设输入空间为 n 维向量的集合，输出空间为类标记集合 $c_1, c_2, \cdots, c_k$. 输入为特征向量 x 属于输入空间, 输出为类标记 y 属于输出空间.Xs 是定义在输入空间上的随机变量,Y 是定义在在输出空间上的随机向量.$P(X,Y)$ 是 X 和 Y 的联合概率分布.

训练数据集为:$T=(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$ 由 $P(X,Y)$ 独立同分布产生.

朴素贝叶斯通过训练数据集学习联合概率分布 $P(X,Y)$. 具体地，学习以下先验概率分布及条件概率分布: 先验概率分布 $P(Y=c_k), k=1,2,\cdots,K$, 条件概率分布 $P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k), k=1,2,\cdots,K$

于是学习到联合概率分布 $P(X,Y)$

假设 $x^{(j)}$ 可取值有 S_j 个，则 $x^{(j)}$ 可能取值的集合为 $a_{j1}, a_{j2}, \cdots, a_{jS_j}$, $j=1,2,...n$;Y 可取值有 K 个，那么参数个数为 $K\prod_{j=1}^n S_j$

朴素贝叶斯法对条件概率分布作了条件独立性的假设.

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

朴素贝叶斯法分类时，对给定的输入 x，通过学习到的模型计算后验概率分布 $P(Y=c_k|X=x)$，将后验概率最大的类作为输出. 后验概率计算根据贝叶斯定理进行：

$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^n P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_{k=1}^K P(X=x|Y=c_k)P(Y=c_k)}$

将条件独立性假设代入上式得：

$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^n P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_{k=1}^K P(Y=c_k)\prod_{j=1}^n P(X^{(j)}=x^{(j)}|Y=c_k)}$

这是朴素贝叶斯分类的基本公式. 于是，朴素贝叶斯分类器可表示为：

$y=f(x)=\arg\max\limits_{c_k}\frac{P(Y=c_k)\prod_{j=1}^n P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_{k=1}^K P(Y=c_k)\prod_{j=1}^n P(X^{(j)}=x^{(j)}|Y=c_k)}$

注意到分母对所有 c_k 都是相同的，所以

$y=f(x)=\arg\max\limits_{c_k}P(Y=c_k)\prod_{j=1}^n P(X^{(j)}=x^{(j)}|Y=c_k)$

从上式可以看出，朴素贝叶斯法的学习也就是要估计先验概率 $P(Y=c_k)$ 和条件分布概率 $P(X^{(j)}=x^{(j)}|Y=c_k)$，可以应用极大似然估计法估计相应的概率.

下面写出推导过程:

把 $p(Y=c_k)$ 和 $p(x^{(j)}=a_{jl}|y=c_k)$ 作为参数.

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

$p(x|y=c_k)=\prod_{j=1}^n p(x^j|y=c_k)=\prod_{j=1}^n\prod_{l=1}^{S_j}p(X^{(j)}=a_{jl}|Y=c_k)^{1\{x^{(j)}=a_{jl},y=c_k\}}$

为叙述方便起见，下面以 φ 代表参数集合 $p(Y=c_k), p(x^{(j)}=a_{jl}|y=c_k)$

先写出 log 似然函数

$$
\begin{aligned}
l(\varphi)&=log\prod_{i=1}^N p(x_i,y_i;\varphi)\\
&=log\prod_{i=1}^N p(x_i;y_i;\varphi)p(y_i;\varphi)\\
&=log\prod_{i=1}^N[\prod_{j=1}^n p(x_i^{(j)}|y_i;\varphi)]p(y_i;\varphi)\\
&=\sum_{i=1}^N(logp(y_i,\varphi)+\sum_{j=1}^n logp(x_i^{(j)}|y_i;\varphi))\\
&=\sum_{i=1}^N[\sum_{k=1}^K logp(y=c_k)^{1\{y_i=c_k\}}+\sum_{k=1}^K\sum_{j=1}^n\sum_{l=1}^{S_j}logp(x^j=a_{jl}|y=c_k)^{1\{x_i^{(j)}=a_{jl},y_i=c_k\}}]\\
&=\sum_{i=1}^N[\sum_{k=1}^K 1\{y_i=c_k\}logp(y=c_k)+\sum_{k=1}^K\sum_{j=1}^n\sum_{l=1}^{S_j}1\{x_i^{(j)}=a_{jl},y_i=c_k\}logp(x^j=a_{jl}|y=c_k)]
\end{aligned}
$$

在上式中把 $p(Y=c_k)$和$p(x^{(j)}=a_{jl}|y=c_k)(j=1,2,\cdots,n;l=1,2,\cdots,S_j;k=1,2,\cdots,K)$ 作为参数.

先求先验概率 $p(Y=c_k)$ 的最大似然估计, 因为存在约束条件 $\sum_{k=1}^K p(y=c_k)=1$，所以下面开始用拉格朗日乘数法分别求最大似然估计 (条件极值):

上式中只有前半段含有 $p(Y=c_k)$，所以在求先验概率估计值时就只管前半部分.

令 $F=\sum\limits_{i=1}^{N}\Big\{\big(\sum\limits_{k=1}^{K}1\{y_i=c_k\}logp(y=c_k)\big)+\lambda(1-\sum\limits_{k=1}^{K}p(y=c_k))\Big\}$

分别对 $p(y=c_k)(k=1,2,\cdots,K)$和λ 求导:

$$\begin{cases}\dfrac{\partial F}{\partial p(y=c_1)}=\sum\limits_{i=1}^{N}\{\dfrac{1\{y_i=c_1\}}{p(y=c_1)}-\lambda\}=0\\ \dfrac{\partial F}{\partial p(y=c_2)}=\sum\limits_{i=1}^{N}\{\dfrac{1\{y_i=c_2\}}{p(y=c_2)}-\lambda\}=0\\ \vdots\\ \dfrac{\partial F}{\partial p(y=c_K)}=\sum\limits_{i=1}^{N}\{\dfrac{1\{y_i=c_K\}}{p(y=c_K)}-\lambda\}=0\\ \dfrac{\partial F}{\partial \lambda}=\sum\limits_{i=1}^{N}\{1-\sum\limits_{k=1}^{K}p(y=c_k)\}=0\end{cases}\tag{1}$$

则由前面面 K 个式子可得:

$$\begin{cases}p(y=c_1)=\frac{\sum\limits_{i=1}^{N}1\{y_i=c_1\}}{N\lambda}\\ p(y=c_2)=\frac{\sum\limits_{i=1}^{N}1\{y_i=c_2\}}{N\lambda}\\ \vdots\\ p(y=c_K)=\frac{\sum\limits_{i=1}^{N}1\{y_i=c_K\}}{N\lambda}\end{cases}\tag{2}$$

由于 $\sum\limits_{k=1}^{K}p(y=c_k)=1$, 则将上面左边全部式子加起来, 可以得到

$$1=\sum_{k=1}^{K}p(y=c_k)=\frac{\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{N}1\{y_i=c_k\}}{N\lambda}=\frac{N}{N\lambda}$$

即 $\lambda=1$, 代入方程组 (2), 可得 $p(y=c_k)$ 的极大似然估计为:

$p(y=c_k)=\frac{\sum\limits_{i=1}^{N}1\{y_i=c_2\}}{N}(k=1,2,\cdots,K)$

下面开始求 $p(x^{(j)}=a_{jl}|y=c_k)$ 的极大似然估计:

已知 log 似然函数为:

$$l(\varphi)=\sum_{i=1}^{N}[\sum_{k=1}^{K}1\{y_i=c_k\}logp(y=c_k)+\sum_{k=1}^{K}\sum_{j=1}^{n}\sum_{l=1}^{S_j}1\{x_i^{(j)}=a_{jl},y_i=c_k\}logp(x^{(j)}=a_{jl}|y=c_k)]$$

只需对式子后面部分求偏导即可. 由于存在约束条件 $\sum\limits_{l=1}^{S_j}p(x^{(j)}=a_{jl}|y=c_k)=1$, 所以也可用拉格朗日乘数法求极大似然估计:

令

$$G=\sum_{i=1}^{N}\Big\{\sum_{k=1}^{K}\sum_{j=1}^{n}\Big(\big(\sum_{l=1}^{S_j}1\{x_i^{(j)}=a_{jl},y_i=c_k\}logp(x^{(j)}=a_{jl}|y=c_k)\big)+\lambda_{kj}(1-\sum_{l=1}^{S_j}p(x^{(j)}=a_{jl}|y=c_k))\Big)\Big\}$$

注意由于对于每个 k 和 j 都存在约束条件 $\sum\limits_{l=1}^{S_j}p(x^{(j)}=a_{jl}|y=c_k)=1$, 所以总共有 $k\times l$ 个约束条件, 上式中的参数 λ_{kj} 对应的是 k 和 j 固定时的约束条件.

$$\begin{cases} \dfrac{\partial G}{\partial p(x^{(j)}=a_{jl}|y=c_k)} = \sum\limits_{i=1}^{N}\Big\{\dfrac{1\{x_i^{(j)}=a_{jl},y_i=c_k\}}{p(x^j=a_{jl}|y=c_k)}-\lambda_{kj}\Big\}=0 \\ \dfrac{\partial G}{\partial \lambda_{kj}} = \sum\limits_{i=1}^{N}\{1-\sum\limits_{l}^{S_j}p(x^{(j)}=a_{jl}|y=c_k)\}=0 \end{cases} \tag{3}$$

由第 1 个式子可得 $p(x^j=a_{jl}|y=c_k)=\frac{\sum_{i=1}^{N}1\{x_i^{(j)}=a_{jl},y_i=c_k\}}{N\lambda_{kj}}$

由第 2 个式子可得 $\sum\limits_{l}^{S_j}p(x^{(j)}=a_{jl}|y=c_k)=1$

联立两个式子可以得到:

$1=\sum\limits_{l}^{S_j}p(x^{(j)}=a_{jl}|y=c_k)=\frac{\sum_{l}^{S_j}\sum_{i=1}^{N}1\{x_i^{(j)}=a_{jl},y_i=c_k\}}{N\lambda_{kj}}=\frac{\sum_{i=1}^{N}1\{y_i=c_k\}}{N\lambda_{kj}}$

解得:$N\lambda_{kj}=\sum\limits_{i=1}^{N}1\{y_i=c_k\}$

则有:$p(x^j=a_{jl}|y=c_k)=\frac{\sum_{i=1}^{N}1\{x_i^{(j)}=a_{jl},y_i=c_k\}}{\sum_{i=1}^{N}1\{y_i=c_k\}}$

证明完毕.