

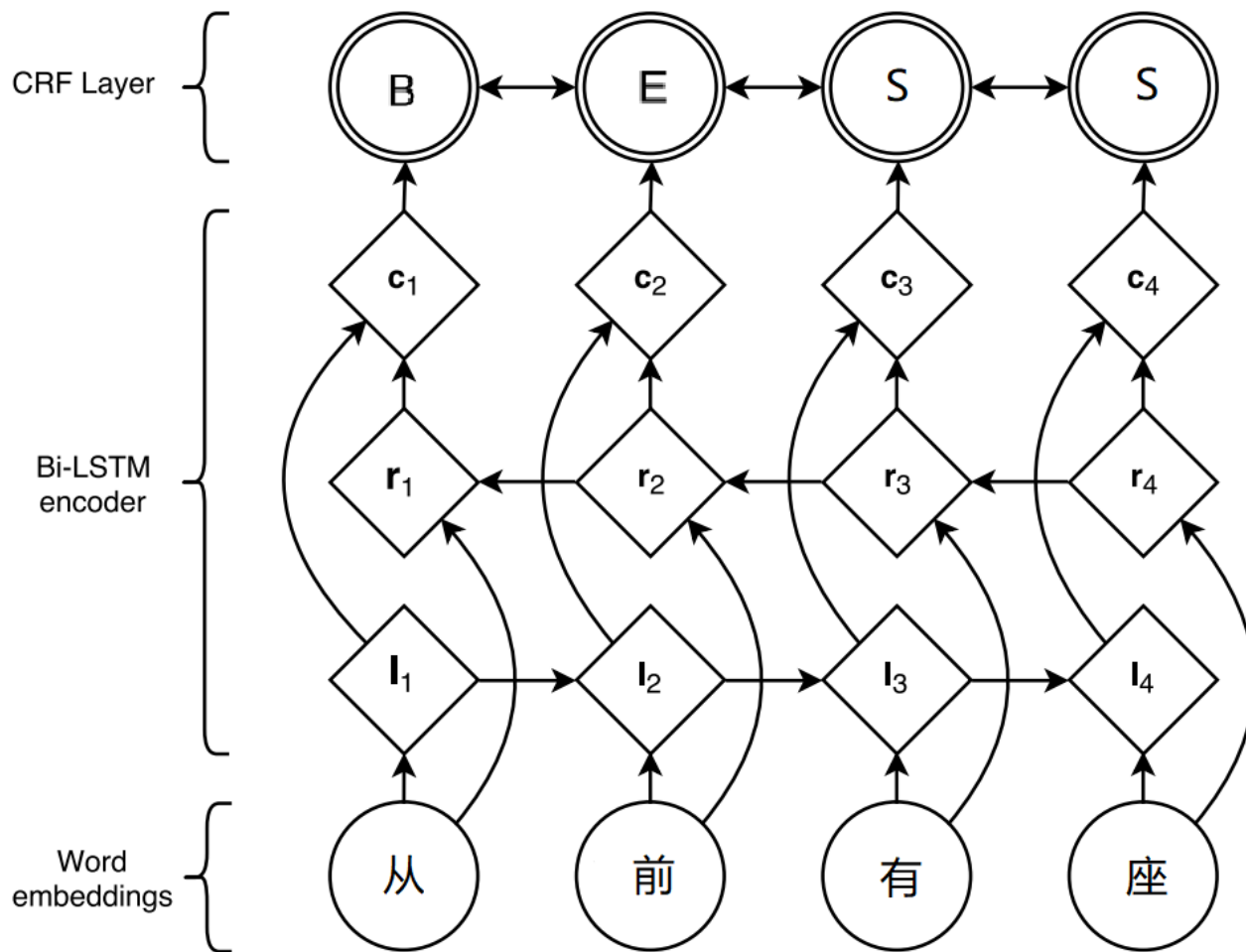
自然语言处理

Natural Language Processing

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn



基于LSTM+CRF的序列标注方法

Lecture 6: 条件随机场

Lecture 6.1 概述

随机场

随机场是由若干个位置组成的整体，当给每一个位置中按照某种分布随机赋予一个值之后，其全体就叫做**随机场**。

以词性标注为例：假如有一个十个词组成的句子需要做词性标注。这十个词每个词的词性可以在已知的词性集合（名词，动词...）中去选择。当我们为每个词选择完词性后，这就形成了一个**随机场**。

马尔可夫随机场

马尔科夫随机场是随机场的特例，它假设随机场中某一个位置的赋值仅仅与和它相邻的位置的赋值有关，和与其不相邻的位置的赋值无关。

条件随机场

条件随机场(conditional random field, CRF)是**马尔科夫随机场**的特例，它假设马尔科夫随机场中只有X和Y两种变量，X一般是给定的，而Y一般是在给定X的条件下的输出。

在词性标注的例子中，X是词，Y是词性。如果我们假设它是一个马尔科夫随机场，那么它也就是一个**条件随机场**。

概述

◆ 提出

条件随机场(conditional random field, CRF)于2001年由 J. Lafferty 等人提出，是用于标注和划分序列结构数据的概率化结构模型，在NLP和图像处理中得到了广泛应用。

基本思路：给定观察序列 X ，输出标记序列 Y ，通过计算 $P(Y|X)$ 求解最优标记序列。

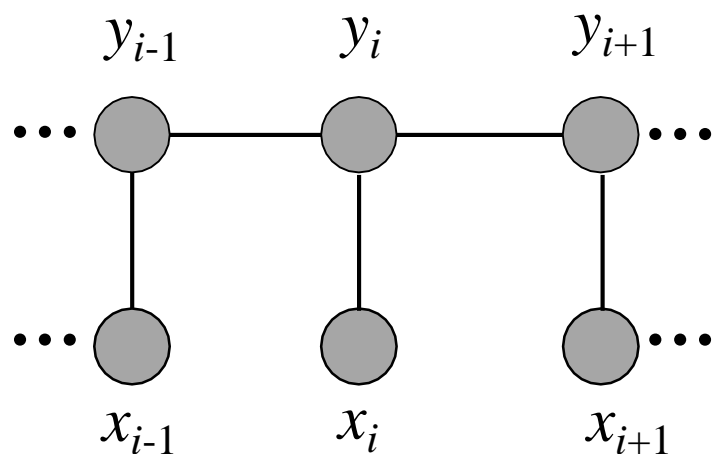
举例

基本思路：给定观察序列 X ，输出标记序列 Y ，通过计算 $P(Y|X)$ 求解最优标记序列。

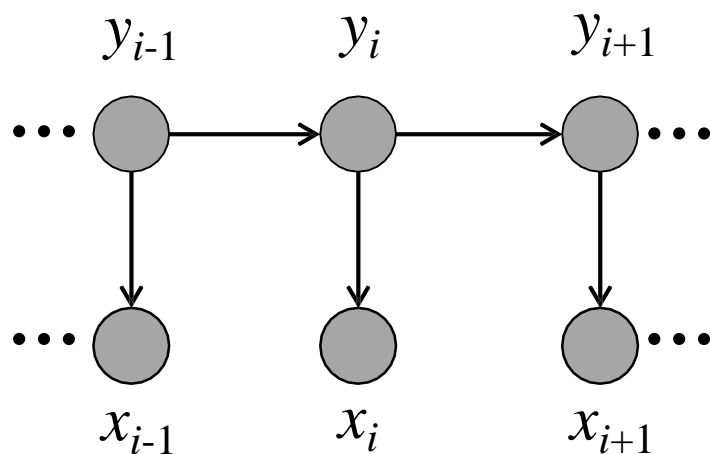
自然语言处理中的 **词性标注**和 **中文分词**就是适合CRF使用的任务，因为它们往往和上下文有关。

图示

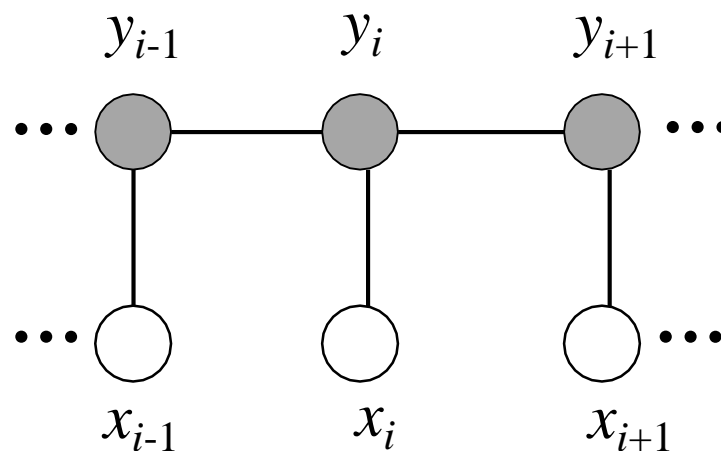
序列标注问题可以建模为简单的链式结构图，结点对应标记序列 Y 中的元素。如下图所示：



HMMs vs. CRFs



HMMs



CRFs

CRFs中的空心节点 x 表示该节点并不是由模型生成的

参数化形式

我们如何将条件随机场转化为机器学习模型呢？

答案：通过特征函数和其权重系数。

参数化形式

我们如何将条件随机场转化为机器学习模型呢？

答案：通过特征函数和其权重系数。

什么是特征函数呢？

这里的特征函数分为两类，一类是定义在Y节点上的节点特征函数，称作状态函数，只和当前节点有关，记为：

$$s_l(y_i, X, i) \quad l = 1, 2, 3, \dots, L$$

其中， L 是定义在该节点的节点特征函数的总个数， i 是当前节点在序列的位置。

参数化形式

什么是特征函数呢？

第二类是定义在Y上下文的转移特征函数，这类特征函数只和当前节点和上一个节点有关，记为：

$$t_k(y_{i-1}, y_i, X, i) \quad k = 1, 2, 3, \dots, K$$

其中, K 是定义在该节点的转移特征函数的总个数, i 是当前节点在序列的位置。之所以只有上下文相关的转移特征函数, 没有不相邻节点之间的特征函数, 是因为CRF满足马尔科夫性质。

参数化形式

- ❖ 无论是状态函数还是转移函数，它们的取值只能是0或者1,即满足特征条件或者不满足特征条件。
- ❖ 同时，我们可以为每个特征函数赋予一个权值，用以表达这个特征函数的重要性。
- ❖ 假设 t_k 的权重系数是 λ_k ， s_l 的权重系数是 μ_l ，则CRF的参数化形式为：

$$P(Y/X) = \frac{1}{Z(X)} \exp\left(\sum_{i,k} \lambda_k t_k(Y_{i-1}, Y_i, X, i) + \sum_{i,l} \mu_l s_l(Y_i, X, i)\right)$$

参数化形式

其中， $Z(X)$ 为归一化因子：

$$Z(X) = \sum_Y \exp\left(\sum_{i,k} \lambda_k t_k(Y_{i-1}, Y_i, X, i) + \sum_{i,l} \mu_l s_l(Y_i, X, i)\right)$$

Lecture 6.2 模型训练

CRFs及其应用

◆ 应用举例：基于字标注的分词方法

基本思想：将分词过程看作是字的分类问题，每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。一般情况下，每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)。

CRFs及其应用

乒乓球拍卖完了。

(1) 乒乓球/拍/卖/完/了/。/

(2) 乒/B 乓/M 球/M 拍/E 卖/S 完/S 了/S 。/S

在字标注过程中，对所有的字根据预定义的特征进行词位特征学习，获得一个概率模型，然后在待切分字符串上，根据字与字之间的结合紧密程度，得到一个词位的分类结果，最后根据词位定义直接获得最终的分词结果。

CRFs及其应用

乒/B 乒/M 球/E 拍/S 卖完了。



B, E, M, S ?

特征：

- 当前字的前后 n 个字
- 当前字左边字的标记
- 当前字在词中的位置
-

模型训练

实现 CRFs 也需要解决如下三个问题：

① 特征选取

② 参数训练

③ 解码

模型训练

①特征选取

- 一元特征：当前字、当前字的前一个字、当前字的后一个字
- 二元特征：各标记间的转移特征

$$s_1(y_i, X, i) = \begin{cases} 1 & \text{如果当前字是“拍”，当前字的标记 } y_i \text{ 是 M} \\ 0 & \text{否则} \end{cases}$$

$$s_2(y_i, X, i) = \begin{cases} 1 & \text{如果当前字是“拍”，当前字 } y_i \text{ 的标记是 E} \\ 0 & \text{否则} \end{cases}$$

.....

乒/B 乒/M 球/E 拍/S 卖? 完了。

模型训练

①特征选取

对应转移函数的特征：

$$t_1(y_{i-1}, y_i, X, i) = \begin{cases} 1 & \text{如果前一个字的标记} y_{i-1} \text{ 是B, 当前字的标记} y_i \text{ 是M} \\ 0 & \text{否则} \end{cases}$$

$$t_2(y_{i-1}, y_i, X, i) = \begin{cases} 1 & \text{如果前一个字的标记} y_{i-1} \text{ 是M, 当前字的标记} y_i \text{ 是M} \\ 0 & \text{否则} \end{cases}$$

.....

乒/B 兵/M 球/E 拍/S 卖? 完了。

模型训练

②参数训练

通过训练语料估计特征权重 λ_j ,使其在给定一个观察序列 X 的条件下,找到一个最有可能的标记序列 Y ,即条件概率 $P(Y|X)$ 最大。

条件概率已由上文给出:

$$P(Y/X) = \frac{1}{Z(X)} \exp\left(\sum_{i,k} \lambda_k t_k(Y_{i-1}, Y_i, X, i) + \sum_{i,l} \mu_l s_l(Y_i, X, i)\right)$$

模型训练

为了训练特征权重 λ_j ，需要计算模型的损失和梯度。由梯度更新 λ_j ，直到 λ_j 收敛。

- 损失函数定义为负对数似然函数：

$$L(\lambda) = -\log p(Y | X, \lambda) + \frac{\varepsilon}{2} \lambda^2$$

采用随机梯度下降优化该损失函数！

模型训练

③解码

条件随机场解码的过程就是根据模型求解的过程,可以由维特比(Viterbi)算法完成。维特比算法是一个动态规划算法,动态规划要求局部路径也是最优路径的一部分。

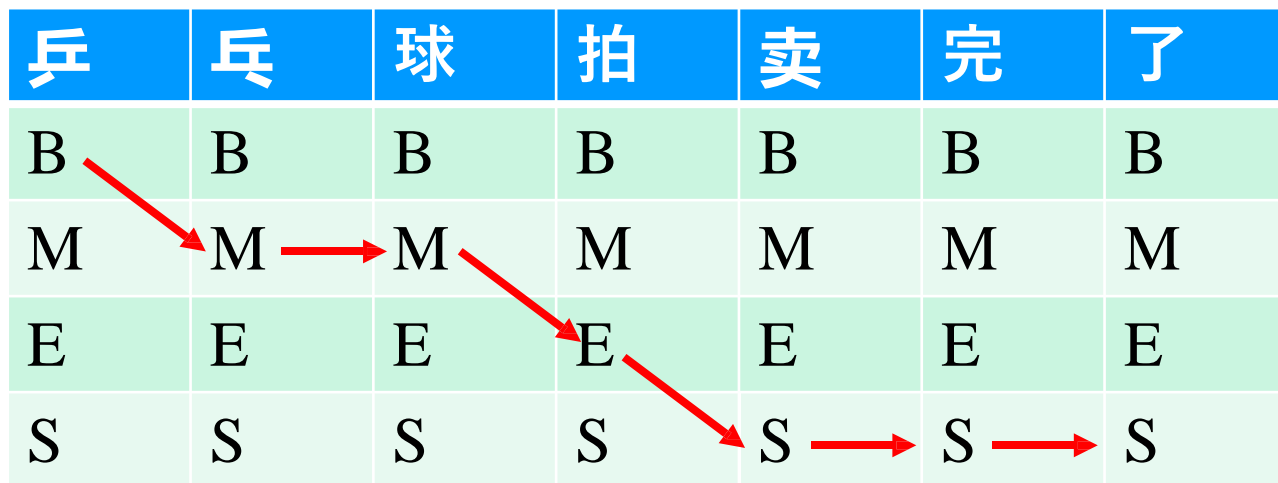
模型训练

③ 解码

以中文分词为例：乒 乓 球 拍 卖 完 了

维特比算法就是在下面由标记组成的矩阵中搜索一条最优的路径。

乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S



分词结果：乒/B 乓/M 球/M 拍/E 卖/S 完/S 了/S

模型训练

到达每个标记的分数由以下三部分组成：

- 标记的一元特征权重 W ：分别用 W_1^B 表示第一个字被标记为B的权重，用 W_1^S 表示第一个字被标记为S的权重，等等。
- 标记的路径得分 R ：分别用 R_2^B 表示第二个字被标记为B时的路径得分，用 R_2^E 表示第二个字被标记为E的路径得分，等等。
- 前一个字的标记到当前字标记转移的特征权重 T ：用 T_{BM} 表示由标记B到M的转移特征权重，类似地，其他转移特征权重分别记为： T_{BE} 、 T_{MM} 、 T_{ME} 、 T_{EB} 、 T_{ES} 、 T_{SB} 和 T_{SS} 等。

模型训练

- 利用下式迭代计算每一字被标记为每一种标记的分数：

$$\mathbf{R}_{i+1}^B = \max\{ \mathbf{T}_{EB} \times \mathbf{R}_i^E, \mathbf{T}_{SB} \times \mathbf{R}_i^S \} \times \mathbf{W}_{i+1}^B$$

$$\mathbf{R}_{i+1}^E = \max\{ \mathbf{T}_{BE} \times \mathbf{R}_i^B, \mathbf{T}_{SE} \times \mathbf{R}_i^E \} \times \mathbf{W}_{i+1}^E$$

$$\mathbf{R}_{i+1}^S = \max\{ \mathbf{T}_{ES} \times \mathbf{R}_i^E, \mathbf{T}_{SS} \times \mathbf{R}_i^S \} \times \mathbf{W}_{i+1}^S$$

....

$$P(Y/X) = \frac{1}{Z(X)} \exp\left(\sum_{i,k} \lambda_k t_k(Y_{i-1}, Y_i, X, i) + \sum_{i,l} \mu_l s_l(Y_i, X, i) \right)$$

举例

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

第一步：计算第一个字“乒”的标记分数（以标记B为例）。由于不存在转移特征，故路径权重 R_1^B 为：

$$R_1^B = W_1^B = \lambda_1 \times f(\text{null}, \text{乒}, B) + \lambda_2 \times f(\text{乒}, B) + \lambda_3 \times f(\text{乒}, B, \text{乒})$$

$f(\bullet)$ 表示特征，其中 $f(\text{null}, \text{乒}, B)$ 表示当前字“乒”被标记为B，前一个字为空； $f(\text{乒}, B)$ 表示当前字“乒”被标记为B； $f(\text{乒}, B, \text{乒})$ 表示当前字“乒”被标记为B，且后一个字为“乒”。特征的权重 λ_1 、 λ_2 和 λ_3 都可以从训练中得到（参数训练部分）。

举例

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

第二步：计算第二个字“乓”的标记分数（以标记B为例）。
首先计算一元权重 W_2^B ，继而由上一个字的路径权重计算当前路径权重 R_2^B 为：

$$R_2^B = \max\{T_{EB} \times R_1^E, T_{SB} \times R_1^S\} \times W_2^B$$

同样，对于“乓”字的标记S、M和E分别计算 R_2^M 、 R_2^E 和 R_2^S 。

举例

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

第三步：依据第二步迭代计算直至最后一个“了”字，得到 R_7^E , R_7^S 。比较这两个值，确定最优路径，然后以该值的标记点为起始点回溯，得到整个句子的最优路径。回溯过程：

由： $\max\{R_7^E, R_7^S\} = R_7^S$ ，可推出“了”字标记为S；

由： $R_7^S = \max\{T_{ES} \times R_6^E, T_{SS} \times R_6^S\} \times W_7^S = T_{SS} \times R_6^S \times W_7^S$ ，

可推出“完”字标记为S；依次回溯至第一个字，解码完毕。

Thank you!

权小军 中山大学数据科学与计算机学院