第七章：逻辑斯谛回归和最大熵模型

Ch7: Logistic regression & Maximum entropy model

概述

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**。

本章首先介绍逻辑斯谛回归模型，然后介绍最大熵模型，最后讲述逻辑斯谛回归与最大熵模型的学习算法。

回归和分类

BASIS FOR COMPARISON	CLASSIFICATION	REGRESSION
Basic	The discovery of model or functions where the mapping of objects is done into predefined **classes**.	A devised model in which the mapping of objects is done into **values**.
Involves prediction of	Discrete values	Continuous values
Algorithms	Decision tree, logistic regression, etc.	Regression tree (Random forest), linear regression, etc.
Nature of the predicted data	Unordered	Ordered
Method of calculation	Measuring accuracy	Measurement of root mean square error

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Logistic distribution，

分布函数：$F(x)=P(X\leqslant x)=\dfrac{1}{1+\mathrm{e}^{-(x-\mu)/\gamma}}$

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

μ为位置参数，γ>0为形状参数

关于 (μ, 1/2)中心对称：$F(-x+\mu)-\dfrac{1}{2}=-F(x-\mu)+\dfrac{1}{2}$

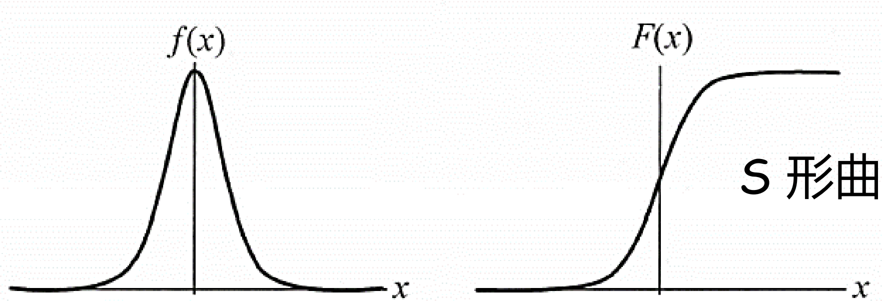


图 6.1　逻辑斯谛分布的密度函数与分布函数

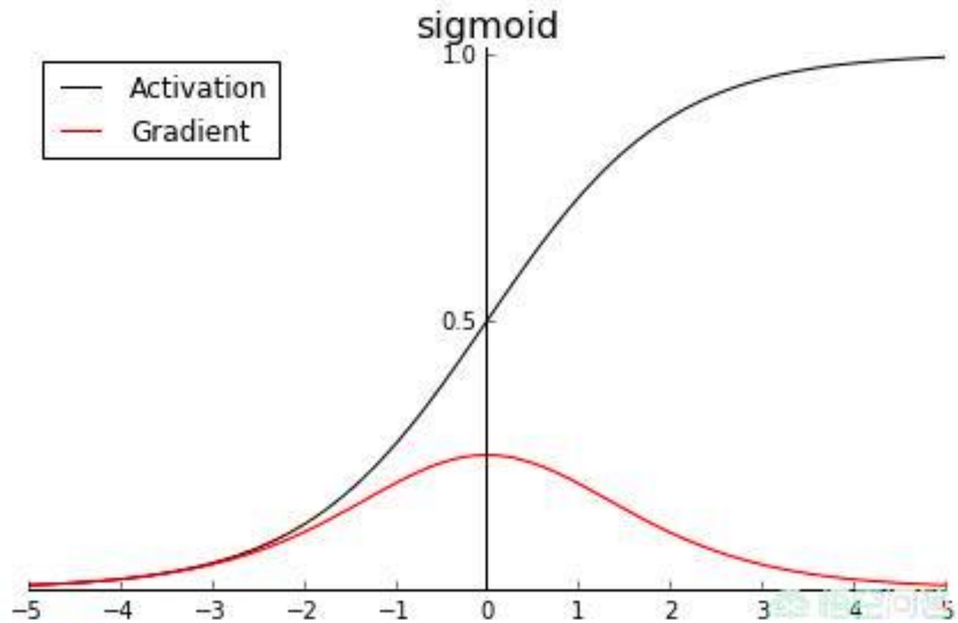
S 形曲线-Sigmoid curve

逻辑斯谛分布

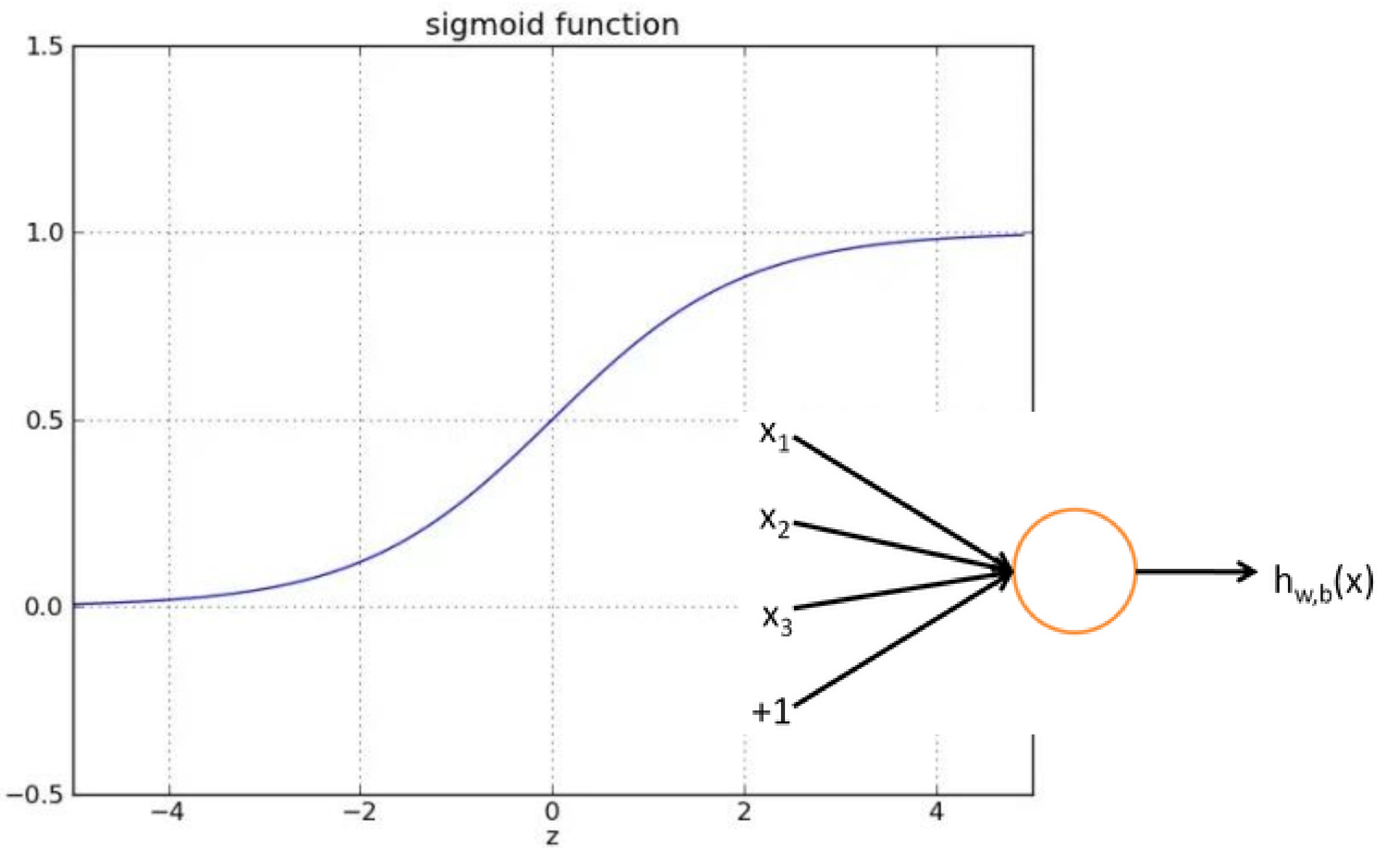
- **Sigmoid:**

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

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



逻辑斯谛分布



Sigmoid function:

$$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{3} W_i x_i + b), \quad f(z) = \frac{1}{1+\exp(-z)}$$

```
def sigmoid(X):
    return 1.0/(1+exp(-X))
```

二项逻辑斯蒂回归

Binomial logistic regression model

由条件概率*P*(*Y*|*X*)表示的分类模型，形式化为**Logistic distribution**

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：

$$P(Y=1|x)=\frac{\exp(w \bullet x+b)}{1+\exp(w \bullet x+b)} \tag{6.3}$$

$$P(Y=0|x)=\frac{1}{1+\exp(w \bullet x+b)} \tag{6.4}$$

这里，$x \in \mathbf{R}^n$ 是输入，$Y \in \{0,1\}$ 是输出，$w \in \mathbf{R}^n$ 和 $b \in \mathbf{R}$ 是参数，w 称为权值向量，b 称为偏置，$w \cdot x$ 为 w 和 x 的内积。

$$w=(w^{(1)},w^{(2)},\cdots,w^{(n)},b)^{\mathrm{T}} \quad x=(x^{(1)},x^{(2)},\cdots,x^{(n)},1)^{\mathrm{T}}$$

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

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

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

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

Logistic Regression vs. Linear Regression

Linear Regression: 输出一个标量wx+b，这个值是连续值，所以可以用来处理回归问题。

Logistic Regression: 把上面的wx+b通过sigmoid函数映射到(0，1)上，并划分一个阈值，大于阈值的分为一类，小于等于分为另一类，可以用来处理二分类问题。

二项逻辑斯蒂回归

事件的几率（odds）：事件发生与事件不发生的概率之比为：

$$\boxed{\frac{p}{1-p}}$$

称为事件的发生比 (the odds of experiencing an event)

对数几率：$\boxed{\operatorname{logit}(p)=\log\frac{p}{1-p}}$

对逻辑斯蒂回归：$\boxed{\log\frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)}=w \cdot x}$

问题: 如何确定w值呢?

答案：通过极大似然函数估计获得，并且Y~f(x;w)

似然函数

- **似然函数**是统计模型中**参数的函数**。给定输出x时，关于参数θ的似然函数$L(\theta|x)$（在数值上）等于给定参数θ后变量X的概率：

$$L(\theta|x) = P(X=x|\theta)$$

- 似然函数的重要性不是它的取值，而是当**参数变化时，概率密度函数到底是变大还是变小**。
- 极大似然函数：**似然函数取得最大值表示相应的参数能够使得统计模型最为合理。**

似然函数

逻辑斯谛回归模型学习时，对于给定的训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$，其中 $x_i \in \mathrm{R}^n$， $y_i \in \{0,1\}$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：$P(Y=1|x)=\pi(x), \quad P(Y=0|x)=1-\pi(x)$

其联合概率密度函数，即似然函数$L(w)$为：

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

目标：求出使这一**似然函数的值最大**的参数，$w_1,w_2,\dots,w_n$，使得$L(w)$取得最大值。

模型参数估计

对数似然函数：$L(w)=\sum_{i=1}^{N}[y_i\log\pi(x_i)+(1-y_i)\log(1-\pi(x_i))]$

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

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

对$L(w)$求极大值，得到w的估计值。

通常采用**梯度下降法**及**拟牛顿法**，学到的模型：

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

多项logistic回归

- 设Y的取值集合为：$\{1,2,\cdots,K\}$

- 多项Logistic回归模型

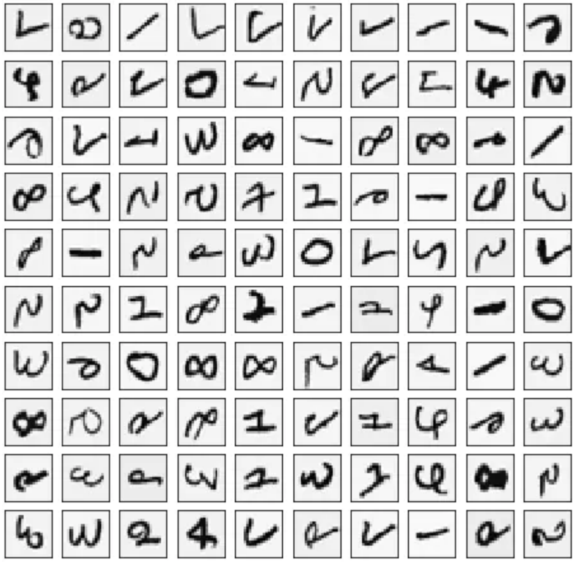
$$P(Y=k \mid x)=\frac{\exp(w_k \cdot x)}{1+\sum_{k=1}^{K-1}\exp(w_k \cdot x)}, \quad k=1,2,\cdots,K-1$$

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

多项logistic回归

对于N分类问题，则是先得到N组w值不同的wx+b，然后归一化，比如用softmax函数，最后变成N个类上的概率，可以处理多分类问题。

Softmax函数：$f(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$



应用：数字手写识别

https://zhuanlan.zhihu.com/p/43686373

最大熵模型

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

最大熵原理：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，表述为在满足约束条件的模型集合中选取熵最大的模型。

假设离散随机变量X的概率分布是P(X)，

熵：$H(P) = -\sum_{x} P(x) \log P(x)$

且：$0 \leqslant H(P) \leqslant \log |X|$

|X|是X的取值个数，X均匀分布时右边等号成立。

例子

假设随机变量X有5个取值{A，B，C，D，E}，估计各个值的概率。

解：满足， $P(A)+P(B)+P(C)+P(D)+P(E)=1$

等概率估计：$P(A)=P(B)=P(C)=P(D)=P(E)=\frac{1}{5}$

加入一些先验：
$$P(A)+P(B)=\frac{3}{10}$$
$$P(A)+P(B)+P(C)+P(D)+P(E)=1$$

于是：
$$P(A)=P(B)=\frac{3}{20}$$
$$P(C)=P(D)=P(E)=\frac{7}{30}$$

如果还有第3个约束条件：
$$P(A)+P(C)=\frac{1}{2}$$
$$P(A)+P(B)=\frac{3}{10}$$
$$P(A)+P(B)+P(C)+P(D)+P(E)=1$$

最大熵模型

X和Y分别是输入和输出的集合，这个模型表示的是对于给定的输入X，以条件概率P(Y|X)输出Y。

给定数据集：$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

联合分布P(X, Y)的经验分布：$\tilde{P}(X,Y) \Rightarrow \tilde{P}(X=x,Y=y)=\frac{\nu(X=x,Y=y)}{N}$

边缘分布P(X)的经验分布：$\tilde{P}(X) \Rightarrow \tilde{P}(X=x)=\frac{\nu(X=x)}{N}$

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

最大熵模型

特征函数$f(x, y)$关于经验分布 $\tilde{P}(X,Y)$ 的期望值:

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

特征函数$f(x, y)$关于模型$P(Y|X)$与经验分布 $\tilde{P}(X)$ 的期望值:

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

如果模型能够获取训练数据中的信息，那么就可以假设这两个期望值相等，即：

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

假设有n个特征函数： $f_i(x,y)$， $i=1,2,\cdots,n$

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C中条件熵H(P)最大的模型称为最大熵模型

条件熵的回顾：

H(X|Y)定义为在给定条件Y下，X的条件概率分布的熵对Y的数学期望：

$$\begin{aligned} H(X|Y) &= \sum_y p(y) H(X|Y=y) \\ &= -\sum_y p(y) \sum_x p(x|y) log(p(x|y)) \\ &= -\sum_y \sum_x p(x,y) log(p(x|y)) \\ &= -\sum_{x,y} p(x,y) log(p(x|y)) \end{aligned}$$

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。
- 对于给定的数据集以及特征函数：$f_i(x,y)$
- 最大熵模型的学习等价于约束最优化问题:

$$
\begin{aligned}
\max_{P\in\mathbf{C}} \quad & H(P) = -\sum_{x,y} \tilde{P}(x)P(y\mid x)\log P(y\mid x) \\
\text{s.t.} \quad & E_P(f_i) = E_{\tilde{P}}(f_i), i = 1,2,\cdots,n \\
& \sum_y P(y\mid x) = 1
\end{aligned}
$$

$$
\begin{aligned}
& \min_{P\in\mathbf{C}} -H(P) = \sum_{x,y} \tilde{P}(x)P(y\mid x)\log P(y\mid x) \\
& \text{s.t.} E_P(f_i) - E_{\tilde{P}}(f_i) = 0, i = 1,2,\cdots,n \\
& \sum_y P(y\mid x) = 1
\end{aligned}
$$

最大熵模型的学习

- 约束最优化的原始问题→无约束的原始问题(拉格朗日问题)

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

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

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

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

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

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

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

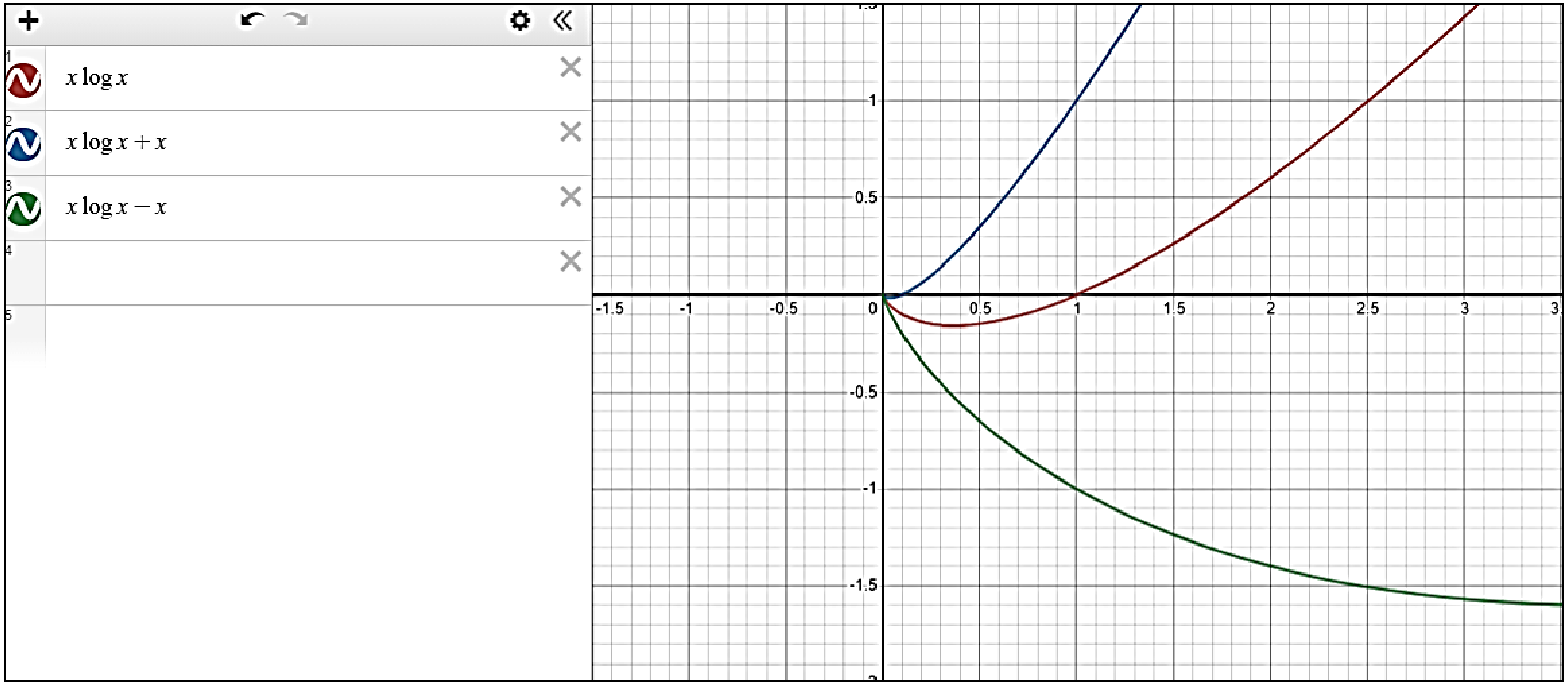
$$+\sum_{i=1}^n w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x,y}\tilde{P}(x)P(y\mid x)f_i(x,y)\right)$$

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

由于拉格朗日函数*L*(P，ω)是P的凸函数，所以原始问题的解与对偶问题的解是等价的。 **(回顾SVM模型)**

最大熵模型的学习

$L(P，\omega)$是P的凸函数



最大熵模型的学习

- 最优化原始问题到对偶问题：

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

- 先求内部极小化问题：$\min_{P\in\mathbf{C}} L(P,w)$ 是w的函数，

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

- $\psi(\omega)$ 称为对偶函数。同时，将其解记作

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

最大熵模型的学习

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

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

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

求L(P, w)对P(y|x)的偏导数:

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

$$= \sum_{x,y}\tilde{P}(x)\left(\log P(y\mid x) + 1 - w_0 - \sum_{i=1}^{n} w_i f_i(x,y)\right) = 0$$

得:

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

最大熵模型的学习

由：$\sum_y P(y|x)=1$

得：$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^n w_i f_i(x,y)\right) \qquad (6.22)$$

规范化因子：$$Z_w(x)=\sum_y \exp\left(\sum_{i=1}^n w_i f_i(x,y)\right) \qquad (6.23)$$

模型 $P_w = P_w(y|x)$ 就是**最大熵模型，** **w为参数向量**

求解对偶问题外部的极大化问题：

$$\max_w \Psi(w)$$

解记为，$w^* = \arg\max_w \Psi(w)$

最优模型，$P^* = P_{w^*} = P_{w^*}(y|x)$

例子

- 原例子中的最大熵模型：

$$\min \quad -H(P)=\sum_{i=1}^{5}P(y_i)\log P(y_i)$$

$$\text{s.t.} \quad P(y_1)+P(y_2)=\tilde{P}(y_1)+\tilde{P}(y_2)=\frac{3}{10}$$

$$\sum_{i=1}^{5}P(y_i)=\sum_{i=1}^{5}\tilde{P}(y_i)=1$$

$$L(P,w)=\sum_{i=1}^{5}P(y_i)\log P(y_i)+w_1\left(P(y_1)+P(y_2)-\frac{3}{10}\right)+w_0\left(\sum_{i=1}^{5}P(y_i)-1\right)$$

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

例子

$$\frac{\partial L(P,w)}{\partial P(y_1)}=1+\log P(y_1)+w_1+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_2)}=1+\log P(y_2)+w_1+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_3)}=1+\log P(y_3)+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_4)}=1+\log P(y_4)+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_5)}=1+\log P(y_5)+w_0$$

解得：

$$P(y_1)=P(y_2)=\mathrm{e}^{-w_1-w_0-1}$$

$$P(y_3)=P(y_4)=P(y_5)=\mathrm{e}^{-w_0-1}$$

例子

- 得：

$$\min_P L(P,w) = L(P_w,w) = -2e^{-w_1-w_0-1} - 3e^{-w_0-1} - \frac{3}{10}w_1 - w_0$$

$$\max_w L(P_w,w) = -2e^{-w_1-w_0-1} - 3e^{-w_0-1} - \frac{3}{10}w_1 - w_0$$

- 对w_i求偏导并令为0：

$$\boxed{\begin{aligned} e^{-w_1-w_0-1} &= \frac{3}{20} \\ e^{-w_0-1} &= \frac{7}{30} \end{aligned}} \Rightarrow \boxed{\begin{aligned} P(y_1) = P(y_2) &= \frac{3}{20} \\ P(y_3) = P(y_4) = P(y_5) &= \frac{7}{30} \end{aligned}}$$

极大似然估计

最大熵模型就是(6.22), (6.23)表示的条件概率分布，

$$P_w(y|x) = \frac{1}{Z_w(x)} \exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right) \tag{6.22}$$

$$Z_w(x) = \sum_y \exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right) \tag{6.23}$$

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：

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

极大似然估计

- **证明：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$P(Y|X)$的对数似然函数表示为：

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

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

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

$$\sum_{x,y}\tilde{P}(x,y)\log Z_w(x)$$

$$=\sum_{x,y}\tilde{P}(x)P(y/x)\log Z_w(x)$$

$$=\sum_{x}\tilde{P}(x)\boxed{\sum_{y}P(y/x)}\log Z_w(x)$$

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)\quad(6.22)$$

$$Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)\quad(6.23)$$

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

极大似然估计

而对偶函数：

$$\Psi(w)=\sum_{x,y}\tilde{P}(x)P_w(y\mid x)\log P_w(y\mid x) \qquad w_0(1-\sum_y P_w(y|x))$$

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

(参见 6.17)

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

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

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x) \qquad \textbf{(6.27)}$$

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

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

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x) \qquad (6.26)$$

最大熵模型的极大似然估计

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

最大熵模型的对偶函数极大化 $\Longleftrightarrow$ **最大熵模型的极大似然估计**

极大似然估计

最大熵模型的学习问题可转换为具体求解**对数似然函数极大化**或**对偶函数极大化**的问题。

最大熵模型更一般的形式：

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right) \tag{6.28}$$

其中，

$$Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right) \tag{6.29}$$

这里，$x\in\mathbf{R}^n$ 为输入，$y\in\{1,2,\cdots,K\}$ 为输出，$w\in\mathbf{R}^n$ 为权值向量，$f_i(x,y)$，$i=1,2,\cdots,n$ 为任意实值特征函数。

最大熵模型与逻辑斯谛回归模型有类似的形式，它们又称为**对数线性模型**(log linear model)。模型学习就是在给定的训练数据条件下对模型进行极大似然估计或正则化的极大似然估计。

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：
  - 梯度下降法
  - 改进的迭代尺度法
  - 牛顿法
  - 拟牛顿法

梯度下降法

- 梯度下降法(gradient descent)
- 最速下降法(steepest descent)
- 梯度下降法是一种迭代算法。选取适当的初值$x^{(0)}$，不断迭代，更新x的值，进行目标函数的极小化，直到收敛。由于负梯度方向是使函数值下降最快的方向，在迭代的每一步，以负梯度方向更新x的值，从而达到减少函数值的目的。

梯度下降法

假设$f(x)$具有一阶连续偏导数的函数：$\min_{x\in\mathbf{R}^n} f(x)$

一阶泰勒展开：$f(x)=f(x^{(k)})+g_k^{\mathrm{T}}(x-x^{(k)})$

$f(x)$在$x^{(k)}$的梯度值： $g_k=g(x^{(k)})=\nabla f(x^{(k)})$

$$x^{(k+1)} \leftarrow x^{(k)}+\lambda_k p_k$$

负梯度方向：$p_k=-\nabla f(x^{(k)})$ ，由一维搜索确定$\boldsymbol{\lambda}_k$

$$f(x^{(k)}+\lambda_k p_k)=\min_{\lambda\geqslant 0} f(x^{(k)}+\lambda p_k)$$

当 $\|f(x^{(k+1)})-f(x^{(k)})\|<\varepsilon$ 或 $\|x^{(k+1)}-x^{(k)}\|<\varepsilon$ 时，停止迭代，令 $x^*=x^{(k+1)}$。