

条件随机场

条件随机场 (conditional random field, CRF) 是给定一组随机变量条件下另一组输出随机变量的条件概率分布模型。条件随机场可以用于不同的预测问题，本文主要参考《统计学习方法》条件随机场，主要说明用于标注问题的线性链(linear chain)条件随机场。

概率无向图模型

概率无向图模型(probabilistic undirected graphical model)，又称马尔科夫随机场(Markov random field)，是一个可以由无向图表示的联合概率分布。

模型定义

概率图模型：由图 $G = (V, E)$ 表示的概率分布，图中节点 $v \in V$ 表示一个随机变量 Y_v ，边 $e \in E$ 表示随机变量之间的概率依赖关系。如果联合概率分布满足马尔科夫性，称此联合概率分布为概率无向图模型，或马尔科夫随机场。

《统计学习方法》条件随机场解释了三种马尔科夫性，成对，局部，全局马尔科夫性，这里就不赘述，核心思想为如果图中两个节点（或节点的集合）A,B之间没有直接的边相连，那么A,B关于另外的节点（或节点的集合）C条件独立，数学表示为： $P(Y_A, Y_B | Y_C) = P(Y_A | Y_C)P(Y_B | Y_C)$

概率无向图模型的因子分解

因子（势函数）：定义在变量子集上的非负实函数，主要用于定义概率分布函数。

团(clique)：无向图中任何两个节点均有边连接的节点子集。

最大团：若C是无向图中的一个团，并且不能再加入任何一个节点使其成为一个更大的团，则称此C为最大团(maximal clique)

将概率无向图模型的联合概率分布表示为其最大团上的随机变量的函数的乘积形式的操作，称为因子分解（联合概率分布定义）（根据马尔科夫假设，不存在直接相连，就没有关系）

定义无向图模型G，C为G上的最大团， Y_C 表示对应的随机变量，无向图的联合概率分布可写做最大团C上的函数 $\Phi_C(Y_C)$ ，则有：

$$P(Y) = \frac{1}{Z} \prod_C \Phi_C(Y_C)$$

Z为规范化因子， $Z = \sum_Y \prod_C \Phi_C(Y_C)$ ，保证Y构成一个概率分布。 $\Phi_C(Y_C)$ 称为势函数。要求必须为正的，因此，通常写成指数形式：

势函数

势函数 $\phi_Q(x_Q)$ 定量刻画变量集 x_Q 中变量之间的相关关系，它应该是非负的，在所编号的变量取值上有较大的函数值。为了满足非负性，指数函数常被用于定义势函数，即：

$$\Phi_C(x_C) = e^{-H_C(x_C)}$$

$H_C(x_C)$ 是一个定义在变量 x_C 上的实值函数，常见形式为：

$$H_C(x_C) = \sum_{u,v \in C, u \neq v} \alpha_{u,v} x_u, x_v + \sum_{v \in C} \beta_v x_v$$

其中 $\alpha_{u,v}, \beta_v$ 为参数。

条件随机场的定义与形式

条件随机场是在给定随机变量X条件下，随机变量Y的马尔科夫场，这里主要介绍线性链条件随机场，可以用于标注问题。

条件概率模型 $P(Y|X)$, Y是输出变量，表示标记序列，X是输入变量，表示需要标注的观察序列，标记序列也称为状态序列。利用训练数据集通过极大似然估计和正则化极大似然估计得到条件概率模型 $\hat{P}(Y|X)$

条件随机场：若随机变量Y构成一个由无向图 $G = (V, E)$ 表示的马尔科夫随机场，即：

$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w - v)$, v 只与与其直接相连的 w 有关，与其他变量无关。

对任意节点 v 都成立，则称条件概率分布 $P(Y|X)$ 为条件随机场。

线性链条件随机场：设 $X = (X_1, X_2 \dots X_n)$, $Y = (Y_1, Y_2 \dots Y_n)$ 均为线性链表示的随机变量序列，若给定随机变量序列X的条件下，随机变量序列Y的条件概率分布构成条件随机场，即：

$P(Y_i|X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1}), i = 1, 2, 3 \dots n$

则称 $P(Y|X)$ 为线性链条件随机场。标注问题中X观察序列，Y是标注序列。

条件随机场的参数化形式：

在随机变量X取x的条件下，随机变量Y取值为y的条件概率具有

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

（指数内部加法为概率乘法）

第一项为最大团，其中 t_k, s_l 为特征函数， λ_k, μ_l 为对应的权值。

t_k 是定义在观测序列的两个相邻标记位置上的转移特征函数(transition feature function)，用于刻画相邻标记变量之间的相关关系以及观测序列对它们的影响。

s_l 是定义在观察序列的标记位置i上的状态特征函数(status feature function)，用于刻画观察序列对标记变量的影响

t_k, s_l 都依赖于位置，是局部特征。通常取0或1，满足条件取1，否则取0。

条件随机场完全有特征函数和权值决定。

条件随机场的简化形式：

同一个特征在各个位置都有定义，可以对同一个特征在各个位置求和，将局部特征函数转化为一个全局特征函数，从而将条件随机场写成权值向量与特征向量的内积的形式，设有 K_1 个转移特征， K_2 个状态特征， $K = K_1 + K_2$

$$f_k(y_{i-1}, y_i, x, i) = \begin{cases} t_k(y_{i-1}, y_i, x, i), k = 1, 2, \dots, K_1 \\ s_l(y_i, x, i), k = K_1 + l; l = 1, 2, \dots, K_2 \end{cases}$$

一个特征函数对于一个权值，因此可以将转移和状态特征在各个位置i求和，记作：

$$f_k(y, x) = \sum_{i=1}^n f_k(y_{i-1}, y_i, x, i), k = 1, 2, \dots, K$$

用 w_k 表示特征 $f_k(y, x)$ 的权值，即：

$$w_k = \begin{cases} \lambda_k, k = 1, 2, \dots, K_1 \\ \mu_l, k = K_1 + l; l = 1, 2, \dots, K_2 \end{cases}$$

于是条件随机场的形式可以表示为

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

$$Z(x) = \sum_y \exp(\sum_{k=1}^K w_k f_k(y, x))$$

$$\text{权值向量} : w = (w_1, w_2 \dots w_K)^T$$

$$\text{全局特征向量} : F(y, x) = (f_1(y, x), f_2(y, x) \dots f_K(y, x))^T$$

则条件随机场可以写出向量的w与F(y, x)的内积形式。

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

$$Z_w(x) = \sum_y \exp(w \cdot F(y, x))$$

条件随机场的矩阵形式：

引进特殊的起点和终点状态标记 $y_0 = start, y_{n+1} = stop$, 可以用矩阵表示条件随机场。

在观测序列x的每个位置 $i = 1, 2, \dots, n+1$, 定义一个m阶矩阵 (m是 y_i 取值的个数)

$$M_i(x) = [M_i(y_{i-1}, y_i | x)]$$

$M_i(x)_{[j,k]}$ 表示 $y_{i-1} = j, y_i = k$ 的特征函数加权求和的指数。

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

表示同一个位置，不同特征函数之和

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

$$P(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{n+1} (\sum_{k=1}^K w_k f_k(y_{i-1}, y_i, x, i))) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{n+1} W_i(y_{i-1}, y_i | x)) = \frac{1}{Z(x)} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i, x, i)$$

$$Z_w(x) = (M_1(x) M_2(x) \dots M_{n+1}(x))_{start, stop}$$

条件概率的概率计算问题

条件随机场的概率计算问题，是给定条件随机场 $P(Y|X)$, 输入序列为x和输出序列y，计算条件概率 $P(Y_i = y_i | x)$ ， $P(Y_{i-1} = y_{i-1}, Y_i = y_i | x)$ 以及相应的数学期望问题。

前向-后向算法

$\alpha_i(y_i | x)$ 表示在位置i的标记是 y_i 的前部分标记序列的非规范化概率， y_i 取值可能又m个，为一个m维列向量。

$$\alpha_0(y|x) = \begin{cases} 1, y = start \\ 0, otherwise \end{cases}$$

$$\alpha_i^T(x) = \alpha_{i-1}^T(x) M_i(x)$$

同理定义后向概率

$\beta_i(y_i | x)$ 表示位置i的标记为 y_i ，从i+1到n的后半部分标记序列的非规范化概率。

$$\beta_{n+1}(y_{n+1} | x) = \begin{cases} 1, y_{n+1} = end \\ 0, otherwise \end{cases}$$

$$\beta_i(x) = M_{i+1}(x) \beta_{i+1}(x)$$

$$Z(x) = \alpha_n^T(x) \cdot 1 = 1^T \cdot \beta_1(x)$$

概率计算

标记序列在位置*i*是 y_i 的条件概率

$$P(Y_i = y_i | x) = \frac{\alpha_i^T(y_i | x) \beta_i(y_i | x)}{Z(x)}$$

在位置*i* - 1和位置*i* , 标记为 y_{i-1}, y_i 的条件概率为：

$$P(Y_i = y_i, Y_{i-1} = y_{i-1} | x) = \frac{\alpha_{i-1}^T M_i(y_{i-1}, y_i | x) (y_i | x) \beta_i(y_i | x)}{Z(x)}$$

条件随机场的学习算法

已知训练集 , 可以得到经验概率分布 $\tilde{P}(X, Y)$, 可以通过极大化训练数据的对数似然函数来求模型参数：

$$L(a) = \sum_{x,y} \tilde{P}(x, y) \log P(y | x) = \sum_{i=1}^n \sum_{k=1}^K w_k f_k(y_i, x_i) - \sum_{i=1}^n \log Z(x_i)$$

上述问题为关于 w_i 的无约束最优化问题 , 因此 , 可以直接采用梯度下降法就行求解 , 由于上式为凸函数 , 因此 , 可以采用更快速的牛顿法或拟牛顿法进行求解。

条件随机场的预测算法

条件随机场的预测问题是给定条件随机场 $P(Y|X)$ 和输入序列x , 求条件概率最大的输出概率 y^* 的问题。

$$y^* = \operatorname{argmax}_y P(y | x) = \operatorname{argmax}_y \frac{\exp(w \cdot F(y, x))}{Z(x)} = \operatorname{argmax}_y \exp(w \cdot F(y, x)) = \operatorname{argmax}_y w \cdot F(y, x)$$

$w = (w_1, w_2, \dots, w_K)^T$ 不同特征函数的权重

$$F(y, x) = (f_1(y, x), f_2(y, x), \dots, f_K(y, x))^T$$

因此

$$y^* = \sum_{i=1}^n w F_i(y_{i-1}, y_i, x)$$

其中 $F_i(y_{i-1}, y_i, x) = (f_1(y_{i-1}, y_i, x, i), f_2(y_{i-1}, y_i, x, i), f_3(y_{i-1}, y_i, x, i), \dots, f_K(y_{i-1}, y_i, x, i),)$

同一个位置 , 不同特征函数组成的向量。

由于上式满足最优子结构 , 因此可以采用动态规划方法进行求解。

初始化：

求出位置1标记为j=1,2...m的非规范化概率

$$\delta_1(j) = w \cdot F_1(y_0 = \text{start}, y_1 = j, x)$$

递归公式：

$$\delta_i(l) = \max_{1 \leq j \leq m} (\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x)) \quad i=1, 2, \dots, m$$

$$\Phi_i(l) = \operatorname{argmax}_{1 \leq j \leq m} (\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x))$$

$$y_n^* = \operatorname{argmax}_{1 \leq j \leq m} \delta_n(j)$$

回溯

$$y_i^* = \Phi_{i+1}(y_{i+1}^*), i = n-1, \dots, 1$$