四、10、条件随机场

2016年3月27日 星期日　　下午9:25

一、简介

CRF、conditional random field、条件随机场。

是在随机变量X情况下，随机变量Y的马尔可夫随机场。通常我们在序列标注问题中用到CRF，在序列标注问题中随机变量X是输入序列（标注序列），这是已知条件；随机变量Y的是输出序列（状态序列），待预测序列。也就是在条件X情况下Y的随机场，所以叫条件随机场。

或者说

CRF是给定一组随机变量情况下，另一组随机变量的条件概率分布模型。他的特点是，假设输出随机变量构成马尔可夫随机场。

二、先验知识

1、概率图模型

probabilistic graphical model，是用图表示的概率分布。其中，结点（node）表示随机变量，边（edge）表示随机变量之间的概率依赖关系。

2、马尔可夫随机场

概率无向图模型（probabilistic undirected graphical model）、又称马尔可夫随机场（Markov random field）

概率图模型中的边是没有方向的，这样的模型叫做概率无向图模型，又称马尔可夫随机场。

概率无向图模型满足以下性质（也是判定定理，因为是等价的？）：

1、成对马尔可夫性

2、局部马尔可夫性

3、全局马尔可夫性

看起来性质很高深，简单理解，我认为就是“没有边连接的结点之间是条件独立的”。

3、马尔可夫性

1、成对马尔可夫性（pairwise Markov property）

任意两个没有边连接的结点u，v与其他所有节点O有：

$$P(Y_u, Y_v \mid Y_O) = P(Y_u \mid Y_O)P(Y_v \mid Y_O)$$

2、局部马尔可夫性（loca Markov property）

v是任意结点，W是与v有边连接的所有结点，O是v、W之外的所有结点，有：

$$P(Y_v, Y_O \mid Y_W) = P(Y_v \mid Y_W)P(Y_O \mid Y_W)$$

等价的有

$$P(Y_v \mid Y_W) = P(Y_v \mid Y_W, Y_O)$$

3、全局马尔可夫性（global Markov property）

集合A、B是在图中被C分开的任意结点集合，有：

$$P(Y_A, Y_B \mid Y_C) = P(Y_A \mid Y_C)P(Y_B \mid Y_C)$$

以上三个性质是等价的。

简单理解，我认为就是“没有边连接的结点之间是条件独立的”。

4、最大熵模型

在满足观测数据的基础上，对未知的参数估计时，尽量让其熵最大的模型。通常用于分类问题。详见三、7.

三、思想

1、线性链条件随机场

本文主要介绍线性链条件随机场（linear chain conditional random field）。一般用于序列标注问题中，X是输入序列（观测序列），Y是输出序列（标记序列、状态序列）。

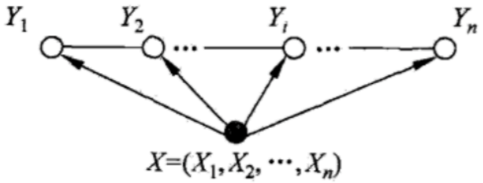

图 11.4　线性链条件随机场

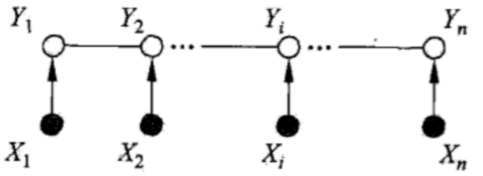

图 11.5　X 和 Y 有相同的图结构的线性链条件随机场

2、整体思想

训练：极大似然估计，预测：预测各可能序列的概率，挑出概率最大的序列作为结果。

3、公式形式表示

输入序列是x，输出序列是y的概率：

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

A 变量解释

Z（x）是归一化项，类似最大熵和softmax，λ和μ是权重，t和s是特征，在书中是0/1的特征，但是实际使用可以是实质特征（是一个数）。

i是序列标号，表示当前是整个序列中的第几个tag。

l和k仅仅是循环的下标，没有实际意义，但是注意他们循环的大小：

转移特征的k的维度K = 输入x的特征维度 * 输出y的类别数 * 输出y的类别数，也就是说转移位置的权重参数λ有K维，也就是说特征的每一维a和前一个词的类别b、当前词的类别c都有一个权重连接，类似W_abc,W[a][b][c]，

同理状态特征l的维度L = 输入x的特征维度 * 输出y的类

别数，也就是说特征的每一维a和当前词的类别b，都有一个权重链接。

B 由来

这个公式源于条件概率公式，最原始的形式是：

P（Y | X）= P（Yi | X ，Y1 ，Y2 ...Yi-1 ，Yi+1...Yn）

由于输出变量Y满足马尔可夫性，而且是线性链，所以

P（Yi | X ，Y1 ，Y2 ...Yi-1 ，Yi+1...Yn）= P（Yi | X，Yi-1 ，Yi+1）

至于相邻时刻的Y有关。

C 概率相乘

上式表示的是输入序列是x情况下，输出序列是y的条件概率。

$$\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

这一项可以拆成多个exp（λ*t）或exp（μ*s）的形式，根据条件概率公式，这每一个乘子都表示一个事件的概率。

但是为何在这里用exp（ ），可以看做是softmax的未归一化概率

D 最大熵

这里参数的表示方式如果不太理解，可以参考最大熵部分的讲解，是很相似的。

4、与最大熵模型

与最大熵模型的形式和求解方法有相似之初，可以简单的理解为，CRF是在最大熵基础上增加了转移特征，详见后面的叙述。

四、算法

五、优化算法

A 改进的迭代尺度法

（见统计学习方法）

与最大熵的优化算法很类似

B 拟牛顿法

（见统计学习方法）

与最大熵的优化算法很类似

C 随机梯度下降

（见四、10.2 条件随机场代码讲解）

六、与最大熵模型的联系