逻辑回归与最大熵模型

> 逻辑回归和最大熵模型都属于**对数线性模型**（log linear model）

6.1 逻辑斯谛分布

6.1.1 逻辑斯谛分布

- **逻辑斯谛分布**（logistic distribution）：设X是连续随机变量，X服从逻辑斯谛分布是指X具有下列分布函数和概率密度函数

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

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

其中μ为位置参数，$\gamma>0$为形状参数；

- 分布函数图形是一条S型曲线（sigmoid curve），该曲线以点$(\mu, \frac{1}{2})$中心对称，即$F(-x+\mu)-\frac{1}{2}=-F(x+\mu)+\frac{1}{2}$；
- 曲线在中心附件增长速度较快，在两端增长速度较慢（斜率接近0）；
- 形参γ的值越小，曲线在中心附近增长得越快；

6.1.2 二项逻辑斯谛回归模型

- 二项逻辑斯谛回归模型（binomial logistic regression model）：分类模型，由条件概率分布$P(Y|X)$表示，随机变量X取值为实数，Y取值为0或1，且满足

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

其中$x\in\mathbb{R}^n$是输入，$Y\in\{0,1\}$是输出，$w\in\mathbb{R}^n, b\in\mathbb{R}$分别是权重向量参数和偏置参数；

- 该模型比较两个条件概率值的大小，将某个实例x分到概率值大的那一类；
- **模型化简**：$w=(w^{(1)},w^{(2)},\dots,w^{(n)},b)^T, x=(x^{(1)},x^{(2)},\dots,x^{(n)},1)^T$；
- **一个事件的几率**（odds）：是指该事件发生的概率与该事件不发生的概率的比值，如果事件发生概率为p，那么几率是$\frac{p}{1-p}$；该事件的对数几率（log odds）或logit函数是

$$\text{logit}(p)=\log\frac{p}{1-p} \tag{3}$$

- 对于逻辑斯谛回归，有

$$\log\frac{P(Y=1|x)}{1-P(Y=1|x)}=w\cdot x \tag{4}$$

也即$Y=1$的对数几率是输入x的线性函数；

- 逻辑斯谛回归模型将线性函数$w\cdot x$转换为概率，当$w\cdot x$接近正无穷时，概率值为1，接近负无穷时概率值为0；

6.1.3 模型参数估计

- 在给定训练数据集下，可以使用极大似然法估计模型参数，设

$$P(Y=1|x)=\pi(x), P(Y=0|x)=1-\pi(x) \tag{5}$$

- 似然函数为：

$$\prod_{i=1}^{N}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{6}$$

- 对数似然函数为（注意其中$w\in\mathbb{R}^{n+1}, x_i\in\mathbb{R}^{n+1}$）：

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

- 通过求解$L(w)$的极大值得参数估计值$\hat{w}$，常用方法是*梯度下降*及*拟牛顿法*;

6.1.4 多项逻辑斯谛回归

- **多项逻辑斯谛回归模型**（multi-nominal logistic regression model）：用于多类分类，假设离散随机变量Y的取值集合是$\{1,2,\dots,K\}$，那么模型可以表达为

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

其中$w_k\in\mathbb{R}^{n+1}, x\in\mathbb{R}^{n+1}$;

6.2 最大熵模型

6.2.1 最大熵原理

- **最大熵原理**是概率模型学习的一个准则，其认为在学习概率模型时，在所有可能的概率模型（分布）中，*熵最大的模型是最好的模型*（在满足约束条件的模型集合中选取熵最大模型）；
- 假设离散随机变量X的概率分布为$P(X)$，则其熵（同时参照5.2.2节信息增益）是

$$H(P)=-\sum_{i=1}^{N}P(x)\log P(x) \tag{7}$$

且满足下列不等式：

$$0 \leq H(P) \leq \log N \tag{8}$$

当且仅当X的分布是均匀分布时右边的等号成立，即*均匀分布的熵最大*；

- 在没有更多信息的情况下（事实约束），不确定部分都是“等可能的”，最大熵原理通过熵最大化来表示等可能性；

6.2.2 最大熵模型的定义

- 将最大熵原理应用到分类得到**最大熵模型**；
- 假设分类模型是一个条件概率分布$P(Y|X)$，考虑该模型需要满足的条件，给定训练集T，可以确定联合分布$P(X,Y)$的*经验分布*和边缘分布$P(X)$的*经验分布*：

$$\tilde{P}(X=x,Y=y)=\frac{\nu(X=x,Y=y)}{N} \tag{9}$$

$$\tilde{P}(X=x)=\frac{\nu(X=x)}{N} \tag{10}$$

其中$\nu(X=x,Y=y)$训练集中样本(x,y)出现频数，$\nu(X=x)$输入x出现的频数，N是训练集样本容量；

- **特征函数**（feature function）$f(x,y)$描述输入x和输出y之间的某一个*事实*，定义为

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

- 特征函数关于经验分布$\tilde{P}(X,Y)$的期望值，用$\mathbb{E}_{\tilde{P}}(f)$表示：

$$\mathbb{E}_{\tilde{P}}(f)=\sum_{x,y}\tilde{P}(x,y)f(x,y) \tag{11}$$

- 特征函数关于模型$P(Y|X)$与经验分布$\tilde{P}(X)$的期望值，用$\mathbb{E}_{P}(f)$表示：

$$\mathbb{E}_{P}(f)=\sum_{x,y}\tilde{P}(x)P(y|x)f(x,y) \tag{12}$$

- 如果模型能够获取训练数据中的信息，那么可以假设两个期望值相等：$\mathbb{E}_{\tilde{P}}(f)=\mathbb{E}_P(f)$，因此可以作为模型学习的约束条件；
- 假如有n个特征函数$f_i(x,y),i=1,2,\dots,n$，那么就有n个约束条件；

- **最大熵模型**：假设满足所有约束条件的模型集合为

$$\mathcal{C}=\{P\in\mathcal{P}|\mathbb{E}_P(f_i)=\mathbb{E}_{\tilde{P}}(f_i),i=1,2,\dots,n\} \tag{13}$$

定义在条件概率分布$P(Y|X)$上的条件熵为

$$H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x) \tag{14}$$

则模型集合$\mathcal{C}$中条件熵最大的模型称为最大熵模型，式中对数为自然对数；

6.2.3 最大熵模型的学习

- 最大熵模型的学习可以形式化为*约束最优化问题*；
- 给定训练数据集T及特征函数，学习等价的约束最优化问题为：

$$\begin{aligned}\max_{P\in\mathcal{C}}\quad & H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)\\ \text{s.t.}\quad & \mathbb{E}_{\tilde{P}}(f_i)=\mathbb{E}_P(f_i),\quad i=1,2,\dots,n\\ & \textstyle\sum_y P(y|x)=1\end{aligned}$$

- 将求最大值问题改写成等价求最小值问题：

$$\begin{aligned}\min_{P\in\mathcal{C}}\quad & -H(P)=\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)\\ \text{s.t.}\quad & \mathbb{E}_{\tilde{P}}(f_i)-\mathbb{E}_P(f_i)\quad =0,\quad i=1,2,\dots,n\\ & 1-\textstyle\sum_y P(y|x)\quad =0\end{aligned}$$

- 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**$w_0,w_1,\dots,w_n$，定义拉格朗日函数$L(P,w)$：

$$L(P,w)=-H(P)+w_0\left(1-\sum_y P(y|x)\right)+\sum_{i=1}^{n}w_i(\mathbb{E}_{\tilde{P}}(f_i)-\mathbb{E}_P(f_i)) \tag{15}$$

- 最优化原始问题：$\min\limits_{P\in\mathcal{C}}\max\limits_{w}L(P,w)$；对偶问题：$\max\limits_{w}\min\limits_{P\in\mathcal{C}}L(P,w)$；由于拉格朗日函数$L(P,w)$是$P$的凸函数，所以原始问题和对偶问题的解是等价的；

1.求解对偶问题内部的极小化问题

- $\min\limits_{P\in\mathcal{C}} L(P,w)$是$w$的函数，记作

$$\Psi(w)=\min_{P\in\mathcal{C}} L(P,w)=L(P_w,w) \tag{16}$$

$\Psi(w)$称为对偶函数，将其解记作

$$P_w=\arg\min_{P\in\mathcal{C}} L(P,w)=P_w(y|x) \tag{17}$$

- 具体地，求$L(P,w)$对$P(y|x)$的偏导数：

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

令偏导等于0，在$\tilde{P}(x)>0$情况下，解得：

$$P(y|x)=\exp\left(\sum_{i=1}^n w_i f_i(x,y)+w_0-1\right)=\frac{\exp\left(\sum_{i=1}^n w_i f_i(x,y)\right)}{\exp(1-w_0)} \tag{18}$$

- 由于$\sum_y P(y|x)=1$，得

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^n w_i f_i(x,y)\right) \tag{19}$$

其中$Z_w(x)=\sum_y \exp\left(\sum_{i=1}^n w_i f_i(x,y)\right)$，是归一化因子；

- 模型$P_w=P_w(y|x)$就是最大熵模型，其中w是最大熵模型中的参数向量；

2.求解对偶问题外部的极大化问题

- 进一步求解极大化问题$\max\limits_w \Psi(w)$，将其解记为：

$$w^*=\arg\max_w \Psi(w) \tag{20}$$

- 可以应用最优化算法求极大化，得到w^*用来表示$P^*\in\mathcal{C}$，
  $P^*=P_{w^*}=P_{w^*}(y|x)$是学习到的最优模型

6.2.4 极大似然估计

- 对偶函数的极大化等价于*最大熵模型的极大似然估计*;
- 已知训练数据的经验概率分布$\tilde{P}(X,Y)$，条件概率分布$P(Y|X)$的对数似然函数为:

$$L_{\tilde{P}}(P_w) = \log \prod_{x,y} P(y|x)^{\tilde{P}(x,y)} = \sum_{x,y} \tilde{P}(x,y) \log P(y|x) \tag{21}$$

- 当条件概率分布$P(y|x)$是最大熵模型时，对数似然函数$L_{\tilde{P}}(P_w)$为:

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

- 而对偶函数$\Psi(w)$为:

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

- 可见$\Psi(w) = L_{\tilde{P}}(P_w)$，对偶函数等价于对数似然函数，即证明了最大熵模型学习中对偶函数极大化等价最大熵模型的极大似然估计;

6.3 模型学习的最优化算法

> - 上述两个模型最后都可以归结为以似然函数为目标函数的最优化问题，通常采用*迭代算法*求解;
> - 多种最优化方法都适用，常用的有**改进的迭代尺度法**、**梯度下降法**、**牛顿法**或**拟牛顿法**，牛顿法或拟牛顿法收敛速度更快;

6.3.1 改进的迭代尺度法

- **改进的迭代尺度法**（improved iterative scaling， IIS）是一种最大熵模型学习最优化算法；
- 已知最大熵模型为：

$$P_w(y|x) = \frac{1}{Z_w(x)} \exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right) \tag{22}$$
$$Z_w(x) = \sum_y \exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

对数似然函数为：

$$L(w) = \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n} w_i f_i(x,y) - \sum_x \tilde{P}(x) \log Z_w(x) \tag{23}$$

目标是通过极大似然估计学习模型参数w^*；

- **IIS思路**：假设当前参数向量是$w = (w_1, w_2, \ldots, w_n)^T$，希望找到一个新的参数向量$w + \delta = (w_1 + \delta_1, \ldots, w_n + \delta_n)^T$使得模型的对数似然函数值增大，如果该参数更新方法$\tau: w \to w + \delta$存在，那就可以重复使用直到找到最大值；
- 改进的迭代尺度算法IIS：
  - 输入特征函数$f_1, f_2, \ldots, f_n$，经验分布$\tilde{P}(X,Y)$，模型$P_w(y|x)$；
  - 初始化参数值全部为0，$w_i = 0, i = 1, 2, \ldots, n$；
  - 对于每一个$i \in \{1, 2, \ldots, n\}$：
    1. 令δ_i是方程$\sum_{x,y} \tilde{P}(x) P(y|x) f_1(x,y) \exp(\delta_i f^{\#}(x,y)) = \mathbb{E}_{\tilde{P}}(f_i)$的解，其中$\mathbb{E}_{\tilde{P}}(f_i) = \sum_{x,y} \tilde{P}(x,y) f_i(x,y)$，而$f^{\#}(x,y) = \sum_{i=1}^{n} f_i(x,y)$表示所有特征在$(x,y)$出现的次数；
    2. 更新w_i值：$w_i \leftarrow w_i + \delta_i$；
  - 如果不是所有的w_i都收敛，重复上述步骤；
- 如果$f^{\#}(x,y) = M$对于所有的(x,y)都是常数，那么δ_i可以显式表示为

$$\delta_i = \frac{1}{M} \log \frac{\mathbb{E}_{\tilde{P}}(f_i)}{\mathbb{E}_P(f_i)} \tag{24}$$

- 如果$f^{\#}(x,y)$不是常数，必须通过数值计算求δ_i；简单有效的方法是*牛顿法*，通过迭代求得δ_i^*：

$$\delta_i^{(k+1)} = \delta_i^{(k)} - \frac{g(\delta_i^{(k)})}{g'(\delta_i^{(k)})} \tag{25}$$

其中$g(\delta_i) = \sum_{x,y} \tilde{P}(x)P(y|x)f_1(x,y)\exp(\delta_i f^{\#}(x,y)) - \mathbb{E}_{\tilde{P}}(f_i)$，只要选取的初值$\delta_i^{(0)}$适当，由于方程有单根，因此牛顿法恒收敛，$g(\delta_i^*) = 0$；

6.3.3 拟牛顿法

- **拟牛顿法BFGS算法**：
  1. 输入特征函数$f_1, f_2, \ldots, f_n$，经验分布$\tilde{P}(X,Y)$，目标函数$f(w) = -L(w)$，梯度$g(w) = \nabla f(w)$，精度要求ϵ；
  2. 选定初始点$w^{(0)}$，取B_0为正定对称矩阵，$k = 0$；
  3. 如果$\|g(w^{(k)})\| < \epsilon$，停止计算，得$w^* = w^{(k)}$；否则进入下一步；
  4. 由$B_k p_k = -g(w^{(k)})$求得p_k；
  5. 一维搜索：求λ_k使得$f(w^{(k)} + \lambda_k p_k) = \min\limits_{\lambda \geq 0} f(w^{(k)} + \lambda p_k)$；
  6. 置$w^{(k+1)} = w^{(k)} + \lambda_k p_k$；
  7. 如果$\|g(w^{(k+1)})\| < \epsilon$，停止计算，得$w^* = w^{(k+1)}$；否则按照下式求出$B_{k+1}$：

     $$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T \delta_k} - \frac{B_k \delta_k \delta_k^T B_k}{\delta_k^T B_k \delta_k} \tag{26}$$

     其中$y_k = g(w^{(k+1)}) - g(w^{(k)}), \delta_k = w^{(k+1)} - w^{(k)}$；
  8. $k = k + 1$，回到步骤（4）；
- 目标函数：

$$f(w) = \sum_x \tilde{P}(x) \log Z_w(x) - \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^n w_i f_i(x,y) \tag{27}$$

- 梯度：

$$g(w) = \left(\frac{\partial f(w)}{\partial w_1}, \frac{\partial f(w)}{\partial w_2}, \ldots, \frac{\partial f(w)}{\partial w_n} \right)^T \tag{28}$$

其中

$$\frac{\partial f(w)}{\partial w_i} = \sum_{x,y} \tilde{P}(x) P_w(y|x) f_i(x,y) - \mathbb{E}_{\tilde{P}}(f_i), i = 1, 2, \ldots, n \tag{29}$$