

最大熵模型中的对数似然函数表示法解释



苍狼

人工智能学习

5 人赞同了该文章

书中最大熵模型的极大似然估计这部分中，条件概率分布的对数似然函数表示为

$$L_{\tilde{P}}(P_w) = \log \prod_{x,y} P(y|x)^{\tilde{P}(x,y)} = \sum_{x,y} \tilde{P}(x,y) \log P(y|x) \tag{1}$$

感觉似然函数应该表示为：

$$L_{\tilde{P}}(P_w) = \log \prod_{x,y} P(y|x) \tag{2}$$

但是式（1）中似然函数出现了指数形式。

其实公式（2）是错误的。似然函数定义为样本集中各个样本的联合概率，最大熵模型的训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ，因此，似然函数应该为：

$$L_{\tilde{P}}(P_w) = \log \prod_{i=1}^n P(x_i, y_i) \tag{3}$$

博文《最大熵模型中的对数似然函数的解释》给出了解释，推导过程如下：

训练数据集 T 中包含 n 个样本， (x_i, y_i) 为其中一个样本。样本集中会有很多值相同的样本，即 $(x_i, y_i) = (x_j, y_j) (i \neq j)$ 。假设样本的取值为 k 个，分别为 $\{(v_1, w_1), (v_2, w_2), \dots, (v_k, w_k)\}$ 。用 $C[(X, Y) = (v_i, w_i)]$ 表示训练集中样本值为 (v_i, w_i) 的频数，则公式（3）可以表示为：

$$L_{\tilde{P}}(P_w) = \log \prod_{i=1}^k P(v_i, w_i)^{C[(X,Y)=(v_i,w_i)]} \tag{4}$$

等号两边同时开 n 次方，可得

$$L_{\tilde{P}}(P_w)^{\frac{1}{n}} = \log \prod_{i=1}^k P(v_i, w_i)^{\frac{C[(X,Y)=(v_i,w_i)]}{n}} \tag{5}$$

而经验概率分布 $\tilde{P}(X = v_i, Y = w_i) = \frac{C[(X,Y) = (v_i, w_i)]}{n}$ ，所以公式（5）可表示为

$$L_{\tilde{P}}(P_w)^{\frac{1}{n}} = \log \prod_{i=1}^k P(v_i, w_i)^{\tilde{P}(v_i, w_i)} \tag{6}$$

博文《最大熵模型中的对数似然函数的解释》中描述说将公式（6）简化写成

$$L_{\tilde{P}}(P_w)^{\frac{1}{n}} = \log \prod_x P(x, y)^{\tilde{P}(x,y)} \tag{7}$$

这个地方有待商榷，因为公式（7）中连乘是针对 n 个样本，6 式中连乘是针对 k 个样本值，一般 $k < n$ ，显然两式的结果不一样，公式 7 的结果会小于 6 的结果。

浙江工业大学 教师 Jerry 在回答知乎提问《为什么最大熵模型的极大似然估计中带有指数》时，提出引用交叉熵的概念解释进行解释，解释如下：

交叉熵 (CrossEntropy) 用来度量真实分布 $P(x, y)$ 与经验分布 $\tilde{P}(x, y)$ 在 (X, Y) 的差异性，我们希望经验分布于真实分布之间的差异最小化，即**交叉熵最小**。

$$\begin{aligned} CrossEntropy &= -\log \prod_{x,y} p(x, y)^{\tilde{P}(x,y)} \\ &= -\sum_{x,y} \hat{p}(x, y) \log p(x, y) \\ &= -\sum_{x,y} \hat{p}(x, y) \log [\hat{p}(x) p(y|x)] \\ &= -\sum_{x,y} \hat{p}(x, y) \log p(y|x) + \sum_{x,y} \hat{p}(x, y) \log \hat{p}(x) \end{aligned} \tag{8}$$

由于公式（8）结果中的第 2 项为常数，所以极小化 **CrossEntropy** 等价于极大化 $\sum_{x,y} \tilde{P}(x, y) \log P(y|x)$ ，即公式（1）中对数似然函数，这个就是书中为什么出现指数型似然函数的原因。

显然这个解释更有说服力。


参考文献

- 【1】最大熵模型中的对数似然函数的解释
- 【2】为什么最大熵模型的极大似然估计中带有指数

发布于 2019-09-24

统计学 概率论 熵

文章被以下专栏收录



统计学习方法笔记

推荐阅读



信息熵数值计算

Dreisteine



最大熵模型与最大似然估计

南京汤包没有馅



信息论 (自信息 信息熵 联合熵 条件熵 互信息)

1、约定数学表达形式大写 X \quad Y \quad Z 表示随机变量小写 x \quad y \quad z 表示随机变量具体取值花体 \mathcal{X} \quad \mathcal{Y} \quad \mathcal{Z} 表集合集合的势，即集合中...
xods



最小二乘法和最大似然法异同

请参考以下链接：
http://blog.sina.com.cn/s/blog_4t
http://blog.csdn.net/luo86106/arti
对于最小二乘法，当从模型总体随机抽取 n 组样本观测值后...
傲慢偏见

5 条评论

切换为时间排序

写下你的评论...



efresmeat

03-20

(6)跟（7）是一个意思吧，（7）中连乘下的x代表x的k种可能取值，而不是样本中的n个x，所以（7）中也是k个值相乘

赞

长行 回复 efresmeat

06-15

是这样的，(7)连乘下写成x,y似乎表达得更清楚

赞

刃枫

03-08

MaxEntropy里的交叉熵为什么和KL里的交叉熵不一样

赞

jayy

2020-11-11

为什么是针对联合概率的最大似然，而不是对后验概率的最大似然

赞

Z N

2020-04-07

我觉得第一种说法是对的，我在看条件随机场的迭代尺度法也是带了指数，说明应该在更广义的方面解释，而且连乘的时候重复的都合在一起了，体现在了指数上

赞