最大熵算法

一、最大熵模型MEM

1. 设随机变量 X 的概率分布为 $P(X)$，熵为：$H(P)=-\sum_X P(X)\log P(X)$ 。

   可以证明：$0\leq H(P)\leq \log|X|$ ，其中 $|X|$ 为 X 的取值的个数。

   当且仅当 X 的分布为均匀分布时有 $H(P)=\log|X|$ 。即 $P(X)=\frac{1}{|X|}$ 时熵最大。

1.1 最大熵原理

1. 最大熵`Max Entropy`原理：学习概率模型时，在所有可能的概率模型（即概率分布）中，熵最大的模型是最好的模型。
   - 通常还有其他已知条件来确定概率模型的集合，因此最大熵原理为：在满足已知条件的情况下，选取熵最大的模型。
   - 在满足已知条件前提下，如果没有更多的信息，则那些不确定部分都是“等可能的”。而`等可能性`通过熵最大化来刻画。
2. 最大熵原理选取熵最大的模型，而决策树的划分目标选取熵最小的划分。原因在于：
   - 最大熵原理认为在满足已知条件之后，选择不确定性最大（即：不确定的部分是等可能的）的模型。也就是不应该再施加任何额外的约束。

     因此这是一个求最大不确定性的过程，所以选择熵最大的模型。
   - 决策树的划分目标是为了通过不断的划分从而不断的降低实例所属的类的不确定性，最终给实例一个合适的分类。因此这是一个不确定性不断减小的过程，所以选取熵最小的划分。

1.2 期望的约束

1. 一种常见的约束为期望的约束：$\mathbb{E}[f(X)]=\sum_X P(X)f(X)=\tau$ ，其中 $f(\cdot)$ 代表随机变量 X 的某个函数（其结果是另一个随机变量）。
   - 其物理意义为：随机变量 $\tilde{X}=f(X)$ 的期望是一个常数。
   - 示例：当 $f(X)=X$ 时，约束条件为：$\mathbb{E}[X]=\tau$ ，即随机变量 X 的期望为常数。
2. 如果有多个这样的约束条件：

$$\begin{aligned}\mathbb{E}[f_1(X)]&=\sum_X P(X)f_1(X)=\tau_1\\&\vdots\\\mathbb{E}[f_k(X)]&=\sum_X P(X)f_k(X)=\tau_k\end{aligned}$$

   则需要求解约束最优化问题：

$$\max_{P(X)} -\sum_X P(X)\log P(X)$$
$$s.t. \sum_X P(X) = 1, 0 \le P \le 1$$
$$\mathbb{E}[f_1(X)] = \sum_X P(X) f_1(X) = \tau_1$$
$$\vdots$$
$$\mathbb{E}[f_k(X)] = \sum_X P(X) f_k(X) = \tau_k$$

3. 给出拉格朗日函数：

$$L(P) = -\sum_X P(X)\log P(X) + \lambda_0\left(\sum_X P(X) - 1\right) + \lambda_1\left(\sum_X P(X) f_1(X) - \tau_1\right) + \cdots + \lambda_k\left(\sum_X P(X) f_k(X) - \tau_k\right)$$

可以求得：

$$P(X) = \frac{1}{Z}\exp\left(-\sum_{i=1}^k \lambda_i f_i(X)\right), \quad Z = \sum_X \exp\left(-\sum_{i=1}^k \lambda_i f_i(X)\right)$$

- 将 $P(X)$ 代入 ，有：

$$\mathbb{E}[f_1(X)] = \sum_X \frac{f_1(X)}{Z}\exp\left(-\sum_{i=1}^k \lambda_i f_i(X)\right) = \tau_1$$
$$\vdots$$
$$\mathbb{E}[f_k(X)] = \sum_X \frac{f_k(X)}{Z}\exp\left(-\sum_{i=1}^k \lambda_i f_i(X)\right) = \tau_k$$

  则可以求解出各个 λ_i 。

- 该式子并没有解析解，而且数值求解也相当困难。

4. 当只有一个约束 $f(X) = X$ 时，表示约束了变量的期望，即 $\sum_X P(X)X = \tau$ 。此时有：

$$P(X) = \frac{\exp(-\lambda X)}{\sum_X \exp(-\lambda X)}$$

代入 $\sum_X P(X)X = \tau$ ，解得：$P(X) = \frac{1}{\tau}\exp\left(-\frac{X}{\tau}\right)$ 。

即：约束了随机变量期望的分布为指数分布。

5. 当有两个约束 $f_1(X) = X, \quad f_2(X) = X^2$ 时，表示约束了变量的期望和方差。即：

$$\sum_X P(X)X = \tau_1, \quad \sum_X P(X)X^2 = \tau_2$$

此时有：

$$P(X) = \frac{\exp(-\lambda_1 X - \lambda_2 X)}{\sum_X \exp(-\lambda_1 X - \lambda_2 X)}$$

代入约束可以解得：

$$P(X)=\sqrt{\frac{1}{2\pi(\tau_2-\tau_1^2)}}\exp\left(-\frac{(X-\tau_1)^2}{2(\tau_2-\tau_1^2)}\right)$$

它是均值为 τ_1 ，方差为 $\tau_2-\tau_1^2$ 的正态分布。即：约束了随机变量期望、方差的分布为正态分布。

二、分类任务最大熵模型

1. 设分类模型是一个条件概率分布 $P(Y\mid X=\vec{\mathbf{x}}),X\in\mathcal{X}\subseteq\mathbb{R}^n$ 为输入， $Y\in\mathcal{Y}$ 为输出。

   给定一个训练数据集 $\mathbb{D}=\{(\vec{\mathbf{x}}_1,y_1),(\vec{\mathbf{x}}_2,y_2),\cdots,(\vec{\mathbf{x}}_N,y_N)\}$，学习的目标是用最大熵原理选取最好的分类模型。

2.1 最大熵模型

1. 根据训练集 $\mathbb{D}$，可以得到联合分布 $P(X,Y)$ 的经验分布 $\tilde{P}(X,Y)$ 和 $P(X)$ 的经验分布 $\tilde{P}(X)$：

   $$\tilde{P}(X=\vec{\mathbf{x}},Y=y)=\frac{\upsilon(X=\vec{\mathbf{x}},Y=y)}{N},\quad\vec{\mathbf{x}}\in\mathcal{X},y\in\mathcal{Y}$$
   $$\tilde{P}(X)=\frac{\upsilon(X=\vec{\mathbf{x}})}{N},\quad\vec{\mathbf{x}}\in\mathcal{X}$$

   其中 N 为样本数量， υ 为频数。

2. 用特征函数 $f(\vec{\mathbf{x}},y)$ 描述输入 $\vec{\mathbf{x}}$ 和输出 y 之间的某个事实：

   $$f(\vec{\mathbf{x}},y)=\begin{cases}1, & \text{if } \vec{\mathbf{x}},y \text{ statisfy the fact.}\\ 0, & \text{or else.}\end{cases}$$

   - 特征函数是一个二值函数，但是理论上它也可以取任意值。
   - 特征函数 $f(\vec{\mathbf{x}},y)$ 关于经验分布 $\tilde{P}(X,Y)$ 的期望定义为 $\mathbb{E}_{\tilde{P}}[f]$ ：$\mathbb{E}_{\tilde{P}}[f]=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}},y)f(\vec{\mathbf{x}},y)$ 。

     这个期望其实就是约束 f 在训练集上的统计结果的均值（也就是约束 f 出现的期望的估计量）。

     - 如果 f 取值为二值 `0,1` ，则表示约束 f 在训练集上出现的次数的均值。
     - 如果 f 取值为任意值，则表示约束 f 在训练集上累计的结果的均值。
   - 特征函数 $f(\vec{\mathbf{x}},y)$ 关于模型 $P(Y\mid X)$ 与经验分布 $\tilde{P}(X)$ 的期望用 $\mathbb{E}_P[f]$ 表示：

     $$\mathbb{E}_P[f]=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})f(\vec{\mathbf{x}},y)$$

     理论上 $\mathbb{E}_P[f]=\sum_{\vec{\mathbf{x}},y}P(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})f(\vec{\mathbf{x}},y)$ ，这里使用 $\tilde{P}(\vec{\mathbf{x}})$ 作为 $P(\vec{\mathbf{x}})$ 的估计。
   - 可以假设这两个期望相等，即：$\mathbb{E}_{\tilde{P}}[f]=\mathbb{E}_P[f]$。
     - $\tilde{P}(\vec{\mathbf{x}},y)$ 在 $(\vec{\mathbf{x}},y)\notin\mathbb{D}$ 时为 0，在 $(\vec{\mathbf{x}},y)\in\mathbb{D}$ 才有可能非 0。因此 $\mathbb{E}_{\tilde{P}}[f]=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}},y)f(\vec{\mathbf{x}},y)$ 仅仅在 $(\vec{\mathbf{x}},y)\in\mathbb{D}$ 上累加。
     - $\tilde{P}(\vec{\mathbf{x}})$ 在 $\vec{\mathbf{x}}\notin\mathbb{D}$ 时为 0，在 $\vec{\mathbf{x}}\in\mathbb{D}$ 才有可能非 0。因此 $\mathbb{E}_P[f]=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})f(\vec{\mathbf{x}},y)$ 仅在 $\vec{\mathbf{x}}\in\mathbb{D}$ 上累加。

3. 理论上，由于 $P(y\mid\vec{\mathbf{x}})=\frac{P(\vec{\mathbf{x}},y)}{P(\vec{\mathbf{x}})}$ ，看起来可以使用 $\frac{\tilde{P}(\vec{\mathbf{x}},y)}{\tilde{P}(\vec{\mathbf{x}})}$ 作为 $P(y\mid\vec{\mathbf{x}})$ 的一个估计。

   但是这个估计只考虑某个点 $(\vec{\mathbf{x}},y)$ 上的估计，并未考虑任何约束。所以这里通过特征函数的两种期望相等来构建在数据集整体上的最优估计。

4. 最大熵模型：假设有 n 个约束条件 $f_i(\vec{\mathbf{x}},y),i=1,2,\cdots,n$，满足所有约束条件的模型集合为：$\mathcal{C}=\{P\in\mathcal{P}\mid\mathbb{E}_P[f_i]=\mathbb{E}_{\tilde{P}}[f_i],i=1,2,\cdots,n\}$ 。

定义在条件概率分布 $P(Y\mid X)$ 上的条件熵为：

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

则模型集合 $\mathcal{C}$ 中条件熵最大的模型称为最大熵模型。

2.2 词性标注约束案例

1. 在词性标注任务中，给定单词序列 $\{v_1,v_2,v_3,\cdots,v_d\}$，需要给出每个单词对应的词性 $\{s_1,s_2,s_3,\cdots,s_d\}$。如：`{他们 吃 苹果}` 对应的标注序列为 `{代词 动词 名词}`。

   假设标注仅仅与当前单词有关，与前面、后面的单词无关，也无前面、后面的标注有关。即：标注 s_i 由单词 v_i 唯一决定。

   则统计文本中所有单词及其词性，得到训练集 $\mathbb{D}=\{(v_1,s_1),\cdots,(v_N,s_N)\}$，其中 N 为单词数量。

2. 假设没有任何约束，则每个单词取得任何词性的概率都是等可能的。现在发现：`苹果` 这个单词的词性标记结果中，大部分都是`名词`，因此可以定义特征函数：

   $$f(v,s)=\begin{cases}1, & \text{当}v\text{为苹果,且}s\text{为名词时}\\0, & \text{其他情况}\end{cases}$$

   统计满足特征函数的样本的个数 N_f，除以样本总数 N。则可以认为：当数据足够多时，这个商就是统计意义下的结果：

   $$\mathbb{E}_{\tilde{P}}[f(v,s)]=\sum_{v,s}\tilde{P}(v,s)f(v,s)=\frac{N_f}{N}$$

   其中：

   - $\tilde{P}(v,s)=\frac{N_{v,s}}{N}$，$N_{v,s}$ 为二元对 (v,s) 出现的次数。
   - 满足特征函数的样本出现总数为：$\sum_{v,s\text{ where }f(v,s)=1}N_{v,s}=N_f$。

3. 事实上对于任意单词 $v\in\mathbb{V}$，其中 $\mathbb{V}=\{\mathbf{v}_1,\cdots,\mathbf{v}_V\}$ 为所有单词的词汇表，V 为词汇表大小；以及对任意词性 $s\in\mathbb{S}$，其中 $\mathbb{S}=\{\mathbf{s}_1,\cdots,\mathbf{s}_S\}$ 为词性集合（如名词、动词、形容词....），S 为词性表大小。可以任意选择搭配从而构造非常庞大的特征函数：

   $$f_{1,1}(v,s)=\begin{cases}1, & v=\mathbf{v}_1,\text{and }s=\mathbf{s}_1\\0, & other\end{cases}$$
   $$f_{1,2}(v,s)=\begin{cases}1, & v=\mathbf{v}_1,\text{and }s=\mathbf{s}_2\\0, & other\end{cases}$$
   $$\vdots$$
   $$f_{i,j}(v,s)=\begin{cases}1, & v=\mathbf{v}_i,\text{and }s=\mathbf{s}_j\\0, & other\end{cases}$$
   $$\vdots$$
   $$\mathbf{v}_i\in\mathbb{V},\mathbf{s}_j\in\mathbb{S}$$

   以及约束条件：$\mathbb{E}_{\tilde{P}}[f_{i,j}(v,s)]=\sum_{v,s}\tilde{P}(v,s)f_{i,j}(v,s)=\frac{N_{f_{i,j}}}{N}$。其中$N_{f_{i,j}}$ 为满足特征函数 $f_{i,j}$ 的样本个数。

   - 如果 $N_{f_{i,j}}$ 较大，则说明该约束指定的 `单词,词性` 搭配的可能性很高。
   - 如果 $N_{f_{i,j}}$ 较小，则说明该约束指定的 `单词,词性` 搭配的可能性很低。
   - 如果 $N_{f_{i,j}}$ 为 0，则说明该约束指定的 `单词,词性` 搭配几乎不可能出现。

4. 待求的模型为 $P(s\mid v)$ 。以矩阵的形式描述为：

$$\mathbf{P}=\begin{bmatrix}P_{1,1}&P_{1,2}&\cdots&P_{1,S}\\P_{2,1}&P_{2,2}&\cdots&P_{2,S}\\\vdots&\vdots&\ddots&\vdots\\P_{V,1}&P_{V,2}&\cdots&P_{V,S}\end{bmatrix},\quad P_{i,j}\geq 0,\quad \sum_{j=1}^{S}P_{i,j}=1,i=1,2,\cdots,V$$

其中 $P_{i,j}=P(s=\mathbf{s}_j\mid v=\mathbf{v}_i)$，即单词 $\mathbf{v}_i$ 的词性为 $\mathbf{s}_j$ 的概率。

- 设单词 $\mathbf{v}_i$ 在 $\mathbb{D}$ 中出现的次数为 N_i ，则有：$\tilde{P}(\mathbf{v}_i)=\frac{N_i}{N}$ 。则有：

$$\mathbb{E}_P[f_{i,j}]=\sum_{v,s}\tilde{P}(v)P(s\mid v)f_{i,j}(v,s)=\frac{N_i}{N}\times P_{i,j}$$

- 考虑到 $\mathbb{E}_{\tilde{P}}[f_{i,j}(v,s)]=\frac{N_{f_{i,j}}}{N}$ ，则根据 $\mathbb{E}_{\tilde{P}}[f_{i,j}(v,s)]=\mathbb{E}_P[f_{i,j}]$ 有：

$$P_{i,j}=\frac{N_{f_{i,j}}}{N_i}$$

  - 其物理意义为：单词 $\mathbf{v}_i$ 的词性为 $\mathbf{s}_j$ 的概率 = 数据集 $\mathbb{D}$ 中单词 $\mathbf{v}_i$ 的词性为 $\mathbf{s}_j$ 出现的次数 / 数据集 $\mathbb{D}$ 中单词 $\mathbf{v}_i$ 出现的次数。
  - 由于 $\tilde{P}(v=\mathbf{v}_i,s=\mathbf{s}_j)=\frac{N_{f_{i,j}}}{N}$ ， $\tilde{P}(v=\mathbf{v}_i)=\frac{N_i}{N}$ ，因此可以发现有：

$$\frac{\tilde{P}(v=\mathbf{v}_i,s=\mathbf{s}_j)}{\tilde{P}(v=\mathbf{v}_i)}=\frac{N_{f_{i,j}}}{N_i}=P_{i,j}=P(s=\mathbf{s}_j\mid v=\mathbf{v}_i)$$

    因此在这个特殊的情形下，$\frac{\tilde{P}(v,s)}{\tilde{P}(v)}$ 是 $\tilde{P}(s\mid v)$ 的估计。

5. 事实上，真实的词性标注还需要考虑前后单词的词性的影响。比如：不可能出现连续的三个动词，也不可能出现连续的五个代词。

   当需要考虑前后文影响时，需要使用 `HMM` 模型 或者 `CRF` 模型。

2.3 模型求解

1. 对给定的训练数据集 $\mathbb{D}=\{(\vec{\mathbf{x}}_1,y_1),(\vec{\mathbf{x}}_2,y_2),\cdots,(\vec{\mathbf{x}}_N,y_N)\}$，以及特征函数 $f_i(\vec{\mathbf{x}},y),i=1,2,\cdots,n$ ，最大熵模型的学习等价于约束最优化问题：

$$\max_{P\in\mathcal{C}}H(P)=-\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})\log P(y\mid\vec{\mathbf{x}})$$
$$s.t.\ \mathbb{E}_P[f_i]=\mathbb{E}_{\tilde{P}}[f_i],i=1,2,\cdots,n$$
$$\sum_y P(y\mid\vec{\mathbf{x}})=1$$

2. 将其转化为最小化问题：

$$\min_{P\in\mathcal{C}}-H(P)=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})\log P(y\mid\vec{\mathbf{x}})$$
$$s.t.\ \mathbb{E}_P[f_i]-\mathbb{E}_{\tilde{P}}[f_i]=0,i=1,2,\cdots,n$$
$$\sum_y P(y\mid\vec{\mathbf{x}})=1$$

   其中：

- $\tilde{P}(\vec{\mathbf{x}}), \mathbb{E}_{\tilde{P}}[f_i]=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}},y)f_i(\vec{\mathbf{x}},y)$ 是已知的。
- $P(y\mid\vec{\mathbf{x}}), \mathbb{E}_{P}[f_i]=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})f_i(\vec{\mathbf{x}},y)$ 是未知的。

3. 将约束最优化的原始问题转换为无约束最优化的对偶问题，通过求解对偶问题来求解原始问题。

   引入拉格朗日乘子 $w_0,w_1,\cdots,w_n$ ，定义拉格朗日函数 $L(P,\vec{\mathbf{w}})$ ：

$$\begin{aligned}L(P,\vec{\mathbf{w}})&=-H(P)+w_0(1-\sum_y P(y\mid\vec{\mathbf{x}}))+\sum_{i=1}^n w_i(\mathbb{E}_{\tilde{P}}[f_i]-E_P(f_i))\\&=\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})\log P(y\mid\vec{\mathbf{x}})+w_0\left(1-\sum_y P(y\mid\vec{\mathbf{x}})\right)\\&+\sum_{i=1}^n w_i\left(\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}},y)f_i(\vec{\mathbf{x}},y)-\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}})P(y\mid\vec{\mathbf{x}})f_i(\vec{\mathbf{x}},y)\right)\end{aligned}$$

   - 最优化的原始问题是：$\min_{P\in\mathcal{C}}\max_{\vec{\mathbf{w}}}L(P,\vec{\mathbf{w}})$ ，对偶问题是 $\max_{\vec{\mathbf{w}}}\min_{P\in\mathcal{C}}L(P,\vec{\mathbf{w}})$ 。
   - 由于拉格朗日函数 $L(P,\vec{\mathbf{w}})$ 是凸函数，因此原始问题的解与对偶问题的解是等价的。
   - 求解对偶问题：先求解内部的极小化问题，之后求解对偶问题外部的极大化问题。

4. 先求解内部的极小化问题：$\min_{P\in\mathcal{C}}L(P,\vec{\mathbf{w}})$ 。

   它是一个 $\vec{\mathbf{w}}$ 的函数，将其记作：$\Psi(\vec{\mathbf{w}})=\min_{P\in\mathcal{C}}L(P,\vec{\mathbf{w}})=L(P_{\vec{\mathbf{w}}},\vec{\mathbf{w}})$ 。

   - 先用 $L(P,\vec{\mathbf{w}})$ 对 $P(y\mid\vec{\mathbf{x}})$ 求偏导数：

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

     令偏导数为 0 。在 $\tilde{P}(\vec{\mathbf{x}})>0$ 时，解得：

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

   - 由于 $\sum_y P(y\mid\vec{\mathbf{x}})=1$ ，则有：$\sum_y\frac{\exp(\sum_{i=1}^n w_i f_i(\vec{\mathbf{x}},y))}{\exp(1-w_0)}=1$ 。因此有：

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

   - 定义 $Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})=\sum_y\exp(\sum_{i=1}^n w_i f_i(\vec{\mathbf{x}},y))$ 为规范因子，则：

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

     由该式表示的模型 $P_{\vec{\mathbf{w}}}=P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})$ 就是最大熵模型。

5. 再求解对偶问题外部的极大化问题：$\max_{\vec{\mathbf{w}}}\Psi(\vec{\mathbf{w}})$ 。

   - 将其解记作 $\vec{\mathbf{w}}^*$ ，即：$\vec{\mathbf{w}}^*=\arg\max_{\vec{\mathbf{w}}}\Psi(\vec{\mathbf{w}})$ 。
   - 求得 $\vec{\mathbf{w}}^*$ 之后，用它来表示 $P_{\vec{\mathbf{w}}}=P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})$ ，得到 $P^*=P_{\vec{\mathbf{w}}^*}=P_{\vec{\mathbf{w}}^*}(y\mid\vec{\mathbf{x}})$ ，即得到最大熵模型。

6. 上述过程总结为：

   - 先求对偶问题的内部极小化，得到 $\Psi(\vec{\mathbf{w}})$ 函数，以及极值点 $P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})$ 。
   - 再求 $\Psi(\vec{\mathbf{w}})$ 函数的极大值，得到 $\vec{\mathbf{w}}^*$ 。

- 最后将 $\vec{\mathbf{w}}^*$ 代入 $P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})$ 得到最终模型 P^*。

7. 可以证明： $\Psi(\vec{\mathbf{w}})$ 函数的最大化，等价于最大熵模型的极大似然估计。

   证明如下：已知训练数据 $\mathbb{D}$ 中，$(\vec{\mathbf{x}},y)$ 出现的频次为 $k_{\vec{\mathbf{x}},y}$。则条件概率分布 $P(y\mid\vec{\mathbf{x}})$ 的对数似然函数为：

$$\log\prod_{\vec{\mathbf{x}},y}P(y\mid\vec{\mathbf{x}})^{k_{\vec{\mathbf{x}},y}}=\sum_{\vec{\mathbf{x}},y}k_{\vec{\mathbf{x}},y}\log P(y\mid\vec{\mathbf{x}})$$

   将对数似然函数除以常数 N，考虑到 $\frac{k_{\vec{\mathbf{x}},y}}{N}=\tilde{P}(\vec{\mathbf{x}},y)$ ，其中 $\tilde{P}(\vec{\mathbf{x}},y)$ 为经验概率分布。则 $P(y\mid\vec{\mathbf{x}})$ 的对数似然函数为：

$$\sum_{\vec{\mathbf{x}},y}\tilde{P}(\vec{\mathbf{x}},y)\log P(y\mid\vec{\mathbf{x}})$$

   再利用 ：

$$P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})=\frac{1}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y)\right)$$

   代入，最后化简合并，最终发现它就是 $\Psi(\vec{\mathbf{w}})$。

2.4 最大熵与逻辑回归

1. 设 $\vec{\mathbf{x}}=(x_1,x_2,\cdots,x_n)^T$为 n 维变量，对于二类分类问题，定义 n 个约束：

$$f_i(\vec{\mathbf{x}},y)=\begin{cases}x_i, & y=1\\ 0, & y=0\end{cases},\quad i=1,2,\cdots,n$$

2. 根据最大熵的结论，有：

$$\begin{aligned}Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})&=\sum_y\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y)\right)=\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y=0)\right)+\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y=1)\right)\\&=1+\exp\left(\sum_{i=1}^{n}w_ix_i\right)=1+\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})\end{aligned}$$

   以及：

$$P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})=\frac{1}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y)\right)=\frac{1}{1+\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})}\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y)\right)$$

   - 当 $y=1$ 时有：

$$\begin{aligned}P_{\vec{\mathbf{w}}}(y=1\mid\vec{\mathbf{x}})&=\frac{1}{1+\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})}\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y=1)\right)\\&=\frac{1}{1+\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})}\exp\left(\sum_{i=1}^{n}w_ix_i\right)=\frac{\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})}{1+\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})}\end{aligned}$$

   - 当 $y=0$ 时有：

$$P_{\vec{\mathbf{w}}}(y=0\mid\vec{\mathbf{x}})=\frac{1}{1+\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})}\exp\left(\sum_{i=1}^{n}w_if_i(\vec{\mathbf{x}},y=0)\right)=\frac{1}{1+\exp(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}})}$$

   最终得到：

$$\log \frac{P_{\vec{\mathbf{w}}}(y=1 \mid \vec{\mathbf{x}})}{P_{\vec{\mathbf{w}}}(y=0 \mid \vec{\mathbf{x}})}=\vec{\mathbf{w}} \cdot \vec{\mathbf{x}}$$

这就是逻辑回归模型。

三、最大熵的学习

1. 最大熵模型的学习就是在给定训练数据集 $\mathbb{D}$ 时，对模型进行极大似然估计或者正则化的极大似然估计。
2. 最大熵模型与 `logistic` 回归模型有类似的形式，它们又称为对数线性模型。
   - 它们的目标函数具有很好的性质：光滑的凸函数。因此有多种最优化方法可用，且保证能得到全局最优解。
   - 最常用的方法有：改进的迭代尺度法、梯度下降法、牛顿法、拟牛顿法。

3.1 改进的迭代尺度法

1. 改进的迭代尺度法 `Improved Iterative Scaling:IIS` 是一种最大熵模型学习的最优化算法。
2. 已知最大熵模型为：

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

   其中

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

   对数似然函数为：

$$L(\vec{\mathbf{w}})=\log \prod_{\vec{\mathbf{x}}, y} P_{\vec{\mathbf{w}}}(y \mid \vec{\mathbf{x}})^{\tilde{P}(\vec{\mathbf{x}}, y)}=\sum_{\vec{\mathbf{x}}, y}\left[\tilde{P}(\vec{\mathbf{x}}, y) \log P_{\vec{\mathbf{w}}}(y \mid \vec{\mathbf{x}})\right]$$

$$=\sum_{\vec{\mathbf{x}}, y}\left(\tilde{P}(\vec{\mathbf{x}}, y) \sum_{i=1}^{n} w_i f_i(\vec{\mathbf{x}}, y)\right)-\sum_{\vec{\mathbf{x}}}\left(\tilde{P}(\vec{\mathbf{x}}) \log Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})\right)$$

   最大熵模型的目标是：通过极大化似然函数学习模型参数，求出使得对数似然函数最大的参数 $\hat{\vec{\mathbf{w}}}$ 。

3. `IIS` 原理：假设最大熵模型当前的参数向量是 $\vec{\mathbf{w}}=(w_1, w_2, \cdots, w_n)^T$ ，希望找到一个新的参数向量 $\vec{\mathbf{w}}+\vec{\delta}=(w_1+\delta_1, w_2+\delta_2, \cdots, w_n+\delta_n)^T$ ，使得模型的对数似然函数值增大。
   - 若能找到这样的新参数向量，则更新 $\vec{\mathbf{w}} \leftarrow \vec{\mathbf{w}}+\vec{\delta}$ 。
   - 重复这一过程，直到找到对数似然函数的最大值。
4. 对于给定的经验分布 $\tilde{P}(\vec{\mathbf{x}}, y)$ ，模型参数从 $\vec{\mathbf{w}}$ 到 $\vec{\mathbf{w}}+\vec{\delta}$ 之间，对数似然函数的改变量为：

$$L(\vec{\mathbf{w}}+\vec{\delta})-L(\vec{\mathbf{w}})=\sum_{\vec{\mathbf{x}}, y}\left(\tilde{P}(\vec{\mathbf{x}}, y) \sum_{i=1}^{n} \delta_i f_i(\vec{\mathbf{x}}, y)\right)-\sum_{\vec{\mathbf{x}}}\left(\tilde{P}(\vec{\mathbf{x}}) \log \frac{Z_{\vec{\mathbf{w}}+\vec{\delta}}(\vec{\mathbf{x}})}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\right)$$

   - 利用不等式：当$\alpha>0$时 $-\log \alpha \geq 1-\alpha$, 有：

$$L(\vec{\mathbf{w}}+\vec{\delta})-L(\vec{\mathbf{w}})\geq\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)+\sum_{\vec{\mathbf{x}}}\left[\tilde{P}(\vec{\mathbf{x}})\left(1-\frac{Z_{\vec{\mathbf{w}}+\vec{\delta}}(\vec{\mathbf{x}})}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\right)\right]$$
$$=\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)+\sum_{\vec{\mathbf{x}}}\tilde{P}(\vec{\mathbf{x}})-\sum_{\vec{\mathbf{x}}}\left(\tilde{P}(\vec{\mathbf{x}})\frac{Z_{\vec{\mathbf{w}}+\vec{\delta}}(\vec{\mathbf{x}})}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\right)$$

- 考虑到 $\sum_{\vec{\mathbf{x}}}\tilde{P}(\vec{\mathbf{x}})=1$，以及：

$$\frac{Z_{\vec{\mathbf{w}}+\vec{\delta}}(\vec{\mathbf{x}})}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}=\frac{\sum_y\exp(\sum_{i=1}^{n}(w_i+\delta_i)f_i(\vec{\mathbf{x}},y))}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}$$
$$=\frac{1}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\sum_y\left[\exp\left(\sum_{i=1}^{n}w_i f_i(\vec{\mathbf{x}},y)\right)\cdot\exp\left(\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)\right]$$
$$=\sum_y\left[\frac{1}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\cdot\exp\left(\sum_{i=1}^{n}w_i f_i(\vec{\mathbf{x}},y)\right)\cdot\exp\left(\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)\right]$$

  根据$P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})=\frac{1}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}\exp(\sum_{i=1}^{n}w_i f_i(\vec{\mathbf{x}},y))$ 有：

$$\frac{Z_{\vec{\mathbf{w}}+\vec{\delta}}(\vec{\mathbf{x}})}{Z_{\vec{\mathbf{w}}}(\vec{\mathbf{x}})}=\sum_y\left[P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})\cdot\exp\left(\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)\right]$$

  则有：

$$L(\vec{\mathbf{w}}+\vec{\delta})-L(\vec{\mathbf{w}})\geq\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)+1$$
$$-\sum_{\vec{\mathbf{x}}}\left[\tilde{P}(\vec{\mathbf{x}})\sum_y\left(P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})\exp\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)\right]$$

- 令

$$A(\vec{\delta}\mid\vec{\mathbf{w}})=\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)+1-\sum_{\vec{\mathbf{x}}}\left[\tilde{P}(\vec{\mathbf{x}})\sum_y\left(P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})\exp\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)\right]$$

  则 $L(\vec{\mathbf{w}}+\vec{\delta})-L(\vec{\mathbf{w}})\geq A(\vec{\delta}\mid\vec{\mathbf{w}})$ 。

5. 如果能找到合适的 $\vec{\delta}$ 使得 $A(\vec{\delta}\mid\vec{\mathbf{w}})$ 提高，则对数似然函数也会提高。但是 $\vec{\delta}$ 是个向量，不容易同时优化。
   - 一个解决方案是：每次只优化一个变量 δ_i 。
   - 为达到这个目的，引入一个变量 $f^o(\vec{\mathbf{x}},y)=\sum_{i=1}^{n}f_i(\vec{\mathbf{x}},y)$。
6. $A(\vec{\delta}\mid\vec{\mathbf{w}})$ 改写为：

$$A(\vec{\delta}\mid\vec{\mathbf{w}})=\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^{n}\delta_i f_i(\vec{\mathbf{x}},y)\right)+1$$
$$-\sum_{\vec{\mathbf{x}}}\left[\tilde{P}(\vec{\mathbf{x}})\sum_y\left(P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})\exp\left(f^o(\vec{\mathbf{x}},y)\sum_{i=1}^{n}\frac{\delta_i f_i(\vec{\mathbf{x}},y)}{f^o(\vec{\mathbf{x}},y)}\right)\right)\right]$$

   - 利用指数函数的凸性，根据

$$\frac{f_i(\vec{\mathbf{x}},y)}{f^o(\vec{\mathbf{x}},y)}\geq 0,\quad\sum_{i=1}^{n}\frac{f_i(\vec{\mathbf{x}},y)}{f^o(\vec{\mathbf{x}},y)}=1$$

以及 `Jensen` 不等式有：

$$\exp\left(f^o(\vec{\mathbf{x}},y)\sum_{i=1}^n\frac{\delta_i f_i(\vec{\mathbf{x}},y)}{f^o(\vec{\mathbf{x}},y)}\right)\le\sum_{i=1}^n\left(\frac{f_i(\vec{\mathbf{x}},y)}{f^o(\vec{\mathbf{x}},y)}\exp(\delta_i f^o(\vec{\mathbf{x}},y))\right)$$

于是：

$$A(\vec{\delta}\mid\vec{\mathbf{w}})\ge\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^n\delta_i f_i(\vec{\mathbf{x}},y)\right)+1$$
$$-\sum_{\vec{\mathbf{x}}}\left[\tilde{P}(\vec{\mathbf{x}})\sum_y\left(P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})\sum_{i=1}^n\left(\frac{f_i(\vec{\mathbf{x}},y)}{f^o(\vec{\mathbf{x}},y)}\exp(\delta_i f^o(\vec{\mathbf{x}},y))\right)\right)\right]$$

- 令

$$B(\vec{\delta}\mid\vec{\mathbf{w}})=\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^n\delta_i f_i(\vec{\mathbf{x}},y)\right)+1$$
$$-\sum_{\vec{\mathbf{x}}}\left[\tilde{P}(\vec{\mathbf{x}})\sum_y\left(P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})\sum_{i=1}^n\left(\frac{f_i(\vec{\mathbf{x}},y)}{f^o(\vec{\mathbf{x}},y)}\exp(\delta_i f^o(\vec{\mathbf{x}},y))\right)\right)\right]$$

则：$L(\vec{\mathbf{w}}+\vec{\delta})-L(\vec{\mathbf{w}})\ge B(\vec{\delta}\mid\vec{\mathbf{w}})$。这里 $B(\vec{\delta}\mid\vec{\mathbf{w}})$ 是对数似然函数改变量的一个新的（相对不那么紧）的下界。

7. 求 $B(\vec{\delta}\mid\vec{\mathbf{w}})$ 对 δ_i 的偏导数：

$$\frac{\partial B(\vec{\delta}\mid\vec{\mathbf{w}})}{\partial\delta_i}=\sum_{\vec{\mathbf{x}},y}[\tilde{P}(\vec{\mathbf{x}},y)f_i(\vec{\mathbf{x}},y)]-\sum_{\vec{\mathbf{x}}}\left(\tilde{P}(\vec{\mathbf{x}})\sum_y[P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})f_i(\vec{\mathbf{x}},y)\exp(\delta_i f^o(\vec{\mathbf{x}},y))]\right)=0$$

令偏导数为 0 即可得到 δ_i ：

$$\sum_{\vec{\mathbf{x}}}\left(\tilde{P}(\vec{\mathbf{x}})\sum_y[P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})f_i(\vec{\mathbf{x}},y)\exp(\delta_i f^o(\vec{\mathbf{x}},y))]\right)=\mathbb{E}_{\tilde{P}}[f_i]$$

最终根据 $\delta_1,\cdots,\delta_n$ 可以得到 $\vec{\delta}$ 。

8. `IIS` 算法：
   - 输入：
     - 特征函数 $f_1,f_2,\cdots,f_n$
     - 经验分布 $\tilde{P}(\vec{\mathbf{x}},y),\tilde{P}(\vec{\mathbf{x}})$
     - 模型 $P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})$
   - 输出：
     - 最优参数 w_i^*
     - 最优模型 $P_{\vec{\mathbf{w}}^*}(y\mid\vec{\mathbf{x}})$
   - 算法步骤：
     - 初始化：取 $w_i=0,i=1,2,\cdots,n$ 。
     - 迭代，迭代停止条件为：所有 w_i 均收敛。迭代步骤为：
       - 求解 $\vec{\delta}=(\delta_1,\cdots,\delta_n)^T$，求解方法为：对每一个 $i,i=1,2,\cdots,n$ ：

        - 求解 δ_i 。其中δ_i 是方程：$\sum_{\vec{\mathbf{x}}}\left(\tilde{P}(\vec{\mathbf{x}})\sum_y[P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})f_i(\vec{\mathbf{x}},y)\exp(\delta_i,f^o(\vec{\mathbf{x}},y))]\right)=\mathbb{E}_{\tilde{P}}[f_i]$ 的解，其中：$f^o(\vec{\mathbf{x}},y)=\sum_{i=1}^n f_i(\vec{\mathbf{x}},y)$ 。
        - 更新 $w_i\leftarrow w_i+\delta_i$ 。
    - 判定迭代停止条件。若不满足停止条件，则继续迭代。

3.2 拟牛顿法

1. 若对数似然函数 $L(\vec{\mathbf{w}})$ 最大，则 $-L(\vec{\mathbf{w}})$ 最小。

   令 $F(\vec{\mathbf{w}})=-L(\vec{\mathbf{w}})$ ，则最优化目标修改为：

   $$\min_{\vec{\mathbf{w}}\in\mathbb{R}^n}F(\vec{\mathbf{w}})=\min_{\vec{\mathbf{w}}\in\mathbb{R}^n}\sum_{\vec{\mathbf{x}}}\left(\tilde{P}(\vec{\mathbf{x}})\log\sum_y\exp\left(\sum_{i=1}^n w_i f_i(\vec{\mathbf{x}},y)\right)\right)-\sum_{\vec{\mathbf{x}},y}\left(\tilde{P}(\vec{\mathbf{x}},y)\sum_{i=1}^n w_i f_i(\vec{\mathbf{x}},y)\right)$$

   计算梯度:

   $$\vec{g}(\vec{\mathbf{w}})=\left(\frac{\partial F(\vec{\mathbf{w}})}{\partial w_1},\frac{\partial F(\vec{\mathbf{w}})}{\partial w_2},\cdots,\frac{\partial F(\vec{\mathbf{w}})}{\partial w_n}\right)^T,$$

   $$\frac{\partial F(\vec{\mathbf{w}})}{\partial w_i}=\sum_{\vec{\mathbf{x}}}[\tilde{P}(\vec{\mathbf{x}})P_{\vec{\mathbf{w}}}(y\mid\vec{\mathbf{x}})f_i(\vec{\mathbf{x}},y)]-\mathbb{E}_{\tilde{P}}[f_i],\quad i=1,2,\cdots,n$$

2. 最大熵模型学习的 `BFGS` 算法：

   - 输入：
     - 特征函数 $f_1,f_2,\cdots,f_n$
     - 经验分布 $\tilde{P}(\vec{\mathbf{x}},y),\tilde{P}(\vec{\mathbf{x}})$
     - 目标函数 $F(\vec{\mathbf{w}})$
     - 梯度 $\vec{g}(\vec{\mathbf{w}})=\nabla F(\vec{\mathbf{w}})$
     - 精度要求 ε
   - 输出：
     - 最优参数值 $\vec{\mathbf{w}}^*$
     - 最优模型 $P_{\vec{\mathbf{w}}^*}(y\mid\vec{\mathbf{w}})$
   - 算法步骤：
     - 选定初始点 $\vec{\mathbf{w}}^{<0>}$ ，取 $\mathbf{B}_0$ 为正定对阵矩阵，迭代计数器 $k=0$ 。
     - 计算 $\vec{\mathbf{g}}_k=\vec{g}(\vec{\mathbf{w}}^{<k>})$ ：
       - 若 $|\vec{\mathbf{g}}_k|<\varepsilon$ ，停止计算，得到 $\vec{\mathbf{w}}^*=\vec{\mathbf{w}}^{<k>}$
       - 若 $|\vec{\mathbf{g}}_k|\geq\varepsilon$ ：
         - 由 $\mathbf{B}_k\vec{\mathbf{p}}_k=-\vec{\mathbf{g}}_k$ 求得 $\vec{\mathbf{p}}_k$
         - 一维搜索：求出 λ_k ： $\lambda_k=\arg\min_{\lambda\geq 0}F(\vec{\mathbf{w}}^{<k>}+\lambda_k\vec{\mathbf{p}}_k)$
         - 置 $\vec{\mathbf{w}}^{<k+1>}=\vec{\mathbf{w}}^{<k>}+\lambda_k\vec{\mathbf{p}}_k$
         - 计算 $\vec{\mathbf{g}}_{k+1}=\vec{g}(\vec{\mathbf{w}}^{<k+1>})$。若 $|\vec{\mathbf{g}}_{k+1}|<\varepsilon$ ，停止计算，得到 $\vec{\mathbf{w}}^*=\vec{\mathbf{w}}^{<k+1>}$ 。
         - 否则计算 $\mathbf{B}_{k+1}$ ：

           $$\mathbf{B}_{k+1}=\mathbf{B}_k+\frac{\vec{\mathbf{y}}_k\vec{\mathbf{y}}_k^T}{\vec{\mathbf{y}}_k^T\vec{\delta}_k}-\frac{\mathbf{B}_k\vec{\delta}_k\vec{\delta}_k^T\mathbf{B}_k}{\vec{\delta}_k^T\mathbf{B}_k\vec{\delta}_k}$$

其中：$\vec{\mathbf{y}}_k = \vec{\mathbf{g}}_{k+1} - \vec{\mathbf{g}}_k$，$\vec{\delta}_k = \vec{\mathbf{w}}^{<k+1>} - \vec{\mathbf{w}}^{<k>}$ 。

- 置 $k = k + 1$ ，继续迭代。