Author: 陆星宇 Date: 2021/2/15

第6章 逻辑斯谛回归与最大熵模型

一、交流讨论

问题1：结合逻辑斯谛分布的来源，探讨使用逻辑斯谛回归模型的优点是什么？

逻辑斯谛分布的分布函数是sigmoid函数。我认为选择这种分布有以下几点：

1. 定义域：(−∞,+∞)；值域：(−1,1)；满足我们对于二分类模型输入输出的需求。
2. 函数图像形状，曲线在中心附近增长速度较快，在两端增长速度较慢．形状参数γ的值越小，曲线在中心附近增长得越快。其实逻辑斯谛学习与线性模型的区别不大，本质上都是寻找一个切分点，但是这种切分点附近函数值变化陡峭的函数要求提高切分点的精度。
3. 一种感性的认识是：大多数情况下，并没有办法知道未知事件的概率分布形式，此时正态分布是一个最好的选择。Sigmoid函数和正态分布函数的积分形式形状非常类似。但计算正态分布的积分函数，计算代价非常大，而Sigmoid函数计算量非常的小，因此被选为替代函数。

问题2：为何说在满足相同约束的情况下，熵最大的模型是最优的模型？

熵最大的时候，说明随机变量最不确定。最大熵原理的实质就是，在已知部分知识的前提下，关于未知分布最合理的推断就是符合已知知识最不确定或最随机的推断，这是我们可以作出的不偏不倚的选择，任何其它的选择都意味着我们增加了其它的约束和假设，这些约束和假设根据我们掌握的信息无法作出。

问题3：书中将最大熵模型和逻辑斯谛回归放在一起讲，两者的相似之处与不同之处是什么？

首先，最大熵模型和逻辑斯谛回归模型都是通过极大似然估计进行参数学习的分类方法。其次，最大熵模型与逻辑斯谛回归模型有类似的形式，都被称为对数线性模型。它们都有共同的特征就是加权线性求和高维特征向量的各分量，将和的值作为分类依据，为了提高分类精度，再代入对数变换。不同之处在于，其条件概率分布的表示形式不同。

问题4：逻辑斯蒂回归为何采用sigmoid函数，这个函数有独特的优点吗？有没有不足之处呢？

函数优点同问题1，不足之处有如下几点：

1. 在趋向无穷的地方，函数值变化很小，容易缺失梯度；
2. sigmoid函数的输出不是0均值的，会导致后层的神经元的输入是非0均值的信号，这会对梯度产生影响。
3. 计算复杂度高，因为sigmoid函数是指数形式。

当然前两点是相对于深层神经网络来说的，对本章的模型并无影响。

二、内容概要

2.1 逻辑斯蒂分布

$$F(x)=\frac{1}{1+e^{-(x-\mu)/\gamma}}$$

$$f(x)=\frac{e^{-(x-\mu)/\gamma}}{\gamma(1+e^{-(x-\mu)/\gamma})^2}$$

2.2 二项逻辑斯谛回归模型

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

2.3 参数估计

对数似然函数为：

$$L(w)=ln(\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i})$$
$$=\sum_{i=1}^{N}[y_i ln\pi(x_i)+(1-y_i)ln(1-\pi(x_i))]$$
$$=\sum_{i=1}^{N}[y_i ln\frac{\pi(x_i)}{1-\pi(x_i)}+ln(1-\pi(x_i))]$$
$$=\sum_{i=1}^{N}[y_i(w\cdot x_i)-log(1+exp(w\cdot x_i))]$$

2.4 多项逻辑斯谛回归

$$P(Y=k|x)=\frac{exp(w_k\cdot x)}{1+\sum_{k=1}^{K-1}exp(w_k\cdot x)},\ k=1,2,\dots,K-1$$
$$P(Y=K|x)=\frac{1}{1+\sum_{k=1}^{K-1}exp(w_k\cdot x)}$$

2.5 最大熵模型

2.5.1 最大熵原理

熵最大的模型是最好的模型。

设离散型随机变量的概率分布是$P(X)$，则其熵是

$$H(P)=-\sum_{x}P(x)lnP(x)$$

熵满足下列不等式

$$0\le H(P)\le ln|X|$$

其中|X|的是X的取值个数，当且仅当X服从均匀分布时，等号成立即熵最大

2.5.2 最大熵模型的定义

特征函数：

$$f(n)=\begin{cases}1, & x与y满足某一事实\\ 0, & 否则\end{cases}$$

特征分布关于经验分布的期望：

$$E_{\tilde{P}}(f)=\sum_{x,y}\tilde{P}(x,y)f(x,y)$$

特征函数关于模型与经验分布的期望：

$$E_P(f) = \sum_{x,y} \tilde{P}(x)P(y|x)f(x,y)$$

如果模型能够获取训练数据中的信息，那么假设两个期望值相等。即

$$E_{\tilde{P}}(f) = E_P(f) \ or \sum_{x,y} \tilde{P}(x,y)f(x,y) = \sum_{x,y} \tilde{P}(x)P(y|x)f(x,y)$$

定义最大熵模型为：

假设满足所有约束条件的模型集合为$C={P|E_{\tilde P}(f_i)=E_P(f_i),i=1,2,...,n}$

定义在条件概率分布上的条件熵为$H(P)=-\sum_{x,y}\tilde P(x)P(y|x)lnP(y|x)$

则模型集合C中条件熵H(P)最大的模型称为最大熵模型。