第5章

朴素贝叶斯分类器

本章目标

- 理解极大似然估计；
- 理解并掌握朴素贝叶斯分类；
- 了解拉普拉斯平滑；
- 了解朴素贝叶斯分类器和极大似然估计之间的联系；
- 实现简单的朴素贝叶斯分类器完成垃圾信息分类问题。

朴素贝叶斯分类器是一种有监督的统计学过滤器，在垃圾邮件过滤、信息检索等领域十分常用。通过本章的介绍，读者将会了解朴素贝叶斯分类器因何得名、其与贝叶斯公式的联系，以及其与极大似然估计的关系。

5.1 极大似然估计

对于工厂生产的某一批灯泡，质检部门希望检测其合格率。设 m 表示产品总数，随机变量 $X_i \in \{0,1\}$ 表示编号为 i 的产品是否合格。由于这些产品都是同一批生产的，不妨假设：

$$X_1, X_2, \cdots, X_m \overset{i.i.d.}{\sim} \text{Bern}(p) \tag{5-1}$$

其中，p 表示产品合格的概率，也就是质检部门希望得到的数据。根据经典概率模型有

$$p \approx \frac{1}{m}\sum_{i=1}^{m} X_i \tag{5-2}$$

但是式(5-2)为什么成立？这就需要使用极大似然估计来证明了。

极大似然估计的思想是：找到这样一个参数 p，它使所有随机变量的联合概率最大。例中，联合概率表示为

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) = \prod_{i=1}^{m} P(X_i = x_i) = \prod_{i=1}^{m} p^{x_i}(1-p)^{1-x_i} \tag{5-3}$$

最大化联合概率等价于求

$$\begin{aligned} p^* &= \underset{p}{\operatorname{argmax}} \log \prod_{i=1}^{m} p^{x_i}(1-p)^{1-x_i} \\ &= \underset{p}{\operatorname{argmax}} \sum_{i=1}^{m} (x_i \log p + (1-x_i)\log(1-p)) \\ &= \underset{p}{\operatorname{argmax}}\, m\log(1-p) + \log\frac{p}{1-p}\sum_{i=1}^{m} x_i \end{aligned} \tag{5-4}$$

根据微积分知识容易证明式(5-2)：

$$p^* = \frac{1}{m}\sum_{i=1}^{m} X_i \tag{5-5}$$

形式化地说，已知整体的概率分布模型 $f(x;\theta)$，但是模型的参数 θ 未知时，可以使用极大似然估计来估计 θ 的值。这里的概率分布模型既可以是连续的(概率密度函数)也可以是离散的(概率质量函数)。假设在一次随机实验中，我们独立同分布地抽到了 m 个样本 x_1，x_2，…，x_m 组成的样本集合。似然函数，也就是联合概率分布：

$$L(\theta) = f_m(x_1, x_2, \cdots, x_m; \theta) = \prod_{i=1}^{m} f(x_i;\theta) \tag{5-6}$$

表示当前样本集合出现的可能性。令似然函数 $L(\theta)$对参数 θ 的导数为 0，可以得到 θ 的最优解。但是运算中涉及乘法运算及乘法的求导等，往往计算上存在不便性。而对似然函数取对数并不影响似然函数的单调性，即

$$L(\theta_1) > L(\theta_2) \Rightarrow \log L(\theta_1) > \log L(\theta_2) \tag{5-7}$$

所以最大化对数似然函数：

$$l(\theta) = \log L(\theta) = \log \prod_{i=1}^{m} p(x_i;\theta) = \sum_{i=1}^{m} \log(p(x_i;\theta)) \tag{5-8}$$

可以在保证最优解与似然函数相同的条件下，大大减少计算量。

极大似然估计通过求解参数 θ 使得 $f_N(x_1, x_2, \cdots, x_N; \theta)$最大，这是一种很朴素的思想：既然从总体中随机抽样得到了当前样本集合，那么当前样本集合出现的可能性极大。

5.2 朴素贝叶斯分类

在概率论中，贝叶斯公式的描述如下：

$$P(Y_i \mid X) = \frac{P(X, Y_i)}{P(X)} = \frac{P(Y_i)P(X \mid Y_i)}{\sum_{j=1}^{K} P(Y_j)P(X \mid Y_j)} \tag{5-9}$$

其中 $Y_1, Y_2, \cdots, Y_K$ 为一个完备事件组，$P(Y_i)$称为先验概率，$P(Y_i \mid X)$称为后验概率。设 $X=(X^1, X^2, \cdots, X^n)$表示 n 维(离散)样本特征，$Y \in \{c_1, c_2, \cdots, c_K\}$表示样本类别。由于一个样本只能属于这 K 个类别中的一个，所以 $Y_1, Y_2, \cdots, Y_K$ 一定是完备的。

给定样本集合 $D=\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$，我们希望估计 $P(Y \mid X)$。根据贝叶斯公式，对于任意样本 $x=(x^1, x^2, \cdots, x^n)$，其标签为 c_k 的概率为

$$P(Y=c_k \mid X=x)=\frac{P(Y=c_k)P(X=x \mid Y=c_k)}{P(X=x)} \tag{5-10}$$

假设随机变量 $X^1,X^2,\cdots,X^n$ 相互独立，则有

$$\begin{aligned}P(X=x \mid Y=c_k)&=P(X^1=x^1,X^2=x^2,\cdots,X^n=x^n \mid Y=c_k)\\&=\prod_{i=1}^{n}P(X^i=x^i \mid Y=c_k)\end{aligned} \tag{5-11}$$

代入式(5-10)得

$$P(Y=c_k \mid X=x)=\frac{P(Y=c_k)\prod_{i=1}^{n}P(X^i=x^i \mid Y=c_k)}{P(X=x)} \tag{5-12}$$

在实际进行分类任务时，不需要计算出 $P(Y|X)$ 的精确值，只需要求出 k^* 即可。

$$k^*=\operatorname{argmax}_k P(Y=c_k \mid X=x) \tag{5-13}$$

不难看出，式(5-12)右侧的分母部分与 k 无关。因此

$$k^*=\operatorname{argmax}_k P(Y=c_k)\prod_{i=1}^{n}P(X^i=x^i \mid Y=c_k) \tag{5-14}$$

式中所有项都可以用频率代替概率在样本集合上进行估计：

$$\begin{cases}P(Y=c_k)\approx\dfrac{N_k}{m}\\[2ex]P(X^i=x^i \mid Y=c_k)\approx\dfrac{\sum_{j=1}^{m}I\{x_j^i=x^i,y_j=c_k\}}{N_k}\end{cases} \tag{5-15}$$

其中，N_k 表示 D 中标签为 c_k 的样本数量。

5.3　拉普拉斯平滑

当样本集合不够大时，可能无法覆盖特征的所有可能取值。也就是说，可能存在某个 c_k 和 x^i 使

$$P(X^i=x^i \mid Y=c_k)=0 \tag{5-16}$$

此时，无论其他特征分量的取值为何，都一定有

$$P(Y=c_k)\prod_{i=1}^{n}P(X^i=x^i \mid Y=c_k)=0 \tag{5-17}$$

为了避免这样的问题，实际应用中常采用平滑处理。典型的平滑处理就是拉普拉斯平滑：

$$\begin{cases}P(Y=c_k)\approx\dfrac{N_k+1}{m+K}\\[2ex]P(X^i=x^i \mid Y=c_k)\approx\dfrac{\sum_{j=1}^{m}I\{x_j^i=x^i,y_j=c_k\}+1}{N_k+A_i}\end{cases} \tag{5-18}$$

其中，A_i 表示 X^i 的所有可能取值的个数。

基于上述讨论，完整的朴素贝叶斯分类器的算法描述见算法 5-1。

算法 5-1　朴素贝叶斯分类器

输入：样本集合 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$；待预测样本 x；样本标记的所有可能取值$\{c_1,c_2,\cdots,c_K\}$；样本输入变量 X 的每个属性变量 X^i 的所有可能取值$\{a_{i1},a_{i2},\cdots,a_{iA_i}\}$

输出：待预测样本 x 所属的类别

1. 计算标记为 c_k 的样本出现的概率

$$P(Y=c_k)=\frac{N_k+1}{m+K},\quad k=1,2,\cdots,K$$

2. 计算标记为 c_k 的样本，其 X^i 分量的属性值为 a_{ip} 的概率

$$P(X^i=a_{ip}\mid Y=c_k)=\frac{\sum_{j=1}^{N_k}I(x_j^i=a_{ip},y_j=c_k)+1}{N_k+A_i}$$

3. 根据上面的估计值计算 x 属于所有 y_k 的概率值，并选择概率最大的作为输出

$$y=\underset{k=1,2,\cdots,K}{\operatorname{argmax}}(P(Y=c_k\mid X=x))$$

$$=\underset{k=1,2,\cdots,K}{\operatorname{argmax}}\left(P(Y=c_k)\prod_{i=1}^{n}P(X^i=x^i\mid Y=c_k)\right)$$

Return y

5.4　朴素贝叶斯分类器的极大似然估计解释

朴素贝叶斯思想的本质是极大似然估计，$P(Y=c_k)$和 $P(X^i=x^i\mid Y=c_k)$是我们要估计的概率值。以 $P(Y=c_k)$为例，令 $\theta_k=P(Y=c_k)$，则似然函数为

$$L(\theta)=\prod_{i=1}^{m}P(Y=y_i)=\prod_{k=1}^{K}\theta_k^{N_k}\tag{5-19}$$

根据极大似然估计，求 θ 等价于求解下面的优化问题：

$$\max_{\theta}\quad l(\theta)=\sum_{k=1}^{K}N_k\ln\theta_k$$

$$\text{s.t.}\quad\sum_{k=1}^{K}\theta_k=1\tag{5-20}$$

使用拉格朗日乘子法求解。首先构造拉格朗日乘数为

$$\text{Lag}(\theta)=\sum_{k=1}^{K}N_k\ln\theta_k+\lambda\left(\sum_{k=1}^{K}\theta_k-1\right)\tag{5-21}$$

令拉格朗日函数对 θ_k 的偏导为 0，有

$$\frac{\partial\text{Lag}(\theta)}{\partial\theta_k}=\frac{N_k}{\theta_k}+\lambda=0\Rightarrow N_k=-\lambda\theta_k\tag{5-22}$$

于是

$$\sum_{k=1}^{K}N_k=-\lambda\left(\sum_{k=1}^{K}\theta_k\right)=-\lambda\tag{5-23}$$

解得

$$\begin{cases}\lambda = -m \\ \theta_k = \dfrac{N_k}{\lambda} = \dfrac{N_k}{m}\end{cases} \tag{5-24}$$

这样便得到了 $P(Y=c_k)$的极大似然估计。对 $P(X^i=x^i \mid Y=c_k)$的极大似然估计求解过程类似，留给读者自行推导。

5.5　实例：基于朴素贝叶斯实现垃圾短信分类

本节以一个例子来阐述朴素贝叶斯分类器在垃圾短信分类中的应用。SMS Spam Collection Data Set 是一个垃圾短信分类数据集，包含了 5574 条短信，其中有 747 条垃圾短信。数据集以纯文本的形式存储，其中每行对应一条短信。每行的第一个单词是 spam 或 ham，表示该行的短信是否为垃圾短信。随后记录了短信的内容，内容和标签之间以制表符分隔。

该数据集没有收录进 sklearn.datasets，所以需要自行加载，如代码清单 5-1 所示。

代码清单 5-1　加载 SMS 垃圾短信数据集

```
with open('./SMSSpamCollection.txt', 'r', encoding = 'utf8') as f:
    sms = [line.split('\t') for line in f]
y, x = zip(*sms)
```

加载完成后，x 和 y 分别是长为 5574 的字符串列表。其中 y 的每个元素只可能是 spam 或 ham，分别表示垃圾短信和正常短信。x 的每个元素表示对应短信的内容。在训练贝叶斯分类器前，需要先将 x 和 y 转换成适于训练的数值表示形式，这个过程称为特征提取，如代码清单 5-2 所示。

代码清单 5-2　SMS 垃圾短信数据集特征提取

```
from sklearn.feature_extraction.text import CountVectorizer as CV
from sklearn.model_selection import train_test_split
y = [label == 'spam' for label in y]
x_train, x_test, y_train, y_test = train_test_split(x, y)
counter = CV(token_pattern = '[a-zA-Z]{2,}')
x_train = counter.fit_transform(x_train)
x_test = counter.transform(x_test)
```

特征提取的结果存储在(x_train, y_train)以及(x_test, y_test)中。其中 x_train 和 x_test 分别是 4180×6595 和 1394×6595 的稀疏矩阵。不难看出，两个矩阵的行数之和等于 5574，也就是完整数据集的大小。因此两个矩阵的每行应该代表一个样例，那么每列代表什么呢？查看 counter 的 vocabulary_属性就会发现，其大小恰好是 6595，也就是所有短信中出现过的不同单词的个数。例如短信“Go until jurong point, go”中一共有 5 个单词，但是由于 go 出现了两次，所以不同单词的个数只有 4 个。x_train 和 x_test 中的第(i,j)个元素就表示第 j 个单词在第 i 条短信中出现的次数。

最后就是朴素贝叶斯分类器的构造与训练，如代码清单 5-3 所示。我们首先基于训练集训练朴素贝叶斯分类器，然后分别在训练集和测试集上进行测试。测试结果显示，模型在

训练集上的分类准确率达到0.993,在测试集上的分类准确率为0.986。可见朴素贝叶斯分类器达到了良好的分类效果。

代码清单 5-3　朴素贝叶斯分类器的构造与训练

```
from sklearn.naive_bayes import MultinomialNB as NB
model = NB()
model.fit(x_train, y_train)
train_score = model.score(x_train, y_train)
test_score = model.score(x_test, y_test)
print("train score:", train_score)
print("test score:", test_score)
```

朴素贝叶斯分类器假设样本特征之间相互独立。这一假设非常强,以至于几乎不可能满足。但是在实际应用中,朴素贝叶斯分类器往往表现良好,特别是在垃圾邮件过滤、信息检索等场景下往往表现优异。


题库

习题 5

一、选择题

1. 朴素贝叶斯分类器的训练过程是基于训练集 D 来估计(　　)。

A. 先验概率　　B. 后验概率　　C. 概率分布函数　D. 概率密度函数

2. 下列哪种情况不能用朴素贝叶斯分类器?(　　)

A. 训练数据集较大

B. 实例具有几个属性

C. 给定分类参数,描述实例的属性应该是条件独立的

D. 要求有较高的分类精度

3. 贝叶斯分类器的训练中,最大似然法估计参数的过程包括(　　)。

A. 求导数,令偏导数为0,得到似然方程组

B. 对似然函数取对数,并整理

C. 解似然方程组,得到所有参数即为所求

D. 以上所有

4. 朴素贝叶斯是一种特殊的Bayes分类器,特征变量是 X,类别标签是 Y,它的一个假定是(　　)。

A. 各类别的先验概率 $P(Y)$ 是相等的

B. 特征变量 X 的各个维度是类别条件独立随机变量

C. 以0为均值,sqrt(2)/2为标准差的正态分布

D. $P(X|Y)$ 是高斯分布

5. 表5-1中列出了14个日期中天气、温度、湿度和风力四个因素和小明是否攀岩的关系。基于这14个观测数据,采用朴素贝叶斯分类方法计算出实例 <天气=晴天,温度=凉爽,湿度=高,风力=强>时"休息"的概率为(　　)。

表 5-1　观测数据

日　期	天　气	温　度	湿　度	风　力	攀　岩
D1	晴天	热	高	弱	休息
D2	晴天	热	高	强	休息
D3	阴天	热	高	弱	攀岩
D4	下雨	温和	高	弱	攀岩
D5	下雨	凉爽	正常	弱	攀岩
D6	下雨	凉爽	正常	强	休息
D7	阴天	凉爽	正常	强	攀岩
D8	晴天	温和	高	弱	休息
D9	晴天	凉爽	正常	弱	攀岩
D10	下雨	温和	正常	弱	攀岩
D11	晴天	温和	正常	强	攀岩
D12	阴天	温和	高	强	攀岩
D13	阴天	热	正常	弱	攀岩
D14	下雨	温和	高	强	休息

A. 0.0795　　B. 0.0205　　C. 0.64　　D. 0.33

二、判断题

1. 贝叶斯的思想是“由因推果”。（　　）
2. 可以用极大似然估计法解贝叶斯分类器。（　　）
3. 贝叶斯分类器可以解决无监督学习的问题。（　　）
4. 朴素贝叶斯分类器不存在数据平滑问题。（　　）
5. 贝叶斯分类器是一种基于贝叶斯公式的分类器。（　　）

三、填空题

1. 朴素贝叶斯分类算法假设属性之间相互_____。
2. 贝叶斯分类器可以解决_____学习的问题。
3. 贝叶斯分类器是基于_____概率，推导出_____概率。
4. 假定某同学使用贝叶斯分类模型时，由于失误操作，致使训练数据中两个维度重复表示，那么模型效果精度会_____。
5. 贝叶斯定理中，如果描述随机事件 A 和 B 的条件概率的定理，表达式是_____。

四、问答题

1. 简述朴素贝叶斯的优缺点。
2. 简述朴素贝叶斯与 LR 的区别。
3. 简述朴素贝叶斯基本原理和预测过程。
4. 朴素贝叶斯中有没有超参数可以调？

五、应用题

1. 已知样本的属性和标签如表 5-2 所示，当某样本属性为($a2$，$b2$，$c2$)时，采用朴素贝叶斯方法，求非归一化的 $P(L3|a2,b2,c2)P(L3|a2,b2,c2)$值。

表 5-2 样本的属性和标签

属性 1	属性 2	属性 3	标　　签
$a2$	$b1$	$c3$	$L2$
$a1$	$b1$	$c2$	$L3$
$a1$	$b1$	$c1$	$L1$
$a3$	$b3$	$c1$	$L3$
$a1$	$b3$	$c2$	$L3$
$a3$	$b1$	$c3$	$L1$
$a2$	$b2$	$c1$	$L3$
$a1$	$b2$	$c1$	$L3$
$a2$	$b3$	$c3$	$L3$
$a2$	$b2$	$c3$	$L1$

2. 通过 sklearn.datasets 生成两种类别数据,使用朴素贝叶斯进行分类并展示结果。

3. 通过数据集(https://www.kaggle.com/c/sf-crime/data)对旧金山犯罪进行分类预测。