Author：Liu Jian

Time：2020-06-06

机器学习10-最大熵原理与最大熵模型

1 最大熵原理

1.1 问题的构建

最大熵原理：在满足约束条件的模型集合中选取熵最大的模型 (可以这样理解，熵越大，编码也就越长，模型的表示能力也就越强)。下面我们从最大熵原理出发，学习某个概率分布 $P(y|\boldsymbol{x})$ (即 $P_{model}(y|\boldsymbol{x})$ ，只不过这一模型的形式我们没有指定，而是根据最大熵原理推得)， $\boldsymbol{x}\in\mathbb{X}\subseteq\mathbb{R}^d, y\in\mathbb{Y}\subseteq\mathbb{R}$ 。样本数据 $D=\{(\boldsymbol{x}_i,y_i)\}_1^N$ ，通过统计频率我们可以得到经验分布 $\tilde{P}(\boldsymbol{x},y)$ (即 $P_{data}(\boldsymbol{x},y)$ ，当然也可以得到 $P_{data}(\boldsymbol{x})$ 、 $P_{data}(y|\boldsymbol{x})$ 等)。

首先，我们给出学习的约束条件。给定特征函数 $f(\boldsymbol{x},y)$ ，基于经验分布我们可以计算特征函数的期望：

$$\mathbb{E}_{\tilde{P}}[f(\boldsymbol{x},y)]=\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)f(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y=\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})\tilde{P}(y|\boldsymbol{x})f(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

我们将上式中的 $\tilde{P}(y|\boldsymbol{x})$ 替换为我们所要求的分布 $P(y|\boldsymbol{x})$ ：

$$\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})f(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\triangleq\mathbb{E}_P[f(\boldsymbol{x},y)]$$

自然地，我们希望替换前后二式相等，于是我们令 $\mathbb{E}_{\tilde{P}}[f(\boldsymbol{x},y)]=\mathbb{E}_P[f(\boldsymbol{x},y)]$ ，这也就得到了最大熵模型的约束条件。

接下来，我们给出优化的目标函数。我们可以选择最大化如下的条件熵：

$$H(y|\boldsymbol{x})=-\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})\ln P(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

综上，我们要求解如下的约束优化问题：

$$\begin{gathered}\max_P H(y|\boldsymbol{x})\Leftrightarrow\min_P\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})\ln P(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y\\ \mathrm{s.t.}\ \ \mathbb{E}_{\tilde{P}}[f_i(\boldsymbol{x},y)]=\mathbb{E}_P[f_i(\boldsymbol{x},y)],\ \ i=1,\cdots,n\\ \int_{\mathbb{Y}}P(y|\boldsymbol{x})\mathrm{d}y=1\end{gathered}$$

其中，$\tilde{P}$ 由训练样本 D 给出，$f_i,\ i=1,\cdots,n$ 为给定的特征函数。

最后，我们指出：

1. 若我们要求解的对象是 $P(\boldsymbol{x},y)$，在构建约束时，我们令 $\mathbb{E}_{\tilde{P}}[f(\boldsymbol{x},y)]=\mathbb{E}_P[f(\boldsymbol{x},y)]$，可得：

$$\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)f(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y=\int_{\mathbb{X}\times\mathbb{Y}}P(\boldsymbol{x},y)f(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

但这里我们要求解的对象是 $P(y|\boldsymbol{x})$，我们令：

$$\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)f(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y=\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})f(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

即从等号左边到右边，我们只替换其中的 $\tilde{P}(y|\boldsymbol{x})$ 为 $P(y|\boldsymbol{x})$，即对于等式右边，除了我们要求解的对象--条件概率 $y|\boldsymbol{x}$ 取 $P(y|\boldsymbol{x})$ (即 $P_{model}(y|\boldsymbol{x})$) 外，其他概率均取经验分布。

2. 类似地，这里我们选择最大化条件概率 $P(y|\boldsymbol{x})$ 的交叉熵：

$$-\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})\ln P(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

其中，除了我们要求解的对象--条件概率 $y|\boldsymbol{x}$ 取 $P(y|\boldsymbol{x})$ (即 $P_{model}(y|\boldsymbol{x})$) 外，其他概率均取经验分布。若我们要求解的对象是 $P(\boldsymbol{x},y)$，则我们最大化信息熵：

$$-\int_{\mathbb{X}\times\mathbb{Y}}P(\boldsymbol{x},y)\ln P(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

1.2 对偶问题

原问题：

$$\begin{aligned}&\min_P\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})\ln P(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y\\&\text{s.t. }\mathbb{E}_{\tilde{P}}[f_i(\boldsymbol{x},y)]=\mathbb{E}_P[f_i(\boldsymbol{x},y)],\ \ i=1,\cdots,n\\&\int_{\mathbb{Y}}P(y|\boldsymbol{x})\mathrm{d}y=1\end{aligned}$$

上述是一个泛函问题。

构造拉格朗日函数如下：

$$\begin{aligned}L(P,\mu,\boldsymbol{\lambda})=&\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})\ln P(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y+\mu(\boldsymbol{x})\left(1-\int_{\mathbb{Y}}P(y|\boldsymbol{x})\mathrm{d}y\right)\\&+\sum_{i=1}^{n}\lambda_i\int_{\mathbb{X}\times\mathbb{Y}}\Big(\tilde{P}(\boldsymbol{x},y)-\tilde{P}(\boldsymbol{x})P(y|\boldsymbol{x})\Big)f_i(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\end{aligned}$$

其中，$\mu(\boldsymbol{x}),\boldsymbol{\lambda}$ 为拉格朗日乘子，约束条件 $\int_{\mathbb{Y}}P(y|\boldsymbol{x})\mathrm{d}y=1$ 是一个关于自变量 $\boldsymbol{x}$ 的函数约束，因此其拉格朗日乘子是一个函数 $\mu(\boldsymbol{x})$ 。可以这样理解，约束条件 $\int_{\mathbb{Y}}P(y|\boldsymbol{x})\mathrm{d}y=1$ 在空间 $\mathbb{X}$ 中的每个点都对应了一个约束，也就对应了一个拉格朗日乘子，这些拉格朗日乘子可统一地用 $\mu(\boldsymbol{x})$ 表示。拉格朗日函数 $L(P,\lambda)$ 是关于 P 的凸函数，因此主问题 $\min_P\max_{\mu,\boldsymbol{\lambda}}L(P,\mu,\boldsymbol{\lambda})$ 与对偶问题 $\max_{\mu,\boldsymbol{\lambda}}\min_P L(P,\mu,\boldsymbol{\lambda})$ 等价，我们通过求解对偶问题来得到原问题的解。

首先是 $\min_P L(P,\mu,\boldsymbol{\lambda})$，我们可以视 $P(\cdot|\star)$ 在输入输出空间 $\mathbb{X}\times\mathbb{Y}$ 每个点上的值为一个参数来进行优化，求导并令导数为 0 (KKT 条件之 stationarity)：

$$\frac{\partial L(P,\mu,\boldsymbol{\lambda})}{\partial P(y|\boldsymbol{x})}=0$$

$$\Updownarrow$$

$$\tilde{P}(\boldsymbol{x})(\ln P(y|\boldsymbol{x})+1)-\mu(\boldsymbol{x})-\tilde{P}(\boldsymbol{x})\sum_{i=1}^{n}\lambda_i f_i(\boldsymbol{x},y)=0$$

$$P(y|\boldsymbol{x})=\frac{\exp\left(\sum_{i=1}^{n}\lambda_i f_i(\boldsymbol{x},y)\right)}{\exp\left(1-\mu(\boldsymbol{x})/\tilde{P}(\boldsymbol{x})\right)}\triangleq P_{\mu,\lambda}(y|\boldsymbol{x})$$

其中， $f_i(\boldsymbol{x},y),\tilde{P}(\boldsymbol{x})$ 是已知的量。因此，对偶问题：

$$\max_{\mu,\boldsymbol{\lambda}}\min_{P}L(P,\mu,\boldsymbol{\lambda})=\max_{\mu,\boldsymbol{\lambda}}L(P_{\mu,\lambda},\mu,\boldsymbol{\lambda})$$

事实上，我们可以作进一步化简，消去变量 $\mu(\boldsymbol{x})$ ，只保留变量 $\boldsymbol{\lambda}$ 。因为由 $\int_{\mathbb{Y}}P(y|\boldsymbol{x})\mathrm{d}y=1$ (KKT条件之primal feasibility)， $\mu(\boldsymbol{x}),\boldsymbol{\lambda}$ 之间存在如下的约束条件：

$$\exp\left(1-\mu(\boldsymbol{x})/\tilde{P}(\boldsymbol{x})\right)=\int_{\mathbb{Y}}\exp\left(\sum_{i=1}^{n}\lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}y\triangleq Z(\boldsymbol{x};\boldsymbol{\lambda})$$

使用规范化因子 $Z(\boldsymbol{x};\boldsymbol{\lambda})$ ，我们记：

$$P_\lambda(y|\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x};\boldsymbol{\lambda})}\exp\left(\sum_{i=1}^{n}\lambda_i f_i(\boldsymbol{x},y)\right)$$

则：

$$\begin{aligned}&\max_{\mu,\boldsymbol{\lambda}}L(P_{\mu,\lambda},\mu,\boldsymbol{\lambda})\\&=\max_{\boldsymbol{\lambda}}\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\ln P_\lambda(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y\\&+0+\sum_{i=1}^{n}\lambda_i\int_{\mathbb{X}\times\mathbb{Y}}\left(\tilde{P}(\boldsymbol{x},y)-\tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\right)f_i(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\\&\triangleq\max_{\boldsymbol{\lambda}}\Psi(\boldsymbol{\lambda})\end{aligned}$$

通过求解 $\max_{\boldsymbol{\lambda}}\Psi(\boldsymbol{\lambda})$ 得到 $\boldsymbol{\lambda}$ 的解，代入到 $P_\lambda(y|\boldsymbol{x})$ 中即可得 $P(y|\boldsymbol{x})$ 解。此外，还可以根据 $\mu(\boldsymbol{x}),\boldsymbol{\lambda}$ 之间的关系得到 $\mu(\boldsymbol{x})$ 的解。

2 最大熵模型

2.1 模型描述

根据上一章的内容，我们抽象出如下的模型，并称之为最大熵模型：

$$P_\lambda(y|\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x};\boldsymbol{\lambda})}\exp\left(\sum_{i=1}^{n}\lambda_i f_i(\boldsymbol{x},y)\right)$$

其中， $f_i(\boldsymbol{x},y)$ 为给定的特征函数， $Z(\boldsymbol{x};\boldsymbol{\lambda})$ 为规范化因子，待学习的量为参数 $\boldsymbol{\lambda}$ 。可以看到，和线性模型、决策树等模型一样，最大熵模型的形式给定，我们需要对其中的待估参数进行学习。最大熵模型属于对数线性模型。

2.2 模型学习--极大似然估计

给定一个含有待估参数的机器学习模型和采样数据 D ， 我们会采用极大似然估计对未知参数进行估计：

$$\max_{\lambda} -\sum_{i=1}^{N} \ln P_{\lambda}(y_i|\boldsymbol{x}_i)$$

而事实上，上式等价于 $\max_{\lambda} \Psi(\lambda)$，接下来，我们来证明这一点。

1. 我们知道，极大似然估计：

$$\max -\sum_{i=1}^{N} \ln P_{model}(y_i|\boldsymbol{x}_i)$$

$$\Updownarrow$$

$$\max -\int_{\mathbb{X}\times\mathbb{Y}} P_{data}(\boldsymbol{x},y) \ln P_{model}(y|\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}y$$

因此，

$$\max_{\lambda} -\sum_{i=1}^{N} \ln P_{\lambda}(y_i|\boldsymbol{x}_i)$$

$$\Updownarrow$$

$$\min_{\lambda} \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y) \ln P_{\lambda}(y|\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}y$$

代入模型 $P_{\lambda}(y|\boldsymbol{x})$ 进行化简：

$$\begin{aligned}
&\int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y) \ln P_{\lambda}(y|\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y) \left(\sum_{i=1}^{n} \lambda_i f_i(\boldsymbol{x},y) - \ln Z(\boldsymbol{x};\boldsymbol{\lambda}) \right) \mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y) \left(\sum_{i=1}^{n} \lambda_i f_i(\boldsymbol{x},y) \right) \mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y) \ln Z(\boldsymbol{x};\boldsymbol{\lambda}) \mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y) \left(\sum_{i=1}^{n} \lambda_i f_i(\boldsymbol{x},y) \right) \mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathbb{X}} \tilde{P}(\boldsymbol{x}) \ln Z(\boldsymbol{x};\boldsymbol{\lambda}) \mathrm{d}\boldsymbol{x}
\end{aligned}$$

2. 对 $\Psi(\lambda)$ 变形：

$$\Psi(\lambda) =$$

$$\int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\ln P_\lambda(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y + \int_{\mathbb{X}\times\mathbb{Y}} \left(\tilde{P}(\boldsymbol{x},y) - \tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\right)\left(\sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

$$= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

$$+ \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\ln P_\lambda(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\left(\sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

$$= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

$$+ \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\left(\ln P_\lambda(y|\boldsymbol{x}) - \sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

$$= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y + \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})(-\ln Z(\boldsymbol{x};\boldsymbol{\lambda}))\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

$$= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathbb{X}} \tilde{P}(\boldsymbol{x})\ln Z(\boldsymbol{x};\boldsymbol{\lambda})\mathrm{d}\boldsymbol{x}$$

3. 可见**对最大熵模型 (存在具体的形式和待定参数) 进行极大似然估计就等价于依据最大熵原理 (对 $P(y|\boldsymbol{x})$ 的形式没有作任何假设) 构建模型。**

3 数值优化算法

由前可知，我们要解决如下的优化问题：

$$\max_{\boldsymbol{\lambda}} \Psi(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda}} \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathbb{X}} \tilde{P}(\boldsymbol{x})\ln Z(\boldsymbol{x};\boldsymbol{\lambda})\mathrm{d}\boldsymbol{x}$$

其中，除待优化变量 $\boldsymbol{\lambda}$ 未知外，其他均已知。接下来，我们介绍两种数值优化方法--改进的迭代尺度法 (improved iterative scaling, IIS) 和拟牛顿法。

3.1 改进的迭代尺度法 (improved iterative scaling, IIS)

IIS 的思路是：假设当前参数向量是 $\boldsymbol{\lambda}$，我们希望找到一个新的参数向量 $\boldsymbol{\lambda}+\boldsymbol{\delta}$，使得目标函数增大，重复这一过程，直到找到最大值。为此，我们构造 $\Psi(\boldsymbol{\lambda}+\boldsymbol{\delta})-\Psi(\boldsymbol{\lambda})$ 的下界，也就是下文中的 $B(\boldsymbol{\delta};\boldsymbol{\lambda})$，$\Psi(\boldsymbol{\lambda}+\boldsymbol{\delta})-\Psi(\boldsymbol{\lambda}) \geqslant B(\boldsymbol{\delta};\boldsymbol{\lambda})$，通过不断优化下界的值来提到目标函数的值。

1. 我们有：

$$\Psi(\boldsymbol{\lambda}+\boldsymbol{\delta})-\Psi(\boldsymbol{\lambda}) = \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathbb{X}} \tilde{P}(\boldsymbol{x})\ln\frac{Z(\boldsymbol{x};\boldsymbol{\lambda}+\boldsymbol{\delta})}{Z(\boldsymbol{x};\boldsymbol{\lambda})}\mathrm{d}\boldsymbol{x}$$

$$\geqslant \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y + \int_{\mathbb{X}} \tilde{P}(\boldsymbol{x})\left(1-\frac{Z(\boldsymbol{x};\boldsymbol{\lambda}+\boldsymbol{\delta})}{Z(\boldsymbol{x};\boldsymbol{\lambda})}\right)\mathrm{d}\boldsymbol{x}$$

$$(\text{不等式 } -\ln x \geqslant 1-x,\ x>0)$$

$$= \int_{\mathbb{X}\times\mathbb{Y}} \tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n \delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y + 1 - \int_{\mathbb{X}} \tilde{P}(\boldsymbol{x})\frac{Z(\boldsymbol{x};\boldsymbol{\lambda}+\boldsymbol{\delta})}{Z(\boldsymbol{x};\boldsymbol{\lambda})}\mathrm{d}\boldsymbol{x}$$

又由前可知：

$$Z(\boldsymbol{x};\boldsymbol{\lambda})=\frac{1}{P_\lambda(y|\boldsymbol{x})}\exp\left(\sum_{i=1}^n\lambda_i f_i(\boldsymbol{x},y)\right)=\int_{\mathbb{Y}}\exp\left(\sum_{i=1}^n\lambda_i f_i(\boldsymbol{x},y)\right)\mathrm{d}y$$

我们有：

$$\begin{aligned}Z(\boldsymbol{x};\boldsymbol{\lambda}+\boldsymbol{\delta})&=\int_{\mathbb{Y}}\exp\left(\sum_{i=1}^n\lambda_i f_i(\boldsymbol{x},y)\right)\exp\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}y\\&=\int_{\mathbb{Y}}Z(\boldsymbol{x};\boldsymbol{\lambda})P_\lambda(y|\boldsymbol{x})\exp\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}y\\&\Updownarrow\\\frac{Z(\boldsymbol{x};\boldsymbol{\lambda}+\boldsymbol{\delta})}{Z(\boldsymbol{x};\boldsymbol{\lambda})}&=\int_{\mathbb{Y}}P_\lambda(y|\boldsymbol{x})\exp\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}y\end{aligned}$$

则

$$\begin{aligned}&\Psi(\boldsymbol{\lambda}+\boldsymbol{\delta})-\Psi(\boldsymbol{\lambda})\\\geqslant&\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y+1-\int_{\mathbb{X}}\tilde{P}(\boldsymbol{x})\frac{Z(\boldsymbol{x};\boldsymbol{\lambda}+\boldsymbol{\delta})}{Z(\boldsymbol{x};\boldsymbol{\lambda})}\mathrm{d}\boldsymbol{x}\\=&\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y+1-\int_{\mathbb{X}}\tilde{P}(\boldsymbol{x})\int_{\mathbb{Y}}P_\lambda(y|\boldsymbol{x})\exp\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}y\mathrm{d}\boldsymbol{x}\\=&\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y+1-\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\exp\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y\\\triangleq&A(\boldsymbol{\delta};\boldsymbol{\lambda})\end{aligned}$$

至此，我们找到了一个下界 $A(\boldsymbol{\delta};\boldsymbol{\lambda})$，但对 $A(\boldsymbol{\delta};\boldsymbol{\lambda})$ 的优化依然稍显复杂，为此，下面我们作进一步缩放。

2. 记 $f^\#(\boldsymbol{x},y)=\sum_{i=1}^n f_i(\boldsymbol{x},y)$，由 Jensen 不等式：

$$\begin{aligned}\exp\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)&=\exp\left(\sum_{i=1}^n\frac{f_i(\boldsymbol{x},y)}{f^\#(\boldsymbol{x},y)}\delta_i f^\#(\boldsymbol{x},y)\right)\\&\leqslant\sum_{i=1}^n\frac{f_i(\boldsymbol{x},y)}{f^\#(\boldsymbol{x},y)}\exp\left(\delta_i f^\#(\boldsymbol{x},y)\right)\end{aligned}$$

则：

$$\begin{aligned}A(\boldsymbol{\delta};\boldsymbol{\lambda})\leqslant&\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)\left(\sum_{i=1}^n\delta_i f_i(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y+1\\&-\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})\sum_{i=1}^n\frac{f_i(\boldsymbol{x},y)}{f^\#(\boldsymbol{x},y)}\exp\left(\delta_i f^\#(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y\triangleq B(\boldsymbol{\delta};\boldsymbol{\lambda})\end{aligned}$$

至此，我们得到了下界 $B(\boldsymbol{\delta};\boldsymbol{\lambda})$，即：

$$\Psi(\boldsymbol{\lambda}+\boldsymbol{\delta})-\Psi(\boldsymbol{\lambda})\geqslant B(\boldsymbol{\delta};\boldsymbol{\lambda})$$

3. 下面我们说明为什么优化下界 $B(\boldsymbol{\delta};\boldsymbol{\lambda})$ 比优化下界 $A(\boldsymbol{\delta};\boldsymbol{\lambda})$ 更简单，及如何对 $B(\boldsymbol{\delta};\boldsymbol{\lambda})$ 进行优化。将 $B(\boldsymbol{\delta};\boldsymbol{\lambda})$ 对任意分量 δ_i 求导，并令导数为 0，我们有：

$$\frac{\partial B(\boldsymbol{\delta};\boldsymbol{\lambda})}{\partial\delta_i}=\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)f_i(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y-\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})f_i(\boldsymbol{x},y)\exp\left(\delta_i f^\#(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y=0$$

可以看到，上式中的未知量只有分量 δ_i，而含有其他分量 $\delta_j\ (j\neq i)$，因此对每个分量 δ_i，单独求解对应的方程即可；但若选择优化下界 $A(\boldsymbol{\delta};\boldsymbol{\lambda})$，为了求解 $\boldsymbol{\delta}$，我们需求解的是一个联立的方程组，难度更大。

此外，若对任意的 $\boldsymbol{x},y$，$f^{\#}(\boldsymbol{x},y)\equiv C$，则 δ_i 有如下的解析解：

$$\delta_i=\frac{1}{C}\ln\frac{\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)f_i(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y}{\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})f_i(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y}=\frac{1}{C}\ln\frac{\mathbb{E}_{\tilde{P}}[f_i]}{\mathbb{E}_P[f_i]}$$

否则，我们可以采用牛顿法迭代求解方程：

$$\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x},y)f_i(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y-\int_{\mathbb{X}\times\mathbb{Y}}\tilde{P}(\boldsymbol{x})P_\lambda(y|\boldsymbol{x})f_i(\boldsymbol{x},y)\exp\left(\delta_i f^{\#}(\boldsymbol{x},y)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}y=0$$

上述方程有单根，牛顿法恒收敛且收敛速度很快。

3.2 拟牛顿法

按照拟牛顿法的格式进行求解即可，可参见笔记《7-数值优化方法》中的相关内容，《统计学习方法》中给出的是 BFGS 算法，不再赘述。