概率图模型

一、概率图模型

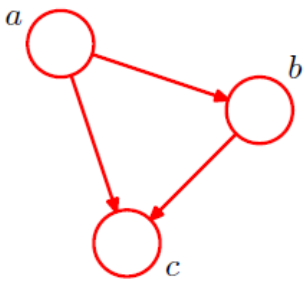
1. 考虑三个随机变量 a,b,c ，其联合概率分布为：

$$P(a,b,c)=P(c\mid a,b)P(a,b)=P(c\mid a,b)P(b\mid a)P(a)$$

   - 对每个随机变量引入一个节点，然后为每个节点关联上式右侧对应的条件概率。
   - 对于每个条件概率分布，在图中添加一个链接（箭头）：箭头的起点是条件概率的条件代表的结点。

     对于因子 $P(a)$ ，因为它不是条件概率，因此没有输入的链接。
   - 如果存在一个从结点 a 到结点 b 的链接，则称结点 a 是结点 b 的父节点，结点 b 是结点 a 的子节点。
   - 可以看到，上式的左侧关于随机变量 a,b,c 是对称的，但是右侧不是。

     实际上通过对 $P(a,b,c)$ 的分解，隐式的选择了一个特定的顺序（即 a,b,c ）。如果选择一个不同的顺序，则得到一个不同的分解方式，因此也就得到一个不同的图的表现形式。



2. 对于 K 个随机变量的联合概率分布，有：

$$P(X_1,X_2,\cdots,X_K)=P(X_K\mid X_1,X_2,\cdots,X_{K-1})\cdots P(X_2\mid X_1)P(X_1)$$

   - 它对应于一个具有 K 个结点的有向图。
     - 每个结点对应于公式右侧的一个条件概率分布。
     - 每个结点的输入链接包含了所有的编号低于它的结点。
   - 这个有向图是全链接的，因为每对结点之间都存在一个链接。

     实际应用中，真正有意义的信息是图中的链接的缺失，因为：
     - 全链接的计算量太大。
     - 链接的缺失代表了某些随机变量之间的不相关或者条件不相关。
   - 设节点 X_i 的父节点集合为 Ψ_{X_i} ，则所有随机变量的联合概率分布为：

$$P(X_1,X_2,\cdots,X_K)=\prod_{k=1}^{K}P(X_k\mid\Psi_{X_k})$$

3. 前面讨论的是：每个结点对应于一个变量。可以很容易的推广到每个结点代表一个变量的集合（或者关联到一个向量）的情形。

   可以证明：如果上式右侧的每一个条件概率分布都是归一化的，则这个表示方法整体总是归一化的。

4. 概率图模型`probabilistic graphical model`就是一类用图来表达随机变量相关关系的概率模型：
   - 用一个结点表示一个或者一组随机变量。
   - 结点之间的边表示变量间的概率相关关系。

   概率图描述了：联合概率分布在所有随机变量上能够分解为一组因子的乘积的形式，而每个因子只依赖于随机变量的一个子集。
5. 根据边的性质不同，概率图模型可以大致分为两类：
   - 使用有向无环图表示随机变量间的依赖关系，称作有向图模型或者贝叶斯网络`Bayesian network`。

     有向图对于表达随机变量之间的因果关系很有用。
   - 使用无向图表示随机变量间的相关关系，称作无向图模型或者马尔可夫网络`Markov network`。

     无向图对于表达随机变量之间的软限制比较有用。
6. 概率图模型的优点：
   - 提供了一个简单的方式将概率模型的结构可视化。
   - 通过观察图形，可以更深刻的认识模型的性质，包括条件独立性。
   - 高级模型的推断和学习过程中的复杂计算可以利用图计算来表达，图隐式的承载了背后的数学表达式。

二、贝叶斯网络

1. 贝叶斯网络`Bayesian network`借助于有向无环图来刻画特征之间的依赖关系，并使用条件概率表`Conditional Probability Table:CPT`来描述特征的联合概率分布。

   这里每个特征代表一个随机变量，特征的具体取值就是随机变量的采样值。

2.1 条件独立性

1. 一个贝叶斯网 $\mathcal{B}$ 由结构 $\mathcal{G}$ 和参数 Θ 两部分组成，即 $\mathcal{B}=(\mathcal{G},\Theta)$：
   - 网络结构 $\mathcal{G}$ 是一个有向无环图，其中每个结点对应于一个特征。

     若两个特征之间有直接依赖关系，则他们用一条边相连。
   - 参数 Θ 定量描述特征间的这种依赖关系。设特征 X_i 在 $\mathcal{G}$ 中父节点的集合为 Ψ_{X_i}， 则 Θ 包含了该特征的条件概率表：

$$\theta_{X_i|\Psi_{X_i}}=P(X_i\mid\Psi_{X_i})$$

2. 贝叶斯网结构有效地表达了特征间的条件独立性。

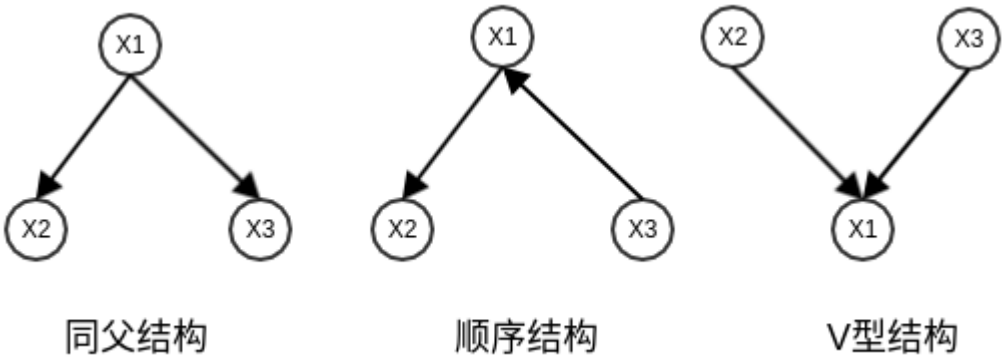
   给定父节点集，贝叶斯网络假设每个特征与它的非后裔结点表达的特征是相互独立的。于是有：

$$P(\mathbb{X})=P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i\mid\Psi_{X_i})=\prod_{i=1}^{n}\theta_{X_i|\Psi_{X_i}}$$

   推导过程：

$$\begin{aligned}P(\mathbb{X})=P(X_1,X_2,\cdots,X_n)&=P(X_1\mid X_2,\cdots,X_n)P(X_2,\cdots,X_n)\\&=P(X_1\mid\Psi_{X_1})P(X_2,\cdots,X_n)\\&=P(X_1\mid\Psi_{X_1})P(X_2\mid X_3,\cdots,X_n)P(X_3,\cdots,X_n)\\&=P(X_1\mid\Psi_{X_1})P(X_2\mid\Psi_{X_2})P(X_3,\cdots,X_n)=\cdots\end{aligned}$$

3. 贝叶斯网络中三个结点之间典型依赖关系如下图：



- 同父结构：给定父节点 X_1 的取值， 则 X_2 与 X_3 条件独立，即：

$$P(X_2, X_3 \mid X_1) = P(X_2 \mid X_1)P(X_3 \mid X_1)$$

- 顺序结构：给定中间节点 X_1 的取值， 则 X_2 与 X_3 条件独立，即：

$$P(X_2, X_3 \mid X_1) = P(X_2 \mid X_1)P(X_3 \mid X_1)$$

  即：在 X_1 给定的条件下， X_2, X_3 之间被阻断。因此它们关于 X_1 条件独立。

- `V` 型结构：给定子节点 X_1 的取值，则 X_2 与 X_3 必定不是条件独立的。即：$P(X_2, X_3 \mid X_1) \neq P(X_2 \mid X_1)P(X_3 \mid X_1)$ 。

  事实上 X_2 与 X_3 是独立的（但不是条件独立的），即 $P(X_2, X_3) = P(X_2)P(X_3)$ 。

4. 为了分析有向图中节点之间的条件独立性，可以使用有向分离技术：
   - 找出有向图中的所有 `V` 型结构，在 `V` 型结构的两个父节点之间加上一条无向边。
   - 将所有的有向边改成无向边。

   这样产生的无向图称作道德图 `moral graph` 。父节点相连的过程称作道德化 `moralization` 。基于道德图能直观、迅速的找到结点之间的条件独立性。

2.2 网络的学习

1. 贝叶斯网络的学习可以分为参数学习和结构学习两部分
   - 参数学习比较简单。只需要通过对训练样本“计数”，估计出每个结点的条件概率表即可。

     但是前提是必须知道网络结构。
   - 结构学习比较复杂，结构学习被证明是 `NP` 难问题。
2. 贝叶斯网络的结构学习通常采用 `评分搜索` 来求解。
   - 先定义一个评分函数，以此评估贝叶斯网络与训练数据的契合程度。然后基于这个评分函数寻找结构最优的贝叶斯网。
   - 最常用的评分函数基于信息论准则：将结构学习问题视作一个数据压缩任务。
     - 学习的目标是找到一个能以最短编码长度描述训练集数据集的模型。这就是 `最小描述长度 Minimal Description Length:MDL` 准则。
     - 此时的编码长度包括了：描述模型自身所需要的字节长度，和使用该模型描述数据所需要的字节长度。
3. 给定训练集 $\mathbb{D} = \{(\vec{\mathbf{x}}_1, y_1), (\vec{\mathbf{x}}_2, y_2), \cdots, (\vec{\mathbf{x}}_N, y_N)\}$，贝叶斯网络 $\mathcal{B} = (\mathcal{G}, \Theta)$ 在 $\mathbb{D}$ 上的评分函数定义为：

$$score(\mathcal{B} \mid \mathbb{D}) = f(\theta)|\mathcal{B}| - L(\mathcal{B} \mid \mathbb{D})$$

其中： $f(\theta)$ 表示描述每个参数 θ 所需的字节数； $|\mathcal{B}|$ 是贝叶斯网络的参数个数；

$$L(\mathcal{B}\mid\mathbb{D})=\sum_{i=1}^{N}\log P(\vec{\mathbf{x}}_i)$$

是贝叶斯网 $\mathcal{B}$ 的对数似然。因此：

- 第一项 $f(\theta)|\mathcal{B}|$ 是计算编码贝叶斯网络 $\mathcal{B}$ 所需要的字节数。
- 第二项 $-\sum_{i=1}^{N}\log P(\vec{\mathbf{x}}_i)$ 是计算 $\mathcal{B}$ 所对应的概率分布 P 需要多少字节来描述 $\mathbb{D}$ 。

4. 现在结构学习任务转换为一个优化任务，即寻找一个贝叶斯网络 $\mathcal{B}$ 使评分函数 $score(\mathcal{B}\mid\mathbb{D})$ 最小。

   问题是，从所有可能的网络结构空间中搜索最优贝叶斯网络结构是个 `NP` 难问题，难以快速求解。

   有两种方法可以在有限时间内求得近似解：

   - 贪心算法。如从某个网络结构出发，每次调整一条边，直到评分函数不再降低为止。
   - 增加约束。通过给网络结构增加约束来缩小搜索空间，如将网络结构限定为树形结构等。

5. 贝叶斯网络训练好之后就能够用来进行未知样本的预测。

   最理想的是直接根据贝叶斯网络定义的联合概率分布来精确计算后验概率，但问题是这样的“精确推断”已经被证明是 `NP` 难的。

   此时需要借助“近似推断”，通过降低精度要求从而在有限时间内求得近似解，常用的近似推断为吉布斯采样(`Gibbs sampling`)。

三、马尔可夫随机场

1. 根据前面的介绍，有向图模型可以将一组变量上的联合概率分布分解为局部条件概率分布的乘积。无向图模型也可以表示一个分解形式。

   马尔可夫随机场 `Markov Random Field:MRF` 是一种著名的无向图模型。

2. 现实任务中，可能只知道两个变量之间存在相关关系，但是并不知道具体怎样相关，也就无法得到变量之间的依赖关系。

   - 贝叶斯网络需要知道变量之间的依赖关系，从而对依赖关系（即条件概率）建模。
   - 马尔科夫随机场并不需要知道变量之间的依赖关系。它通过变量之间的联合概率分布来直接描述变量之间的关系。

     如： X_1,X_2 两个变量的联合概率分布为：

X_1,X_2	$P(X_1,X_2)$
$X_1=0,X_2=0$	1000
$X_1=0,X_2=1$	10
$X_1=1,X_2=0$	20
$X_1=1,X_2=1$	2000

     则这个分布表示： X_1 和 X_2 取值相同的概率很大。

   - 事实上这里的 $P(\cdot)$ 就是后面介绍的势函数。
     - 它们的总和不一定为 1 。即：这个表格并未定义一个概率分布，它只是告诉我们某些配置具有更高的可能性。

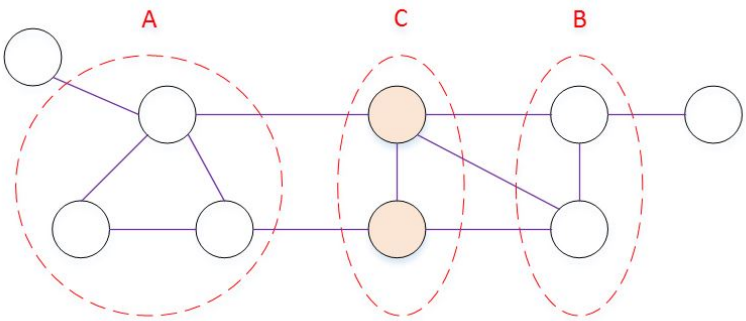
    - 它们并没有条件关系，它涉及到变量的联合分布的比例。
3. 马尔可夫随机场可以应用于图像问题中。
    - 每个像素都表示成一个节点，相邻像素之间相互影响。
    - 像素之间并不存在因果关系，它们之间的作用是对称的。因此使用无向图概率模型，而不是有向图概率模型。

3.1 马尔科夫性

1. 对结点 A, B, C ，若去掉结点 C 之后 A, B 分属于两个联通分支，则称结点 A, B 关于结点 C 条件独立，记作 $A \perp B \mid C$ 。

   这一概念可以推广到集合。

2. 分离集`separating set`：如下图所示，若从结点集`A`中的结点到结点集`B`中的结点都必须经过结点集`C`中的结点，则称结点集`A`和结点集`B`被结点集`C`分离，`C`称作分离集。



3. 马尔可夫随机场有三个马尔可夫性定义：
    - 全局马尔科夫性。
    - 局部马尔科夫性。
    - 成对马尔科夫性。
4. 全局马尔可夫性`global Markov property`：给定两个变量子集和它们的分离集，则这两个变量子集关于分离集条件独立。

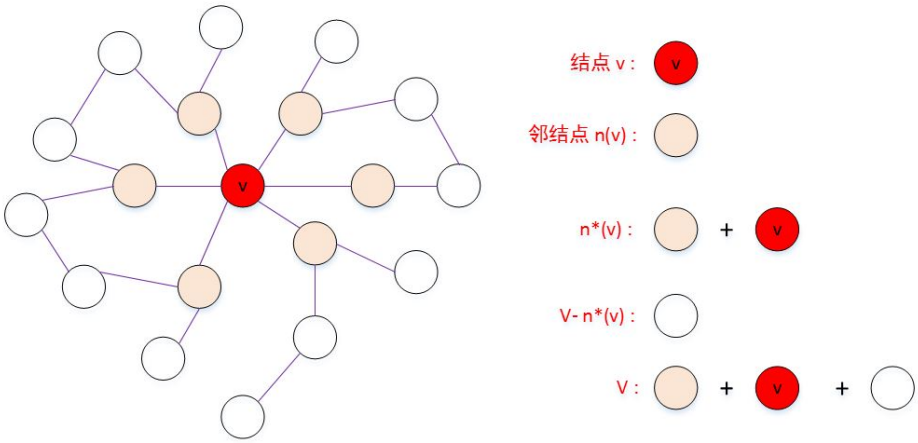
   如上图，令结点集`A`、`B`、`C`对应的变量集分别为 $\mathbb{X}_A, \mathbb{X}_B, \mathbb{X}_C$， 则 $\mathbb{X}_A$ 和 $\mathbb{X}_B$ 在给定 $\mathbb{X}_C$ 的条件下独立，记作：$\mathbb{X}_A \perp \mathbb{X}_B \mid \mathbb{X}_C$ 。

5. 局部马尔可夫性`local Markov property`：给定某变量的邻接变量，则该变量与其他变量（既不是该变量本身，也不是邻接变量）关于邻接变量条件独立。

   即：令 $\mathbb{V}$ 为图的结点集， $n(v)$ 为结点 v 在图上的邻接结点， $n^*(v) = n(v) \bigcup \{v\}$ ，则有：

$$\mathbb{X}_v \perp \mathbb{X}_{\mathbb{V}-n^*(v)} \mid \mathbb{X}_{n(v)}$$

> 这是根据全局马尔可夫性推导而来

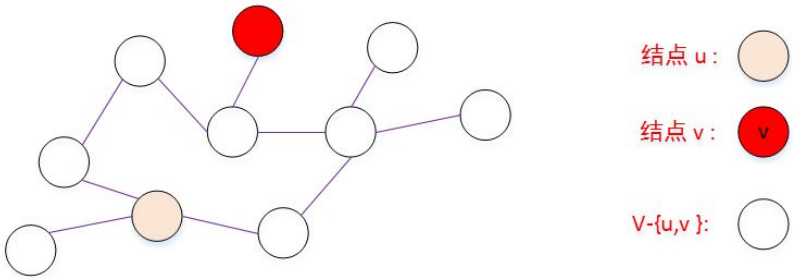


6. 成对马尔可夫性`pairwise Markov property`：给定两个非邻接变量，则这两个变量关于其他变量（即不是这两个变量的任何其他变量）条件独立。

   即：令 $\mathbb{V}$ 为图的结点集，令 $\mathbb{E}$ 为图的边集。对图中的两个结点 u,v，若 $(u,v)\notin\mathbb{E}$，则有：

$$\mathbb{X}_u \perp \mathbb{X}_v \mid \mathbb{X}_{\mathbb{V}-\{u,v\}}$$

> 这也是根据全局马尔可夫性推导而来



3.2 极大团

1. 考虑两个结点 X_i,X_j，如果它们之间不存在链接，则给定图中其他所有结点，那么这两个结点一定是条件独立的
   - 因为这两个结点之间没有直接的路径，并且所有其他路径都通过了观测的结点。
   - 该条件独立性表示为：

$$P(X_i,X_j\mid\mathbb{X}-\{X_i,X_j\})=P(X_i\mid\mathbb{X}-\{X_i,X_j\})P(X_j\mid\mathbb{X}-\{X_i,X_j\})$$

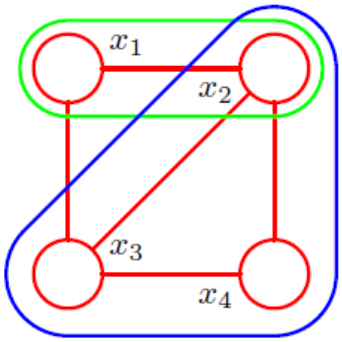
   - 对于联合概率分布的分解，则一定要让 X_i,X_j 不能出现在同一个因子中，从而让属于这个图的所有可能的概率分布都满足条件独立性质。
2. 这里引入团的概念：
   - 对于图中结点的一个子集，如果其中任意两个结点之间都有边连接，则称该结点子集为一个团`clique`。即：团中的结点集合是全连接的。
   - 若在一个团中加入团外的任何一个结点都不再形成团，则称该团为极大团`maximal clique`。

     即：极大团就是不能被其他团所包含的团。

- 显然，每个结点至少出现在一个极大团中。

  如下图所示

  - 所有的团有：
    $\{X_1,X_2\},\{X_2,X_3\},\{X_3,X_4\},\{X_4,X_2\},\{X_1,X_3\},\{X_1,X_2,X_3\},\{X_2,X_3,X_4\}$
  - 极大团有：$\{X_1,X_2,X_3\},\{X_2,X_3,X_4\}$



3. 可以将联合概率分布分解的因子定义为团中变量的函数，也称作势函数。

   它是定义在随机变量子集上的非负实函数，主要用于定义概率分布函数。

4. 在马尔可夫随机场中，多个变量之间的联合概率分布能够基于团分解为多个因子的乘积，每个因子仅和一个团相关。

   对于 n 个随机变量 $\mathbb{X}=\{X_1,X_2,\cdots,X_n\}$，所有团构成的集合为 $\mathcal{C}$， 与团 $\mathbb{Q}\in\mathcal{C}$ 对应的变量集合记作 $\mathbb{X}_{\mathbb{Q}}$，则联合概率 $P(\mathbb{X})$ 定义为：

   $$P(\mathbb{X})=\frac{1}{Z}\prod_{\mathbb{Q}\in\mathcal{C}}\psi_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})$$

   其中：

   - 所有团构成了整个概率图（团包含了结点和连接）， 任意两个团之间不互相包含（但是可以相交）。
   - $\psi_{\mathbb{Q}}$ 为与团 $\mathbb{Q}$ 对应的势函数，用于对 团 $\mathbb{Q}$ 中的变量关系进行建模。
   - $Z=\sum_{\mathbb{X}}\prod_{\mathbb{Q}\in\mathcal{C}}\psi_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})$ 为规范化因子，确保 $P(\mathbb{X})$ 满足概率的定义。
   - 实际应用中，Z 的精确计算非常困难。但是很多任务往往并不需要获得 Z 的精确值。

5. 在上述 $P(\mathbb{X})$ 计算公式中，团的数量会非常多。如：所有相互连接的两个结点都会构成一个团。这意味着有非常多的乘积项。

   注意到：若团 $\mathbb{Q}$ 不是极大团，则它必被一个极大团 $\mathbb{Q}^*$ 所包含。此时有： $\mathbb{X}_{\mathbb{Q}}\subseteq\mathbb{X}_{\mathbb{Q}^*}$ 。

   于是： 随机变量集合 $\mathbb{X}_{\mathbb{Q}}$内部随机变量之间的关系不仅体现在势函数 $\psi_{\mathbb{Q}}$ 中，也体现在势函数 $\psi_{\mathbb{Q}^*}$ 中（这是根据势函数的定义得到的结论）。

   于是： 联合概率 $P(\mathbb{X})$ 可以基于极大团来定义。假定所有极大团构成的集合为 $\mathcal{C}^*$， 则有`Hammersley-Clifford` 定理：

   $$P(\mathbb{X})=\frac{1}{Z^*}\prod_{\mathbb{Q}\in\mathcal{C}^*}\psi_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})$$

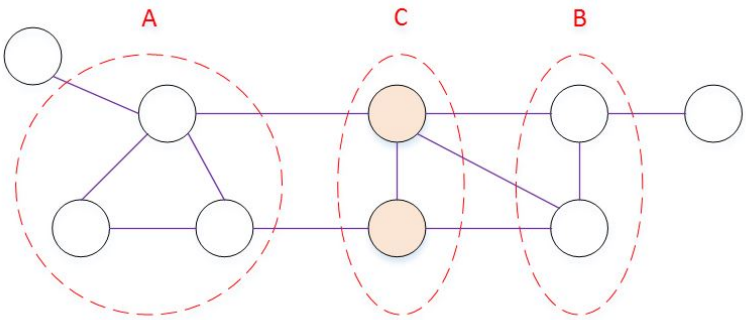
   其中：$Z^*=\sum_{\mathbb{X}}\prod_{\mathbb{Q}\in\mathcal{C}^*}\psi_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})$ 为规范化因子，确保 $P(\mathbb{X})$ 满足概率的定义 。

6. 通常贝叶斯网络可以将因子定义成表格形态，而马尔可夫随机场将因子定义为势函数。因为马尔可夫随机场无法将因子表格化。

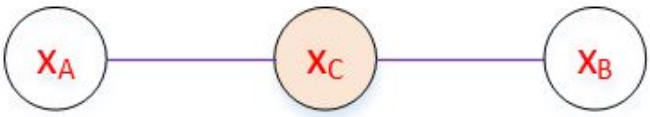
   假设有 n 个随机变量 $X_1,X_2,\cdots,X_n$ ， 它们的取值都是 $\{0,1\}$ 。假设马尔可夫随机场中它们是全连接的，则其联合概率分布需要 $O(2^n)$ 个参数。

如果表达成表格形态，横轴表示连接的一个端点、纵轴表示连接的另一个端点，则需要 $O(n^2)$ 个参数。当 n 较大的时候， $O(n^2) < O(2^n)$ ，因此表格无法完全描述马尔可夫随机场的参数。

7. 全局马尔可夫性的一个证明：



将上图简化为如下所示：



- 最大团有两个：$\{\mathbb{X}_A, \mathbb{X}_C\}, \{\mathbb{X}_B, \mathbb{X}_C\}$ ，因此联合概率为：

$$P(\mathbb{X}_A, \mathbb{X}_B, \mathbb{X}_C) = \frac{1}{Z}\psi_{AC}(\mathbb{X}_A, \mathbb{X}_C)\psi_{BC}(\mathbb{X}_B, \mathbb{X}_C)$$

- 基于条件概率的定义有：

$$P(\mathbb{X}_A, \mathbb{X}_B \mid \mathbb{X}_C) = \frac{P(\mathbb{X}_A, \mathbb{X}_B, \mathbb{X}_C)}{P(\mathbb{X}_C)}$$

根据：

$$P(\mathbb{X}_C) = \sum_{\mathbb{X}'_A}\sum_{\mathbb{X}'_B} P(\mathbb{X}'_A, \mathbb{X}'_B, \mathbb{X}_C) = \sum_{\mathbb{X}'_A}\sum_{\mathbb{X}'_B} \frac{1}{Z}\psi_{AC}(\mathbb{X}'_A, \mathbb{X}_C)\psi_{BC}(\mathbb{X}'_B, \mathbb{X}_C)$$

将 $P(\mathbb{X}_C)$ 和 $P(\mathbb{X}_A, \mathbb{X}_B, \mathbb{X}_C)$ 代入，有：

$$\begin{aligned} P(\mathbb{X}_A, \mathbb{X}_B \mid \mathbb{X}_C) &= \frac{\psi_{AC}(\mathbb{X}_A, \mathbb{X}_C)\psi_{BC}(\mathbb{X}_B, \mathbb{X}_C)}{\sum_{\mathbb{X}'_A}\sum_{\mathbb{X}'_B}\psi_{AC}(\mathbb{X}'_A, \mathbb{X}_C)\psi_{BC}(\mathbb{X}'_B, \mathbb{X}_C)} \\ &= \frac{\psi_{AC}(\mathbb{X}_A, \mathbb{X}_C)}{\sum_{\mathbb{X}'_A}\psi_{AC}(\mathbb{X}'_A, \mathbb{X}_C)} \cdot \frac{\psi_{BC}(\mathbb{X}_B, \mathbb{X}_C)}{\sum_{\mathbb{X}'_B}\psi_{BC}(\mathbb{X}'_B, \mathbb{X}_C)} \end{aligned}$$

- 考虑 $P(\mathbb{X}_A \mid \mathbb{X}_C)$：

$$P(\mathbb{X}_A \mid \mathbb{X}_C)=\frac{P(\mathbb{X}_A,\mathbb{X}_C)}{P(\mathbb{X}_C)}=\frac{\sum_{\mathbb{X}'_B}P(\mathbb{X}_A,\mathbb{X}'_B,\mathbb{X}_C)}{\sum_{\mathbb{X}'_A}\sum_{\mathbb{X}'_B}P(\mathbb{X}'_A,\mathbb{X}'_B,\mathbb{X}_C)}$$
$$=\frac{\sum_{\mathbb{X}'_B}\frac{1}{Z}\psi_{AC}(\mathbb{X}_A,\mathbb{X}_C)\psi_{BC}(\mathbb{X}'_B,\mathbb{X}_C)}{\sum_{\mathbb{X}'_A}\sum_{\mathbb{X}'_B}\frac{1}{Z}\psi_{AC}(\mathbb{X}'_A,\mathbb{X}_C)\psi_{BC}(\mathbb{X}'_B,\mathbb{X}_C)}$$
$$=\frac{\psi_{AC}(\mathbb{X}_A,\mathbb{X}_C)\sum_{\mathbb{X}'_B}\psi_{BC}(\mathbb{X}'_B,\mathbb{X}_C)}{\left(\sum_{\mathbb{X}'_A}\psi_{AC}(\mathbb{X}'_A,\mathbb{X}_C)\right)\left(\sum_{\mathbb{X}'_B}\psi_{BC}(\mathbb{X}'_B,\mathbb{X}_C)\right)}$$
$$=\frac{\psi_{AC}(\mathbb{X}_A,\mathbb{X}_C)}{\sum_{\mathbb{X}'_A}\psi_{AC}(\mathbb{X}'_A,\mathbb{X}_C)}$$

- 同理，可以推导出：

$$P(\mathbb{X}_B \mid \mathbb{X}_C)=\frac{\psi_{BC}(\mathbb{X}_B,\mathbb{X}_C)}{\sum_{\mathbb{X}'_B}\psi_{BC}(\mathbb{X}'_B,\mathbb{X}_C)}$$

- 于是有：

$$P(\mathbb{X}_A,\mathbb{X}_B \mid \mathbb{X}_C)=P(\mathbb{X}_A \mid \mathbb{X}_C)\cdot P(\mathbb{X}_B \mid \mathbb{X}_C)$$

8. 有向图和无向图模型都将复杂的联合分布分解为多个因子的乘积：
   - 无向图模型的因子是势函数，需要全局归一化。

     优点是：势函数设计不受概率分布的约束，设计灵活。
   - 有向图模型的因子是概率分布，不需要全局归一化。

     优点是：训练相对高效。

3.3 势函数

1. 势函数 $\psi_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})$ 的作用是刻画变量集 $\mathbb{X}_{\mathbb{Q}}$ 中变量之间的相关关系。
2. 与有向图的联合分布的因子不同，无向图中的势函数没有一个具体的概率意义。
   - 这可以使得势函数的选择具有更大的灵活性，但是也产生一个问题：对于具体任务来说，如何选择势函数。
   - 可以这样理解：将势函数看做一种度量：它表示局部变量的哪种配置优于其他配置。
3. 势函数必须是非负函数（确保概率非负），且在所偏好的变量取值上具有较大的函数值。 如：

$$\psi_{AC}(X_A,X_C)=\begin{cases}2.0, & \text{if} \quad X_A=X_C\\ 0.1, & \text{otherwise}\end{cases}$$
$$\psi_{BC}(X_B,X_C)=\begin{cases}0.1, & \text{if} \quad X_B=X_C\\ 1.5, & \text{otherwise}\end{cases}$$

   - 该模型偏好变量 X_A,X_C 拥有相同的取值；偏好 X_B,X_C 拥有不同的取值。
   - 如果想获取较高的联合概率，则可以令 X_A 和 X_C 相同，且 X_B 和 X_C 不同。
4. 通常使用指数函数来定义势函数：

$$\psi_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})=e^{-H_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})}$$

其中 $H_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})$ 是一个定义在变量集 $\mathbb{X}_{\mathbb{Q}}$ 上的实值函数，称作能量函数。

- 指数分布被称作玻尔兹曼分布。

- 联合概率分布被定义为势函数的乘积，因此总能量可以通过将每个最大团中的能量相加得到。

  这就是采取指数函数的原因，指数将势函数的乘积转换为能量函数的相加。

5. $H_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}})$ 常见形式为：

$$H_{\mathbb{Q}}(\mathbb{X}_{\mathbb{Q}}) = \sum_{u,v\in\mathbb{Q},u\neq v}\alpha_{u,v}t_{u,v}(u,v) + \sum_{v\in\mathbb{Q}}\beta_v s_v(v)$$

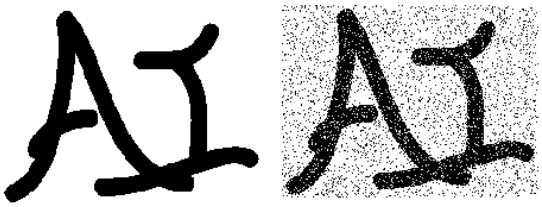
其中：

- $\alpha_{u,v},\beta_v$ 表示系数； $t_{u,v}(u,v),s_v(v)$ 表示约束条件。
- 上式第一项考虑每一对结点之间的关系；第二项考虑单个结点。

3.4 图像降噪应用

1. 马尔可夫随机场的一个应用是图像降噪。

   如下图所示，左侧图片为原始图像，右侧图片为添加了一定噪音（假设噪音比例不超过 10%） 的噪音图像。现在给定噪音图像，需要得到原始图像。

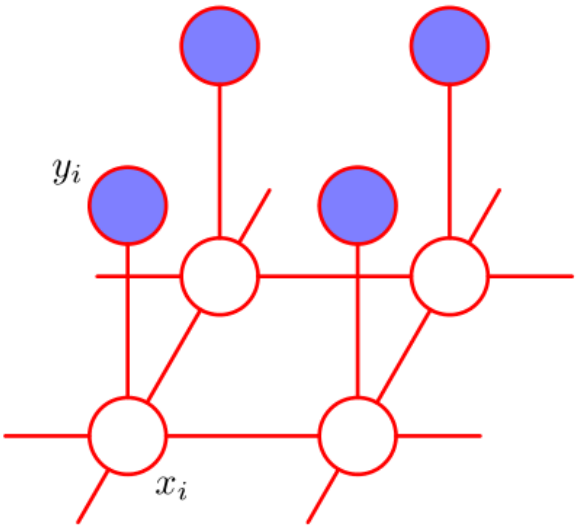


2. 设随机变量 Y_i 表示噪音图像中的像素，随机变量 X_i 表示原始图像中的像素。其中：
   - i 代表图片上的每个位置。
   - $Y_i,X_i\in\{+1,-1\}$ 。当它们取 `+1` 时，表示黑色；取 `-1` 时，表示白色。
3. 由于已知噪音图像，因此 Y_i 的分布是已知的。原始图像未知，则 X_i 的分布待求解。
   - 由于噪音图像是从原始图像添加噪音而来，因此我们认为：Y_i 和 X_i 具有较强的关联。
   - 由于原始图像中，每个像素和它周围的像素值比较接近，因此 X_i 与它相邻的像素也存在较强的关联。

     因此我们假设：X_i 只和它直接相邻的像素有联系（即：条件独立性质）。

   因此得到一个具备局部马尔可夫性质的概率图模型。模型中具有两类团：
   - 团 $\{X_i,Y_i\}$ ：原始图像的像素和噪音图像的像素。
   - 团 $\{X_i,X_j\}$：原始图像的像素和其直接相邻的像素。

   这两类团就是模型中的最大团。



4. 定义能量函数：
   - 对于团 $\{X_i, Y_i\}$，定义能量函数：$H_1(X_i, Y_i) = -\eta X_i Y_i$。即：$X_i, Y_i$ 相同时，能量较低；X_i, Y_i 不同时，能量较高。
   - 对于团 $\{X_i, X_j\}$，定义能量函数：$H_2(X_i, X_j) = -\beta X_i X_j$。即：$X_i, X_j$ 相同时，能量较低；X_i, X_j 不同时，能量较高。
   - 另外对于团 $\{X_i, Y_i\}$，$\{X_i, X_j\}$ 这个整体，定义能量函数：$H_3(X_i) = hX_i$。即：X_i 较大时，能量较高；X_i 较小时，能量较低。

   于是得到整体的能量函数为：

$$H(\mathbb{X}, \mathbb{Y}) = h\sum_i X_i - \beta \sum_{(i,j)\in\mathbb{E}} X_i X_j - \eta \sum_i X_i Y_i$$

   其中 $\mathbb{E}$ 为原始图像的相邻像素连接得到的边。

   考虑到 $P(\mathbb{X}, \mathbb{Y}) = \frac{1}{Z^*} e^{-H(\mathbb{X},\mathbb{Y})}$，根据最大似然准则，则模型优化目标是：

$$\min_{X_i} H(\mathbb{X}, \mathbb{Y}) = \min_{X_i} h\sum_i X_i - \beta \sum_{(i,j)\in\mathbb{E}} X_i X_j - \eta \sum_i X_i Y_i$$

5. 对于能量函数最小化这个最优化问题，由于每个位置的 X_i 都可以取2 个值 $\{+1, -1\}$，因此有 2^N 种取值策略，N 为原始图像的像素数量。如果 N 较大，则参数的搜索空间非常巨大。

   实际任务中通过 `ICM` 算法、模拟退火算法、或者 `graph cuts` 算法来解决这个参数搜索问题。

四、条件随机场 CRF

1. 生成式概率图模型是直接对联合分布进行建模，如隐马尔可夫模型和马尔可夫随机场都是生成式模型。

   判别式概率图模型是对条件分布进行建模，如条件随机场 `Conditional Random Field:CRF`。

2. 条件随机场试图对多个随机变量（它们代表标记序列）在给定观测序列的值之后的条件概率进行建模：

   令 $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ 为观测变量序列，$\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ 为对应的标记变量序列。条件随机场的目标是构建条件概率模型 $P(\mathbf{Y} \mid \mathbf{X})$。

   即：已知观测变量序列的条件下，标记序列发生的概率。

3. 标记随机变量序列 $\mathbf{Y}$ 的成员之间可能具有某种结构：

- 在自然语言处理的词性标注任务中，观测数据为单词序列，标记为对应的词性序列（即动词、名词等词性的序列），标记序列具有线性的序列结构。
- 在自然语言处理的语法分析任务中，观测数据为单词序列，标记序列是语法树，标记序列具有树形结构。

4. 令 $\mathcal{G}=<\mathbb{V},\mathbb{E}>$ 表示与观测变量序列 $\mathbf{X}$ 和标记变量序列 $\mathbf{Y}$ 对应的无向图， Y_v 表示与结点 v 对应的标记随机变量， $n(v)$ 表示结点 v 的邻接结点集。若图 $\mathcal{G}$ 中结点对应的每个变量 Y_v 都满足马尔可夫性，即：

$$P(Y_v \mid \mathbf{X},\mathbf{Y}_{\mathbb{V}-\{v\}})=P(Y_v \mid \mathbf{X},Y_{n(v)})$$

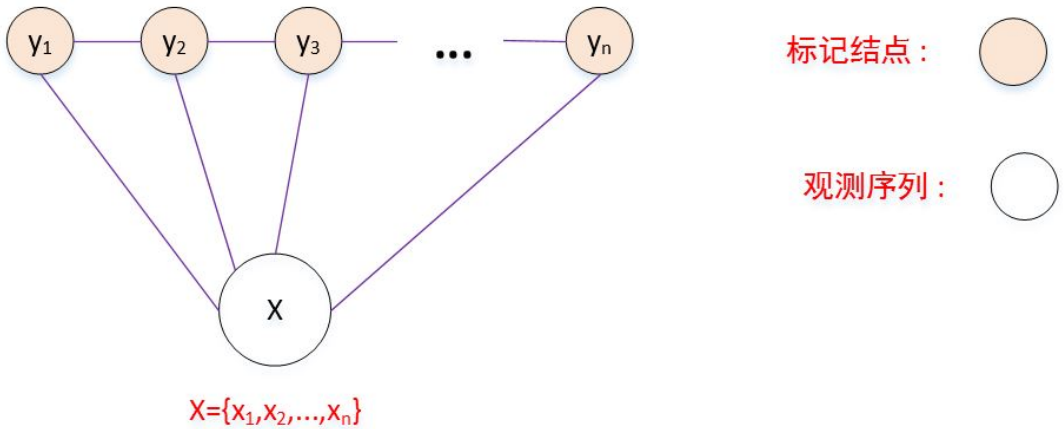
则 $(\mathbf{Y},\mathbf{X})$ 构成了一个条件随机场。

4.1 链式条件随机场

1. 理论上讲，图 $\mathcal{G}$ 可以具有任意结构，只要能表示标记变量之间的条件独立性关系即可。

   但在现实应用中，尤其是对标记序列建模时，最常用的是链式结构，即链式条件随机场`chain-structured CRF`。

   如果没有特殊说明，这里讨论是基于链式条件随机场。



2. 给定观测变量序列 $\mathbf{X}=\{X_1,X_2,\cdots,X_n\}$，链式条件随机场主要包含两种关于标记变量的团：
   - 单个标记变量与 $\mathbf{X}$ 构成的团： $\{Y_i,\mathbf{X}\},i=1,2,\cdots,n$。
   - 相邻标记变量与 $\mathbf{X}$ 构成的团： $\{Y_{i-1},Y_i,\mathbf{X}\},i=2,\cdots,n$。
3. 与马尔可夫随机场定义联合概率的方式类似，条件随机场使用势函数和团来定义条件概率 $P(\mathbf{Y}\mid\mathbf{X})$。

   采用指数势函数，并引入特征函数`feature function`，定义条件概率：

$$P(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=1}^{n-1}\lambda_j t_j(Y_i,Y_{i+1},\mathbf{X},i)+\sum_{k=1}^{K_2}\sum_{i=1}^{n}\mu_k s_k(Y_i,\mathbf{X},i)\right)$$

   其中：
   - $t_j(Y_i,Y_{i+1},\mathbf{X},i)$：在已知观测序列情况下，两个相邻标记位置上的转移特征函数`transition feature function`。
     - 它刻画了相邻标记变量之间的相关关系，以及观察序列 $\mathbf{X}$ 对它们的影响。
     - 位置变量 i 也对势函数有影响。比如：已知观测序列情况下，相邻标记取值`(代词，动词)`出现在序列头部可能性较高，而`(动词，代词)`出现在序列头部的可能性较低。
   - $s_k(Y_i,\mathbf{X},i)$：在已知观察序列情况下，标记位置 i 上的状态特征函数`status feature function`。

        - 它刻画了观测序列 $\mathbf{X}$ 对于标记变量的影响。
        - 位置变量 i 也对势函数有影响。比如：已知观测序列情况下，标记取值 `名词` 出现在序列头部可能性较高，而 `动词` 出现在序列头部的可能性较低。
    - λ_j,μ_k 为参数， Z 为规范化因子（它用于确保上式满足概率的定义）。

      K_1 为转移特征函数的个数，K_2 为状态特征函数的个数。
4. 特征函数通常是实值函数，用来刻画数据的一些很可能成立或者预期成立的经验特性。

   一个特征函数的例子（词性标注）：

$$t_j(Y_i,Y_{i+1},\mathbf{X},i)=\begin{cases}1, & \text{if}\quad Y_{i+1}=[\mathrm{P}],\ Y_i=[\mathrm{V}]\ \text{and}\ X_i=\text{"knock"}\\ 0, & \text{otherwise}\end{cases}$$

$$s_k(Y_i,\mathbf{X},i)=\begin{cases}1, & \text{if}\quad Y_i=[\mathrm{V}]\ \text{and}\ X_i=\text{"knock"}\\ 0, & \text{otherwise}\end{cases}$$

    - 转移特征函数刻画的是：第 i 个观测值 X_i 为单词 `"knock"` 时，相应的标记 Y_i 和 Y_{i+1} 很可能分别为 `[V]` 和 `[P]` 。
    - 状态特征函数刻画的是： 第 i 个观测值 X_i 为单词 `"knock"` 时，标记 Y_i 很可能为 `[V]` 。
5. 条件随机场与马尔可夫随机场均使用团上的势函数定义概率，二者在形式上没有显著区别。

   条件随机场处理的是条件概率，马尔可夫随机场处理的是联合概率。
6. $P(\mathbf{Y}\mid\mathbf{X})$ 的形式类似于逻辑回归。事实上，条件随机场是逻辑回归的序列化版本。
    - 逻辑回归是用于分类问题的对数线性模型。
    - 条件随机场是用于序列化标注的对数线性模型。

4.1.1 CRF 的简化形式

1. 注意到条件随机场中的同一个特征函数在各个位置都有定义，因此可以对同一个特征在各个位置求和，将局部特征函数转化为一个全局特征函数。

   这样就可以将条件随机场写成权值向量和特征向量的内积形式，即条件随机场的简化形式。
2. 设有 K_1 个转移特征函数， K_2 个状态特征函数。令 $K=K_1+K_2$ ，定义：

$$f_k(Y_i,Y_{i+1},\mathbf{X},i)=\begin{cases}t_k(Y_i,Y_{i+1},\mathbf{X},i), & k=1,2,\cdots,K_1\\ s_l(Y_i,\mathbf{X},i), & k=K_1+l;l=1,2,\cdots,K_2\end{cases}$$

   > 注意：t_k要求 $1\le i<n$， 而 s_l 要求 $1\le i\le n$ ，因此对于边界条件要特殊处理。

   对转移与状态函数在各个位置 i 求和，记作：

$$f_k(\mathbf{Y},\mathbf{X})=\sum_{i=1}^{n}f_k(Y_i,Y_{i+1},\mathbf{X},i),\quad k=1,2,\cdots,K$$

   其中 $\mathbf{Y}=\{Y_1,Y_2,\cdots,Y_n\}$ 为标记变量序列， $\mathbf{X}=\{X_1,X_2,\cdots,X_n\}$ 为观测变量序列。

   该式子刻画的是特征函数在所有位置上的和，可以理解为特征函数在所有位置上的得的总分。
3. 用 w_k 表示特征 $f_k(\mathbf{Y},\mathbf{X})$ 的权值，即：

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

   则条件随机场可以简化为：

$$P(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=1}^{n-1}\lambda_j t_j(Y_i,Y_{i+1},\mathbf{X},i)+\sum_{k=1}^{K_2}\sum_{i=1}^{n}\mu_k s_k(Y_i,\mathbf{X},i)\right)$$
$$=\frac{1}{Z}\exp\left(\sum_{k=1}^{K_1}\lambda_k\sum_{i=1}^{n-1} t_k(Y_i,Y_{i+1},\mathbf{X},i)+\sum_{l=1}^{K_2}\mu_l\sum_{i=1}^{n} s_l(Y_i,\mathbf{X},i)\right)$$
$$=\frac{1}{Z}\exp\left(\sum_{k=1}^{K} w_k f_k(\mathbf{Y},\mathbf{X})\right)$$

其中 $Z=\sum_{\mathbf{Y}}\exp\left(\sum_{k=1}^{K} w_k f_k(\mathbf{Y},\mathbf{X})\right)$， $\sum_{\mathbf{Y}}$ 表示对所有可能的标记序列进行求和。

4. 定义权值向量为 $\vec{\mathbf{w}}=(w_1,w_2,\cdots,w_K)^T$ ，定义全局特征向量为：
$$\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X})=(f_1(\mathbf{Y},\mathbf{X}),f_2(\mathbf{Y},\mathbf{X}),\cdots,f_K(\mathbf{Y},\mathbf{X}))^T$$

则条件随机场可以简化为：
$$P(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z}\exp\left(\sum_{k=1}^{K} w_k f_k(\mathbf{Y},\mathbf{X})\right)=\frac{1}{Z}\exp\left(\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X})\right)$$

其中 $Z=\sum_{\mathbf{Y}}\exp\left(\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X})\right)$， $\sum_{\mathbf{Y}}$ 表示对所有可能的标记序列进行求和。

$\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X})$ 的物理意义为：已知序列 $\mathbf{X}$ 的条件下，标记序列为 $\mathbf{Y}$ 的未归一化的概率。它就是每个特征函数的总分的加权和（权重为特征函数的权重）。

4.1.2 CRF 的矩阵形式

1. 假设标记变量 $Y_i,i=1,2,\cdots,n$ 的取值集合为 $\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_m\}$， 其中 m 是标记的取值个数。

   对于观测变量序列和标记变量序列的每个位置 $i=0,1,2,\cdots,n$ ，定义一个 m 阶矩阵：
$$\mathbf{M}_i(\mathbf{X})=\begin{bmatrix} M_i(\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_1\mid\mathbf{X}) & M_i(\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2\mid\mathbf{X}) & \cdots & M_i(\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_m\mid\mathbf{X}) \\ M_i(\tilde{\mathbf{y}}_2,\tilde{\mathbf{y}}_1\mid\mathbf{X}) & M_i(\tilde{\mathbf{y}}_2,\tilde{\mathbf{y}}_2\mid\mathbf{X}) & \cdots & M_i(\tilde{\mathbf{y}}_2,\tilde{\mathbf{y}}_m\mid\mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ M_i(\tilde{\mathbf{y}}_m,\tilde{\mathbf{y}}_1\mid\mathbf{X}) & M_i(\tilde{\mathbf{y}}_m,\tilde{\mathbf{y}}_2\mid\mathbf{X}) & \cdots & M_i(\tilde{\mathbf{y}}_m,\tilde{\mathbf{y}}_m\mid\mathbf{X}) \end{bmatrix}_{m\times m}$$

   其中：$M_i(\tilde{\mathbf{y}}_u,\tilde{\mathbf{y}}_v\mid\mathbf{X})=M_i(Y_i=\tilde{\mathbf{y}}_u,Y_{i+1}=\tilde{\mathbf{y}}_v\mid\mathbf{X}),u,v=1,2,\cdots,m$ 。i其中：
$$M_i(Y_i=\tilde{\mathbf{y}}_u,Y_{i+1}=\tilde{\mathbf{y}}_v\mid\mathbf{X})=\exp\left(\sum_{k=1}^{K} w_k f_k(Y_i=\tilde{\mathbf{y}}_u,Y_{i+1}=\tilde{\mathbf{y}}_v,\mathbf{X},i)\right)$$

   $\left(\sum_{k=1}^{K} w_k f_k(\cdot)\right)$ 物理意义是：在位置 i ，所有转移特征函数值加上所有状态特征函数值。对其取指数则是非规范化概率。

2. 引入两个特殊的状态标记：起点状态标记 $Y_0=start$ 表示起始符，终点状态标记 $Y_{n+1}=stop$ 表示终止符。

   - $M_0(\tilde{\mathbf{y}}_u,\tilde{\mathbf{y}}_v\mid\mathbf{X})$ 表示第 0 个位置的标记为 $\tilde{\mathbf{y}}_u$ ， 第 1个位置的标记为 $\tilde{\mathbf{y}}_v$ 。

     第 0 个位置是一个虚拟符号，表示这是一个新的序列。因为 Y_0 状态取值只能是 $start$ ， 则：

$$\mathbf{M}_0(\mathbf{X})=\begin{bmatrix} M_0(start,\tilde{\mathbf{y}}_1\mid\mathbf{X}) & M_0(start,\tilde{\mathbf{y}}_2\mid\mathbf{X}) & \cdots & M_0(start,\tilde{\mathbf{y}}_m\mid\mathbf{X}) \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}_{m\times m}$$

- $M_n(\tilde{\mathbf{y}}_u,\tilde{\mathbf{y}}_v\mid\mathbf{X})$ 表示第 n 个位置的标记为 $\tilde{\mathbf{y}}_u$ ，第 $n+1$ 个位置的标记为 $\tilde{\mathbf{y}}_v$ 。由于序列长度为 n，因此第 $n+1$ 个位置是一个虚拟符号，表示该序列结束。因为 Y_{n+1} 的取值只能是 $stop$ ，则：

$$\mathbf{M}_n(\mathbf{X})=\begin{bmatrix} M_n(\tilde{\mathbf{y}}_1,stop\mid\mathbf{X}) & 0 & 0 & \cdots & 0 \\ M_n(\tilde{\mathbf{y}}_2,stop\mid\mathbf{X}) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_n(\tilde{\mathbf{y}}_m,stop\mid\mathbf{X}) & 0 & 0 & \cdots & 0 \end{bmatrix}_{m\times m}$$

  因此 $\mathbf{M}_0(\mathbf{X})$ 和 $\mathbf{M}_n(\mathbf{X})$ 中包含大量的 0 。

3. 给定观测变量序列 $\mathbf{X}=\{X_1,X_2,\cdots,X_n\}$ ， 标记变量序列 $\mathbf{Y}=\{Y_1,Y_2,\cdots,Y_n\}$ 可以这样产生：
   - 首先位于起点状态 Y_0 。
   - 然后从 Y_i 转移到 Y_{i+1}， $i=0,1,2,\cdots,n$ 。
   - 最后到达终点状态 Y_{n+1} 。

   于是条件概率为：

$$P(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z}\prod_{i=0}^{n}M_i(Y_i,Y_{i+1}\mid\mathbf{X})$$

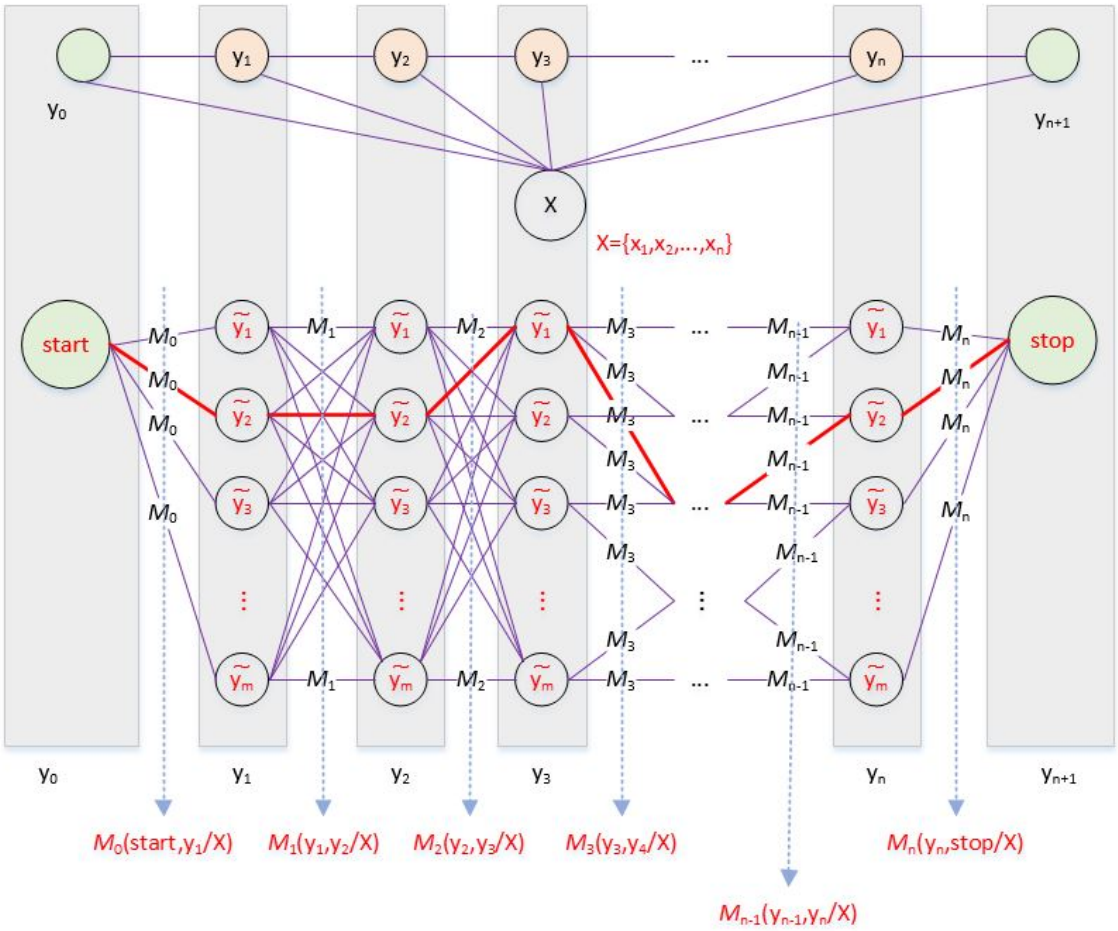
   其中 Z 是以 $start$ 为起点，以 $stop$ 为终点的所有标记路径的非规范化概率 $\prod_{i=0}^{n}M_i(Y_i,Y_{i+1}\mid\mathbf{X})$ 之和：

$$Z=\sum_{Y_i\in\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_m\}}\ \sum_{Y_{i+1}\in\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_m\}}\prod_{i=0}^{n}M_i(Y_i,Y_{i+1}\mid\mathbf{X})$$

   它也等于矩阵乘积 $(\mathbf{M}_0(\mathbf{X})\mathbf{M}_1(\mathbf{X})\cdots\mathbf{M}_n(\mathbf{X}))$ 的结果的第一个元素（位于第一行第一列）的元素。

   > 根据 $\mathbf{M}_0(\mathbf{X})$ 和 $\mathbf{M}_n(\mathbf{X})$ 的性质，该结果矩阵只有第一个元素非零，其他所有元素都为 0）。

4. 如下图所示，上半部分是条件随机场的示意图，下半部分是条件随机场所有可能的路径。
   - 除了起点和终点外，每个节点都有 m 个可能的取值。
   - 起点取值只能是 $start$， 终点取值只能是 $stop$ 。
   - 红色粗实线给出了从起点到终点的一条可能的路径。



5. 矩阵形式和前述形式是统一的：

$$P(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z}\prod_{i=0}^{n}M_i(Y_i,Y_{i+1}\mid\mathbf{X})=\frac{1}{Z}\prod_{i=0}^{n}\exp\left(\sum_{k=1}^{K}w_kf_k(Y_i,Y_{i+1},\mathbf{X},i)\right)$$
$$=\frac{1}{Z}\exp\left(\sum_{i=0}^{n}\sum_{k=1}^{K}w_kf_k(Y_i,Y_{i+1},\mathbf{X},i)\right)$$

与前述形式区别在于：这里由于引入了两个特殊的状态标记：起点状态标记、终点状态标记，从而累加的区间为 $\sum_{i=0}^{n}$ 。

由于引入的状态标记是人为引入且状态固定的，因此相当于引入了常量。它会相应的改变 Z 的值，最终结果是统一的。

4.2 概率计算问题

4.2.1 概率计算

1. 条件随机场的概率计算问题是：已知条件随机场 $P(\mathbf{Y}\mid\mathbf{X})$，其中 Y_i 的取值集合为 $\mathcal{Y}=\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_m\}$ ， 给定观测值序列 $\tilde{\mathbf{X}}=\{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\cdots,\tilde{\mathbf{x}}_n\}$， 给定标记值序列 $\tilde{\mathbf{Y}}=\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_n\}$，其中 $\tilde{\mathbf{y}}_i\in\mathcal{Y}$：
   - 计算条件概率： $P(Y_i=\tilde{\mathbf{y}}_i\mid\tilde{\mathbf{X}})$ 。
   - 计算条件概率： $P(Y_i=\tilde{\mathbf{y}}_i,Y_{i+1}=\tilde{\mathbf{y}}_{i+1}\mid\tilde{\mathbf{X}})$ 。

   类似隐马尔可夫模型，可以通过前向-后向算法解决条件随机场的概率计算问题。

2. 采用 CRF 的矩阵形式，对于每个指标 $i=0,1,\cdots,n+1$，（包括了起点标记和终点标记）：
   - 定义前向概率 $\alpha_i(Y_i\mid\tilde{\mathbf{X}})$：表示在位置 i 的标记是 Y_i，并且到位置 i 的前半部分标记序列的非规范化概率。

     由于 Y_i 的取值有 m 个，因此前向概率向量 $\vec{\alpha}_i(\tilde{\mathbf{X}})$ 是 m 维列向量：
     $$\vec{\alpha}_i(\tilde{\mathbf{X}})=\begin{bmatrix}\alpha_i(Y_i=\mathbf{y}_1\mid\tilde{\mathbf{X}})\\ \alpha_i(Y_i=\mathbf{y}_2\mid\tilde{\mathbf{X}})\\ \vdots\\ \alpha_i(Y_i=\mathbf{y}_m\mid\tilde{\mathbf{X}})\end{bmatrix}$$
   - 定义后向概率 $\beta_i(Y_i\mid\tilde{\mathbf{X}})$ 表示在位置 i 的标记是 Y_i，并且从位置 $i+1$ 的后半部分标记序列的非规范化概率。

     由于 Y_i 的取值有 m 个，因此后向概率向量 $\vec{\beta}_i(\tilde{\mathbf{X}})$ 也是 m 维列向量：
     $$\vec{\beta}_i(\tilde{\mathbf{X}})=\begin{bmatrix}\beta_i(Y_i=\mathbf{y}_1\mid\tilde{\mathbf{X}})\\ \beta_i(Y_i=\mathbf{y}_2\mid\tilde{\mathbf{X}})\\ \vdots\\ \beta_i(Y_i=\mathbf{y}_m\mid\tilde{\mathbf{X}})\end{bmatrix}$$

3. 根据 CRF 的矩阵形式， 前向概率 $\alpha_i(Y_i\mid\tilde{\mathbf{X}})$ 的递推形式为：
   $$\alpha_0(Y_0\mid\tilde{\mathbf{X}})=\begin{cases}1, & \text{if}\quad Y_0=start\\ 0, & \text{otherwise}\end{cases}$$
   $$\alpha_{i+1}(Y_{i+1}\mid\tilde{\mathbf{X}})=\sum_{Y_i\in\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_m\}}\alpha_i(Y_i\mid\tilde{\mathbf{X}})M_i(Y_i,Y_{i+1}\mid\tilde{\mathbf{X}}),\quad i=0,1,\cdots,n$$
   - 其物理意义是：所有从起点到 Y_i 的路径再通过 $M_i(Y_i,Y_{i+1}\mid\tilde{\mathbf{X}})$ 转移到 Y_{i+1}。
   - 根据前向概率向量 $\vec{\alpha}_i(\tilde{\mathbf{X}})$ 的定义，则也可以表示为：
     $$\vec{\alpha}_{i+1}^T(\tilde{\mathbf{X}})=\vec{\alpha}_i^T(\tilde{\mathbf{X}})M_i(\tilde{\mathbf{X}}),\quad i=0,1,\cdots,n$$

4. 根据 CRF 的矩阵形式，后向概率 $\beta_i(\mathbf{y}_i\mid\tilde{\mathbf{X}})$ 的递归形式为：
   $$\beta_{n+1}(Y_{n+1}\mid\tilde{\mathbf{X}})=\begin{cases}1, & \text{if}\quad Y_{n+1}=stop\\ 0, & \text{otherwise}\end{cases}$$
   $$\beta_i(Y_i\mid\tilde{\mathbf{X}})=\sum_{Y_{i+1}\in\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_m\}}\beta_{i+1}(Y_{i+1}\mid\tilde{\mathbf{X}})M_i(Y_i,Y_{i+1}\mid\tilde{\mathbf{X}}),\quad i=0,1,\cdots,n$$
   - 其物理意义可以这样理解：从 Y_i 到终点的路径可以这样分解：
     - 先通过 $M_i(Y_i,Y_{i+1}\mid\tilde{\mathbf{X}})$ 从 Y_i 到 Y_{i+1}。
     - 再通过 Y_{i+1} 到终点。
     - 对所有可能的 Y_{i+1} 累加，即得到从 Y_i 到终点的路径。
   - 根据后向概率向量 $\vec{\beta}_i(\tilde{\mathbf{X}})$ 的定义，则也可以表示为：
     $$\vec{\beta}_i(\tilde{\mathbf{X}})=M_i(\tilde{\mathbf{X}})\vec{\beta}_{i+1}(\tilde{\mathbf{X}}),i=0,1,\cdots,n$$

5. 根据前向-后向向量定义可以得到：$Z=\vec{\alpha}_{n+1}(\tilde{\mathbf{X}})\cdot\vec{\mathbf{1}}=\vec{\beta}_0(\tilde{\mathbf{X}})\cdot\vec{\mathbf{1}}$。其中：
   - Z 为 CRF 的矩阵形式中的归一化因子。
   - $\vec{\mathbf{1}}$ 为元素均为 1 的 m 维列向量。

6. 概率计算： 给定观测序列 $\tilde{\mathbf{X}}=\{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\cdots,\tilde{\mathbf{x}}_n\}$， 给定标记值序列 $\tilde{\mathbf{Y}}=\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_n\}$， 其中 $\tilde{\mathbf{y}}_i\in\mathcal{Y}$ ，据前向-后向向量的定义，有：

   - 标记序列在位置 i 处标记 $Y_i=\tilde{\mathbf{y}}_i$ 的条件概率为：

$$P(Y_i=\tilde{\mathbf{y}}_i\mid\tilde{\mathbf{X}})=\frac{\alpha_i(Y_i=\tilde{\mathbf{y}}_i\mid\tilde{\mathbf{X}})\beta_i(Y_i=\tilde{\mathbf{y}}_i\mid\tilde{\mathbf{X}})}{Z}$$

   - 标记序列在位置 i 处标记 $Y_i=\tilde{\mathbf{y}}_i$ ， 且在位置 $i+1$ 处标记 $Y_{i+1}=\tilde{\mathbf{y}}_{i+1}$ 的条件概率为：

$$\begin{aligned}&P(Y_i=\tilde{\mathbf{y}}_i,Y_{i+1}=\tilde{\mathbf{y}}_{i+1}\mid\tilde{\mathbf{X}})\\=&\frac{\alpha_i(Y_i=\tilde{\mathbf{y}}_i\mid\tilde{\mathbf{X}})M_i(Y_i=\tilde{\mathbf{y}}_i,Y_{i+1}=\tilde{\mathbf{y}}_{i+1}\mid\tilde{\mathbf{X}})\beta_{i+1}(Y_{i+1}=\tilde{\mathbf{y}}_{i+1}\mid\tilde{\mathbf{X}})}{Z}\end{aligned}$$

     其中： $Z=\vec{\alpha}_{n+1}(\tilde{\mathbf{X}})\cdot\vec{\mathbf{1}}=\vec{\beta}_0(\tilde{\mathbf{X}})\cdot\vec{\mathbf{1}}$ 。

4.2.2 期望值计算

1. 利用前向-后向向量可以计算特征函数 $f_k(Y_i,Y_{i+1},\mathbf{X},i),k=1,2,\cdots,K$关于联合分布$P(\mathbf{X},\mathbf{Y})$ 和条件分布 $P(\mathbf{Y}\mid\mathbf{X})$ 的数学期望。

   - 特征函数 $f_k(Y_i,Y_{i+1},\mathbf{X},i)$ 关于条件分布 $P(\mathbf{Y}\mid\mathbf{X})$ 的数学期望为：

$$\begin{aligned}&\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[f_k(Y_i,Y_{i+1},\mathbf{X},i)]=\sum_{\mathbf{Y}}P(\mathbf{Y}\mid\mathbf{X})f_k(Y_i,Y_{i+1},\mathbf{X},i)\\=&\sum_{Y_i,Y_{i+1}}f_k(Y_i,Y_{i+1},\mathbf{X},i)\frac{\alpha_i(Y_i\mid\mathbf{X})M_i(Y_i,Y_{i+1}\mid\mathbf{X})\beta_{i+1}(Y_{i+1}\mid\mathbf{X})}{Z}\end{aligned}$$

     其中 $Z=\vec{\alpha}_{n+1}(\mathbf{X})\cdot\vec{\mathbf{1}}=\vec{\beta}_0(\mathbf{X})\cdot\vec{\mathbf{1}}$ 。

     如果合并所有的位置 i ， 则有特征函数 $f_k(\mathbf{Y},\mathbf{X})$ 的期望为：

$$\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[f_k(\mathbf{Y},\mathbf{X})]=\sum_{i=0}^{n}\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[f_k(Y_i,Y_{i+1},\mathbf{X},i)]$$

     其物理意义为：在指定观测序列 $\mathbf{X}$ 的条件下，特征 $f_k(\mathbf{Y},\mathbf{X})$ 的均值。

   - 假设 $\mathbf{X}$ 的经验分布为 $\tilde{P}(\mathbf{X})$ ， 则特征函数 $f_k(Y_i,Y_{i+1},\mathbf{X},i)$ 关于联合分布$P(\mathbf{X},\mathbf{Y})$ 的数学期望为：

$$\begin{aligned}&\mathbb{E}_{P(\mathbf{X},\mathbf{Y})}[f_k(Y_i,Y_{i+1},\mathbf{X},i)]=\sum_{\mathbf{X},\mathbf{Y}}P(\mathbf{X},\mathbf{Y})f_k(f_k(Y_i,Y_{i+1},\mathbf{X},i)\\=&\sum_{\mathbf{X}}\tilde{P}(\mathbf{X})\sum_{\mathbf{Y}}P(\mathbf{Y}\mid\mathbf{X})f_k(Y_i,Y_{i+1},\mathbf{X},i)\\=&\sum_{\mathbf{X}}\tilde{P}(\mathbf{X})\sum_{Y_i,Y_{i+1}}f_k(Y_i,Y_{i+1},\mathbf{X},i)\frac{\alpha_i(Y_i\mid\mathbf{X})M_i(Y_i,Y_{i+1}\mid\mathbf{X})\beta_{i+1}(Y_{i+1}\mid\mathbf{X})}{Z}\end{aligned}$$

     如果合并所有的位置 i ， 则有特征函数 $f_k(\mathbf{Y},\mathbf{X})$ 的期望为：

$$\mathbb{E}_{P(\mathbf{X},\mathbf{Y})}[f_k(\mathbf{Y},\mathbf{X},i)]=\sum_{i=0}^{n}\mathbb{E}_{P(\mathbf{X},\mathbf{Y})}[f_k(Y_i,Y_{i+1},\mathbf{X},i)]$$

     其物理意义为：在所有可能观测序列和标记序列中， $f_k(\mathbf{Y},\mathbf{X})$ 预期发生的次数。

2. 上述两个式子是特征函数数学期望的一般计算公式。

- 对于转移特征函数 $t_k(Y_i,Y_{i+1},\mathbf{X},i),k=1,2,\cdots,K_1$ ， 可以将上式中的 f_k 替换成 t_k，即得到转移特征函数的两个期望。
- 对于状态特征函数 $s_l(Y_i,\mathbf{X},i),l=1,2,\cdots,K_2$ ，可以将 f_k 替换成 $s_l,k=K_1+l$ ，即得到状态特征函数的两个期望。

4.3 CRF 的学习算法

1. 条件随机场模型实际上是定义在时序数据上的对数线性模型，其学习方法包括：极大似然估计、正则化的极大似然估计。

   具体的实现算法有：改进的迭代尺度法`Improved Iterative Scaling:IIS`、 梯度下降法、拟牛顿法。

2. 给定训练数据集 $\mathbb{D}=\{(\tilde{\mathbf{X}}_1,\tilde{\mathbf{Y}}_1),(\tilde{\mathbf{X}}_2,\tilde{\mathbf{Y}}_2),\cdots,(\tilde{\mathbf{X}}_N,\tilde{\mathbf{Y}}_N)\}$，其中：

   - $\tilde{\mathbf{X}}_i=\{\tilde{\mathbf{x}}_{i,1},\tilde{\mathbf{x}}_{i,2},\cdots,\tilde{\mathbf{x}}_{i,n_i}\}$ 代表第 i 个观测序列，其长度为 n_i 。

     $\tilde{\mathbf{x}}_{i,j}$ 代表第 i 个观测序列的第 j 个位置的值。

   - $\tilde{\mathbf{Y}}_i=\{\tilde{\mathbf{y}}_{i,1},\tilde{\mathbf{y}}_{i,2},\cdots,\tilde{\mathbf{y}}_{i,n_i}\}$ 代表第 i 个标记序列，其长度为 n_i 。

     $\tilde{\mathbf{y}}_{i,j}$ 代表第 i 个标记序列的第 j 个位置的值，其中 $\tilde{\mathbf{y}}_{i,j}\in\mathcal{Y}=\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_m\}$ 。

   - 同一组观测序列和标记序列的长度相同，不同组的观测序列之间的长度可以不同。

   给定 K 个特征函数 $f_k(\mathbf{Y},\mathbf{X})=\sum_{i=1}^n f_k(Y_i,Y_{i+1},\mathbf{X},i),\quad k=1,2,\cdots,K$ ，其中：

   $$f_k(Y_i,Y_{i+1},\mathbf{X},i)=\begin{cases}t_k(Y_i,Y_{i+1},\mathbf{X},i), & k=1,2,\cdots,K_1\\ s_l(Y_i,\mathbf{X},i), & k=K_1+l;l=1,2,\cdots,K_2\end{cases}$$

   而 $t_k(Y_i,Y_{i+1},\mathbf{X},i)$ 为转移特征函数，$s_l(Y_i,\mathbf{X},i)$ 为状态特征函数。

   条件随机场的学习问题是：给定训练数据集 $\mathbb{D}$ 和 K 个特征函数，估计条件随机场的模型参数。即模型中的参数 $w_k,k=1,2,\cdots,K$ ：

   $$P(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z}\exp\left(\sum_{k=1}^K w_k f_k(\mathbf{Y},\mathbf{X})\right)$$

   其中 $Z=\sum_{\mathbf{Y}}\exp\left(\sum_{k=1}^K w_k f_k(\mathbf{Y},\mathbf{X})\right)$ 为归一化参数。

   注意到：Z 包含参数 w_k ，同时 Z 也受 $\mathbf{X}$ 影响，因此 将 Z 记作 $Z_{\vec{\mathbf{w}}}(\mathbf{X})$ 。因此将待求解的模型重新写作：

   $$P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z_{\vec{\mathbf{w}}}(\mathbf{X})}\exp\left(\sum_{k=1}^K w_k f_k(\mathbf{Y},\mathbf{X})\right)$$

   > Z 不受 $\mathbf{Y}$ 的影响，因为 $\sum_{\mathbf{Y}}$ 将变量 $\mathbf{Y}$ 吸收掉。

3. 考虑观测序列和标记序列 $(\tilde{\mathbf{X}}_i,\tilde{\mathbf{Y}}_i)$， 根据经验分布 $\tilde{P}(\mathbf{X},\mathbf{Y})$ ， 该对序列出现的次数为 $N\times\tilde{P}(\mathbf{X}=\tilde{\mathbf{X}}_i,\mathbf{Y}=\tilde{\mathbf{Y}}_i)$ 。

   - 利用条件概率和经验分布 $\tilde{P}(\mathbf{X})$ ，出现如此多次数的 $(\mathbb{X}_s,\mathbf{Y}_s)$ 概率为：

     $$\left[\tilde{P}(\mathbf{X}=\tilde{\mathbf{X}}_i)P_{\vec{\mathbf{w}}}(\mathbf{Y}=\tilde{\mathbf{Y}}_i\mid\mathbf{X}=\tilde{\mathbf{X}}_i)\right]^{N\times\tilde{P}(\mathbf{X}=\tilde{\mathbf{X}}_i,\mathbf{Y}=\tilde{\mathbf{Y}}_i)}$$

   - 考虑整个训练集，则训练数据集 $\mathbb{D}$ 发生的概率为：$\prod_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\right]^{N\times\tilde{P}(\mathbf{X},\mathbf{Y})}$ 。

     其对数似然函数为：

$$L_{\vec{\mathbf{w}}}=\log\prod_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\right]^{N\tilde{P}(\mathbf{X},\mathbf{Y})}=\sum_{\mathbf{X},\mathbf{Y}}\log\left[\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\right]^{N\tilde{P}(\mathbf{X},\mathbf{Y})}$$
$$=\sum_{\mathbf{X},\mathbf{Y}}\left[N\tilde{P}(\mathbf{X},\mathbf{Y})\log\left[\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\right]\right]$$
$$=\sum_{\mathbf{X},\mathbf{Y}}\left[N\tilde{P}(\mathbf{X},\mathbf{Y})\log\tilde{P}(\mathbf{X})\right]+\sum_{\mathbf{X},\mathbf{Y}}\left[N\tilde{P}(\mathbf{X},\mathbf{Y})\log P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\right]$$

- 因为最终需要求解对数似然的最大值，考虑到 $\sum_{\mathbf{X},\mathbf{Y}}\left[N\tilde{P}(\mathbf{X},\mathbf{Y})\log\tilde{P}(\mathbf{X})\right]$ 为常数，所以去掉该项，则有：

  $$L_{\vec{\mathbf{w}}}=\sum_{\mathbf{X},\mathbf{Y}}\left[N\tilde{P}(\mathbf{X},\mathbf{Y})\log P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\right]$$

  忽略约去常系数 N，则有：$L_{\vec{\mathbf{w}}}=\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\log P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\right]$ 。

  > 与最大熵情况类似，这里使用了经验分布 $\tilde{P}(\mathbf{X},\mathbf{Y}),\tilde{P}(\mathbf{X})$ ，但是并没有使用 $\frac{\tilde{P}(\mathbf{X},\mathbf{Y})}{\tilde{P}(\mathbf{X})}$ 来作为 $P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})$ 的估计值，因为我们是针对该条件概率建模。

- 根据 $P_{\vec{\mathbf{w}}}$ 的定义，有：

  $$L_{\vec{\mathbf{w}}}=\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\log\frac{1}{Z_{\vec{\mathbf{w}}}(\mathbf{X})}\exp\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)\right]$$
  $$=\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)-\tilde{P}(\mathbf{X},\mathbf{Y})\log Z_{\vec{\mathbf{w}}}(\mathbf{X})\right]$$
  $$=\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)\right]-\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\log Z_{\vec{\mathbf{w}}}(\mathbf{X})\right]$$
  $$=\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)\right]-\sum_{\mathbf{X}}\left[\tilde{P}(\mathbf{X})\log Z_{\vec{\mathbf{w}}}(\mathbf{X})\right]$$

  其形式和最大熵算法完全一致，因此可以直接使用最大熵算法的学习算法。

  > 这也说明了，从最大熵原理可以推导出条件随机场的条件概率表示形式。

4. 给定训练数据集 $\mathbb{D}$ ，则可以获取经验概率分布 $\tilde{P}(\mathbf{X},\mathbf{Y})$ 和 $\tilde{P}(\mathbf{X})$，从而可以通过极大化训练数据的对数似然函数 $L_{\vec{\mathbf{w}}}$ 来求解模型。

4.3.1 改进的迭代尺度法

1. `IIS` 算法通过迭代的方法不断优化对数似然函数增量的下界，达到极大化对数似然函数的目的。具体推导可以参看最大熵算法。
2. 假设模型的当前参数向量是 $\vec{\mathbf{w}}=(w_1,w_2,\cdots,w_K)^T$， 参数向量的增量为 $\vec{\delta}=(\delta_1,\delta_2,\cdots,\delta_K)^T$。更新的参数向量为 $\vec{\mathbf{w}}+\vec{\delta}=(w_1+\delta_1,w_2+\delta_2,\cdots,w_K+\delta_K)^T$ 。

   `IIS` 推导的结果为：每一轮迭代中， δ_k 满足：

   $$\sum_{\mathbf{X}}\left(\tilde{P}(\mathbf{X})\sum_{\mathbf{Y}}[P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_kf^o(\mathbf{Y},\mathbf{X}))]\right)=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

   将 $\tilde{P}(\mathbf{X})$ 放到 $\sum_{\mathbf{Y}}$ 右侧，重新整理为：

$$\sum_{\mathbf{X},\mathbf{Y}}\left(\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k f^o(\mathbf{Y},\mathbf{X}))\right)=E_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

其中：

- $f^o(\mathbf{Y},\mathbf{X})=\sum_{k=1}^K f_k(\mathbf{Y},\mathbf{X})=\sum_{k=1}^K\sum_{i=0}^n f_k(Y_i,Y_{i+1},\mathbf{X},i)$：为所有特征函数在序列 $(\mathbf{X},\mathbf{Y})$ 的所有位置的总和。
- $\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)=\sum_{\mathbf{X},\mathbf{Y}}[\tilde{P}(\mathbf{X},\mathbf{Y})f_k(\mathbf{Y},\mathbf{X})]=\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\sum_{i=0}^n f_k(Y_i,Y_{i+1},\mathbf{X},i)\right]$：为特征函数 f_k 在训练集 $\mathbb{D}$ 中对所有序列样本的所有位置上的求和。

3. `CRF` 学习的改进迭代尺度算法 `IIS`：
   - 输入：
     - 特征函数 f_k
     - 经验分布 $\tilde{P}(\mathbf{X},\mathbf{Y})$
     - 经验分布 $\tilde{P}(\mathbf{X})$
   - 输出：
     - 参数估计值 $\hat{\vec{\mathbf{w}}}$
     - 模型 $P_{\hat{\vec{\mathbf{w}}}}(\mathbf{Y}\mid\mathbf{X})$
   - 算法步骤：
     - 初始化：对所有的 $k=1,2,\cdots,K$，取值 $w_k=0$。
     - 迭代，迭代的停止条件是：所有 w_k 都收敛。迭代步骤为：
       - 对每一个 $k=1,2,\cdots,K$，从方程中计算出 δ_k：

$$\sum_{\mathbf{X}}\left(\tilde{P}(\mathbf{X})\sum_{\mathbf{Y}}[P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k f^o(\mathbf{Y},\mathbf{X}))]\right)=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

       - 更新 w_k 的值：$w_k\leftarrow w_k+\delta_k$。如果不是所有 w_k 都收敛，继续迭代。
     - 迭代终止时，$\hat{\vec{\mathbf{w}}}=(w_1,w_2,\cdots,w_K)^T$。

4.3.1.1 算法 S

1. 在 `IIS` 算法中，$f^o(\mathbf{Y},\mathbf{X})$ 为所有特征函数在序列 $(\mathbf{X},\mathbf{Y})$ 的所有位置的总和。对于不同的数据 $(\mathbf{X},\mathbf{Y})$，$f^o(\mathbf{Y},\mathbf{X})$ 很有可能不同。

   选择一个常数 S，定义松弛特征：$s(\mathbf{Y},\mathbf{X})=S-f^o(\mathbf{Y},\mathbf{X})$。
   - 通常选择足够大的常数 S，使得对于训练数据集的所有数据 $(\mathbf{X},\mathbf{Y})\in\mathbb{D}$，$s(\mathbf{Y},\mathbf{X})=S-f^o(\mathbf{Y},\mathbf{X})\geq 0$ 成立。
   - 当每个特征函数的取值范围都是 $\{0,1\}$ 时，则可以将 S 选取为：$K\times\sum_{i=1}^N n_i$。其物理意义为：所有潜在的特征函数取 `1` 的位置的总数，乘以特征函数的数量。
2. 将松弛特征代入，有：

$$\begin{aligned}&\sum_{\mathbf{X}}\left(\tilde{P}(\mathbf{X})\sum_{\mathbf{Y}}[P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k f^o(\mathbf{Y},\mathbf{X}))]\right)\\&=\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k(S-s(\mathbf{Y},\mathbf{X})))\end{aligned}$$

考虑到 $s(\mathbf{Y},\mathbf{X})\geq 0$ 在 $\mathbb{D}$ 上恒成立，因此有：

$$\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k(S-s(\mathbf{Y},\mathbf{X})))\leq\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k S)$$

因此对于下面两个方程的解：

$$\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k^{(1)}f^o(\mathbf{Y},\mathbf{X}))=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

$$\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k^{(2)}S)=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

必然有： $\delta_k^{(2)}\leq\delta_k^{(1)}$ 。

> 如果 $\delta_k^{(2)}>\delta_k^{(1)}$ ，则有：
>
> $$\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)=\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k^{(2)}S)$$
>
> $$>\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k^{(1)}f^o(\mathbf{Y},\mathbf{X}))=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

因此可以将 $\delta_k^{(2)}$ 作为 $\delta_k^{(1)}$ 的一个近似。其物理意义为：更新 w_k 的值（ $w_k\leftarrow w_k+\delta_k$ ）时，选择一个较小的更新幅度。

3. 对于方程 $\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k S)=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$ ，其解为：

$$\delta_k=\frac{1}{S}\log\frac{\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)}{\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})}=\frac{1}{S}\log\frac{\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)}{\mathbb{E}_{P(\mathbf{X},\mathbf{Y})}(f_k)}$$

其中 $\mathbb{E}_{P(\mathbf{X},\mathbf{Y})}(f_k)=\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})$ 。

4. `CRF` 学习的算法 `S` ：
   - 输入：
     - 特征函数 f_k
     - 经验分布 $\tilde{P}(\mathbf{X},\mathbf{Y})$
     - 经验分布 $\tilde{P}(\mathbf{X})$
   - 输出：
     - 参数估计值 $\hat{\vec{\mathbf{w}}}$
     - 模型 $P_{\hat{\vec{\mathbf{w}}}}(\mathbf{Y}\mid\mathbf{X})$
   - 算法步骤：
     - 初始化：对所有的 $k=1,2,\cdots,K$ ，取值 $w_k=0$ 。
     - 迭代，迭代的停止条件是：所有 w_k 都收敛。迭代步骤为：
       - 对每一个 $k=1,2,\cdots,K$ ，计算 δ_k ：

         $$\delta_k=\frac{1}{S}\log\frac{\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)}{\mathbb{E}_{P(\mathbf{X},\mathbf{Y})}(f_k)}$$

         其中 $\mathbb{E}_{P(\mathbf{X},\mathbf{Y})}(f_k)=\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})$ 。
       - 更新 w_k 的值： $w_k\leftarrow w_k+\delta_k$ 。如果不是所有 w_k 都收敛，继续迭代。
     - 迭代终止时， $\hat{\vec{\mathbf{w}}}=(w_1,w_2,\cdots,w_K)^T$ 。

4.3.1.2 算法 T

1. 在算法 `S` 中，通常需要使常数 S 取得足够大，此时每一步迭代的增量向量会较小，算法收敛会变慢。

算法 T 试图解决这个问题。

2. 算法 T 对每个观测序列 $\tilde{\mathbf{X}}_i=\{\tilde{\mathbf{x}}_{i,1},\tilde{\mathbf{x}}_{i,2},\cdots,\tilde{\mathbf{x}}_{i,n_i}\},i=1,2,\cdots,N$，计算其特征总数最大值 $f^o(\tilde{\mathbf{X}}_i)$：

$$f^o(\tilde{\mathbf{X}}_i)=\max_{\mathbf{Y}}f^o(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}}_i)=\max_{\mathbf{Y}}\sum_{k=1}^{K}\sum_{l=0}^{n_i}f_k(Y_l,Y_{l+1},\mathbf{X}=\tilde{\mathbf{X}}_i,l)$$

记 $T(\mathbf{X})=\max_{\mathbf{Y}}f^o(\mathbf{Y},\mathbf{X})$，由于 $T(\mathbf{X})\geq f^o(\mathbf{Y},\mathbf{X})$，则对于下面两个方程的解：

$$\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k^{(1)}f^o(\mathbf{Y},\mathbf{X}))=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

$$\sum_{\mathbf{X},\mathbf{Y}}\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_k^{(2)}T(\mathbf{X}))=\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)$$

必然有：$\delta_k^{(2)}\leq\delta_k^{(1)}$，原因与算法 S 相同。

因此可以将 $\delta_k^{(2)}$ 作为 $\delta_k^{(1)}$ 的一个近似。其物理意义为：更新 w_k 的值（ $w_k\leftarrow w_k+\delta_k$ ）时，选择一个较小的更新幅度。

由于 $T(\mathbf{X})\leq S$，因此算法 T 求解的 δ 会相对于算法 S 更大，使得算法收敛的更快。

3. 对于方程 $\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)=\sum_{\mathbf{X},\mathbf{Y}}\left(\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})\exp(\delta_kT(\mathbf{X}))\right)$，有：

$$\begin{aligned}\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)&=\sum_{\mathbf{X},\mathbf{Y}}\left(\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})\sum_{i=0}^{n}f_k(Y_i,Y_{i+1},\mathbf{X},i)\exp(\delta_kT(\mathbf{X}))\right)\\&=\sum_{\mathbf{X}}\left[\tilde{P}(\mathbf{X})\exp(\delta_kT(\mathbf{X}))\sum_{\mathbf{Y}}\sum_{i=0}^{n}(P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(Y_i,Y_{i+1},\mathbf{X},i))\right]\end{aligned}$$

令：$a_{k,\mathbf{X}}=\sum_{\mathbf{Y}}P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})=\sum_{\mathbf{Y}}\sum_{i=0}^{n}(P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(Y_i,Y_{i+1},\mathbf{X},i))$，则有：

$$\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)=\sum_{\mathbf{X}}\tilde{P}(\mathbf{X})\exp(\delta_kT(\mathbf{X}))a_{k,\mathbf{X}}$$

它就是一个以 δ_k 为变量的非线性方程，求解该方程即可得到 δ_k 的解。

4. CRF 学习的算法 T：
   - 输入：
     - 特征函数 f_k
     - 经验分布 $\tilde{P}(\mathbf{X},\mathbf{Y})$
     - 经验分布 $\tilde{P}(\mathbf{X})$
   - 输出：
     - 参数估计值 $\hat{\vec{\mathbf{w}}}$
     - 模型 $P_{\hat{\vec{\mathbf{w}}}}(\mathbf{Y}\mid\mathbf{X})$
   - 算法步骤：
     - 初始化：对所有的 $k=1,2,\cdots,K$，取值 $w_k=0$。
     - 迭代，迭代的停止条件是：所有 w_k 都收敛。迭代步骤为：
       - 对每一个 $k=1,2,\cdots,K$，从方程中计算出 δ_k：

$$\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k)=\sum_{\mathbf{X}}\tilde{P}(\mathbf{X})\exp(\delta_kT(\mathbf{X}))a_{k,\mathbf{X}}$$

       - 更新 w_k 的值：$w_k\leftarrow w_k+\delta_k$。如果不是所有 w_k 都收敛，继续迭代。

- 迭代终止时，$\hat{\vec{\mathbf{w}}}=(w_1,w_2,\cdots,w_K)^T$ 。

4.3.2 拟牛顿法

1. 条件随机场的对数似然函数为：

$$L_{\vec{\mathbf{w}}}=\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)\right]-\sum_{\mathbf{X}}\left[\tilde{P}(\mathbf{X})\log Z_{\vec{\mathbf{w}}}(\mathbf{X})\right]$$

其中：$Z_{\vec{\mathbf{w}}}=\sum_{\mathbf{Y}}\exp\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)$ 。

学习的优化目标是最大化对数似然函数 $L_{\vec{\mathbf{w}}}$ 。

令：

$$F(\vec{\mathbf{w}})=-L_{\vec{\mathbf{w}}}$$
$$=\sum_{\mathbf{X}}\left[\tilde{P}(\mathbf{X})\log\sum_{\mathbf{Y}}\exp\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)\right]-\sum_{\mathbf{X},\mathbf{Y}}\left[\tilde{P}(\mathbf{X},\mathbf{Y})\left(\sum_{k=1}^{K}w_kf_k(\mathbf{Y},\mathbf{X})\right)\right]$$

于是学习优化目标是最小化 F。

2. 计算梯度:

$$\vec{g}(\vec{\mathbf{w}})=(\frac{\partial F(\vec{\mathbf{w}})}{\partial w_1},\frac{\partial F(\vec{\mathbf{w}})}{\partial w_2},\cdots,\frac{\partial F(\vec{\mathbf{w}})}{\partial w_K})^T,$$
$$\frac{\partial F(\vec{\mathbf{w}})}{\partial w_k}=\sum_{\mathbf{X},\mathbf{Y}}[\tilde{P}(\mathbf{X})P_{\vec{\mathbf{w}}}(\mathbf{Y}\mid\mathbf{X})f_k(\mathbf{Y},\mathbf{X})]-\mathbb{E}_{\tilde{P}(\mathbf{X},\mathbf{Y})}(f_k),\quad k=1,2,\cdots,K$$

3. `CRF` 学习的 `BFGS` 算法：
   - 输入：
     - 特征函数 $f_1,f_2,\cdots,f_K$
     - 经验分布 $\tilde{P}(\mathbf{X},\mathbf{Y}),\tilde{P}(\mathbf{X})$
     - 目标函数 $F(\vec{\mathbf{w}})$
     - 梯度 $\vec{g}(\vec{\mathbf{w}})=\nabla F(\vec{\mathbf{w}})$
     - 精度要求 ε
   - 输出：
     - 最优参数值 $\vec{\mathbf{w}}^*$
     - 最优模型 $P_{\vec{\mathbf{w}}^*}(\mathbf{Y}\mid\mathbf{X})$
   - 算法步骤：
     - 选定初始点 $\vec{\mathbf{w}}^{<0>}$，取 $\mathbf{B}_0$ 为正定对阵矩阵，置 $k=0$
     - 计算 $\vec{\mathbf{g}}_k=\vec{g}(\vec{\mathbf{w}}^{<k>})$:
       - 若 $|\vec{\mathbf{g}}_k|<\varepsilon$，停止计算，得到 $\vec{\mathbf{w}}^*=\vec{\mathbf{w}}^{<k>}$
       - 若 $|\vec{\mathbf{g}}_k|\geq\varepsilon$:
         - 由 $\mathbf{B}_k\vec{\mathbf{p}}_k=-\vec{\mathbf{g}}_k$ 求得 $\vec{\mathbf{p}}_k$
         - 一维搜索：求出 λ_k ： $\lambda_k=\min_{\lambda\geq0}F(\vec{\mathbf{w}}^{<k>}+\lambda\vec{\mathbf{p}}_k)$
         - 置 $\vec{\mathbf{w}}^{<k+1>}=\vec{\mathbf{w}}^{<k>}+\lambda_k\vec{\mathbf{p}}_k$
         - 计算 $\vec{\mathbf{g}}_{k+1}=\vec{g}(\vec{\mathbf{w}}^{<k+1>})$。若 $|\vec{\mathbf{g}}_{k+1}|<\varepsilon$，停止计算，得到 $\vec{\mathbf{w}}^*=\vec{\mathbf{w}}^{<k+1>}$

- 否则计算 $\mathbf{B}_{k+1}$：

$$\mathbf{B}_{k+1}=\mathbf{B}_k+\frac{\vec{\mathbf{y}}_k\vec{\mathbf{y}}_k^T}{\vec{\mathbf{y}}_k^T\vec{\delta}_k}-\frac{\mathbf{B}_k\vec{\delta}_k\vec{\delta}_k^T\mathbf{B}_k}{\vec{\delta}_k^TB_k\vec{\delta}_k}$$

其中：$\vec{\mathbf{y}}_k=\vec{\mathbf{g}}_{k+1}-\vec{\mathbf{g}}_k,\quad \vec{\delta}_k=\vec{\mathbf{w}}^{<k+1>}-\vec{\mathbf{w}}^{<k>}$

- 置 $k=k+1$，继续迭代。

4.4 预测算法

1. 给定条件随机场 $P(\mathbf{Y}\mid\mathbf{X})$ 以及输入序列（观测序列） $\tilde{\mathbf{X}}=\{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\cdots,\tilde{\mathbf{x}}_n\}$ 的情况下，求条件概率最大的输出序列（标记序列） $\mathbf{Y}^*=\{\tilde{\mathbf{y}}_1^*,\tilde{\mathbf{y}}_2^*,\cdots,\tilde{\mathbf{y}}_n^*\}$，其中 $\tilde{\mathbf{y}}_i^*\in\mathcal{Y}=\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_m\}$ 。
2. 根据条件随机场的简化形式：

$$P(\mathbf{Y}\mid\mathbf{X})=\frac{1}{Z}\exp\left(\sum_{k=1}^K w_kf_k(\mathbf{Y},\mathbf{X})\right)=\frac{1}{Z}\exp\Big(\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X})\Big)$$

其中 $Z=\sum_{\mathbf{Y}}\exp\Big(\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X})\Big)$ 。

于是：

$$\mathbf{Y}^*=\arg\max_{\mathbf{Y}}P(\mathbf{Y}\mid\mathbf{X}=\tilde{\mathbf{X}})=\arg\max_{\mathbf{Y}}\frac{1}{Z}\exp\Big(\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}})\Big)$$

考虑到 Z 与 $\mathbf{Y}$ 无关，于是：

$$\mathbf{Y}^*=\arg\max_{\mathbf{Y}}\exp\Big(\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}})\Big)=\arg\max_{\mathbf{Y}}\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}})$$

于是：条件随机场的预测问题就成为求非规范化概率最大的最优路径问题。

其中：

$$\begin{aligned}\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}})&=\sum_{k=1}^K w_kf_k(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}})\\&=\sum_{k=1}^K w_k\sum_{i=0}^n f_k(Y_i,Y_{i+1},\mathbf{X}=\tilde{\mathbf{X}},i)\end{aligned}$$

其中就是非规范化概率。

3. 定义局部特征向量：

$$\vec{\mathbf{F}}_i(Y_i,Y_{i+1},\mathbf{X}=\tilde{\mathbf{X}})=(f_1(Y_i,Y_{i+1},\mathbf{X}=\tilde{\mathbf{X}},i),\cdots,f_K(Y_i,Y_{i+1},\mathbf{X}=\tilde{\mathbf{X}},i))^T$$

其物理意义为：每个特征函数在 $\mathbf{X}=\tilde{\mathbf{X}}$ 的条件下，在位置 i 处的取值组成的向量。

于是：$\mathbf{Y}^*=\arg\max_{\mathbf{Y}}\sum_{i=0}^n\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}_i(Y_i,Y_{i+1},\mathbf{X}=\tilde{\mathbf{X}})$ 。

为了便于统一描述，这里引入了两个人工标记：$\mathbf{y}_0=start,\mathbf{y}_{n+1}=stop$ 。它们具有唯一的、固定的取值（不是随机变量）。

4. 维特比算法用动态规划来求解条件随机场的预测问题。它用动态规划求解概率最大路径（最优路径），这时一条路径对应着一个标记序列。
   - 根据动态规划原理，最优路径具有这样的特性：如果最优路径在位置 i 通过结点 $\tilde{\mathbf{y}}_i^*$， 则这一路径从结点 $\tilde{\mathbf{y}}_i^*$ 到终点 $\tilde{\mathbf{y}}_n^*$ 的部分路径，对于从 $\tilde{\mathbf{y}}_i^*$ 到 $\tilde{\mathbf{y}}_n^*$ 的所有可能路径来说，也必须是最优的。

- 只需要从位置 $i=1$ 开始，递推地计算从位置 1 到位置 i 且位置 i 的标记为 $\tilde{\mathbf{y}}_i^*$ 的各条部分路径的最大非规范化概率。

  于是在位置 $i=n$ 的最大非规范化概率即为最优路径的非规范化概率 P^* ，最优路径的终结点 $\tilde{\mathbf{y}}_n^*$ 也同时得到。

- 之后为了找出最优路径的各个结点，从终结点 $\tilde{\mathbf{y}}_n^*$ 开始，由后向前逐步求得结点 $\tilde{\mathbf{y}}_{n-1}^*,\cdots,\tilde{\mathbf{y}}_1^*$，得到最优路径 $\mathbf{Y}^*=(\tilde{\mathbf{y}}_1^*,\tilde{\mathbf{y}}_2^*,\cdots,\tilde{\mathbf{y}}_n^*)$。

5. 假设标记 Y 的取值集合为 $\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_m\}$， 其中 m 是标记的取值个数。

   - 首先求出位置 `1` 的各个标记值的非规范化概率：

   $$\delta_1(j)=\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}_0(Y_0=start,Y_1=\mathbf{y}_j,\mathbf{X}=\tilde{\mathbf{X}}),j=1,2,\cdots,m$$

   - 根据递推公式，求出到位置 i 的各个标记取值的非规范化概率的最大值，同时记录非规范化概率最大值的路径：

   $$\delta_i(l)=\max_{1\le j\le m}\{\delta_{i-1}(j)+\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}_i(Y_i=\mathbf{y}_j,Y_{i+1}=\mathbf{y}_l,\mathbf{X}=\tilde{\mathbf{X}})\}$$

   $$\Psi_i(l)=\arg\max_{1\le j\le m}\{\delta_{i-1}(j)+\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}_i(Y_i=\mathbf{y}_j,Y_{i+1}=\mathbf{y}_l,\mathbf{X}=\tilde{\mathbf{X}})\}$$

   $$l=1,2,\cdots,m;\quad i=1,2,\cdots,n$$

     - 其中 $\delta_i(l)$ 为：到位置 i 、且标记取值为 $\mathbf{y}_l$ 的非规范化概率的最大值。
     - $\Psi_i(l)$ 为：到位置 i 且标记取值为 $\mathbf{y}_l$ 的最优路径中，位置 $i-1$ 处的标记的取值的编号。

   - 到 $i=n$ 时，这时求得非规范化概率的最大值，以及最优路径的终点：

   $$\max_{\mathbf{Y}}\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}})=\max_{1\le j\le m}\delta_n(j),\quad \text{node}_n=\arg\max_{1\le j\le m}\delta_n(j),\quad \tilde{\mathbf{y}}_n^*=\mathbf{y}_{\text{node}_n}$$

   - 根据最优路径得到：$\text{node}_i=\Psi_{i+1}(\text{node}_{i+1}),i=n-1,n-2,\cdots,1$ 。

     最终得到最优路径：$\mathbf{Y}^*=\{\mathbf{y}_{\text{node}_1},\mathbf{y}_{\text{node}_2},\cdots,\mathbf{y}_{\text{node}_n}\}$ 。

6. `CRF` 预测的维特比算法：

   - 输入：
     - 模型特征向量 $\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X})$ ，其中标记 Y 的取值集合为 $\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_m\}$
     - 权值向量 $\vec{\mathbf{w}}$
     - 观测序列 $\tilde{\mathbf{X}}=\{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\cdots,\tilde{\mathbf{x}}_n\}$
   - 输出： 最优路径 $\mathbf{Y}^*=\{\tilde{\mathbf{y}}_1^*,\tilde{\mathbf{y}}_2^*,\cdots,\tilde{\mathbf{y}}_n^*\}$
   - 算法步骤：
     - 初始化：$\delta_1(j)=\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}_0(Y_0=start,Y_1=\mathbf{y}_j,\mathbf{X}=\tilde{\mathbf{X}}),j=1,2,\cdots,m$ 。
     - 递推：对 $i=1,2,\cdots,n$ ：

     $$\delta_i(l)=\max_{1\le j\le m}\{\delta_{i-1}(j)+\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}_i(Y_i=\mathbf{y}_j,Y_{i+1}=\mathbf{y}_l,\mathbf{X}=\tilde{\mathbf{X}})\}$$

     $$\Psi_i(l)=\arg\max_{1\le j\le m}\{\delta_{i-1}(j)+\vec{\mathbf{w}}\cdot\vec{\mathbf{F}}_i(Y_i=\mathbf{y}_j,Y_{i+1}=\mathbf{y}_l,\mathbf{X}=\tilde{\mathbf{X}})\}$$

     $$l=1,2,\cdots,m;\quad i=1,2,\cdots,n$$

     - 终止：

$$\max_{\mathbf{Y}} \vec{\mathbf{w}}\cdot\vec{\mathbf{F}}(\mathbf{Y},\mathbf{X}=\tilde{\mathbf{X}})=\max_{1\le j\le m}\delta_n(j),\quad \text{node}_n=\arg\max_{1\le j\le m}\delta_n(j),\quad \tilde{\mathbf{y}}_n^*=\mathbf{y}_{\text{node}_n}$$

- 回溯路径：$\text{node}_i=\Psi_{i+1}(\text{node}_{i+1}),i=n-1,n-2,\cdots,1$ 。
- 返回路径：$\mathbf{Y}^*=\{\mathbf{y}_{\text{node}_1},\mathbf{y}_{\text{node}_2},\cdots,\mathbf{y}_{\text{node}_n}\}$ 。

4.5 词性标注任务

1. 在词性标注任务中，给定单词序列 $\tilde{\mathbf{X}}=\{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\cdots,\tilde{\mathbf{x}}_n\}$，需要给出每个单词对应的词性 $\tilde{\mathbf{Y}}=\{\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,\cdots,\tilde{\mathbf{y}}_n\}$ 。如：`他们 吃 苹果` 对应的标注序列为 `代词 动词 名词` 。
   - 给定一个单词序列，可选的标注序列有很多种，需要从中挑选出一个最靠谱的作为标注结果。
   - 标注序列是否靠谱，是通过利用特征函数来对序列打分来获得。序列的得分越高（意味着非规范化概率越大），则越靠谱。
2. `CRF` 中的特征函数接受四个参数：
   - 单词序列 $\mathbf{X}$ 。
   - 位置 i ，表示序列 $\mathbf{X}$ 中第 i 个单词。
   - 标记 Y_i ，表示第 i 个单词的标注词性。
   - 标记 Y_{i-1} ，表示第 $i-1$ 个单词的标注词性。

   特征函数的输出值为 $\{0,1\}$，表示满足/不满足这个特征。
3. 每个特征函数对应一个权重 w 。
   - 若权重为正，则说明该特征贡献一个正的分数；如果权重为负，则说明该特征贡献一个负的分数。
   - 权重越大，则说明该特征靠谱的可能性越大（对于正的权重），或者该特征不靠谱的程度越大（对于负的权重）。
   - 该权重是模型的参数，从训练集中训练得到。

4.6 CRF 与 HMM 模型

1. 设已知隐马尔科夫模型的参数，则给定观察序列 $\mathbf{O}=\{o_1,\cdots,o_T\}$，以及标记序列 $\mathbf{I}=\{i_1,\cdots,i_T\}$，其出现的概率为：

$$P(\mathbf{I},\mathbf{O})=P(i_1)\times\left(\prod_{t=1}^{T-1}P(i_{t+1}\mid i_t)\right)\times\prod_{t=1}^{T}P(o_t\mid fi_t)$$

$$\rightarrow\log P(\mathbf{I},\mathbf{O})=\log P(i_1)+\sum_{t=1}^{T-1}\log P(i_{t+1}\mid i_t)+\sum_{t=1}^{T}\log P(o_t\mid i_t)$$

   - 对于 `HMM` 中的每一个转移概率 $P(i_{t+1}=j\mid i_t=i)$，定义特征函数：

$$s_{i,j}(\mathbf{O},t,i_t,i_{t+1})=\begin{cases}1, & i_{t+1}=j \text{ and } i_t=i\\ 0, & \text{else}\end{cases}$$

     定义其权重为：$w_{i,j}^{(s)}=\log P(i_{t+1}=j\mid i_t=i)$ 。
   - 对于 `HMM` 中每一个发射概率 $P(o_t=k\mid i_t=j)$，定义特征函数：

$$t_{j,k}(\mathbf{O},t,i_t)=\begin{cases}1, & i_t=j \text{ and } o_t=k\\ 0, & \text{else}\end{cases}$$

     定义其权重为：$w_{j,k}^{(t)}=\log P(o_t=k\mid i_t=j)$ 。
   - 则有：

$$\begin{aligned}\log P(\mathbf{I},\mathbf{O}) &= \log P(i_1)+\sum_{t=1}^{T-1}\log P(i_{t+1}\mid i_t)+\sum_{t=1}^{T}\log P(o_t\mid i_t)\\ &=\sum_{t=0}^{T-1}\log P(i_{t+1}\mid i_t)+\sum_{t=1}^{T}\log P(o_t\mid i_t)\\ &=\sum_{t=0}^{T-1}w_{i,j}^{(s)}s_{i,j}(\mathbf{O},t,i_t,i_{t+1})+\sum_{t=1}^{T}w_{j,k}^{(t)}t_{j,k}(\mathbf{O},t,i_t)\end{aligned}$$

其中 i_0 为人工引入的 $start$ 结点。

因此 $\log P(\mathbf{I},\mathbf{O})$ 的物理意义为：所有特征函数在所有位置之和。将其归一化之后就得到了 `CRF` 的公式。

2. 从推导中可以看出：每一个 `HMM` 模型都等价于某个 `CRF` 模型。
   - `CRF` 模型比 `HMM` 模型更强大。

     因为 `CRF` 模型能定义数量更多、更丰富多样的特征函数，能使用任意形式的权重。
   - `HMM` 模型中当前的标注结果只依赖于当前的单词，以及前一个标注结果。
   - `HMM` 模型转换过来的形式中，权重是条件概率的对数，因此权重都是负的。
3. `HMM` 是生成式模型，而 `CRF` 是判别式模型
   - 生成式模型对联合概率建模，可以生成样本。

     生成式模型定义的是联合概率，必须列举所有观察序列的可能值。在很多场景下，这是很困难的。
   - 判别式模型对条件概率建模，无法生成样本，只能判断分类。