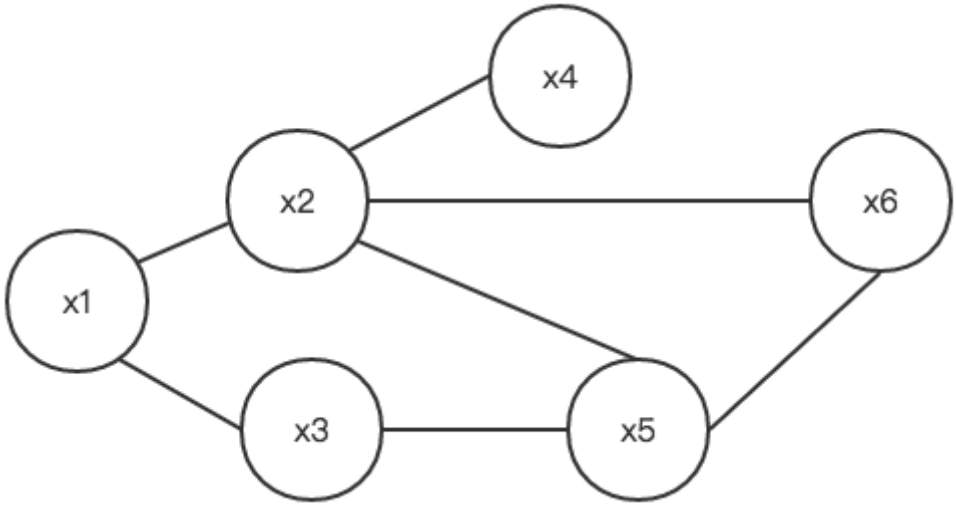
CRF

机器学习的一些算法

0. 补充知识

马尔可夫随机场

之前我们学习的概率有向图模型，比如HMM，称之为贝叶斯网络，那么对应的我们将概率无向图模型称之为马尔可夫网络或者马尔可夫随机场（MRF）。



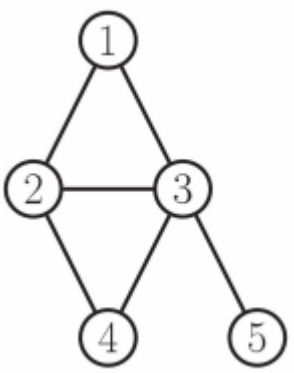
团

团（Clique）：图中节点的子集，其中任意两个节点之间都有边连接。

极大团

极大团：一个团，其中加入任何一个其他的节点都不能再形成团。

比如：



上图中，团有{1,2}, {1,3}, {2,3}, {2,4},{3,4}, {3,5},{1,2,3}, {2,3,4}
极大团有：{1,2,3}, {2,3,4}, {3,5}

势函数

势函数（Potential Function，又称为因子 Factor）：是定义在变量子集上的非负实函数，用于定义概率分布函数。

马尔可夫随机场中，**多个变量之间的联合概率分布可以基于团分解为多个势函数的乘积，每个势函数仅与一个团相关。**

Hammersley-Clifford 定理

对于 N 个变量的马尔可夫随机场，其变量为 $X = (X_1, X_2, \dots X_N)$（上图例子中 N=5）。

设所有团构成的集合为 C，与团 $Q \in C$ 对应的变量集合记作 X_Q，则联合概率为：

$$P(X) = \frac{1}{Z} \prod_{Q \in C} \Psi_Q(X_Q)$$

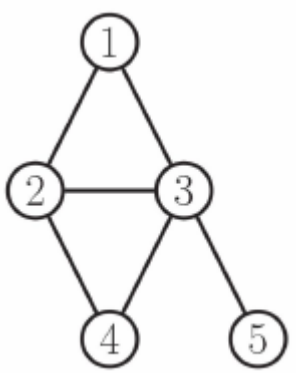
其中Ψ_Q为与团 Q 对应的势函数，用于对团 Q 中的变量关系进行建模。

Z 为规范化因子，很多时候要计算它很困难，不过好在大多数情况下，我们无须计算出 Z 的精确值。

当团 Q 不是极大团的时候，它必然属于某个极大团——实际上每一个非极大团都是如此，此时我们完全可以只用极大团来计算 P(X)：$P(X) = \frac{1}{Z^*} \prod_{Q \in C^*} \Psi_Q(X_Q)$，其中 C^*为所有极大团的集合。

这叫做 Hammersley-Clifford 定理，是随机场（Random Fields）的基础定理，它给出了一个马尔可夫随机场被表达为**正概率分布**的充分必要条件。

举个例子：
比如刚刚看过的这张图：

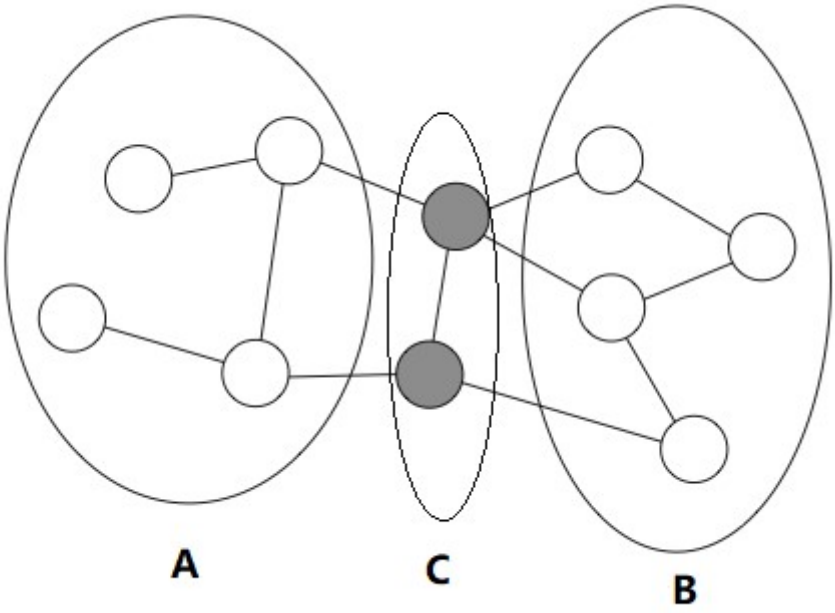


它的表达就应该是：

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

分离

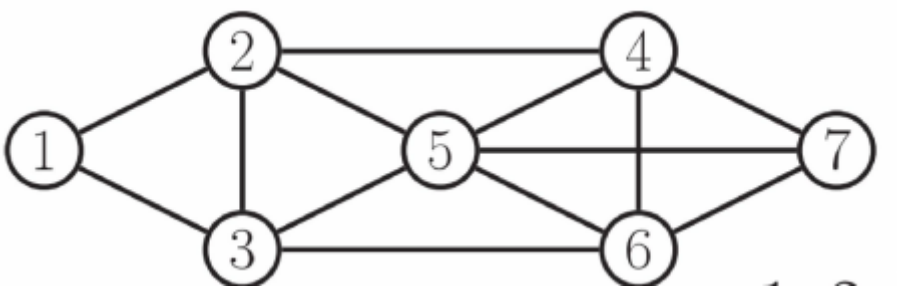
设 A、B、C 都是马尔可夫随机场中的节点集合，若从 A 中的节点到 B 中的节点都必须经过 C 中的节点，则称 A 和 B 被 C 分离，C 称为分离集（Separating Set）。参考下图：



马尔可夫性

马尔可夫性的原始定义为：当一个随机过程在给定当前状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态；换句话说，在给定当前状态时，它与过去状态（即该过程的历史路径）是条件独立的，那么此随机过程即具有**马尔可夫性。**

我们把马尔可夫性引入到马尔可夫随机场中，把当前状态看作无向图中的一个节点，过去状态看作与当前状态节点有历史路径（边）连接的其他节点。



- **全局马尔可夫性**：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。
  比如上图中：$1,2 \perp 6,7|3,4,5$
- **局部马尔可夫性**：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Y_w的条件下，随机变量Y_v和Y_o条件独立。
  比如上图中：$1 \perp \text{rest}|2,3$
- **成对马尔可夫性**：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。
  比如上图中：$1 \perp 7|\text{rest}$

实际上通过全局可以得到局部，通过局部可以得到成对，通过成对可以得到全局，所以这三者实际上是等价的。

实际上这个三个性质对MRF有决定性作用，满足三者其一的无向图称之为MRF。

特征函数

通常是一些实值函数，用来刻画数据的一些很可能成立或期望成立的经验特性。

比如下面就是一个特征函数：

$$f_i(x,y) = \begin{cases} 1 & \text{if } y = \text{ name and } x = \text{ Mister} \\ 0 & \text{otherwise} \end{cases}$$

比如在字标注的中文分词中，有：

```
If(y==' B' &&x==' 人' ) return 1 else return 0;

If(y==' E' &&x==' 人' ) return 1 else return 0;

If(y==' M' &&x==' 人' ) return 1 else return 0;

If(y==' S' &&x==' 人' ) return 1 else return 0;
```

1. CRF

CRF是一种无向图模型，它和马尔可夫随机场的不同之处在于，MRF是生成模式，而CRF是判别式模式，对条件分布进行建模。

两者又是相关的，CRF是有条件的马尔可夫随机场。CRF是给定随机变量X的条件下，随机变量Y的马尔可夫随机场。

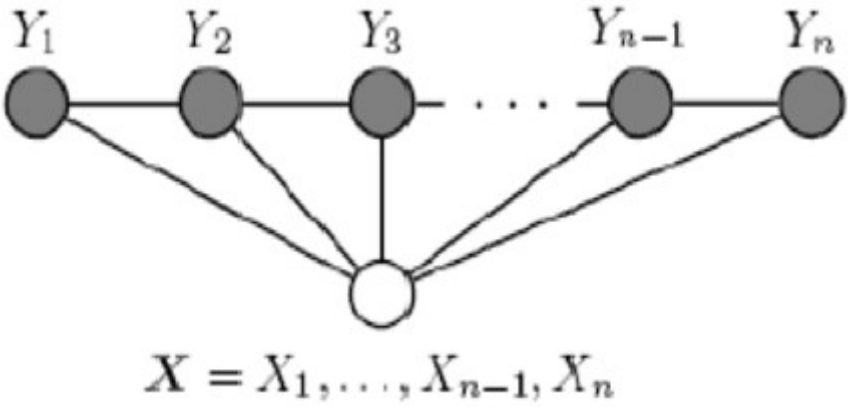
定义：设X和Y是随机变量，$P(Y|X)$是给定X条件下Y的条件概率分布。如果随机变量Y构成了一个由无向图$G=<V,E>$表示的**马尔可夫随机场**，则称条件概率分布$P(Y|X)$为**CRF**。

对于这个定义我们换一种简单的表达形式，设 X 和 Y是随机变量，$P(Y|X)$ 是给定 X 条件下Y 的条件概率分布。如果随机变量 Y 构成一个无向图 $G=<V，E>$，且图 G 中每一个变量Y_v都满足马尔可夫性——$P(Y_v|X,Y_Z)=P(Y_v|X,Y_W)$其中 Z 表示无向图中节点$v$ 以外所有点的集合，W 表示无向图中与节点 v 有边连接的所有节点集合（**也就是说，不相连表示独立**）——则 $P(Y|X)$为CRF。
其中X 是输入变量，表示需要标注的观测序列，比如在中文分词中分的每个句子的每个字；Y是输出变量，表示状态（或称标记）序列，比如在中文分词中标注的**BEMS**这些。

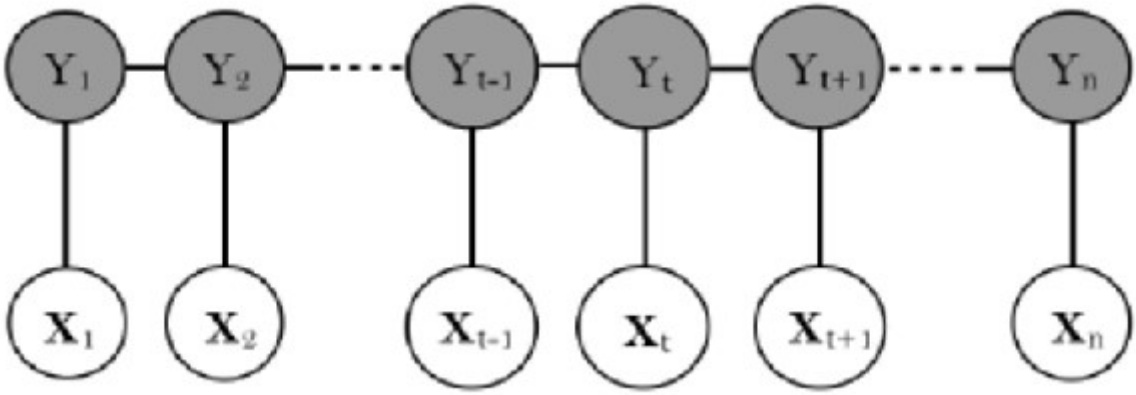
注：从定义的层面，一般不要求X和Y具有相同的结构，不过在实际运行中一般假设X和Y具有相同图结构。

1.1 CRF线性链

在现实应用中，最常被用到的 CRF 是线性链（Linear Chain）CRF，其结构如下：



当 X 和 Y 具备相同结构时，其形如下：



上图中，X 为观测序列，Y 为状态序列。
如果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CRF，那么随机变量Y也满足马尔可夫性。即：

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

(说白了状态只和相连的两个有关，而与其他独立，表达出了一种线性连接的关系)
则称$P(Y|X)$为条件随机场，X为输入/观测序列，Y为输出/标记/状态序列。

则我们选用指数势函数，并引入特征函数，则条件概率被定义为：

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_j \sum_{i=1}^{n-1} \lambda_j t_j \left(y_{i+1}, y_i, \mathbf{x}, i\right) + \sum_k \sum_{i=1}^{n} \mu_k s_k \left(y_i, \mathbf{x}, i\right) \right)$$

其中

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

参数说明：

- 特征函数：$t_k \left(y_{i-1}, y_i, x, i\right)$　$s_l \left(y_i, x, i\right)$
  - 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, 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
- 特征函数对应的权值λ_k, μ_l
- Z(x)为规范化因子，保证$\mathbf{P(Y|X)}$为概率分布。

1.2 举个例子

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的取值范围为$\{1, 2\}$。
对于第一条连接边，假设特征和权值如下:
$t_1 = t_1 \left(y_{i-1} = 1, y_i = 2, x, i\right), \quad i = 2, 3, \quad \lambda_1 = 1$
对应的特征函数是：

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

下同：

t1=t1(y1=1,y2=2,x,2) $\lambda_1=1$

t1=t1(y2=1,y3=1,x,3) $\lambda_1=1$

t2=t2(y1=1,y2=1,x,2) $\lambda_2=0.5$

t3=t3(y2=2,y3=1,x,3) $\lambda_3=1$

t4=t4(y1=2,y2=1,x,2) $\lambda_4=1$

t5=t5(y2=2,y3=2,x,3) $\lambda_5=0.2$

s1=s1(y1=1,x,1) $\mu_1=1$

s2=s2(y2=2,x,i) $\mu 2=0.5$

s3=s3(y1=1,x,i) $\mu_3=0.8$

s4=s4(y3=2,x,i) $\mu_4=0.5$

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

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x,i\right)+\sum_{i,l}\mu_l s_l\left(y_i,x,i\right)\right)$$
$$\propto\exp\left(\sum_{k=1}^{5}\lambda_k\sum_{i=2}^{3}t_k\left(y_{i-1},y_i,x,i\right)+\sum_{l=1}^{4}\mu_l\sum_{i=1}^{3}s_l\left(y_i,x,i\right)\right)$$
$$=\exp(3.2)$$

1.3 简化版形式

之前的形式表达太过于复杂，我们用一种简单一点的形式来表达。为方便起见，将转移特征和状态特征及其权值同统一的符号表示

设有K1个转移特征，K2个状态特征，K=K1+K2，则

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

这样一来，在所有位置对转移/状态特征在位置i进行求和，就变成了

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

用w_k表示特征$f_k(y,x)$的权值，即

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

于是条件随机场可以表示为：

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

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

若以w表示权重向量，

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

以F(x,y)表示全局特征向量，

$$F(y,x)=(f_1(y,x),f_2(y,x),\ldots,f_K(y,x))^T$$

则可以写成向量内积的形式

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

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

1.4 三个问题

相比较HMM，CRF同样也有三个问题。

概率计算问题。

问题名称：概率计算问题。
已知信息：给定 CRF

- P(Y|X)
- 观测序列 x
- 状态序列 y

求解目标：求条件概率 $P(Y_i=y_i|x),P(Y_{i-1}=y_{i-1},Y_i=y_i|x)$ 以及相应的数学期望。
求解方法：前向后向算法。

预测问题。

问题名称：预测问题。
已知信息：给定 CRF

- P(Y|X)
- 观测序列 x

求解目标：求条件概率最大的状态序列 y^*，也就是对观测序列进行的标注。
求解方法：Viterbi算法

学习问题

问题名称：学习问题。
已知信息：训练数据集。
求解目标：求 CRF 模型的参数。
求解方法：IIS：改进的迭代尺度算法

建议大家参考简书：如何轻松愉快地理解条件随机场（CRF）？，查看CRF如何解决一个实际的问题。

参考

简书：如何轻松愉快地理解条件随机场（CRF）？
掘金：条件随机场实现命名实体识别
机器学习——周志华
统计学习方法
classical probabilistic models and conditional random fields