$$P(B|A)=\frac{P(B)P(A|B)}{P(A)} \tag{6-3}$$

现在假设输入空间 $\mathcal{X}\subseteq R^n$ ，为 n 维向量的集合，输出空间为类标记 $\mathcal{Y}=\{c_1,c_2,...,c_K\}$ ，输入为特征向量 $x\in\mathcal{X}$ ，输出为类标记 $y\in\mathcal{Y}$ 。同时， X 是定义在输入空间 $\mathcal{X}$ 上的随机变量， Y 是定义在输出空间 $\mathcal{Y}$ 上的随机变量，也就是说 X 是一个 $m\times n$ 的矩阵， y 为类标签。 $P(X,Y)$ 是 X 和 Y 的联合概率分布，训练集 $T=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$ 由 $P(X,Y)$ 独立同分布产生。

根据上面的分析可知，可以通过学习数据的先验分布，再学习数据的条件概率分布，即可得到联合概率分布 $P(X,Y)$ 。具体地，对于每个类别来说其先验概率分布为

$$P(Y=c_k)=\frac{\#c_k}{m},k=1,2,...,K \tag{6-4}$$

其中 $\#c_k$ 表示该类别一共有多少个样本， m 表示样本总数。

同时，对于已知类标下的条件概率分布为

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},...,X^{(n)}=x^{(n)}|Y=c_k) \tag{6-5}$$

其中 $x^{(i)}$ 表示第 i 个特征的取值。

从式(6-5)可知，在实际情况中想要知道其条件概率是不能的。因此朴素贝叶斯对条件概率分布又做了条件独立性假设，即 $P(AB|D)=P(A|D)P(B|D)$ ，而这也是“朴素”一词的由来。故式(6-5)可改写为

$$P(X=x|Y=c_k)=\prod_{i=1}^{n}P(X^{(i)}=x^{(i)}|Y=c_k) \tag{6-6}$$

由此，根据公式(6-3)的分析可知，对于已知特征属性 $X=x$ 的条件下，其属于类别 $Y=c_k$ 的后验概率为

$$P(Y=c_k \mid X=x)=\frac{P(X=x \mid Y=c_k)P(Y=c_k)}{\sum_{k=1}^{K}P(X=x \mid Y=c_k)P(Y=c_k)} \tag{6-7}$$

进一步，将式(6-6)代入(6-7)可得

$$P(Y=c_k \mid X=x)=\frac{P(Y=c_k)\prod_{i=1}^{n}P(X^{(i)}=x^{(i)} \mid Y=c_k)}{\sum_{k=1}^{K}P(Y=c_k)\prod_{i=1}^{n}P(X^{(i)}=x^{(i)} \mid Y=c_k)} \tag{6-8}$$

于是，朴素贝叶斯分类器可以表示为

$$y=\arg\max_{c_k}=\frac{P(Y=c_k)\prod_{i=1}^{n}P(X^{(i)}=x^{(i)} \mid Y=c_k)}{\sum_{k=1}^{K}P(Y=c_k)\prod_{i=1}^{n}P(X^{(i)}=x^{(i)} \mid Y=c_k)} \tag{6-9}$$

即通过计算出任意样本属于类别c_k的概率后，选择其中概率最大者作为其分类的类标。但是，我们发现在式(6-9)中，对于每个样本的后验概率的计算来说，其都有相同的分母。因此，式(6-9)可进一步简化为

$$y=\arg\max_{c_k} P(Y=c_k)\prod_{i=1}^{n}P(X^{(i)}=x^{(i)} \mid Y=c_k) \tag{6-10}$$

注意：$\arg\max_{c_k}$ 的含义是，使得 y 取最大值时 c_k 的取值

虽然朴素贝叶斯算法看似做了一个及其简单的假设，但是其在实际的运用过程中却都有着不错的效果，尤其是在文档分类和垃圾邮件分类场景下仅需要少量数据集就能获得不错的效果[①]。

6.1.3 计算示例

通过 6.1.2 节内容的介绍， 朴素贝叶斯算法的整个原理过程就算是介绍完了。下面再来通过一个实际的计算示例来体会一下朴素贝叶斯分类器的计算流程。

如表 6-1 所示为一个信用卡审批数据集，其中 $X^{(1)}\in A_1=\{0,1\}$ 表示有无工作，$X^{(2)}\in A_2=\{0,1\}$ 表示是否有房，$X^{(3)}\in A_3=\{D,S,T\}$ 表示学历等级，$Y\in C=\{0,1\}$ 表示是否审批通过的类标记。试由表 6-1 中的训练集学习一个朴素贝叶斯分类器，并确定 $x=(0,1,D)$ 的类标记 Y。

表 6-1 示例计算数据

样本	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
$X^{(2)}$	1	1	1	0	1	0	0	0	0	0	1	1	1	0	0
$X^{(3)}$	T	S	S	T	T	T	D	T	T	D	D	T	T	S	S
Y	1	1	1	0	0	0	0	1	1	1	1	1	1	0	1

根据式(6-4)，由表 6-1 易知，各个类别的先验概率为

① Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

$$P(Y=0)=\frac{5}{15},\ P(Y=1)=\frac{10}{15} \tag{6-11}$$

条件概率为

$$\begin{aligned}
&P(X^{(1)}=0|Y=0)=\frac{4}{5},P(X^{(1)}=1|Y=0)=\frac{1}{5}\\
&P(X^{(2)}=0|Y=0)=\frac{4}{5},P(X^{(2)}=1|Y=0)=\frac{1}{5}\\
&P(X^{(3)}=D|Y=0)=\frac{1}{5},P(X^{(3)}=S|Y=0)=\frac{1}{5}\\
&P(X^{(3)}=T|Y=0)=\frac{3}{5},P(X^{(1)}=0|Y=1)=\frac{3}{10}\\
&P(X^{(1)}=1|Y=1)=\frac{7}{10},P(X^{(2)}=0|Y=1)=\frac{4}{10}\\
&P(X^{(2)}=1|Y=1)=\frac{6}{10},P(X^{(3)}=D|Y=1)=\frac{2}{10}\\
&P(X^{(3)}=S|Y=1)=\frac{3}{10},P(X^{(3)}=T|Y=1)=\frac{5}{10}
\end{aligned} \tag{6-12}$$

以上计算过程便是根据训练集训练朴素贝叶斯分类器模型参数的过程。根据这些参数，便可以对给定的 $x=(0,1,D)$ 进行预测。

首先根据式(6-10)分别计算出其属于各个类别的后验概率为

$$\begin{aligned}
&P(Y=0|X=x)\\
&=P(Y=0)\cdot P(X^{(1)}=0|Y=0)\cdot P(X^{(2)}=1|Y=0)\cdot P(X^{(3)}=D|Y=0)\\
&=\frac{5}{15}\cdot\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}=\frac{4}{375}
\end{aligned} \tag{6-13}$$

$$\begin{aligned}
&P(Y=1|X=x)\\
&=P(Y=1)\cdot P(X^{(1)}=0|Y=1)\cdot P(X^{(2)}=1|Y=1)\cdot P(X^{(3)}=D|Y=1)\\
&=\frac{10}{15}\cdot\frac{3}{10}\cdot\frac{6}{10}\cdot\frac{2}{10}=\frac{3}{125}
\end{aligned} \tag{6-14}$$

于是便可以知道，样本 $x=(0,1,D)$ 属于 $y=1$ 的可能性最大。

6.1.4 求解步骤

根据上面两节的介绍，可以将朴素贝叶斯分类算法的求解步骤总结如下：

输入：训练数据 $T=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$，其中 $x_i=(x_i^{(1)},x_i^{(2)},...,x_i^{(n)})^T$，$x_i^{(j)}$ 是第 i 个样本的第 j 维特征，$x_i^{(j)}\in\{a_{j1},a_{j2},...,a_{jS_j}\}$，$a_{jl}$ 是第 j 维特征可能取得的第 l 个值。同时，$j=1,2,...,n$，$l=1,2,...,S_j$，$y_i\in\{c_1,c_2,...,c_K\}$；以及实例 x；

输出：实例 x 的分类[①]。

1) 计算先验概率与条件概率

根据极大似然估计，用给定的数据集来计算各类别的先验概率和条件概率。

[①]李航，统计机器学习，清华大学出版社

$$P(Y=c_k)=\frac{\sum_{i=1}^{m}I(y_i=c_k)}{m},k=1,2,...,K \tag{6-15}$$

$$P(X^{(j)}=a_{jl}\,|\,Y=c_k)=\frac{\sum_{i=1}^{m}I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{m}I(y_i=c_k)} \tag{6-16}$$

$$j=1,2,...,n;\ l=1,2,...,S_j;\ k=1,2,...,K$$

其中 $I(\cdot)$ 指示函数，当 $y_i=c_k$ 时输出值为 1，反之则为 0。

2) 计算各特征取值下的后验概率

$$P(Y=c_k)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\,|\,Y=c_k),k=1,2,...,K \tag{6-17}$$

3) 极大化后验概率确定类别

$$y=\arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\,|\,Y=c_k) \tag{6-18}$$

到此，对于朴素贝叶斯算法的原理及计算过程就介绍完了。根据 6.1.4 节的介绍可以知道，朴素贝叶斯算法所接受的特征输入都是离散型特征（Discrete Features），也就非连续性的特征取值，例如基于词袋模型的文本特征表示等。因此，对于这部分的示例代码将放在第 7 章中进行介绍。

6.1.5 小结

在本节中，笔者首先介绍了朴素贝叶斯算法中的几个基本概念；然后详细介绍了朴素贝叶斯算法的原理，知道了“朴素”一词的含义以及为什么可以通过贝叶斯算法来完成分类任务；最后对朴素贝叶斯算法的具体计算流程进行了总结。

6.2 贝叶斯估计

在介绍完 6.1 节中的内容后算是对朴素贝叶斯算法的原理有了清楚的认识，但还有一个不能忽略的问题就是，当训练集不充分的情况下，某个维度的条件概率缺失时该怎么处理。例如在 6.1.3 节的示例中，如果条件概率 $P(X^{(3)}=D\,|\,Y=1)=0$，即训练集中不存在这一情况，而在测试的数据样本中却存在这种情况。如果此时仍旧将这种情况下的条件概率看作是是 0，那么在预测的时候将会产生很大的错差。面对这样的情况该怎么办呢？

6.2.1 平滑处理

通常，解决这类问题的一个有效办法就是在各个估计中加入一个平滑项（Smoothing Parameter）。那么，此时先验概率和条件概率的计算方法为

$$P_\lambda(Y=c_k)=\frac{\sum_{i=1}^{m}I(y_i=c_k)+\lambda}{m+K\lambda} \tag{6-19}$$

$$P_\lambda(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{m}I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{m}I(y_i=c_k)+S_j\lambda} \tag{6-20}$$

其中 K 表示数据集分类的类别数；S_j 表示第 j 维特征的取值情况数；$\lambda\geq 0$，且当 $\lambda=1$ 时称为拉普拉斯平滑（Laplace Smoothing），这也是常用的做法。

同时，当 $\lambda>0$ 时分别称式(6-19)(6-20)为先验概率和条件概率的贝叶斯估计。并且可以发现，当 $\lambda=0$ 时，就是极大似然估计；

6.2.2 计算示例

接下来，将 6.1.3 节中的数据用拉普拉斯平滑（$\lambda=1$）再来计算一次。在计算之前我们知道，此时类别数 $K=2$，$S_1=2,S_2=2,S_3=3$。

根据式表 6-1 和式(6-19)易知，各类别的先验概率分别为

$$P(Y=0)=\frac{6}{17},\ P(Y=1)=\frac{11}{17} \tag{6-21}$$

条件概率为

$$\begin{aligned}
&P(X^{(1)}=0\mid Y=0)=\frac{5}{7},P(X^{(1)}=1\mid Y=0)=\frac{2}{7}\\
&P(X^{(2)}=0\mid Y=0)=\frac{5}{7},P(X^{(2)}=1\mid Y=0)=\frac{2}{7}\\
&P(X^{(3)}=D\mid Y=0)=\frac{2}{8},P(X^{(3)}=S\mid Y=0)=\frac{2}{8}\\
&P(X^{(3)}=T\mid Y=0)=\frac{4}{8},P(X^{(1)}=0\mid Y=1)=\frac{4}{12}\\
&P(X^{(1)}=1\mid Y=1)=\frac{8}{12},P(X^{(2)}=0\mid Y=1)=\frac{5}{12}\\
&P(X^{(2)}=1\mid Y=1)=\frac{7}{12},P(X^{(3)}=D\mid Y=1)=\frac{3}{13}\\
&P(X^{(3)}=S\mid Y=1)=\frac{4}{13},P(X^{(3)}=T\mid Y=1)=\frac{6}{13}
\end{aligned} \tag{6-22}$$

计算出属于各个类别的后验概率为

$$\begin{aligned}
&P(Y=0\mid X=x)\\
&=P(Y=0)\cdot P(X^{(1)}=0\mid Y=0)\cdot P(X^{(2)}=1\mid Y=0)\cdot P(X^{(3)}=D\mid Y=0)\\
&=\frac{6}{17}\cdot\frac{5}{7}\cdot\frac{2}{7}\cdot\frac{2}{8}\approx 0.02
\end{aligned} \tag{6-23}$$

$$\begin{aligned}
&P(Y=1\mid X=x)\\
&=P(Y=1)\cdot P(X^{(1)}=0\mid Y=1)\cdot P(X^{(2)}=1\mid Y=1)\cdot P(X^{(3)}=D\mid Y=1)\\
&=\frac{11}{17}\cdot\frac{4}{12}\cdot\frac{7}{12}\cdot\frac{3}{13}\approx 0.03
\end{aligned} \tag{6-24}$$

于是我们同样可以得出，样本 $x=(0,1,D)$ 属于 $y=1$ 的可能性最大。

到此，对于朴素贝叶斯算法的原理及计算过程就介绍完了。对于这部分的 sklearn 示例代码也将在第 7 章中进行介绍。由于在不同的书中对于一些算法原理有着不同的称谓，这也导致读者在初学翻阅各种资料时候发现一会儿又多了这个概念，一会儿又多了那个概念极为苦恼。不过名称并不太重要，重要的是要知道具体指代的具体东西。如图 6-1 所示是笔者对遇到的各种“叫法”进行的总结，仅供参考。

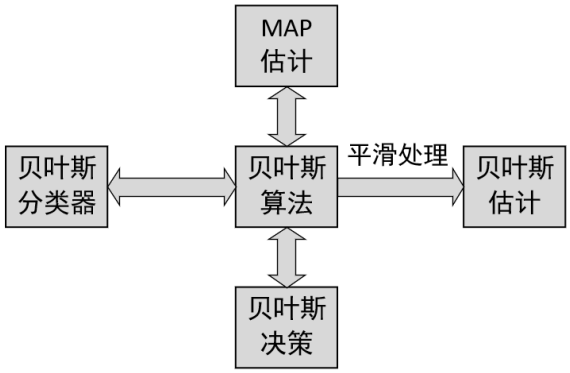

图 6-1 概念辨析图

6.2.3 小结

在本节中，笔者介绍了如何处理在贝叶斯算法中条件概率为 0 时的处理方法，即贝叶斯估计；然后也辨析了几个在贝叶斯算法中容易混淆的概念。值得一提的是，其实平滑处理这种做法不仅仅可以用于此处，在其它任何类似的情况中都可以借鉴这种做法。例如在下一章将要介绍的 TF-IDF 中同样也会用到。亦或是编写含有除运算的程序中，为了防止分母出现零的情况，都可以采用这样的做法。

总结一下，在本章中笔者首先介绍了朴素贝叶斯算法中的几个基本概念，包括先验概率、后验概率、极大后验概率和极大似然估计等，因为只有在对这些概率有了感性的认识才更加有利于对后续算法原理的理解；接着笔者介绍了朴素贝叶斯算法的基本原理，并且还以一个真实的示例对整个算法计算过程进行了演示；然后介绍了以平滑处理的方式来处理贝叶斯算法中可能存在的条件概率为 0 的情况，即贝叶斯估计；最后还对贝叶斯算法中几种常见的算法名称进行了总结。

本次内容就到此结束，感谢您的阅读！如果你觉得上述内容对你有所帮助，欢迎分享至一位你的朋友！若有任何疑问与建议，请添加笔者微信'nulls8'或加群进行交流。**青山不改，绿水长流，我们月来客栈见！**

扫码关注@月来客栈可获得更多优质内容！

代码仓库： https://github.com/moon-hotel/MachineLearningWithMe

2021年

第一章：机器学习环境安装　　PDF内容

Python版本为3.6，各个Python包版本见`requirements.txt`，使用如下命令即可安装：

```
pip install -r requirements.txt
```

第二章：从零认识线性回归　　代码　　PDF内容

第三章：从零认识逻辑回归　　代码　　PDF内容

第四章：模型的改善与泛化　　代码　　PDF内容

第五章：K 近邻算法与原理　　代码　　PDF内容

第六章：朴素贝叶斯算法

第七章：文本特征提取与模型复用

第八章：决策树与集成模型

第九章：支持向量机

第十章：聚类算法