

# 条件随机场 conditional random fields

## 条件随机场概述


条件随机场模型是Lafferty于2001年，在最大熵模型和隐马尔科夫模型的基础上，提出的一种判别式概率无向图学习模型，是一种用于标注和切分有序数据的条件概率模型。

CRF最早是针对序列数据分析提出的，现已成功应用于自然语言处理（Natural Language Processing，NLP）、生物信息学、机器视觉及网络智能等领域。

## 序列标注

标注：人名 地名 组织名

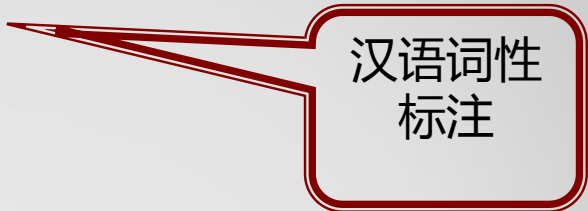
观察序列：毛泽东



实体命名  
识别

标注：名词 动词 助词 形容词 副词 .....

观察序列：今天天气非常好！



汉语词性  
标注

一、产生式模型和判别式模型 ( Generative model vs. Discriminative model )

二、概率图模型 ( Graphical Models )

三、朴素贝叶斯分类器 ( Naive Bayes Classifier )

四、隐马尔可夫模型 ( Hidden Markov Model , HMM )

五、最大熵模型 ( Maximum Entropy Model , MEM )

六、最大熵马尔可夫模型 ( MEMM )

七、条件随机场 ( conditional random fields , CRF )

# 一、产生式模型和判别式模型 ( Generative model vs. Discriminative model )

$o$ 和 $s$ 分别代表观察序列和标记序列

- **产生式模型**：构建 $o$ 和 $s$ 的联合分布 $p(s,o)$ ，因可以根据联合概率来生成样本，如HMM，BNs，MRF。
- **判别式模型**：构建 $o$ 和 $s$ 的条件分布 $p(s|o)$ ，因为没有 $s$ 的知识，无法生成样本，只能判断分类，如SVM，CRF，MEMM。

产生式模型：无穷样本 ==》 概率密度模型 = 产生模型 ==》 预测

判别式模型：有限样本 ==》 判别函数 = 预测模型 ==》 预测

一个举例：

$(1,0), (1,0), (2,0), (2, 1)$

产生式模型：

$P(x, y):$

$P(1, 0) = 1/2, P(1, 1) = 0, P(2, 0) = 1/4, P(2, 1) = 1/4.$

判别式模型：

$P(y | x):$

$P(0|1) = 1, P(1|1) = 0, P(0|2) = 1/2, P(1|2) = 1/2$

## 两种模型比较：

**Generative model**：从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度，不关心判别边界。

### 优点:

- 实际上带的信息要比判别模型丰富，研究单类问题比判别模型灵活性强
- 能更充分的利用先验知识
- 模型可以通过增量学习得到

### 缺点：

- 学习过程比较复杂
- 在目标分类问题中易产生较大的错误率

**Discriminative model** : 寻找不同类别之间的最优分类面，反映的是异类数据之间的差异。

### 优点:

- 分类边界更灵活，比使用纯概率方法或生产模型得到的更高级。
- 能清晰的分辨出多类或某一类与其他类之间的差异特征
- 在聚类、viewpoint changes, partial occlusion and scale variations中的效果较好
- 适用于较多类别的识别

### 缺点：

- 不能反映训练数据本身的特性。
- 能力有限，可以告诉你的是1还是2，但没有办法把整个场景描述出来。

**二者关系**：由生成模型可以得到判别模型，但由判别模型得不到生成模型。



## 二、概率图模型 ( Graphical Models )

**概率图模型**：是一类用图的形式表示随机变量之间条件依赖关系的概率模型，是概率论与图论的结合。图中的节点表示随机变量，缺少边表示条件独立假设。

$$G = (V, E)$$

$V$ ：顶点/节点，表示随机变量

$E$ ：边/弧

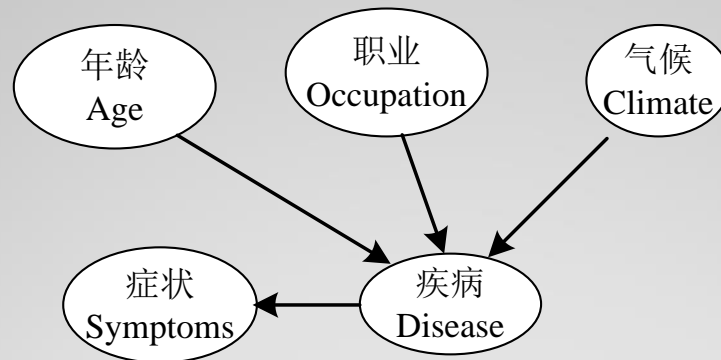
**两个节点邻接**：两个节点之间存在边，记为  $X_i \sim X_j$ ，不存在边，表示条件独立

**路径**：若对每个  $i$ ，都有  $X_{i-1} \square X_i$ ，则称序列  $X_1, \dots, X_N$  为一条路径

根据图中边有无方向，常用的概率图模型分为两类：

有向图：最基本的是贝叶斯网络(Bayesian Networks ,BNs)

举例



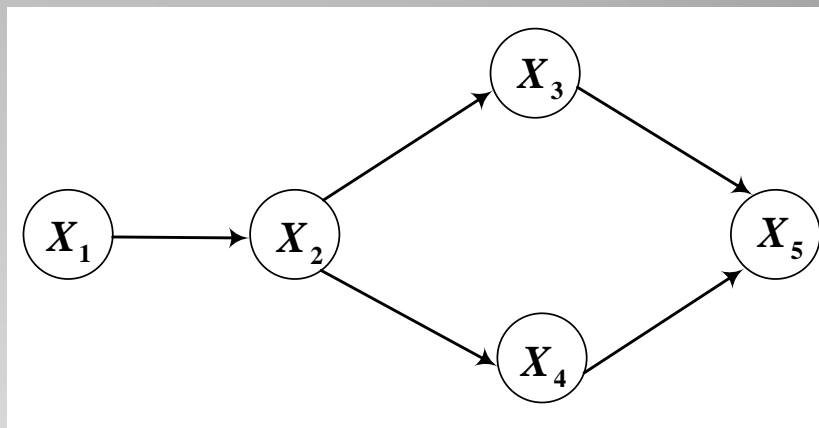
$$P(A,O,C,D,S|M) =$$

$$P(A|M)P(O|M)P(C|M)P(D|A,O,C,M)P(S|D,M)$$

## 有向图模型的联合概率分解

每个节点的条件概率分布表示为：

$P(\text{当前节点}|\text{它的父节点})$



联合分布：

$$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^N p(X_i | \pi(X_i))$$

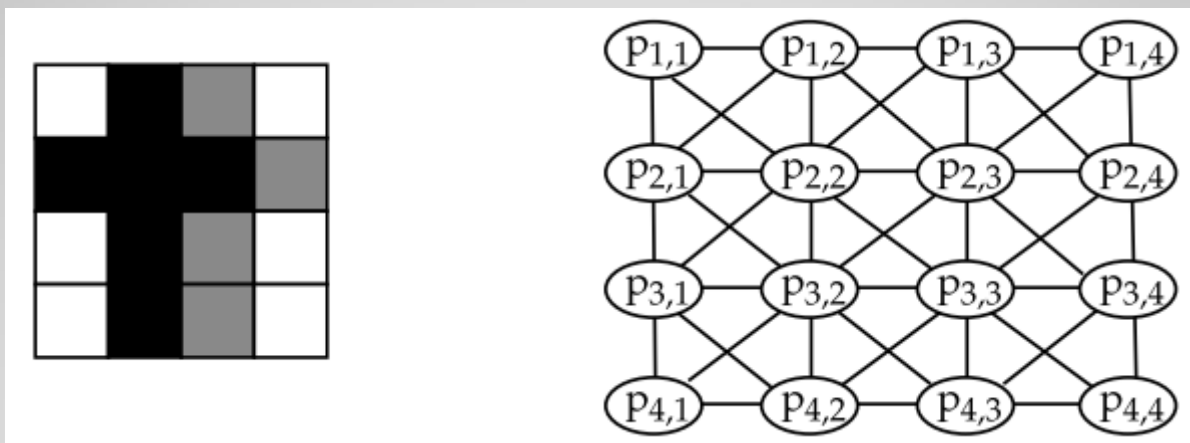
$$P(X_1, X_2, \dots, X_5) = p(X_1)p(X_2|X_1)p(X_3|X_2)p(X_4|X_2)p(X_5|X_3X_4)$$

无向图：马尔可夫随机场(Markov Random Fields, MRF)

马尔可夫随机场模型中包含了一组具有马尔可夫性质的随机变量，这些变量之间的关系用无向图来表示

马尔可夫性：
$$p(x_i | x_j, j \neq i) = p(x_i | x_j, x_i \perp x_j)$$

举例



团(clique)：任何一个全连通（任意两个顶点间都有边相连）的子图

最大团(maximal clique)：不能被其它团所包含的团

例如右图的团有 $C1=\{X_1, X_2, X_3\}$ 和 $C2=\{X_2, X_3, X_4\}$

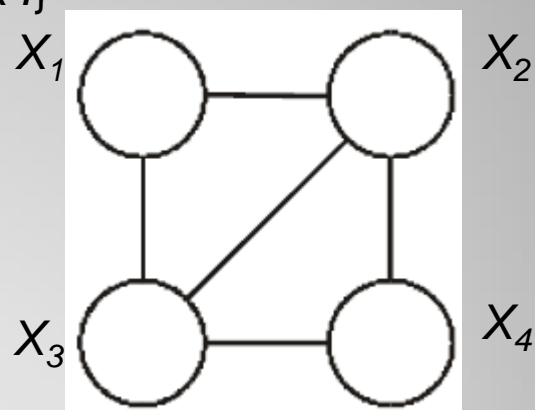
无向图模型的联合概率分解

$$P(X_1, X_2, \dots, X_N) = \frac{1}{Z} \prod_{i=1}^N \Psi_i(C_i)$$

$$Z = \sum_{\{X_1, X_2, \dots, X_N\}} \left\{ \prod_{i=1}^N \Psi_i(C_i) \right\}$$

势函数 ( potential function )

$\Psi_i(C_i)$ ：是关于 $C_i$ 上 随机变量的函数



$$p(X_1, X_2, X_3, X_4) = \frac{\Phi_1(X_1, X_2, X_3) \Phi_2(X_2, X_3, X_4)}{\sum_{\{X_1, X_2, X_3, X_4\}} \{ \Phi_1(X_1, X_2, X_3) \Phi_2(X_2, X_3, X_4) \}}$$

### 三、朴素贝叶斯分类器 ( Naive Bayes Classifier )

设 $x \in \Omega$ 是一个类别未知的数据样本， $Y$ 为类别集合，若数据样本 $x$ 属于一个特定的类别 $y_j$ ，那么分类问题就是决定 $P(y_j|x)$ ，即在获得数据样本 $x$ 时，确定 $x$ 的最佳分类。所谓最佳分类，一种办法是把它定义为在给定数据集中不同类别 $y_j$ 先验概率的条件下最可能的分类。贝叶斯理论提供了计算这种可能性的一种直接方法。

$$p(y_j | x) = \frac{p(x | y_j) p(y_j)}{p(x)}$$

$P(y_j)$ 代表还没有训练数据前， $y_j$ 拥有的初始概率。 $P(y_j)$ 常被称为 $y_j$ 的先验概率(prior probability)，它反映了我们所拥有的关于 $y_j$ 是正确分类机会的背景知识，它应该是独立于样本的。

如果没有这一先验知识，那么可以简单地将每一候选类别赋予相同的先验概率。不过通常我们可以用样例中属于 $y_j$ 的样例数 $|y_j|$ 比上总样例数 $|D|$ 来近似，即

$$P(y_j) = \frac{|y_j|}{|D|}$$

$$p(y_j | x) = \frac{p(x | y_j) p(y_j)}{p(x)}$$

$p(x | y_j)$  是联合概率，指当已知类别为  $y_j$  的条件下，看到样本  $x$  出现的概率。

若设  $x = (a_1, a_2, \dots, a_m)$

则  $p(x | y_j) = p(a_1, a_2, \dots, a_m | y_j)$



条件独立性：

$$p(a, b | c) = p(a | c) p(b | c)$$

在给定随机变量C时，a，b条件独立。

假定：在给定目标值  $y_j$  时，x的属性值之间相互条件独立。

$$p(x | y_j) = p(a_1, a_2, \dots, a_m | y_j) = \prod_{i=1}^m p(a_i | y_j)$$

$$p(y_j | x) = \frac{p(x | y_j) p(y_j)}{p(x)}$$

$p(y_j | x)$  是后验概率，即给定数据样本 $x$ 时 $y_j$ 成立的概率，而这正是我们所感兴趣的。

$P(y_j|x)$  被称为 $Y$ 的后验概率 ( posterior probability )，因为它反映了在看到数据样本 $x$ 后 $y_j$ 成立的置信度。

后验概率

$$p(y_j | x) = \frac{p(y_j)p(x|y_j)}{p(x)} \quad j = 1, \dots, |Y|$$

$$\arg \max_j p(y_j | x) = \arg \max_j p(y_j | x_1, x_2, x_3)$$

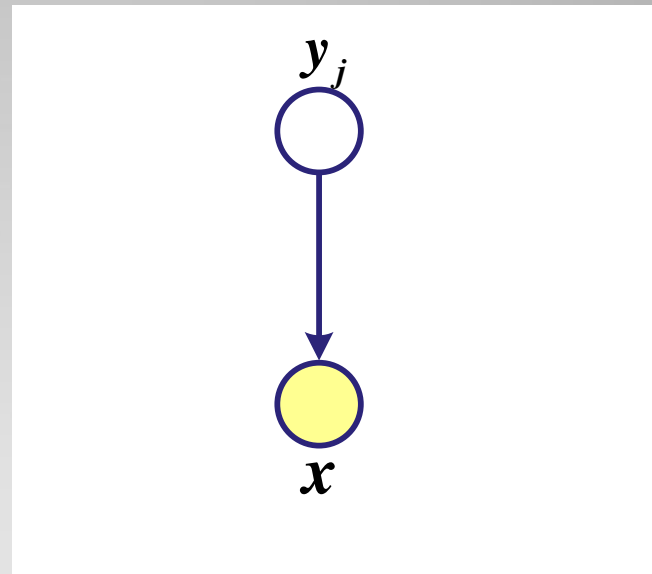
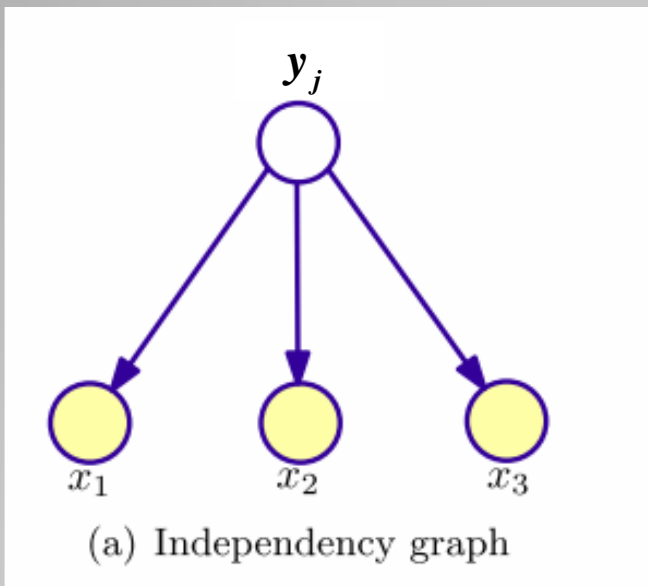
$$= \arg \max_j \frac{p(x_1, x_2, x_3 | y_j)p(y_j)}{p(x_1, x_2, x_3)}$$

$$= \arg \max_j p(x_1, x_2, x_3, y_j)$$

基本假设

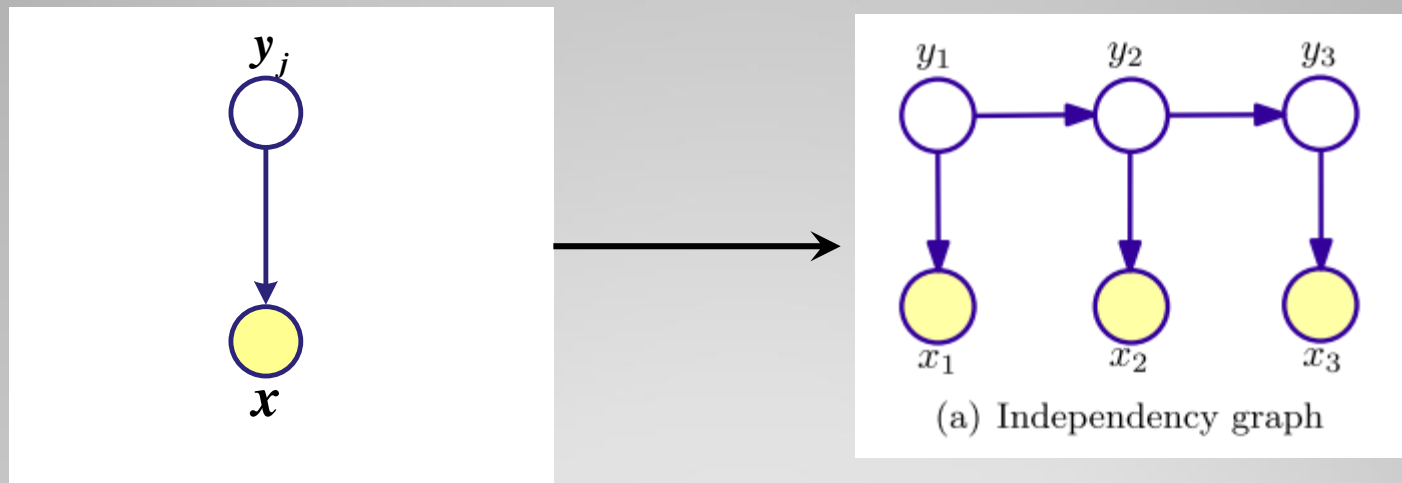
$$= \arg \max_j \prod_{i=1}^3 p(x_i | y_j)p(y_j)$$

## 朴素贝叶斯分类器的概率图表示



$$P(x_1, x_2, x_3, y_j) = p(y_j) p(x_1 | y_j) p(x_2 | y_j) p(x_3 | y_j)$$

## 隐马尔可夫模型的概率图表示



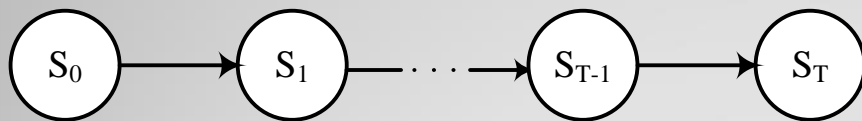
$$p(\vec{y}, \vec{x}) = \prod_{i=1}^n p(y_i | y_{i-1}) p(x_i | y_i)$$

### 三、隐马尔可夫模型 ( Hidden Markov Model, HMM )

马尔可夫模型：是一个三元组  $\lambda = (S, \Pi, A)$

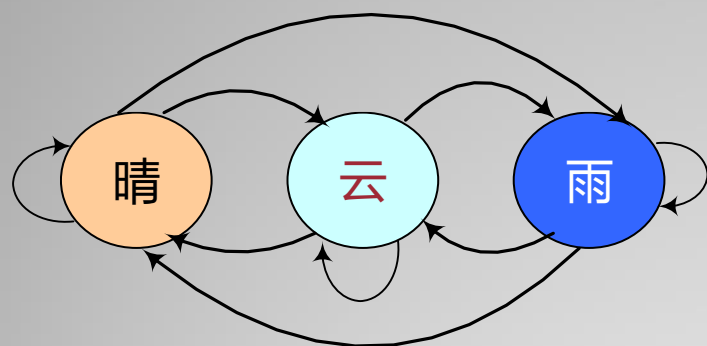
其中  $S$  是状态的集合， $\Pi$  是初始状态的概率， $A$  是状态间的转移概率。

一阶马尔可夫链



$$p(S_0, S_1, \dots, S_T) = \prod_{t=1}^T p(S_t | S_{t-1}) p(S_0)$$

## 一阶马尔可夫模型的例子



晴 云 雨

$$S = \{s_1, s_2, s_3\}$$

$$\pi = (1, 0, 0)$$

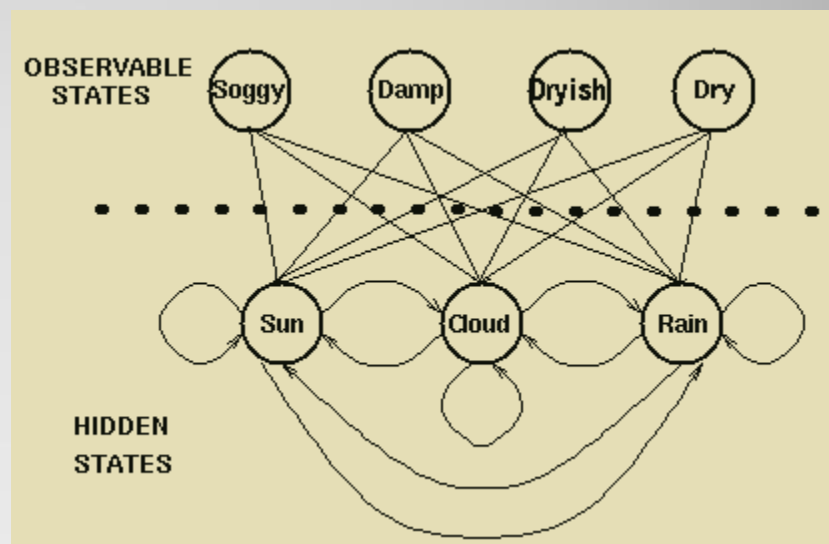
		<i>today</i>		
		<i>sun</i>	<i>cloud</i>	<i>rain</i>
<i>yesterday</i>	<i>sun</i>	$\begin{bmatrix} 0.50 & 0.375 & 0.125 \\ 0.25 & 0.125 & 0.625 \\ 0.25 & 0.375 & 0.375 \end{bmatrix}$		
	<i>cloud</i>			
	<i>rain</i>			

**问题**：假设今天是晴天，请问未来三天的天气呈现云雨晴的概率是多少？

## 隐马尔可夫模型(HMM)

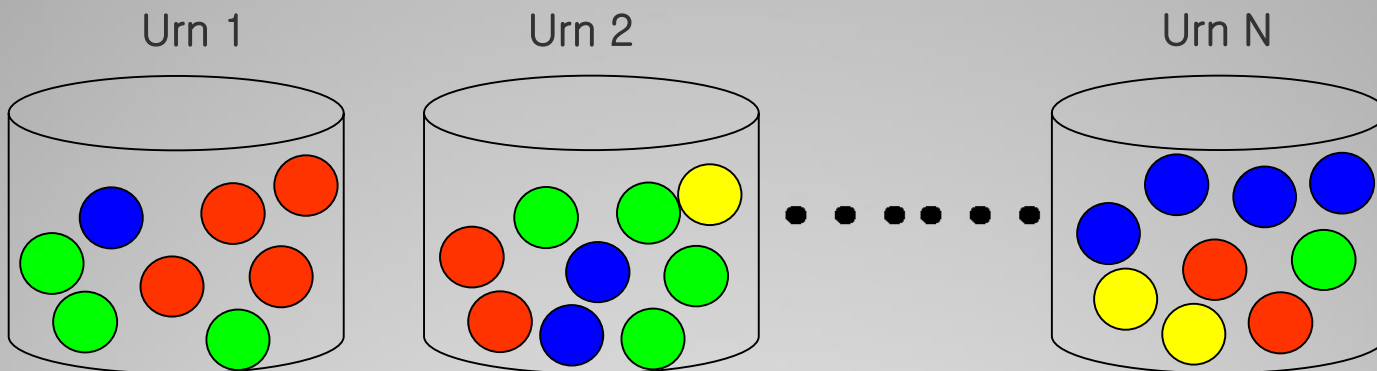
HMM是一个五元组  $\lambda = (Y, X, \Pi, A, B)$  , 其中  $Y$  是隐状态 (输出变量) 的集合,  $X$  是观察值 (输入) 集合,  $\Pi$  是初始状态的概率,  $A$  是状态转移概率矩阵,  $B$  是输出观察值概率矩阵。

		<i>today</i>			
		<i>sun</i>	<i>cloud</i>	<i>rain</i>	
<i>yesterday</i>	<i>sun</i>	$\begin{bmatrix} 0.50 & 0.375 & 0.125 \\ 0.25 & 0.125 & 0.625 \\ 0.25 & 0.375 & 0.375 \end{bmatrix}$			
	<i>cloud</i>				
	<i>rain</i>				
		<i>soggy</i>	<i>damp</i>	<i>dryish</i>	<i>dry</i>
<i>sun</i>	$\begin{bmatrix} 0.05 & 0.15 & 0.20 & 0.60 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.35 & 0.10 & 0.05 \end{bmatrix}$				
<i>cloud</i>					
<i>rain</i>					





## HMM实例



实验进行方式如下：

- 根据初始概率分布，随机选择 $N$ 个缸中的一个开始实验
- 根据缸中球颜色的概率分布，随机选择一个球，记球的颜色为 $x_1$ ，并把球放回缸中
- 根据缸的转移概率分布，随机选择下一口缸，重复以上步骤。

最后得到一个描述球的颜色序列 $x_1, x_2, \dots$ 称为观察值序列 $X$ 。

Observed Ball Sequence         

## 评价问题

**问题1**：给定观察序列  $X = \{x_1, x_2, \dots, x_T\}$  以及模型  $\lambda = (\pi, A, B)$ ，计算  $P(X|\lambda)$

## 解码问题

**问题2**：给定观察序列  $X = \{x_1, x_2, \dots, x_T\}$  以及模型  $\lambda$ ，如何选择一个对应的状态序列  $Y = (y_1, y_2, \dots, y_N)$ ，使得  $Y$  能够最为合理的解释观察序列  $X$ ？

## 参数学习问题

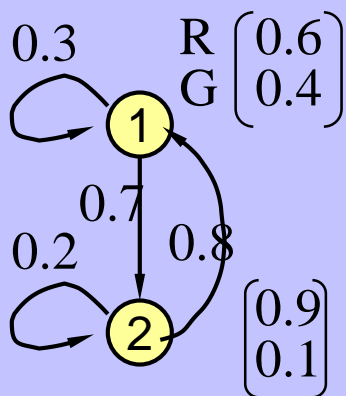
**问题3**：给定观察序列  $X = \{x_1, x_2, \dots, x_T\}$ ，调整模型参数  $\lambda = (\pi, A, B)$ ，使  $P(X|\lambda)$  最大？

**问题1**：给定观察序列  $X = \{x_1, x_2, \dots, x_T\}$  以及模型  $\lambda = (\pi, A, B)$ ，计算  $P(X|\lambda)$

基本算法：

$$P(X / \lambda) = \sum_{\text{所有} Y} P(X / Y, \lambda) P(Y / \lambda)$$

$$\pi = [0.5 \quad 0.5]^T$$



R

R

G

①

①

①

$$0.5 \times 0.3 \times 0.3 \times 0.6 \times 0.6 \times 0.4$$

①

①

②

①

②

①

⋮

## 前向算法：

定义前向变量： $\alpha_t(i) = P(x_1, x_2, \dots, x_t, y_t = i | \lambda) \quad 1 \leq t \leq T$

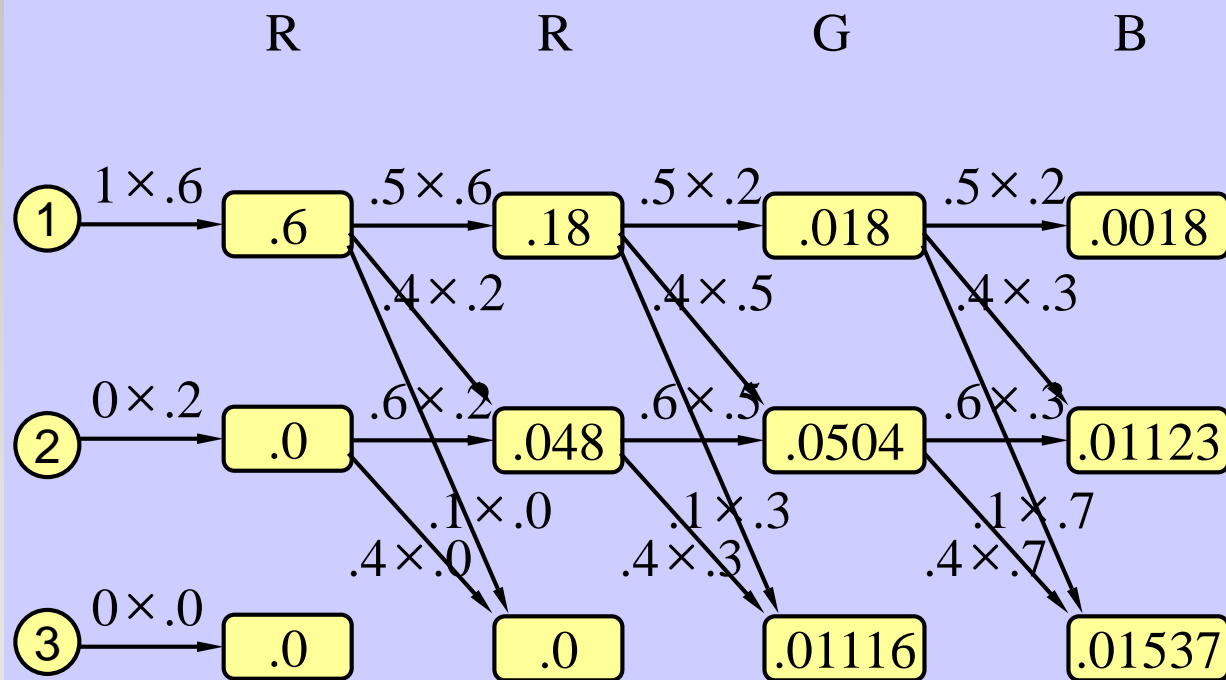
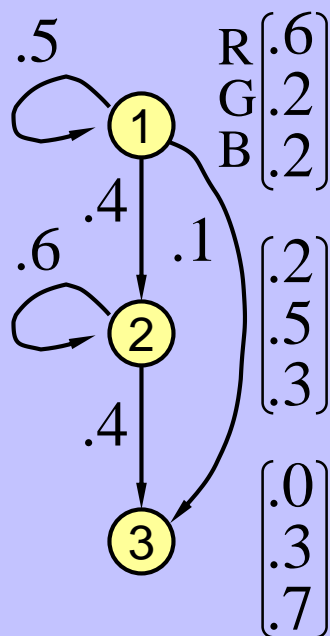
初始化： $\alpha_1(i) = \pi_i b_i(x_1) \quad 1 \leq i \leq N$

递归： $\alpha_{t+1}(j) = [\sum_{i=1}^N \alpha_t(i) a_{ij}] b_j(x_{t+1}) \quad 1 \leq t \leq T-1, 1 \leq j \leq N$

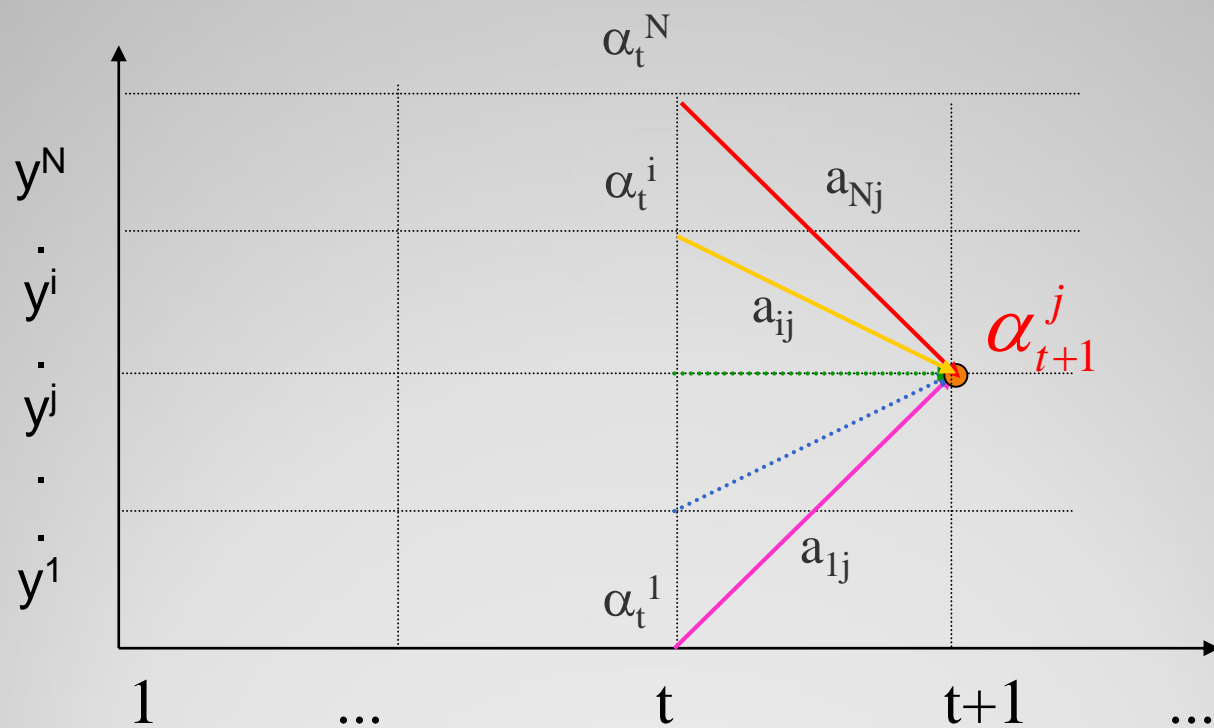
终结： $P(X / \lambda) = \sum_{i=1}^N \alpha_T(i)$

## 前向算法举例：

$$\pi = [1 \ 0 \ 0]^T$$



## 前向法示意图



## 后向法

定义后向变量

$$\beta_t(i) = P(x_{t+1}, x_{t+2}, \dots, x_T, y_t = i / \lambda) \quad 1 \leq t \leq T-1$$

初始化：  $\beta_T(i) = 1 \quad 1 \leq i \leq N$

递归：  $\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(x_{t+1}) \beta_{t+1}(j) \quad t = T-1, T-2, \dots, 1, 1 \leq i \leq N$

终结：  $P(X / \lambda) = \sum_{i=1}^N \beta_1(i)$

**问题2**：给定观察序列  $X = \{x_1, x_2, \dots, x_T\}$  以及模型  $\lambda$ ，如何选择一个对应的状态序列  $Y = (y_1, y_2, \dots, y_N)$ ，使得  $Y$  能够最为合理的解释观察序列  $X$ ？

## Viterbi 算法：

定义：

$$\delta_t(i) = P[y_1 y_2 \dots y_{t-1}, y_t = i, x_1, x_2, \dots, x_t \mid \lambda]$$

要找的就是  $T$  时刻  $\delta_T(i)$  所代表的那个状态序列

$$\arg \max_Y P(Y \mid X)$$



## Viterbi 算法：

初始化  $\delta_1(i) = \pi_i b_i(x_1) \quad \psi_1(i) = 0$

递归  $\delta_{t+1}(j) = \max_{1 \leq i \leq N} \delta_t(i) a_{ij} b_j(x_{t+1})$

$$\psi_{t+1}(j) = \arg \max_{1 \leq i \leq N} \delta_t(i) a_{ij} b_j(x_{t+1})$$

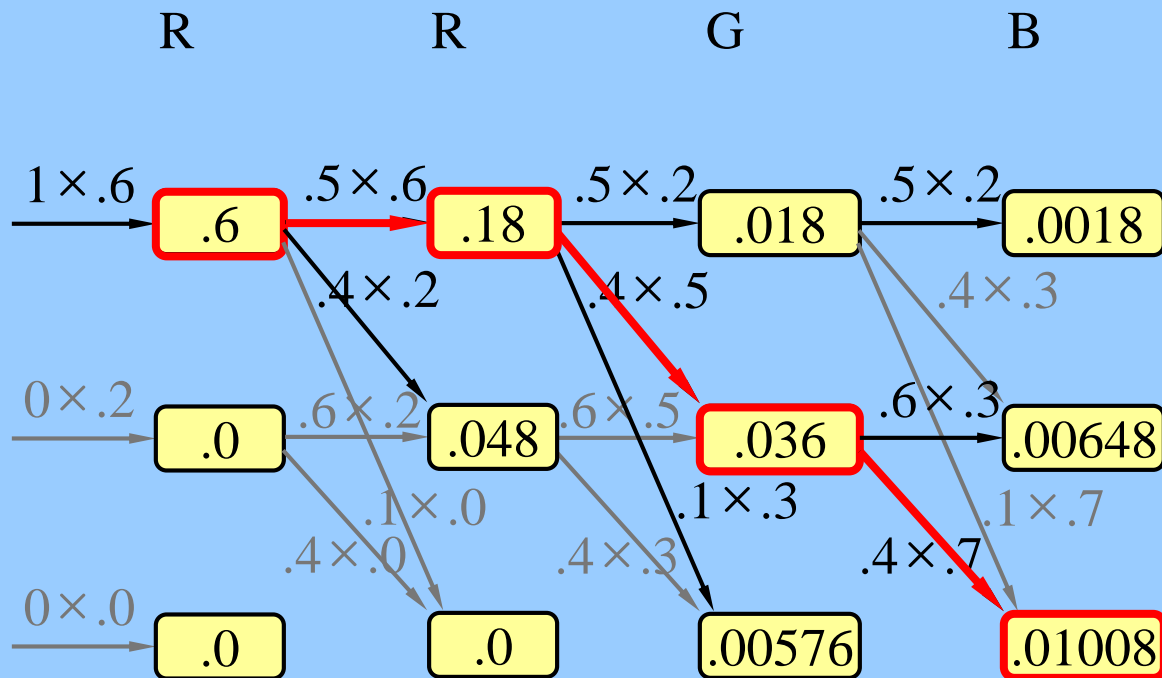
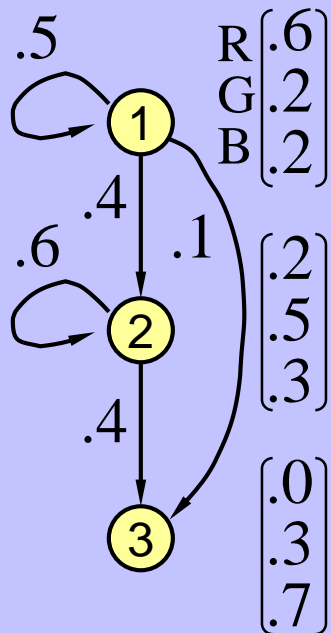
结束  $P^* = \max_{1 \leq i \leq N} \delta_T(i)$

$$y_T^* = \arg \max_{1 \leq i \leq N} \delta_T(i)$$

得到最优路径  $y_t^* = \psi_{t+1}(y_{t+1}^*), \quad t = T-1, \dots, 1$

## Viterbi 算法举例：

$$\pi = [1 \ 0 \ 0]^T$$



**问题3**：给定观察序列  $X = \{x_1, x_2, \dots, x_T\}$ ，调整模型参数  $\lambda = (\pi, A, B)$ ，使  $P(X|\lambda)$  最大？

**思想**：给定一个模型和输出字符序列，任意设定初始参数值，通过不断循环更新参数的方法，设法达到最优。

Baum 1970

**算法步骤**：

1. 初始模型（待训练模型） $\lambda_0$ ，
2. 基于 $\lambda_0$ 以及观察值序列 $X$ ，训练新模型 $\lambda$ ；
3. 如果  $\log P(X|\lambda) - \log(P(X|\lambda_0)) < \Delta$ ，说明训练已经达到预期效果，算法结束。
4. 否则，令 $\lambda_0 = \lambda$ ，继续第2步工作

# Baum-Welch算法

定义：

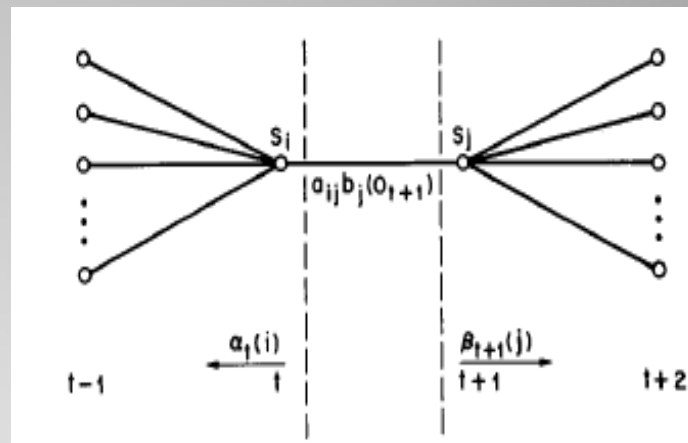
给定模型 $\lambda$ 和观察序列条件下，从 $i$ 到 $j$ 的转移概率定义为 $\xi_t(i, j)$

$$\begin{aligned}\xi_t(i, j) &= P(y_t = i, y_{t+1} = j | X, \lambda) \\ &= \frac{\alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)}\end{aligned}$$

$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j) \quad t\text{时刻处于状态}y_i\text{的概率}$$

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{整个过程中从状态}y_i\text{转出的次数 (number of time) 的预期}$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{从}y_i\text{跳转到}y_j\text{次数的预期}$$



## 重新估计

$$\hat{a}_{ij} = \frac{\sum_t \xi_t(i, j)}{\sum_t \sum_j \xi_t(i, j)}$$

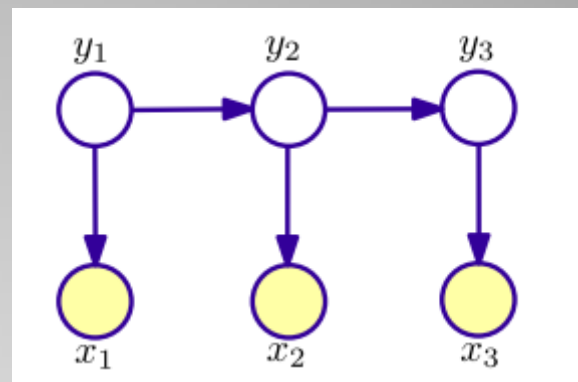
$$\hat{b}_j(k) = \frac{\sum_{t, x_t=k} \gamma_t(j)}{\sum_t \gamma_t(j)}$$

$\pi_i$  = 当 $t=1$ 时处于 $S_i$ 的概率 =  $\gamma_1(i)$

- 该算法又称为向前向后算法 ( *Forward-backward algorithm* )
- 经常得到局部最优解

HMMs等生产式模型存在的问题：

$$P(X) = \sum_{\text{所有的Y}} \prod_{i=1}^T p(y_i | y_{i-1}) p(x_i | y_i)$$



1. 由于生成模型定义的是联合概率，必须列举所有观察序列的可能值，这对多数领域来说是比较困难的。

2. 基于观察序列中的每个元素都相互条件独立。即在任何时刻观察值仅仅与状态（即要标注的标签）有关。对于简单的数据集，这个假设倒是合理。但大多数现实世界中的真实观察序列是由多个相互作用的特征和观察序列中较长范围内的元素之间的依赖而形成的。

## 四、最大熵模型 ( Maximum Entropy Model , MEM )

- 最大熵模型主要是在已有的一些限制条件下估计未知的概率分布。

熵的计算公式：
$$H(X) = - \sum_{x \in X} p(x) \log p(x)$$

熵的性质：
$$0 \leq H(X) \leq \log |X|$$

- 其中X在离散分布时是随机变量的个数；
  - 当X为确定值，即没有变化的可能时，左边等式成立；
  - 可以证明，当X服从均匀分布时，右边等式成立，即均匀分布时熵最大。
- 最大熵的原理认为，从不完整的信息（例如有限数量的训练数据）推导出的唯一合理的概率分布应该在满足这些信息提供的约束条件下拥有最大熵值。求解这样的分布是一个典型的约束优化问题。

定义条件熵  $H(y|x) = - \sum_{(x,y) \in Z} p(y,x) \log p(y|x)$

模型目的  $p^*(y|x) = \arg \max_{p(y|x) \in P} H(y|x)$

定义特征函数  $f_i(x, y) \in \{0, 1\} \quad i = 1, 2, \dots, m$

约束条件  $\sum_{y \in Y} p(y|x) = 1 \quad (1)$

$$E(f_i) = \tilde{E}(f_i) \quad i = 1, 2, \dots, m \quad (2)$$

$$\tilde{E}(f_i) = \sum_{(x,y) \in Z} \tilde{p}(x,y) f_i(x,y) = \frac{1}{N} \sum_{(x,y) \in T} f_i(x,y) \quad N = |T|$$

$$\begin{aligned} E(f_i) &= \sum_{(x,y) \in Z} p(x,y) f_i(x,y) = \sum_{(x,y) \in Z} p(x) p(y|x) f_i(x,y) \\ &= \frac{1}{N} \sum_{x \in T} \sum_{y \in Y} p(y|x) f_i(x,y) \end{aligned}$$



该条件约束优化问题的Lagrange函数

$$\Lambda(p, \vec{\lambda}) = H(y|x) + \sum_{i=1}^m \lambda_i \left( E(f_i) - \tilde{E}(f_i) \right) + \lambda_{m+1} \left( \sum_{y \in Y} p(y|x) - 1 \right)$$

最大熵模型：

$$p_{\vec{\lambda}}^*(y|x) = \frac{1}{Z_{\vec{\lambda}}(x)} \exp \left( \sum_{i=1}^m \lambda_i f_i(x, y) \right),$$

$$Z_{\vec{\lambda}}(x) = \sum_{y \in \mathcal{Y}} \exp \left( \sum_{i=1}^m \lambda_i f_i(x, y) \right).$$

$$p_{\vec{\lambda}}^*(y|x) = \frac{1}{Z_{\vec{\lambda}}(x)} \exp \left( \sum_{i=1}^m \lambda_i f_i(x, y) \right),$$

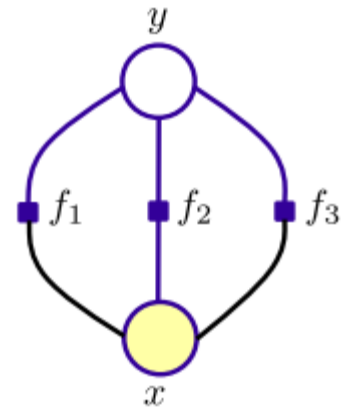
## 最大熵模型的概率图

$$p_{\vec{\lambda}}(y|x) = \frac{1}{Z_{\vec{\lambda}}(x)} \prod_{i=1}^m \exp(\lambda_i f_i(x, y))$$

$$P(X_1, X_2, \dots, X_N) = \frac{1}{Z} \prod_{i=1}^N \Psi_i(C_i)$$



(a) Independence graph



(b) Factor graph

# 有向图模型和无向图模型的对比

## 1 共同之处

将复杂的联合分布分解为多个因子的乘积

## 2 不同之处

无向图模型因子是势函数，需要全局归一

有向图模型因子是概率分布、无需全局归一

## 3 优缺点

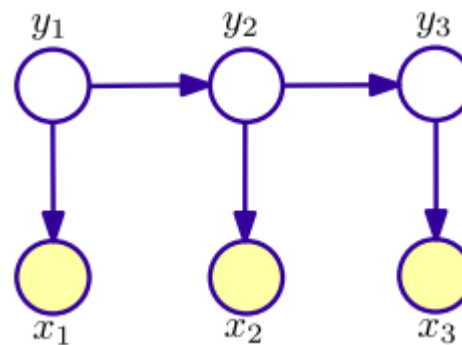
无向图模型中势函数设计不受概率分布约束，设计灵活，但全局归一代价高

有向图模型无需全局归一、训练相对高效



NBs

序列



HMMs



MEMs

序列

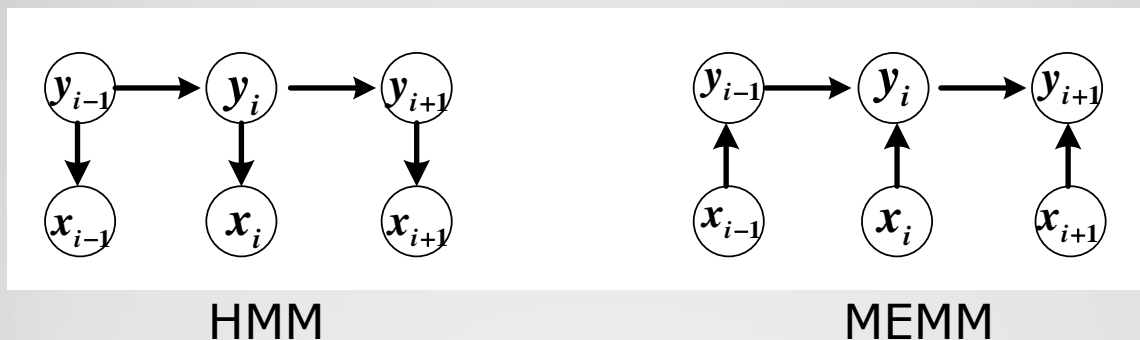


## 六、最大熵马尔可夫模型 ( MEMM )

HMM：状态集合 $Y$ ，观察值集合 $X$ ，两个状态转移概率：从 $y_{i-1}$ 到 $y_i$ 的条件概率分布 $P(y_i | y_{i-1})$ ，状态 $y_i$ 的输出观察值概率 $P(x_i | y_i)$ ，初始概率 $P_0(y)$ 。

MEMM：用一个 $P(y_i | y_{i-1}, x_i)$ 分布来替代HMM中的两个条件概率分布，它表示从先前状态，在观察值下得到当前状态的概率，即根据前一状态和当前观察预测当前状态。每个这样的分布函数都是一个服从最大熵的指数模型。

$$p_{y_{i-1}}(y_i | x_i) = \frac{1}{Z(x_i, y_{i-1})} \exp \left\{ \sum_a \lambda_a f_a(x_i, y_i) \right\} \quad i = 1, 2, \dots, T$$



## 参数学习

目的：通过学习 $\lambda_a$ 使得MEMM中的每个转换函数达到最大熵。

GIS ( Generalized Iterative Scaling ) 算法

## 编码问题

Viterbi算法的思想

## MEMM存在的问题：标记偏见 ( Label Bias Problem ) 问题

最大熵马尔科夫模型与 HMMs 不同，它不是一个生产模型，而是一个基于下状态分类器的有限状态模型，但是，它却存在一个缺点就是标记偏见问题(label bias problem)，如图 1 所示。

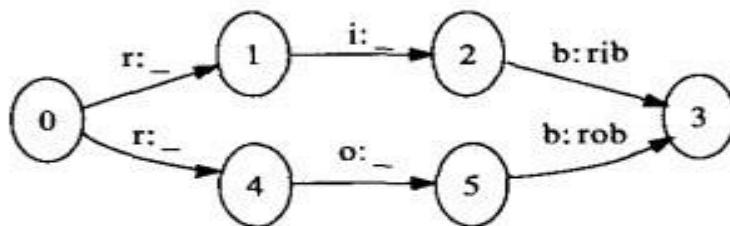


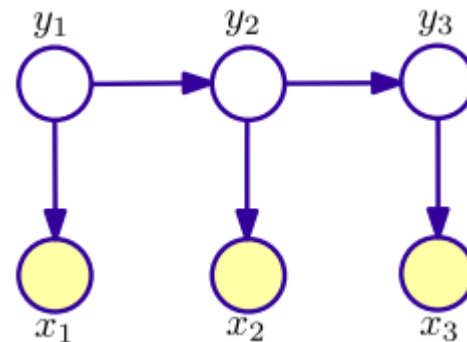
图 1 标记偏见问题

图 1 为一个简单的有限状态机用来区别单词 rob 和 rib。从状态 0 到 1 或 4 的概率相同(他们面对相同的观测 r)从状态 1 到 2 以及从状态 4 到 5 均只有惟一的转换，故而这个有限状态机无法区分 rob 和 rib。产生这种问题的原因就在于 MEMMs 对于状态序列的计算是局部的。而 CRFs 则是对状态序列进行全局的计算，从而克服了标记偏见问题。



NBs

序列

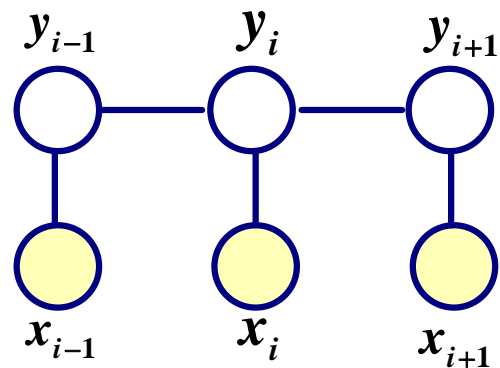


HMMs



MEMs

序列



linear-chain CRF



## 五、条件随机场 ( conditional random fields , CRF )

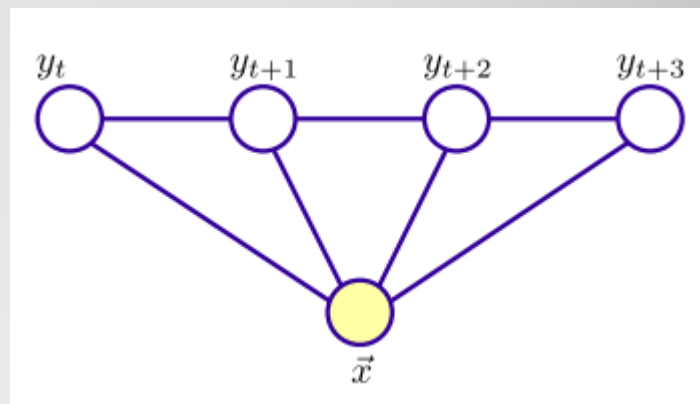
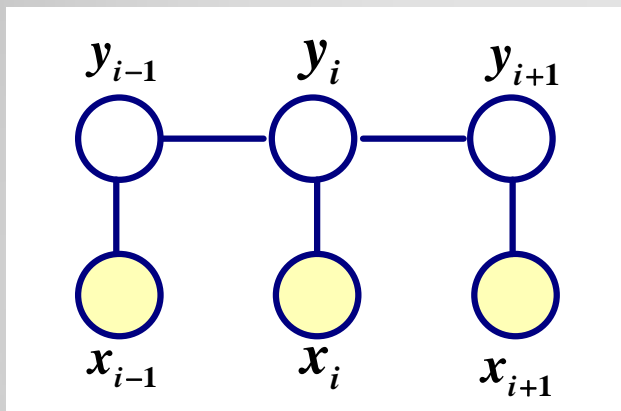
- 简单地讲，随机场可以看成是一组随机变量的集合（这组随机变量对应同一个样本空间）。当给每一个位置按照某种分布随机赋予一个值之后，其全体就叫做**随机场**。
- 当然，这些随机变量之间可能有依赖关系，一般来说，也只有当这些变量之间有依赖关系的时候，我们将其单独拿出来看成一个随机场才有实际意义。
- **马尔科夫随机场 ( MRF )** 对应一个无向图。这个无向图上的每一个节点对应一个随机变量，节点之间的边表示节点对应的随机变量之间有概率依赖关系。因此，MRF的结构本质上反应了我们的先验知识——哪些变量之间有依赖关系需要考虑，而哪些可以忽略。
- 具有马尔科夫性质：离当前因素比较遥远(这个遥远要根据具体情况自己定义)的因素对当前因素的性质影响不大。

- 现在，如果给定的MRF中每个随机变量下面还有观察值，我们要确定的是给定观察集合下，这个MRF的分布，也就是条件分布，那么这个MRF就称为CRF。它的条件分布形式完全类似于MRF的分布形式，只不过多了一个观察集合 $x$ 。
- 最通用角度来看，CRF本质上是给定了观察值 (observations)集合的MRF。

## CRF定义：

设  $G = (V, E)$  是一个无向图， $Y = \{Y_v | v \in V\}$  是以  $G$  中节点  $v$  为索引的随机变量  $Y_v$  构成的集合。在给定  $X$  的条件下，如果每个随机变量  $Y_v$  服从马尔可夫属性，即  $p(Y_v | X, Y_u, u \neq v) = p(Y_v | X, Y_u, u \sqcap v)$ ，则  $(X, Y)$  就构成一个条件随机场。

最简单且最常用的是一阶链式结构，即线性链结构（Linear-chain CRFs）



## Linear-chain CRFs 模型：

令  $x = \{x_1, x_2, \dots, x_n\}$  表示观察序列， $y = \{y_1, y_2, \dots, y_n\}$  是有限状态的集合，

根据随机场的基本理论：

$$p(y|x, \lambda) \propto \exp \left( \sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i) \right)$$

$t_j(y_{i-1}, y_i, x, i)$ ：对于观察序列的标记位置  $i-1$  与  $i$  之间的转移特征函数

$s_k(y_i, x, i)$ ：观察序列的  $i$  位置的状态特征函数

将两个特征函数统一为： $f_j(y_{i-1}, y_i, x, i)$

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp \left( \sum_{i=1}^n \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i) \right)$$

$$Z(x) = \sum_j \exp \left( \sum_{i=1}^n \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i) \right)$$

# 关键问题

## 1. 特征函数的选择

特征函数的选取直接关系模型的性能。

## 2. 参数估计

从已经标注好的训练数据集学习条件随机场模型的参数，即各特征函数的权重向量 $\lambda$ 。

## 3. 模型推断

在给定条件随机场模型参数 $\lambda$ 下，预测出最可能的状态序列。

# 1.特征函数的选择

CRFs模型中特征函数的形式定义： $f_j(y_{j-1}, y_i, x, i)$

它是状态特征函数和转移特征函数的统一形式表示。特征函数通常是二值函数，取值要么为1要么为0。

在定义特征函数的时候，首先构建观察值上的真实特征 $b(x, i)$ 的集合，即所有 $i$ 时刻的观察值 $x$ 的真实特征，结合其对应的标注结果，就可以获得模型的特征函数集。

$$b(x, i) = \begin{cases} 1 & \text{如果时刻 } i \text{ 观察值 } x \text{ 是大写开头} \\ 0 & \text{否则} \end{cases}$$

$$f(y_{i-1}, y_i, x, i) = \begin{cases} b(x, i) & \text{if } y_{i-1} = \text{< title >, } y_i = \text{< author >} \\ 0 & \text{otherwise} \end{cases}$$

## 2. 参数估计

### 极大似然估计 ( Maximum Likelihood Estimation , MLE)

假定对于训练数据有一组样本集合  $D = \{x^{(j)}, y^{(j)}\}, \forall j = 1, \dots, N$ ,

样本是相互独立的,  $\tilde{p}(x, y)$  为训练样本中  $(x, y)$  的经验概率,

对于某个条件模型  $p(y|x, \Theta)$ , 训练数据D的似然函数公式为:

$$L(\Theta) = \prod_{x,y} p(y|x, \Theta)^{\tilde{p}(x,y)}$$

取对数形式:

$$L(\Theta) = \sum_{x,y} \tilde{p}(x, y) \log p(y|x, \Theta)$$

CRFs模型中极大似然函数：

$$L(\lambda) = \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^n \left( \sum_j \lambda_j f_j((y_{i-1}, y_i, x, i)) \right) - \sum_x \tilde{p}(x) \log Z(x)$$

$$L(\lambda) = \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^n \lambda f - \sum_x \tilde{p}(x) \log Z(x)$$

对  $\lambda_j$  求导：

$$\begin{aligned} \frac{\partial L(\lambda)}{\lambda_j} &= \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^n f_j(y_{i-1}, y_i, x) - \sum_{x,y} \tilde{p}(x) p(y|x, \lambda) \sum_{i=1}^n f_j(y_{i-1}, y_i, x) \\ &= E_{\tilde{p}(x,y)} [f_j(x, y)] - \sum_k E_{p(y|x^{(k)}, \lambda)} [f_j(x^{(k)}, y)] \end{aligned}$$

令上式等于0，求 $\lambda$

模型分布中特征的期望等于经验分布中的期望值——最大熵原理



## 1、迭代缩放

Lafferty提出两个迭代缩放的算法用于估计条件随机场的极大似然参数

- GIS算法 ( Generalised Iterative Scaling )
- IIS算法 ( Improved Iterative Scaling )

**迭代缩放**是一种通过更新规则以更新模型中的参数，通过迭代改善联合或条件模型分布的方法。更新规则如下：

$$\lambda_j \leftarrow \lambda_j + \delta\lambda_j$$

其中更新值  $\delta\lambda_j$  使得新的值  $\lambda_j$  比原来的值  $\lambda_j$  更接近极大似然值。

## 迭代缩放的基本原理

假定我们有一个以  $\lambda = \{\lambda_1, \lambda_2, \dots\}$  为参数的模型  $p(y|x, \lambda)$ ，并且要找到一组新的参数： $\lambda + \Delta = \{\lambda_1 + \delta\lambda_1, \lambda_2 + \delta\lambda_2, \dots\}$  使得在该参数条件下的模型具有更高的对数似然值。通过迭代，使之最终达到收敛。

对于条件随机场对数似然值的变化可以表示为：

$$\begin{aligned} L(\lambda + \Delta) - L(\lambda) &= \sum_{x,y} \tilde{p}(x,y) \log p(y|x, \lambda + \Delta) - \sum_{x,y} \tilde{p}(x,y) \log p(y|x, \lambda) \\ &= \sum_{x,y} \tilde{p}(x,y) \left[ \sum_{i=1}^n \sum_j \delta\lambda_j f_j(y_{i-1}, y_i, x) \right] - \sum_x \tilde{p}(x) \log \frac{Z_{\lambda+\Delta}(x)}{Z_{\lambda}(x)} \end{aligned}$$

引入辅助函数：

$$A(\lambda, \Delta) = \sum_{x, y} \tilde{p}(x, y) \left[ \sum_{i=1}^n \sum_j \delta \lambda_j f_j(y_{i-1}, y_i, x) \right] + 1 \\ - \sum_x \tilde{p}(x) p(y|x, \lambda) \left[ \sum_{i=1}^n \sum_j \left( \frac{f_j(y_{i-1}, y_i, x)}{T(x, y)} \right) \exp(\delta \lambda_j T(x, y)) \right]$$

$$T(x, y) = \sum_{i=1}^n \sum_j f_j(y_{i-1}, y_i, x)$$

定义为在观察序列和标记序列为 $(x, y)$ 的条件下，特征值为1的特征的个数。

根据  $L(\lambda + \Delta) - L(\lambda) \geq A(\lambda, \Delta)$ ，寻找使  $A(\lambda, \Delta)$  最大化的  $\Delta$ ，  
使用迭代算法计算最大似然参数集。

迭代过程：

( A ) 将每个  $\lambda_j$  设初始值；

( B ) 对于每个  $\lambda_j$ ，计算  $\frac{\partial A(\lambda, \Delta)}{\partial \delta \lambda_j} = 0$ ，即

$$\frac{\partial A(\lambda, \Delta)}{\partial \delta \lambda_j} = \sum_{x, y} \tilde{p}(x, y) \sum_{i=1}^n f_j(y_{i-1}, y_i, x)$$

$$- \sum_x \tilde{p}(x) \sum_y p(y|x, \lambda) \sum_{i=1}^n f_j(y_{i-1}, y_i, x) \exp(\delta \lambda_j T(x, y)) = 0$$

应用更新规则  $\lambda_j \leftarrow \lambda_j + \delta \lambda_j$ ，更新每个参数，直到收敛。

## GIS算法：

GIS是迭代缩放的一种，为了确保参数收敛的结果达到全局最优，GIS需要对特征集进行约束，即令每个训练数据中的事件  $T(x, y) = C$ 。

定义了一个全局修正特征  $S(x, y)$ ：

$$S(x, y) \square C - \sum_{i=1}^n \sum_j f_j(y_{i-1}, y_i, x)$$

其中C是训练语料中所有的x和y情况下  $T(x, y)$  的最大值，即等于最大可能的特征个数，特征  $S(x, y)$  的加入确保了  $T(x, y) = C$ 。

假定对于所有的事件，条件随机场选定的特征的总和是常量C。

更新值按下式计算

$$\delta \lambda_j = \frac{1}{C} \log \left( \frac{E_{\tilde{p}(x, y)} [f_k]}{E_{p(y|x, \lambda)} [f_k]} \right)$$

$$\delta\lambda_j = \frac{1}{C} \log \left( \frac{E_{\tilde{p}(x,y)}[f_k]}{E_{p(y|x,\lambda)}[f_k]} \right)$$

$$E_{\tilde{p}(x,y)}[f_k] = \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^n f_j(y_{i-1}, y_i, x)$$

$$E_{p(y|x,\lambda)}[f_k] = \sum_x \tilde{p}(x) \sum_y p(y|x,\lambda) \sum_{i=1}^n f_j(y_{i-1}, y_i, x) \exp(\delta\lambda_j T(x,y))$$

**问题：**

1. GIS算法的收敛速度由计算更新值的步长确定。C值越大，步长越小，收敛速度就越慢；反之C值越小，步长越大，收敛的速度也就越快。

2. GIS算法是依赖于一个额外的全局修正特征 $S(x,y)$ ，以确保对于每个 $(x,y)$ 对的有效特征的总和是一个常量。但是一旦加入这个新的特征，就认为这个特征和特征集中所有其他的特征之间是相互独立的，并且它的参数也需要使用上式来更新。计算期望需要对所有可能的标记序列求和，这将是一个指数级的计算过程。

IIS算法：

重新定义：
$$T(x, y) \approx T(x) \sqcap \max_y T(x, y)$$

将每个对观察序列和标记序列对(x,y)起作用的特征值的和近似等于对于观察序列x的最大可能的观察特征的和

$$E_{p(y|x, \lambda)}[f_k] = \sum_{m=0}^{T_{\max}} a_{k,m} \exp(\delta \lambda_j)^m$$

$$a_{k,m} = \sum_x \tilde{p}(x) \sum_y p(y|x, \lambda) \sum_{i=1}^n f_k(y_{i-1}, y_i, x) \delta(m, T(x))$$

使用牛顿—拉夫森方法求解

## 2、梯度算法

L-BFGS算法：

$$\frac{\partial L(\lambda)}{\lambda_j} = E_{\tilde{p}(x,y)} [f_j(x,y)] - \sum_k E_{p(y|x^{(k)},\lambda)} [f_j(x^{(k)},y)] - \frac{\lambda_k}{\sigma^2}$$

Dong C. Liu and Jorge Nocedal : 【On The Limited Memory BFGS Method For Large Scale Optimization】

Jorge Nocedal用Fortran语言实现了L-BFGS工具包来进行条件随机场的参数估计与训练，该数学工具包可从

<http://www.ece.northwestern.edu/~nocedal/>下载。

另外，Taku Kudo实现了L-BFGS算法的c语言版本，该工具集成在了其开发的CRF++工具包中，网址为

<http://www.chasen.org/~taku/software/CRF++/>。



### 3.模型推断

常见的两个问题：一、在模型训练中，需要边际分布  $p(y_t, y_{t-1} | x)$  和  $Z(x)$  ；  
二、对于未标记的序列，求其最可能的标记。

第一个问题采用前向后向法解决；

第二个问题通过Viterbi算法解决。Viterbi算法是一种动态规划算法，其思想精髓在于将全局最佳解的计算过程分解为阶段最佳解的计算。

## 最大熵马尔科夫模型举例——基于文本的网络地址信息抽取

<P>公司:青岛银河钢木家具厂</P><P>地址:青岛市重庆南路 247 号</P>

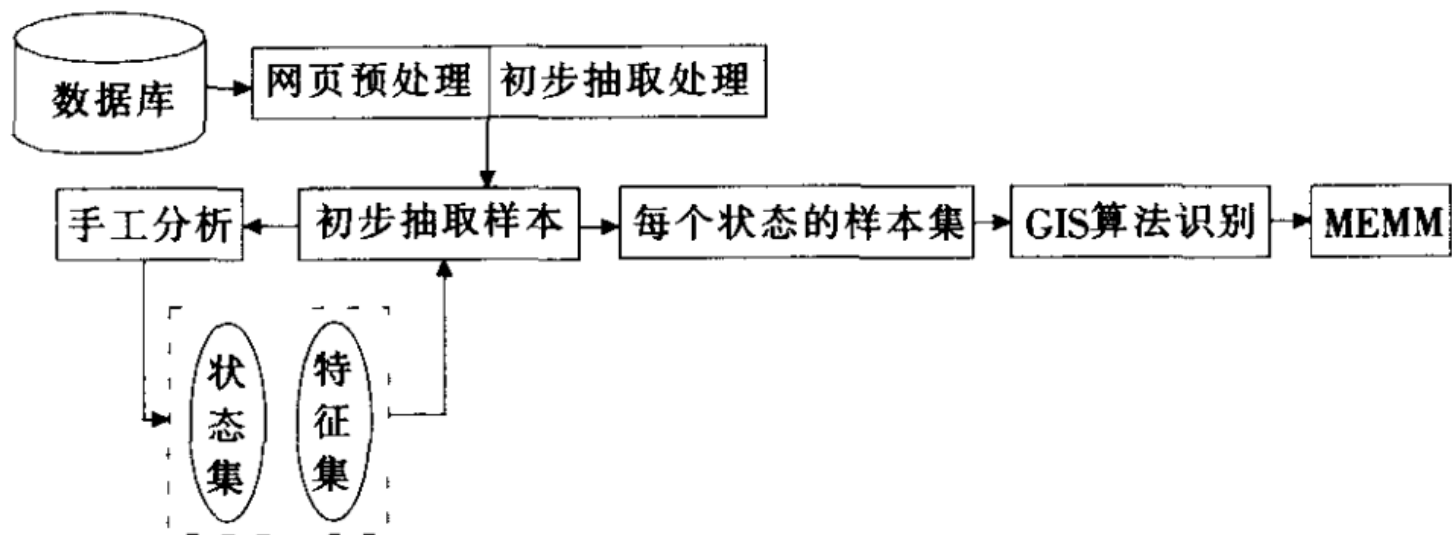
<P>电话:0532-5032359</P><P>传真:0532-5032263</P>

<P>Email:<A HREF="mailto:daiguoli@yhfnt.com">daiguoli@yhfnt.com</FONT></A></P>

<div align="center">地址:胶南市泰薛路中端 邮编:266401

电话:0532-6151696 传真:0532-6151333 手机:13589367999<br><div>

任务:完成地址,电话,传真,E-mail 等信息的识别和抽取



流程图

## 页面预处理

#地址:胶南市泰薛路中端# 邮编:266401#电话:0532-6151696#  
传真:0532-6151333# 手机:13589367999#<http://www.hengrun-qd.cn>#  
E-mail:qdjntlx@sohu.com#hengrun@hengrun-qd.cn#

页面文本中加入#用于保留结构信息和页面内容的自然划分,便于对文本页面的进一步处理。

## 模型建立

确定状态集合Y，观察值（特征）集合X

状态集合包含：邮编、电话、电邮、地址、联系人、账号、手机、网址、传真，对于其他可能出现的状态定义了“other”来代表。

特征集合包含：“具有@符号”

“最大数字串长度为6”

“最大数字串长度为11”

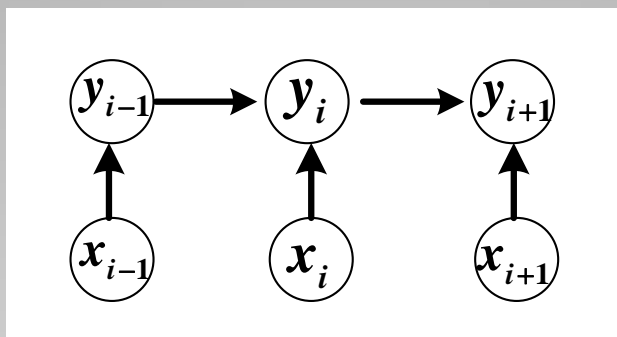
“最大数字串长度介于6到11”

“最大数字长度大于15”

“最大数字长度小于6，字符串总长度大于30”

“最大数字长度小于6，字符串总长度介于8到30”

“最大数字长度小于6，字符串总长度小于6”，.....



$$p_{y_{i-1}}(y_i | x_i) = \frac{1}{Z(x_i, y_{i-1})} \exp \left\{ \sum_a \lambda_a f_a(x_i, y_i) \right\}$$

特征函数  $f_a(x, y)$  表示数据集  $\langle X, Y \rangle$  的特性：

$$f_a(x, y) = \begin{cases} 1 & \text{如果 } x \text{ 只含有6位数字 \& } y = \text{邮编} \\ 0 & \text{其他} \end{cases}$$

进一步引入一系列的特征函数  $\{f_1, f_2, \dots, f_n\}$

参数学习

$$p_{y_{i-1}}(y_i | x_i) = \frac{1}{Z(x_i, y_{i-1})} \exp \left\{ \sum_a \lambda_a f_a(x_i, y_i) \right\}$$

用上述的状态和特征集对初步抽取样本进行统计，得到每个状态所对应的样本集，通过对于每个这样的样本集合采用 GIS算法进行参数学习，最终得到 MEMM。

说明：

GIS算法要求对于每一个  $\langle x, y \rangle$ ，特征之和达到一个常数  $C$ ，即有

$$\sum_{i=1}^n f_i(x, y) = C$$

如果不满足，则令  $C = \max_{\langle x, y \rangle} \sum_{i=1}^n f_i(x, y)$

并加入一个修正函数，使得  $f_{n+1}(x, y) = C - \sum_{i=1}^n f_i(x, y)$

## GIS算法的步骤：

1. 初始  $\lambda_a^{(0)} = 1, a \in \{1, 2, \dots, n+1\}$

2.  $\forall a \in \{1, 2, \dots, n+1\},$

(a) 计算每个特征的  $\tilde{E}_a = \sum_{x,y} f_a(x_t, y_t);$

(b)  $p_{y_{i-1}}^{(j)}(y_i | x_t) = \frac{1}{Z(x_t, y_{i-1})} \exp \left\{ \sum_a \lambda_a f_a(x_t, y_i) \right\}$

(c) 用当前的  $\lambda_a$  值计算  $E_a^j = \sum_{x_t, y_i} p_{y_{i-1}}^{(j)}(y_i | x_t) f_a(x_t, y_i)$

(d) 更新  $\lambda_a^{(j+1)} = \lambda_a^{(j)} + \frac{1}{C} \log \left( \frac{\tilde{E}_a}{E_a^{(j)}} \right)$

(e) 满足收敛条件，结束；否则转到(b)

通过GIS算法得到状态转移函数，这些状态转移函数的集合组成了MEMM模型

## 识别和抽取

### 改进的Viterbi算法

( 1 ) 输入观察值序列  $x_1, x_2, \dots, x_T$

( 2 ) 递归

$$V_t(y_i) = \max_{1 \leq r \leq N} V_{t-1}(y_r) p_{y_r}(y_i | o_t)$$

$$Q_t(y_i) = \arg \max_{1 \leq r \leq N} V_{t-1}(y_r) p_{y_r}(y_i | x_t)$$

( 3 ) 结束  $V^* = \max_{1 \leq i \leq N} V_T(y_i)$

$$Q^* = \arg \max_{1 \leq i \leq N} Q_T(y_i)$$



## 评测指标

$$\text{召回率 ( Recall )} = \frac{\text{正确识别出的实体个数}}{\text{标准结果中实体的总数}} \times 100\%$$

$$\text{精确率 ( Precision )} = \frac{\text{正确识别出的实体个数}}{\text{识别出的实体总数}} \times 100\%$$

**关键：**特征的选择

## 条件随机场模型举例——中文命名实体识别

在中文信息处理领域，命名实体识别是各种自然语言处理技术的重要基础。

命名实体：人名、地名、组织名三类

模型形式

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp \left( \sum_{i=1}^n \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i) \right)$$

$$Z(x) = \sum_j \exp \left( \sum_{i=1}^n \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i) \right)$$

**关键：**特征函数的确定

适用于人名的特征模板

“上下文”，指的是包括当前词 $w_0$ 及其前后若干个词的一个“观察窗口” ( $w_{-n}, w_{-n+1}, \dots, w_0, \dots, w_n$ )。理论上来说，窗口越大，可利用的上下文信息越多，但窗口开得过大除了会严重降低运行效率，还会产生过拟合现象；而窗口过小，特征利用的就不够充分，会由于过于简单而丢失重要信息。

通过一些模板来筛选特征。模板是对上下文的特定位置和特定信息的考虑。

“人名的指界词”：主要包括称谓词、动词和副词等，句首位置和标点符号也可。

根据指界词与人名同现的概率的大小，将人名的左右指界词各分为两级，生成4个人名指界词列表：

类型	级别	列表名称	举例
左指界词	1 级	PBW1	记者、纪念
	2 级	PBW2	称赞、叮咛
右指界词	1 级	PAW1	报道、会见
	2 级	PAW2	供认、坚决

还建立了若干个资源列表，包括：中国人名姓氏用表、中国人名名字用表、欧美俄人名常用字表、日本人名常用字表。

定义了用于人名识别特征的原子模板，每个模板都只考虑了一种因素：

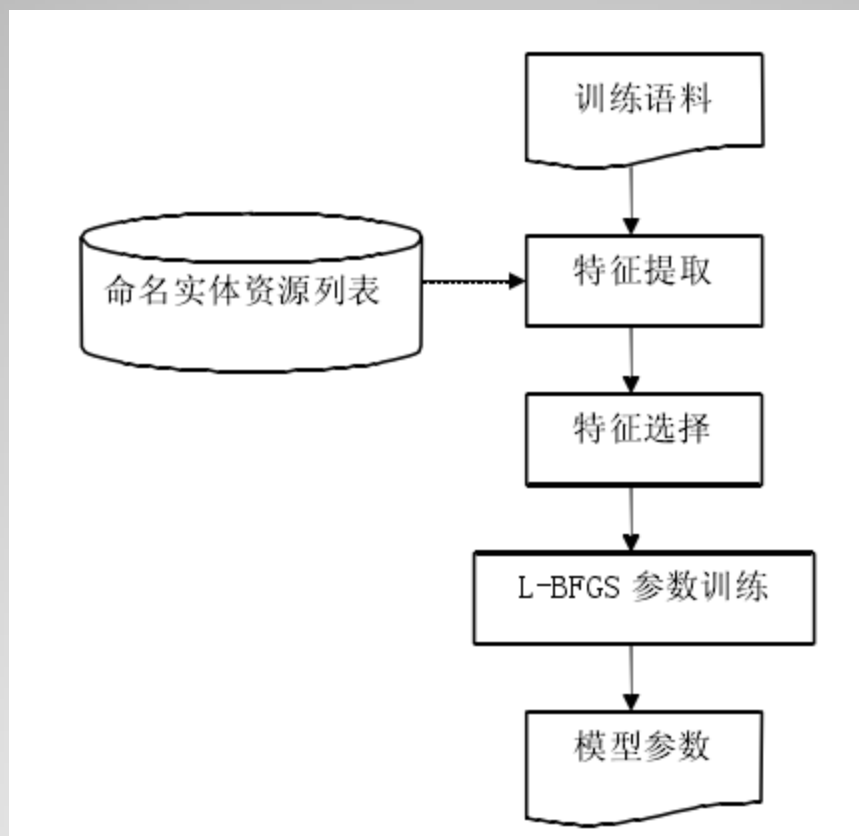
序号	原子模板	意义
P1	ChSurName	当前词是否为中国人名姓氏用字
P2	ChLastName	当前词是否为中国人名名字用字
P3	EurName	当前词是否为欧美俄人名常用字
P4	JapName	当前词是否为日本人名常用字
P5	PerFirRightBoundary	当前词后面第一个词是否为右指界词（1、2级）
P6	PerSecRightBoundary	当前词后面第二个词是否为右指界词（1、2级）
P7	PerFirLeftBoundary	当前词前面第一个词是否为左指界词（1、2级）
P8	PerSecLeftBoundary	当前词前面第二个词是否为左指界词（1、2级）

当特征函数取特定值时，特征模板被实例化就可以得到具体的特征。

“当前词的前一个词 $w_{-1}$ 在人名1级左指界词列表中出现”

$$f_i(x, y) = \begin{cases} 1 & \text{If PBW1}(w_{-1}) = \text{ture and } y = \text{person} \\ 0 & \text{else} \end{cases}$$

类似的，做地名、组织名的特征提取和选择，并将其实例化，得到所有的特征函数。



模型训练流程图

## 评测指标

$$\text{召回率 ( Recall )} = \frac{\text{正确识别的命名实体首部 ( 尾部 ) 的个数}}{\text{标准结果中命名实体首部 ( 尾部 ) 的总数}} \times 100\%$$

$$\text{精确率 ( Precision )} = \frac{\text{正确识别的命名实体首部 ( 尾部 ) 的个数}}{\text{识别出的命名实体首部 ( 尾部 ) 的总数}} \times 100\%$$

$$\text{F-值} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}}$$

## 整体评价：

**优点：**条件随机场模型既具有判别式模型的优点，又具有产生式模型考虑到上下文标记间的**转移概率**，以序列化形式进行**全局参数优化**和解码的特点，解决了其他判别式模型(如最大熵马尔科夫模型)难以避免的**标记偏见**问题。

**缺点：**模型训练时收敛速度比较慢



## CRF的研究进展：

2001 年，卡耐基·梅隆大学的 Lafferty 教授针对序列数据处理提出了 CRF 模型。【Conditional random fields- Probabilistic models for segmenting and labeling sequence data】

2003 年，Kumar 博士将 CRF 模型扩展到 2-维格型结构，开始将其引入到图像分析领域，吸引了学术界的高度关注。

Ariadna Quattoni   Michael Collins   Trevor Darrell  
【Conditional Random Fields for Object Recognition】

Asela Gunawardana等人

【Hidden Conditional Random Fields for Phone Classification】

2007年，Charles Sutton , Andrew McCallum

【Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data】

## 基础的参考文献：

### 概率图

【*An Introduction to Variational Methods for Graphical models*】

【*Classical Probabilistic Models and Conditional Random Fields*】

### 经典概率模型与CRF

【*An Introduction to Conditional Random Fields for Relational Learning*】

【*Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*】

【*Operations for learning with Graphical models*】

## 有用的参考文献：

【*Efficient Training of Conditional Random Fields*】

【*Efficiently Inducing features of random fields*】

【*A maximum entropy approach to natural language processing*】

【*Multiscale Conditional Random Fields for Image Labeling*】

【*Training Conditional Random Fields via Gradient Tree Boosting*】

# CRF的发展方向

## 1. 复杂拓扑结构的CRF

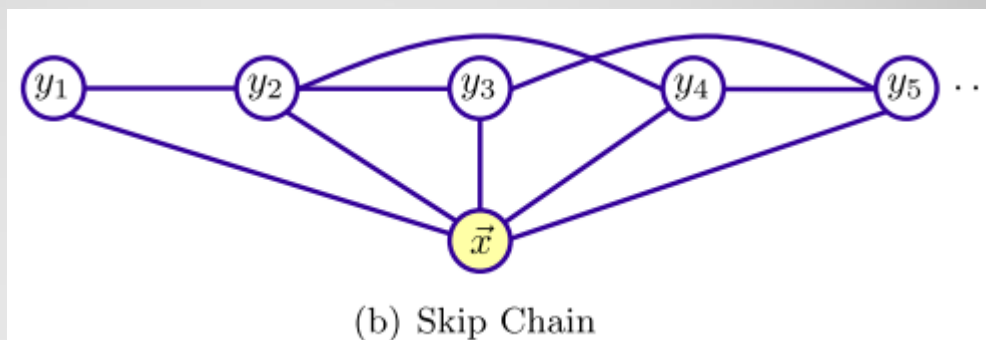
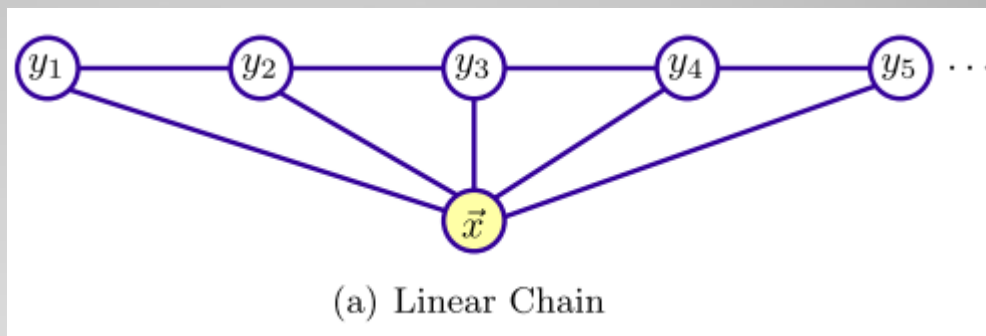
Linear Chain



Skip Chain



Arbitrarily structured CRFs



# CRF的发展方向

1. 复杂拓扑结构的CRF
2. 模型训练和推断的快速算法
3. CRF模型特征的选择和归纳