逻辑斯蒂回归与最大熵模型

W.J.Z

1 逻辑斯蒂回归

1.1 逻辑斯蒂分布

设 X 为连续随机变量，X 服从逻辑斯蒂分布是指 X 具有以下的分布函数和密度函数：

$$F(x)=P(X\leq x)=\frac{1}{1+e^{\frac{-(x-\mu)}{\gamma}}} \tag{1}$$

$$f(x)=F'(x)=\frac{e^{\frac{-(x-\mu)}{\gamma}}}{\gamma\left(1+e^{\frac{-(x-\mu)}{\gamma}}\right)^2} \tag{2}$$

μ 为位置参数，$\gamma>0$ 为形状参数。

1.2 逻辑斯蒂回归模型

设 Y 的取值为 0 或者 1，$w=(w^1,w^2,\ldots,w^n,b)^T$，$x=(x^1,x^2,\ldots,x^n,1)^T$，则二项逻辑斯蒂回归模型为如下的条件概率分布：

$$P(Y=1|x)=\frac{exp(w\cdot x)}{1+exp(w\cdot x)} \tag{3}$$

$$P(Y=0|x)=\frac{1}{1+exp(w\cdot x)} \tag{4}$$

一件事情发生的几率是指改事件发生的概率与该事件不发生概率的比值，对于逻辑斯蒂回归而言，改事件的对数几率为：

$$\log\frac{P(Y=1|x))}{1-P(Y=1|x)}=w\cdot x \tag{5}$$

输出 $Y=1$ 的对数几率是输入 x 的线性函数表示的模型，线性函数的值越接近正无穷，概率值就越接近 1；线性函数的值越接近负无穷，概率值就越接近 0.

假设离散型随机变量 Y 的取值集合为 $1,2,3,...,K$, 多项逻辑斯蒂回归模型是：

$$P\left(Y=k|x\right)=\frac{exp\left(w_k\cdot x\right)}{1+\sum_{k=1}^{K-1}exp\left(w_k\cdot x\right)} \tag{6}$$

$$P\left(Y=K|x\right)=\frac{1}{1+\sum_{k=1}^{K-1}exp\left(w_k\cdot x\right)} \tag{7}$$

1.3 模型参数估计

设 $y_i\epsilon\left\{0,1\right\}$ 应用极大似然估计法评估模型参数，设：

$$P\left(Y=1|x\right)=\pi\left(x\right),P\left(Y=0|x\right)=1-\pi\left(x\right), \tag{8}$$

似然函数为：

$$\prod_{i=1}^{N}\left[\pi\left(x_i\right)\right]^{y_i}\left[1-\pi\left(x_i\right)\right]^{1-y_i} \tag{9}$$

对数似然函数为：

$$L\left(w\right)=\sum_{i=1}^{N}\left[y_i\log\pi\left(x_i\right)+\left(1-y_i\log\left(1-\pi\left(x_i\right)\right)\right)\right] \tag{10}$$

$$=\sum_{i=1}^{N}\left[y_i\log\frac{\pi\left(x_i\right)}{1-\pi\left(x_i\right)}+\log\left(1-\pi\left(x_i\right)\right)\right] \tag{11}$$

$$=\sum_{i=1}^{N}\left[y_i\left(w\cdot x_i\right)-\log\left(1+exp\left(w\cdot x_i\right)\right)\right] \tag{12}$$

使用梯度下降法求极大值。

2 最大熵模型

2.1 最大熵原理

最大熵原理认为要选择的概率模型首先必须满足已有的事实，即约束条件，在没有更多信息的情况下，那些不确定的部分都是” 等可能的”。

2.2 最大熵模型

设联合分布的经验分布和边缘分布的经验分布为：

$$P\left(X=x,Y=y\right)=\frac{v\left(X=x,Y=y\right)}{N} \tag{13}$$

$$P\left(X=x\right)=\frac{v\left(X=x\right)}{N} \tag{14}$$

$v(X=x,Y=y)$ 表示训练数据中样本 (x,y) 出现的频数，$v(X=x)$ 表示训练数据中输入 x 的频数，N 表示训练样本容量。

用特征函数表示输入 x 与输出 y 之间的某一个事实，其定义为：

$$f(x,y)=\begin{cases}1, & \text{x 与 y 满足某一事实}\\ 0, & \text{其他}\end{cases} \tag{15}$$

特征函数 $f(x,y)$ 关于联合经验分布的期望值为：

$$E_{p(x,y)}=\sum_{x,y}P\left(x,y\right)f\left(x,y\right) \tag{16}$$

特征函数关于模型 $P(Y|X)$ 与边缘经验分布的期望值为：

$$E_{p(y|x)}=\sum_{x,y}P\left(x\right)P\left(y|x\right)f\left(x,y\right) \tag{17}$$

假设两个期望值相等，则：

$$E_{p(x,y)}=E_{p(y|x)} \tag{18}$$

公式 18将作为模型学习的约束条件。定义在条件概率分布的条件熵为：

$$H\left(P\right)=-\sum_{x,y}P\left(x\right)P\left(y|x\right)\log P\left(y|x\right) \tag{19}$$

公式 19为最大熵模型，式中的对数为自然对数。

2.3 最大熵模型的学习

最大熵模型的学习等价于约束最优化问题。

$$\max H\left(P\right)=-\sum_{x,y}P\left(x\right)P\left(y|x\right)\log P\left(y|x\right)$$

$$E_{p(x,y)} = E_{p(y|x)}$$

$$\sum_{y} P\left(y|x\right) = 1$$

按照最优化的习惯，将求最大值问题改写为等价的求最小值问题。

$$\min -H\left(P\right) = \sum_{x,y} P\left(x\right) P\left(y|x\right) \log P\left(y|x\right)$$

$$E_{p(x,y)} - E_{p(y|x)} = 0$$

$$\sum_{y} P\left(y|x\right) = 1$$

将约束最优化的原始问题转换为无约束最优化的对偶问题，引入拉格朗日乘子，得：

$$\begin{aligned} L\left(P,w\right) &= -H\left(P\right) + w_0 \left(1 - \sum_{y} P\left(y|x\right)\right) + \sum_{i=1}^{n} w_i \left(E_{p(x,y)} - E_{p(y|x)}\right) \\ &= \sum_{x,y} P\left(x\right) P\left(y|x\right) \log P\left(y|x\right) + w_0 \left(1 - \sum_{y} P\left(y|x\right)\right) + \sum_{i=1}^{n} w_i \left(E_{p(x,y)} - E_{p(y|x)}\right) \end{aligned}$$

对 $L(P,w)$ 对 $P(y|x)$ 求偏导得：

$$P\left(y|x\right) = \frac{exp\left(\sum_{i=1}^{n} w_i f_i\left(x,y\right)\right)}{exp\left(1 - w_0\right)} \tag{20}$$

将 $P(y|x)$ 带入 $L(P,w)$ 对 w 求偏导，得出 w_i 的值进而得出 $P(y|x)$ 的值。