最大熵模型

覃一发

2021 年 9 月 13 日

摘要

本文介绍了最大熵原理的出发点、理论推导、算法伪代码以及样例演示结果。

1 最大熵原理

所谓最大熵，指的是在估计目标概率分布时，在满足约束条件前提下选取熵最大者。

2 最大熵模型

X 和 Y 是输入、输出集合。最大熵模型表示的是对于给定的输入 $x \in X$，输出条件概率 $p(y|x)$, $y \in Y$。

给定数据集 $T = \{(x_1, y_1), (x_1, y_1), ..., (x_N, y_N)\}$。

联合分布 $p(x, y)$ 的经验分布, 边缘分布 $P(x)$ 的经验分布如下。$v(x)$ 为数据集里输入量取值为 x 的样本个数，$v(x, y)$ 同理。

$$\begin{aligned} \tilde{P}(x, y) &= \frac{v(x, y)}{N} \\ \tilde{P}(x) &= \frac{v(x)}{N} \end{aligned} \tag{1}$$

特征函数是以 x 和 y 为输入量的实值函数，用于对模型施加约束。该类函数通常输出 0 或 1，但也可以是任意实数值。

$$f(x, y) = \begin{cases} 1, if \quad x, y \quad satisfy \quad some \quad constrain \\ 0, else \end{cases} \tag{2}$$

特征函数 $f(x,y)$ 关于经验分布,$\tilde{P}(x,y)$ 的期望值如下。

$$E_{\tilde{P}}[f] = \sum_{x,y} \tilde{P}(x,y) f(x,y) \tag{3}$$

特征函数 $f(x,y)$ 关于模型 $P(y|x)$ 和经验分布 $\tilde{P}(x)$ 的期望值如下。

$$E_P[f] = \sum_{x,y} \tilde{P}(x) P(y|x) f(x,y) \tag{4}$$

如果模型 $P(y|x)$ 能够反映出训练数据中的分布规律，那么对于单个特征函数 $f(x,y)$，$E_{\tilde{P}}[f] = E_P[f]$。

$$\sum_{x,y} \tilde{P}(x,y) f(x,y) = \sum_{x,y} \tilde{P}(x) P(y|x) f(x,y) \tag{5}$$

如果有 n 个特征函数，那么就有 n 条如同公式 5的约束。

定义（最大熵模型）假设满足所有约束条件的模型集合为

$$\mathcal{C} \equiv \{P \in \mathcal{P} | E_{\tilde{P}}[f] = E_P[f]\} \tag{6}$$

定义在条件概率分布 $P(y|x)$ 上的条件熵如下。

$$H(P) \equiv -\sum_{x,y} \tilde{P}(x,y) \ln P(y|x) \tag{7}$$

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

3 最大熵模型的学习

最大熵模型的学习过程即求解使条件熵最大的条件概率模型。可以形式化为最优化问题。使熵最大等价于负熵最小。最小化的形式是为了符合最优化理论的习惯。

$$\min_{P \in \mathcal{C}} -H(P) = -\sum_{x,y} P(y|x) \ln P(y|x) \tag{8}$$

$$\begin{aligned} s.t. \quad & E_{\tilde{P}}[f_i] = E_P[f_i] \\ & \sum_x \tilde{P}(x) \sum_y P(y|x) = 1 \end{aligned} \tag{9}$$

求解带约束的最小化问题，可以引进拉格朗日乘子 $\omega_0, \omega_1, ..., \omega_n$，定义拉格朗日函数 $L(P,\omega)$。这里 ω_0 的约束写法和李航的《统计学习方法》不

太一样，因为其实 $\sum_y P(y|x) = 1$ 不是 1 条约束而是多条约束，有多少个可取的 x 就有多少条约束，所以不能用一个权重分配给多条约束。

$$\begin{aligned} L(P,\omega) \equiv & -H(P) + \omega_0(1 - \sum_x \tilde{P}(x) \sum_y P(y|x)) + \sum_{i=1}^{n} \omega_i (E_{\tilde{P}}[f_i] - E_P[f_i]) \\ = & \sum_{x,y} \tilde{P}(x) P(y|x) \ln P(y|x) + \omega_0 (1 - \sum_x \tilde{P}(x) \sum_y P(y|x)) + \\ & \sum_{i=1}^{n} \omega_i (\sum_{x,y} \tilde{P}(x,y) f_i(x,y) - \sum_{x,y} \tilde{P}(x) P(y|x) f_i(x,y)) \end{aligned} \tag{10}$$

最优化的原始问题是 Eq.11。

$$\min_{P \in \mathcal{C}} \max_{\omega} L(P,\omega) \tag{11}$$

$\min\max$ 的形式可以保证不遵守约束的候选 $L(P,\omega)$ 会导致 $\max_\omega L(P,\omega)$ 中直接变成 $+\infty$,而若遵守约束 $\max_\omega L(P,\omega)$ 仍为有限值,从而在 $\min_{P\in\mathcal{C}} \max_\omega L(P,\omega)$ 中被“选中”。

对偶问题如 Eq.12。

$$\max_{\omega} \min_{P \in \mathcal{C}} L(P,\omega) \tag{12}$$

由于 $L(P,\omega)$ 是 P 的凸函数，原始问题 11的解等价于对偶问题公式 12的解。所以可以通过求解对偶问题来得到原始问题的解。

先求解对偶问题内部的极小化问题 $\min_{P\in\mathcal{C}} L(P,\omega)$，相当于先固定住 ω 而变化 $P(y|x)$ 使得 $L(P,\omega)$ 最小。求 $\frac{\partial L(P,\omega)}{\partial P}$，令其为零。

$$\begin{aligned} \frac{\partial L(P,\omega)}{\partial P} = & \tilde{P}(x)(1 + \log P(y|x)) - \omega_0 \tilde{P}(x) - \\ & \sum_{i=1}^{n} \omega_i \tilde{P}(x) f_i(x,y) \\ = & \tilde{P}(x)(1 + \log P(y|x) - \omega_0 - \sum_{i=1}^{n} \omega_i f_i(x,y)) = \quad 0 \end{aligned} \tag{13}$$

显然，由于 $\tilde{P}(x) > 0$，下边必然成立。

$$1 + \log P(y|x) - \omega_0 - \sum_{i=1}^{n} \omega_i f_i(x,y) = 0 \tag{14}$$

消掉对数符号，得到 $P(y|x)$。

$$P(y|x) = \frac{\exp(\sum_{i=1}^{n} \omega_i f_i(x,y))}{\exp(1-\omega_0)} \tag{15}$$

到目前为止，我们在约束中其实并没有充分体现 $P(y|x)$ 是概率这一事实。既然是概率那么可以通过边际概率公式这“隐藏条件”来缩小求解空间。

$$\sum_y P(y|x) = \frac{\sum_y \exp(\sum_{i=1}^{n} \omega_i f_i(x,y))}{\exp(1-\omega_0)} = 1 \tag{16}$$

容易得到下式。

$$\exp(1-\omega_0) = \sum_y \exp(\sum_{i=1}^{n} \omega_i f_i(x,y)) \tag{17}$$

那么使 $L(P,\omega)$ 取极小值的解如下。

$$\begin{aligned} P_\omega(y|x) &= \frac{\exp(\sum_{i=1}^{n} \omega_i f_i(x,y))}{Z_\omega} \\ Z_\omega &= \sum_y \exp(\sum_{i=1}^{n} \omega_i f_i(x,y)) \end{aligned} \tag{18}$$

我们把 $min_{P\in\mathcal{C}}L(P,\omega)$ 记为 $\Psi(\omega)$

$$
\begin{aligned}
\Psi(\omega) =& \sum_{x,y}\tilde{P}(x)P_\omega(y|x)\log P_\omega(y|x) + \omega_0(\sum_x \tilde{P}(x)\sum_y P(y|x) - 1) + \\
& \sum_{i=1}^{n}\omega_i(\sum_{x,y}\tilde{P}(x,y)f_i(x,y) - \sum_{x,y}\tilde{P}(x)P_\omega(y|x)f_i(x,y)) \\
=& \sum_{x,y}\tilde{P}(x)P_\omega(y|x)\log P_\omega(y|x) + 0 + \\
& \sum_{i=1}^{n}\omega_i(\sum_{x,y}\tilde{P}(x,y)f_i(x,y) - \sum_{x,y}\tilde{P}(x)P_\omega(y|x)f_i(x,y)) \\
=& \sum_{x,y}\tilde{P}(x)P_\omega(y|x)(\sum_{i=1}^{n}\omega_i f_i(x,y) - \log Z_\omega) + \\
& \sum_{i=1}^{n}\omega_i(\sum_{x,y}\tilde{P}(x,y)f_i(x,y) - \sum_{x,y}\tilde{P}(x)P_\omega(y|x)f_i(x,y)) \\
=& \sum_{x,y}\tilde{P}(x)P_\omega(y|x)(-\log Z_\omega) + \sum_{i=1}^{n}\omega_i(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)) \\
=& \sum_{x,y}\tilde{P}(x,y)\sum_i \omega_i f_i(x,y) - \sum_{x,y}\tilde{P}(x)P_\omega(y|x)\log Z_\omega \\
=& \sum_{x,y}\tilde{P}(x,y)\sum_i \omega_i f_i(x,y) - \sum_x \tilde{P}(x)\log Z_\omega
\end{aligned}
\tag{19}
$$

由于在求解 P_ω 过程中我们已经用掉了 $\sum_y P(y|x) = 1$ 这一隐藏条件，并且 $\sum_x \tilde{P}(x) = 1$，所以公式 19中 ω_0 项直接为 0。上式第二步利用了 $\log P_\omega(y|x) = \sum_{i=1}^{n}\omega_i f_i(x,y) - \log Z_\omega$ 。第三步有抵消项。第六步同样利用了 $\sum_y P_\omega(y|x) = 1$。

那么求解最大熵模型，现在就剩下寻找一组最优的 ω 使得 $\Psi(\omega)$ 最大。

4 最大熵与最大似然

最大熵模型具有和逻辑回归模型类似的形式，它们都属于广义线性模型。当条件概率模型采用公式 20的形式时，利用最大似然策略得到的模型与最大熵模型是一致的。训练集经验分布为 $\tilde{P}(x,y)$。

$$
P(y|x) = \frac{\exp\left(\sum_i \omega_i f_i(x,y)\right)}{Z\omega} \tag{20}
$$

那么对数似然函数展开后如 Eq.21所示，与 Eq.19一致。

$$
\begin{aligned}
LogLikelihood(P(y|x)) &= \log \prod_{x,y} P(y|x)^{\tilde{P}(x,y)} \\
&= \sum_{x,y} \tilde{P}(x,y) \log P(y|x) \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{i} \omega_i f_i(x,y) - \sum_{x,y} \tilde{P}(x,y) \log Z_\omega \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{i} \omega_i f_i(x,y) - \sum_{x} \tilde{P}(x) \sum_{y} P(y|x) \log Z_\omega \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{i} \omega_i f_i(x,y) - \sum_{x} \tilde{P}(x) \log Z_\omega
\end{aligned}
\tag{21}
$$

求一组 ω 使得对数似然函数最大化实质上和是对 $\Psi(\omega)$ 最大化一样的。其背后的意义是最大熵模型寻找满足约束条件下，最平均最等可能最公正的模型，但同时也是在寻找与数据最像的逻辑斯特回归模型。

5 最大熵模型具体求解

最大熵模型求解可以通过梯度下降、牛顿法、拟牛顿法 (BFP, BFGS) 等进行求解，目标函数为 $\Psi(\omega)$。也可以使用改进迭代尺度法 (Improved Iterative Scaling, IIS) 进行。本节主要介绍 IIS 方法。

5.1 改进的迭代尺度法 IIS

IIS 的思路是根据现有权重向量 $\omega = \{\omega_1, ..., \omega_n\}$，找到一个新的修正向量 $\omega + \delta = \{\omega_1 + \delta_1, ..., \omega_n + \delta_n\}$，使得模型的对数似然函数增大，重复以上步骤直至找到对数似然最大值。

那么修正向量带来的似然函数增量 ΔL 如下。

$$
\begin{aligned}
\Delta L =& L(P_\omega, \omega + \delta) - L(P_\omega, \omega) \\
=& \sum_{x,y} \tilde{P}(x,y) \sum_{i} (\omega_i + \delta_i) f_i(x,y) - \sum_{x} \tilde{P}(x) \log Z_{\omega+\delta} - \\
& \sum_{x,y} \tilde{P}(x,y) \sum_{i} \delta_i f_i(x,y) + \sum_{x} \tilde{P}(x) \log Z_\omega \\
=& \sum_{x,y} \tilde{P}(x,y) \sum_{i} \delta_i f_i(x,y) - \sum_{x} \tilde{P}(x) \log \frac{Z_{\omega+\delta}}{Z_\omega}
\end{aligned}
\tag{22}
$$

观察 Eq.22，难点显然集中在第二项。该项实质上是对数的线性组合，那么可考虑相关不等式，例如不等式 Eq.23。

$$-\log\alpha > 1-\alpha \tag{23}$$

那么继续化简 ΔL。

$$\Delta L > \sum_{x,y}\tilde{P}(x,y)\sum_i \delta_i f_i(x,y) + \sum_x \tilde{P}(x)(1-\frac{Z_{\omega+\delta}}{Z_\omega}) \tag{24}$$

先化简 $\frac{Z_{\omega+\delta}}{Z_\omega}$。

$$\begin{aligned}\frac{Z_{\omega+\delta}}{Z_\omega} =& \frac{\sum_y \exp\sum_i(\omega_i+\delta_i)f_i(x,y)}{\sum_y \exp\sum_i \omega_i f_i(x,y)} \\ =& \sum_y \frac{\exp\sum_i(\omega_i+\delta_i)f_i(x,y)}{\sum_y \exp\sum_i \omega_i f_i(x,y)} \\ =& \sum_y \frac{\exp\sum_i \omega_i f_i(x,y)}{\sum_y \exp\sum_i \omega_i f_i(x,y)} \exp\sum_i \delta_i f_i(x,y) \\ =& \sum_y P_\omega(y|x)\exp\sum_i \delta_i f_i(x,y)\end{aligned} \tag{25}$$

回到 ΔL。

$$\begin{aligned}\Delta L &> \sum_{x,y}\tilde{P}(x,y)\sum_i \delta_i f_i(x,y) + \sum_x \tilde{P}(x)(1-\frac{Z_{\omega+\delta}}{Z_\omega}) \\ &= \sum_{x,y}\tilde{P}(x,y)\sum_i \delta_i f_i(x,y) + \sum_x \tilde{P}(x)(1-\sum_y P_\omega(y|x)\exp\sum_i \delta_i f_i(x,y)) \\ &= \sum_{x,y}\tilde{P}(x,y)\sum_i \delta_i f_i(x,y) + 1 - \sum_x \tilde{P}(x)\sum_y P_\omega(y|x)\exp\sum_i \delta_i f_i(x,y)\end{aligned} \tag{26}$$

这里我们得到了 ΔL 的下界，记为 $A(\delta|\omega)$。

$$A(\delta|\omega) = \sum_{x,y}\tilde{P}(x,y)\sum_i \delta_i f_i(x,y) + 1 - \sum_x \tilde{P}(x)\sum_y P_\omega(y|x)\exp\sum_i \delta_i f_i(x,y) \tag{27}$$

似乎到这里再一次卡住了，能否再用一次不等式呢？记 $f^{\#}(x,y) = \sum_i f_i(x,y)$，意义是点 (x,y) 在所有特征函数中的响应程度（有多少特征函数搭理它）。代入 Eq.26，得到下边式子。

$$\begin{aligned}
&\Delta L > A(\delta|\omega) \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_i \delta_i f_i(x,y) + 1 - \sum_x \tilde{P}(x) \sum_y P_\omega(y|x) \exp\left(\sum_i \frac{f_i(x,y)}{f^{\#}(x,y)} \delta_i f^{\#}(x,y)\right) \\
&> \sum_{x,y} \tilde{P}(x,y) \sum_i \delta_i f_i(x,y) + 1 - \sum_x \tilde{P}(x) \sum_y P_\omega(y|x) \sum_i \left(\frac{f_i(x,y)}{f^{\#}(x,y)} \exp\left(\delta_i f^{\#}(x,y)\right)\right)
\end{aligned} \tag{28}$$

得到了下界的下界，记为 $B(\delta|\omega)$。为了使其最大化，我们让 $B(\delta|\omega)$ 对 δ_i 求导，使导数为 0，得到 Eq.30。

$$B(\delta|\omega) = \sum_{x,y} \tilde{P}(x,y) \sum_i \delta_i f_i(x,y) + 1 - \sum_x \tilde{P}(x) \sum_y P_\omega(y|x) \sum_i \left(\frac{f_i(x,y)}{f^{\#}(x,y)} \exp\left(\delta_i f^{\#}(x,y)\right)\right) \tag{29}$$

$$\begin{aligned}
\frac{\partial B(\delta|\omega)}{\partial \delta_i} =& \sum_{x,y} \tilde{P}(x,y) f_i(x,y) - \sum_{x,y} \tilde{P}(x) P_\omega(y|x) f_i(x,y) \exp\left(\delta_i f^{\#}(x,y)\right) \\
=& 0
\end{aligned} \tag{30}$$

如果对于任意 (x,y) 都有 $f^{\#}(x,y) = C$，即同样的常数，那么公式 Eq.30实际上可以写成如下。该公式的意义是权重参数的改变量正比与经验分布于模型分布之比的对数，即如果某特征函数的模型分布期望值低于经验分布，那么增大其权重参数。11

$$\begin{aligned}
&E_{\tilde{P}}[f_i] = E_P[f_i] exp(\delta_i M) \\
&\delta_i = \frac{1}{C} \log \frac{E_{\tilde{P}}[f_i]}{E_P[f_i]}
\end{aligned} \tag{31}$$

如果该条件不成立，那么可以使用牛顿法来求解 Eq.30。记 $g(\delta_i)$ 为新下界 $B(\delta|\omega)$ 的梯度，那么 δ_i 的改变量乘以现有梯度的梯度后应该能够和现有梯度大小相同，方向相反。这正是牛顿法的思路–自变量应该朝着梯度

消失的方向去增加。

$$\delta_i^{(k+1)} - \delta_i^{(k)} = -\frac{g(\delta_i^k)}{g'(\delta_i^k)} \tag{32}$$

IIS 通过最大化参数微调后与微调前的增量，来推动模型参数的估计；增量不方便最大化，则最大化其下界，甚至下界的下界。这一通过下界、上界来实现优化目的，把困难问题用更简单问题进行近似的思路，非常值得反复体会。