Natural Language Process

最大熵模型

原理及例子

目录

1 什么是最大熵原理

- “不要将鸡蛋放到一个篮子里”，可以降低风险。这里面就包含了最大熵原理。
- 最大熵原理是概率模型学习的一个准则。最大熵原理认为，在学习概率模型时，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。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。

1 什么是最大熵原理

- 例子1：假设随机变量X有5个取值{A,B,C,D,E}，要估计各个值的概率P(A),P(B),…,P(E).
- 这些概率值满足条件**P(A)+P(B)+P(C)+P(D)+P(E)=1**
- 但是满足这个条件的概率分布有无数个。如果没有其他信息，一个可行的办法就是认为他们的概率都相等，均为0.2。
- 如果再加一个条件**P(A) + P(B) = 0.3**，那么各个值的概率为多少?

2 数学知识-拉格朗日乘子法

问题: 求函数 $z = f(\mathbf{x})$ 在条件 $\phi_i(\mathbf{x}) = 0$ $(i = 1, 2, \cdots, m)$ 下的可能极值点, 其中 $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$.

利用 Lagrange 乘子法, 可将上述**带约束的极值问题**转化为**无约束极值问题**来进行求解, 具体求解步骤如下:

1. 构造函数 $L(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \phi_i(\mathbf{x})$, 其中 λ_i $(i = 1, 2, \cdots, m)$ 为 **Lagrange 乘子**.

2. 求解方程组

$$\begin{cases} \frac{\partial L}{\partial \mathbf{x}} = \mathbf{0}, \\ \phi_i(\mathbf{x}) = 0, \quad i = 1, 2, \cdots, m \end{cases} \tag{1.1}$$

其中 $\frac{\partial L}{\partial \mathbf{x}} = (\frac{\partial L}{\partial x_1}, \frac{\partial L}{\partial x_2}, \cdots, \frac{\partial L}{\partial x_n})^T$ 表示 L 关于 $\mathbf{x}$ 的梯度.

2 数学知识-Bayes定理

- Bayes定理用来描述两个条件概率之间的关系。若计P(A)和P(B)分别表示事件A和事件B发生的概率，P(A|B)表示事件B发生的情况下事件A发生的概率，P(A,B)表示事件A和B同时发生的概率，则有：

$$P(A|B) = \frac{P(A,B)}{P(B)}, \tag{1.2}$$

$$P(B|A) = \frac{P(A,B)}{P(A)}, \tag{1.3}$$

- 利用(1.2)和(1.3)可以进一步得到贝叶斯公式：

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}, \tag{1.4}$$

2 数学知识-熵

- 熵(entropy)是热力学中的概念，由香浓引入到信息论中。在信息论和概率统计中，熵用来表示随机变量不确定性的度量。

定义 1.1 设 $X \in \{x_1, x_2, \cdots, x_n\}$ 为一个离散随机变量，其概率分布为 $p(X = x_i) = p_i$, $i = 1, 2, \cdots, n$, 则 X 的熵为

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i, \tag{1.5}$$

其中，当 $p_i = 0$ 时，定义 $0 \log 0 = 0$.

- H(x)依赖于X的分布，而与X的具体值无关。H(X)越大，表示X的不确定性越大。

2 数学知识-条件熵

定义 1.2 设 $X \in \{x_1, x_2, \cdots, x_n\}$, $Y \in \{y_1, y_2, \cdots, y_m\}$ 为离散随机变量. 在已知 X 的条件下, Y 的**条件熵** *(conditional entropy)* 可定义为

$$H(Y|X) = \sum_{i=1}^{n} p(x_i) H(Y|X = x_i) = -\sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j|x_i) \log p(y_j|x_i), \tag{1.8}$$

它表示已知 X 的条件下, Y 的条件概率分布的熵对 X 的数学期望.

3 最大熵模型的定义

- 最大熵原理是统计学习的一般原理，将它应用到分类就得到了最大熵模型
- 假设分类模型是一个条件概率分布P(Y|X)，X表示输入，Y表示输出。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，我们的目标就是利用最大熵原理选择最好的分类模型。

$$\mathcal{T} = \{(x_1, y_1), (x_2, y_2), \cdots (x_N, y_N)\}, \tag{3.1}$$

- 按照最大熵原理，我们应该优先保证模型满足已知的所有约束。那么如何得到这些约束呢?
- 思路是：从训练数据T中抽取若干特征，然后要求这些特征在T上关于经验分布的期望与它们在模型中关于p(x,y)的数学期望相等，这样，一个特征就对应一个约束。

3.1 特征函数

- 特征选取是机器学习中一个重要的问题。特征通常用特征函数来表示，对于一个给定的样本(x,y)，特征函数可以定义为任意的实值函数。它表示输入x和输出y之间的某一事实，其定义为:

$$f(x,y) = \begin{cases} 1, & \text{若 } x,y \text{ 满足某个事实;} \\ 0, & \text{否则.} \end{cases} \tag{3.2}$$

3.1 特征函数

例 3.1 假设我们需要判断“打”字是动词还是量词，已知的训练数据有

$$(x_1, y_1) = (\text{一打火柴}, \text{量词}),$$
$$(x_2, y_2) = (\text{三打啤酒}, \text{量词}),$$
$$(x_3, y_3) = (\text{五打塑料袋}, \text{量词}),$$
$$(x_4, y_4) = (\text{打电话}, \text{动词}),$$
$$(x_5, y_5) = (\text{打篮球}, \text{动词}),$$
$$\cdots$$

通过观察，我们发现，“打”前面为数字时，“打”是量词，“打”后面为名词时，“打”是动词，这就是从训练数据中提取的两个特征，可分别用特征函数表示为

$$f_1(x, y) = \begin{cases} 1, & \text{若“打”前面为数字;} \\ 0, & \text{否则.} \end{cases}$$

$$f_2(x, y) = \begin{cases} 1, & \text{若“打”后面为名词;} \\ 0, & \text{否则.} \end{cases}$$

定义了这两个特征函数后，对于训练数据，我们便有

$$f_1(x_1, y_1) = f_1(x_2, y_2) = f_1(x_3, y_3) = 1; f_1(x_4, y_4) = f_1(x_5, y_5) = 0; \cdots$$
$$f_2(x_1, y_1) = f_2(x_2, y_2) = f_2(x_3, y_3) = 0; f_2(x_4, y_4) = f_2(x_5, y_5) = 1; \cdots$$

3.2 经验分布

- 经验分布是指通过训练数据T上进行统计得到的分布。我们需要考察两个经验分布，分别是x，y的联合经验分布以及x的分布。其定义如下：

$$\tilde{p}(x,y) = \frac{count(x,y)}{N}, \quad \tilde{p}(x) = \frac{count(x)}{N}, \tag{3.3}$$

- (3.3)中count(x,y)表示(x,y)在数据T中出现的次数，count(x)表示x在数据T中出现的次数。

3.3 约束条件

- 对于任意的特征函数f，记$E_{\tilde{p}}(f)$表示f在训练数据T上关于$\tilde{p}(x,y)$的数学期望。$E_p(f)$表示f在模型上关于p(x,y)的数学期望。按照期望的定义，有：

$$E_{\tilde{p}}(f) = \sum_{x,y} \tilde{p}(x,y) f(x,y), \tag{3.4}$$

$$E_p(f) = \sum_{x,y} p(x,y) f(x,y), \tag{3.5}$$

- 我们需要注意的是公式(3.5)中的p(x,y)是未知的。并且我们建模的目标是p(y|x)，因此我们利用Bayes定理得到p(x,y)=p(x)p(y|x)。此时，p(x)也还是未知，我们可以使用经验分布对p(x)进行近似。

$$E_p(f) = \sum_{x,y} \tilde{p}(x) p(y|x) f(x,y). \tag{3.6}$$

3.3 约束条件

- 对于概率分布p(y|x)，我们希望特征f的期望应该和从训练数据中得到的特征期望是一样的。因此，可以提出约束:

$$E_p(f) = E_{\tilde{p}}(f), \tag{3.7}$$

$$\sum_{x,y} \tilde{p}(x)p(y|x)f(x,y) = \sum_{x,y} \tilde{p}(x,y)f(x,y). \tag{3.8}$$

- 假设从训练数据中抽取了n个特征，相应的便有n个特征函数以及n个约束条件。

$$\mathcal{C}_i : E_p(f_i) = E_{\tilde{p}}(f_i) := \tau_i, \ i = 1, 2, \cdots, n. \tag{3.9}$$

3.4 最大熵模型

- 给定数据集T，我们的目标就是根据最大熵原理选择一个最优的分类器。
- 已知特征函数和约束条件，我们将熵的概念应用到条件分布上面去。我们采用条件熵。

$$H(p(y|x)) = -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x)...(3.11)$$

- 至此，我们可以给出最大熵模型的完整描述了。对于给定的数据集T，特征函数$f_i(x,y),i=1,\ldots,n$，最大熵模型就是求解模型集合C中条件熵最大的模型：

$$\min_{p\in C} -H(p) = (\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x)),....(3.12)$$

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

$$\sum_{y} p(y|x) = 1.....(3.13)$$

4 最大熵模型的学习

- 最大熵模型的学习过程就是求解最大熵模型的过程。求解约束最优化问题(3.12)，(3.13)所得的解就是最大熵模型学习的解。思路如下：

- 利用拉格朗日乘子法将最大熵模型由一个带约束的最优化问题转化为一个与之等价的无约束的最优化问题，它是一个min max问题。

- 利用对偶问题的等价性，将原始问题转换为一个max min问题。

4.1 原始问题和对偶问题

- 利用拉格朗日乘子法定义关于(3.7)、(3.12)和(3.13)的拉格朗日函数如下：

$$
\begin{aligned}
L(p,\lambda) &= -H(p)+\lambda_0\left(1-\sum_y p(y|x)\right)+\sum_{i=1}^{n}\lambda_i(\tau_i-E_p(f_i)) \\
&= \sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x)+\lambda_0\left(1-\sum_y p(y|x)\right) \\
&\quad +\sum_{i=1}^{n}\lambda_i\left(\tau_i-\sum_{x,y}\tilde{p}(x)p(y|x)f_i(x,y)\right) \qquad (4.1)
\end{aligned}
$$

- 利用拉格朗日对偶性，(3.6)、(3.12)和(3.13)定义的最大熵模型等价于求解：

$$
\min_{p\in\mathcal{C}}\max_{\lambda} L(p,\lambda), \qquad (4.2)
$$

4.1 原始问题和对偶问题

- 通过交换极大和极小的位置，可以得到公式(4.2)的对偶问题：

$$\max_{\lambda} \min_{p \in \mathcal{C}} L(p, \lambda). \tag{4.3}$$

- 经过两次等价转换，求解最大熵模型，就是求解对偶问题(4.3)就可以了。

4.2 极小问题求解

- 对偶问题(4.3)内部的极小问题是关于参数lamba的问题

$$\Psi(\lambda) = \min_{p \in \mathcal{C}} L(p, \lambda) = L(p_\lambda, \lambda), \tag{4.4}$$

$$p_\lambda = \operatorname{argmin}_{p \in \mathcal{C}} L(p, \lambda), \tag{4.5}$$

- 我们可以利用拉格朗日乘子法获取p。

4.2 极小问题求解

- 首先计算拉格朗日函数L对p(y|x)的偏导数。

$$
\begin{aligned}
&\frac{\partial L(p,\lambda)}{\partial p(y|x)} \\
=\ & \sum_{x,y}\tilde{p}(x)(\log p(y|x)+1)-\sum_{y}\lambda_0-\sum_{i=1}^{n}\lambda_i\left(\sum_{x,y}\tilde{p}(x)f_i(x,y)\right) \\
=\ & \sum_{x,y}\tilde{p}(x)(\log p(y|x)+1)-\sum_{x}\tilde{p}(x)\sum_{y}\lambda_0-\sum_{x,y}\tilde{p}(x)\sum_{i=1}^{n}\lambda_i f_i(x,y) \quad (\text{利用}\sum_{x}\tilde{p}(x)=1) \\
=\ & \sum_{x,y}\tilde{p}(x)(\log p(y|x)+1)-\sum_{x,y}\tilde{p}(x)\lambda_0-\sum_{x,y}\tilde{p}(x)\sum_{i=1}^{n}\lambda_i f_i(x,y) \\
=\ & \sum_{x,y}\tilde{p}(x)\left(\log p(y|x)+1-\lambda_0-\sum_{i=1}^{n}\lambda_i f_i(x,y)\right) \quad (\text{提出}\sum_{x,y}\tilde{p}(x))
\end{aligned}
$$

4.2 极小问题求解

- 令上面的公式等于0，可以得到：

$$\log p(y|x) + 1 - \lambda_0 - \sum_{i=1}^{n} \lambda_i f_i(x, y) = 0,$$

- 进一步可以解得：

$$p(y|x) = e^{\lambda_0 - 1} \cdot e^{\sum_{i=1}^{n} \lambda_i f_i(x,y)}, \tag{4.6}$$

- 将上面的公式带入(3.13)，可以得到

$$\sum_{y} p(y|x) = e^{\lambda_0 - 1} \cdot \sum_{y} e^{\sum_{i=1}^{n} \lambda_i f_i(x,y)} = 1,$$

4.2 极小问题求解

- 进一步可得：

$$e^{\lambda_0 - 1} = \frac{1}{\sum_y e^{\sum_{i=1}^{n} \lambda_i f_i(x,y)}}. \tag{4.7}$$

- 将(4.7)带回(4.6)，可以得到：

$$p_\lambda = \frac{1}{Z_\lambda(x)} e^{\sum_{i=1}^{n} \lambda_i f_i(x,y)}, \tag{4.8}$$

$$Z_\lambda(x) = \sum_y e^{\sum_{i=1}^{n} \lambda_i f_i(x,y)} \tag{4.9}$$

- (4.9)称为规范化因子。(4.8)中的p是最大熵模型的解，可以看到他具有指数的形式。

4.3 最大似然估计

- 得到对偶问题(4.3)内部的极小问题的解p之后，需要进一步求解外层的极大值问题。

$$\max_{\lambda} \Psi(\lambda), \tag{4.10}$$

$$\lambda^* = \operatorname*{argmax}_{\lambda} \Psi(\lambda), \tag{4.11}$$

$$\begin{aligned}
& \Psi(\lambda) \\
= \; & L(p_\lambda, \lambda) \quad \text{(由(4.4)式)} \\
= \; & \sum_{x,y} \tilde{p}(x) p_\lambda(y|x) \log p_\lambda(y|x) + \sum_{i=1}^{n} \lambda_i \left(\tau_i - \sum_{x,y} \tilde{p}(x) p_\lambda(y|x) f_i(x,y) \right) \quad \text{(由(4.1)式)} \\
= \; & \sum_{i=1}^{n} \lambda_i \tau_i + \sum_{x,y} \tilde{p}(x) p_\lambda(y|x) \left(\log p_\lambda(y|x) - \sum_{i=1}^{n} \lambda_i f_i(x,y) \right)
\end{aligned}$$

4.3 最大似然估计

- 由(4.8)可以得到：

$$\log p_\lambda(y|x) = \sum_{i=1}^{n} \lambda_i f_i(x,y) - \log Z_\lambda(x) \tag{4.12}$$

- 将(4.12)带入到Ψ中，可以得到：

$$\begin{aligned}
\Psi(\lambda) &= \sum_{i=1}^{n} \lambda_i \tau_i - \sum_{x,y} \tilde{p}(x) p_\lambda(y|x) \log Z_\lambda(x) \\
&= \sum_{i=1}^{n} \lambda_i \tau_i - \sum_{x} \tilde{p}(x) \log Z_\lambda(x) \sum_{y} p_\lambda(y|x) \\
&= \sum_{i=1}^{n} \lambda_i \tau_i - \sum_{x} \tilde{p}(x) \log Z_\lambda(x) \quad (\text{利用} \sum_{y} p_\lambda(y|x) = 1)
\end{aligned} \tag{4.13}$$

4.4 例子

- 题：假设随机变量X有5个取值{A,B,C,D,E}，且满足条件P(A)+P(B)=0.3且P(A)+P(B)+P(C)+P(D)+P(E)=1。求最大熵模型。
- 为了方便，分别用y_1~y_5表示A~E于是最大熵模型的最优化问题是：

$$\min -H(p)=\sum_{i=1}^{5} p(y_i)\log p(y_i)$$

$$s.t. p(y_1)+p(y_2)=\tilde{p}(y_1)+\tilde{p}(y_2)=\frac{3}{10}$$

$$\sum_{i=1}^{5} p(y_i)=\sum_{i=1}^{5}\tilde{p}(y_i)=1$$

4.4 例子

- 引进拉格朗日乘子w0和w1，定义拉格朗日函数如下：

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

- 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。所以求解max min L(p,w)首先需要求解关于p的极小化问题。为此需要固定w0和w1。求偏导数：

$$\frac{\partial L(p,w)}{\partial p(y_1)}=1+\log p(y_1)+w_1+w_0$$

$$\frac{\partial L(p,w)}{\partial p(y_2)}=1+\log p(y_2)+w_1+w_0$$

$$\frac{\partial L(p,w)}{\partial p(y_3)}=1+\log p(y_3)+w_0$$

$$\frac{\partial L(p,w)}{\partial p(y_4)}=1+\log p(y_4)+w_0$$

$$\frac{\partial L(p,w)}{\partial p(y_5)}=1+\log p(y_5)+w_0$$

4.4 例子

- 令各个偏导数为0，可以得到：

$$p(y_1) = p(y_2) = e^{-w_1 - w_0 - 1}$$
$$p(y_3) = p(y_4) = p(y_5) = e^{-w_0 - 1}$$

so:

$$\min_p L(p,w) = L(p_w,w) = -2e^{-w_1-w_0-1} - 3e^{-w_0-1} - \frac{3}{10}w_1 - w_0$$

- 再求L(p,w)关于w的极大化问题：

$$\max_w L(p_w,w) = -2e^{-w_1-w_0-1} - 3e^{-w_0-1} - \frac{3}{10}w_1 - w_0$$

4.4 例子

- 分别对w0和w1求偏导，并令其等于0，可以得到

$$e^{-w_1-w_0-1}=\frac{3}{20}$$

$$e^{-w_0-1}=\frac{7}{30}$$

$$so:$$

$$p(y_1)=p(y_2)=\frac{3}{20}$$

$$p(y_3)=p(y_4)=p(y_5)=\frac{7}{30}$$

5 最优化算法

- 公式(4.11)没有显式的解析解，因此需要借助于其他的方法。由于目标函数是一个 凸函数，所以可以借助多种优化方法来进行求解，并且能保证得到全局最优解。
- 为最大熵模型量身定制的两个最优化方法分别是通用迭代尺度法(GIS)和改进的迭代尺度法(IIS)。

5.1 GIS算法

算法 5.1 (GIS)

Step 1 初始化参数. 令 $\lambda := \mathbf{0}$.

Step 2 计算 $E_{\tilde{p}}(f_i)$, $i = 1, 2, \cdots, n$.

Step 3 执行一次迭代, 对参数做一次刷新.

计算 $E_{p_\lambda}(f_i)$, $i = 1, 2, \cdots, n$.

FOR $i = 1, 2, \cdots, n$ **DO**

{

$$\lambda_i := \lambda_i + \eta \log \frac{E_{\tilde{p}}(f_i)}{E_{p_\lambda}(f_i)}$$

}

Step 4 检查收敛条件, 若达到收敛条件则算法结束; 否则转至 *Step 3*.

5.2 IIS算法

算法 5.2 (IIS)

Step 1 初始化参数. 令 $\lambda := 0$.

Step 2 执行一次迭代, 对参数做一次刷新.

FOR $i = 1, 2, \cdots, n$ **DO**

{

2.1 求解方程

$$\sum_{x,y} \tilde{p}(x) p(y|x) f_i(x,y) e^{\Delta\lambda_i \sum_{i=1}^{n} f_i(x,y)} = \tilde{p}(f_i)$$

得到 $\Delta\lambda_i$.

2.2 令 $\lambda_i := \lambda_i + \Delta\lambda_i$.

}

Step 3 检查收敛条件, 若达到收敛条件则算法结束; 否则转至 *Step 2*.

参考资料

- 李航. 统计学习方法[M]. 北京：清华大学出版社，2012
- 吴军. 数学之美[M]. 北京：人民邮电出版社，2012
- 最大熵学习笔记
- 关于最大熵模型的严重困惑：为什么没有解析解?
- 最大熵-IIS（Improved Iterative Scaling）训练算法的Java实现
- 如何理解最大熵模型里面的特征?