机器学习 作业三

Hugo Zhang

1

设朴素贝叶斯分类中一共有K个类，输入特征为n维，第j维特征有m_j种取值，一共有N个数据，属于第k类的数据有N_k个，第k类中第j维特征为l的数据有N_k^{jl}个。使用条件概率公式计算后验概率，并引入全概率公式 $P(x)=\sum\limits_{k=1}^{K}P(x|Y=c_k)P(Y=c_k)$可得贝叶斯公式:

$$P(Y=c_i|x)=\frac{P(Y=c_i,x)}{P(x)}=\frac{P(x|Y=c_i)P(Y=c_i)}{\sum\limits_{k=1}^{K}P(x|Y=c_k)P(Y=c_k)}$$

朴素贝叶斯方法里假设类已确定的条件下各维数据特征是条件独立的，即:

$$P(X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},...,X^{(n)}=x^{(n)}|Y=c_k)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

我们一共需要估计$K(1+\sum\limits_{j=1}^{n}m_j)$个参数，即K个类先验概率 $p(y=c_k)$以及在每个类的条件下输入的每个维度每个取值的条件概率 $p(x^{(j)}=a_{jl}|y=c_k)$。设参数集合为 θ。

1.1

使用极大似然估计，对数似然函数如下，需注意最后两个式子中已经带入了之前提到的参数。

$$\begin{aligned}
L(\theta)&=\ln\prod_{i=1}^{N}p(x_i,y_i;\theta)\\
&=\ln\prod_{i=1}^{N}p(x_i|y_i;\theta)p(y_i;\theta)\\
&=\ln\prod_{i=1}^{N}(\prod_{j=1}^{n}p(x_i^{(j)}|y_i;\theta))p(y_i;\theta)\\
&=\sum_{i=1}^{N}(\ln p(y_i;\theta)+\sum_{j=1}^{n}\ln p(x_i^{(j)}|y_i;\theta))\\
&=\sum_{i=1}^{N}(\sum_{k=1}^{K}\ln p(y=c_k)^{I(y_i=c_k)}+\sum_{j=1}^{n}\sum_{l=1}^{m_j}\sum_{k=1}^{K}\ln p(x^{(j)}=a_{jl}|y=c_k)^{I(x_i^{(j)}=a_{jl},y_i=c_k)})\\
&=\sum_{i=1}^{N}(\sum_{k=1}^{K}I(y_i=c_k)\ln p(y=c_k)+\sum_{j=1}^{n}\sum_{l=1}^{m_j}\sum_{k=1}^{K}I(x_i^{(j)}=a_{jl},y_i=c_k)\ln p(x^{(j)}=a_{jl}|y=c_k))
\end{aligned}$$

此处的参数满足概率和为1的约束，因此可以使用拉格朗日乘子法，也可以将约束带入上式直接由极值条件求解。此处我们使用拉格朗日乘子法。可知约束为:

$$\sum_{k=1}^{K}p(y=c_k)=1$$

$$\sum_{l=1}^{m_j} p(x^{(j)} = a_{jl}|y = c_k) = 1, j = 1, 2, ..., n; k = 1, 2, ..., K$$

使用拉格朗日乘子法，得到拉格朗日函数:

$$\mathcal{L}(\theta, \alpha, \beta) = L(\theta) + \alpha(\sum_{k=1}^{K} p(y = c_k) - 1) + \sum_{j=1}^{n}\sum_{k=1}^{K} \beta_{jk}(\sum_{l=1}^{m_j} p(x^{(j)} = a_{jl}|y = c_k) - 1)$$

首先求 $p(y = c_k)$参数，由

$$0 = \frac{\partial \mathcal{L}(\theta, \alpha, \beta)}{\partial p(y = c_k)} = \sum_{i=1}^{N} \frac{I(y_i = c_k)}{p(y = c_k)} + \alpha = \frac{N_k}{p(y = c_k)} + \alpha, k = 1, 2, ..., K$$

可得

$$p(y = c_k) = -\frac{N_k}{\alpha}, k = 1, 2, ..., K$$

带入约束

$$1 = \sum_{k=1}^{K} p(y = c_k) = \sum_{k=1}^{K} -\frac{N_k}{\alpha} = -\frac{N}{\alpha}$$

解得

$$\alpha = -N$$

$$p(y = c_k) = \frac{N_k}{N}, k = 1, 2, ..., K$$

再求 $p(x^{(j)} = a_{jl}|y = c_k)$参数，由

$$0 = \frac{\partial \mathcal{L}(\theta, \alpha, \beta)}{\partial p(x^{(j)} = a_{jl}|y = c_k)} = \sum_{i=1}^{N} \frac{I(x_i^{(j)} = a_{jl}, y_i = c_k)}{p(x^{(j)} = a_{jl}|y = c_k)} + \beta_{jk} = \frac{N_k^{jl}}{p(x^{(j)} = a_{jl}|y = c_k)} + \beta_{jk}$$

可得

$$p(x^{(j)} = a_{jl}|y = c_k) = -\frac{N_k^{jl}}{\beta_{jk}}, l = 1, 2, ..., m_j; j = 1, 2, ..., n; k = 1, 2, ..., K$$

带入约束

$$1 = \sum_{l=1}^{m_j} p(x^{(j)} = a_{jl}|y = c_k) = \sum_{l=1}^{m_j} -\frac{N_k^{jl}}{\beta_{jk}} = -\frac{N_k}{\beta_{jk}}, j = 1, 2, ..., n; k = 1, 2, ..., K$$

解得

$$\beta_{jk} = -N_k, j = 1, 2, ..., n; k = 1, 2, ..., K$$

$$p(x^{(j)} = a_{jl}|y = c_k) = \frac{N_k^{jl}}{N_k}, l = 1, 2, ..., m_j; j = 1, 2, ..., n; k = 1, 2, ..., K$$

综上所述，极大似然估计中的参数估计值分别为

$$p(y = c_k) = \frac{N_k}{N}, k = 1, 2, ..., K$$

$$p(x^{(j)} = a_{jl}|y = c_k) = \frac{N_k^{jl}}{N_k}, l = 1, 2, ..., m_j; j = 1, 2, ..., n; k = 1, 2, ..., K$$

1.2

使用贝叶斯估计，考虑用参数后验概率的期望作为估计值。此处要估计的参数都可以视为多项分布的参数，将 $p(y=c_k), k=1,2,...,K$作为1组多项分布的参数、$p(x^{(j)}=a_{jl}|y=c_k), l=1,2,...,m_j$作为$nK$组多项分布的参数来分别估计。先验分布设为狄利克雷分布。

接下来分四步进行推导，第一步推导狄利克雷分布中向量的每一个分量的期望，第二步推导说明以狄利克雷分布为先验分布的多项分布参数的后验分布仍为狄利克雷分布，第三步推导说明朴素贝叶斯算法中的每一组参数都是多项分布的参数，最后综合以上结论给出参数估计值。

第一步，推导狄利克雷分布里随机向量每一维分量的期望。假设一个向量$X=(X_1,X_2,...,X_N)$服从参数为α的狄利克雷分布，即概率密度函数为$f(X_1,X_2,...,X_N;\alpha_1,\alpha_2,...,\alpha_N)=\frac{1}{B(\alpha)}\prod_{i=1}^{N}X_i^{\alpha_i-1}$，其中$B$为Beta函数。第$j$维分量$X_j$的期望如下，推导中用到了 $B(\alpha)=\frac{\prod\Gamma(\alpha_i)}{\Gamma(\sum\alpha_i)}$ 以及 $\Gamma(x+1)=x\Gamma(x)$:

$$
\begin{aligned}
E(X_j) &= \int\int...\int X_j f(X;\alpha)dx_1dx_2...dx_N \\
&= \frac{\Gamma(\sum_{i=1}^{N}\alpha_i)}{\prod_{i=1}^{N}\Gamma(\alpha_i)}\int\int...\int X_j^{\alpha_j}\prod_{i\neq j}^{N}X_i^{\alpha_i-1}dx_1dx_2...dx_N \\
&= \frac{\Gamma(\sum_{i=1}^{N}\alpha_i)}{\prod_{i=1}^{N}\Gamma(\alpha_i)}\cdot\frac{\Gamma(\alpha_j+1)\prod_{i\neq j}^{N}\Gamma(\alpha_i)}{\Gamma(1+\sum_{i=1}^{N}\alpha_i)} \\
&= \frac{\Gamma(\sum_{i=1}^{N}\alpha_i)\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)\Gamma(1+\sum_{i=1}^{N}\alpha_i)} \\
&= \frac{\Gamma(\sum_{i=1}^{N}\alpha_i)\alpha_j\Gamma(\alpha_j)}{\Gamma(\alpha_j)(\sum_{i=1}^{N}\alpha_i)\Gamma(\sum_{i=1}^{N}\alpha_i)} \\
&= \frac{\alpha_j}{\sum_{i=1}^{N}\alpha_i}
\end{aligned}
$$

第二步，推导以狄利克雷分布为先验分布的多项分布参数的后验分布。设参数的先验分布（狄利克雷分布）的概率密度函数为 $h(p_1,p_2,...,p_K;\alpha_1,\alpha_2,...,\alpha_K)$，给定的数据$X$服从多项分布$X\sim PN(N:p_1,p_2,...,p_K)$，概率密度函数为$m(x_1,x_2,...,x_K;p_1,p_2,...,p_K)=\frac{(\sum_{i=1}^{K}x_i)!}{\prod_{i=1}^{K}x_i!}\prod_{i=1}^{K}p_i^{x_i}$，由贝叶斯公式:

$$P(p_1,p_2,...,p_K|X)=\frac{P(X|p_1,p_2,...,p_K)P(p_1,p_2,...,p_K)}{\int\int...\int P(X|p_1,p_2,...,p_K)P(p_1,p_2,...,p_K)dp_1dp_2...dp_K}$$

带入概率密度函数，得到后验分布的概率密度函数为

$$
\begin{aligned}
p(p_1,p_2,...,p_K|X) &= \frac{m(x_1,x_2,...,x_K;p_1,p_2,...,p_K)h(p_1,p_2,...,p_K;\alpha_1,\alpha_2,...,\alpha_K)}{\int\int...\int m(x_1,x_2,...,x_K;p_1,p_2,...,p_K)h(p_1,p_2,...,p_K;\alpha_1,\alpha_2,...,\alpha_K)dp_1dp_2...dp_K} \\
&= \frac{\prod_{i=1}^{K} p_i^{x_i+\alpha_i-1}}{\int\int...\int \prod_{i=1}^{K} p_i^{x_i+\alpha_i-1}dp_1dp_2...dp_K} \\
&= \frac{\Gamma(\sum_{i=1}^{K}(x_i+\alpha_i))}{\prod_{i=1}^{K}\Gamma(x_i+\alpha_i)}\prod_{i=1}^{K} p_i^{x_i+\alpha_i-1}
\end{aligned}
$$

可见该后验分布为狄利克雷分布，该狄利克雷分布的参数为 $x_i+\alpha_i, i=1,2,...,K$。

第三步，说明朴素贝叶斯算法中的每一组参数都是多项分布的参数。参数 $p(y=c_k), k=1,2,...K$满足 $\sum_{k=1}^{K} p(y=c_k)=1$，相当于多项分布$PN(N: p(y=c_1), p(y=c_2), ..., p(y=c_k))$；每一个$j$和$k$确定的一组参数 $p(x^{(j)}=a_{jl}|y=c_k), l=1,2,...,m_j$满足 $\sum_{l=1}^{m_j} p(x^{(j)}=a_{jl}|y=c_k)=1$，相当于多项分布$PN(N_k: p(x^{(j)}=a_{j1}|y=c_k), p(x^{(j)}=a_{j2}|y=c_k), ..., p(x^{(j)}=a_{j,m_j}|y=c_k))$。

最后一步，综合上述结论得到参数估计值。对于参数 $p(y=c_k), k=1,2,...K$，其后验概率分布为参数$x_i+\alpha_i, i=1,2,...$的狄利克雷分布，值得注意的是此处的x_i实为总样本中类别为i的样本数量即N_i，而α_i为先验分布的参数，因此$p(y=c_k)$的后验概率的期望为

$$
E(p(y=c_k)|X) = \frac{N_k+\alpha_k}{\sum_{i=1}^{K}(N_i+\alpha_i)} = \frac{N_k+\alpha_k}{N+\sum\alpha}, k=1,2,...,K
$$

对于参数 $p(x^{(j)}=a_{jl}|y=c_k)$，其后验概率分布为参数$x_i+\alpha_i, i=1,2,...$的狄利克雷分布，值得注意的是此处的x_i实为总样本中类别为k且输入特征第j维取值为i的样本数量即N_k^{ji}，而α_i为先验分布的参数，因此$p(x^{(j)}=a_{jl}|y=c_k)$的后验概率的期望为

$$
E(p(x^{(j)}=a_{jl}|y=c_k)|X) = \frac{N_k^{jl}+\alpha_l}{\sum_{i=1}^{K}(N_k^{ji}+\alpha_i)} = \frac{N_k^{jl}+\alpha_l}{N_k+\sum\alpha}, l=1,2,...,m_j; j=1,2,...,n; k=1,2,...,K
$$

这里我们不妨取对称的狄利克雷分布作为所有参数组的先验概率分布，且设先验分布的参数皆为λ，将所有参数的后验概率分布的期望作为参数估计值，最后可以得到：

$$
p(y=c_k) = \frac{N_k+\lambda}{N+K\lambda}, k=1,2,...,K
$$

$$
p(x^{(j)}=a_{jl}|y=c_k) = \frac{N_k^{jl}+\lambda}{N_k+m_j\lambda}, l=1,2,...,m_j; j=1,2,...,n; k=1,2,...,K
$$

2

这三种方法都基于近邻法，都使用距离来表示点的相似程度，都使用代表点来表示数据区域或者单元的原型，都根据与测试数据接近的点的信息来进行预测。他们的不同在于如何从训练数据中得到代表点、用多少代表点来进行预测（即k取多少）。详细情况如下：

方法	k-近邻法	基于K-means的分类方法	学习向量量化方法
代表点	所有的训练数据	每一类k个代表点	每一类k个代表点
训练过程	无	使用所有的同类点更新代表点	使用一个任意点更新最近的代表点
用于预测的点数	k个点占优投票	1个点	1个点