深度解读条件随机场*

Yang Fu

December 6, 2022

Abstract

CRF 是概率图模型中极具代表性的概率无向图模型，常用于命名实体识别等序列标注任务。CRF 模型的表达式是对数线性函数，参数包括特征函数及其权重。训练时，基于前向-后向算法计算正则化后的概率值，然后基于极大似然估计训练模型参数；推理时，基于维特比算法计算给定观测序列下条件概率最大的标签序列。CRF 的实现版本有两种，传统机器学习实现版本支持人工显式地定义特征模板，深度学习实现版本在所有位置上共享特征函数及其权重。相比 Softmax，CRF 接受标签转移约束，基于全局归一化预测，彻底解决了标签不一致的问题，缺点是归一化因子计算量较大，同时不能很好处理嵌套和不连续实体。

1 理论部分

1.1 模型表达式

给定序列 x，CRF 预测标签序列 y 的条件概率 $P(y|x)$ 为：

$$P(y|x) = \frac{1}{Z(x)} exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\Big) \tag{1}$$

$$Z(x) = \sum_{y} exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\Big) \tag{2}$$

- $Z(x)$ 是归一化因子，使得 $\sum_y p(y|x) = 1$，实现了全局归一化；
- exp() 的角色是势函数；
- t_k 和 s_l 分别是状态转移特征函数和状态特征函数，取值为 0 或 1，λ_k 和 μ_l 是对应特征函数的权重；

接下来从概率无向图模型开始，逐步推导 CRF 模型表达式。

*在自然语言处理领域，由于文本具有序列性，实际只讨论条件随机场（Conditional Random Field, CRF）的一个特例，即线性链条件随机场（Linear Chain CRF）。

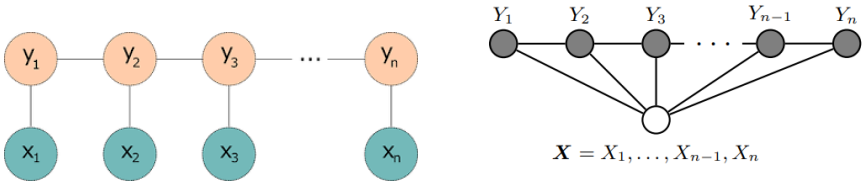
1.2 推导过程

概率无向图模型又称马尔可夫随机场，它要求 $P(Y)$ 必须满足成对、局部、全局马尔可夫性，然后根据 Hammersley-Clifford 定理，概率无向图模型的联合概率分布 $P(Y)$ 可以分解成规范化的最大团的势函数乘积[1]。

在马尔可夫随机场的基础上引入随机变量 X，Y 保持为马尔可夫随机场，则给定 X 条件下 Y 的条件概率 $P(Y|X)$ 即表示条件随机场。特殊地，如果 Y 具有链式结构，便得到线性链条件随机场。由于 Y 仍旧满足马尔可夫性，有 $P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_v, w \sim v)$，其中 $\sim$ 表示相邻。则线性链条件随机场的条件概率可表示为：

$$P(Y_i|X, Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1}) \tag{3}$$

Fig 1 展示了两种形式的线性链条件随机场。由于线性链条件随机场的最大团就是相邻两个结点的集合。而通过对这两张图进行内部比较，可以发现左边的图更为特殊，因为 X 和 Y 具有相同的结构。现实中，比如对于 NER 问题，对一个文本序列进行序列标注，都一般假设 X 和 Y 具有相同结构，满足左边的那张图。



(a) X 和 Y 都具有链式结构　　(b) 仅 Y 具有链式结构

Figure 1: CRF 架构图

基于两种形式的最大团，CRF 分别定义状态转移特征函数 $t_k(y_{i-1}, y_i, x, i)$ 和状态特征函数 $s_l(y_i, x, i)$，并给它们分配了权重。这两种特征函数分别描述了状态转移概率和发射概率。以特征函数 t_1 为例：

$$t1(y_{i-1} = 1, y_i = 2, x, i) = 1, \quad i = 2, 3 \tag{4}$$

特征函数 t_1 的含义是当 $i = 2$ 或 3 时，如果 $y_{i-1} = 1, y_i = 2$ 成立，特征函数 t_1 取 1，否则 t_1 取 0.

在 Hammersley-Clifford 定理中，势函数必须是非负函数，CRF 采用 exp 作为势函数，最大团的势函数的乘积就变成指数上的累加。

于是，便可得到线性链条件随机场的模型表达式：

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)) \tag{5}$$

[1]这里暂不深究 Hammersley-Clifford 定理，只需明确概率无向图模型的联合概率是由图中最大团的势函数相乘得到的。关于最大团的定义可以查阅图论。

这里有一个细节值得注意：i 和 k/l 分别是位置编号和特征函数编号，它们是分开的！换句话说，每个位置上都存在多种特征函数，而一个特征函数又作用在整个序列 y 上。

1.3 CRF 的矩阵表达形式

将两种特征函数统一为 $f_k(y,x)$，得到上一节模型表达式的简化表达：

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{k=1}^{K} w_k f_k(y,x)) \tag{6}$$

$$f_k(y,x) = \sum_{i=1}^{L} f_k(y_{i-1}, y_i, x, i) \tag{7}$$

可以看到，$f_k(y,x)$ 是单个特征函数在整个序列上的取值，$P(y|x)$ 是所有特征函数与权重的加权和。这里的特征函数有 K 个，是两种特征函数的数量和，位置则是从 1 到 L。

接下来，尝试将其转化成向量形式[2]：

$$P_w(y|x) = \frac{exp(w \cdot F(y,x))}{Z_w(x)} \tag{8}$$

$$w = (w_1, w_2, ..., w_K)^T \tag{9}$$

$$F(y,x) = (f_1(y,x), f_2(y,x), ..., f_K(y,x))^T \tag{10}$$

将表达式继续简化为矩阵表示形式。假设标签集合为 $\{v_1, v_2, ..., v_c\}$，在位置 i 处定义一个 c 阶矩阵随机变量：

$$M_i(x) = [M_i(y_{i-1}, y_i, x)]$$
$$= \begin{bmatrix} M_i(y_{i-1}=v_1, y_i=v_1|x) & M_i(y_{i-1}=v_1, y_i=v_2|x) & \cdots & M_i(y_{i-1}=v_1, y_i=v_c|x) \\ M_i(y_{i-1}=v_2, y_i=v_1|x) & M_i(y_{i-1}=v_2, y_i=v_2|x) & \cdots & M_i(y_{i-1}=v_2, y_i=v_c|x) \\ \vdots & \vdots & \ddots & \vdots \\ M_i(y_{i-1}=v_m, y_i=v_1|x) & M_i(y_{i-1}=v_c, y_i=v_2|x) & \cdots & M_i(y_{i-1}=v_m, y_i=v_c|x) \end{bmatrix}$$

其中

$$M_i(y_{i-1}, y_i, x) = exp(W_i(y_{i-1}, y_i|x)) \tag{11}$$

$$W_i(y_{i-1}, y_i|x) = \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, x, i) \tag{12}$$

[2]这里的 $\cdot$ 表示内积。

如果考虑起始和终止标签,标签序列 $y = \{y_0, y_1, ..., y_{L+1}\} = \{start, y_1, y_2, ..., y_L, stop\}$ 的条件概率为:

$$P_w(y|x) = \frac{1}{Z_w(x)} \prod_{i=1}^{L+1} M_i(y_{i-1}, y_i|x) \tag{13}$$

这里的分子就是当前标签序列 y 的预测分数。所有可能的标签序列共有 L^C 种，归一化因子 $Z_w(x)$ 是以 start 为起点，以 stop 为终点通过状态的所有路径的非规范化概率 $\prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|x)$ 之和。

综上，考虑 start 和 stop 的 CRF 模型的参数量是 $(L+1) * C * C$，但是实际上，tensorflow 和 pytorch 实现代码中，$M_i(x)$ 在所有位置上共享，因而参数量为 $C * C$.

2 训练

3 推导

4 实现

References