1.什么是熵

1.1 术语定义

2. 最大熵模型

2.1 无偏原则

2.2 最大熵模型的表示

2.3 对偶问题极大化的指数解

2.4 最大熵模型的极大似然估计

3. 最优化算法

3.1 改进的迭代尺度法（IIS)

为了更好的理解本文，需要了解的概率必备知识有：

1. 大写字母X表示随机变量，小写字母x表示随机变量X的某个具体的取值；

2. P(X)表示随机变量X的概率分布，P(X,Y)表示随机变量X、Y的联合概率分布，P(Y|X)表示已知随机变量X的情况下随机变量Y的条件概率分布；

3. p(X = x)表示随机变量X取某个具体值的概率，简记为p(x)；

4. p(X = x, Y = y) 表示联合概率，简记为p(x,y)，p(Y = y|X = x)表示条件概率，简记为p(y|x)，且有：p(x,y) = p(x) * p(y|x)。

需要了解的有关函数求导、求极值的知识点有：

5. 如果函数y=f(x)在[a, b]上连续，且其在(a,b)上可导，如果其导数f′ (x) >0，则代表函数f(x)在[a,b]上单调递增，否则单调递减；如果函数的二阶导f''(x) > 0，则函数在[a,b]上是凹的，反之，如果二阶导f''(x) < 0，则函数在[a,b]上是凸的。

6. 设函数f(x)在x0处可导，且在x处取得极值，则函数的导数F′ (x0) = 0。

7. 以二元函数z = f(x,y)为例，固定其中的y，把x看做唯一的自变量，此时，函数对x的导数称为二元函数z=f(x,y)对x的偏导数。

8. 为了把原带约束的极值问题转换为无约束的极值问题，一般引入拉格朗日乘子，建立拉格朗日函数，然后对拉格朗日函数求导，令求导结果等于0，得到极值。

1.什么是熵

熵用来表示随机变量的不确定性。一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。所以说，信息熵可以被认为是系统有序化程度的一个度量。

1.1 术语定义

熵：如果一个随机变量X的可能取值为X = {x1, x2,…, xk}，其概率分布为P(X = xi) = pi（i = 1,2, ..., n），则随机变量X的熵定义为：

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

把最前面的负号放到最后，便成了：

$$H(X)=\sum_{x} p(x)\log \frac{1}{p(x)}$$

联合熵：两个随机变量X,Y的联合分布，可以形成联合熵Joint Entropy,用H(X,Y)表示。

条件熵：在随机变量X发生的前提下，随机变量Y发生所新带来的熵定义为Y的条件熵，用H(Y|X)表示，用来衡量在已知随机变量X的条件下随机变量Y的不确定性。

且有此式子成立：H(Y|X) = H(X,Y) - H(X)，整个式子表示(X,Y)发生所包含的熵减去X单独发生包含的熵。推导如下：

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

简单解释下上面的推导过程。整个式子共6行，其中

- 第二行推到第三行的依据是边缘分布p(x)等于联合分布p(x,y)的和；
- 第三行推到第四行的依据是把公因子logp(x)乘进去，然后把x,y写在一起；
- 第四行推到第五行的依据是：因为两个sigma都有p(x,y)，故提取公因子p(x,y)放到外边，然后把里边的-（log p(x,y) - log p(x)）写成- log (p(x,y)/p(x))；
- 第五行推到第六行的依据是：p(x,y) = p(x) * p(y|x)，故p(x,y) / p(x) = p(y|x)。

相对熵：又称互熵，交叉熵，鉴别信息，Kullback熵等。设p(x)、q(x)是X中取值的两个概率分布，则p对q的相对熵是：

$$
D(p\,\|\,q)=\sum_{x}p(x)\log\frac{p(x)}{q(x)}=E_{p(x)}\log\frac{p(x)}{q(x)}
$$

在一定程度上，相对熵可以度量两个随机变量的“距离”，且有D(p||q) ≠D(q||p)。另外，值得一提的是，D(p||q)是必然大于等于0的。

互信息：两个随机变量X,Y的互信息定义为X,Y的联合分布和各自独立分布乘机的相对熵，用I(X,Y)表示

$$I(X,Y)=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

且有I(X,Y)=D(P(X,Y) || P(X)P(Y))。下面，咱们来计算下H(Y)-I(X,Y)的结果，如下：

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

通过上面的计算过程，我们发现竟然有H(Y)-I(X,Y) = H(Y|X)。故通过条件熵的定义，有：H(Y|X) = H(X,Y) - H(X)，而根据互信息定义展开得到H(Y|X) = H(Y) - I(X,Y)，把前者跟后者结合起来，便有I(X,Y)= H(X) + H(Y) - H(X,Y)，此结论被多数文献作为互信息的定义。

2. 最大熵模型

2.1 无偏原则

最大熵原理认为，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。最大熵模型的本质，已知X，计算Y的概率，且尽可能让Y的概率最大，即最大化H(Y|X)：

$$\max H(Y \mid X) = \sum_{\substack{x \in \{x_1, x_2\} \\ y \in \{y_1, y_2, y_3, y_4\}}} p(x, y) \log \frac{1}{p(y \mid x)}$$

且满足以下4个约束条件：

$$p(x_1) + p(x_2) = 1$$

$$p(y_1) + p(y_2) + p(y_3) + p(y_4) = 1$$

$$p(y_4) = 0.05$$

$$p(y_2 \mid x_1) = 0.95$$

2.2 最大熵模型的表示

一般表达式：

$$\max_{p \in P} H(Y \mid X) = \sum_{(x,y)} p(x, y) \log \frac{1}{p(y \mid x)}$$

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

给定一个训练数据集

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

学习的目标是用最大熵原理选择最好的分类模型．

首先考虑模型应该满足的条件．给定训练数据集，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，分别以 $\tilde{P}(X,Y)$ 和 $\tilde{P}(X)$ 表示．这里，

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

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

其中，$\nu(X=x,Y=y)$ 表示训练数据中样本 (x,y) 出现的频数，$\nu(X=x)$ 表示训练数据中输入 x 出现的频数，N 表示训练样本容量．

用特征函数（feature function）$f(x,y)$ 描述输入 x 和输出 y 之间的某一个事实．其定义是

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

它是一个二值函数[8]，当 x 和 y 满足这个事实时取值为 1，否则取值为 0．

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

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

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，用 $E_P(f)$ 表示．

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

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

$$E_P(f)=E_{\tilde{P}}(f) \tag{6.10}$$

或

$$\sum_{x,y}\tilde{P}(x)P(y|x)f(x,y)=\sum_{x,y}\tilde{P}(x,y)f(x,y) \tag{6.11}$$

我们将式 (6.10) 或式 (6.11) 作为模型学习的约束条件．假如有 n 个特征函数 $f_i(x,y)$，$i=1,2,\cdots,n$，那么就有 n 个约束条件．

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

$$\mathcal{C}\equiv\{P\in\mathcal{P}\,|\,E_P(f_i)=E_{\tilde{P}}(f_i),\ \ i=1,2,\cdots,n\} \tag{6.12}$$

定义在条件概率分布 $P(Y|X)$ 上的条件熵为

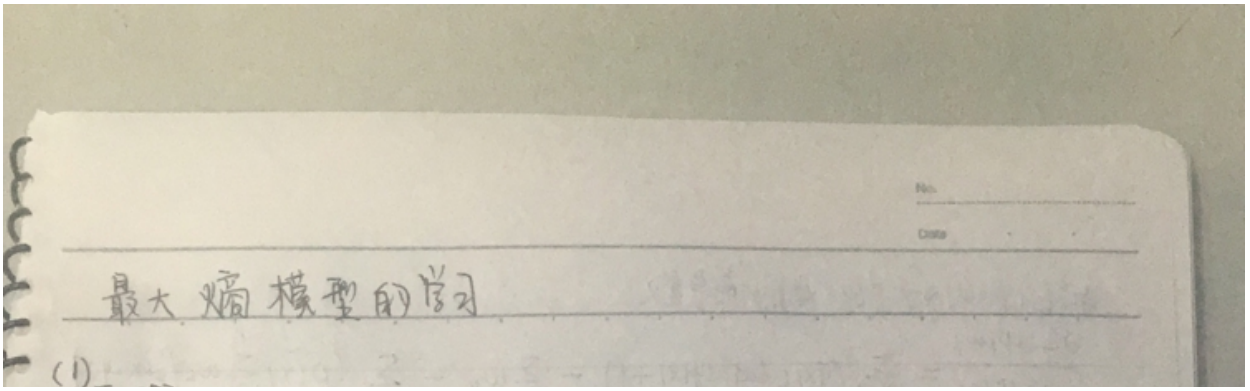
$$H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x) \tag{6.13}$$

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．式中的对数为自然对数．

该问题已知若干条件，要求若干变量的值使到目标函数（熵）最大，其数学本质是最优化问题（Optimization Problem），其约束条件是线性的等式，而目标函数是非线性的，所以该问题属于非线性规划（线性约束）(non-linear programming with linear constraints)问题，故可通过引入Lagrange函数将原带约束的最优化问题转换为无约束的最优化的对偶问题。

$$\Lambda(p,\vec{\lambda})=H(y|x)+\sum_{i=1}^{m}\lambda_i\left(E(f_i)-\tilde{E}(f_i)\right)+\lambda_{m+1}\left(\sum_{y\in Y}p(y|x)-1\right)$$

2.3 对偶问题极大化的指数解



原问题：

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

$$s.t. \quad \sum_{x,y} p(y|x)\,\tilde{p}(x)\, f(x,y) = \sum_{x,y} \tilde{p}(x,y) f(x,y)$$

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

按照最优化问题的习惯，将求最大值问题改写为求解最小值问题：

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

$$s.t. \quad \sum_{x,y} \tilde{p}(x)\, p(y|x) f(x,y) - \sum_{x,y} \tilde{p}(x,y) f(x,y) = 0 \quad \left(E_p(f) - E_{\tilde{p}}(f) = 0\right)$$

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

对偶问题：

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

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

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

将 $\min_{p\in C} \max_w L(p,w)$ 转化为 $\max_w \min_{p\in C} L(p,w)$

(2) 首先，求 $\min_{p\in C} L(p,w)$

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

记它的解 $p_w = \operatorname{argmin}_{p\in C} L(p,w) = p_w(y|x)$

求 $L(p,w)$ 对 $p(y|x)$ 的偏导数

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

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

$$= 0$$

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

因为 $\sum_y p(y|x) = 1$,

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

所以即 $p_w(y|x) = \dfrac{\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)}{\sum_y \exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)}$ ✓

令分母 $Z_w(x) = \sum_y \exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$ 称为规范化因子 ✓

$f_i(x,y)$ 是特征函数，w_i 是特征权值.

w 是最大熵模型中的参数向量.

(3) 将 $p_w(y|x)$ 代回拉格朗日函数

得：

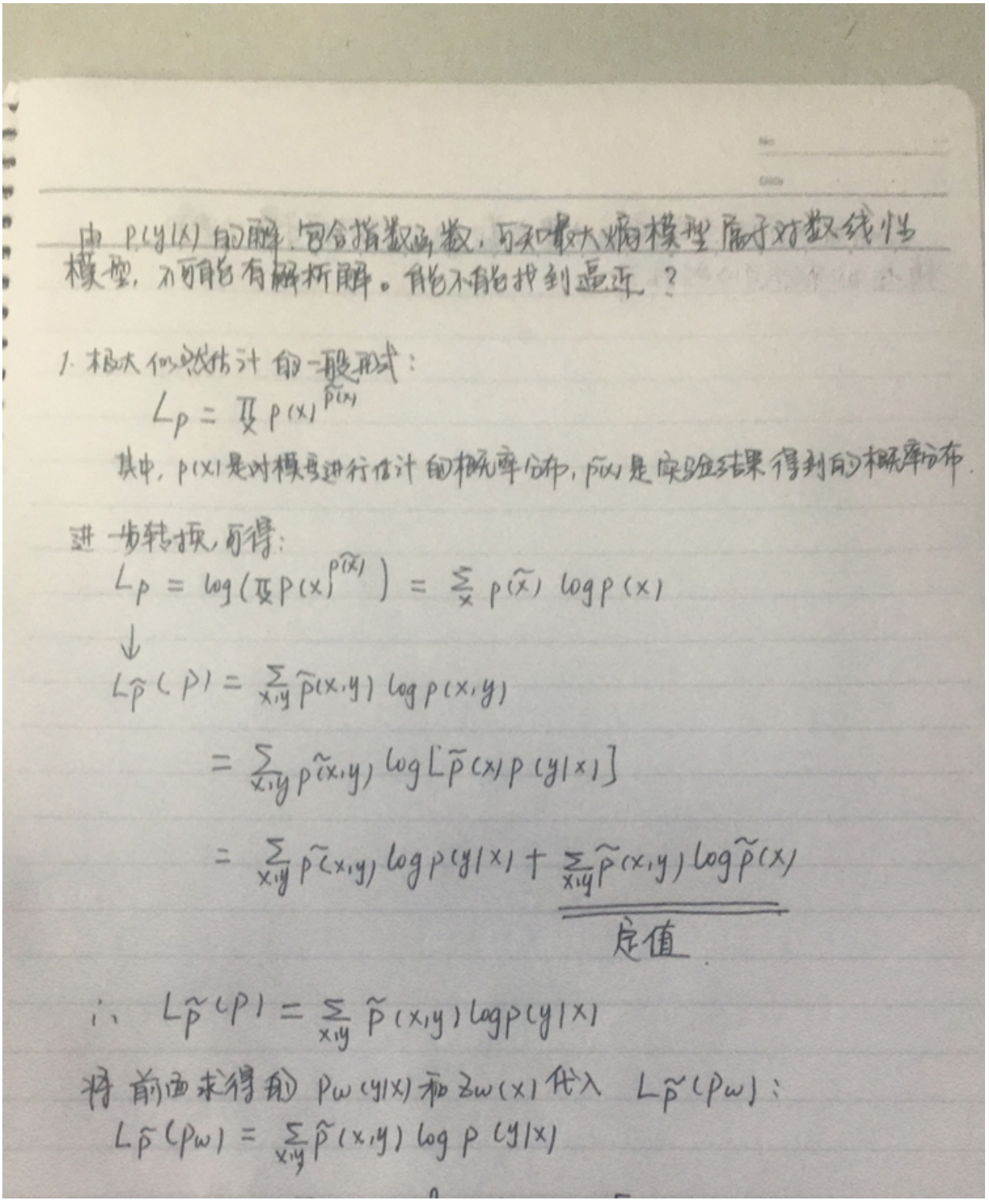
~~$L(p,w) = -\sum_{x,y}\tilde{p}(x)p_w(y|x)\log Z_w(x) + \sum_{i=1}^{n} p(x,y)\sum_{x,y} w_i f_i(x,y)$~~

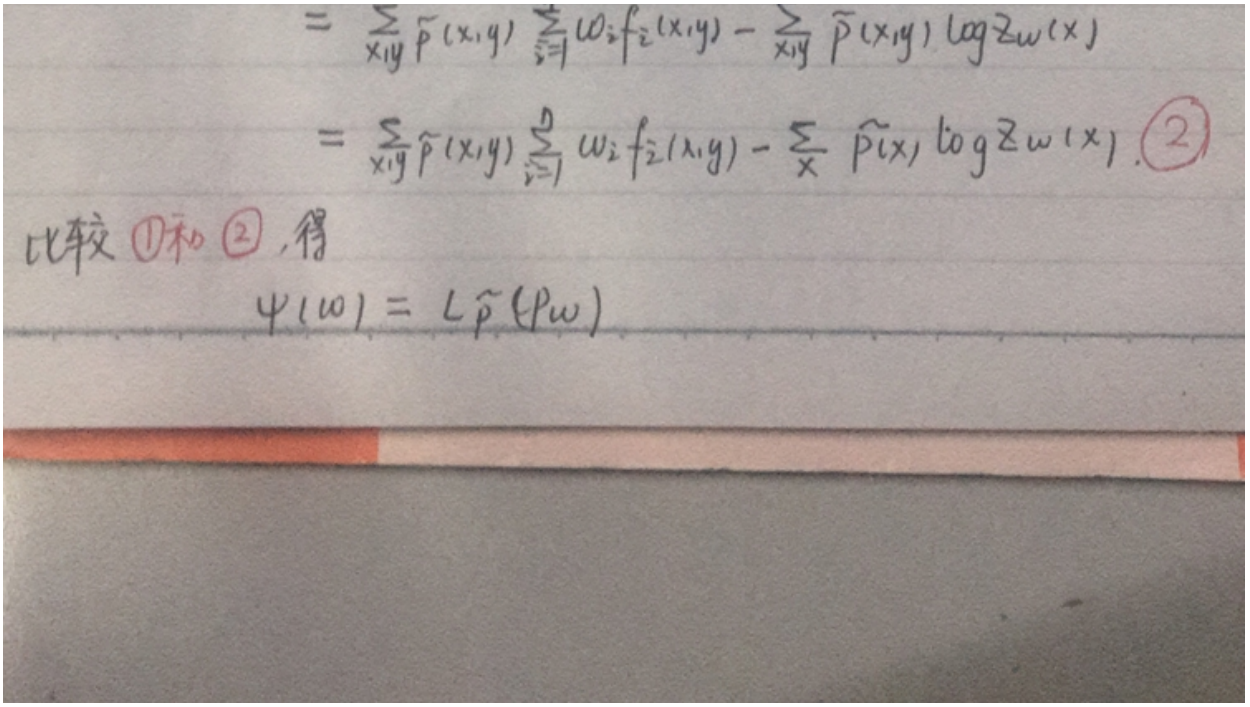
~~最后~~

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

2.4 最大熵模型的极大似然估计

所谓最大似然，即最大可能，在“模型已定，参数 θ 未知”的情况下，通过观测数据估计参数 θ 的一种思想或方法，换言之，解决的是取怎样的参数 θ 使得产生已得观测数据的概率最大的问题。





证明了最大熵模型学习中的对偶函数极大化等价于最大熵模型的极大似然估计。

且根据MLE的正确性，可以断定：最大熵的解（无偏的对待不确定性）同时是最符合样本数据分布的解，进一步证明了最大熵模型的合理性。两相对比，熵是表示不确定性的度量，似然表示的是与知识的吻合程度，进一步，最大熵模型是对不确定度的无偏分配，最大似然估计则是对知识的无偏理解。

最大熵模型与逻辑斯蒂回归模型有类似的形式，它们又称为对数线性模型。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。

3. 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．从最优化的观点看，这时的目标函数具有很好的性质．它是光滑的凸函数，因此多种最优化的方法都适用，保证能找到全局最优解．常用的方法有改进的迭代尺度法、梯度下降法、牛顿法或拟牛顿法．牛顿法或拟牛顿法一般收敛速度更快．

3.1 改进的迭代尺度法（IIS)

现在的问题转换成：通过极大似然函数求解最大熵模型的参数，即求上述对数似然函数参数w的极大值。此时，通常通过迭代算法求解，比如改进的迭代尺度法IIS、梯度下降法、牛顿法或拟牛顿法。这里主要介绍下其中的改进的迭代尺度法IIS。

改进的迭代尺度法IIS的**核心思想**是：假设最大熵模型当前的参数向量是w，希望找到一个新的参数向量w+δ，使得当前模型的对数似然函数值L增加。重复这一过程，直至找到对数似然函数的最大值。

下面，咱们来计算下参数λ 变到λ+δ的过程中，对数似然函数的增加量，用L(λ+δ)-L(λ)表示，同时利用不等式：-lnx ≥1-x ，x>0，可得到对数似然函数增加量的下界，如下：

$$
\begin{aligned}
&L(\lambda+\delta)-L(\lambda)\\
&=\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)-\sum_{x}\overline{p}(x)\log\frac{Z_{\lambda+\delta}(x)}{Z_{\lambda}(x)}\\
&\geq\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\frac{Z_{\lambda+\delta}(x)}{Z_{\lambda}(x)}\\
&=\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\frac{Z_{\lambda+\delta}(x)}{Z_{\lambda}(x)}\\
&=\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\exp\left(\sum_{i=1}^{n}\delta_i f_i(x,y)\right)\\
&=\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\sum_{y}p_{\lambda}(y|x)\exp\left(\sum_{i=1}^{n}\delta_i f_i(x,y)\right)
\end{aligned}
$$

将上述求得的下界结果记为A(δ | λ)，为了进一步降低这个下界，即缩小A(δ | λ)的值，引入一个变量：

$$f^{\#}(x,y)=\sum_{i}f_i(x,y)$$

其中，f 是一个二值函数，故f#(x, y)表示的是所有特征(x, y)出现的次数，然后利用Jason不等式，可得：

$$A(\delta|\lambda)=\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\sum_{y}p_\lambda(y|x)\exp\left(\sum_{i=1}^{n}\delta_i f_i(x,y)\right)$$

$$=\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\sum_{y}p_\lambda(y|x)\exp\left(f^{\#}(x,y)\sum_{i=1}^{n}\frac{\delta_i f_i(x,y)}{f^{\#}(x,y)}\right)$$

$$\geq\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\sum_{y}p_\lambda(y|x)\sum_{i=1}^{n}\frac{f_i(x,y)}{f^{\#}(x,y)}\exp\left(\delta_i f^{\#}(x,y)\right)$$

Jensen不等式：

$$f\left(\sum_{i=1}^{n}\lambda_i x_i\right)\leq\sum_{i=1}^{n}\lambda_i f(x_i)$$

$$f\left(\frac{\sum_{i=1}^{n}p_i x_i}{\sum_{j=1}^{n}p_j}\right)\leq\frac{\sum_{i=1}^{n}p_i f(x_i)}{\sum_{j=1}^{n}p_j}$$

我们把上述式子求得的A(δ | λ)的下界记为B(δ | λ)：

$$B(\delta|\lambda)=\sum_{x,y}\overline{p}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\overline{p}(x)\sum_{y}p_\lambda(y|x)\sum_{i=1}^{n}\frac{f_i(x,y)}{f^{\#}(x,y)}\exp\left(\delta_i f^{\#}(x,y)\right)$$

相当于B(δ | λ)是对数似然函数增加量的一个新的下界，可记作：L(λ+δ)-L(λ) >= B(δ | λ)。

接下来，对B(δ | λ)求偏导，得：

$$\frac{\partial B(\delta|\lambda)}{\partial\delta_i}=\sum_{x,y}\overline{p}(x,y)f_i(x,y)-\sum_{x}\overline{p}(x)\sum_{y}p_\lambda(y|x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right)$$

$$=\sum_{x,y}\overline{p}(x,y)f_i(x,y)-\sum_{x,y}\overline{p}(x)p_\lambda(y|x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right)$$

$$=E_{\overline{p}}(f_i)-\sum_{x,y}\overline{p}(x)p_\lambda(y|x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right)$$

此时得到的偏导结果只含δ，除δ之外不再含其它变量，令其为0，可得：

$$\sum_{x,y}\overline{p}(x)p_\lambda(y|x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right)=E_{\overline{p}}(f_i)$$

从而求得δ，问题得解。

值得一提的是，在求解δ的过程中，如果若f#(x,y)=M为常数，则

$$\delta_i=\frac{1}{M}\log\frac{E_{\overline{p}}(f_i)}{E_p(f_i)}$$

否则，用牛顿法解决：

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

求得了δ，便相当于求得权值λ，最终将λ　回代到下式中：

$$p^*(y|x)=\frac{1}{Z_\lambda(x)}e^{\sum_i\lambda_i f_i(x,y)}$$

即得到最大熵模型的最优估计。