七、 CRF

Lafferty, J., McCallum, A., Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data".

应用领域：text processing, text classification, computer vision, bioinformatics…

Discriminative Model

概率无向图模型 PUGM MRF

- 有向图模型，又称作贝叶斯网络(Directed Graphical Models, DGM, Bayesian Network)。

- 使用没有方向的无向边，形成了无向图模型(Undirected Graphical Model, UGM)，又被称为马尔科夫随机场或者马尔科夫网络(Markov Random Field, MRF or Markov network)

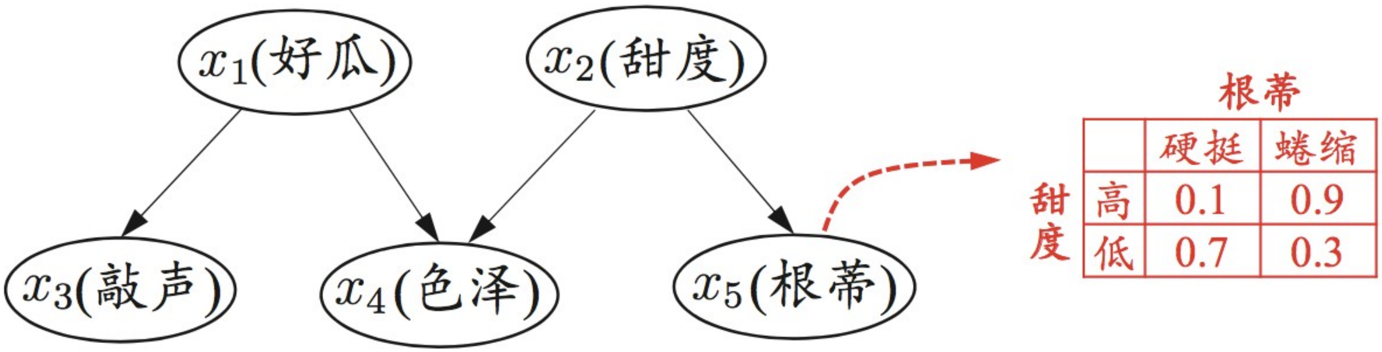
MRF 和 CRF

- 设 $X=(X_1,X_2 \ldots X_n)$ 和 $Y=(Y_1,Y_2 \ldots Y_m)$ 都是联合随机变量，若随机变量 Y 构成一个无向图 G=(V, E) 表示的马尔科夫随机场(MRF)，则条件概率分布 P(Y|X) 称为条件随机场 (Conditional Random Field, CRF)

- *注：大量文献将MRF和CRF混用，包括经典著作。*

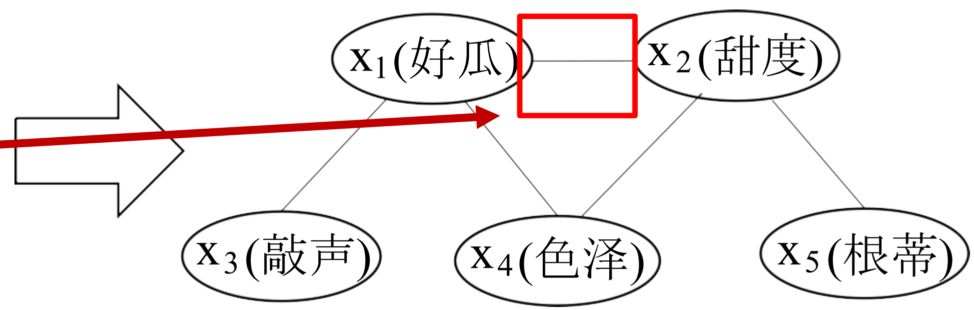
有向分离

“有向分离”(D-separation)



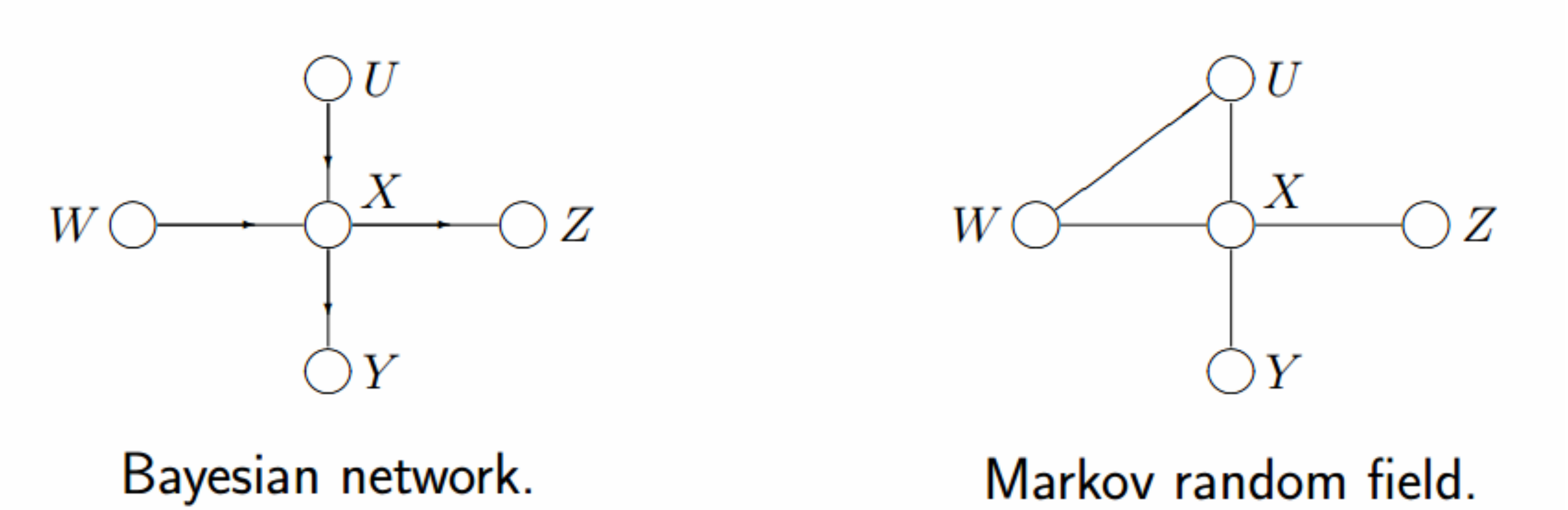
先将有向图转变为无向图

- V 型结构父结点相连
- 有向边变成无向边

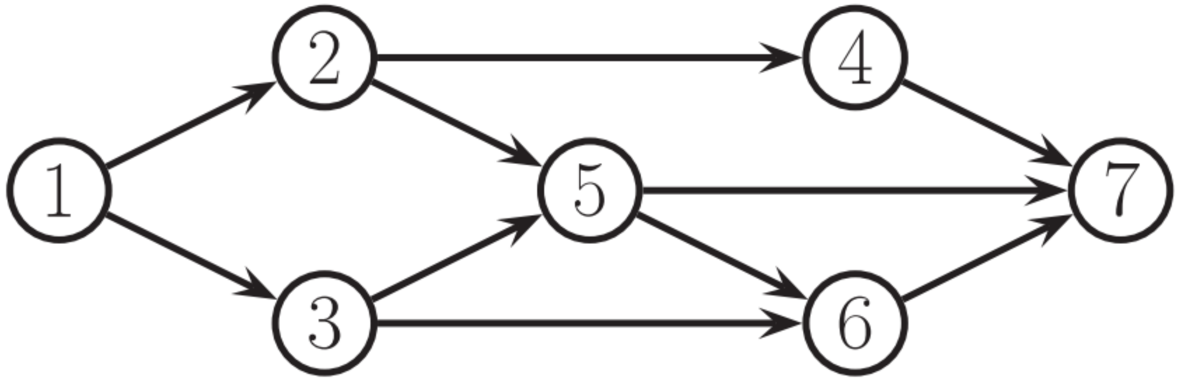


DGM转换成UGM

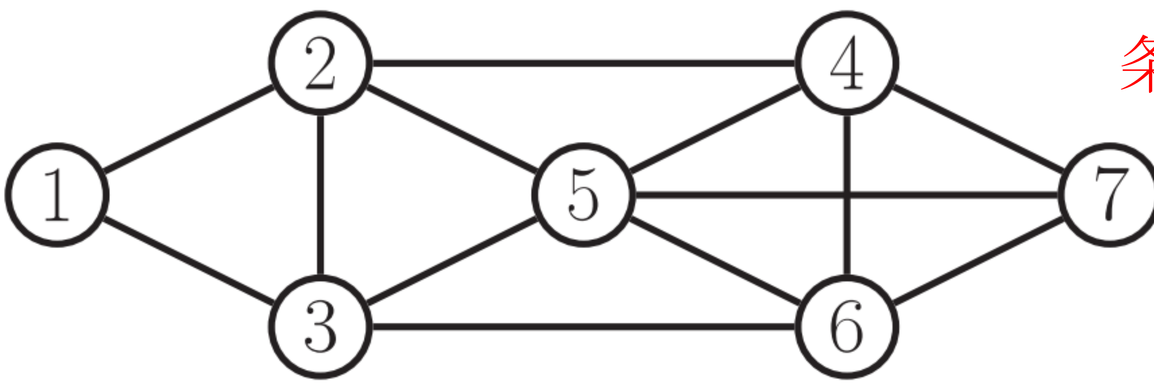
- V 型结构父结点相连
- 有向边变成无向边



DGM转换成UGM



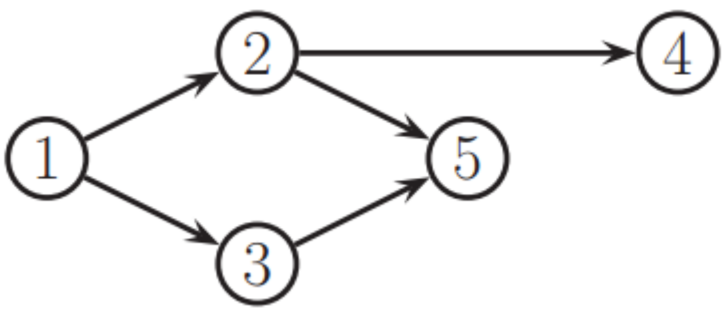
$$4 \perp 5 \mid 2$$

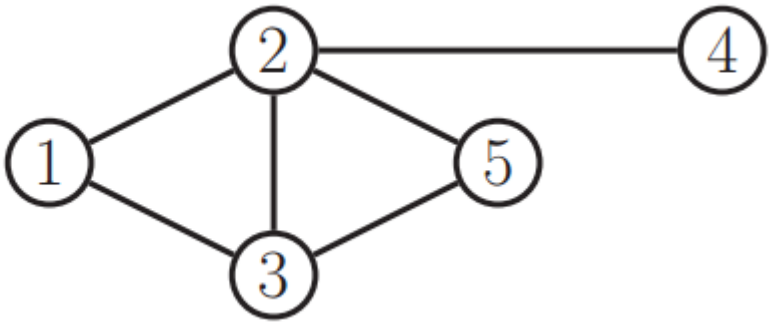


条件独立的破坏

条件独立的破坏

- 靠考察是否有 $A \perp B|C$ ，则计算 U 的祖先图(ancestral graph) : $U = A \cup B \cup C$



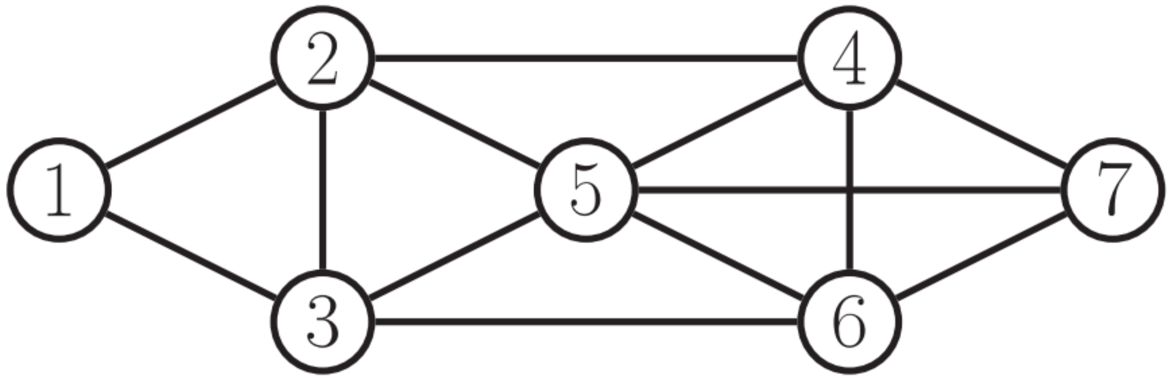


$$4 \perp 5|2$$

MRF的性质

- 成对马尔科夫性
  - parewise Markov property
- 局部马尔科夫性
  - local Markov property
- 全局马尔科夫性
  - global Markov property
- 表述说明：随机变量 $Y=(Y_1,Y_2...Y_m)$ 构成无向图$G=(V,E)$，结点 v 对应的随机变量是 Y_v。
- (**概率无向图模型**) 设联合概率分布 P(Y)，由无向图 G=(V,E) 表示，节点表示随机变量，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、局部或全局马尔科夫性，就称此联合概率分布为概率无向图模型 (Probability Undirected Graphical Model, PUGM)，或者马尔科夫随机场（马尔科夫网络）。

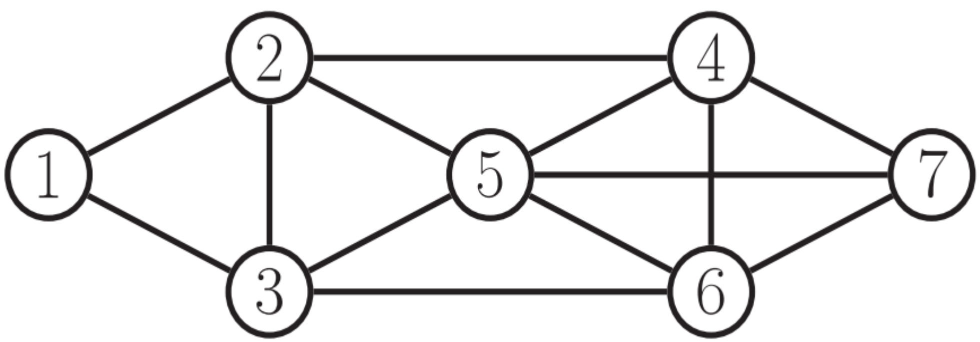
考察结点间的独立性



Pairwise	$1 \perp 7 \mid \text{rest}$
Local	$1 \perp \text{rest} \mid 2, 3$
Global	$1, 2 \perp 6, 7 \mid 3, 4, 5$

成对马尔科夫性

- 设 u 和 v 是无向图 G 中任意两个没有边直接连接的结点，G 中其他结点的集合记做 O；则在给定随机变量 Yo 的条件下，随机变量 Yu 和 Yv 条件独立。
- 即：P(Yu,Yv|Yo)= P(Yu|Yo)* P(Yv|Yo)



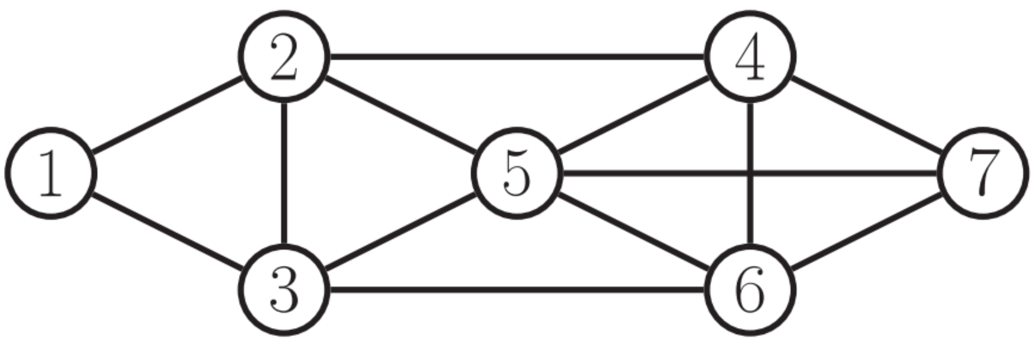
Pairwise $1 \perp 7 | \text{rest}$

Local $1 \perp \text{rest} | 2, 3$

Global $1, 2 \perp 6, 7 | 3, 4, 5$

局部马尔科夫性

- 设 v 是无向图 G 中任意一个结点， W 是与 v 有边相连的所有结点， G 中其他结点记做 O ；则在给定随机变量 Yw 的条件下，随机变量 Yv 和 Yo 条件独立。
- 即：P(Yv,Yo|Yw)= P(Yv|Yw)* P(Yo|Yw)



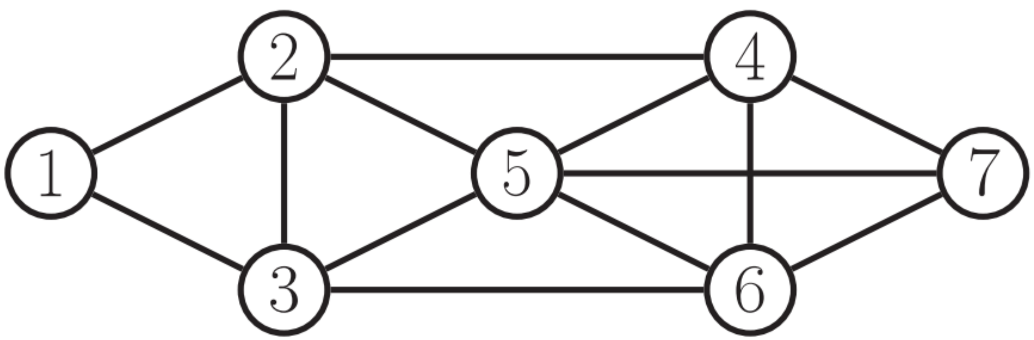
Pairwise $1 \perp 7 | \text{rest}$

Local $1 \perp \text{rest} | 2, 3$

Global $1, 2 \perp 6, 7 | 3, 4, 5$

全局马尔科夫性

- 设结点集合 A， B 是在无向图 G 中被结点集合 C 分开的任意结点集合，则在给定随机变量 Y_C 的条件下，随机变量 Y_A 和 Y_B 条件独立。
- 即：$P(Y_A, Y_B \mid Y_C) = P(Y_A \mid Y_C) * P(Y_B \mid Y_C)$



Pairwise $1 \perp 7 \mid \text{rest}$

Local $1 \perp \text{rest} \mid 2, 3$

Global $1, 2 \perp 6, 7 \mid 3, 4, 5$

三个性质的等价性

$$\begin{tikzcd} G \arrow[r, Rightarrow] & L \arrow[r, Rightarrow] & P \arrow[ll, Rightarrow, bend left, "p(x) > 0"] \end{tikzcd}$$

- 根据全局马尔科夫性，能够得到局部马尔科夫性；
- 根据局部马尔科夫性，能够得到成对马尔科夫性；
- 根据成对马尔科夫性，能够得到全局马尔科夫性；

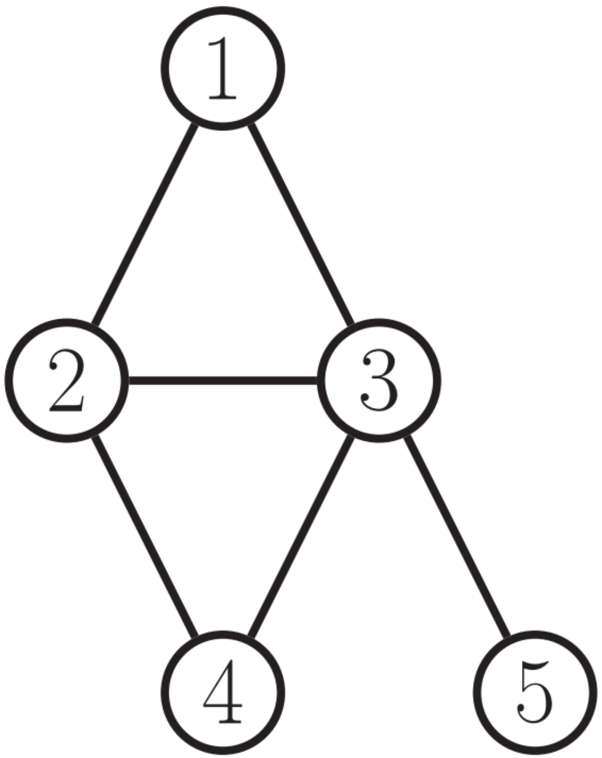
- 满足这三个性质(或其一)的无向图，称为概率无向图模型。

团

- 无向图 G 中任何两个结点均有边连接的子集，称作 G 的团(Clique)。

- 若 C 是 G 的一个团，并且不能再加入任何一个 G 的结点使其成为团，则 C 称作 G 的最大团(Maximal Clique)。

下图的最大团是什么？



- 无向图 G 中任何两个结点均有边连接的子集，称作G的团(Clique)。
- 若 C 是 G 的一个团，并且不能再加入任何一个 G 的结点使其称为团，则 C 称作 G 的最大团(Maximal Clique)。
- C1 = {1, 2, 3};
- C2 = {2, 3, 4}

Hammersley-Clifford定理

- UGM 的联合分布可以表示成最大团上的随机变量的函数的乘积的形式；这个操作叫做UGM的因子分解(Factorization)。
- UGM 的联合概率分布 P(Y) 可以表示成如下形式：

$$P(Y)=\frac{1}{Z}\prod_{c}\psi_{c}(Y_{c})$$

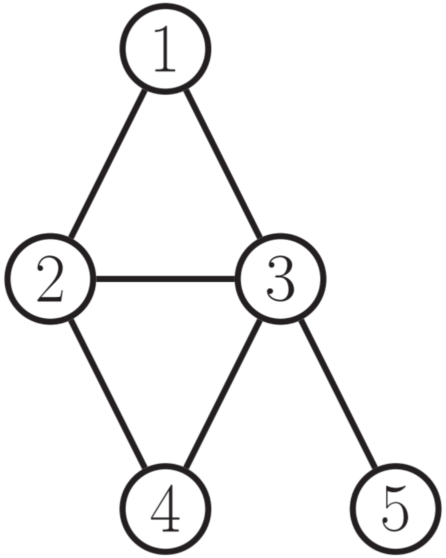
$$Z=\sum_{Y}\prod_{c}\psi_{c}(Y_{c})$$

MRF: Generative Model

- 其中，C 是 G 的最大团，$\psi_c(Y_c)$ 是 C 上定义的严格正函数（通常定义为指数函数），乘积是在 UGM 所有的最大团上进行的，被称作势函数。(Potential Function)。

Hammersley-Clifford定理

- UGM 的联合分布可以表示成最大团上的随机变量的函数的乘积的形式；这个操作叫做UGM的因子分解(Factorization)。



$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\underline{\psi_{123}(y_1, y_2, y_3)}\,\underline{\psi_{234}(y_2, y_3, y_4)}\,\underline{\psi_{35}(y_3, y_5)}$$

条件随机场

- (条件随机场) 设 $X=(X_1,X_2\ldots X_n)$ 和 $Y=(Y_1,Y_2\ldots Y_m)$ 都是随机变量，若随机变量 Y 构成一个无向图 G=(V, E) 表示的马尔科夫随机场，即：

$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v)$$

CRF: Discriminative Model

对任意节点 v 成立，则条件概率分布 P(Y|X) 称为条件随机场(Conditional Random Field, CRF)。其中， w~v 表示与结点 v 相连的所有结点 w，$w \neq v$ 表示 v 以外的所有节点。

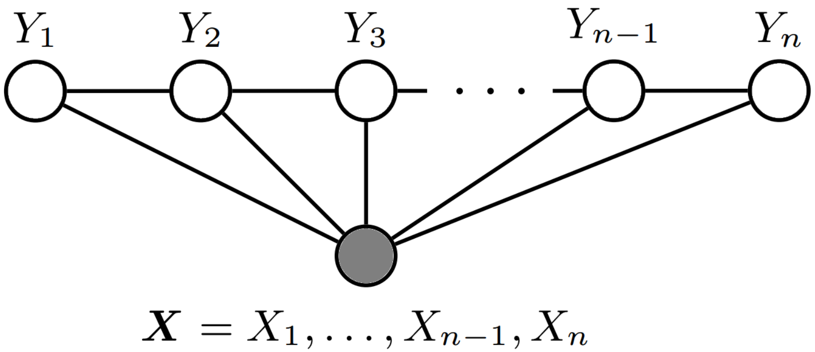
- 定义中并没有要求 X 和 Y 具有相同的结构。

条件随机场

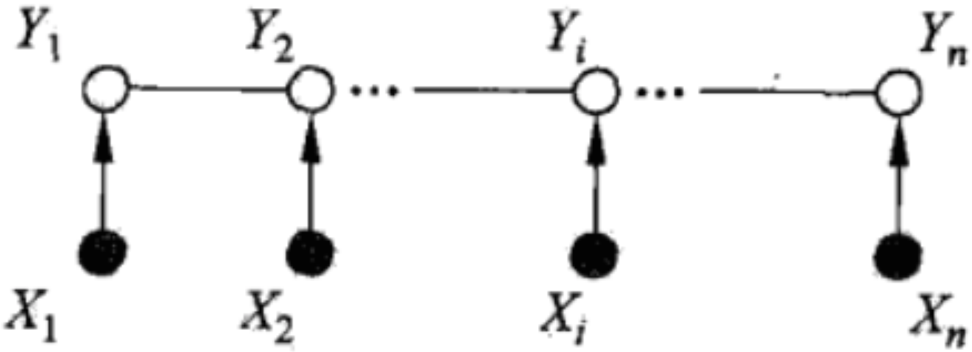
$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v)$$

- 这时，条件概率 P(Y|X) 中，Y 表示标记序列（或称状态序列），X 是需要标注的观测序列。
- 线性链条件随机场的无向图模型

最大团是相邻两个节点的集合



线性链条件随机场



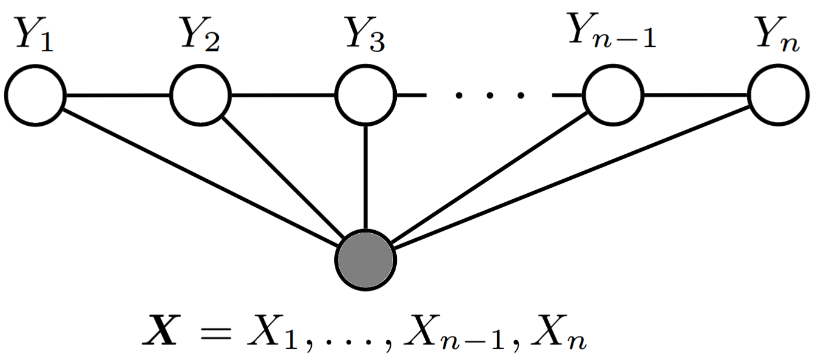
X和Y有相同的图结构的线性链条件随机场

线性链条件随机场的定义

- 设 $X=(X_1,X_2\ldots X_n)$，$Y=(Y_1,Y_2\ldots Y_n)$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，即满足马尔科夫性

$$P\left(Y_i \middle| X, Y_1, Y_2 \cdots Y_n\right) = P\left(Y_i \middle| X, Y_{i-1}, Y_{i+1}\right)$$

则称 P(Y|X) 为线性链条件随机场。在标注问题中，X 表示观测序列，Y 表示对应的输出标记序列或称状态序列。



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 的条件概率有以下形式 (给定输入序列 x, 对输出序列 y 预测的条件概率)：

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

其中， $Z = \sum_{y}\exp\left(\sum_{j}\sum_{i=1}^{n-1}\lambda_j t_j(y_{i+1}, y_i, x, i) + \sum_{k}\sum_{i=1}^{n}\mu_k s_k(y_i, x, i)\right)$

是规范化因子，求和是在所有可能的输出序列上进行的，t_j 和 s_k 是特征函数，λ_j 和 μ_k 是对应的权值。

条件随机场-特征函数

y	$\{y_1$	y_2	y_3	y_4	y_5	$y_6\}$
	[D]	[N]	[V]	[P]	[D]	[N]
x	$\{x_1$	x_2	x_3	x_4	x_5	$x_6\}$
	The	boy	knocked	at	the	watermelon.

- 条件概率可被定义为：

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

- $t_j(y_{i-1}, y_i, x, i)$是定义在观测序列的两个相邻标记位置上的转移特征函数(transition,边上的特征函数),用于刻画相邻标记变量之间的相关关系以及观测序列对它们的影响

$$t_k(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \text{if } y_{i-1} = [N], y_i = [V], \text{and } x_i = \text{“knocked”} \\ 0, & \text{otherwise.} \end{cases}$$

表示第i个观测值 x_i为单词'knocked'时，相应的标记 y_{i-1}, y_i 很可能分别为$[N]$，$[V]$.

条件随机场-特征函数

y $\{y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad y_6\}$

[D] [N] [V] [P] [D] [N]

x $\{x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6\}$

The boy knocked at the watermelon.

- 条件概率可被定义为：

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

- $s_k(y_i, x, i)$是定义在观测序列的标记位置i上的状态特征函数（status），用于刻画观测序列对标记变量的影响

$$s_k(y_i, \mathbf{x}, i) = \begin{cases} 1, & \text{if } y_i = [V] \text{ and } x_i = \text{“knock”}; \\ 0, & \text{otherwise}, \end{cases}$$

- 表示第i个观测值x_i为单词'knocked'时，它相对应的标记很可能为$[V]$.

参数说明

- t_k 和 s_l 都依赖于位置，是局部特征函数；
- 通常， t_k 和 s_l 取值为1或者0；满足特征条件时取1，否则取0；
- CRF 完全由特征函数 t_k、s_l 和对应的权值 λ_k、μ_l 确定。
- 线性链 CRF 属于对数线性模型 (log linear model)。

CRF的两个问题

- CRF的概率计算问题
  - 前向后向算法

- CRF的学习算法
  - 改进的迭代尺度算法 (IIS, Improved Iterative Scaling)
  - 拟牛顿法

- CRF的预测算法
  - Viterbi算法

参考文献

- Machine Learning:A Probabilistic Perspective, Kevin P. Murphy, The MIT Press, 2012
- Conditional Random Fields: An Introduction, Hanna M. Wallach, 2004
- An Introduction to Conditional Random Fields, Charles Sutton, Andrew McCallum, 2012
- 统计学习方法，李航著，清华大学出版社，2012年
- Pattern Recognition and Machine Learning Chapter 13, Bishop M, Springer-Verlag, 2006
- Radiner L,Juang B. An introduction of hidden markov Models. IEEE ASSP Magazine, January 1986