《数据挖掘技术》

⭐ CH06 逻辑斯谛回归与最大熵模型

➡️ Created by *Wang JingHui*

➡️ Version: *4.0*

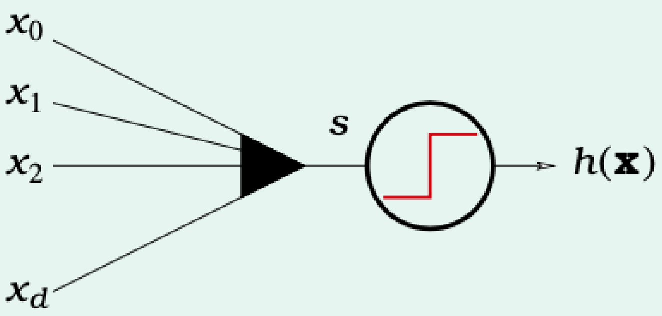
主要内容

1. 逻辑斯谛回归模型
   i. 逻辑斯谛分布
   ii. 二项逻辑斯谛回归模型
   iii. 模型参数估计
   iv. 多项逻辑斯蒂回归模型
2. 最大熵模型
   i. 最大熵原理
   ii. 最大熵模型定义
   iii. 最大熵模型学习
   iv. 极大似然估计
3. 模型学习的最优化算法
   i. 改进的迭代尺度法
   ii. 拟牛顿法

三种线性方式

linear classification

$$h(\mathbf{x}) = \text{sign}(s)$$

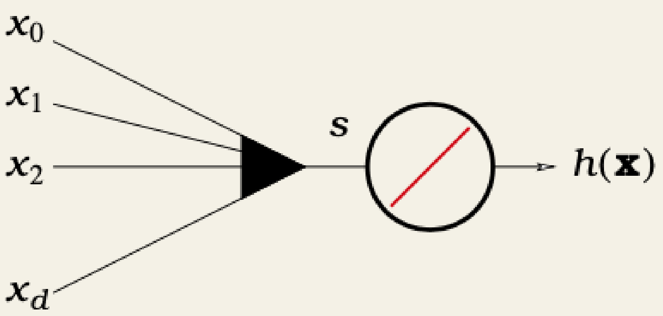


plausible err = 0/1

(small flipping noise)

linear regression

$$h(\mathbf{x}) = s$$

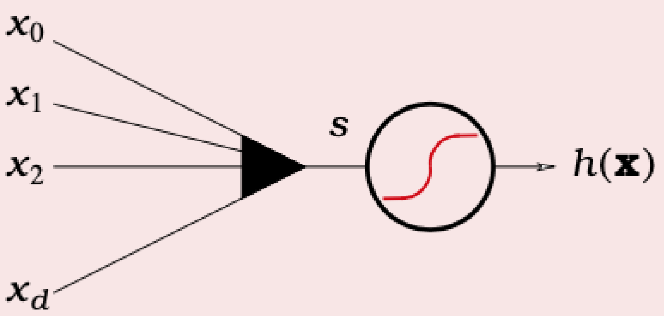


friendly err = squared

(easy to minimize)

logistic regression

$$h(\mathbf{x}) = \theta(s)$$



err = ?

逻辑斯谛回归模型

逻辑斯谛分布

注意：分布函数中关于μ位置参数，γ形状参数的说明，可以大致的和高斯对应理解。

$$F(x)=P(X \leqslant x)=\frac{1}{1+\exp(-(x-\mu)/\gamma)}$$

关于逻辑斯谛， 更常见的一种表达是Logistic function

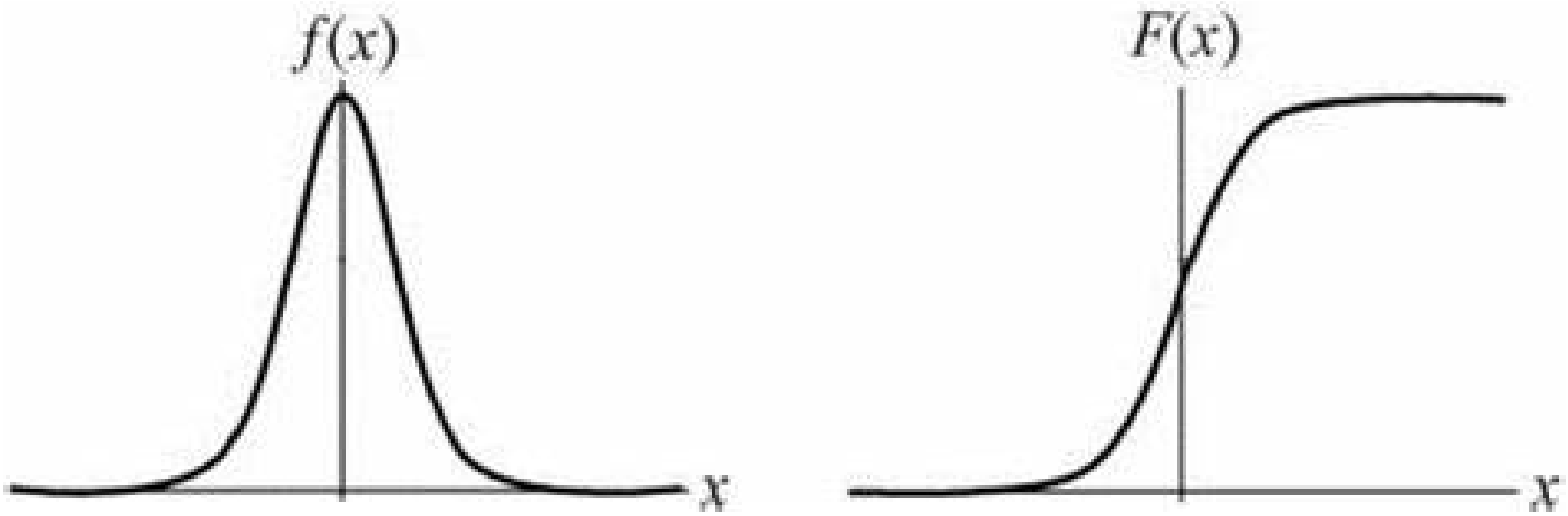
$$\sigma(z)=\frac{1}{1+\exp(-z)}$$

这个函数把实数域映射到(0, 1)区间，这个范围正好是概率的范围， 而且可导, 对于0输入， 得到的是0.5，可以用来表示等可能性。

逻辑斯蒂的密度函数：

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

逻辑斯蒂分布的密度函数与分布函数：



μ为位置参数，$\gamma > 0$为形状参数。

推理

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

$$y = \frac{1}{1+\exp(-(w^T x + b))}$$

得到:

$$\ln \frac{y}{1-y} = w^T x + b \Rightarrow \ln \frac{P(Y=1|x)}{1-P(Y=1|x)} = w^T x + b$$

注: 选择逻辑回归模型，事件$\{Y=1|X\}$发生的对数几率是输入X的线性函数。

比较:

$$y = w^T x + b$$

$$y = sign(w^T x + b)$$

二项逻辑斯谛回归模型

二项逻辑斯谛回归模型是如下的条件概率分布：

$$
\begin{aligned}
P(Y=1|x) &= \frac{\exp(w\cdot x)}{1+\exp(w\cdot x)} \\
&= \frac{\exp(w\cdot x)/\exp(w\cdot x)}{(1+\exp(w\cdot x))/(\exp(w\cdot x))} \\
&= \frac{1}{e^{-(w\cdot x)}+1}
\end{aligned}
$$

$$
\begin{aligned}
P(Y=0|x) &= \frac{1}{1+\exp(w\cdot x)} \\
&= 1-\frac{1}{1+e^{-(w\cdot x)}} \\
&= \frac{e^{-(w\cdot x)}}{1+e^{-(w\cdot x)}}
\end{aligned}
$$

模型参数估计

通过监督学习的方法来估计模型参数，设:

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

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

参数估计这里，似然函数书中的表达

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

这里利用了$y_i \in \{0,1\}$这个特点

对数似然函数

使用对数似然会更简单， 连乘的形式会转换成求和的形式，对数函数为单调递增函数，最大化对数似然等价于最大化似然函数。

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

多项逻辑斯谛回归

假设离散型随机变量Y的取值集合是$1, 2, \ldots, K$, 多项逻辑斯谛回归模型是

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

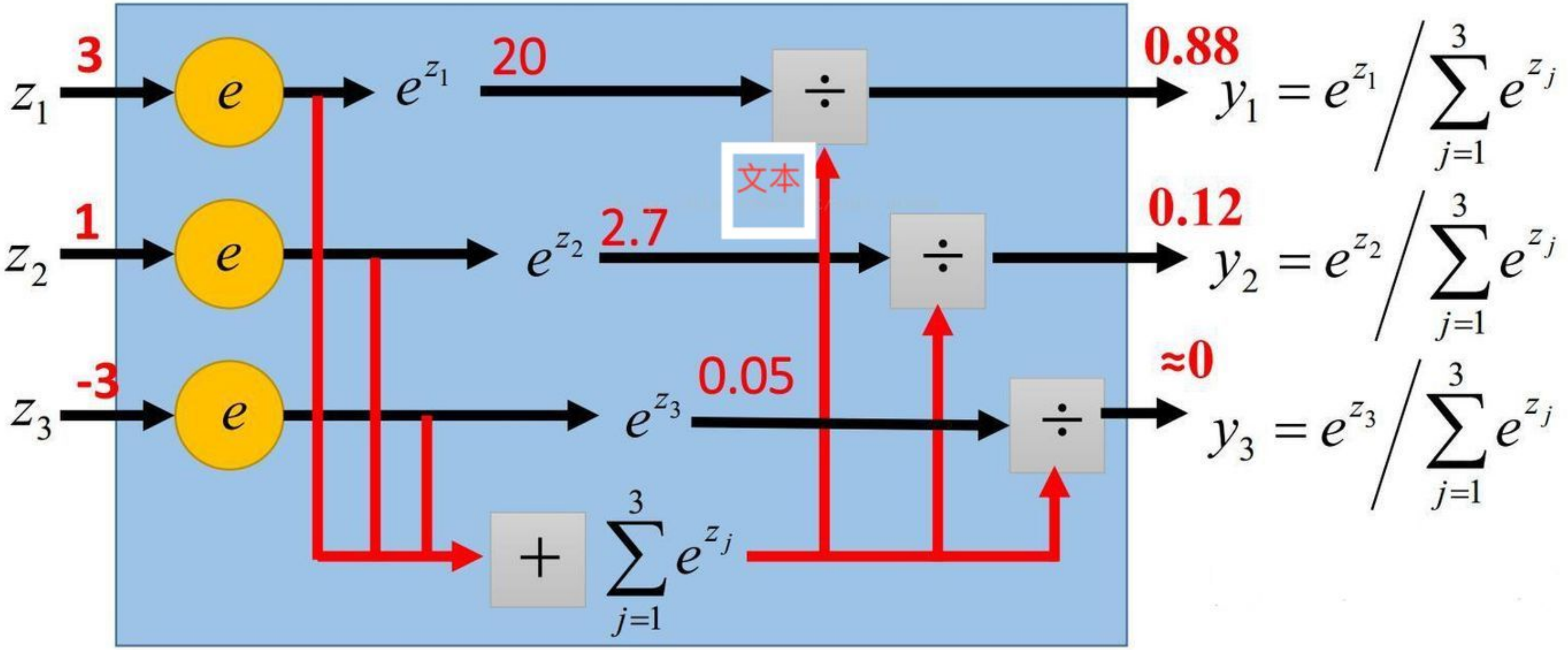
$$P(Y=K|x) = \frac{1}{1+\sum_{k=1}^{K-1}\exp(w_k \cdot x)}$$

- Softmax layer as the output layer

Probability:

- $1 > y_i > 0$
- $\sum_i y_i = 1$

Softmax Layer



熵的相关概念

概念

逻辑斯谛回归模型和最大熵模型，既可以看作是概率模型，又可以看作是非概率模型。

克劳德·艾尔伍德·香农（Claude Elwood Shannon，1916年4月30日—2001年2月24日）

- 美国数学家、信息论的创始人。
- 1948年，划时代的“通信的一个数学理论”分成两部分，在7月和10月的Bell System Technical Journal发表。文章系统论述了信息的定义，怎样数量化信息，怎样更好地对信息进行编码。在这些研究中，概率理论是香农使用的重要工具。香农同时提出了信息熵的概念，用于衡量消息的不确定性。

熵和概率

熵可以从随机变量状态需要的平均信息量角度理解, 也可以从描述统计力学中无序程度的度量角度理解.

> 假设一个发送者想传输一个随机变量x的值给接受者. 在这个过程中, 他们传输的平均信息量可以通过求**信息$h(x)$关于概率分布$p(x)$的期望**得到.
>
> 这个重要的量叫做随机变量x的熵

因为熵的定义把连乘变成了求和, 对数的贡献. 这样可以通过集合的交并来实现熵之间关系的理解.

概率 $\sum_{i=1}^{n} p_i = 1$ $p \in [0,1]$

信息熵是度量样本集合纯度最常用的一种指标.

$$Ent(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

- if $p = 0$, then $p \log_2 p = 0$
- $Ent(D)$越小, D的纯度越高. 非均匀分布比均匀分布熵要小.
- 熵衡量的是不确定性, 概率描述的是确定性

联合熵(相当于并集)

$$
\begin{aligned}
H(X,Y) &= H(X) + H(Y|X) \\
&= H(Y) + H(X|Y) \\
&= H(X|Y) + H(Y|X) + I(X;Y)
\end{aligned}
$$

两个随机变量的联合分布，可以形成**联合熵**，表示为$H(X,Y)$

$$
H(X,Y) = -\sum p(x,y)\log(x,y)
$$

条件熵

条件熵为什么不是这个

$$H(Y|X) = -\sum_{x,y} p(y|x) \log p(y|x)$$

熵是一个多项加权平均值。

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

条件熵$H(Y|X)$

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

$H(X,Y)-H(X)$表示(X,Y)发生所包含的熵，减去X单独发生包含的熵。

条件熵

条件熵是最大熵原理提出的基础

条件熵衡量了条件概率分布的均匀性；

最大熵，就是最大条件熵。

$$\begin{aligned} p^* &= \arg\max_{p \in \mathcal{C}} H(p) \\ &= \arg\max_{p \in \mathcal{C}} \left(-\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x)\right) \end{aligned}$$

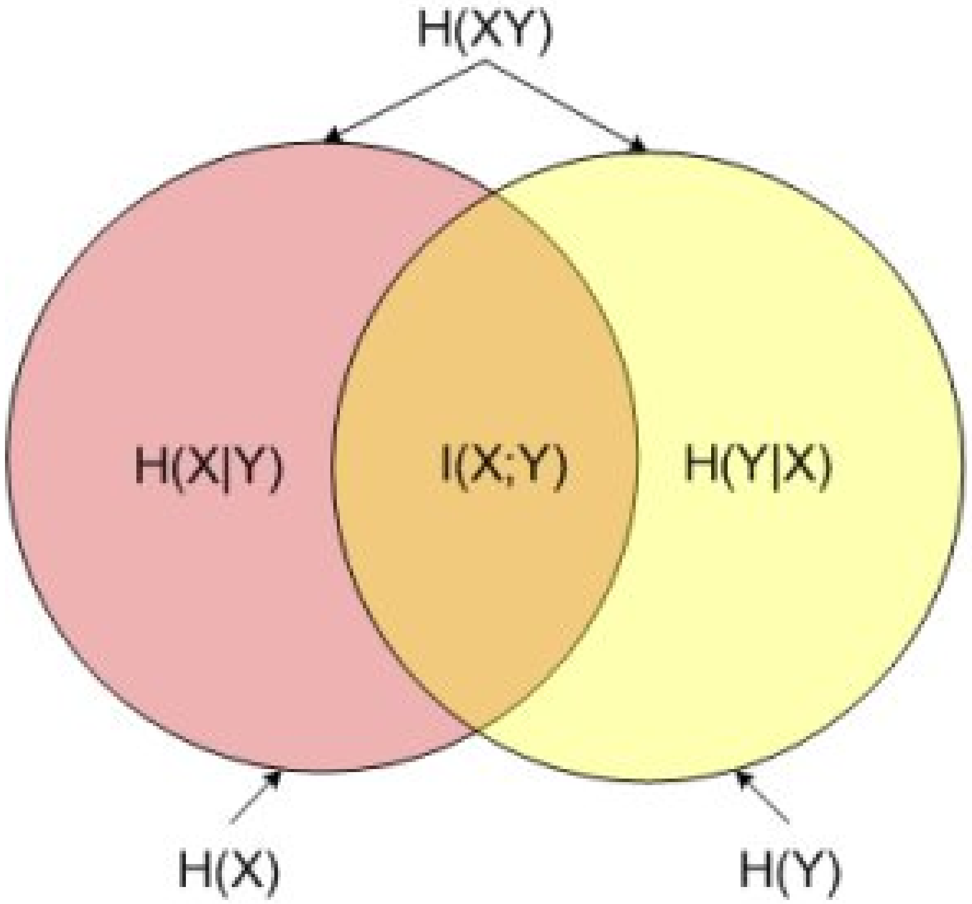
互信息

- 互信息(mutual information), 对应熵里面的交集, 常用来描述差异性
- 一般的, 熵$H(X)$与条件熵$H(Y|X)$之差称为互信息.
- 互信息和条件熵之间的关系

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- 可以把互信息看成由于知道y值而造成的x的不确定性的减小(反之亦然)，也是是信息增益那部分的解释，决策树学习中的信息增益等价于训练数据集中**类**与**特征**的互信息

互信息和条件熵之间的关系



I(X;Y)=H(X)-H(X|Y)

I(X;Y)=H(Y)-H(Y|X)

I(X;Y)=H(X)+H(Y)-H(XY)

交叉熵

刻画两个分布之间的差异

$$
\begin{aligned}
CH(p,q) &= -\sum_{i=1}^{n} p(x_i)\log q(x_i) \\
&= -\sum_{i=1}^{n} p(x_i)\log p(x_i) + \sum_{i=1}^{n} p(x_i)\log p(x_i) - \sum_{i=1}^{n} p(x_i)\log q(x_i) \\
&= H(p) + \sum_{i=1}^{n} p(x_i)\log\frac{p(x_i)}{q(x_i)} \\
&= H(p) + KL(p||q)
\end{aligned}
$$

交叉熵

For two discrete random variables p and q, the cross-entropy is defined as:

$$H(p,q) = -\sum_{x} p(x) \log q(x)$$

This definition is not symmetric. P is intended as the “true” distribution, only partially observed, while Q is intended as the “unnatural” distribution obtained from a constructed statistical model.

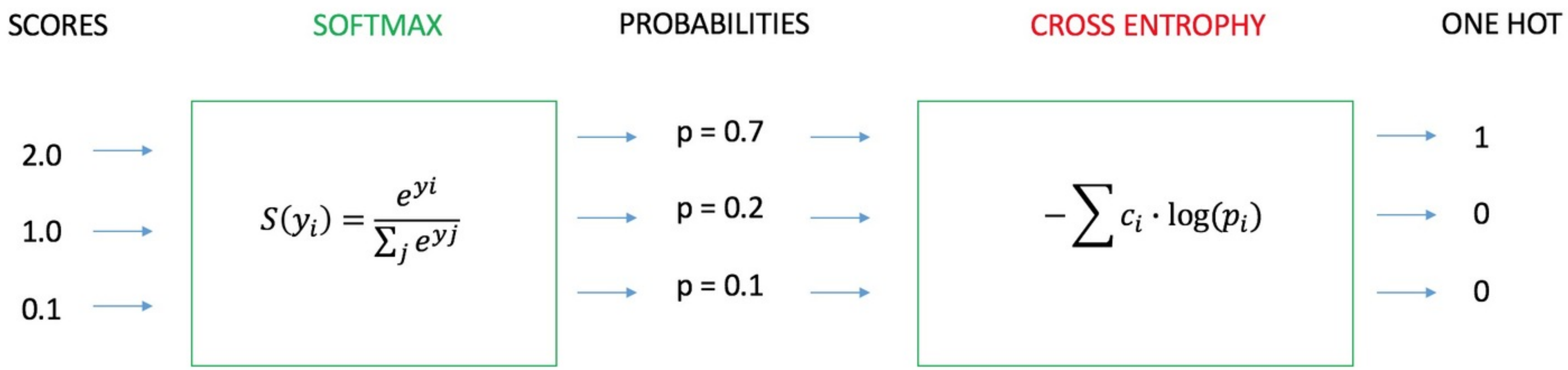
- 引入交叉熵，其用来衡量在给定的真实分布下，使用非真实分布所指定的策略消除系统的不确定性所需要付出的努力的大小。
- 交叉熵刻画的是实际输出（概率）与期望输出（概率）的距离，也就是交叉熵的值越小，两个概率分布就越接近，假设概率分布p为期望输出，概率分布q为实际输出。

softmax 函数和交叉熵-1

softmax 函数，顾名思义是一种 max 函数，max 函数的作用是从一组数据中取最大值作为结果，而 softmax 也起到类似的作用，只是将这组数据进行一些处理，使得计算结果放缩到 (0,1) 区间内。

交叉熵函数与 softmax 函数结合使用，可以得到十分简洁的导数形式，只需将 softmax 的输出结果减 1 再与对应的标签值 相乘即可得到在第 类上的导数，对每个类别分别计算相应的导数，即可得到我们需要的梯度。在许多任务中，标签值往往用 one-hot 形式表示， 一般为 1，那么只需将 softmax 函数的计算结果减 1 即可得到本次传播的第 类的导数值，这使得反向传播中梯度的计算变得十分简单和方便。

softmax 函数和交叉熵-2



最大熵模型

最大熵原理

最大熵原理(Maxent principle)是**概率模型**学习的一个准则.

> Model all that is known and assume nothing about that which is unknown. In other words, given a collection of facts, choose a model which is consistent with all the facts, but otherwise as uniform as possible.
>
> -- Berger, 1996

满足约束条件下使用等概率的方法估计概率分布

最大熵原理很常**见**, 很多原理我们都一直在用, 只是没有上升到理论的高度.

- 等概率表示了对事实的无知, 因为没有更多的信息, 这种判断是合理的.
- 最大熵原理认为要选择的概率模型首先必须满足**已有的事实**, 即**约束条件**
- 最大熵原理根据已有的信息（**约束条件**）, 选择适当的概率模型.
- 最大熵原理认为不确定的部分都是等可能的, 通过熵的最大化来表示**等可能性**.
- 最大熵的原则, 承认已有的, 且对未来无偏
- 最大熵原理并不直接关心特征选择, 但是特征选择是非常重要的, 因为约束可能是成千上万的.

最大熵模型的定义

目标： 假设分类模型是一个条件概率分布， $P(Y|X), X \in \mathcal{X} \subseteq \mathbf{R}^n$

已知： 给定一个训练集 $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$

N是训练样本容量, $x \in \mathbf{R}^n$

分析： 联合分布$P(X,Y)$与边缘分布$P(X)$的经验分布分别为$\widetilde{P}(X,Y)$和$\widetilde{P}(X)$

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

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

上面两个就是不同的数据样本, 在训练数据集中的比例.

如果增加n个**特征函数**, 就可以增加n个**约束条件**

特征函数用来描述$f(x,y)$描述输入x和输出y之间的某一事实，是一个二值函数。

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

例如：

(x_1, y_1) = (一打火柴,量词)

(x_2, y_2) = (打电话,动词)

$$f_1(x,y)=\begin{cases}1 & \text{若”打”前面为数字}\\ 0 & \text{否则}\end{cases}$$

$$f_2(x,y)=\begin{cases}1 & \text{若”打”前面为名词}\\ 0 & \text{否则}\end{cases}$$

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

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

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

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

如果模型能够获取训练数据中的信息, 那么就可以假设这两个期望值相等, 即

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

或

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

上面这个也是模型学习的约束条件，假设有n个特征函数$f_i(x,y), i = 1, 2, ..., n$，那么就有$n$个约束条件。

最大熵模型定义

假设满足所有约束条件的模型集合为

$$\mathcal{C} \equiv \{P \in \mathcal{P} | E_P(f_i) = E_{\widetilde{P}}(f_i), i = 1, 2, \ldots, n\}$$

定义在条件概率分布$P(Y|X)$上的条件熵为

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

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, 上式中对数为自然对数.

最大熵模型的学习

最大熵模型的学习过程就是求解最大熵模型的过程。

最大熵模型的学习可以形式化为约束最优化问题。

$$
\begin{aligned}
\max_{P\in\mathcal{C}} H(P) &= -\sum_{x,y} \widetilde{P}(x)P(y|x)\log P(y|x) \\
s.t. \quad E_P(f_i) &= E_{\widetilde{P}}(f_i),\ i=1,2,\ldots,n \\
\sum_{y} P(y|x) &= 1
\end{aligned}
$$

按照最优化问题的习惯，将求最大值问题改写为等价的求最小值问题。

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

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

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

最优化问题的求解：

引入拉格朗日乘子$w_i, i = 0, 1, \cdots, n$，定义拉格朗日函数$L(P, w)$

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

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

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

求$-H(P)$满足约束条件的极值

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

对偶问题

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

由于拉格朗日函数$L(P, w)$是P的凸函数，所以原始问题的解与对偶问题的结是等价的。

(1).求解对偶问题内部的极小化问题，将其记作：

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

将其解记作

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

(2).求解对偶问题外部的极大化问题

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

将其解记为w^*，即

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

总结过程

$$\max_{P\in\mathcal{C}} H(P) \Rightarrow \min_{P\in\mathcal{C}} -H(P) \Rightarrow \min_{P\in\mathcal{C}} \max_{w} L(P,w) \Rightarrow \max_{w} \min_{P\in\mathcal{C}} L(P,w) \Rightarrow \max_{w} \Psi(w) \Rightarrow w^*$$

最大熵模型的解

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

其中

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

Z_w称为规范化因子；$f_i\left(x,y\right)$是特征函数；w_i是特征的权值。

最大熵模型的极大似然估计

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。

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

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

预测分类原理

这里面重复一下书中的过程, 在$L(P, w)$对P求导并令其为零的情况下解方程能拿到下面公式

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

书中有提到因为$\sum_y P(y|x) = 1$, 然后得到模型

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

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

注意这里面Z_w是归一化因子.

这里面并不是因为概率为1推导出了Z_w的表达式, 而是因为Z_w的位置在分母, 然后对应位置$\exp(1 - w_0)$也在分母, 凑出来这样一个表达式, 意思就是遍历y的所有取值, 求分子表达式的占比.

综上, 如果$f_i(x, y)$只检测是不是存在这种组合, 那么概率就是*归一化的*出现过的特征, 系数求和再取e指数.

模型学习的最优化算法

改进的迭代尺度算法（IIS）

输入：特征函数$f_i, i = 1, 2, \cdots, n$，经验分布$\tilde{P}(x, y)$，模型$P_w(y|x)$

输出：最优参数值w_i^*；最优模型P_{w^*}

(1). 对所有$i \in \{1, 2, \cdots, n\}$，取$w_i = 0$;

(2). 对每一$i \in \{1, 2, \cdots, n\}$

- 令δ_i是方程

$$\sum_{x,y} \tilde{P}(x,y) f_i(x,y) = \sum_{x,y} \tilde{P}(x) P_w(y|x) f_i(x,y) \exp\left(\delta_i f^{\#}(x,y)\right)$$

的解

- 更新w_i的值

$$w_i \leftarrow w_i + \delta_i$$

(3). 如果不是所有w_i都收敛，重复步骤2.

参考

1. Berger,1995, A Brief Maxent Tutorial
2. [数学之美:信息的度量和作用]
3. [数学之美:不要把鸡蛋放在一个篮子里 谈谈最大熵模型]
4. 李航·统计学习方法笔记·第6章 logistic regression与最大熵模型（2）·最大熵模型
5. 最大熵模型与GIS ,IIS算法
6. 关于最大熵模型的严重困惑：为什么没有解析解?
7. 最大熵模型介绍 这个是Berger的文章的翻译.
8. 理论简介 代码实现
9. 另外一份代码

1. 如何理解最大熵模型里面的特征?
2. Iterative Scaling and Coordinate Descent Methods for Maximum Entropy Models
3. [^1]: Generative and discriminative classifiers: Naive Bayes and logistic regression
4. [^2]: On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes
5.
6. [^4]: Multinomial logistic regression

附录：最大熵最优化问题求解

最优化问题的求解：

1. 引入拉格朗日乘子$w_i, i=0,1,\cdots,n$，定义拉格朗日函数$L(P,w)$

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

2. 求$\min_{P\in\mathcal{C}} L\left(P,w\right)$:

记对偶函数$\Psi\left(w\right)=min_{P\in\mathcal{C}}L\left(P,w\right)=L\left(P_w,w\right)$，其解记$P_w=\arg\min_{P\in\mathcal{C}} L\left(P,w\right)=P_w\left(y|x\right)$

$$
\begin{aligned}
\frac{\partial L\left(P,w\right)}{\partial P\left(y|x\right)} &= \sum_{x,y}\tilde{P}\left(x\right)\left(\log P\left(y|x\right)+1\right)-\sum_{y}w_0-\sum_{x,y}\left(\tilde{P}\left(x\right)\sum_{i=1}^{n}w_i f_i\left(x,y\right)\right) \\
&= \sum_{x,y}\tilde{P}\left(x\right)\left(\log P\left(y|x\right)+1\right)-\sum_{x,y}P\left(x\right)w_0-\sum_{x,y}\left(\tilde{P}\left(x\right)\sum_{i=1}^{n}w_i f_i\left(x,y\right)\right) \\
&= \sum_{x,y}\tilde{P}\left(x\right)\left(\log P\left(y|x\right)+1-w_0-\sum_{i=1}^{n}w_i f_i\left(x,y\right)\right)=0
\end{aligned}
$$

由于$\tilde{P}\left(x\right) > 0$，得

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

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

由于

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

则

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

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

代入，得

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

其中

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

Z_w称为规范化因子；$f_i\left(x,y\right)$是特征函数；w_i是特征的权值。
3. 求$\max_w\Psi\left(w\right)$
将其解记为w^*，即

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

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

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

对偶函数

$$
\begin{aligned}
\Psi\left(w\right) &= min_{P\in\mathcal{C}} L\left(P,w\right) = L\left(P_w,w\right) \\
&= -H\left(P_w\right) + w_0\left(1-\sum_y P_w\left(y|x\right)\right) + \sum_{i=1}^{n} w_i\left(E_{\tilde{P}}\left(f_i\right) - E_{P_w}\left(f_i\right)\right) \\
&= \sum_{x,y}\tilde{P}\left(x\right)P_w\left(y|x\right)\log P_w\left(y|x\right) \\
&\qquad + w_0\left(1-\sum_y \frac{1}{Z_w\left(x\right)}\exp\left(\sum_{i=1}^{n} w_i f_i\left(x,y\right)\right)\right) \\
&\qquad + \sum_{i=1}^{n} w_i\left(\sum_{x,y}\tilde{P}\left(x,y\right)f_i\left(x,y\right) - \sum_{x,y}\tilde{P}\left(x\right)P_w\left(y|x\right)f_i\left(x,y\right)\right) \\
&= \sum_{x,y}\tilde{P}\left(x,y\right)\sum_{i=1}^{n} w_i f_i\left(x,y\right) + \sum_{x,y}\tilde{P}\left(x\right)P_w\left(y|x\right)\left(\log P_w\left(y|x\right) - \sum_{i=1}^{n} w_i f_i\left(x,y\right)\right) \\
&= \sum_{x,y}\tilde{P}\left(x,y\right)\sum_{i=1}^{n} w_i f_i\left(x,y\right) - \sum_{x,y}\tilde{P}\left(x,y\right)\log Z_w\left(x\right) \\
&= \sum_{x,y}\tilde{P}\left(x,y\right)\sum_{i=1}^{n} w_i f_i\left(x,y\right) - \sum_{x}\tilde{P}\left(x\right)\log Z_w\left(x\right)
\end{aligned}
$$

得

$$L_{\tilde{P}}\left(P_{w}\right)=\Psi\left(w\right)$$

即，最大熵模型的极大似然估计等价于对偶函数极大化。

↘️ Enjoy your machine learning!

https://github.com/wjssx/

E-mail: csr_dsp@sina.com