

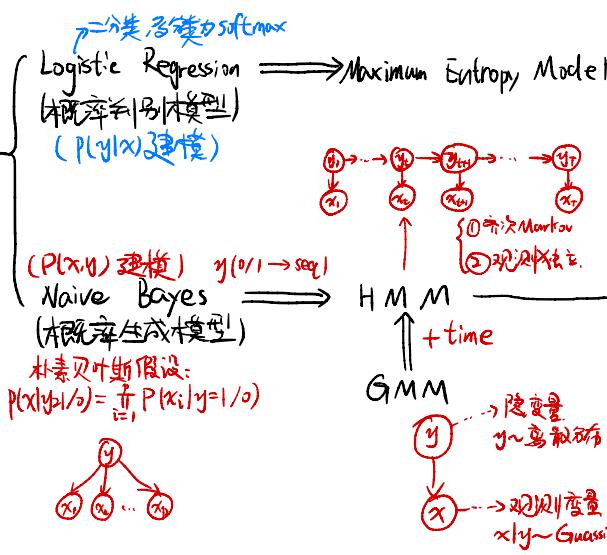
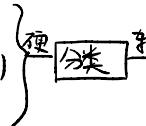
# 条件随机场 (Conditional Random Field)

## 一、背景

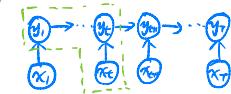
SVM

PLA (感知机)

LDA



Maximum Entropy Markov Model



$P(x_1, x_2)$

MEMM

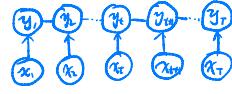
(判别)

克服了MEM的观测独立假设

问题: 标注偏差问题

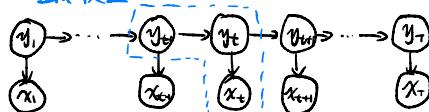
(原因: 局部化)

CRF: 解决了这个问题



## 二. HMM V.S. MEMM

HMM:



$\lambda = (\pi, A, B)$  即给定  $y_t$  的情况下,  $x_t$  与  $x_{t+1}$  无关.

- { ① 齐次一阶 Markov 假设
- ② 观测独立假设

$$P(y_t | y_{1:t-1}, x_{1:t-1}) = P(y_t | y_{t-1})$$

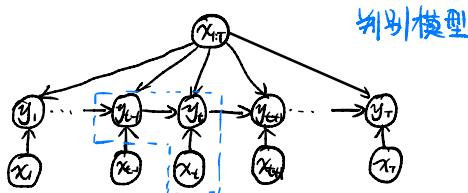
$$P(x_t | y_{1:t-1}, x_{1:t-1}) = P(x_t | y_t)$$

生成式模型的建模对象是联合概率分布  $P(X, Y | \lambda)$

$$P(X, Y | \lambda) = \prod_{i=1}^T P(x_i, y_i | y_{i-1}, \lambda) = \prod_{i=1}^T P(y_i | y_{i-1}, \lambda) \cdot P(x_i | y_i, \lambda)$$

但从直觉上讲, 这两个假设显然是不够合理的.

MEMM:



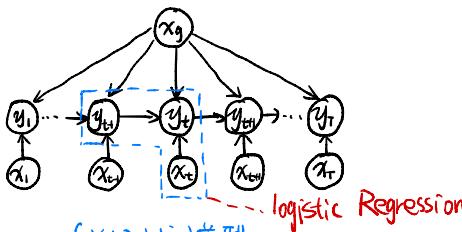
判别模型

给定  $y_t$ ,  $x_t$  和  $x_{t+1}$  不是相互独立的, 因此打破了观测独立假设, 所以更加合理.

建模对象:  $P(Y | X, \lambda)$

$$P(Y | X, \lambda) = \prod_{t=1}^T P(y_t | y_{t-1}, x_{1:T}, \lambda)$$

### 三. MEMM V.S. CRF



优点(和HMM比较):  
 判别式模型  
 打破了观测独立假设

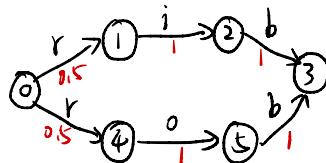
每一个小块可以看成一个 logistic 回归  
 (对每一个小块来说,所有的X都是它的输入),来决定下一状态比类似于  
 $y = f(w^T x + b)$

$$\text{模型: } P(Y|X, \lambda) = \prod_t P(y_t | y_{t-1}, x_{1:T}, \lambda)$$

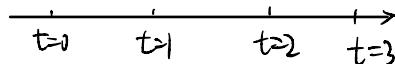
缺点: Label Bias Problem (标注偏差问题)

每一个小块都被归一化了,原因是它是概率分布,概率分布就一定要归一化.  
 直观的解释归一化带来的影响:将整个链看作一根绳子,在 t 时刻抖动这根绳子,绳子的波动一定会对 t+1, t+2, ... 时刻造成影响;但如果进行了归一化,这个波动的影响就会减小,无法正常影响绳子后面,破坏了这个波动的传递性质。

论文中的例子:



此例中, 0, 1, 2, 3, 4, 5 看作是 tagging, r, i, b 看作是观测。



1. 假设在 t=0 时刻, 观测到 r 的情况下由 0  $\rightarrow$  1 和 0  $\rightarrow$  2 的概率相同,  
 由于局部归一化的原因, 导致 1  $\rightarrow$  2 和 4  $\rightarrow$  5 的概率变成了 1, 即:

$$P(2|1, r) = 1 = P(2|1)$$

$$P(5|4, 0) = 1 = P(5|4)$$

可以发现, 由于概率为 1, 输入根本不起作用, 根本没有关注 observation.

2. 假设模型参数由4个样本训练而来，3个rib，1个rib，则可以得到：

$$P(1|0, r) = 0.25, P(4|0, r) = 0.75.$$

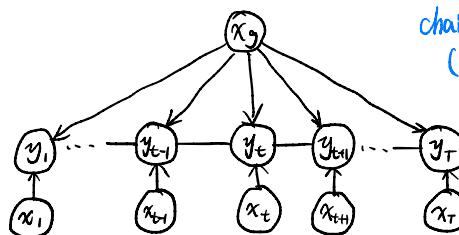
则在Decoding时， $\hat{Y} = \arg \max_{y_1, y_2, y_3} (y_1, y_2, y_3 | \text{rib}) = \max \{0.75 \times 1 \times 1, 0.25 \times 1 \times 1\}$

所以这导致在观测到rib的情况下，所而被解码为0-4-5-3，显然不合适。

由于局部归一化导致 $1 \rightarrow 2$ 概率为1，因此 $1 \rightarrow 2$ 的熵为0。当熵为0时，就不再关注输入是什么，对于无法确定的信息，过多地使用历史信息。

结论：若条件概率分布的熵越小，其关注 observation 的程度就越小，根本原因就是局部归一化。

因此，解决办法就是有向图  $\rightarrow$  无向图，因为无向图的局部不是一个分布，不需要进行局部归一化，只需进行全局归一化。由于其为无向图，因此自然打破了齐次 Markov 假设。



chain-structured CRF.

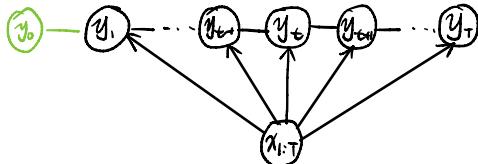
(因为这里关注的是标注问题，因此是  
链性链条件随机场)

克服了 label bias problem

## 四、概率密度函数的参数形式

CRF 的合理性体现在：{ 条件：即条件概率，它是一个判别式模型  
 随机场：即 Markov Field，它是一个无向图模型。 }

CRF 模型图如下：



由于其为无向图，无法直接写出其概率密度函数  $P(Y|X)$

Markov Random Field 因子分解：

$$P(X) = \frac{1}{Z} \prod_{i=1}^K \phi_i(x_{c_i}) = \frac{1}{Z} \prod_{i=1}^K \exp[-E_i(x_{c_i})] = \frac{1}{Z} \exp \sum_{i=1}^K F_i(x_{c_i})$$

能量函数

其中大 K 表示最大团个数， $c_i$  指第 i 个最大团， $\phi_i(\gamma_{c_i})$  表示第 i 个最大团的势函数，其必须为正。但是最大团的个数，应为  $T-1$ ，为了方便见，加上  $y_0$ ，其实是  $T$  个。

对于线性链来讲：

$$\text{因此, } P(Y|X) = \frac{1}{Z} \exp \sum_{i=1}^K F_i(x_{c_i}) = \frac{1}{Z} \exp \sum_{t=1}^T F_t(y_{t+1}, y_t, x_{1:T}) = \frac{1}{Z} \exp \sum_{t=1}^T F_t(y_{t+1}, y_t, x_{1:T})$$

因为线性链每个团的结构相同，没有必要为每个团都设一个不同的  $F_t$ 。

$$F_t(y_{t+1}, y_t, x_{1:T}) = \underbrace{\Delta_{y_{t+1}, x_{1:T}} + \Delta_{y_t, x_{1:T}}}_{\text{状态函数}} + \underbrace{\Delta_{y_{t+1}, y_t, x_{1:T}}}_{\text{转移函数}}$$

由于在  $t$  的上一时刻已经包含了  $\Delta_{y_t}$ ，所以在  $t$  时刻可以将其简化掉，没有必要写两遍。

$$\begin{cases} \Delta_{y_{t+1}, y_t, x_{1:T}} = \sum_{k=1}^K \lambda_k f_k(y_{t+1}, y_t, x_{1:T}) \\ \Delta_{y_t, x_{1:T}} = \sum_{l=1}^L \eta_l g_l(y_t, x_{1:T}) \end{cases}$$

其中  $g_l, f_k$  为特征函数或指示函数， $\lambda_k$  和  $\eta_l$  为参数。

$$\text{故 } P(Y|X) = \frac{1}{Z} \exp \sum_{t=1}^T \left[ \sum_{k=1}^K \lambda_k f_k(y_{t+1}, y_t, x_{1:T}) + \sum_{l=1}^L \eta_l g_l(y_t, x_{1:T}) \right]$$

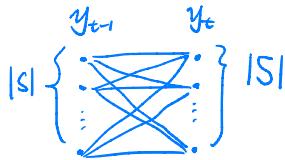
CRF 的 pdf

## 五. 概率密度函数的向量形式

$$P(Y|X) = \frac{1}{Z} \exp \sum_{t=1}^T \left[ \sum_{k=1}^K \lambda_k f_k(y_{t-1}, y_t, x_{t:T}) + \sum_{l=1}^L \eta_l g_l(y_t, x_{t:T}) \right]$$

上节中 K 和 L 是给定的。

例如对任意时刻 t 而言,  $y_t \in S = \{\text{动, 鸟, 副, 助, ...}\}$ , 大小为  $|S|$ , 则对于  $y_{t-1}$  和  $y_t$  的转移图象:



因此共有  $|S|^2$  种可能, 则  $K \leq |S|^2$ , 这些值片有用还是没用, 有很大用是由  $\lambda_k$  决定的。

对概率密度函数的简化:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix} \quad \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{pmatrix} \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_L \end{pmatrix} \quad f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_K \end{pmatrix} = f(y_{t-1}, y_t, x) \quad g = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_L \end{pmatrix} = g(y_t, x)$$

$\lambda$  与  $y$  无关, 它是归一化因子, 是要乘掉  $y$  的。

$$\begin{aligned} P(Y=y|X=x) &= \frac{1}{Z(x, \lambda, \eta)} \exp \sum_{t=1}^T [\lambda^T f(y_{t-1}, y_t, x) + \eta^T g(y_t, x)] \\ &= \frac{1}{Z(x, \lambda, \eta)} \exp \left[ \lambda^T \sum_{t=1}^T f(y_{t-1}, y_t, x) + \eta^T \sum_{t=1}^T g(y_t, x) \right] \end{aligned}$$

$$\text{令 } \theta = \begin{pmatrix} \lambda \\ \eta \end{pmatrix}_{K+L}, \quad H = \begin{bmatrix} \sum_{t=1}^T f \\ \sum_{t=1}^T g \end{bmatrix}_{K+L}$$

$$\text{因此, } P(Y=y|X=x) = \frac{1}{Z(x, \theta)} \exp \{ \theta^T H(y_{t-1}, y_t, x) \}$$

$$= \frac{1}{Z(x, \theta)} \exp \langle \theta, H \rangle$$

去掉  $H$  的原因是接下来需要 learning 和 inference, 涉及到求导等, 使数学公式更明了。

## 六. CRF要解决的问题

{ Learning : 参数估计  
Inference } 边缘概率  
条件概率  
MAP inference / decoding

### Learning

即参数估计, 利用数据来对模型的参数进行求解.

training data:  $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ ,  $x, y$  均是 T 维,  $N$  为样本个数.

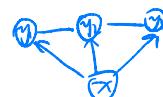
$$\text{即 } \hat{\theta} = \arg \max \prod_{i=1}^N P(y^{(i)} | x^{(i)})$$

### Inference

边缘概率: 求  $P(y_t | \gamma)$

条件概率: 针对生成模型

$$\text{MAP inference / decoding: } \hat{y} = \arg \max P(y | \gamma)$$



下面对边缘概率、Learning 和 decoding 分别讲解.

### 1. 边缘概率

Given  $P(Y=y|X=x)$ , 求  $P(y_t=i|x)$

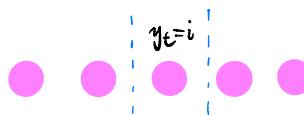
$$P(y|x) = \frac{1}{Z} \prod_{t=1}^T \phi_t(y_{t+1}, y_t, x)$$

$$\text{则 } P(y_t=i|x) = \sum_{y_1, y_2, \dots, y_{t-1}, y_{t+1}, \dots, y_T} P(y|x)$$

$$= \sum_{y_1, \dots, y_{t-1}} \sum_{y_{t+1}, \dots, y_T} \frac{1}{Z} \prod_{t=1}^T \phi_t(y_{t+1}, y_t, x)$$

对于  $y_1, y_2, \dots$  都  $\in S$ ,  $S$  为词性集合, 共  $T-1$  个  $y_t$ , 又因为后面是连乘符号,  
因此复杂度为  $O(|S|^{T-1})$ , 理论上是难以计算的.

观察式子，我们发现每个 $\phi_t$ 只和 2 个 $y$ 相关，因此前面连加号可以移到后面来简化运算，使其变为可解的。



应用变量消除法(VE)：

$$= \frac{1}{2} \cdot \Delta_L \cdot \Delta_R$$

$$\Delta_L = \sum_{y_0, \dots, y_{t-1}} \phi_1(y_0, y_1, x) \cdot \phi_2(y_1, y_2, x) \cdots \phi_{t-1}(y_{t-2}, y_{t-1}, x) \cdot \phi_t(y_{t-1}, y_t=i, x)$$

$$\Delta_R = \sum_{y_{t+1}, \dots, y_T} \phi_{t+1}(y_t=i, y_{t+1}, x) \cdots \phi_T(y_{T-1}, y_T, x)$$

$y_t$  只 ~~只是~~ 是 start  
所以把它看作 SOS.

$$\Rightarrow \Delta_L = \sum_{y_{t+1}} \phi_t(y_{t+1}, y_t=i, x) \left( \sum_{y_{t+2}} \phi_{t+1}(y_{t+2}, y_{t+1}, x) \left( \cdots \left( \sum_{y_T} \phi_T(y_T, y_{T-1}, x) \left( \sum_{y_0} \phi_1(y_0, y_1, x) \right) \right) \right) \right)$$

$$\text{令 } \alpha_t(i) = \Delta_L, \text{ 则 } \alpha_{t+1}(j) = \sum_{y_{t+2}} \phi_{t+1}(y_{t+2}, y_{t+1}=j, x) \left( \cdots \left( \sum_{y_T} \phi_T(y_T, y_{T-1}, x) \left( \sum_{y_0} \phi_1(y_0, y_1, x) \right) \right) \right)$$

$$\text{因此 } \alpha_t(i) = \sum_{y_{t+1}} \phi_t(y_{t+1}=i, y_t=i, x) \cdot \alpha_{t+1}(j)$$

$$= \sum_{j \in S} \phi_t(y_{t+1}=j, y_t=i, x) \cdot \alpha_{t+1}(j)$$

$$\Delta_R = \sum_{y_{t+1}} \phi_{t+1}(y_t=i, y_{t+1}, x) \left( \sum_{y_{t+2}} \phi_{t+2}(y_{t+2}, y_{t+1}, x) \left( \cdots \left( \sum_{y_T} \phi_T(y_T, y_{T-1}, x) \left( \sum_{y_0} \phi_1(y_0, y_1, x) \right) \right) \right) \right) = \beta(i)$$

$$\text{因此 } P(y_t=i|x) = \frac{1}{2} \alpha_t(i) \cdot \beta(i)$$

## CRF Learning

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^N P(y^{(i)} | x^{(i)})$$

$$\hat{y}, \hat{\eta} = \arg \max_{y, \eta} \prod_{i=1}^N P(y^{(i)} | x^{(i)})$$

$$\text{由 } P(Y=y|X=x) = \frac{1}{Z(x, \lambda, \eta)} \exp \sum_{t=1}^T [x^T f(y_{t-1}, y_t, x) + \eta^T g(y_t, x)]$$

$$\text{则 } \hat{y}, \hat{\eta} = \arg \max_{y, \eta} \prod_{i=1}^N P(y^{(i)} | x^{(i)})$$

$$= \arg \max_{\lambda, \eta} \sum_{i=1}^N \log P(y^{(i)} | x^{(i)})$$

$$= \arg \max_{\lambda, \eta} \sum_{i=1}^N (-\log Z(x^{(i)}, \lambda, \eta) + \sum_{t=1}^T [x^T f(y_{t-1}^{(i)}, y_t^{(i)}, x) + \eta^T g(y_t^{(i)}, x)])$$

$$\triangleq \arg \max_{\lambda, \eta} L(\lambda, \eta, x)$$

这是一个优化问题，可以使用梯度上升的方法，即求  $\nabla_\lambda L$ ,  $\nabla_\eta L$ .

求  $\nabla_\lambda L$ :

$$\nabla_\lambda L = \sum_{i=1}^N \left( \sum_{t=1}^T f(y_{t-1}^{(i)}, y_t^{(i)}, x) - \nabla_\lambda \log Z(x^{(i)}, \lambda, \eta) \right) \quad (\text{配分函数})$$

在配分函数分布中，其称为 log-partition function

log-partition function 的导数是充分统计量的边缘值，即  $E[\sum_{t=1}^T f(y_{t-1}, y_t, x)]$

$$\begin{aligned} \text{则 } E[\sum_{t=1}^T f(y_{t-1}, y_t, x)] &= \sum_y P(y|x) \cdot \sum_{t=1}^T f(y_{t-1}, y_t, x) \\ &= \sum_{t=1}^T \left( \sum_y P(y|x) \cdot f(y_{t-1}, y_t, x) \right) \\ &= \sum_{t=1}^T \left( \sum_{y_1, y_2} \sum_{y_3} \sum_{y_4} \dots \sum_{y_T} P(y|x) \cdot f(y_{t-1}, y_t, x) \right) \\ &= \sum_{t=1}^T \sum_{y_1, y_2} \left( \sum_{y_3, y_4, \dots, y_T} P(y|x) \cdot f(y_{t-1}, y_t, x) \right) \\ &= \sum_{t=1}^T \sum_{y_1, y_2} \left( \underbrace{P(y_{t-1}, y_t|x)}_{\text{边缘概率}} \cdot f(y_{t-1}, y_t, x) \right) \end{aligned}$$

$$P(y_{t-1}, y_t|x) = \frac{1}{2} \alpha_{t-1}(i) \cdot \phi_i(y_{t-1}=i, y_t=j|x) \cdot \beta_t(j) = A(y_{t-1}, y_t)$$

$$\text{因此, } \nabla_\lambda L = \sum_{i=1}^N \sum_{t=1}^T \left( f(y_{t-1}^{(i)}, y_t^{(i)}, x) - \sum_{y_1} \sum_{y_2} A(y_{t-1}, y_t) \cdot f(y_{t-1}, y_t, x) \right)$$

$$\text{同理, } \nabla_\eta L = \sum_{i=1}^N \sum_{t=1}^T \left( f(y_{t-1}^{(i)}, y_t^{(i)}, x) - \sum_{y_1} \sum_{y_2} A(y_{t-1}, y_t) \cdot g(y_t, x) \right)$$

$$\text{则 } \begin{cases} \lambda^{t+1} = \lambda^t + \text{step} \cdot \nabla_\lambda L(\lambda^t, \eta^t) \\ \eta^{t+1} = \eta^t + \text{step} \cdot \nabla_\eta L(\lambda^t, \eta^t) \end{cases}$$

在实际中这种方法收敛速度太慢，一般不会使用。