第6章 Logistic回归与最大熵模型

导言

- 逻辑斯谛回归(logistic regression) 是统计学习中的经典分类方法
- 最大熵是概率模型学习的一个准则，将其推广到分类问题得到最大熵模型(maximum entropy model)
- 逻辑斯谛回归模型与最大熵模型都属于对数线性模型
  - 二分类
  - 判别函数变换（对数）后符合线性模型

广义线性模型

- 当导出线性回归时, 一种可以考虑的模型是
  - $y = w \cdot x + \xi$，其中ξ是一个均值为零、方差为σ_ξ^2的正态随机变量。
  - 即，线性基础上增加一个非线性项
- 另一种方法是把y看作随机变量Y的值
  - Y具有均值$w \cdot x$和方差σ_ξ^2 ： $Y \sim N(w \cdot x, \sigma_\xi^2)$
  - 即，线性函数和高斯函数的复合函数
- 更一般的**广义线性模型(Generalized Linear Model, GLM)**推广
  - 用其他参数分布代替正态分布, 并用 $w \cdot x$预测该分布的参数
  - 如逻辑斯谛回归模型，其logit 函数

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

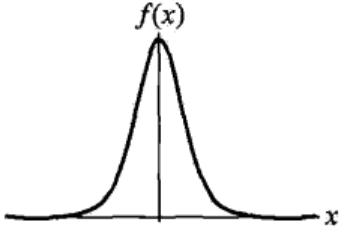
1 Logistic回归模型

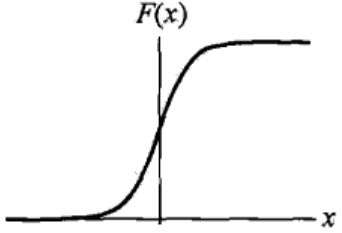
逻辑斯谛分布(Logistic distribution)

【定义6.1 (逻辑斯谛分布)】设X是连续随机变量， X服从Logistic distribution，指X具有分布函数和密度函数

密度函数：$f(x) = F'(x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma\left(1+e^{-(x-\mu)/\gamma}\right)^2}$

分布函数：$F(x) = P(X \le x) = \frac{1}{1+e^{-(x-\mu)/\gamma}} = L(\mu, \gamma)$





其中

μ为位置参数, 散布中心;

$\gamma > 0$为形状参数,表示散布程度，γ越大，散布程度也越大

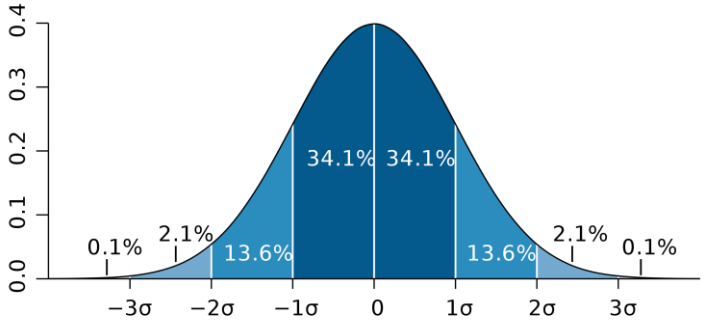
$F(x)$关于$(\mu, 1/2)$中心对称；$f(x)$关于μ对称$f(\mu + x) = f(\mu - x)$

标准的逻辑斯谛分布，记作$L(0,1)$，它的累积分布函数为$F(t) = \frac{1}{1+e^{-t}}$

高斯分布与逻辑斯谛分布

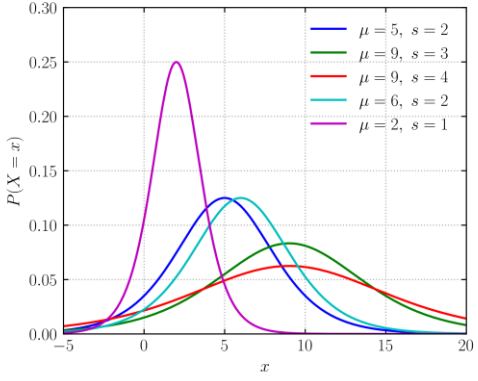
高斯分布(正态分布)

- $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- $X\sim N(\mu,\sigma^2)$
- $f(x)$关于μ对称$f(\mu+x)=f(\mu-x)$
- $N(0,1)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-x^2}$



逻辑斯谛分布

- $f(x)=\frac{\mathrm{e}^{-(x-\mu)/\gamma}}{\gamma\left(1+\mathrm{e}^{-(x-\mu)/\gamma}\right)^2}$
- $X\sim L(\mu,\gamma)$
- $f(x)$关于μ对称$f(\mu+x)=f(\mu-x)$
- $L(0,1)=\frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2}$

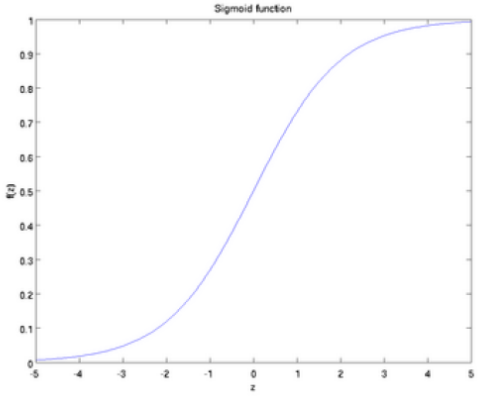


S形函数

- Sigmoid

$$f(z) = \frac{1}{1+\exp(-z)}.$$

$$f'(z) = f(z)(1-f(z))$$

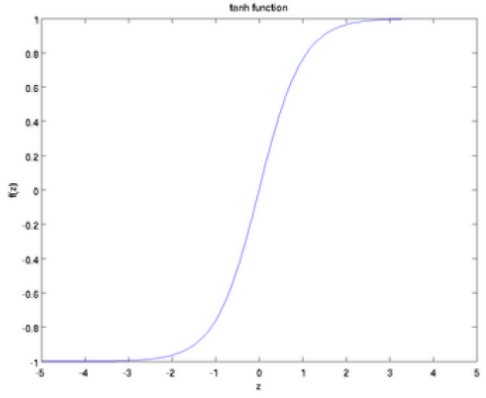


$[0, 1]$

- 双曲正切函数(tanh)

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$f'(z) = 1 - (f(z))^2$$



$[-1, 1]$

逻辑斯谛回归模型

逻辑斯谛回归模型的二分类分布

- 线性函数和逻辑斯谛函数的复核函数
- 的对数几率是线性函数

二项逻辑斯谛回归模型

由条件概率$P(\mathrm{Y}|\mathrm{X})$表示的分类模型，形式化为logistic distribution

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

【定义6.2 (逻辑斯谛回归模型)】 二项逻辑斯谛回归模型是如下的条件概率分布

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

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

这里，$x\in\mathbf{R}^n, Y\in\{0,1\}, w\in\mathbf{R}^n, b\in\mathbf{R}$，$w$为权值向量，$b$为偏置$P(Y=0\mid x)$

为线性函数和逻辑斯谛分布函数的复合函数：$F(x)=P(X\leq x)=\frac{1}{1+\mathrm{e}^{-\frac{x-\mu}{\gamma}}}$

二项逻辑斯谛回归模型

为了方便，扩充齐次$w=\left(w^{(1)},w^{(2)},\ldots,w^{(n)},b\right)^{\mathrm{T}}$，$x=\left(x^{(1)},x^{(2)},\ldots,x^{(n)},1\right)^{\mathrm{T}}$

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

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

二项逻辑斯谛回归模型

- **事件的几率**odds(事件发生与事件不发生的概率之比)：$\frac{p}{1-p}$
- **对数几率**(logit函数)： $\text{logit}(p) = \log\frac{p}{1-p}$
- 逻辑斯谛回归的logit 函数：

$$\text{logit}\big(P(Y=1 \mid x)\big) = \log\frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)} = w \cdot x$$

- 输出$Y=1$的对数几率，是由输入x的线性函数表示的模型，即逻辑斯谛回归模型

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

模型参数估计 - 似然函数

【算法】逻辑斯谛回归模型，训练集$T=\{(x_1,y_1),\cdots,(x_N,y_N)\}$，$x_i\in\mathbf{R}^n, y_i\in\{0,1\}$，极大似然估计法估计模型参数（$\pi(x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)}$中的$w$）

设$P(Y=1\mid x)=\pi(x)$, $P(Y=0\mid x)=1-\pi(x)$，似然函数为

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

模型估计目标：求出使这一似然函数的值最大的参数估计$\hat{w}$

【注】出现(x_i,y_i)的概率的统一形式：$[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$

$$y_i=1,\quad \pi(x_i)=[\pi(x_i)]^{y_i}=[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$$

$$y_i=0,\quad 1-\pi(x_i)=[1-\pi(x_i)]^{1-y_i}=[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$$

模型参数估计

对数似然函数$L(w)$：

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

【注】第二步第一项包含对数几率

对$L(w)$求极大值，得到w的估计值。采用梯度下降及拟牛顿法【常规优化问题】

设w的极大似然估计值为$\hat{w}$，模型条件概率为

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

多项logistic回归

【定义】设Y的取值集合为$\{1,2,\cdots,K\}$，**多项logistic回归模型**

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

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

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

2 最大熵模型

熵最大的模型是最好的模型

模型：分类模型以条件概率$P(Y \mid X)$ 输出 Y

满足约束条件的，$H(P) = -\sum_{x,y} \tilde{P}(x)P(y \mid x)\log P(y \mid x)$最大熵的模型

最大熵模型

【定义】最大熵模型(Maximum Entropy Model)由最大熵原理推导实现

【定义】最大熵原理

- 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
- 即，在满足约束条件的模型集合中，应选取熵最大的模型

- **【注】在没有更多信息的情况下，最大的不确定性（熵最大）为各种情况“等可能”（对各种情况都考虑到，且没有偏差，因此通用性最强）**
  - **最大熵原理通过熵的最大化来表示等可能性。通过最大化熵的数值指标实现“等可能”**

【定义】假设离散随机变量X的概率分布是$P(X)$， X的熵

$$H(P)=\sum_{x}P(x)(-\log P(x))=\mathrm{E}(-logP(x))$$

【性质】 $0\leqslant H(P)\leqslant \log|X|$，$|X|$表示$X$个数，当且仅当$X$的分布是均匀分布时右边等号成立

例6.1

最大熵模型的定义

【算法】输入$X \in \mathcal{X} \subseteq \mathbf{R}^n$, $\mathcal{X}$为输入集合；输出$Y \in \mathcal{Y}$，$\mathcal{Y}$为输出集合

分类模型：对于给定的输入 X，**分类模型以条件概率$P(Y \mid X)$ 输出 Y**

训练数据集：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$；

学习目标：用最大熵原理选择最好的分类模型$P(y \mid x)$

1)首先，确定联合分布 $P(X, Y)$ 的经验分布$\tilde{P}(X, Y)$和边缘分布 $P(X)$ 的经验分布$\tilde{P}(X)$

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

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

其中, $\nu(X = x, Y = y)$ 表示 (x, y) 出现的频数, $\nu(X = x)$ 表示x 出现的频数

最大熵模型的定义

2) 用特征函数 (feature function)$f(x,y)$来描述输入和输出之间的约束

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

用特征函数来评估模型

特征函数$f(x,y)$在模型$\tilde{P}$上关于经验分布$\tilde{P}(X,Y)$的期望值(根据训练数据得到的经验特征期望)

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

特征函数$f(x,y)$关于$P(x)P(y|x)$的期望值(模型上的理论特征期望)

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

$E_P(f)=\sum_{x,y}P(x,y)f(x,y)\ =\sum_{x,y}P(x)P(y\mid x)f(x,y)$，$P(x)$未知，用$\tilde{P}(x)$近似

如果某个模型能够获取训练数据中的信息，两个期望值相等

$$E_P(f)=E_{\tilde{P}}(f)$$

如果有n个特征函数$f_i(x,y),i=1,2,\cdots,n$，那么对应n个约束条件

【注】约束不能保证完全满足，所以需要采用期望。

最大熵模型

【定义6.3 最大熵模型】假设满足所有约束条件的模型(形为$P(Y|X)$) 集合为

$$\mathcal{C} \equiv \{P \in \mathcal{P} \mid E_P(f_i) = E_{\tilde{P}}(f_i), i = 1,2,\cdots,n\}$$

定义在条件概率分布$P(Y|X)$上的条件熵：

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

则模型集合$\mathcal{C}$中条件熵$H(P)$最大的模型称为**最大熵模型**

【注】 模型求解即为计算$P(y \mid x)$

【注】条件熵为条件信息量$(-\log P(y \mid x))$的期望

$$\sum_{x,y} P(x,y)(-\log P(y \mid x))$$

最大熵模型学习求解

最大值带约束=>最小值带约束=>无约束对偶问题

拉格朗日对偶问题求解

最大熵模型学习 - 原始最大值问题(带约束)

【原始问题】对于给定的数据集以及特征函数f_i，最大熵模型的学习等价于约束最优化问题

$$\max_{P \in \mathbf{C}} \quad H(P) = -\sum_{x,y} \tilde{P}(x) P(y \mid x) \log P(y \mid x)$$

$$\text{s.t.} \quad E_P(f_i) = E_{\tilde{P}}(f_i), i = 1,2,\cdots,n$$

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

求解技术路线：**最大值带约束**=>最小值带约束=>无约束对偶问题

最大熵模型学习 - 最小值带约束

求解技术路线：最大值带约束=>**最小值带约束**=>无约束对偶问题

按照最优化问题惯例改写为最小值问题

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

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

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

最大熵模型学习 - 拉格朗日函数(无约束)

求解技术路线：最大值带约束=>最小值带约束=>**无约束对偶问题**

引进拉格朗日乘子$w_0, w_1, \cdots, w_n$，定义拉格朗日函数

$$L(P,w) \equiv -H(P) + w_0\left(1 - \sum_y P(y \mid x)\right) + \sum_{i=1}^{n} w_i\left(E_{\tilde{P}}(f_i) - E_P(f_i)\right)$$

$$= \sum_{x,y} \tilde{P}(x)P(y \mid x)\log P(y \mid x) + w_0\left(1 - \sum_y P(y \mid x)\right) +$$

$$\sum_{i=1}^{n} w_i\left(\sum_{x,y} \tilde{P}(x,y)f_i(x,y) - \sum_{x,y} \tilde{P}(x)P(y \mid x)f_i(x,y)\right)$$

可以证明(附录C)，**约束最优化的原始问题可以转换成无约束最优化问题**

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

最大熵模型学习 - 拉格朗日对偶问题

*拉格朗日原始问题转换到对偶问题：$\min_{P\in\mathbf{C}}\max_{w} L(P,w) \Rightarrow \max_{w}\min_{P\in\mathbf{C}} L(P,w)$

【注】为什么要转成对偶问题？拉格朗日原始问题求偏导，回到带约束的原始问题

$L(P,w)$是P的凸函数，原始问题和对偶问题的解是等价的(附录C)

拉格朗日对偶问题 - $\min_{\mathbf{P}\in\mathbf{C}}\mathbf{L}(\mathbf{P},\mathbf{w})$

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

1)先求极小化问题$\min_{P\in\mathbf{C}}L(P,w)$(结果为$w$的函数)

记$\Psi(w)=\min_{P\in\mathbf{C}}L(P,w)=L(P_w,w)$，$\Psi(w)$称为对偶函数

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

拉格朗日对偶问题求解 - $\min_{P \in C} L(P, w)$

$$L(P,w) = \sum_{x,y} \tilde{P}(x) P(y \mid x) \log P(y \mid x) + w_0 \left(1 - \sum_y P(y \mid x)\right) + \sum_{i=1}^{n} w_i \left(\sum_{x,y} \tilde{P}(x,y) f_i(x,y) - \sum_{x,y} \tilde{P}(x) P(y \mid x) f_i(x,y)\right)$$

固定w_i，对每个分类Y，$L(P,w)$对$P(y \mid x)$偏导

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

$$= \sum_{x,y} \tilde{P}(x) \left(\log P(y \mid x) + 1 - w_0 - \sum_{i=1}^{n} w_i f_i(x,y)\right)$$

令对$P(y \mid x)$偏导为0，在$\tilde{P}(x) > 0$时：【注】??此处存疑

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

【注1】见例6. 2， $P(y \mid x)$的求解应计算所有的$P(y_j \mid x_i)$，即对所有$P(y_j \mid x_i)$求偏导

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

令$\frac{\partial L(P,w)}{\partial P(y|x)} = 0$，因为$\tilde{P}(x) > 0$，所以$\log P(y \mid x) + 1 - w_0 - \sum_{i=1}^{n} w_i f_i(x,y) = 0$

【注2】 $\sum_{x,y} \tilde{P}(x)(w_0) = \sum_y w_0$， $\exp(w_0 - 1) = \frac{1}{\exp(1 - w_0)}$

拉格朗日对偶问题求解 - $\min_{P \in C} L(P, w)$

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

上式等式两侧对y 求和，由$\sum_y P(y \mid x) = 1$，得

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

$$\exp(1 - w_0) = \sum_y \exp\left(\sum_{i=1}^{n} w_i f_i(x, y)\right)$$

$$P(y \mid x) = \frac{\exp(\sum_{i=1}^{n} w_i f_i(x, y))}{\sum_y \exp(\sum_{i=1}^{n} w_i f_i(x, y))}$$

因此，得到最大熵模型 $P_w(y \mid x) = \frac{1}{Z_w(x)} \exp(\sum_{i=1}^{n} w_i f_i(x, y))$

它是w_i的函数，其中规范化因子 $Z_w(x) = \sum_y \exp(\sum_{i=1}^{n} w_i f_i(x, y))$

拉格朗日对偶问题求解 - $\max_{w} \min_{P \in C} L(P, w)$

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

2) 将$P_w(y \mid x)$代入对偶函数$\Psi(w) = \min_{P \in \mathbf{C}} L(P, w)$

求解极大化问题：$\max_{w} \Psi(w)$

其解$w^* = \arg \max_{w} \Psi(w)$

$P^* = P_{w^*} = P_{w^*}(y \mid x)$是学习到的最优模型(最大熵模型)，即

$$P_{w^*}(y \mid x) = \frac{1}{Z_{w^*}(x)} \exp\left(\sum_{i=1}^{n} w_i^* f_i(x, y)\right)$$

其中，$Z_{w^*}(x) = \sum_y \exp(\sum_{i=1}^{n} w_i^* f_i(x, y))$

【注】最大熵模型就是此处的$P_{w^*}(y \mid x)$

例6.2

似然函数与概率

【定义】似然函数(likelihood function）是一种关于统计模型中的参数的函数，表示模型参数中的似然性(likelihood），指**某种事件发生的可能性**

给定联合样本值 $\mathbf{x}$ 下关于 (未知)参数 θ 的函数 $L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta)$

这里的 $\mathbf{x}$ 是指联合样本随机变量 $\mathbf{X}$ 取到的值，即 $\mathbf{X} = \mathbf{x}$； θ 是指未知参数，它属于参数空间；

$f(\mathbf{x} \mid \theta)$ 是一个密度函数，表示(给定) θ下关于联合样本值$\mathbf{x}$的联合密度函数。

在数学中，概率(probability)符合柯尔莫果洛夫公理 (Kolmogorov axioms)的一种数学对象。在数理统计中

- “概率”描述了给定模型参数后，描述结果的合理性，而不涉及任何观察到的数据
- “似然”描述了给定了特定观测值后，描述模型参数是否合理。

最大熵模型参数求解：对偶函数的极大化与极大似然估计

【最大熵模型】根据最大熵模型定义，得约束最优化问题，转换为拉格朗日函数无约束问题。再转换为对偶问题的最优化。求解分两步：

首先，求解带有参数w的条件概率，即带参数的最大熵模型。 $P_w = \arg \min_{P \in \mathbf{C}} L(P, w) = P_w(y \mid x)$

然后，求解参数w，代入P_w。参数求解有两种方法

【算法1-对偶函数极大化】根据P_w，计算对偶函数$\Psi(w) = \min_{P \in \mathbf{C}} L(P, w) = L(P_w, w)$。然后，对偶函数极大化： $w^* = \arg \max_w \Psi(w)$，求出参数$w^*$

【算法2-最大熵模型的极大似然估计】计算$P_w(y \mid x)$的对数似然函数$L_{\tilde{P}}(P_w)$，然后似然函数极大化，求解参数，求出参数w^*

【性质】对偶函数的极大化等价于极大似然估计，即：$\Psi(w) = L_{\tilde{P}}(P_w)$

最大熵模型的极大似然估计

【定义】最大熵模型的极大似然估计：极大化$P_w(y \mid x)$的对数似然函数，求解参数

首先，计算极大似然估计(以$P(Y \mid X)$作为观测值概率的似然函数）。已知训练数据的经验概率分布$\tilde{P}(X,Y)$，条件概率分布$P(Y \mid X)$的对数似然函数$L_{\tilde{P}}(P_w)$ （推导见下页）

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

当条件概率分布$P(Y \mid X)$是最大熵模型时，即$P(Y \mid X)$为$P_{w^*}(y \mid x)$，代入$L_{\tilde{P}}(P_w)$

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

对数似然函数计算推导

【推导】$L_{\tilde{P}}(P_w) = \log \prod_{x,y} P(y \mid x)^{\tilde{P}(x,y)} = \sum_{x,y} \tilde{P}(x,y) \log P(y \mid x)$

训练集$T = \{(x_1, y_1), \ldots, (x_n, y_n)\}$，似然函数：$L_{\tilde{P}}(P_w) = \log \prod_{x,y} P(y \mid x)^{\tilde{P}(x,y)}$

设T中存在k个不同值样本$\{(v_i, w_i)\}$，$C[(X,Y) = (v_i, w_i)]$表示样本值(v_i, w_i)的频数，似然函数$L_{\tilde{P}}(P_w) = \log \prod_{i=1}^{k} P(w_i \mid v_i)^{C[(X,Y)=(v_i,w_i)]}$

等号两边同时开n次方，可得：

$$L_{\tilde{P}}(P_w)^{\frac{1}{n}} = \log \prod_{i=1}^{k} P(w_i \mid v_i)^{\frac{C[(X,Y)=(v_i,w_i)]}{n}}$$

而经验概率分布$\tilde{P}(X = v_i, Y = w_i) = \frac{C[(X,Y)=(v_i,w_i)]}{n}$，上式可表示为

$$L_{\tilde{P}}(P_w)^{\frac{1}{n}} = \log \prod_{i=1}^{k} P(w_i \mid v_i)^{\frac{C[(X,Y)=(v_i,w_i)]}{n}} = \log \prod_{x,y} P(y \mid x)^{\tilde{P}(x,y)}$$

对偶函数$\Psi(w)$的优化方式

对偶函数：$\Psi(w)=\min_{P\in\mathbf{C}} L(P,w)=L(P_w,w)$

$$\Psi(w)=\sum_{x,y}\tilde{P}(x)P_w(y\mid x)\log P_w(y\mid x)+$$

$$\sum_{i=1}^{n}w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x,y}\tilde{P}(x)P_w(y\mid x)f_i(x,y)\right)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_if_i(x,y)+\sum_{x,y}\tilde{P}(x)P_w(y\mid x)\left(\log P_w(y\mid x)-\sum_{i=1}^{n}w_if_i(x,y)\right.$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_if_i(x,y)-\sum_{x,y}\tilde{P}(x)P_w(y\mid x)\log Z_w(x)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_if_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x)$$

【注】最后一步用到$\sum_y P(y\mid x)=1$

所以，对偶函数$\Psi(w)$等价于对数似然函数$L_{\tilde{P}}(P_w)$；

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

最大熵模型

- 将最大熵模型写成更一般的形式：

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

- 最大熵模型与逻辑斯缔回归模型形式类似，对数扩展的线性模型
  - 模型学习在给定的训练数据集进行极大似然估计

3 模型学习的最优化算法

模型学习的最优化算法

- 逻辑斯谛回归模型、最大熵模型学习归结为以似然函数为目标函数的最优化问题，通常通过迭代算法求解，它是光滑的凸函数，因此多种最优化的方法都适用
- 常用的方法有【附录】
  - 改进的迭代尺度法
  - 梯度下降法
  - 牛顿法
  - 拟牛顿法