朴素贝叶斯法

目录
Contents

01 朴素贝叶斯法的学习与分类

- 概念回顾
- 基本方法
- 后验概率最大化的含义

朴素贝叶斯的思想：对于给出的待分类项X，求出在此项X出现的条件下属于各个类别的概率，选出概率最大的类别，就认为此待分类项X属于这个类别。

朴素贝叶斯(*Naive Bayes*)法是基于贝叶斯定理与特征条件假设独立的分类方法。

1、联合概率

假设有随机变量X和Y，此时P(X=a,Y=b)用于表示X=a且Y=b同时发生的概率。这类包含多个条件且所有条件同时成立的概率称为**联合概率**。

2、边缘概率

P(X=a)或P(Y=b)这类仅与单个随机变量有关的概率称为**边缘概率**。

联合概率和边缘概率的联系：

$$P(X=a)=\sum_{i=1}^{k}P(X=a,Y=b_i)$$

3、条件概率

条件概率表示在事件B成立的情况下，事件A的概率，记作P(A|B)，或者说条件概率是指事件A在另外一个事件B已经发生条件下的发生概率。

联合概率、边缘概率和条件概率的关系：

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

4、全概率公式

如果事件A1，A2，A3，…，An构成一个完备事件组，即它们两两互不相容（互斥），其和为全集；并且P(Ai)大于0，则对任意事件B有

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n)$$

$$= \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

5、独立性

设A，B是两个事件，如果事件A，B独立，则

$$P(AB) = P(A) * P(B)$$

6、贝叶斯公式

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

- P(A)是A的**先验概率或边缘概率**，表示事件A发生的置信度。
- P(B|A)是已知A发生后B的**条件概率**。
- P(B)是B的**先验概率或边缘概率**。右边分母使用全概率公式展开。

训练数据集 $$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

其中，输入的特征向量 x_i 的维度是n，y_i的取值是$\{c_1,c_2,\ldots,c_K\}$

接着，朴素贝叶斯法通过训练数据集学习

- **先验概率分布** $$P(Y=c_k),\quad k=1,2,\cdots,K \tag{4.1}$$
- **条件概率分布**

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\ldots,X^{(n)}=x^{(n)}|Y=c_k),\quad k=1,2,\ldots,K \tag{4.2}$$

假设 $x^{(j)}$可取值有S_j个，j=1,2,...,n。Y可取值有K个，那么参数个数为 $K\prod_{j=1}^{n}S_j$

参数个数是指数量级，不可行，需要减少参数。

特征条件独立假设：输入特征 x 在类的确定条件下，各个维度互相独立。

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$$

$$=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k) \tag{4.3}$$

$$=P(X^{(1)}=x^{(1)}|Y=c_k)*P(X^{(2)}=x^{(2)}|Y=c_k)$$
$$*\ldots*P(X^{(n)}=x^{(n)}|Y=c_k)$$

参数个数从 $K\prod_{j=1}^{n}S_j$ 降为 $K\sum_{j=1}^{n}S_j$

贝叶斯定理计算后验概率

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)} \tag{4.4}$$

将式 (4.3) 代入式 (4.4)，有

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)},\quad k=1,2,\cdots,K \tag{4.5}$$

这是朴素贝叶斯法分类的基本公式。于是，朴素贝叶斯分类器可表示为

$$y=f(x)=\arg\max_{c_k}\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)} \tag{4.6}$$

朴素贝叶斯分类器

$$y = f(x) = \arg\max_{c_k} \frac{P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)}{\sum_k P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)} \tag{4.6}$$

在式 (4.6) 中分母对所有 c_k 都是相同的，所以，

$$y = \arg\max_{c_k} P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k) \tag{4.7}$$

朴素贝叶斯分类器计算出X=x的条件下的后验概率 $P(Y = c_k|X = x)$

朴素贝叶斯分类器是认为后验概率最大的所对应的类别就是输入特征x的分类。

1、选择0-1损失函数。

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

- Y:输入样本的标签
- f(X):分类器的预测X所属类别

2、期望风险函数

$$R_{exp}(f) = E[L(Y, f(X))]$$

3、条件期望

$$E[g(X_1) \mid X_2 = x_2] = \sum_{x_1} g(x_1) f(x_1 \mid x_2)$$

期望风险可以写为：

$$R_{exp}(f) = E[L(Y, f(X))] = E_x[L(c_k, f(X))|X] = E_x \sum_{k=1}^{K}[L(c_k, f(X))]P(c_k|X)$$

$$= L(c_1, f(X))P(c_1|X) + L(c_2, f(X))P(c_2, |X) + ... + L(c_K, f(X))P(c_K, |X)$$

举例：假设分类器预测X的类别为c_1,即$f(X)=c_1$。

$$E(f = c_1) = 0 * P(c_1|X) + 1 * P(c_2|X) + ... + 1 * P(c_K|X) = P(c_2|X) + ... + P(c_K|X)$$

$$= 1 - P(c_1|X)$$

3、条件期望

由上面这个例子，可以推广至更一般的结果：

$$E(f = c_k) = 1 - P(c_k|X)$$

我们想要条件期望最小化，就要选择max{$P(c_k|X)$}

这样一来，根据期望风险最小化准则就得到了后验概率最大化准则：

$$f(x) = \arg\max_{c_k} P(c_k|X = x)$$

即朴素贝叶斯法所采用的原理。

02 朴素贝叶斯法的参数估计

- 极大似然估计
- 贝叶斯估计

朴素贝叶斯法学习意味着要估计 $P(y=c_k)$ 和 $P(X^{(j)}=x^{(j)}|y=c_k)$

极大似然估计(MLE)：就是利用已知的样本结果信息，反推具有最大有可能（即最大概率）导致这些样本结果出现的模型参数值。

数学描述：假设$x_1,x_2,\ldots,x_n$是一组独立等分布的抽样，那么MLE对参数θ的估计如下：

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MLE}} &= \arg\max P(X;\theta) \\
&= \arg\max P(x_1;\theta)P(x_2;\theta)\cdots\cdot P(x_n;\theta) \\
&= \arg\max \log \prod_{i=1}^{n} P(x_i;\theta) \\
&= \arg\max \sum_{i=1}^{n} \log P(x_i;\theta)
\end{aligned}
$$

例子：盒中有黑白两种颜色的球，球的数目不知道，颜色比例未知。每次从盒中那一个出来记录颜色，再放入盒中摇匀。这个过程重复100次，假设有70次是白球，问白球所占比例最有可能是多少？

取球可以表示为参数θ的伯努利分布，其中$x_i=1$表示取出的是白球

$$P(x_i;\theta)=\begin{cases}\theta & x_i=1\\ 1-\theta & x_i=0\end{cases}=\theta^{x_i}(1-\theta)^{1-x_i}$$

似然函数可以这样表示：我们取了n次

$$\mathrm{NLL}=-\sum_{i=1}^{n}\log P(x_i;\theta)=-\sum_{i=1}^{n}\log\theta^{x_i}(1-\theta)^{1-x_i}$$

为了求得最大值，似然函数对参数θ求导：

$$\mathrm{NLL}' = -\sum_{i=1}^{n}\left(\frac{x_i}{\theta} + (1-x_i)\frac{-1}{1-\theta}\right) = 0$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$

白球的比例等于取出白球的次数除以总次数。

所以先验概率$P(Y=c_k)$的极大似然估计是：

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N},\quad k=1,2,\cdots,K \tag{4.8}$$

条件概率：设第 j 个特征 $x^{(j)}$ 可能取值的集合为 $\{a_{j1},a_{j2},\cdots,a_{jS_j}\}$，条件概率 $P(X^{(j)}=a_{jl}|Y=c_k)$ 的极大似然估计是

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

$$j=1,2,\cdots,n;\quad l=1,2,\cdots,S_j;\quad k=1,2,\cdots,K \tag{4.9}$$

式中，$x_i^{(j)}$ 是第 i 个样本的第 j 个特征；a_{jl} 是第 j 个特征可能取的第 l 个值；I 为指示函数。

极大似然估计有可能出现分子为0的情况。于是极大似然估计基础上引入参数λ(λ>=0). λ=0,是极大似然估计；λ=1，拉普拉斯平滑。

先验概率：

$$P_\lambda(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+\lambda}{N+K\lambda}$$

条件概率：

$$P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda} \tag{4.10}$$

02 朴素贝叶斯法的参数估计
part2 贝叶斯估计

表 4.1 训练数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	−1	−1	1	1	−1	−1	−1	1	1	1	1	1	1	1	−1

例 4.2 问题同例 4.1，按照拉普拉斯平滑估计概率，即取 $\lambda = 1$。

解 $A_1 = \{1, 2, 3\}$，$A_2 = \{S, M, L\}$，$C = \{1, -1\}$。按照式 (4.10) 和式 (4.11) 计算下列概率：

$$P(Y=1) = \frac{10}{17}, \quad P(Y=-1) = \frac{7}{17} \quad = \ (6+1) \ / \ (15+2*1) \qquad P_\lambda(Y=c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k) + \lambda}{N + K\lambda}$$

$$P(X^{(1)}=1|Y=1) = \frac{3}{12}, \quad P(X^{(1)}=2|Y=1) = \frac{4}{12}, \quad P(X^{(1)}=3|Y=1) = \frac{5}{12} = (4+1)/(9+3*1)$$

对于给定的 $x = (2, S)^{\mathrm{T}}$，计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1) = \frac{10}{17} \cdot \frac{4}{12} \cdot \frac{2}{12} = \frac{5}{153} = 0.0327$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1) = \frac{7}{17} \cdot \frac{3}{9} \cdot \frac{4}{9} = \frac{28}{459} = 0.0610$$

$$P_\lambda(X^{(j)}=a_{jl}|Y=c_k) = \frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k) + \lambda}{\sum_{i=1}^{N} I(y_i=c_k) + S_j\lambda}$$

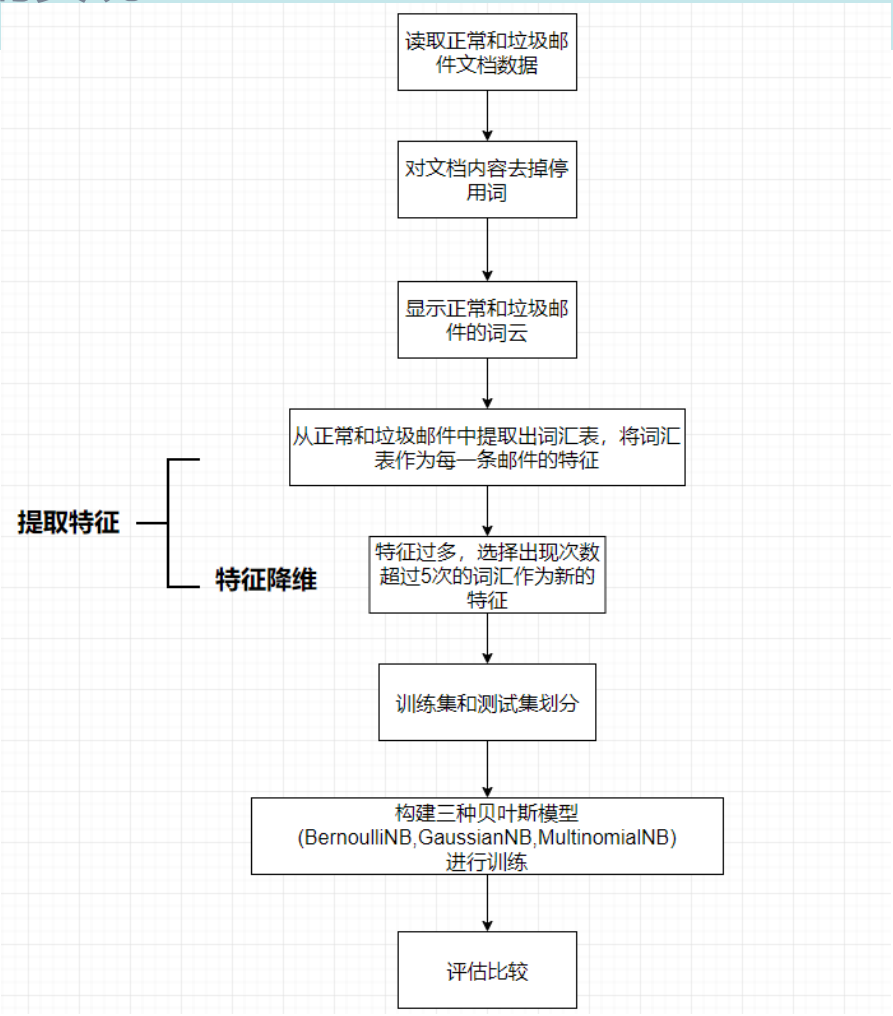
由于 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$。

03 朴素贝叶斯的实现

- 实现流程
- 显示数据集
- 划分数据集和构建模型
- 评估比较

03 朴素贝叶斯法的实现

part1 实现流程



03 朴素贝叶斯法的实现
part2 显示数据集

正常邮件词云(选取200个)

垃圾邮件词云(选取200个)

03 朴素贝叶斯法的实现
part3 划分数据集和构建模型

```
'''
step 5:划分训练集和测试集
'''
train,test,trainlabel,testlabel = train_test_split(textMatrix,labels,test_size=0.3, random_state=5)

'''
step6:构建三种模型 , 并且评估比较
'''
# 伯努利模型
clf = BernoulliNB()
clfmodel = clf.fit(train,trainlabel)
clfpred = clfmodel.predict(test)
print('BernoulliNB的准确率、召回率和f1值')
print(classification_report(testlabel, clfpred))
print('BernoulliNB模型测试集的准去率: ',end='')
print(clfmodel.score(test,testlabel))
```

	aa	abnewkoko	access	action	address	admin	...	默契	默默	鼓励	齐全	齐联	龙信
4634	0	0	0	0	0	0	...	0	0	0	0	0	0
416	0	0	0	0	0	0	...	0	0	0	0	0	0
6369	0	0	0	0	0	0	...	0	0	0	0	0	0
7952	0	0	0	0	0	0	...	0	0	0	0	0	0
4066	0	0	0	0	0	0	...	0	0	0	0	0	0
...	..	...	...	...	...	...	...	..	..	..	..	..	..
3046	0	0	0	0	0	0	...	0	0	0	0	0	0
9917	0	0	0	0	0	0	...	0	0	0	0	0	0
4079	0	0	0	0	0	0	...	0	0	0	0	0	0
2254	0	0	0	0	0	0	...	0	0	0	0	0	0
2915	0	0	0	0	0	0	...	0	0	0	0	0	0

训练数据集：每一行表示一条邮件，每一列表示一个特征，值代表特征在这条特征中出现的次数

03 朴素贝叶斯法的实现
part4 评估比较

BernoulliNB的准确率、召回率和f1值

	precision	recall	f1-score	support
0.0	0.98	0.99	0.99	1483
1.0	0.99	0.98	0.99	1518
accuracy			0.99	3001
macro avg	0.99	0.99	0.99	3001
weighted avg	0.99	0.99	0.99	3001

BernoulliNB模型测试集的准去率: 0.9873375541486171

MultinomialNB的准确率、召回率和f1值

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	1483
1.0	0.99	0.99	0.99	1518
accuracy			0.99	3001
macro avg	0.99	0.99	0.99	3001
weighted avg	0.99	0.99	0.99	3001

MultinomialNB模型测试集的准去率: 0.9893368877040987

GaussianNB的准确率、召回率和f1值

	precision	recall	f1-score	support
0.0	0.99	0.98	0.99	1483
1.0	0.98	0.99	0.99	1518
accuracy			0.99	3001
macro avg	0.99	0.99	0.99	3001
weighted avg	0.99	0.99	0.99	3001

GaussianNB模型测试集的准去率: 0.988003998667111

- support：某类别在测试数据中的样本个数；
- precision：模型预测的结果中有多少是预测正确的；
- macro avg：每个类别评估指标未加权的平均值，比如伯努利模型准确率的 macro avg：(0.98+0.99)/2=0.99
- weighted avg：加权平均

Thank you