条件随机场

Conditional Random Fields (CRF)

2020/1/16

背景

Conditional random fields: Probabilistic models for segmenting and labeling sequence data

J Lafferty, A McCallum, FCN Pereira - 2001 - repository.upenn.edu

We present conditional random fields, a framework for building probabilistic models to segment and label sequence data. Conditional random fields offer several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Conditional random fields also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased ...

☆ 99 被引用次数: 13204 相关文章 所有 77 个版本 »

John Lafferty

http://www.cs.cmu.edu/afs/cs/usr/lafferty/www/people.html

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data， ICML2001

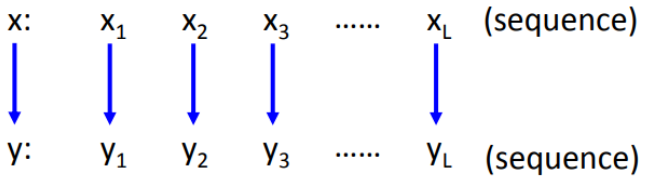
应用场景： sequence labeling （NLP, Voice Recognition）

$$f : X \rightarrow Y$$

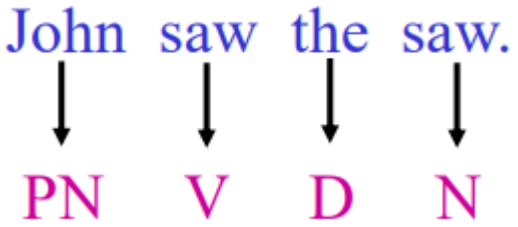
Sequence Sequence

- **Name entity recognition**

Harry Potter is a student of Hogwarts and lived on Privet Drive.



- **POS tagging**

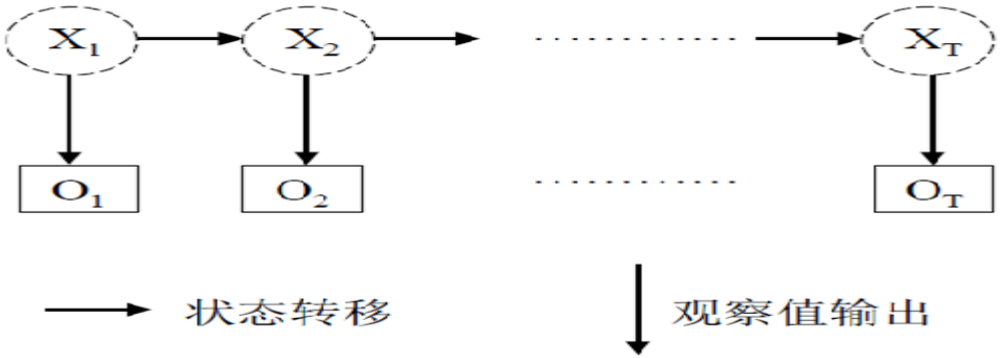


动机

再看HMM：

两个基本假设：

1. 独立输出假设，输出值是严格独立的
2. 时刻t的状态只与t-1时状态有关



Drawbacks：

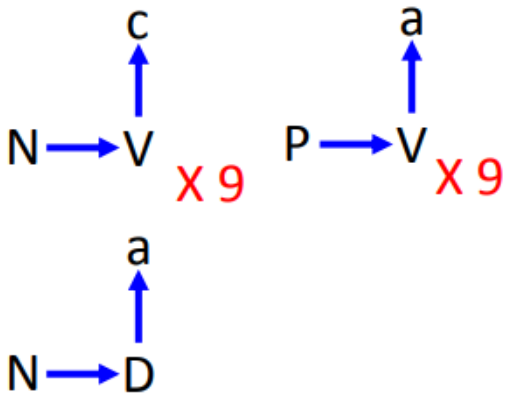
1、The (x, y) never seen in the training data can have large probability *P(x, y)*.

2、观测值独立假设

动机

HMM Drawbacks：

1、The (x, y) never seen in the training data can have large probability *P(x, y)*.

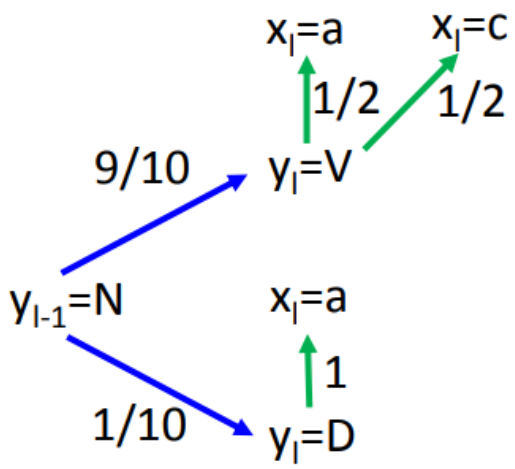


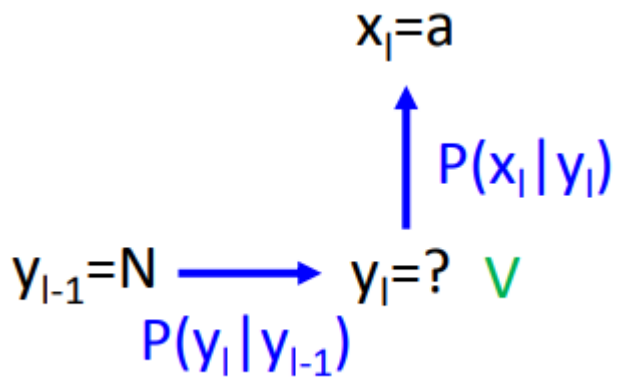
Transition probability:

P(V|N)=9/10　P(D|N)=1/10 ……

Emission probability:

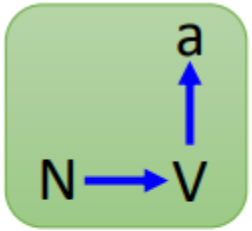
P(a|V)=1/2　　P(a|D)=1 ……





P(V|N)*P(a | V)=0.9*0.5=0.45

P(D|N)*P(a | D)=0.1*1=0.1

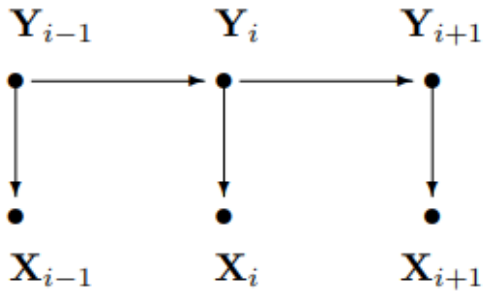


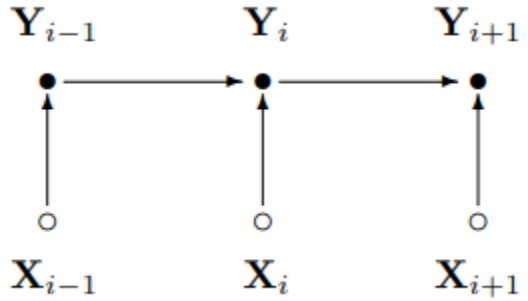
High probability for HMM

动机

HMM Drawbacks：

1、观测值独立假设





HMM

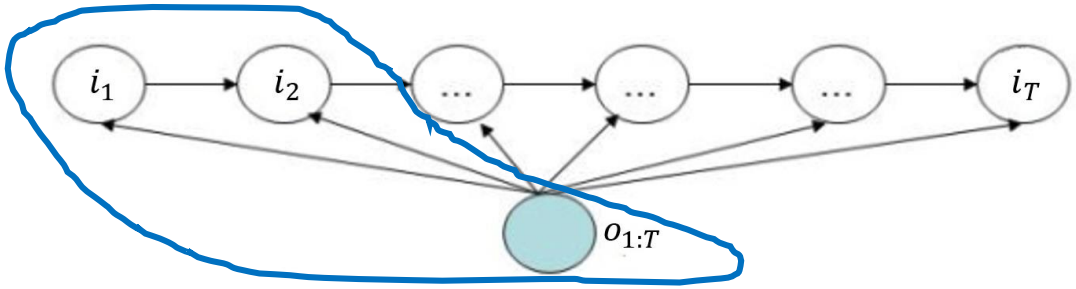
MEMM

HMM：给定Y_i，X_i和X_{i-1}相互独立

MEMM：给定Y_i，X_i和X_{i-1}必不独立

动机

MEMM：



局部归一化

Drawbacks：

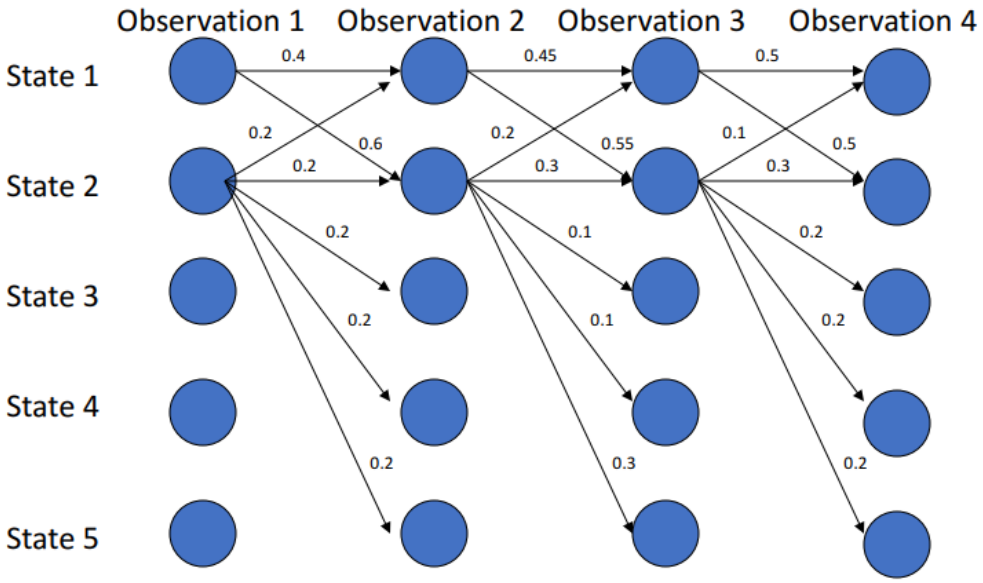
1、Label Bias Problem （标注偏置问题）

即 HMM (MEMM)倾向于选择拥有更少转移分支的状态。

动机

MEMM Drawbacks：

2、Label Bias Problem （标注偏置问题）



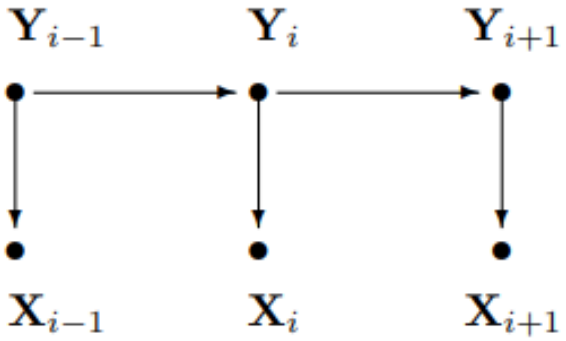
从上图能够看出状态1倾向于转移到状态2，

状态2倾向于停留在状态2.

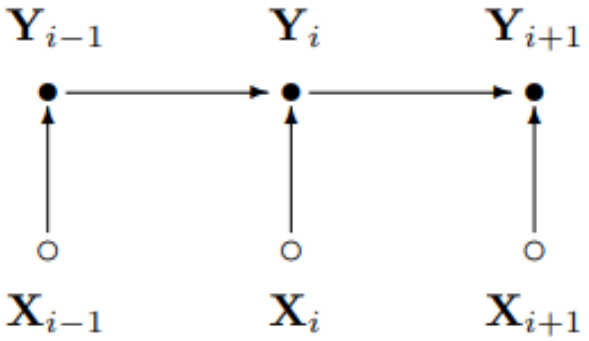
- P(1->1->1->1) = 0.4*0.45*0.5 = 0.090
- P(2->2->2->2) = 0.2*0.3*0.3 = 0.018
- P(1->2->2->2) = 0.6*0.3*0.3 = 0.054
- P(2->1->1->1) = 0.2*0.45*0.5 = 0.450

根据维特比算法，HMM会选择 1->1->1->1 这条路径

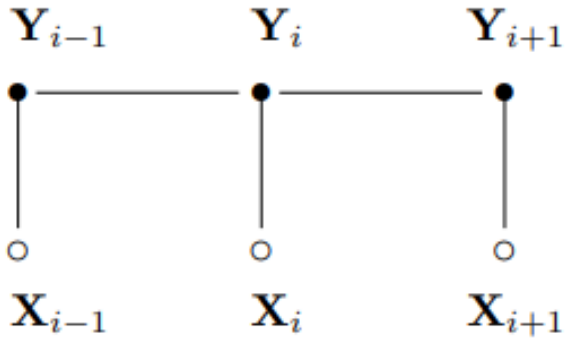
HMM->MEMM->CRF



HMM



MEMM

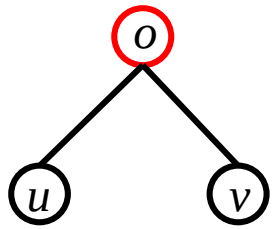


CRF

概率无向图模型（马尔可夫随机场）

定义：设有联合概率分布$P(Y)$,由无向图$G=(V, E)$表示，在图G中，结点表示随机变量，边表示随机变量之间的依赖关系, 如果联合概率分布$P(Y)$满足成对、局部或全局马尔可夫性，就称此联合概率分布为概率无向图模型，或马尔可夫随机场。

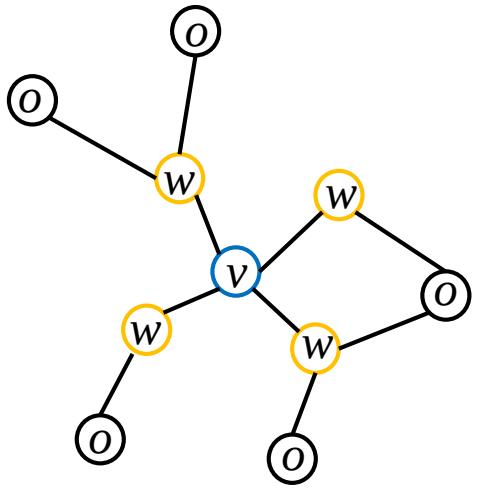
- 成对马尔可夫性(pairwise Markov property)
- 局部马尔可夫性(local Markov properly)
- 全局马尔可夫性(global Markov property)



$$P(Y_u, Y_v \mid Y_o) = P(Y_u \mid Y_o) * P(Y_v \mid Y_o)$$

概率无向图模型

- 成对马尔可夫性(pairwise Markov property)
- 局部马尔可夫性(local Markov properly)
- 全局马尔可夫性(global Markov property)



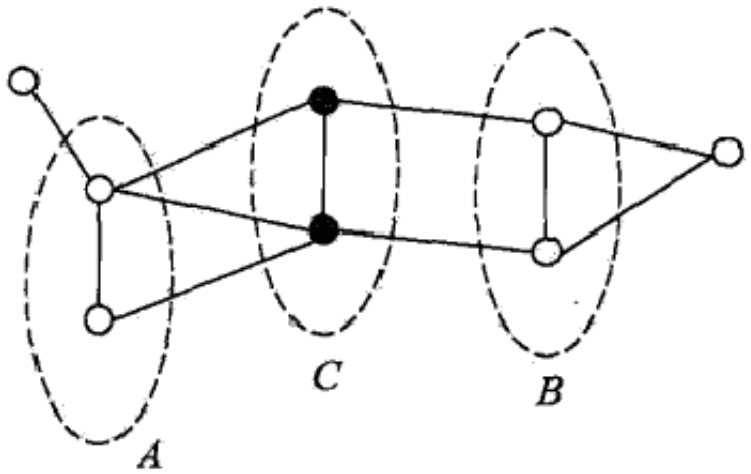
$$P(Y_v, Y_o \mid Y_w) = P(Y_v \mid Y_w) * P(Y_o \mid Y_w)$$

$$P(Y_v \mid Y_w) = P(Y_v \mid Y_w, Y_o)$$

概率无向图模型

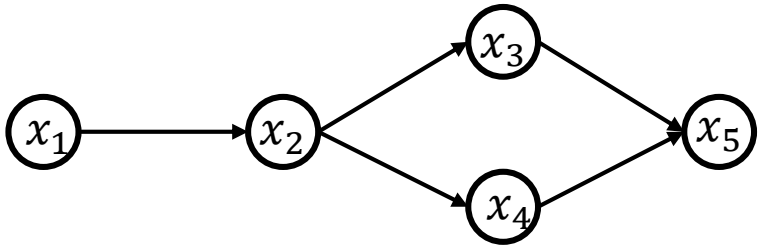
- 成对马尔可夫性(pairwise Markov property)
- 局部马尔可夫性(local Markov properly)
- 全局马尔可夫性(global Markov property)

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



概率无向图模型的联合概率

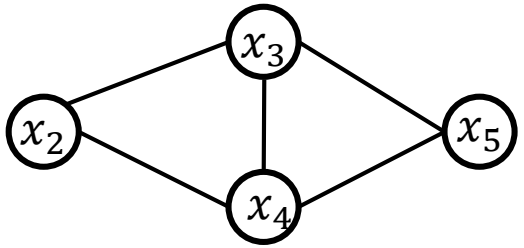
有向图：



联合概率：

$$P(x_1, x_2, \dots x_5) = P(x_1) * P(x_2|x_1) * p(x_3|x_2) * p(x_4|x_2) * p(x_5|x_3, x_4)$$

无向图：



联合概率：？

因子分解

无向图因子分解

因子分解：将整体的联合概率表示为其最大团上的随机变量的函数的乘积

定义 11.2（团与最大团） 无向图 G 中任何两个结点均有边连接的结点子集称为团（clique）. 若 C 是无向图 G 的一个团，并且不能再加进任何一个 G 的结点使其成为一个更大的团，则称此 C 为最大团（maximal clique）.

图 11.3 表示由 4 个结点组成的无向图. 图中由 2 个结点组成的团有 5 个：$\{Y_1, Y_2\}$，$\{Y_2, Y_3\}$，$\{Y_3, Y_4\}$，$\{Y_4, Y_2\}$ 和 $\{Y_1, Y_3\}$. 有 2 个最大团：$\{Y_1, Y_2, Y_3\}$ 和 $\{Y_2, Y_3, Y_4\}$. 而 $\{Y_1, Y_2, Y_3, Y_4\}$ 不是一个团，因为 Y_1 和 Y_4 没有边连接.

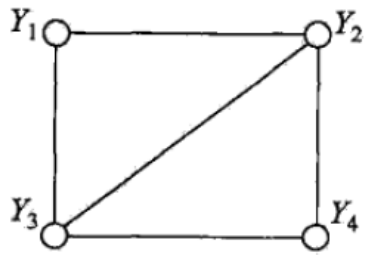


图 11.3　无向图的团和最大团

无向图因子分解

给定概率无向图模型，设其无向图为G，C为G上的最大团，Y_c表示C对应的随机变量，那么概率无向图模型的联合概率分布$P(Y)$可写作图中所有最大团C上的函数$\psi_C(Y_C)$的乘积形式，即

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

Z 为规范化因子，保证$P(Y)$ 构成一个概率分布。

$$Z = \sum_Y \prod_C \psi_C(Y_C)$$

函数$\psi_C(Y_C)$称为势函数，要求是严格正的，所以通常定义为指数函数:

$$\psi_C(Y_C) = e^{-E(Y_C)}$$

为什么可以因子分解？ Hammersley-Clifford定理

条件随机场CRF

定义 11.3（条件随机场） 设 X 与 Y 是随机变量，$P(Y|X)$ 是在给定 X 的条件下 Y 的条件概率分布．若随机变量 Y 构成一个由无向图 $G=(V,E)$ 表示的马尔可夫随机场，即

$$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v) \tag{11.8}$$

对任意结点 v 成立，则称条件概率分布 $P(Y|X)$ 为条件随机场．式中 $w\sim v$ 表示在图 $G=(V,E)$ 中与结点 v 有边连接的所有结点 w，$w\neq v$ 表示结点 v 以外的所有结点，Y_v，Y_u 与 Y_w 为结点 v，u 与 w 对应的随机变量．

定义 11.4（线性链条件随机场） 设 $X=(X_1,X_2,\cdots,X_n)$，$Y=(Y_1,Y_2,\cdots,Y_n)$ 均为线性链表示的随机变量序列，若在给定随机变量序列 X 的条件下，随机变量序列 Y 的条件概率分布 $P(Y|X)$ 构成条件随机场，即满足马尔可夫性

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

$$i=1,2,\cdots,n\ （在 i=1 和 n 时只考虑单边） \tag{11.9}$$

则称 $P(Y|X)$ 为线性链条件随机场．在标注问题中，X 表示输入观测序列，Y 表示对应的输出标记序列或状态序列．

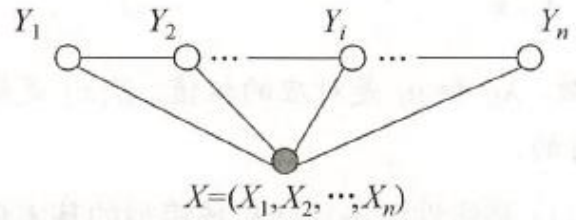


图 11.4　线性链条件随机场

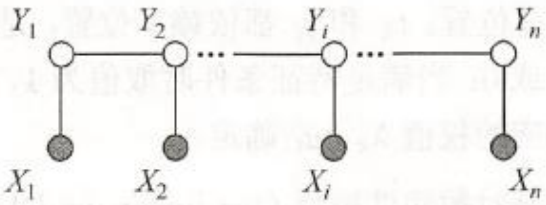


图 11.5　X 和 Y 有相同的图结构的线性链条件随机场

条件随机场的参数化形式

判别模型→对 $p(y|x)$ 建模

定理 11.2（线性链条件随机场的参数化形式） 设 $P(Y|X)$ 为线性链条件随机场，则在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率具有如下形式：

转移特征　　状态特征

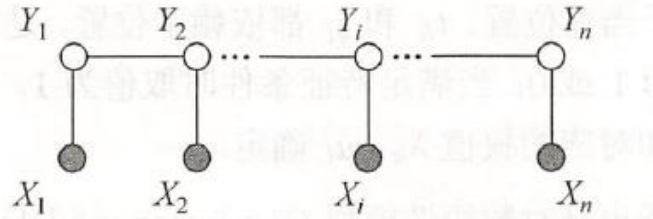
$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \tag{11.10}$$

其中，

$$Z(x) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \tag{11.11}$$

式中，t_k 和 s_l 是特征函数，λ_k 和 μ_l 是对应的权值。$Z(x)$ 是规范化因子，求和是在所有可能的输出序列上进行的。

式 (11.10) 和式 (11.11) 是线性链条件随机场模型的基本形式，表示给定输入序列 x，对输出序列 y 预测的条件概率。式 (11.10) 和式 (11.11) 中，t_k 是定义在边上的特征函数，称为转移特征，依赖于当前和前一个位置；s_l 是定义在结点上的特征函数，称为状态特征，依赖于当前位置。t_k 和 s_l 都依赖于位置，是局部特征函数。通常，特征函数 t_k 和 s_l 取值为 1 或 0；当满足特征条件时取值为 1，否则为 0。条件随机场完全由特征函数 t_k，s_l 和对应的权值 λ_k，μ_l 确定。



CRF也是对数线性模型

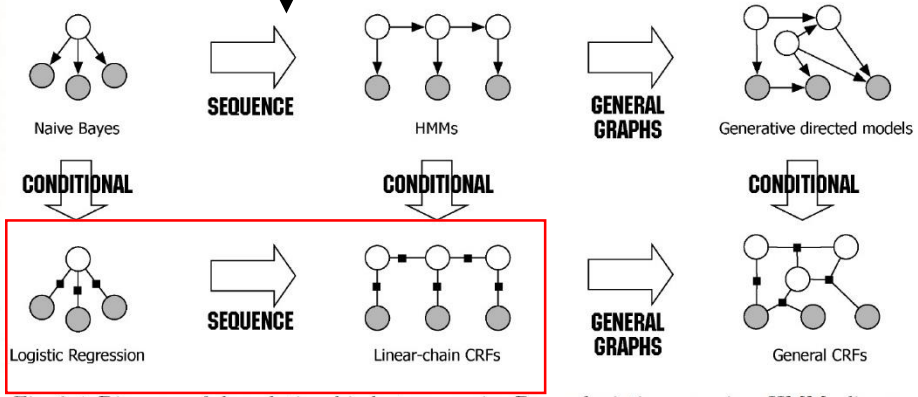


Fig. 2.4 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.

例子

例 设有一标注问题：输入观测序列为 $X=(X_1,X_2,X_3)$，输出标记序列为 $Y=(Y_1,Y_2,Y_3)$，Y_1,Y_2,Y_3 取值于 $\mathcal{Y}=\{1,2\}$。

假设特征 t_k,s_l 和对应的权值 λ_k,μ_l 如下：

$$\begin{cases} t_1(y_1=1,y_2=2,x,2)=1 \\ t_1(y_2=1,y_3=2,x,3)=1 \end{cases}$$

$$t_1=t_1(y_{i-1}=1,y_i=2,x,i),\quad i=2,3,\quad \lambda_1=1$$

这里只注明特征取值为 1 的条件，取值为 0 的条件省略，即

$$t_1(y_{i-1},y_i,x,i)=\begin{cases} 1, & y_{i-1}=1,y_i=2,x,i,(i=2,3) \\ 0, & \text{其他} \end{cases}$$

下同。

$$\begin{aligned}
t_2&=t_2(y_1=1,y_2=1,x,2) && \lambda_2=0.6 \\
t_3&=t_3(y_2=2,y_3=1,x,3) && \lambda_3=1 \\
t_4&=t_4(y_1=2,y_2=1,x,2), && \lambda_4=1 \\
t_5&=t_5(y_2=2,y_3=2,x,3), && \lambda_5=0.2 \\
s_1&=s_1(y_1=1,x,1), && \mu_1=1 \\
s_2&=s_2(y_i=2,x,i),\ i=1,2 && \mu_2=0.5 \\
s_3&=s_3(y_i=1,x,i),\ i=2,3 && \mu_3=0.8 \\
s_4&=s_4(y_3=2,x,3), && \mu_4=0.5
\end{aligned}$$

对给定的观测序列 x，求标记序列为 $y=(y_1,y_2,y_3)=(1,2,2)$ 的非规范化条件概率（即没有除以规范化因子的条件概率）。

解 由式 (11.10)，线性链条件随机场模型为

$$P(y|x)\propto\exp\left[\sum_{k=1}^{5}\lambda_k\sum_{i=2}^{3}t_k(y_{i-1},y_i,x,i)+\sum_{k=1}^{4}\mu_k\sum_{i=1}^{3}s_k(y_i,x,i)\right]$$

$$\lambda_1[t_1(y_1=1,y_2=2,x,2)+t_1(y_2=2,y_3=2,x,3)]+\lambda_2[t_2(y_1=1,y_2=2,x,2)+t_2(y_2=2,y_3=2,x,3)]+\lambda_3[t_3(y_1=1,y_2=2,x,2)+t_3(y_2=2,y_3=2,x,3)]+\lambda_4[t_4(y_1=1,y_2=2,x,2)+t_4(y_2=2,y_3=2,x,3)]+\lambda_5[t_5(y_1=1,y_2=2,x,2)+t_5(y_2=2,y_3=2,x,3)]+\mu_1[s_1(y_1=1,x,1)+s_1(y_2=2,x,2)+s_1(y_3=2,x,3)]+\mu_2[s_2(y_1=1,x,1)+s_2(y_2=2,x,2)+s_2(y_3=2,x,3)]+\mu_3[s_3(y_1=1,x,1)+s_3(y_2=2,x,2)+s_3(y_3=2,x,3)]+\mu_4[s_4(y_1=1,x,1)+s_4(y_2=2,x,2)+s_4(y_3=2,x,3)]=1+0.2+1+0.5+0.5=3.2$$

$$P(y_1=1,y_2=2,y_3=2|x)\propto\exp(3.2)$$

CRF的简化形式

为简便起见，首先将转移特征和状态特征及其权值用统一的符号表示。设有 K_1 个转移特征，K_2 个状态特征，$K = K_1 + K_2$，记

$$f_k(y_{i-1}, y_i, x, i) = \begin{cases} t_k(y_{i-1}, y_i, x, i), & k = 1, 2, \cdots, K_1 \\ s_l(y_i, x, i), & k = K_1 + l;\ l = 1, 2, \cdots, K_2 \end{cases} \tag{11.12}$$

然后，对转移与状态特征在各个位置 i 求和，记作

$$f_k(y, x) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, x, i), \quad k = 1, 2, \cdots, K \tag{11.13}$$

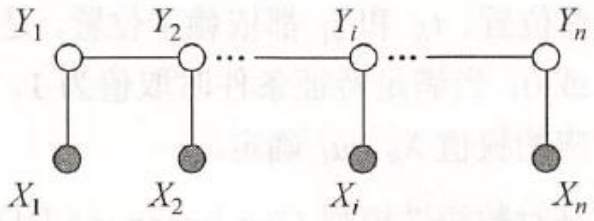
用 w_k 表示特征 $f_k(y, x)$ 的权值，即

$$w_k = \begin{cases} \lambda_k, & k = 1, 2, \cdots, K_1 \\ \mu_l, & k = K_1 + l; l = 1, 2, \cdots, K_2 \end{cases} \tag{11.14}$$

于是，条件随机场 (11.10)~(11.11) 可表示为

$$P(y|x) = \frac{1}{Z(x)} \exp \sum_{k=1}^{K} w_k f_k(y, x) \tag{11.15}$$

$$Z(x) = \sum_{y} \exp \sum_{k=1}^{K} w_k f_k(y, x) \tag{11.16}$$



对给定的观测序列 x，求标记序列为 $y = (y_1, y_2, y_3) = (1, 2, 2)$ 的非规范化条件概率（即没有除以规范化因子的条件概率）。

解 由式 (11.10)，线性链条件随机场模型为

$$P(y|x) \propto \exp\left[\sum_{k=1}^{5} \lambda_k \sum_{i=2}^{3} t_k(y_{i-1}, y_i, x, i) + \sum_{k=1}^{4} \mu_k \sum_{i=1}^{3} s_k(y_i, x, i)\right]$$

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \tag{11.10}$$

其中，

$$Z(x) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \tag{11.11}$$

CRF的简化形式

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

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

若以 w 表示权值向量，即

$$w = (w_1, w_2, \cdots, w_K)^{\mathrm{T}} \tag{11.17}$$

以 $F(y, x)$ 表示全局特征向量，即

$$F(y, x) = (f_1(y, x), f_2(y, x), \cdots, f_K(y, x))^{\mathrm{T}} \tag{11.18}$$

则条件随机场可以写成向量 w 与 $F(y, x)$ 的内积的形式：

$$P_w(y|x) = \frac{\exp(w \cdot F(y, x))}{Z_w(x)} \tag{11.19}$$

其中，

$$Z_w(x) = \sum_{y} \exp(w \cdot F(y, x)) \tag{11.20}$$

CRF的矩阵形式

对观测序列 x 的每一个位置 $i=1,2,\cdots,n+1$，由于 y_{i-1} 和 y_i 在 m 个标记中取值，可以定义一个 m 阶矩阵随机变量

$$M_i(x)=[M_i(y_{i-1},y_i|x)] \tag{11.21}$$

矩阵随机变量的元素为

$$M_i(y_{i-1},y_i|x)=\exp\left(W_i(y_{i-1},y_i|x)\right) \tag{11.22}$$

$$W_i(y_{i-1},y_i|x)=\sum_{k=1}^{K}w_k f_k(y_{i-1},y_i,x,i) \tag{11.23}$$

这里 w_k 和 f_k 分别由式 (11.14) 和式 (11.12) 给出，y_{i-1} 和 y_i 是标记随机变量 Y_{i-1} 和 Y_i 的取值。

这样，给定观测序列 x，相应标记序列 y 的非规范化概率可以通过该序列 $n+1$ 个矩阵的适当元素的乘积 $\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x)$ 表示。于是，条件概率 $P_w(y|x)$ 是

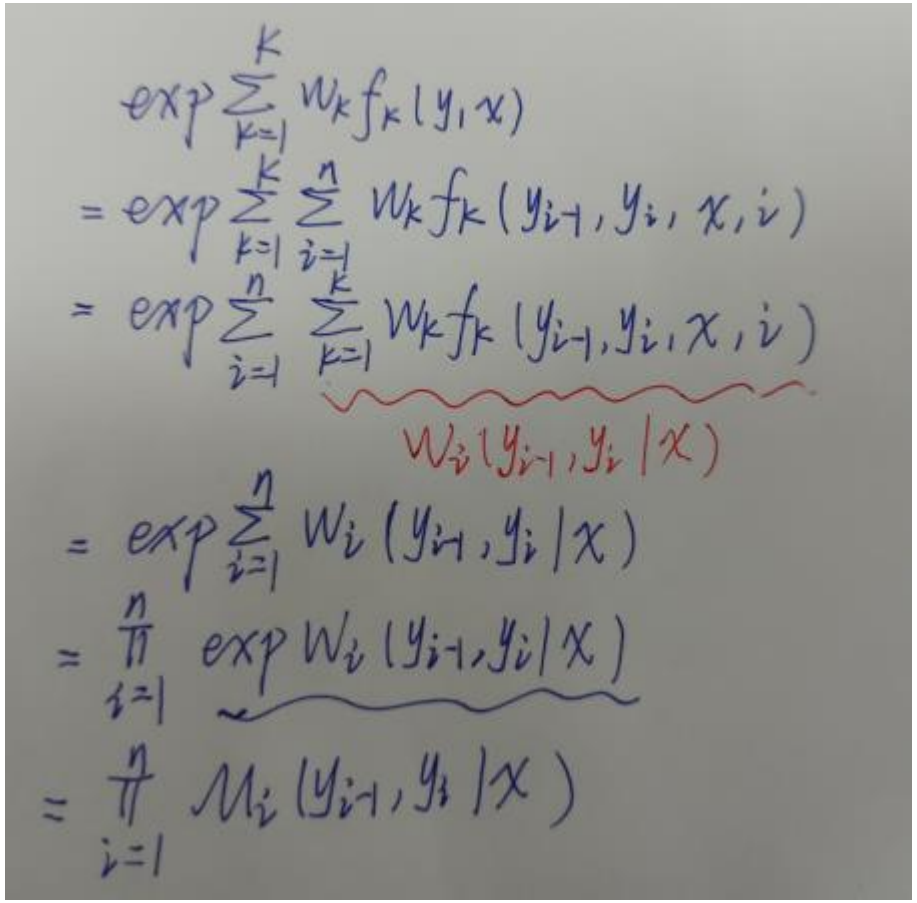
$$P_w(y|x)=\frac{1}{Z_w(x)}\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x) \tag{11.24}$$

其中，$Z_w(x)$ 为规范化因子，是 $n+1$ 个矩阵的乘积的 (start, stop) 元素，即

$$Z_w(x)=[M_1(x)M_2(x)\cdots M_{n+1}(x)]_{\text{start,stop}} \tag{11.25}$$

注意，$y_0=\text{start}$ 与 $y_{n+1}=\text{stop}$ 表示开始状态与终止状态，规范化因子 $Z_w(x)$ 是以 start 为起点 stop 为终点通过状态的所有路径 $y_1y_2\cdots y_n$ 的非规范化概率 $\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x)$ 之和。下面的例子说明了这一事实。

简化形式->矩阵形式



CRF的矩阵形式-例子

例 11.2 给定一个由图 11.6 所示的线性链条件随机场，观测序列 x，状态序列 y，$i=1,2,3$，$n=3$，标记 $y_i \in \{1,2\}$，假设 $y_0=\text{start}=1$，$y_4=\text{stop}=1$，各个位置的随机矩阵 $M_1(x)$，$M_2(x)$，$M_3(x)$，$M_4(x)$ 分别是

$$M_1(x)=\begin{bmatrix} a_{01} & a_{02} \\ 0 & 0 \end{bmatrix}, \quad M_2(x)=\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$M_3(x)=\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}, \quad M_4(x)=\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

试求状态序列 y 以 start 为起点 stop 为终点所有路径的非规范化概率及规范化因子。

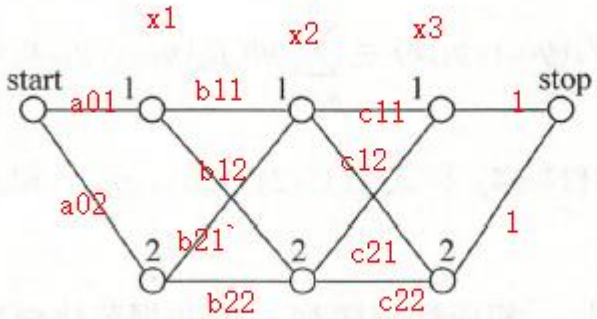


图 11.6 状态路径

解 首先计算图 11.6 中从 start 到 stop 对应于 $y=(1,1,1)$，$y=(1,1,2)$，$\cdots$，$y=(2,2,2)$ 各路径的非规范化概率分别是

$$a_{01}b_{11}c_{11}, \quad a_{01}b_{11}c_{12}, \quad a_{01}b_{12}c_{21}, \quad a_{01}b_{12}c_{22}$$

$$a_{02}b_{21}c_{11}, \quad a_{02}b_{21}c_{12}, \quad a_{02}b_{22}c_{21}, \quad a_{02}b_{22}c_{22}$$

$$Z_w(x)=[M_1(x)M_2(x)\cdots M_{n+1}(x)]_{\text{start,stop}} \tag{11.25}$$

然后按式 (11.25) 求规范化因子。通过计算矩阵乘积 $M_1(x)M_2(x)M_3(x)M_4(x)$ 可知，其第 1 行第 1 列的元素为

$$a_{01}b_{11}c_{11}+a_{02}b_{21}c_{11}+a_{01}b_{12}c_{21}+a_{02}b_{22}c_{22}$$
$$+a_{01}b_{11}c_{12}+a_{02}b_{21}c_{12}+a_{01}b_{12}c_{22}+a_{02}b_{22}c_{21}$$

恰好等于从 start 到 stop 的所有路径的非规范化概率之和，即规范化因子 $Z(x)$。 ■

CRF的三个问题

- 概率计算问题
  - ✓ 前向-后向算法
- 参数学习问题
  - ✓ 改进的迭代尺度算法
- 预测算法（解码）
  - ✓ 维特比算法

CRF的三个问题1：概率计算问题

目标：给定x，计算 $P(Y_i = y_i|x)$ 和 $P(Y_{i-1} = y_{i-1}, Y_i = y_i|x)$

✓ 前向-后向算法

11.3.1 前向-后向算法

对每个指标 $i = 0, 1, \cdots, n+1$，定义前向向量 $\alpha_i(x)$：

$$\alpha_0(y|x) = \begin{cases} 1, & y = \text{start} \\ 0, & \text{否则} \end{cases} \tag{11.26}$$

递推公式为

$$\alpha_i^{\mathrm{T}}(y_i|x) = \alpha_{i-1}^{\mathrm{T}}(y_{i-1}|x)[M_i(y_{i-1}, y_i|x)], \quad i = 1, 2, \cdots, n+1 \tag{11.27}$$

又可表示为

$$\alpha_i^{\mathrm{T}}(x) = \alpha_{i-1}^{\mathrm{T}}(x) M_i(x) \tag{11.28}$$

$\alpha_i(y_i|x)$ 表示在位置 i 的标记是 y_i 并且从 1 到 i 的前部分标记序列的非规范化概率，y_i 可取的值有 m 个，所以 $\alpha_i(x)$ 是 m 维列向量。

同样，对每个指标 $i = 0, 1, \cdots, n+1$，定义后向向量 $\beta_i(x)$：

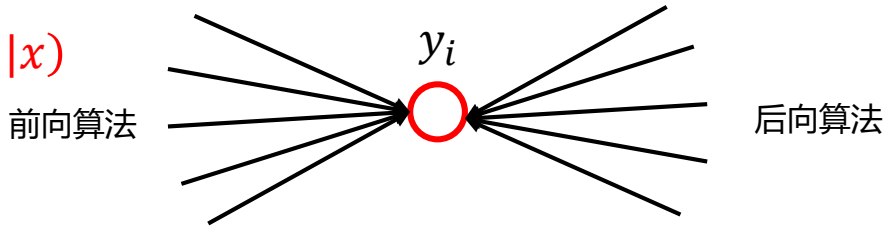
$$\beta_{n+1}(y_{n+1}|x) = \begin{cases} 1, & y_{n+1} = \text{stop} \\ 0, & \text{否则} \end{cases} \tag{11.29}$$

$$\beta_i(y_i|x) = [M_{i+1}(y_i, y_{i+1}|x)]\beta_{i+1}(y_{i+1}|x) \tag{11.30}$$

又可表示为

$$\beta_i(x) = M_{i+1}(x)\beta_{i+1}(x) \tag{11.31}$$

$\beta_i(y_i|x)$ 表示在位置 i 的标记为 y_i 并且从 $i+1$ 到 n 的后部分标记序列的非规范化概率。



按照前向-后向向量的定义，很容易计算标记序列在位置 i 是标记 y_i 的条件概率和在位置 $i-1$ 与 i 是标记 y_{i-1} 和 y_i 的条件概率：

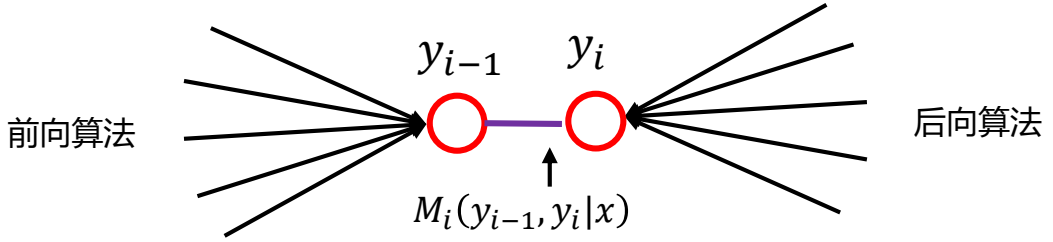
$$P(Y_i = y_i|x) = \frac{\alpha_i^{\mathrm{T}}(y_i|x)\beta_i(y_i|x)}{Z(x)} \tag{11.32}$$

$$P(Y_{i-1} = y_{i-1}, Y_i = y_i|x) = \frac{\alpha_{i-1}^{\mathrm{T}}(y_{i-1}|x) M_i(y_{i-1}, y_i|x)\beta_i(y_i|x)}{Z(x)} \tag{11.33}$$

其中，

$$Z(x) = \alpha_n^{\mathrm{T}}(x)\mathbf{1} = \mathbf{1}\beta_1(x)$$

$\mathbf{1}$ 是元素均为 1 的 m 维列向量。



CRF的三个问题2：参数学习问题

- ✓ 改进的迭代尺度算法

CRF的三个问题3：预测算法（解码）

目标：给定模型和输入序列x，求条件概率最大的输出序列y^*

条件随机场目标函数：

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

$$\begin{aligned} y^* &= \arg\max_y P_w(y|x) \\ &= \arg\max_y \frac{\exp(w \cdot F(y,x))}{Z_w(x)} \\ &= \arg\max_y \exp(w \cdot F(y,x)) \\ &= \arg\max_y (w \cdot F(y,x)) \end{aligned}$$

于是，条件随机场的预测问题成为求非规范化概率最大的最优路径问题

$$\max_y (w \cdot F(y,x))$$

这里，路径表示标记序列。其中，

$$w = (w_1, w_2, \cdots, w_K)^{\mathrm{T}}$$

$$F(y,x) = (f_1(y,x), f_2(y,x), \cdots, f_K(y,x))^{\mathrm{T}}$$

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

注意，这时只需计算非规范化概率，而不必计算概率，可以大大提高效率。为了求解最优路径，将式 (11.52) 写成如下形式：

$$\max_y \sum_{i=1}^{n} w \cdot F_i(y_{i-1}, y_i, x) \tag{11.53}$$

其中，

$$F_i(y_{i-1}, y_i, x) = (f_1(y_{i-1}, y_i, x, i), f_2(y_{i-1}, y_i, x, i), \cdots, f_K(y_{i-1}, y_i, x, i))^{\mathrm{T}}$$

是局部特征向量。

CRF的三个问题3：预测算法（解码）

目标：给定模型和输入序列x，求条件概率最大的输出序列y^*

✓ 维特比算法

算法 11.3（条件随机场预测的维特比算法）

输入：模型特征向量 $F(y,x)$ 和权值向量 w，观测序列 $x=(x_1,x_2,\cdots,x_n)$；

输出：最优路径 $y^*=(y_1^*,y_2^*,\cdots,y_n^*)$。

（1）初始化

$$\delta_1(j)=w\cdot F_1(y_0=\text{start},y_1=j,x),\quad j=1,2,\cdots,m$$

（2）递推。对 $i=2,3,\cdots,n$

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

$$\Psi_i(l)=\arg\max_{1\leqslant j\leqslant m}\{\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x)\},\quad l=1,2,\cdots,m$$

（3）终止

$$\max_y(w\cdot F(y,x))=\max_{1\leqslant j\leqslant m}\delta_n(j)$$

$$y_n^*=\arg\max_{1\leqslant j\leqslant m}\delta_n(j)$$

（4）返回路径

$$y_i^*=\Psi_{i+1}(y_{i+1}^*),\quad i=n-1,n-2,\cdots,1$$

求得最优路径 $y^*=(y_1^*,y_2^*,\cdots,y_n^*)$。

CRF的三个问题3：预测算法（解码）

目标：给定模型和输入序列x，求条件概率最大的输出序列y^*

✓ 维特比算法-例子

例　设有一标注问题：输入观测序列为 $X=(X_1,X_2,X_3)$，输出标记序列为 $Y=(Y_1,Y_2,Y_3)$，Y_1,Y_2,Y_3 取值于 $\mathcal{Y}=\{1,2\}$。

假设特征 t_k,s_l 和对应的权值 λ_k,μ_l 如下：

$$\begin{cases} t_1(y_1=1,y_2=2,x,2)=1 \\ t_1(y_2=1,y_3=2,x,3)=1 \end{cases}$$

$$t_1=t_1(y_{i-1}=1,y_i=2,x,i),\quad i=2,3,\quad \lambda_1=1$$

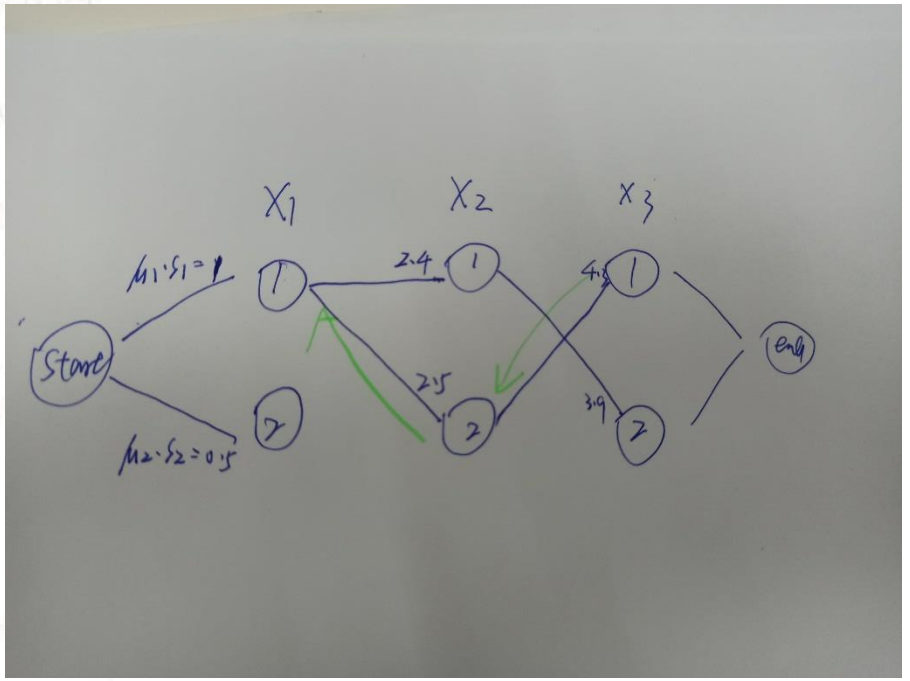
这里只注明特征取值为 1 的条件，取值为 0 的条件省略，即

$$t_1(y_{i-1},y_i,x,i)=\begin{cases} 1, & y_{i-1}=1,y_i=2,x,i,(i=2,3) \\ 0, & \text{其他} \end{cases}$$

下同。

$$\begin{aligned}
t_2&=t_2(y_1=1,y_2=1,x,2) && \lambda_2=0.6 \\
t_3&=t_3(y_2=2,y_3=1,x,3) && \lambda_3=1 \\
t_4&=t_4(y_1=2,y_2=1,x,2), && \lambda_4=1 \\
t_5&=t_5(y_2=2,y_3=2,x,3), && \lambda_5=0.2 \\
s_1&=s_1(y_1=1,x,1), && \mu_1=1 \\
s_2&=s_2(y_i=2,x,i),\ i=1,2 && \mu_2=0.5 \\
s_3&=s_3(y_i=1,x,i),\ i=2,3 && \mu_3=0.8 \\
s_4&=s_4(y_3=2,x,3), && \mu_4=0.5
\end{aligned}$$

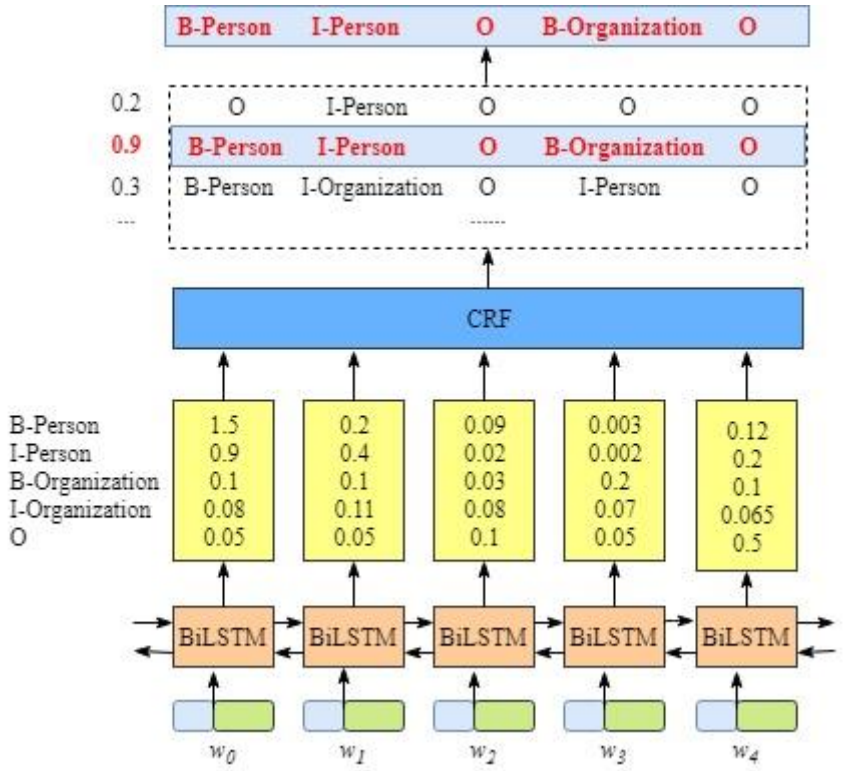
对给定的观测序列 x，求标记序列为 $y=(y_1,y_2,y_3)=(1,2,2)$ 的非规范化条件概率（即没有除以规范化因子的条件概率）。



Sequence labeling

目前深度学习解决序列标注任务常用方法包括LSTM+CRF、BiLSTM+CRF

- Bidirectional LSTM-CRF Models for Sequence Tagging



放假啦!!!