

# 第3章 条件随机场及其应用

宗成庆

中国科学院自动化研究所

[cqzong@nlpr.ia.ac.cn](mailto:cqzong@nlpr.ia.ac.cn)



# 本章内容

---

- ➡ 1. CRFs 定义
- 2. 应用举例
- 3. 习题



# 1. CRFs 定义

## ◆ 模型提出

在NLP和图像处理中有一类问题是进行序列标注和结构划分，而 $n$ -gram是利用当前时刻  $t$  之前已经发生的事件信息。J. Lafferty 等人于2001年提出了条件随机场(**conditional random fields, CRFs**)这一概率化的结构模型。

## ● 基本思想

给定观察序列  $X$ ，输出标识序列  $Y$ ，通过计算  $P(Y|X)$  求解最优标注序列。



# 1. CRFs 定义

## ◆模型定义

设  $G=(V, E)$  为一个无向图,  $V$ 为结点集合,  $E$ 为无向边的集合,  $Y = \{ Y_v | v \in V \}$ , 即 $V$ 中每个结点对应于一个随机变量  $Y_v$ , 其取值范围为可能的标记集合 $\{y\}$ 。如果以观察序列 $X$ 为条件, 每个随机变量  $Y_v$  都满足以下马尔可夫特性:

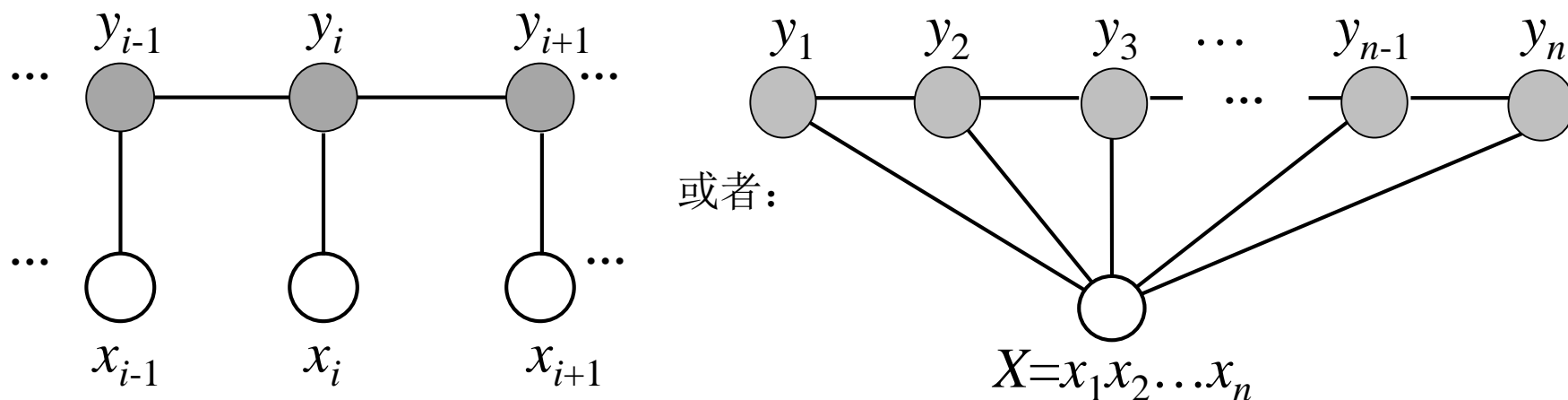
$$p(Y_v / X, Y_w, w \neq v) = p(Y_v / X, Y_w, w \sim v) \quad \dots (5-1)$$

其中,  $w \sim v$  表示两个结点在图中是邻近结点。那么,  $(X, Y)$  为一个条件随机场。

# 1. CRFs 定义

图示：

$$p(Y_v / X, Y_w, w \neq v) = p(Y_v / X, Y_w, w \sim v)$$



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。理论上，只要在标记序列中描述一定的条件独立性， $G$ 的图结构可以任意的。



# 1. CRFs 定义

在CRFs中, 给定观察序列 $X$  时, 某个特定标记序列 $Y$ 的概率可以定义为:

$$p(Y|X) = \frac{1}{Z(X)} \exp \left( \sum_j \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_k \mu_k s_k(y_i, X, i) \right) \quad \dots (5-2)$$

其中,  $t_j(y_{i-1}, y_i, X, i)$ 是转移函数, 表示对于观察序列  $X$  的标注序列从*i-1到i位置上标记的转移概率*。通常把转移函数称作二元特征。 $s_k(y_i, X, i)$ 是状态函数, 表示观察序列 $X$ 在*i位置的标记概率*。通常把状态函数称作一元特征。 $\lambda_j$  和  $\mu_k$  分别是  $t_j$  和  $s_k$  的权重, 需从训练样本中估计出。

$Z(X)$ 为归一化因子:

$$Z(X) = \sum_Y \exp \left( \sum_{i,j} \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_{i,k} \mu_k s_k(y_i, X, i) \right)$$



# 1. CRFs 定义

定义一组关于观察序列的 $\{0, 1\}$  二值特征  $b(X, i)$ , 表示训练样本中某些特征的分布, 如

$$b(X, i) = \begin{cases} 1 & \text{如果 } X \text{ 的 } i \text{ 位置为某个特定的词} \\ 0 & \text{否则} \end{cases}$$

转移函数可以定义为如下形式:

$$t_j(y_{i-1}, y_i, X, i) = \begin{cases} b(X, i) & \text{如果 } y_{i-1} \text{ 和 } y_i \text{ 满足某种搭配条件} \\ 0 & \text{否则} \end{cases}$$

也可以把状态函数写成如下形式:

$$s(y_i, X, i) = s(y_{i-1}, y_i, X, i)$$



# 1. CRFs 定义

由此，特征函数可以统一表示为：

$$F_j(Y, X) = \sum_{i=1}^n f_j(y_{i-1}, y_i, X, i) \quad \dots (5-3)$$

其中，每个局部特征函数  $f_j(y_{i-1}, y_i, X, i)$  表示状态特征  $s(y_{i-1}, y_i, X, i)$  或转移数  $t(y_{i-1}, y_i, X, i)$ 。

条件随机场定义的条件概率可以由下式给出：

$$p(Y | X, \lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right) \quad \dots (5-4)$$

其中， $Z(X)$  为归一化因： $Z(X) = \sum_Y \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$





# 1. CRFs 定义

## ◆回顾第2章中的最大熵模型:

对于求解的问题, 就是估计在条件 $b \in B$ 下(已知知识), 发生某个事件 $a \in A$ (未知分布) 的概率 $p(a|b)$ , 该概率使熵 $H(p(A|B))$ 最大。

$$p^*(a|b) = \frac{1}{Z(b)} \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right)$$

特征函数

特征权重

其中,  $Z(b) = \sum_a \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right)$

$Z(b)$ 为保证对所有 $b$ , 使得  $\sum_a p(a|b) = 1$  的归一常量。



# 1. CRFs 定义

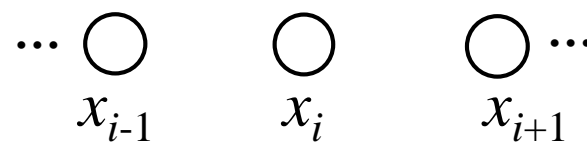
## ◆ ME 模型 与 CRFs 对比:

### ● 相同点:

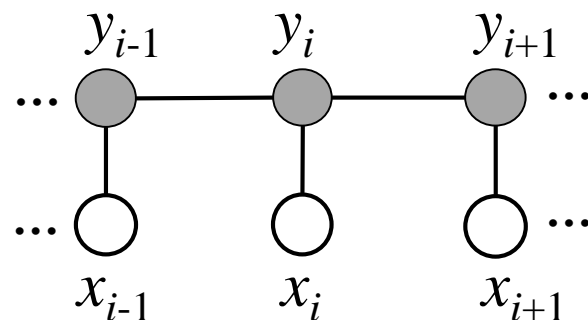
- 都是通过特征函数计算概率，模型形式也一样;
- 可以采用同样的参数训练方法。

### ● 不同点:

- 基于 ME 的分类器对给定输入  $X$  的整体（作为一个单位）或某个点进行分  
类，是解决“点”的问题;
- CRFs 模型是对给定输入  $X$  进行序列  
标注，最终求解的是全局最优标注序  
列  $Y$ 。



(a) ME



(b) CRFs

ME能完成的任务CRFs也能够完成。



# 1. CRFs 定义

## ◆ 模型实现

### ● 需要解决的三个问题：

- ① 特征选取  $\longrightarrow$  参阅最大熵模型。见本章“应用举例”。
- ② 参数训练  $\longrightarrow$  每个特征的权重 $\lambda$  如何确定？
- ③ 解码  $\longrightarrow$  如何快速地搜索最优路径。



# 1. CRFs 定义

## ②参数训练

通过训练语料估计特征权重 $\lambda_j$ ，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。

条件概率已由上文的(5-4)式给出：

$$p(Y | X, \lambda) = \frac{1}{Z(X)} \exp(\sum_j \lambda_j \cdot F_j(Y, X))$$

$$Z(X) = \sum_Y \exp(\sum_j \lambda_j \cdot F_j(Y, X))$$



# 1. CRFs 定义

$$p(Y | X, \lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$

$$Z(X) = \sum_Y \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$

为了训练特征权重 $\lambda_j$ ，需要计算模型的损失和梯度。由梯度更新 $\lambda_j$ ，直到 $\lambda_j$ 收敛。损失函数可定义为负对数似然函数：

$$L(\lambda) = -\log p(Y | X, \lambda) + \frac{\varepsilon}{2} \|\lambda\|_2^2 \quad (\varepsilon \text{取值范围: } 10^{-6} \sim 10^{-3})$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_M \end{bmatrix}$$

$$\|\lambda\|_2 = \sqrt{\sum_i \lambda_i^2}$$



# 1. CRFs 定义

$$p(Y | X, \lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$

$$Z(X) = \sum_Y \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$

$$L(\lambda) = -\log p(Y | X, \lambda) + \frac{\varepsilon}{2} \|\lambda\|_2^2$$

损失函数的梯度为：

$$\frac{\partial L(\lambda)}{\partial \lambda_j} = \frac{\partial \log Z(X)}{\partial \lambda_j} - F_j(Y, X) + \varepsilon \lambda_j$$

$$\lambda_j^{(k+1)} = \lambda_j^{(k)} - \ell \frac{\partial L(\lambda)}{\partial \lambda_j^{(k)}} \quad \dots(5-5)$$

( $k$  为迭代的轮次,  $\ell$  为学习率, 经验值, 可设为0.1等。)



# 1. CRFs 定义

## ● 算法描述

输入：训练集  $\mathcal{D} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^N$ ，学习率  $\ell$ ，正则化系数  $\varepsilon$ ，损失变化上界  $\delta$ 。

1 令  $k = 0$ ，随机初始化特征权重  $\lambda_k$ ， $\Delta L(\lambda^{(k)}) = +\infty$ ； //  $k$  是迭代的轮次  
将训练集划分成若干批。

2 while TURE do:

3 采样一个批次的训练样本  $\{(X^{(i)}, Y^{(i)})\}_{i=1}^m$  ( $m \leq N$ )，计算得到  
该批次样本集的概率  $p(Y^{(i)} | X^{(i)}, \lambda^{(k)})$ ;

$$p(Y | X, \lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$
$$Z(X) = \sum_Y \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$

4 采用负对数似然概函数作为目标函数：

$$L(\lambda^{(k)}) = \sum_{i=1}^m \left( -\log p(Y^{(i)} | X^{(i)}, \lambda^{(k)}) \right) + \frac{\varepsilon}{2} \|\lambda^{(k)}\|_2^2$$

5 if  $k > 1$ :

6 计算损失变化：  $\Delta L(\lambda^{(k)}) = \|L(\lambda^{(k)}) - L(\lambda^{(k-1)})\|$

7 如果  $\Delta L(\lambda^{(k)}) \leq \delta$  结束循环；

8 计算特征权重的梯度值  $\lambda' = \frac{\partial L(\lambda^{(k)})}{\partial \lambda_j^{(k)}}$ ，  $j = 1 \dots M$ ;

9 更新特征权重和迭代轮次计数：  $\lambda^{(k+1)} = \lambda^{(k)} - \ell \lambda'$ ；  $k = k + 1$ ;

10 End

对整个训练集  $\mathcal{D}$  上的所有  
批次样本反复迭代，  
使模型达到整体最优。

# 1. CRFs 定义

## ③ 解码

- **问题描述:** 给定的  $X = x_1x_2...x_T$ , 从1到 $T$ 的每一时刻  $x_i$  都有 $N$ 个可能的标记  $y_1, y_2, ..., y_N$ , 搜索对于 $X$ 的最优标记序列。

## ● 解决思路:

- 将所有可能的路径全部列出来, 依次对比, 选择概率最大的那条路径。

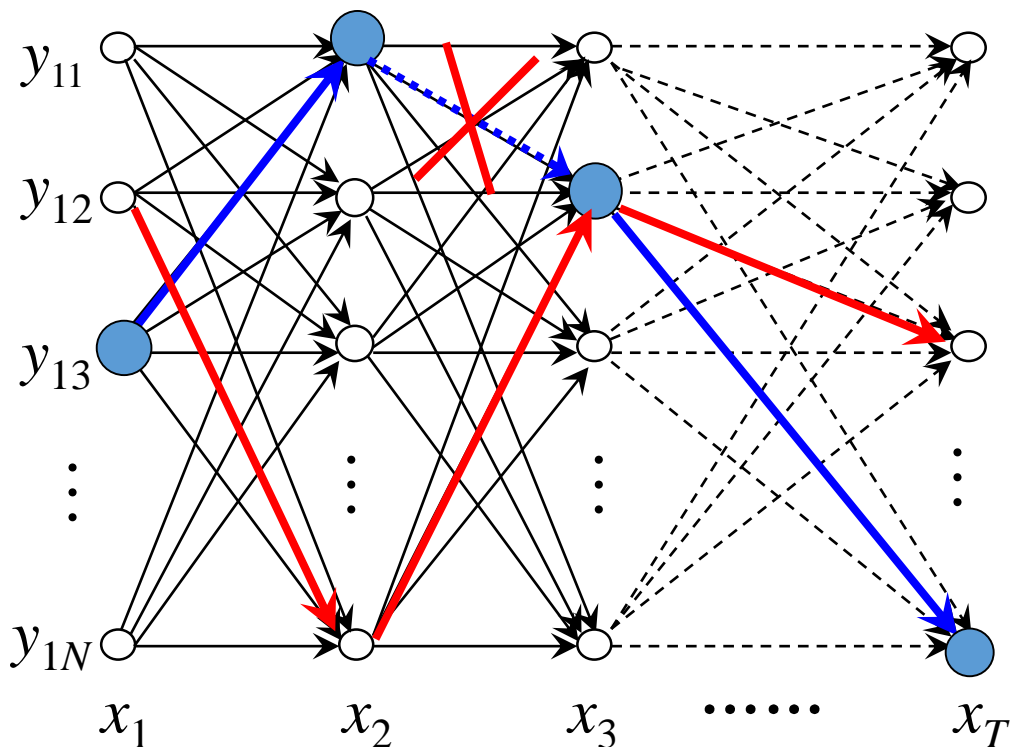
复杂度高:  $N^T$  ✗

- 对于每一时刻, 选择概率最大的标记。

有时标记之间不同现 ✗

- 动态规划, 分而治之。

—— **Viterbi 搜索**







# 1. CRFs 定义

● **Viterbi 搜索算法：** 采用动态规划策略, 搜索全局最优状态序列

➤ **定义：** Viterbi变量 $\delta_t(i)$ 是在 $t$  时刻模型沿着某一条路径到达 $y_i$  的概率（得分）。那么，下一时刻的路径得分为：

$$\delta_{t+1}(j) = \max_{1 \leq k \leq N} \{ \delta_t(k) + \alpha_{\text{path}} \} \quad \dots (5-6)$$

$$\alpha_{\text{path}} = \log(p(Y | X, \lambda)) = \log \left( \frac{\exp(\sum_j \lambda_j \cdot F_j(Y, X))}{\sum_Y \exp(\sum_j \lambda_j \cdot F_j(Y, X))} \right) \quad \dots (5-7)$$

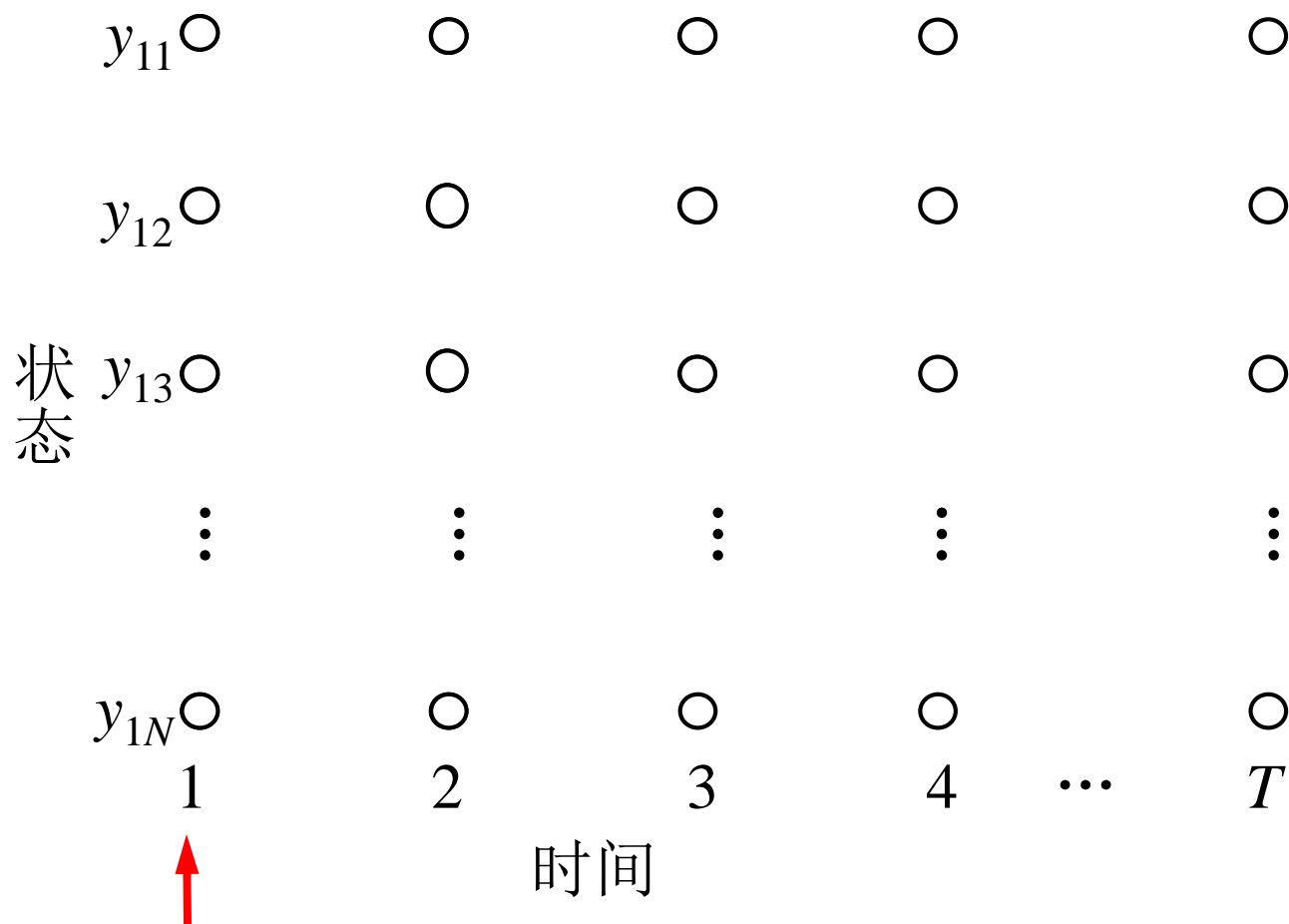
$Z(X)$

$$\propto \sum_j \lambda_j \cdot F_j(Y, X)$$



# 1. CRFs 定义

图解  
Viterbi  
搜索  
过程



# 1. CRFs 定义

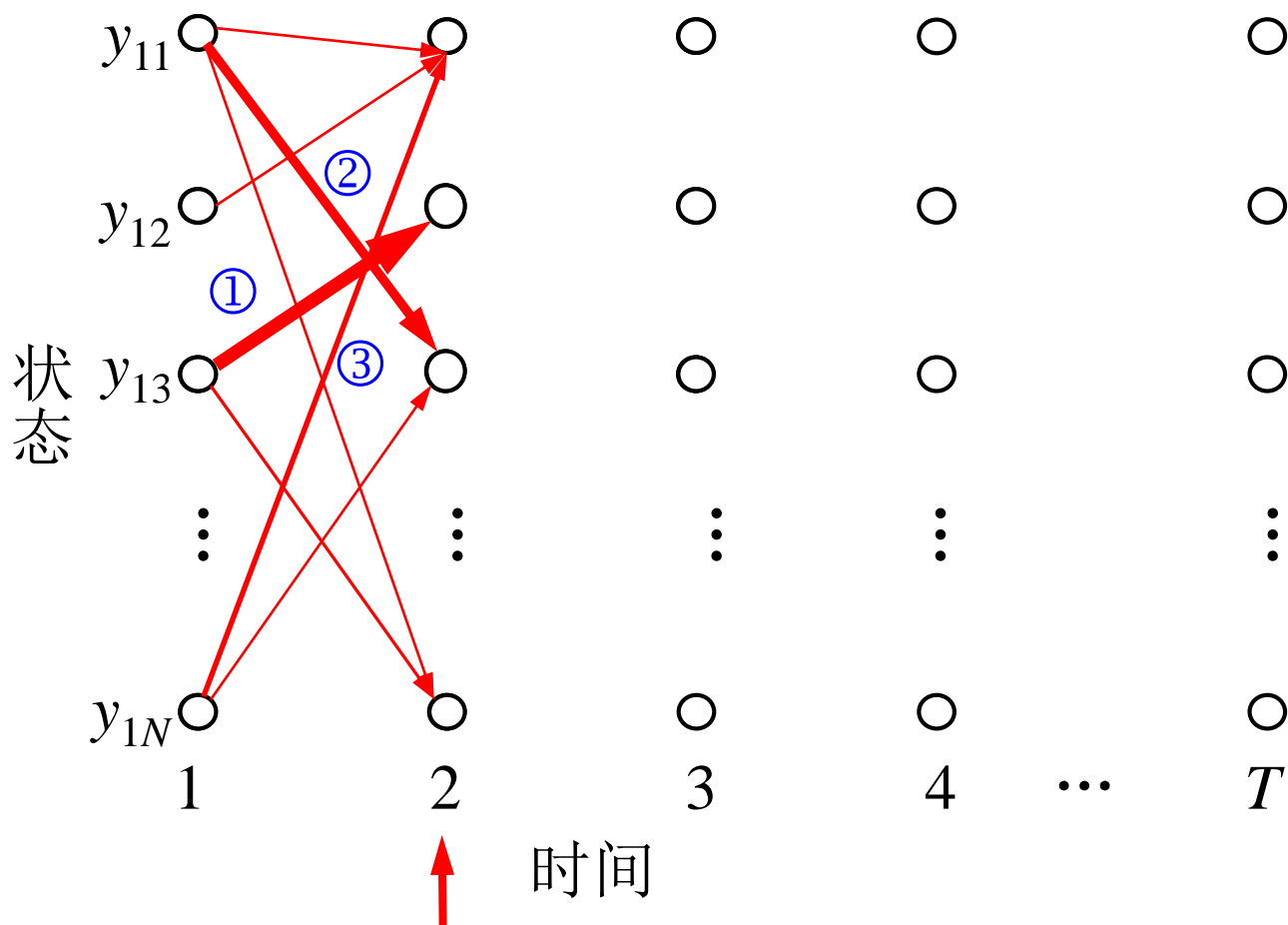
图解  
Viterbi  
搜索  
过程

剪枝策略:

①  $\delta_t(j) \geq \Delta$

②  $NPath \leq \sigma$

(3)



# 1. CRFs 定义

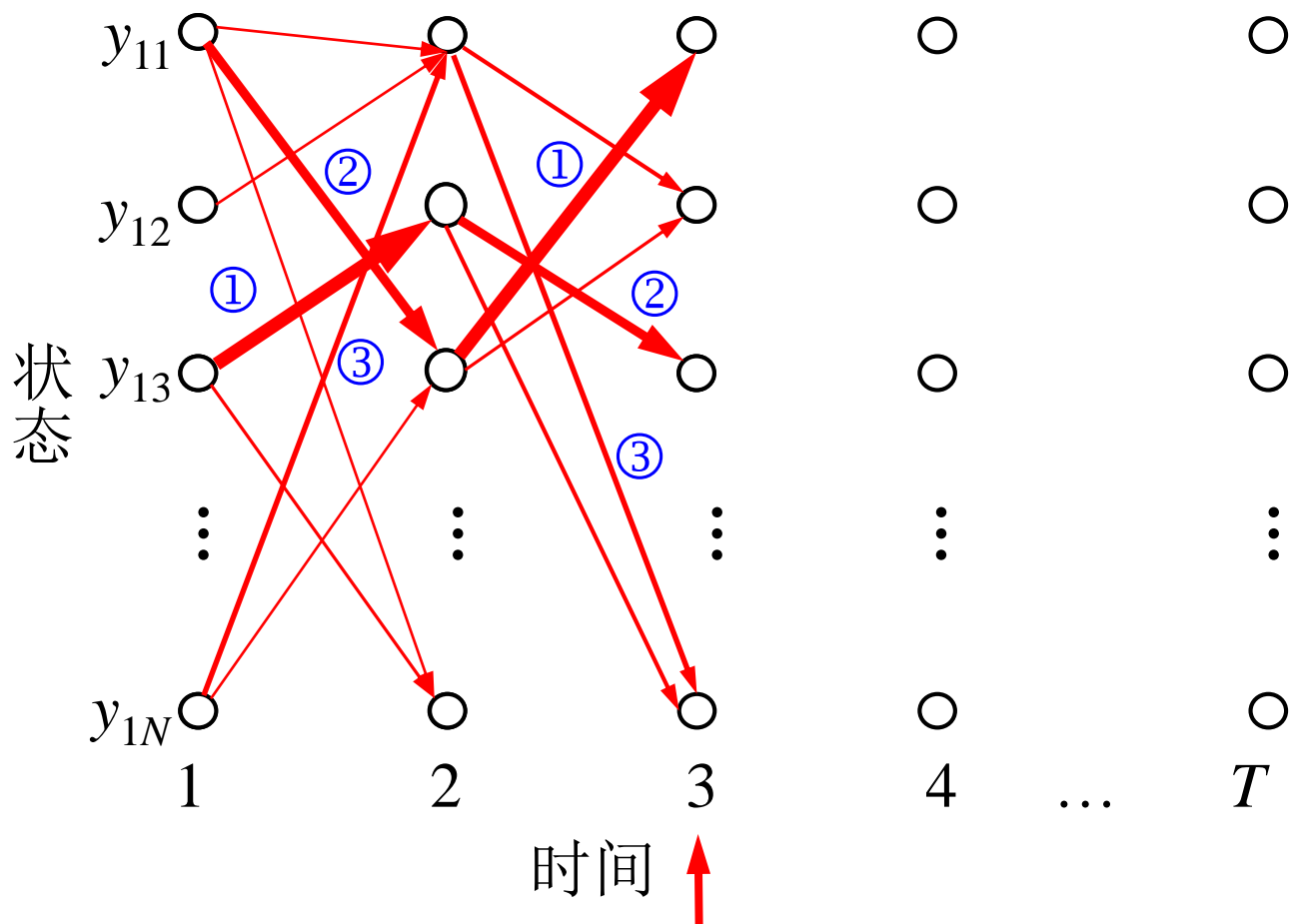
图解  
Viterbi  
搜索  
过程

剪枝策略:

①  $\delta_t(j) \geq \Delta$

②  $NPath \leq \sigma$

(3)



# 1. CRFs 定义

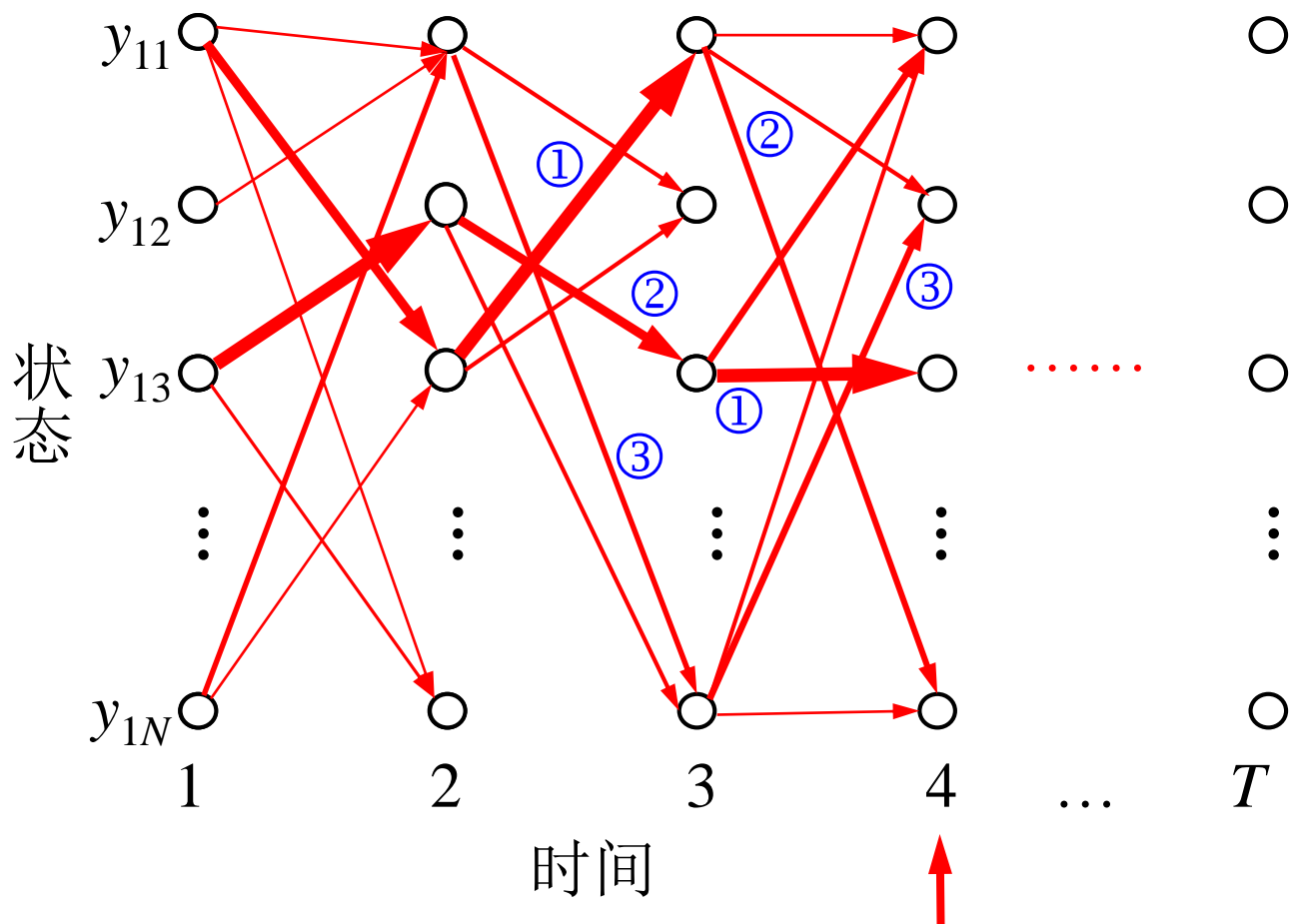
图解  
Viterbi  
搜索  
过程

剪枝策略:

①  $\delta_t(j) \geq \Delta$

②  $NPath \leq \sigma$

(3)





# 1. CRFs 定义

## ● Viterbi 算法描述

(1) 初始化:  $\delta_1(i)$ ,  $1 \leq i \leq N$

(2) 递推计算:

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) + \alpha_{ij}], \quad 2 \leq t \leq T, \quad 1 \leq j \leq N$$

(3) 结束:  $\hat{Y}_T = \arg \max_{1 \leq i \leq N} [\delta_T(i)]$

算法的时间  
复杂度:  $O(N^2T)$



# 本章内容

---

1. CRFs 定义

➡ 2. 应用举例

3. 习题



## 2. 应用举例

### ◆ 由字构词的汉语分词方法

基于字标注的分词方法(Character-based tagging)

#### ● 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)。一般而言，每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)。

该方法由N. Xue (薛念文) 和 S. Converse 提出，首篇论文发表在2002年第一届国际计算语言学学会(ACL)汉语特别兴趣小组 SIGHAN(<http://sighan.cs.uchicago.edu/>) 组织的汉语分词评测研讨会上[Xue and Converse, 2002]。





## 2. 应用举例

例如：乒乓球拍卖完了。

(1) 乒乓球/ 拍/ 卖/ 完/ 了/ 。

(2) 乒乓/ 球拍/ 卖/ 完/ 了/ 。

(3) 乒乓球/ 拍卖/ 完/ 了/ 。

第(3)种切分结果的字位标记形式：

乒/B 兵/M 球/E 拍/B 卖/E 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。

## 2. 应用举例

乒/B 乒/M 球/E 拍/S 卖 完了。

↑ B, E, M, S ?

- 当前字  $x_t$  的前后  $n$  个字  $x_{t\pm n}$  (下面假设  $n=1$ )
- 当前字左边第一个字的标记  $y_{t-1}$
- 当前字的标记  $y_t$

.....

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

分词结果: 乒/B 乒/M 球/E 拍/B 卖/E 完/S 了/S/。/S



## 2. 应用举例

### ① 特征选取

- 一元特征（状态函数）：当前字、当前字的前一个字、当前字的后一个字

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

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

.....

乒/B 乒/M 球/E 拍/S 卖/? 完了。



## 2. 应用举例

➤ 二元特征（转移函数）：

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

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

.....

乒/B 乒/M 球/E 拍/S 卖/? 完了。

假设共选出4个特征：

(a)  $F_1(x_{t-1}, x_t, y_t)$

(b)  $F_2(x_t, y_t)$

(c)  $F_3(x_t, x_{t+1}, y_t)$

(d)  $F_4(y_{t-1}, y_t)$

② 参数训练：获得每个特征的权重  $\lambda_i$  ( $1 \leq i \leq 4$ )。



## 2. 应用举例

③解码:

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了

第1步: 计算第1个字“乒”的标记分数(以标记‘B’为例):

$$\delta_1^B = \delta_0 + \alpha_{\text{path}}$$

$$\delta_0 = 1$$

$$\alpha_{\text{path}} = \lambda_1 F_1(x_0 = \text{'null'}, x_1 = \text{'乒'}, y_1 = \text{'B'})$$

$$+ \lambda_2 F_2(x_1 = \text{'乒'}, y_1 = \text{'B'})$$

$$+ \lambda_3 F_3(x_1 = \text{'乒'}, x_2 = \text{'乓'}, y_1 = \text{'B'})$$

$$+ \lambda_4 F_4(y_0 = \text{'null'}, y_1 = \text{'B'})$$

前一个字为空，当前字“乒”被标记为B的得分。

当前字“乒”被标记为B。

当前字“乒”，且后一个字为“乓”，当前字被标记为B。

前一个字的标签为空，当前字被标记为B。

同样可计算出第1个字“乒”被标记为‘S’的得分。

## 2. 应用举例

第2步：计算第2个字“兵”的标记分数（以标记M为例）。

首先计算在第1个字“兵”的标记为“B”时，第2个字“兵”的标记为“M”的分数：

$$\delta_2^{BM} = \delta_1('B') + \alpha_{\text{path}}$$

$$\begin{aligned} \alpha_{\text{path}} = & \lambda_1 F_1(x_1 = \text{'兵'}, x_2 = \text{'兵'}, y_2 = \text{'M'}) \\ & + \lambda_2 F_2(x_2 = \text{'兵'}, y_2 = \text{'M'}) \\ & + \lambda_3 F_3(x_2 = \text{'兵'}, x_3 = \text{'球'}, y_2 = \text{'M'}) \\ & + \lambda_4 F_4(y_1 = \text{'B'}, y_2 = \text{'M'}) \end{aligned}$$

然后计算在第1个字“兵”的标记为‘S’时，第2个字“兵”的标记为‘M’的分数：

$$\delta_2^{SB} = \delta_1('S') + \alpha_{\text{path}}$$

$$\begin{aligned} \alpha_{\text{path}} = & \lambda_1 F_1(x_1 = \text{'兵'}, x_2 = \text{'兵'}, y_2 = \text{'M'}) \\ & + \lambda_2 F_2(x_2 = \text{'兵'}, y_2 = \text{'M'}) \\ & + \lambda_3 F_3(x_2 = \text{'兵'}, x_3 = \text{'球'}, y_2 = \text{'M'}) \\ & + \lambda_4 F_4(y_1 = \text{'S'}, y_2 = \text{'M'}) \end{aligned}$$

等等其它情况……

## 2. 应用举例

第2步：确定第2个字“兵”被标记为‘M’的得分：

$$\delta_2(M) = \max \{ \delta_2^{BM}, \delta_2^{SM}, \delta_2^{MM}, \delta_2^{EM} \}$$

以此类推，计算第2个字“兵”的标记分别为‘S’、‘B’和‘E’的分数  $\delta_2(S)$ 、 $\delta_2(B)$  和  $\delta_2(E)$ 。

第3步：循环

根据第2个字“兵”的标记分数，计算第3个字“球”的标记分数……根据第6个字“完”的标记分数计算第7个字“了”的标记分数。

**结束**：最终选择分数最高的路径，然后以该路径的标记点为起始点回溯，得到整个句子的路径标记序列。解码完毕。



## 2. 应用举例

### ◆CRFs 的开源代码:

- CRF++ (C++版):

<http://crfpp.googlecode.com/svn/trunk/doc/index.html>

- CRFSuite (C语言版):

<http://www.chokkan.org/software/crfsuite/>

- MALLET (Java版, 通用的自然语言处理工具包, 包括分类、序列标注等机器学习算法):

<http://mallet.cs.umass.edu/>

- NLTK (Python版, 通用的自然语言处理工具包, 很多工具是从MALLET中包装转成的Python接口): <http://nltk.org/>





## 2. 应用举例

---

### ◆CRFs 的经典文献:

- [1] J. Lafferty et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.  
*Proc.ICML'2001*, pages 282-289
- [2] H. M. Wallach. Conditional Random Fields: An Introduction.  
*CIS Technical Report MS-CIS-04-21*, Univ. of Penn., 2004



# 本章小结

---

## ◆条件随机场模型 (CRFs)

模型提出的基本思想

模型定义

## ◆CRFs 实现

(1) 特征选择

(2) 参数( $\lambda$ )训练

(3) 解码搜索最优标记序列: Viterbi 算法

## ◆CRFs 应用举例

以汉语分词为例。



# 本章内容

---

1. 条件随机场定义

2. 应用举例

 3. 习题



### 3. 习题

1. 利用北京大学标注的《人民日报》1998年1月份的分词和词性标注语料，实现基于CRFs的汉语分词方法。
2. 利用北京大学标注的上述语料，实现基于最大熵分类器的由字构词的汉语分词方法，并对切分结果与基于CRFs模型得到的切分结果进行对比分析。
3. 利用北京大学标注的上述语料，实现基于 $n$ 元语法的汉语分词方法，并对切分结果与基于CRFs模型得到的切分结果进行对比分析。
4. 请对比分析基于最大熵的分类模型与CRFs模型。

谢谢!

*Thanks!*

