明理精工 笃学致远

第6章 生成式分类器

GenClassifier

电子工程学院、人工智能学院
college of Electronic Engineering . college of Artificial Intelligence

- 生成式分类器
- 朴素贝叶斯分类器

1. 生成式分类器 vs. 判别式分类器

- 生成式分类器：对所有变量的分布建模$p(\boldsymbol{x}, y)$
  - 得到类先验概率$p(y)$，类条件概率$p(\boldsymbol{x}|y)$
  - 根据贝叶斯公式，得到后验，实现分类
  - 知道所有变量的分布，可以产生数据

- 判别式分类器：直接对后验概率$p(y|\boldsymbol{x})$建模，或者直接对判别边界建模$f(\boldsymbol{x})$
  - Logistic回归：对后验概率$p(y|\boldsymbol{x})$建模
  - SVM：直接对判别函数$f(\boldsymbol{x})$建模

➢ 贝叶斯公式

生成式分类器对所有变量的分布建模 $p(\boldsymbol{x}, y) = p(y)\, p(\boldsymbol{x}|y)$

根据贝叶斯公式，得到后验，实现分类：

$$p(y=c|\boldsymbol{x}) = \frac{p(\boldsymbol{x},y)}{\sum_{c'} p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|y=c)p(y=c)}{\sum_{c'} p(\boldsymbol{x}|y=c')p(y=c')} \propto p(\boldsymbol{x}|y=c)p(y=c)$$

类条件概率　　类先验概率

类先验概率可用多项分布表示： $y \sim \mathrm{Multinoulli}(\boldsymbol{\theta})$

类条件概率 $p(\boldsymbol{x}|y)$：由于$\boldsymbol{x}$为多维向量，条件分布$p(\boldsymbol{x}|y)$建模困难，对其做适当假设

- 朴素贝叶斯： 在给定y的情况下， $\boldsymbol{x}$的各维独立
- 高斯判别分析：在给定y的情况下， $\boldsymbol{x}$为多元高斯分布

2. 朴素贝叶斯分类器 (Naive Bayes Classifier, NBC)

■假设共有 C 个类别$y \in \{1,2,\dots,C\}$ ，类别的先验分布

- 两类： $y\sim\text{Bernoulli}(\theta)$
- 多类： $y\sim\text{Multinoulli}(\boldsymbol{\theta})$

$$\text{Multinoulli}(\boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y_i=c)}$$

■每个样本的特征为： $\boldsymbol{x} = (x_1, x_2, \dots, x_D)^{\mathrm{T}}$

■朴素贝叶斯分类器假设各维特征在给定类别标签的情况下条件独立

$$p(\boldsymbol{x}|y=c) = \prod_{j=1}^{D} p(x_j|y=c)$$

➢ 朴素贝叶斯的训练（模型参数估计）

log似然函数为

类条件概率　　类先验概率

$$\ln p(\mathcal{D}) = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i, y_i) = \sum_{i=1}^{N} \ln\big(p(\boldsymbol{x}_i|y_i)p(y_i)\big) = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i|y_i) + \sum_{i=1}^{N} \ln p(y_i)$$

$$= \sum_{i=1}^{N} \ln\left(\prod_{c=1}^{C}\big(p(\boldsymbol{x}_i|y_i = c)\big)^{\mathbb{I}(y_i=c)}\right) + \sum_{i=1}^{N} \ln\left(\prod_{c=1}^{C}\theta_c^{\mathbb{I}(y_i=c)}\right)$$

$$= \sum_{i=1}^{N} \ln\left(\prod_{c=1}^{C}\left(\prod_{j=1}^{D} p\big(x_{ij}|y_i = c\big)\right)^{\mathbb{I}(y_i=c)}\right) + \sum_{i=1}^{N} \ln\left(\prod_{c=1}^{C}\theta_c^{\mathbb{I}(y_i=c)}\right)$$

$$= \sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{I}(y_i = c)\sum_{j=1}^{D} \ln p\big(x_{ij}|y_i = c\big) + \sum_{i=1}^{N}\sum_{c=1}^{C} \ln \theta_c^{\mathbb{I}(y_i=c)}$$

log似然函数：

$$\ln p(\mathcal{D}) = \sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{j=1}^{D}\mathbb{I}(y_i = c)\ln p\left(x_{ij}|y_i = c\right) + \boxed{\sum_{i=1}^{N}\sum_{c=1}^{C}\ln \theta_c^{\mathbb{I}(y_i=c)}}$$

✓ **类先验概率**

求参数θ_c，需约束条件：$\sum_{c=1}^{C}\theta_c = 1$

采用拉格朗日乘子法求极值：

$$J(\lambda, \boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\ln(\theta_c) + \lambda\left(1 - \sum_{c=1}^{C}\theta_c\right)$$

$$\left.\begin{aligned}\frac{\partial J(\lambda, \boldsymbol{\theta})}{\partial \theta_c} &= \frac{1}{\theta_c}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c) - \lambda = 0 \\ \frac{\partial J(\lambda, \boldsymbol{\theta})}{\partial \lambda} &= 1 - \sum_{c=1}^{C}\theta_c = 0\end{aligned}\right\} \quad \theta_c = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)}{N} = \frac{N_c}{N}$$

第c类的样本占所有样本的比例

✓ 类条件概率

类条件概率

$$\ln p(\mathcal{D}) = \boxed{\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{j=1}^{D} \mathbb{I}(y_i = c) \ln p\left(x_{ij} | y_i = c\right)} + \sum_{i=1}^{N}\sum_{c=1}^{C} \ln \theta_c^{\mathbb{I}(y_i=c)}$$

（1）伯努利分布——NBC二值特征

$p(x_j|Y = c)$可用伯努利分布$\mathrm{Bernoulli}(x_j|Y = c, \theta_{c,j})$表示，

其中参数$\theta_{c,j}$表示在类别$Y = c$的情况下，特征$X_j = 1$的概率。

$$p\left(x_{i,j}|y_i = c\right) = \left(\theta_{c,j}\right)^{x_{i,j}}\left(1 - \theta_{c,j}\right)^{1-x_{i,j}}$$

$$\boxed{p(x_{i,j}|y_i = c) = (\theta_{c,j})^{x_{i,j}}(1-\theta_{c,j})^{1-x_{i,j}}}$$

- log似然函数为

$$\ln p(\mathcal{D}) = \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\sum_{j=1}^{D}\ln p(x_{ij}|y_i = c) + \sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c$$

$$= \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\sum_{j=1}^{D}\ln\left((\theta_{c,j})^{x_{i,j}}(1-\theta_{c,j})^{1-x_{i,j}}\right) + \sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c$$

$$= \sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{j=1}^{D}\mathbb{I}(y_i = c)\left(x_{i,j}\ln\theta_{c,j} + (1-x_{i,j})\ln(1-\theta_{c,j})\right) + \sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c$$

- 类条件分布的参数$\theta_{c,j}$与似然函数中第一项有关

$$\frac{\partial \ln p(\mathcal{D})}{\partial\theta_{c,j}} = \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\frac{x_{i,j}}{\theta_{c,j}} - \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\frac{1-x_{i,j}}{1-\theta_{c,j}}$$

- log似然函数对参数$\theta_{c,j}$的一阶偏导数为0:

$$\frac{\partial \ln p(\mathcal{D})}{\partial p_{c,j}} = \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\frac{x_{i,j}}{\theta_{c,j}} - \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\frac{1-x_{i,j}}{1-\theta_{c,j}} = 0$$

$$\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)x_{i,j}(1-\theta_{c,j}) = \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)(1-x_{i,j})\theta_{c,j}$$

$$\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)x_{i,j} = \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\theta_{c,j}$$

$$\hat{\theta}_{c,j} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)x_{i,j}}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)} = \frac{N_{c,j}}{N_c}$$

$$\frac{\text{Y=c, Xj=1的样本数目}}{\text{第}c\text{类的样本数目}}$$

$$\hat{\theta}_{c,j}=\frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)x_{i,j}}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)}=\frac{N_{c,j}}{N_c}$$

当观测$N_{c,j}=0$时，$\hat{\theta}_{c,j}=0$，从而

$$p(x_{i,j}|y_i=c)=\left(\hat{\theta}_{c,j}\right)^{x_{i,j}}\left(1-\hat{\theta}_{c,j}\right)^{1-x_{i,j}}=0$$

$$p(\boldsymbol{x}|y=c)=\prod_{j=1}^{D}p(x_j|y=c)=0$$

$$p(y=c|\boldsymbol{x})\propto p(\boldsymbol{x}|y=c)p(y=c)=0$$

解决方案：类条件平滑（加入伪计数，每个特征取值的样本数增加平滑因子α）

$$\hat{\theta}_{c,j}=\frac{N_{c,j}+\alpha}{N_c+2\alpha}$$

$\alpha=1$：拉普拉斯平滑

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious	**Y**
D1	1	1	1	1	0	1	1	1	0	0	0	**1**
D2	1	1	0	0	1	1	1	0	1	0	0	**1**
D3	0	0	0	0	0	1	0	0	0	1	1	**0**

V 个二值特征

朴素贝叶斯大大减少了模型的参数量

- 无独立假设：

$$p(y=c|\boldsymbol{x}) = p(\boldsymbol{x}|y)p(y)$$

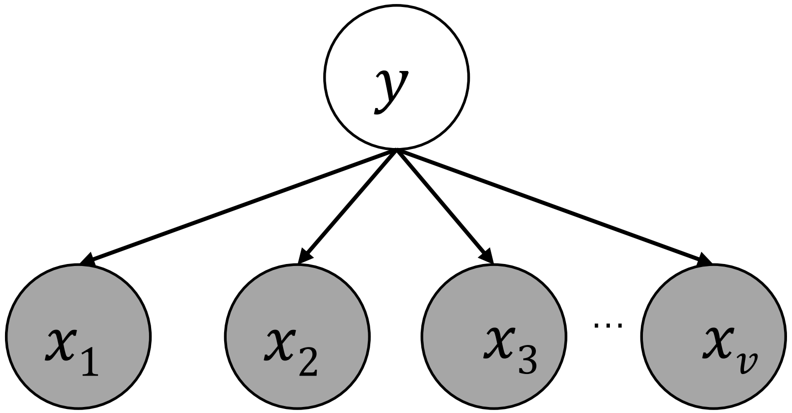
类条件密度参数：$C \times (2^V - 1)$

类先验参数：$C - 1$

- 条件独立假设：

$$p(y=c|\boldsymbol{x}) = p(\boldsymbol{x}|y)p(y)$$

$$= \prod_{j=1}^{D} p(x_j|y)\ p(y)$$



类条件密度参数：$C \times V$

类先验参数：$C - 1$

(2) Multinnoulli分布——NBC离散型特征（类别型特征）

假设特征有M种取值，$p(x_j|Y=c)$可用类别分布$\mathrm{Cat}(x_j|Y=c,\theta_{c,j,m})$表示，其中参数$\theta_{c,j,m}$表示在类别$Y=c$的情况下，特征$X_j=m$的概率。

$$p\left(x_{i,j}=m|y_i=c\right)=\prod_{m=1}^{M}\left(\theta_{c,j,m}\right)^{\mathbb{I}(x_{i,j}=m)}$$

$$p(x_{i,j}=m|y_i=c)=\prod_{m=1}^{M}\left(\theta_{c,j,m}\right)^{\mathbb{I}(x_{i,j}=m)}$$

- log似然函数为

$$\ln p(\mathcal{D})=\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\sum_{j=1}^{D}\ln p\left(x_{ij}|y_i=c\right)+\sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c$$

$$=\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\sum_{j=1}^{D}\ln\left(\prod_{m=1}^{M}\left(\theta_{c,j,m}\right)^{\mathbb{I}(x_{i,j}=m)}\right)+\sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c$$

$$=\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{j=1}^{D}\sum_{m=1}^{M}\mathbb{I}(y_i=c)\mathbb{I}\left(x_{i,j}=m\right)\ln\theta_{c,j,m}+\sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c$$

- 类条件分布的参数$\theta_{c,j,m}$与似然函数中第一项有关

$$\frac{\partial\ln p(\mathcal{D})}{\partial\theta_{c,j,m}}=\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{m=1}^{M}\mathbb{I}(y_i=c)\frac{\mathbb{I}\left(x_{i,j}=m\right)}{\theta_{c,j,m}}$$

- 需约束条件：$\sum_{m=1}^{M}\theta_{j,c,m}=1$，采用拉格朗日乘子法求$\theta_{c,j,m}$的值：

$$J(\lambda,\boldsymbol{\theta})=\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{j=1}^{D}\sum_{m=1}^{M}\mathbb{I}(y_i=c)\mathbb{I}\left(x_{i,j}=m\right)\ln\theta_{c,j,m}+\lambda\left(1-\sum_{m=1}^{M}\theta_{c,j,m}\right)$$

$$\frac{\partial J(\lambda,\boldsymbol{\theta})}{\partial\theta_{c,j,m}}=\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{m=1}^{M}\mathbb{I}(y_i=c)\frac{\mathbb{I}\left(x_{i,j}=m\right)}{\theta_{c,j,m}}-\lambda=0$$

$$\frac{\partial J(\lambda,\boldsymbol{\theta})}{\partial\lambda}=1-\sum_{m=1}^{M}\theta_{c,j,m}=0$$

$$\hat{\theta}_{c,j,m}=\frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{m=1}^{M}\mathbb{I}(y_i=c)\,\mathbb{I}\left(x_{i,j}=m\right)}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)}=\frac{N_{c,j,m}}{N_c}$$

$$\frac{\text{第}c\text{类的所有样本，第}j\text{维特征值为}m\text{的样本数目}}{\text{第}c\text{类的样本数目}}$$

$$\hat{\theta}_{c,j,m}=\frac{N_{c,j,m}+\alpha}{N_c+M_j\alpha}$$

(3) 多项分布

- 特征取值为**序数**（如出现次数），第v维特征表示字典第v个单词出现的次数

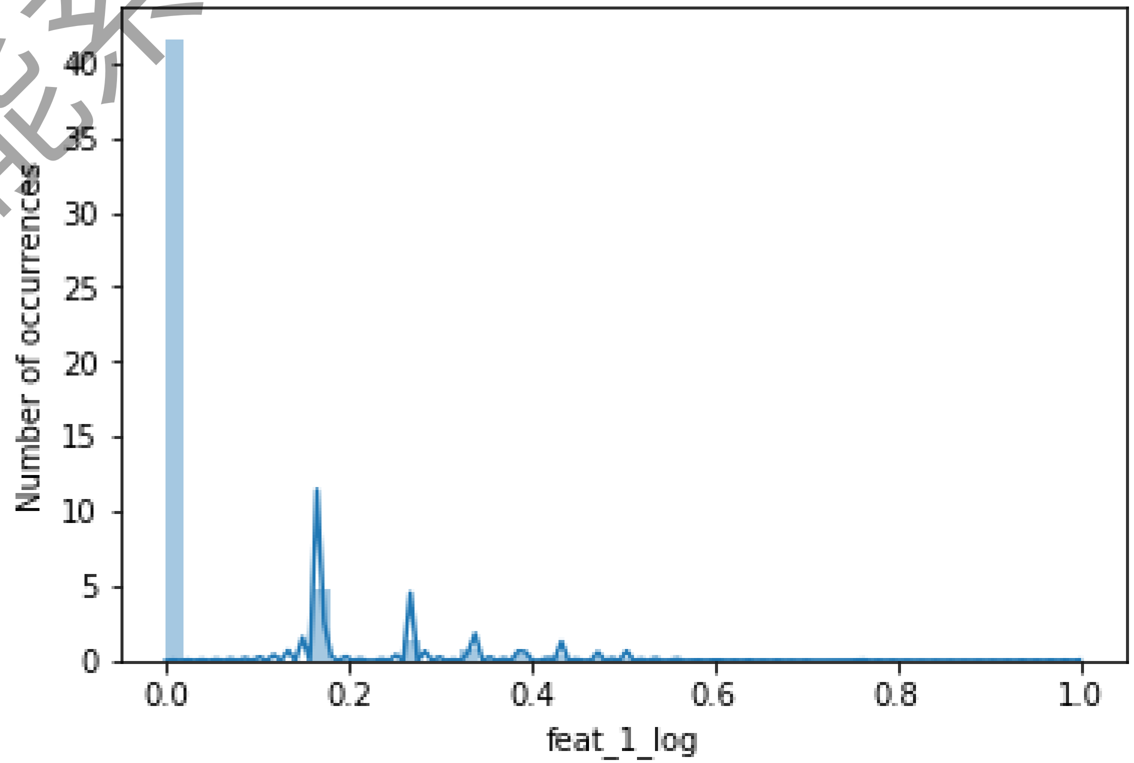
$$\hat{\theta}_{c,v} = \frac{N_{c,v}}{N_c} \qquad \frac{\text{第}c\text{类的所有样本/文档中，第}v\text{个单词出现的次数和}}{\text{第}c\text{类文档中的总单词数目}}$$

Sklearn中的实现：[MultinomialNB]

（4）高斯分布——NBC连续特征

$$p\left(x_{i,j}|y_i=c\right)=N\left(\mu_{c,j},\sigma_{c,j}\right)=\frac{1}{\sqrt{2\pi}\sigma_{c,j}}e^{-\frac{\left(x_{i,j}-\mu_{c,j}\right)^2}{2\left(\sigma_{c,j}\right)^2}}$$

Sklearn中的实现： GaussianNB



- 采用极大似然法来估计模型参数，log似然函数为

$$p(x_{i,j}|y_i=c)=\frac{1}{\sqrt{2\pi}\sigma_{c,j}}e^{-\frac{(x_{i,j}-\mu_{c,j})^2}{2(\sigma_{c,j})^2}}$$

$$\begin{aligned}
\ln p(\mathcal{D}) &= \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\sum_{j=1}^{D}\ln p(x_{ij}|y_i=c)+\sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c \\
&= \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\sum_{j=1}^{D}\ln\left(\frac{1}{\sqrt{2\pi}\sigma_{c,j}}e^{-\frac{(x_{i,j}-\mu_{c,j})^2}{2\sigma^2}}\right)+\sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c \\
&= -\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\frac{D}{2}\ln(2\pi)-\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)D\ln(\sigma_{c,j}) \\
&\quad -\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\sum_{j=1}^{D}\frac{(x_{i,j}-\mu_{c,j})^2}{2(\sigma_{c,j})^2}+\sum_{i=1}^{N}\sum_{c=1}^{C}\ln\theta_c
\end{aligned}$$

- 类条件件分布的参数$\mu_{j,c},\sigma_{j,c}$与似然函数中第2项和第3项有关

- 似然函数对参数求偏导数：

$$\ln p(\mathcal{D}) = -\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)D\ln\left(\sigma_{c,j}\right) - \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\sum_{j=1}^{D}\frac{\left(x_{i,j}-\mu_{c,j}\right)^2}{2\left(\sigma_{j,c}\right)^2}$$

$$\frac{\ln p(\mathcal{D})}{\partial\mu_{c,j}} = \frac{1}{\left(\sigma_{j,c}\right)^2}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\left(x_{i,j} - \mu_{c,j}\right) = 0$$

$$\hat{\mu}_{c,j} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)x_{i,j}}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)} = \bar{x}_{i,j}$$

第c类的所有样本，第j维特征值的均值

- 似然函数对参数求偏导数：

$$\ln p(\mathcal{D}) = -\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)D\ln\left(\sigma_{c,j}\right) - \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\sum_{j=1}^{D}\frac{\left(x_{i,j}-\mu_{j,c}\right)^2}{2\left(\sigma_{c,j}\right)^2}$$

$$\frac{\ln p(\mathcal{D})}{\partial \sigma_{c,j}} = -\frac{1}{\sigma_{c,j}}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c) + \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\frac{\left(x_{i,j}-\mu_{j,c}\right)^2}{\left(\sigma_{c,j}\right)^3} = 0$$

$$\left(\sigma_{c,j}\right)^2 = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\left(x_{i,j}-\mu_{j,c}\right)^2}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)} \quad (5\text{-}10)$$

第c类的所有样本，
第j维特征值的样本方差

例：SNS账号真实性判断　X=[m,m,yes]？

1. 类先验概率：

$$\bar{\theta}_0 = P(R = \text{yes}) = \frac{7}{10}$$

$$\bar{\theta}_1 = P(R = \text{no}) = \frac{3}{10}$$

日志密度L	好友密度F	是否使用真实头像H	账号是否真实R
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

2. 类条件概率：

当类别$R = \text{yes}$时，共有$N_0 = 7$个样本，

特征日志密度L有3种取值：s, l, m，样本数分别为：1，3，3，再加入平滑计数$\alpha = 1$，得到

$$\hat{\theta}_{0,0,s} = \frac{1+1}{7+3} = \frac{1}{5} \qquad \hat{\theta}_{0,0,l} = \frac{2}{5} \qquad \hat{\theta}_{0,0,m} = \frac{2}{5}$$

类似地，得到其他 2 维特征F、H的类条件分布的参数为

$$\bar{\theta}_{0,F,\mathrm{s}}=\frac{1+1}{3+7}=\frac{1}{5},\quad \bar{\theta}_{0,F,\mathrm{m}}=\frac{1+2}{3+7}=\frac{3}{10},\quad \bar{\theta}_{0,F,\mathrm{l}}=\frac{1+4}{3+7}=\frac{1}{2},$$

$$\bar{\theta}_{0,H,\mathrm{no}}=\frac{1+3}{2+7}=\frac{4}{9},\quad \bar{\theta}_{0,H,\mathrm{yes}}=\frac{1+4}{2+7}=\frac{5}{9}。$$

② 当类别标签$R=\mathrm{no}$时，$N_1=3$，特征日志密度L有 s，m，l 这 3 种取值，用 Multinoulli 分布建模，$M_0=3$。在$R=\mathrm{no}$的 3 个样本中，上述 3 种特征取值的样本数分别为 2、1、0，即

$$N_{1,L,\mathrm{s}}=2,\quad N_{1,L,\mathrm{m}}=1,\quad N_{1,L,\mathrm{l}}=0,$$

根据式（5-8），得到

$$\bar{\theta}_{1,L,\mathrm{s}}=\frac{1+2}{3+3}=\frac{1}{2},\quad \bar{\theta}_{1,L,\mathrm{m}}=\frac{1+1}{3+3}=\frac{1}{3},\quad \bar{\theta}_{1,L,\mathrm{l}}=\frac{1+0}{3+3}=\frac{1}{6}。$$

类似地，得到其他 2 维特征F、H的类条件分布的参数为

$$\bar{\theta}_{1,F,\mathrm{s}}=\frac{1+3}{3+3}=\frac{2}{3},\quad \bar{\theta}_{1,F,\mathrm{m}}=\frac{1+0}{3+3}=\frac{1}{6},\quad \bar{\theta}_{1,F,\mathrm{l}}=\frac{1+0}{3+3}=\frac{1}{6},$$

$$\bar{\theta}_{1,H,\mathrm{no}}=\frac{1+1}{2+3}=\frac{2}{5},\quad \bar{\theta}_{1,H,\mathrm{yes}}=\frac{1+2}{2+3}=\frac{3}{5}。$$

3.预测新样本 X=[m,m,yes]

$$P(R=\text{yes}|L=\text{m},F=\text{m},H=\text{yes})$$

$$\propto P(L=\text{m}|R=\text{yes})P(F=\text{m}|R=\text{yes})\ P(H=\text{yes}|R=\text{yes})\ \ P(R=\text{yes})$$

$$=\bar{\theta}_{0,L,\text{m}}\times\bar{\theta}_{0,F,\text{m}}\times\bar{\theta}_{0,H,\text{yes}}\times\bar{\theta}_{0}=\frac{2}{5}\times\frac{3}{10}\times\frac{5}{9}\times\frac{2}{3}=\frac{4}{90},$$

$$P(R=\text{no}|L=\text{m},F=\text{m},H=\text{yes})$$

$$\propto P(L=\text{m}|R=\text{no})P(F=\text{m}|R=\text{no})\ P(H=\text{yes}|R=\text{no})\ \ P(R=\text{no})$$

$$=\bar{\theta}_{1,L,\text{m}}\times\bar{\theta}_{1,F,\text{m}}\times\bar{\theta}_{1,H,\text{yes}}\times\bar{\theta}_{1}=\frac{1}{3}\times\frac{1}{6}\times\frac{3}{5}\times\frac{1}{3}=\frac{1}{90},$$

$$P(R=\text{yes}|L=\text{m},F=\text{m},H=\text{yes})>\ P(R=\text{no}|L=\text{m},F=\text{m},H=\text{yes}),$$

因此该用户的账号真实的可能性更高。

➢ Sklearn中的朴素贝叶斯实现

Scikit-Learn中提供5种朴素贝叶斯的分类算法:

- GaussianNB：特征值为连续值且为高斯分布
- BernoulliNB：特征值为二值
- CategoricalNB：特征值为多个离散值
- MultinomialNB：特征值为多个序数值
- ComplementNB： MultinomialNB 的一种改进，特别适用于不平衡数据集。ComplementNB使用来自每个类的补数的统计数据来计算模型的权重，这样参数估计更稳定，在文本分类任务上，性能通常更好

朴素贝叶斯分类器

$$\boxed{p(y=c|\boldsymbol{x}) = p(\boldsymbol{x}|y=c)p(y=c)}$$

- ✓ 类先验概率 $p(y=c) = \frac{N_c}{N}$
- ✓ 类条件概率 $p(\boldsymbol{x}|y=c)$
  - 伯努利分布 $\hat{\theta}_{c,j} = \frac{N_{c,j}+\alpha}{N_c+2\alpha}$
  - Multinnoulli分布 $\hat{\theta}_{c,j,m} = \frac{N_{c,j,m}+\alpha}{N_c+M_j\alpha}$
  - 多项式分布
  - 高斯分布