第11章 条件随机场

vs 第10章

vs 第6章（对数线性模型 —— 改进迭代 / 拟牛顿

11.1 概率无向图模型

概率图模型：

有向图（贝叶斯网络）

无向图（马尔科夫随机场）

描述无向图中结点之间依赖/独立关系

马尔科夫性

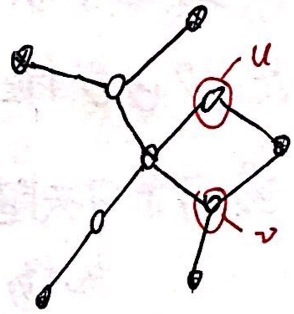
① 成对马尔科夫性

② 局部马尔科夫性

③ 全局马尔科夫性

① 成对马尔科夫性：$P(u, v|O) = P(u|O) \cdot P(v|O)$

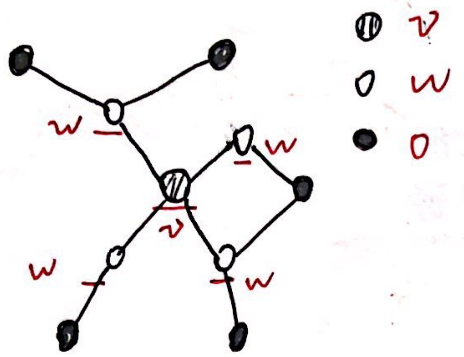
给定O后，u、v互相独立



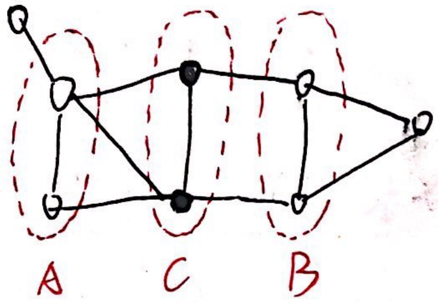
除u、v外其它看成O结点

② 局部马尔科夫性：给定 W，O与V独立

$P(O,V|W)=P(O|W)\cdot P(V|W)$



③ 全局马尔科夫性 给定C后A,B独立 $A \perp\!\!\!\perp B \mid C$



①②③等价关系.

对任一节点来说若满足成对马尔科夫性,则也满足②③

定义11.1

概率无向图模型：设有联合概率分布P(Y)，由无向图G=(V,E)表示，在图G中，结点表示随机变量，边表示随机变量之间的依赖关系。若P(Y)满足①②③中任何一个，就称P(Y)为概率无向图模型，或马尔可夫随机场

概率无向图模型的因子分解

将无向图模型写成 $P(Y)$ 联合概率分布的形式

团的概念：

团内的点，任意两点都有边相联

最大团：再加1个点，就不是团了.(把原来的那个叫为最大团)

定义 11.1

(Hammersley-Clifford定理)

概率无向图模型的联合概率分布 $P(Y)$ 可表示为：

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_C(Y_C)$$

人为给定，需≥0

归一化系数，保证其是个概率分布

$$Z=\sum_{Y}\prod_{C}\Psi_C(Y_C)$$

C 是无向图最大团，Y_C 是 C 的结点对应的随机变量，

$\Psi_C(Y_C)$ 是 C 上~~严格定义的正~~ 定义的严格正函数，乘积是在无向图所有的最大团上进行的

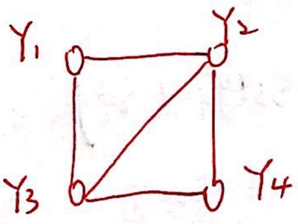
指数保证其≥0

$$\Psi_C(Y_C)=\exp\{-E(Y_C)\}$$

关于 Y_C 的函数（叫能量函数）

也是人为给定

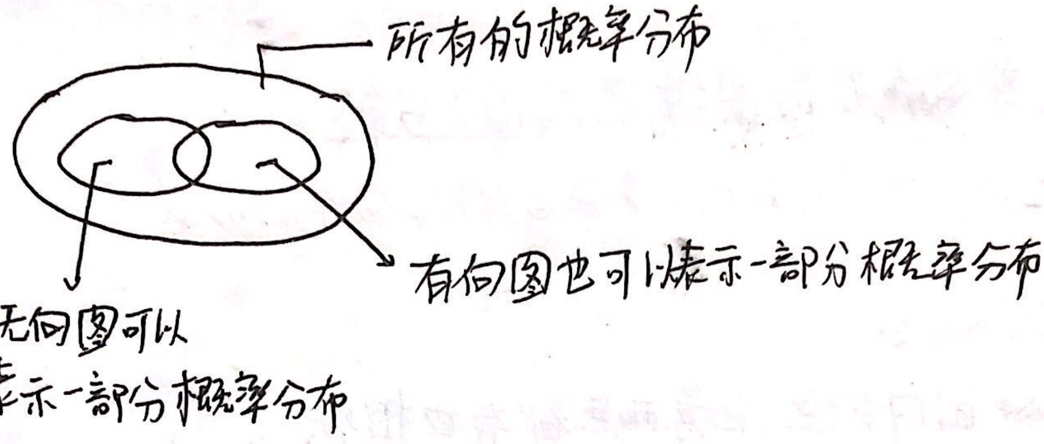
例：



$$P(y_1,y_2,y_3,y_4)=\frac{1}{Z}\Psi_1(y_1,y_2,y_3)\cdot\Psi_2(y_2,y_3,y_4)$$

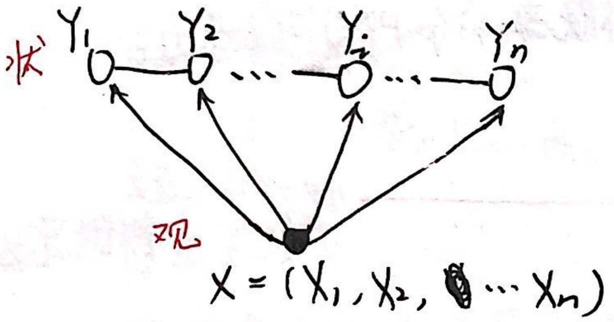
$C_1=\{Y_1,Y_2,Y_3\}$

$C_2=\{Y_2,Y_3,Y_4\}$



11.2 条件随机场的定义与形式

条件随机场



(线性条件随机场)

~ 表示 w 与 v 直接相连的

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

$\left\{ \begin{array}{l} w_1 \sim v \\ w_2 \subset w / w_1 \end{array} \right.$

(局部马尔科夫性)

参数化形式 = 指数相加 最大团乘积

最大团

$Y_1\ Y_2$
$Y_2\ Y_3$
$\vdots$
$Y_{i-1}\ Y_i$

∵有X

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k (\underline{y_{i-1}, y_i}, x, i) + \sum_{i,l} u_l s_l (\underline{y_i}, x, i) \right)$$

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

t_k：转移特征（有k个. 每个特征函数对应一个最大团.

s_l：状态特征

条件随机场的简化形式：(对 y_{i-1}, y_i 求和已经算过了)

将局部特征函数转化为全局特征函数

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

w_k 对应 (λ_k, u_l)；f_k 对应 (t_k, s_l)

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

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

条件随机场的矩阵形式 (对 k 求和已经算过了)

$$P_w(y|x)=\frac{1}{Z_w(x)}\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x)$$

11.3 条件随机场的概率计算问题

利用条件随机场的矩阵形式，计算 $P(Y=y_i|x)$

前向-后向算法

与第10章区别：
- 10章 求观测出现的概率
- 11章：求状态出现的概率

11.4 条件随机场的学习算法

求对数线性模型参数 w (λ_k, u_l)

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

改进的迭代尺度法
拟牛顿法
> 运用在对数线性模型

11.5 条件随机场的预测算法

— 解决标注问题

$$y^*=\arg\max_y P_w(y|x)$$

维特比算法 (DP)

条件随机场的矩阵形式

——与HMM在处理标注问题上区别与联系

先看参数化形式

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

矩阵形式

$$P_w(y|x)=\frac{1}{Z_w(x)}\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x)$$

$$\exp\Big\{\sum_{i,k}\lambda_k t_k(y_{i-1},y_i)+\sum_{i,l}u_l s_l(y_i)\Big\}$$

将i, k分开

$$=\exp\Big\{\sum_i\Big[\sum_k\lambda_k t_k(y_{i-1},y_i)+\sum_l u_l s_l(y_i)\Big]\Big\}$$

$$=\prod_i\exp\Big(\sum_k\lambda_k t_k+\sum_l u_l s_l\Big)$$

$$\underline{M_i(y_{i-1},y_i|x)}=\exp\Big(\sum_k\lambda_k t_k+\sum u_l s_l\Big)$$

矩阵形式是 $i=1\sim n+1$，但参数形式是 $1\sim n$，差了一项

当 $i=1$ 时　$t_k(y_0,y_1)$

↳ 给其一个初值

写成矩阵

$$M_i(x)=\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{m\times m}$$

列：y_{i-1} 取 第1个状态，第2，… m个状态

↓

m为 y_{i-1}，y_i 可取m个值

行 y_i，取 ——

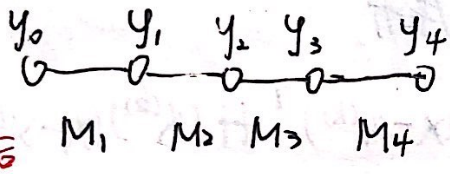
例 11.2

给定图11.6所示线性链条件随机场，观测序列 x，状态序列 y，$i=1,2,3$，$n=3$，标记 $y_i \in \{1,2\}$，假设 $y_0=start=1$，$y_4=stop=1$，各个位置的随机矩阵 $M_1(x)$，$M_2(x)$，$M_3(x)$，$M_4(x)$ 分别是

$$M_1(x)=\begin{bmatrix} a_{01} & a_{02} \\ 0 & 0 \end{bmatrix} \qquad M_2(x)=\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$M_3(x)=\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \qquad M_4(x)=\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

试求状态序列 y 以start为起点、stop为终点所有路径的非规范化概率及规范化因子



M集合相当于HMM的状态转移矩阵

HMM的初始 π 相当于 $M_1(x)$ 即

$P(y_1=1) \propto a_{01}$

$P(y_1=2) \propto a_{02}$

① 路径 $1\to1\to1$

$y_0\to y_1$ M_1 a_{01}

$y_1\to y_2$ M_2 b_{11}

$y_2\to y_3$ M_3 c_{11}

∴为 $a_{01}\cdot b_{11}\cdot c_{11}$

②…⑧ 共8条

∴规范化因子即8条路径概率之和

非规范化概率：求M不考虑规范化因子，即M中每行不为1（概率）

但我们对全体 y_i 看联合概率分布时，要加上 $\frac{1}{Z}$（规范因子）

差别1. 即 局部不规范，全局规范（在马尔科夫中） M

★ 局部规范，全局也规范（在HMM中） A 1

差别2 HMM：状态转移概率A不随位置变化

crf：不对上面这个作假设 ∴更灵活

牛顿法和拟牛顿法

考虑无约束最优化问题

$$\min_{x \in R^n} f(x)$$

假设$f(x)$具有二阶连续偏导数

用迭代法来求x^*

即 $x^{(1)} \to x^{(2)} \to \cdots x^{(k)} \to x^{(k+1)}$ 注$x^{(k)}$表示第k次迭代值为$x^{(k)}$

怎么求?

将$f(x)$在$x^{(k)}$附近进行二阶泰勒展开

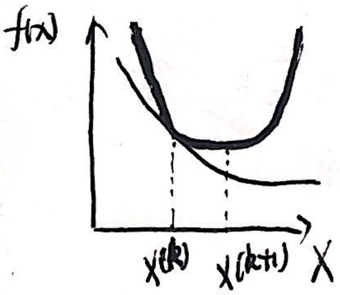
即 $f(x) = f(x^{(k)}) + g_k^T(x-x^{(k)}) + \frac{1}{2}(x-x^{(k)})^T H(x^{(k)})(x-x^{(k)})$

① $g_k = g(x^{(k)}) = \nabla f(x^{(k)})$ 是$f(x)$的梯度向量在$x^{(k)}$的值

② $H(x^{(k)})$是$f(x)$的黑塞矩阵在点$x^{(k)}$的值.

$$H(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{n \times n}$$

当$H(x^{(k)})$是正定矩阵时. $f(x)$的极值为最小值



$$\nabla f(x) = \nabla f(x^{(k)}) + H_k(x-x^{(k)}) = 0$$

$$\Rightarrow x^{(k+1)} = x^{(k)} - H_k^{-1} \nabla f(x^{(k)})$$

$$x^{(k+1)} = x^{(k)} - H_k^{-1} \nabla f(x^{(k)})$$

特点① H_k 正定（拟合的多项式为凸函数）

② 要计算 H_k^{-1}（对不同的 k，H_k 不一样）

∴计算量为 $n \times n$，太大（即牛顿法的缺陷）

$$x^{(k+1)} = x^{(k)} + \lambda P_k$$

在拟牛顿法中 $\lambda = -H_k^{-1} \nabla f(x^{(k)})$（用了二阶导 ∴收敛速度快）

此梯度下降法中：

$x^{(k+1)} = x^{(k)} - \lambda \nabla f(x^{(k)})$（只用到了一阶导）

---

① $f(x)$ x^k 二阶导

② H_k^{-1}（找一个矩阵代替 H_k）（使其求逆比较好求）

拟牛顿法

∴要找近似矩阵，得满足2个条件

① 正定

② $\nabla f(x^{(k+1)}) = \nabla f(x^{(k)}) + H_k (x - x^{(k)})$

$$H_k \delta_k = y_k \Rightarrow \delta_k = H_k^{-1} y_k$$

δ_k: $(x^{(k+1)} - x^{(k)})$　y_k: $\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$

不能直接用∵得知道 $x^{(k+1)}$ 才可 ⇒ δ_k，但我们要求的就是 $x^{(k+1)}$

∴ 将 $H_k \to H_{k+1}$

$H_{k+1} \to x^{k+2}$

2个算法

① DFP $\quad G_k \xrightarrow{替代} H_k^{-1}$

② BFGS $\quad B_k \xrightarrow{替代} H_k$

①

$$G_{k+1} = G_k + \frac{\delta_k \delta_k^T}{\delta_k^T y_k} - \frac{G_k y_k y_k^T G_k}{y_k^T G_k y_k}$$

假设

$G_{k+1} = G_k + \underline{a v} v^T + \underline{b u} u^T$ 去找满足条件的 av, bu

$\Rightarrow H_{k+1}^{-1}$ 替代

∴ $G_k \nabla f(x^{(k)})$ 要比直接用 H_k^{-1} 简单