最大熵模型

凡是知道的，就把它考虑进去，凡是不知道的，通通均匀分布。

2021-08-27

1. 最大熵原理

1.1. 原理简介

最大熵原理：学习概率模型时，在所有可能的概率模型中，熵最大的模型是最好的模型。通常用**约束条件**来确定概率模型的集合，所以，最大熵原理也可以表述为在**满足约束条件的模型集合**中选取**熵最大的模型**。

直观地说，最大熵原理认为要选择的模型需要满足：

1. 已有的事实（约束条件）：必须满足
2. 不确定的部分：等可能

如何表示"等可能"呢？我们知道，均匀分布的熵最大，反过来讲，熵最大时，数据趋向于均匀分布，即等可能。因此，可以使用**熵--可优化的数值目标**，来实现"等可能"的要求。

最大熵原理是统计学习的一般原理，将它应用到分类得到最大熵模型(Maximum Entropy Model)。

1.2. 例题

问题：假设随机变量X有5个取值A, B, C, D, E，估计取各个值的概率$P(A), P(B), P(C), P(D), P(E)$。

解　概率值需要满足如下约束条件：

$$P(A) + P(B) + P(C) + P(D) + P(E) = 1$$

满足这个约束条件的概率分布有无穷多个。按照最大熵原理，在没有其它信息的情况下，需要假定等可能分布，即：

$$P(A) = P(B) = P(C) = P(D) = P(E) = \frac{1}{5}$$

有时，能从一些先验知识中得到一些对概率值的约束条件，如：

$$P(A) + P(B) + P(C) + P(D) + P(E) = 1$$

$$P(A) + P(B) = \frac{3}{10}$$

满足这两个约束条件的概率分布依然有无穷多个。在缺少其它信息的情况下，可以认为A、B是等概率的，C、D、E是等概率的，于是：

$$P(A) = P(B) = \frac{3}{20}$$

$$P(C) = P(D) = P(E) = \frac{7}{20}$$

2. 最大熵模型

2.1. 模型定义

假设分类模型是一个条件概率分布$P(Y|X), X \in R^n, Y \in R$，$X$表示输入，$Y$表示输出。该模型表示的是对于给定的输入$X$，以条件概率$P(Y|X)$输出$Y$。

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

学习目标：使用最大熵原理选择最好的分类模型。

2.1.1. 构造模型条件

对于给定的训练数据集，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**，分别记作$\tilde{P}(X,Y)$和$\tilde{P}(X)$。

$\tilde{P}(X=x,Y=y)=\dfrac{v(X=x,Y=y)}{N}\tilde{P}(X=x)=\dfrac{v(X=x)}{N}v(X=x,Y=y)$表示训练数据中样本$(x,y)$出现的频数$v(X=x)$表示训练样本中

用特征函数$f(x,y)$描述输入x和输出y之间的某一事实，其定义为：

$$f(x,y)=\begin{cases}1, & x\text{与}y\text{ 满足某一事实}\\ 0, & \text{否则}\end{cases} \tag{54}$$

特征函数为二值函数，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：

$$\begin{aligned} E_{\tilde{P}}(f) &= \sum_{x,y}\tilde{P}(x,y)f(x,y) \\ &= \sum_{x,y}\tilde{P}(x)\tilde{P}(y|x)f(x,y) \end{aligned} \tag{55}$$

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：

$$E_P(f)=\sum_{x,y}\tilde{P}(x)P(y|x)f(x,y) \tag{56}$$

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：

$$E_{\tilde{P}}(f)=E_P(f) \tag{57}$$

我们将(7)式作为模型学习的约束条件。假设有m个特征函数$f_i(x,y),\quad i=1,2,\cdots,m$，那么就有$m$个约束条件。

2.1.2. 最大熵模型

假设满足所有约束条件的模型集合为：

$$\mathcal{C}\equiv\left\{P\in\mathcal{P}\mid E_P\left(f_i\right)=E_{\tilde{P}}\left(f_i\right),\quad i=1,2,\cdots,m\right\} \tag{58}$$

定义在模型$P(Y\mid X)$上的条件熵为：

$$H(P)=-\sum_{x,y}\tilde{P}(x)P(y\mid x)\log P(y\mid x) \tag{59}$$

则模型集合$\mathcal{C}$中条件熵最大$H(P)$的模型称为最大熵模型。

2.2. 模型学习

最大熵模型的学习可以形式化为约束最优化问题。

对于给定的训练数据集$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$以及特征函数$f_i(x,y),\quad i=1,2,\cdots,m$，最大熵模型的学习等价于约束最优化问题：

$$\begin{array}{ll} \max_{P\in\mathrm{C}} & H(P)=-\sum_{x,y}\tilde{P}(x)P(y\mid x)\log P(y\mid x) \\ \text{s.t.} & E_P\left(f_i\right)=E_{\tilde{P}}\left(f_i\right),\quad i=1,2,\cdots,m \\ & \sum_y P(y\mid x)=1 \end{array} \tag{60}$$

按照最优化问题的习惯，将求max问题改为等价的求min问题：

$$\begin{array}{ll} \min_{P\in\mathrm{C}} & -H(P)=\sum_{x,y}\tilde{P}(x)P(y\mid x)\log P(y\mid x) \\ \text{s.t.} & E_P\left(f_i\right)-E_{\tilde{P}}\left(f_i\right)=0,\quad i=1,2,\cdots,m \\ & \sum_y P(y\mid x)=1 \end{array} \tag{61}$$

对于上述原始的最优化问题，可以转为无约束最优化的对偶问题，通过求解对偶问题进而求解原始问题。

1. 构建拉格朗日函数

   引入拉格朗日乘子$w_0,w_1,w_2,\cdots,w_m$，定义拉格朗日函数$L(P,w)$：

$$\begin{aligned} L(P,w)\equiv & -H(P)+w_0\left(1-\sum_y P(y\mid x)\right)+\sum_{i=1}^m w_i\left(E_{\tilde{p}}\left(f_i\right)-E_P\left(f_i\right)\right) \\ = & \sum_{x,y}\tilde{P}(x)P(y\mid x)\log P(y\mid x) \\ & +w_0\left(1-\sum_y P(y\mid x)\right) \\ & +\sum_{i=1}^m w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x,y}\tilde{P}(x)P(y\mid x)f_i(x,y)\right) \end{aligned} \tag{62}$$

2. 定义对偶问题

   最优化的原始问题为：

$$\underset{P\in\mathcal{C}}{min}\ \underset{w}{max}\ L(P,w) \tag{63}$$

对偶问题是：

$$\underset{w}{max}\ \underset{P\in\mathcal{C}}{min}\ L(P,w) \tag{64}$$

3. 求对偶问题中的**极小化问题**

首先需要求解对偶问题内部的最小化问题$\underset{P\in\mathcal{C}}{min}\ L(P,w)$，$\underset{P\in\mathcal{C}}{min}\ L(P,w)$是$w$的函数，将其记作：

$$\Psi(w)=\min_{P\in\mathbf{C}}L(P,w)=L\left(P_w,w\right) \tag{65}$$

$\Psi(w)$称为对偶函数，同时，将其解记作：

$$P_w=\arg\min_{P\in\mathbf{C}}L(P,w)=P_w(y\mid x) \tag{66}$$

具体地，为了求$\Psi(w)$的最小值， 求$L(p,w)$对$P(y\mid x)$的偏导数：

$$\begin{aligned}\frac{\partial L(P,w)}{\partial P(y\mid x)}&=\sum_{x,y}\tilde{P}(x)(\log P(y\mid x)+1)-\sum_{y}w_0-\sum_{x,y}\left(\tilde{P}(x)\sum_{i=1}^{m}w_if_i(x,y)\right)\\&=\sum_{x,y}\tilde{P}(x)\left(\log P(y\mid x)+1-w_0-\sum_{i=1}^{m}w_if_i(x,y)\right)\end{aligned} \tag{67}$$

令偏导数等于0，在$\tilde{P}(x)>0$的情况下，解得：

$$P(y\mid x)=\exp\left(\sum_{i=1}^{m}w_if_i(x,y)+w_0-1\right)=\frac{\exp\left(\sum_{i=1}^{m}w_if_i(x,y)\right)}{\exp\left(1-w_0\right)} \tag{68}$$

由于$\sum_y P(y\mid x)=1$,两边求和得：

$$\begin{gathered}P_w(y\mid x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{m}w_if_i(x,y)\right)\\ \text{其中，}Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{m}w_if_i(x,y)\right)\\ Z_w(x)\text{称为规范化因子，}f_i(x,y)\text{是特征函数，}w_i\text{是特征的权值}\end{gathered} \tag{69}$$

$P_w=P_w(y\mid x)$就是最大熵模型，w是最大熵模型中的参数向量。

4. 求对偶问题中的**极大化问题**

求解对偶问题外部的极大化问题：

$$\max_{w}\Psi(w) \tag{70}$$

将其解记为w^*，即：

$$w^*=\arg\max_{w}\Psi(w) \tag{71}$$

因此， 可以使用最优化问题求解对偶函数$\Psi(w)$的极大化，得到w^*，从而可以得到最优化模型P^*，这里，$P^*=P_{w^*}=P_{w^*}(y\mid x)$是学习到的最优化模型（最大熵模型）。

对上述总结，**最大熵模型的一般形式**为：

$$\begin{aligned}&P_w(y\mid x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{m}w_if_i(x,y)\right)\\&\text{其中，}Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{m}w_if_i(x,y)\right)\end{aligned}\quad \text{这里，}x\in\mathbf{R}^n\text{为输入，}y\in\{1,2,\cdots,K\}\text{为输出，}w\in\mathbf{R}^n\text{为权值向量，}f_i(x,y),i=1,2,\cdots,m\text{为任意实}$$

2.3. 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$P(Y\mid X)$的对数似然函数表示为：

$$L_{\tilde{P}}\left(P_w\right)=\log\prod_{x,y}P(y\mid x)^{\tilde{P}(x,y)}=\sum_{x,y}\tilde{P}(x,y)\log P(y\mid x) \tag{73}$$

当条件概率分布$P(y\mid x)$是最大熵模型的解（公式17）时，对数似然函数$L_{\tilde{P}}\left(P_w\right)$为：

$$\begin{aligned}L_{\tilde{P}}\left(P_w\right)&=\sum_{x,y}\tilde{P}(x,y)\log P(y\mid x)\\&=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{m}w_if_i(x,y)-\sum_{x,y}\tilde{P}(x,y)\log Z_w(x)\\&=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{m}w_if_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x)\end{aligned} \tag{74}$$

根据公式17，对偶函数$\Psi(w)$可化简为：

$$\begin{aligned}\Psi(w)=&\sum_{x,y}\tilde{P}(x)P_w(y\mid x)\log P_w(y\mid x)+\\&\sum_{i=1}^{m}w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x,y}\tilde{P}(x)P_w(y\mid x)f_i(x,y)\right)\\=&\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{m}w_if_i(x,y)+\sum_{x,y}\tilde{P}(x)P_w(y\mid x)\left(\log P_w(y\mid x)-\sum_{i=1}^{m}w_if_i(x,y)\right)\end{aligned} \tag{75}$$

$$
\begin{aligned}
&= \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} w_i f_i(x,y) - \sum_{x,y} \tilde{P}(x) P_w(y \mid x) \log Z_w(x) \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} w_i f_i(x,y) - \sum_{x} \tilde{P}(x) \log Z_w(x)
\end{aligned}
$$

比较式(21)和式(22)，可得：

$$
\Psi(w) = L_{\tilde{P}}(P_w) \tag{76}
$$

因此，对偶函数$\Psi(w)$等价于对数似然函数$L_{\tilde{P}}(P_w)$。因此，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。**

2.4. 模型学习的最优化算法

已知最大熵模型为：

$$
\begin{aligned}
P_w(y \mid x) &= \frac{1}{Z_w(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(x,y)\right) \\
\text{其中，} Z_w(x) &= \sum_{y} \exp\left(\sum_{i=1}^{m} w_i f_i(x,y)\right)
\end{aligned} \tag{77}
$$

最大熵模型的对数似然函数为：

$$
L(w) = \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} w_i f_i(x,y) - \sum_{x} \tilde{P}(x) \log Z_w(x) \tag{78}
$$

目标：通过极大似然法估计模型参数，即求使得对数函数达到极大的w^*。

2.4.1. 改进的迭代尺度法

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。

IIS的思想：

1. 初始化最大熵模型的参数向量：$w = (w_1, w_2, \cdots, w_n)^T$。
2. 找到新的参数向量：$w + \delta = (w_1 + \delta_1, w_2 + \delta_2, \cdots, w_n + \delta_n)$，使得模型的对数似然函数值增大。
3. 重复$\text{步骤}2$，迭代更新参数$w = w + \delta$，直到满足退出条件。

推导过程：

1. 寻找下界函数1

   对于给定的经验分布$\tilde{P}(x,y)$，模型参数从w增加到$w+\delta$，对数似然函数的改变量是：

$$
\begin{aligned}
L(w+\delta) - L(w) &= \sum_{x,y} \tilde{P}(x,y) \log P_{w+\delta}(y \mid x) - \sum_{x,y} \tilde{P}(x,y) \log P_w(y \mid x) \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} \delta_i f_i(x,y) - \sum_{x} \tilde{P}(x) \log \frac{Z_{w+\delta}(x)}{Z_w(x)}
\end{aligned} \tag{79}
$$

   利用不等式：$-\log \alpha \geq 1 - \alpha, \quad \alpha > 0$

   建立对数似然函数改变量的下界：

$$
\begin{aligned}
L(w+\delta) - L(w) &\geqslant \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} \delta_i f_i(x,y) + 1 - \sum_{x} \tilde{P}(x) \frac{Z_{w+\delta}(x)}{Z_w(x)} \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} \delta_i f_i(x,y) + 1 - \sum_{x} \tilde{P}(x) \sum_{y} P_w(y \mid x) \exp \sum_{i=1}^{m} \delta_i f_i(x,y)
\end{aligned} \tag{80}
$$

   将右端记为：

$$
A(\delta \mid w) = \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} \delta_i f_i(x,y) + 1 - \sum_{x} \tilde{P}(x) \sum_{y} P_w(y \mid x) \exp \sum_{i=1}^{m} \delta_i f_i(x,y) \tag{81}
$$

   于是有：

$$
L(w+\delta) - L(w) \geq A(\delta \mid w) \tag{82}
$$

   即$A(\delta \mid w)$是对数似然函数改变量的一个下界。

   然而，若对$A(\delta \mid w)$求δ_i的导数，由于$g = \exp \sum_{i=1}^{m} \delta_i f_i(x,y) = \prod_{i=1}^{m} \exp\left(\delta_i f_i(x,y)\right)$，那么$\frac{\partial g}{\partial \delta_i} = \prod_{i=1}^{m} \exp\left(\delta_i f_i(x,y)\right)$的结果耦合了多个$\delta_i$，导致不易优化。为了能够继续优化，IIS试图一次只优化其中一个变量δ_i，而固定其它变量δ_j，$i \neq j$。

2. 寻找下界函数2

   引入变量$f^{\#}(x,y) = \sum_{i=1}^{m} f_i(x,y)$。因为$f_i$是二值函数，故$f^{\#}(x,y)$表示所有特征在$(x,y)$出现的次数。

   将$A(\delta \mid w)$进行改写：

$$
A(\delta \mid w) = \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{m} \delta_i f_i(x,y) + 1 - \sum_{x} \tilde{P}(x) \sum_{y} P_w(y \mid x) \exp\left(f^{\#}(x,y) \sum_{i=1}^{m} \frac{\delta_i f_i(x,y)}{f^{\#}(x,y)}\right) \tag{83}
$$

   利用指数函数的凸性以及对任意i，有$\frac{f_i(x,y)}{f^{\#}(x,y)} \geq 0 \text{且} \sum_{i=1}^{m} \frac{f_i(x,y)}{f^{\#}(x,y)} = 1$，根据Jensen不等式，得到：

$$\exp\left(\sum_{i=1}^{m}\frac{f_i(x,y)}{f^{\#}(x,y)}\delta_i f^{\#}(x,y)\right)\leqslant\sum_{i=1}^{m}\frac{f_i(x,y)}{f^{\#}(x,y)}\exp\left(\delta_i f^{\#}(x,y)\right) \tag{84}$$

于是：

$$\begin{aligned}A(\delta\mid w)\geqslant&\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{m}\delta_i f_i(x,y)+1-\\&\sum_{x}\tilde{P}(x)\sum_{y}P_w(y\mid x)\sum_{i=1}^{m}\left(\frac{f_i(x,y)}{f^{\#}(x,y)}\right)\exp\left(\delta_i f^{\#}(x,y)\right)\end{aligned} \tag{85}$$

记不等式右侧为：

$$\begin{aligned}B(\delta\mid w)=&\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{m}\delta_i f_i(x,y)+1-\\&\sum_{x}\tilde{P}(x)\sum_{y}P_w(y\mid x)\sum_{i=1}^{m}\left(\frac{f_i(x,y)}{f^{\#}(x,y)}\right)\exp\left(\delta_i f^{\#}(x,y)\right)\end{aligned} \tag{86}$$

于是得到：

$$L(w+\delta)-L(w)\geq B(\delta\mid w) \tag{87}$$

这里，$B(\delta\mid w)$为对数似然函数改变量的一个新的下界。

求$B(\delta\mid w)$对δ_i的偏导数：

$$\frac{\partial B(\delta\mid w)}{\partial\delta_i}=\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x}\tilde{P}(x)\sum_{y}P_w(y\mid x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right) \tag{88}$$

在上式中，除δ_i外不含任何其它变量。令偏导数为0得到：

$$\sum_{x,y}\tilde{P}(x)P_w(y\mid x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right)=E_{\tilde{P}}\left(f_i\right) \tag{89}$$

于是，根据上式可依次求解δ_i，求得δ后，可进一步得到w，从而可以重复迭代过程。

基于上述推导，给出改进的迭代尺度法IIS。

算法1.1(改进的迭代尺度法**IIS**)

输入：特征函数$f_1,f_2,\cdots,f_m$; 经验分布$\tilde{P}(X,Y)$; 模型$P_w(y\mid x)$

输出：最优参数值w_i^*; 最优模型P_{w^*}

算法步骤：

1. 对所有的$i\in\{1,2,\cdots,n\}$，取初值$w_i=0$
2. 对每一个$i\in\{1,2,\cdots,n\}$
   1. 令δ_i是方程$\sum_{x,y}\tilde{P}(x)P_w(y\mid x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right)=E_{\tilde{P}}\left(f_i\right)$的解。

      其中，$f^{\#}(x,y)=\sum_{i=1}^{m}f_i(x,y)$
   2. 更新w_i的值：$w_i\leftarrow w_i+\delta_i$
3. 如果不是所有的w_i都收敛，重复步骤(2)

算法的核心是求解δ_i。分以下情况进行讨论：

1. $f^{\#}(x,y)$是常数

   即对任何x,y，有$f^{\#}(x,y)=M$，那么δ_i可以显式地表示为：

$$\delta_i=\frac{1}{M}\log\frac{E_{\tilde{P}}\left(f_i\right)}{E_P\left(f_i\right)} \tag{90}$$

2. $f^{\#}(x,y)$不是常数

   必须通过数值法求解δ_i，简单有效的方法是牛顿法。令$g(\delta_i)=\sum_{x,y}\tilde{P}(x)P_w(y\mid x)f_i(x,y)\exp\left(\delta_i f^{\#}(x,y)\right)=E_{\tilde{P}}\left(f_i\right)$，牛顿法通过迭代求得$\delta_i^*$，使得$g(\delta_i^*)=0$，迭代公式是：

$$\delta_i^{(k+1)}=\delta_i^{(k)}-\frac{g\left(\delta_i^{(k)}\right)}{g'\left(\delta_i^{(k)}\right)} \tag{91}$$

   只要适当选取初始值$\delta_i^{(0)}$，由于$g(\delta_i)$有单根，因此牛顿法恒收敛，而且收敛速度很快。

2.4.2. 拟牛顿法

最大熵模型：

$$P_w(y\mid x)=\frac{\exp\left(\sum_{i=1}^{m}w_i f_i(x,y)\right)}{\sum_y\exp\left(\sum_{i=1}^{m}w_i f_i(x,y)\right)} \tag{92}$$

目标函数：

$$\min_{w\in\mathbf{R}^m} f(w)=\sum_x \tilde{P}(x)\log\sum_y \exp\left(\sum_{i=1} w_i f_i(x,y)\right)-\sum_{x,y}\tilde{P}(x,y)\sum_{i=1} w_i f_i(x,y) \tag{93}$$

梯度:

$$g(w)=\left(\frac{\partial f(w)}{\partial w_1},\frac{\partial f(w)}{\partial w_2},\cdots,\frac{\partial f(w)}{\partial w_m}\right)^{\mathrm{T}} \tag{94}$$

其中，梯度具体为:

$$\frac{\partial f(w)}{\partial w_i}=\sum_{x,y}\tilde{P}(x)P_w(y\mid x)f_i(x,y)-E_{\tilde{P}}(f_i),\quad i=1,2,\cdots,m \tag{95}$$

相应的拟牛顿法BFGS算法如下。

算法1.2(最大熵模型学习的$\mathbf{BFGS}$算法)

输入：特征函数$f_1,f_2,\cdots,f_m$; 经验分布$\tilde{P}(X,Y)$; 目标函数$f(w)$; 梯度$g(w)=\nabla f(w)$, 精度要求 ε

输出：最优参数值w^*; 最优模型$P_{w^*}(y\mid x)$

算法步骤：

1. 选定初始点$w^{(0)}$，取B_0为正定对称矩阵，置$k=0$
2. 计算$g_k=g(w^{(k)})$。若 $\|g_k\|<\varepsilon$, 则停止计算, 得 $w^*=w^{(k)}$; 否则转 (3);
3. 由 $B_k p_k=-g_k$ 求出 p_k;
   一维搜索：求λ_k使得：
4. $f\left(w^{(k)}+\lambda_k p_k\right)=\min_{\lambda\geqslant 0} f\left(w^{(k)}+\lambda p_k\right)$
5. 置 $w^{(k+1)}=w^{(k)}+\lambda_k p_k$;
   计算$g_{k+1}=g\left(w^{(k+1)}\right)$，若$\|g_{k+1}\|<\varepsilon$，则停止计算，得$w^*=w^{(k+1)}$；否则，按下式求出$B_{k+1}$：
6. $$B_{k+1}=B_k+\frac{y_k y_k^{\mathrm{T}}}{y_k^{\mathrm{T}}\delta_k}-\frac{B_k\delta_k\delta_k^{\mathrm{T}}B_k}{\delta_k^{\mathrm{T}}B_k\delta_k},$$
   其中,
   $$y_k=g_{k+1}-g_k,\quad \delta_k=w^{(k+1)}-w^{(k)}$$
7. 置$k=k+1$，转(3)。

2.5. 例题：最大熵模型学习

问题：假设随机变量X有5个取值A,B,C,D,E，已知$P(A)+P(B)=\frac{3}{10}$，估计取各个值的概率$P(A),P(B),P(C),P(D),P(E)$。

解　为了方便，分别以y_1,y_2,y_3,y_4,y_5表示A,B,C,D,E。于是，最大熵模型学习的最优化问题是:

$$\begin{array}{ll}\min & -H(P)=\sum_{i=1}^5 P(y_i)\log P(y_i)\\ \text{s.t.} & P(y_1)+P(y_2)=\tilde{P}(y_1)+\tilde{P}(y_2)=\frac{3}{10}\\ & \sum_{i=1}^5 P(y_i)=\sum_{i=1}^5\tilde{P}(y_i)=1\end{array} \tag{96}$$

引入拉格朗日乘子w_0、w_1，定义拉格朗日函数:

$$L(P,w)=\sum_{i=1}^5 P(y_i)\log P(y_i)+w_1\left(P(y_1)+P(y_2)-\frac{3}{10}\right)+w_0\left(\sum_{i=1}^5 P(y_i)-1\right) \tag{97}$$

根据拉格朗日对偶性，可以通过求解对偶最优化问题得到原始最优化问题的解，因此，接下来求解:

$$\max_w \min_P L(P,w) \tag{98}$$

首先求解$L(P,w)$关于P的极小化问题。为此，固定w_0、w_1，求偏导数:

$$\begin{aligned}\frac{\partial L(P,w)}{\partial P(y_1)}&=1+\log P(y_1)+w_1+w_0\\ \frac{\partial L(P,w)}{\partial P(y_2)}&=1+\log P(y_2)+w_1+w_0\\ \frac{\partial L(P,w)}{\partial P(y_3)}&=1+\log P(y_3)+w_0\\ \frac{\partial L(P,w)}{\partial P(y_4)}&=1+\log P(y_4)+w_0\\ \frac{\partial L(P,w)}{\partial P(y_5)}&=1+\log P(y_5)+w_0\end{aligned} \tag{99}$$

令各偏导数等于0，解得:

$$\begin{aligned}&P(y_1)=P(y_2)=\mathrm{e}^{-w_1-w_0-1}\\ &P(y_3)=P(y_4)=P(y_5)=\mathrm{e}^{-w_0-1}\end{aligned} \tag{100}$$

于是:

$$\min_P L(P,w)=L(P_w,w)=-2\mathrm{e}^{-w_1-w_0-1}-3\mathrm{e}^{-w_0-1}-\frac{3}{10}w_1-w_0 \tag{101}$$

再求解$L(P_w, w)$关于w的极大化问题：

$$\max_w L\left(P_w, w\right) = -2\mathrm{e}^{-w_1-w_0-1} - 3\mathrm{e}^{-w_0-1} - \frac{3}{10} w_1 - w_0 \tag{102}$$

分别求$L(P_w, w)$对w_0、w_1的偏导数并令其为0，得到：

$$\begin{aligned} \mathrm{e}^{-w_1-w_0-1} &= \frac{3}{20} \\ \mathrm{e}^{-w_0-1} &= \frac{7}{30} \end{aligned} \tag{103}$$

于是，得到所要求的概率分布为：

$$\begin{aligned} P\left(y_1\right) = P\left(y_2\right) &= \frac{3}{20} \\ P\left(y_3\right) = P\left(y_4\right) = P\left(y_5\right) &= \frac{7}{30} \end{aligned} \tag{104}$$

3. 参考文档

1. (52条消息) 最大熵模型中的对数似然函数的解释_wkebj的博客-CSDN博客
2. 为什么最大熵模型的极大似然估计中带有指数？ - 知乎 (zhihu.com)
3. 最大熵模型中的对数似然函数表示法解释 - 知乎 (zhihu.com)
4. (52条消息) 最大熵模型中的数学推导_结构之法 算法之道-CSDN博客