逻辑斯谛回归

1 逻辑斯谛回归模型

1.1 逻辑斯谛分布

- **密度函数**

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

- **分布函数**

$$f(x)=F'(x)=\frac{e^{\frac{-(x-\mu)}{\lambda}}}{\lambda(1+e^{\frac{-(x-\mu)}{\lambda}})^2}$$

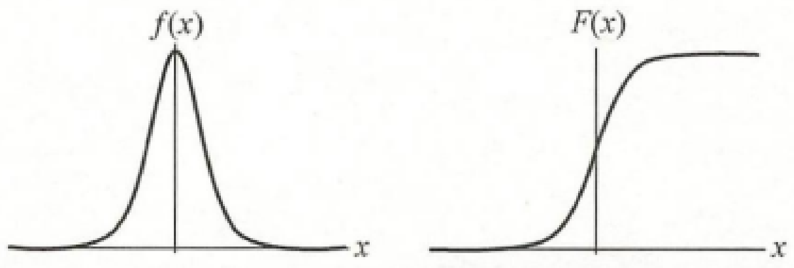

图 6.1 逻辑斯谛分布的密度函数与分布函数

1.2 二项逻辑斯谛回归模型

- **二项逻辑斯谛回归模型是一种分类模型，**$X\in\mathbb{R}^n$**，**$Y\in\{0,1\}$**。**

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

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

- **有时为了方便，**$w=(w^{(1)},w^{(2)},\ldots,w^{(n)},b)^T$**，**$x=(x^{(1)},x^{(2)},\ldots,x^{(n)},1)^T$**。**

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

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

- **假设事件发生的概率是**p**，那么该事件的几率是**$\frac{p}{1-p}$**，该事件的对数几率（log odds）或logit函数是**

$$\mathrm{logit}(p)=\log\frac{p}{1-p}$$

由此能得出

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

1.3 模型参数估计

- $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$，$x_i \in \mathbb{R}^n$，$y_i \in \{0,1\}$，**可以应用极大似然估计法估计模型参数，从而得到逻辑斯谛回归模型。**
- **设：**

$$P(Y=1|x) = \pi(x)$$
$$P(Y=0|x) = 1-\pi(x)$$

则得出似然函数：

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

转换为对数似然函数为：

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

对$L(w)$求极大值，得到w的估计值。
由此，问题就变成了对数似然函数为目标函数的最优化问题。逻辑斯谛回归函数中常用的是梯度下降法和拟牛顿法。

- **若w的极大似然估计值是$\hat{w}$，那么学到的逻辑斯谛回归函数模型为**

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

1.4 多项逻辑斯谛回归

- $Y=\{1,2,\ldots,K\}$

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

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

2 最大熵模型

2.1 最大熵原理

- **熵越大，越混乱**
- **最大熵原理是在满足约束条件的模型集合中选取熵最大的模型**
  假设：

$$H(P)=-\sum_x P(x)\log P(x)$$

  熵需满足：

$$0\leq H(P)\leq \log|x|$$

> **例子：**
> **假设随机变量X有5个取指$\{A,B,C,D,E\}$，估计各个值的概率$P(A),P(B),P(C),P(D),P(E)$。**
>
> **解：**
> $P(A)+P(B)+P(C)+P(D)+P(E)=1$
> **能够满足这个条件的概率分布有无穷个，但是要求最大熵，那各个值就会拥有相等的概率，即：**
> $P(A)=P(B)=P(C)=P(D)=P(E)=\frac{1}{5}$
>
> **那再增加一个约束条件：**
> $P(A)+P(B)=\frac{3}{10}$
> **为了得到最大熵，**
> $P(A)=P(B)=\frac{3}{20}$
> $P(C)=P(D)=P(E)=\frac{7}{20}$

2.2 最大熵模型的定义

假设分类模型是一个条件概率分布$P(Y|X)$，$X\in\chi\subseteq\mathbb{R}^n$表示输入，$Y\in\gamma$表示输出。
$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$

- **经验分布**

$$\tilde{P}(X=x,Y=y)=\frac{\nu(X=x,Y=y)}{N}$$

$$\tilde{P}(X=x)=\frac{\nu(X=x)}{N}$$

  其中，$\nu(X=x,Y=y)$表示训练集中出现$(X=x,Y=y)$的频数。而N为训练集样本个数
- **特征函数**

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

- 期望值

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

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

最大熵模型

假设满足所有约束条件的模型集合为

$$C=\{P\in P|E_p(f_i)=E_{\tilde{P}}(f_i),\quad i=1,2,\dots,n\}$$

在条件概率分布$P(Y|X)$**上的条件熵为**

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

则模型合辑C**条件熵**$H(P)$**最大的模型就是最大熵模型。**

2.3 最大熵模型的学习

给定数据集$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$**以及**$f_i(x,y),i=1,2,\dots,n$**。**

- **约束最优化问题：**

$$\min_{P\in C}\quad -H(P)=\sum_{x,y}\tilde{P}(x)|(y|x)\log P(y|x)$$

$$s.t.\quad E_p(f_i)-E_{\tilde{P}}(f_i)=0,\quad i=1,2,\dots,n$$

$$\sum_y P(y|x)=1$$

此处引进拉格朗日乘子$w_0,w_1,w_2,\dots,w_n$**，定义拉格朗日函数**$L(P,w)$**：**

$$\begin{aligned}L(P,w)&\equiv -H(P)+w_0(1-\sum_y P(y|x))+\sum_{i=1}^n w_i(E_{\tilde{P}}(f_i)-E_P(f_i))\\&=\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)+w_0(1-\sum_y P(y|x))\\&+\sum_{i=1}^n w_i(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x,y}\tilde{P}(x)P(y|x)f_i(x,y))\end{aligned}$$

> **拉格朗日乘数法：**[理解【拉格朗日乘数法】：有等号约束的最优化]

由此，最优化的原始问题是：

$$\min_{P\in C}\max_w L(P,w)$$

其对偶问题便是：

$$\max_{P\in C}\min_{w} L(P,w)$$

由于拉格朗日函数$L(P,w)$是P的凸函数，因此要找到其最小值。

$$\Psi(w)=\min_{P\in C} L(P,w)=L(P_w,w)$$

而其中

$$P_w=\arg\min_{P\in C} L(P,w)=P_w(y|x)$$

因此，对其进行偏导

$$\begin{aligned}\frac{\partial L(P,w)}{\partial P(y|x)}&=\sum_{x,y}\tilde{P}(x)(\log P(y|x)+1)-\sum_{y}w_0-\sum_{x,y}(\tilde{P}(x)\sum_{i=1}^{n}w_if_i(x,y))\\&=\sum_{x,y}\tilde{P}(x)(\log P(y|x)+1-w_0-\sum_{x,y}w_if_i(x,y))\end{aligned}$$

寻找极限值（令偏导数等于0，在$\tilde{P}(x)>0$的情况下）

$$\begin{aligned}P(y|x)&=\exp(\sum_{i=1}^{n}w_if_i(x,y)+w_0-1)\\&=\frac{\exp(\sum_{i=1}^{n}w_if_i(x,y))}{\exp(1-w_0)}\end{aligned}$$

由于$\sum_y P(y|x)=1$，得

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp(\sum_{i=1}^{n}w_if_i(x,y))$$

其中，

$$Z_w(x)=\sum_{y}\exp(\sum_{i=1}^{n}w_if_i(x,y))$$

- $Z_w(x)$**称为规范化因子；**$f_i(x,y)$**是特征函数；**w_i**是特征的权值**
- $P_w=P_w(y|x)$**就是最大熵模型**
  之后，求解对偶问题的外部的极大化问题

  $$\max_{w}\Psi(w)$$

  将其解记为w^*，即

  $$w^*=\arg\max_{w}\Psi(w)$$

应用最优化算大求对偶函数$\Psi(w)$的极大化，得到w^*。而，$P_{w^*}(y|x)$便是最优模型（最大熵模型）。也就是说，最大熵模型的学习归结为对偶函数$\Psi(w)$的极大化。

- $\Psi(w)$为拉格朗日函数的最小值，而最优模型（最大熵模型）为拉格朗日函数的最小值集合里的最大值

> 可以查看116页的例子6.2

2.4 极大似然估计

已知训练数据的经验概率分布$\tilde{P}(X,Y)$，条件概率分布$P(Y|X)$的对数似然函数表示为

$$
\begin{aligned}
L_{\tilde{P}}(P_w) &= \log\prod_{x,y} P(y|x)^{\tilde{P}(x,y)} \\
&= \sum_{x,y}\tilde{P}(x,y)\log P(y|x) \\
&= \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y) - \sum_{x,y}\tilde{P}(x,y)\log Z_w(x) \\
&= \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y) - \sum_{x}\tilde{P}(x)\log Z_w(x)
\end{aligned}
$$

对偶函数$\Psi(w)$

$$
\begin{aligned}
\Psi(w) = & \sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x) + w_0(1-\sum_{y}P(y|x))+ \\
& \sum_{i=1}^{n} w_i(\sum_{x,y}\tilde{P}(x,y)f_i(x,y) - \sum_{x,y}\tilde{P}(x)P(y|x)f_i(x,y)) \\
&= \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y) + \sum_{x,y}\tilde{P}(x)P_w(y|x)(\log P_w(y|x) - \sum_{i=1}^{n} w_i f_i(x,
\\
&= \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y) - \sum_{x,y}\tilde{P}(x)P_w(y|x)\log Z_w(x) \\
&= \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y) - \sum_{x}\tilde{P}(x)\log Z_w(x)
\end{aligned}
$$

由此可得，

$$
\Psi(w) = L_{\tilde{P}}(P_w)
$$

3 模型学习的最优化算法

- 改进的迭代尺度法
- 梯度下降法
- 牛顿法
- 拟牛顿法

> 过于复杂，自行观看