Conditional Random Field

条件随机场

max margin classifier

$\begin{cases} \min \frac{1}{2} w^T w \\ s.t.\ y_i(w^T x_i + b) \geq 1 \end{cases}$

SVM } 硬 — 分类 — 软

PLA

LDA — 线性判别分析

… — 感知机

$f(w) = \mathrm{sign}(w^T x)$

softmax

Logistic Regression ⟹ Maximum Entropy Model

给定均值和方差, Gaussian Dist 熵最大

概率判别模型

$P(y|x)$ 建模

$P(x,y)$ 建模

Native Bayes $\xrightarrow{seq}$ Hidden Markov Model

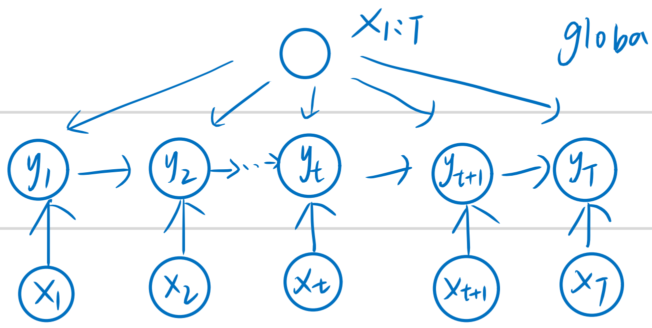
⇑ time

Gaussian Mixture Model

概率生成模型

朴素贝叶斯假设

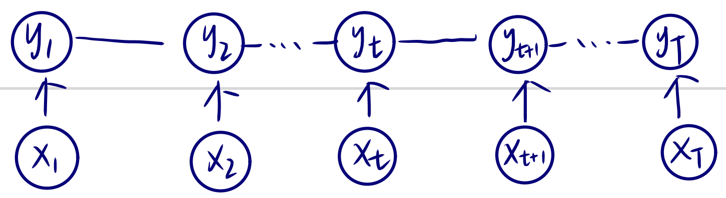
Maximum Entropy Model, Markov → MEMM (判别) → CRF



打破了HMM的观测独立假设

Label bias problem (原因: 局部归一化)

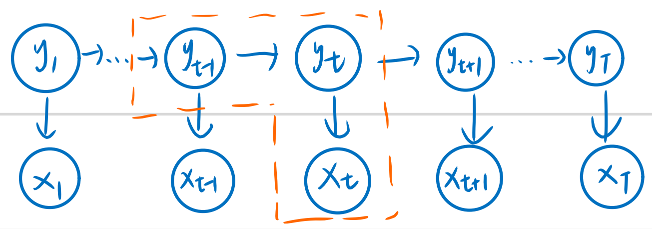
CRF: 判别, $P(Y|X)$

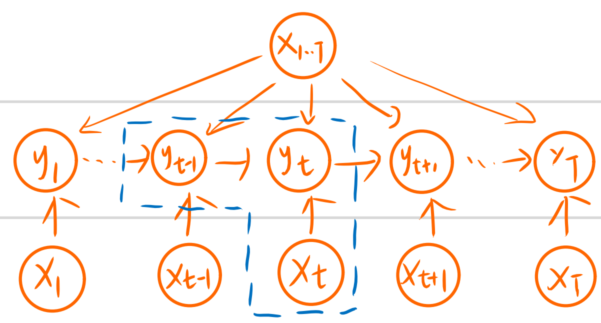


→ 全局归一化

$P(Y|X) = \frac{1}{Z} \sum_{i=1}^{K} \phi_{c_i}(x_{c_i})$

HMM vs MEMM





生成模型 (Naive Bayes, GMM)

{ 齐次一阶 Markov
{ 观测独立假设

两个假设都较折中

$P(y_t | y_{1:t-1}, x_{1:t-1}) = P(y_t | y_{t-1})$

$P(x_t | y_{1:t}, x_{1:t-1}) = P(x_t | y_t)$

建模对象: $P(X, Y | \lambda)$

$P(X, Y | \lambda) = \prod_{t=1}^{T} P(x_t, y_t, y_{t-1} | \lambda)$

$= \prod_{t=1}^{T} P(y_t | y_{t-1}, \lambda) P(x_t | y_t, \lambda)$

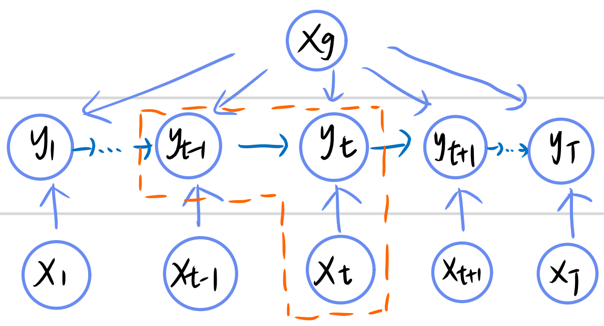
建模对象: $P(Y | X, \lambda)$

$P(Y | X, \lambda) = \prod_{t=1}^{T} P(y_t | y_{t-1}, X, \lambda)$ $\quad X = x_{1:T}$

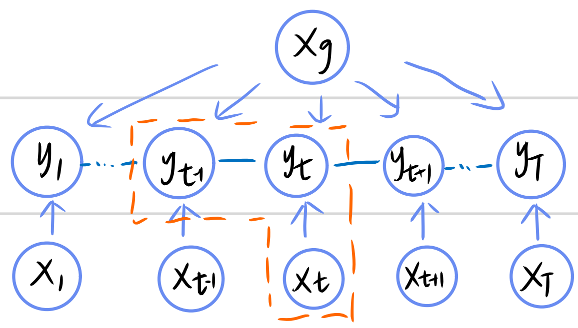
打破了观测独立假设 (更加合理)

联合概率 → 生成模型 条件概率 → 判别模型

MEMM vs CRF



有向 → 无向



mass score 归一化

优点（和HMM比较）

chain-structured CRF

- 判别式模型
- 打破观测独立性

克服了 Label Bias Problem

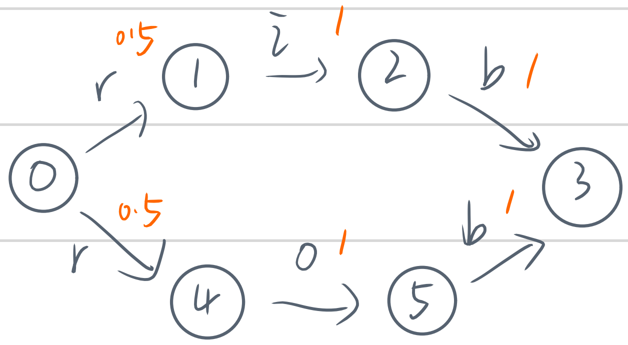
HMM → MEMM → CRF（合理性）

模型：$P(Y|X,\lambda)=\prod_t P(y_t|y_{t-1},x_{1:T},\lambda)$

- 条件 → 判别式模型
- 随机场 → 无向图模型

缺点：Label Bias Problem

MRF 因子分解 $x \in \mathbb{R}^p$



$$P(X)=\frac{1}{Z}\prod_{i=1}^{K}\underbrace{\psi_i(X_{c_i})}_{>0}$$

势函数

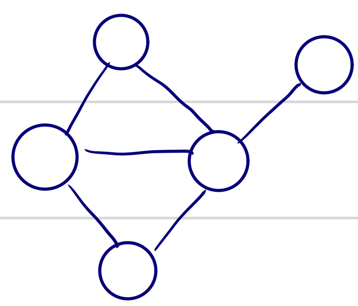
$$=\frac{1}{Z}\prod_{i=1}^{K}\exp[-E_i(X_{c_i})]$$

能量函数

$$=\frac{1}{Z}\exp\sum_{i=1}^{K}F_i(X_{c_i})$$

$P(2|1,i)=1=P(2|1)$

$P(5|4,o)=1=P(5|4)$



K个最大团

C_i：最大团

可以看出从①→②根本没有关注observation

conditional dist with low entropy

take less notice of observation.

$$P(Y|X)=\frac{1}{Z}\exp\sum_{i=1}^{K}F_i(x_{c_i})$$

$$=\frac{1}{Z}\exp\sum_{t=1}^{T}F_t(y_{t-1},y_t,x_{1:T})$$

$\triangle_{y_{t-1},y_t,x_{1:T}}=\sum_{k=1}^{K}\lambda_k f_k(y_{t-1},y_t,x_{1:T})$

f_k：特征函数

$F_t(y_{t-1},y_t,x_{1:T})$

$=\triangle_{y_{t-1},x_{1:T}}+\triangle_{y_t,x_{1:T}}+\triangle_{y_{t-1},y_t,x_{1:T}}$

状态函数　　转移函数

$f_k\{y_{t-1}=名词,y_t=动词,x_{1:T}\}=1$ ⇐

$f_k(y_{t-1}=名词,y_t=动词,x_{1:T}\}=0$

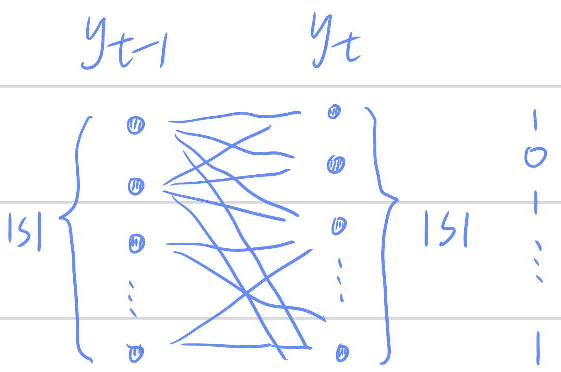
$=\triangle_{y_t,x_{1:T}}+\triangle_{y_{t-1},y_t,x_{1:T}}$

$\triangle_{y_t,x_{1:T}}=\sum_{l=1}^{L}\eta_l g_l(y_t,x_{1:T})$

f_k与g_l是给定的特定函数，λ_k,η_l是参数

$$P(Y|X)=\frac{1}{Z}\exp\sum_{t=1}^{T}\Big[\sum_{k=1}^{K}\lambda_k f_k(y_{t-1},y_t,x_{1:T})+\sum_{l=1}^{L}\eta_l g_l(y_t,x_{1:T})\Big]$$

CRF的条件概率密度函数



$K\le |S|^2$

1或0由λ_k来决定

$$P(Y=y|X=x)=\frac{1}{Z(x,\lambda,\eta)}\exp\sum_{t=1}^{T}\big[\lambda^T f(y_{t-1},y_t,x)+\eta^T g(y_t,x)\big]$$ 向量形式

$$y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_T\end{pmatrix}\quad x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_T\end{pmatrix}\quad \lambda=\begin{pmatrix}\lambda_1\\\lambda_2\\\vdots\\\lambda_K\end{pmatrix}\quad \eta=\begin{pmatrix}\eta_1\\\eta_2\\\vdots\\\eta_L\end{pmatrix}\quad f=\begin{pmatrix}f_1\\f_2\\\vdots\\f_K\end{pmatrix}=f(y_{t-1},y_t,x)\quad g=\begin{pmatrix}g_1\\g_2\\\vdots\\g_L\end{pmatrix}=g(y_t,x)$$

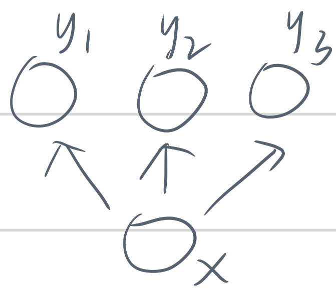
$$=\frac{1}{Z(x,\lambda,\eta)}\exp\Big[\lambda^T\sum_{t=1}^{T}f(y_{t-1},y_t,x)+\eta^T\sum_{t=1}^{T}g(y_t,x)\Big]$$

$$\theta=\begin{pmatrix}\lambda\\\eta\end{pmatrix}_{K+L}\qquad H=\begin{pmatrix}\sum_{t=1}^{T}f\\\sum_{t=1}^{T}g\end{pmatrix}_{K+L}$$

$$P(Y=y|X=x)=\frac{1}{Z(x,\theta)}\exp\,\theta^T\cdot H(y_t,y_{t-1},x)\qquad \langle\theta,H\rangle$$

模型要解决的问题

- learning: parameter estimation — Given training data: $\{(x^{(i)},y^{(i)})\}_{i=1}^{N}$ $x,y\in\mathbb{R}^T$; $\hat{\theta}=\arg\max\prod_{i=1}^{N}P(y^{(i)}|x^{(i)})$
- Inference
  - marginal prob $\Rightarrow P(y_t|X)\rightarrow P(y_t=i|X)$
  - conditional prob 生成模型
  - MAP Inference: decoding $\Rightarrow \hat{y}=\arg\max_{y=y_1,y_2\cdots y_T}P(y|X)$



边缘概率计算

Given $P(Y=y|X=x)$，求 $P(y_t=i|X)$

$$P(y|x)=\frac{1}{Z}\prod_{t=1}^{T}\psi_t(y_{t-1},y_t,x)$$

$y \to y_1,y_2,\cdots y_T$

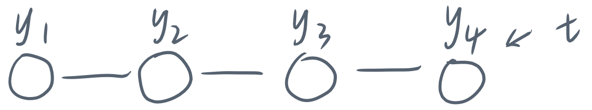
$$P(y_t=i|x)=\sum_{y_1,y_2,\cdots,y_{t-1},y_{t+1},\cdots,y_T}P(y|x)=\sum_{y\langle1,t-1\rangle}\sum_{y\langle t+1,T\rangle}\frac{1}{Z}\prod_{t'=1}^{T}\psi_{t'}(y'_{t-1},y'_t,x)$$

$y_t\in S$

$O(T\cdot|S|^T)=\frac{1}{Z}\Delta_{左}\cdot\Delta_{右}$

$$\Delta_{左}=\sum_{y\langle1:t-1\rangle}\psi_1(y_0,y_1,x)\psi_2(y_1,y_2,x)\cdots\psi_{t-1}(y_{t-2},y_{t-1},x)\cdot\psi_t(y_{t-1},y_t=i,x)$$

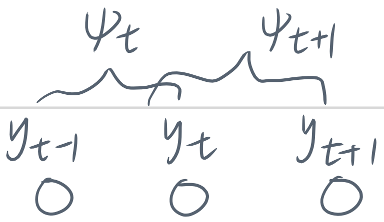
$$\Delta_{右} = \sum_{y<t+1,T>} \psi_{t+1}(y_t=i, y_{t+1}, x) \cdots \psi_T(y_{T-1}, y_T, x)$$



按顺序积分

$$\Delta_{左} = \sum_{y_{t-1}} \psi_t(y_{t-1}, y_t=i, x) \sum_{y_{t-2}} \psi_{t-1}(y_{t-2}, y_{t-1}, x) \cdots \sum_{y_1} \psi_2(y_1, y_2, x) \cdot \sum_{y_0} \psi_1(y_0, y_1, x)$$

$$\Delta_{左} \quad \alpha_t(i): \underbrace{y_0\ y_1\ y_2 \cdots y_{t-1}}_{\text{所有势函数乘积}}\ \underbrace{y_t = i}_{\text{左半部势函数}}$$



$$\alpha_{t-1}(j): \underbrace{y_0\ y_1\ y_2 \cdots y_{t-2}}_{\text{所有势函数}}\ \underbrace{y_{t-1} = j}_{\text{左半部势函数}}$$

右半部势函数

$$\underline{\sum_{y_{t-1}} \psi_t(y_{t-1}=j, y_t=i, x)} \cdot \alpha_{t-1}(j) = \alpha_t(i)$$

$$\sum_{j \in S} \psi_t(y_{t-1}=j, y_t=i, x) \cdot \alpha_{t-1}(j) = \alpha_t(i)$$

$$\begin{aligned} \Delta_{左} &= \alpha_t(i) \\ \Delta_{右} &= \beta_t(i) \end{aligned} \Rightarrow P(y_t=i|X) = \frac{1}{Z}\alpha_t(i)\beta_t(i)$$

参数估计

$$\hat{\theta} = \operatorname{argmax} \prod_{i=1}^{N} P(y^{(i)}|X^{(i)})$$

$$\Downarrow$$

$$\hat{\lambda}, \hat{\eta} = \underset{\lambda,\eta}{\operatorname{argmax}} \prod_{i=1}^{N} P(y^{(i)}|X^{(i)})$$

$$P(y|X) = \frac{1}{Z(x,\lambda,\eta)} \exp \sum_{t=1}^{T} [\lambda^T \cdot f(y_{t-1}, y_t, x) + \eta^T \cdot g(y_t, x)]$$

$$\underset{\lambda,\eta}{\operatorname{argmax}} \log \prod_{i=1}^{N} P(y^{(i)}|X^{(i)}) = \underset{\lambda,\eta}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(y^{(i)}|X^{(i)})$$

$$= \underset{\lambda,\eta}{\operatorname{argmax}} \sum_{i=1}^{N} \left\{ -\log Z(x^{(i)}, \lambda, \eta) + \sum_{t=1}^{T} [\lambda^T \cdot f(y_{t-1}, y_t, x) + \eta^T \cdot g(y_t, x)] \right\}$$

$$\triangleq \underset{\lambda,\eta}{\operatorname{argmax}} L(\lambda, \eta, x^{(i)})$$

方法：梯度上升 $\nabla_\lambda L, \nabla_\eta L$

$$\nabla_\lambda L = \sum_{i=1}^{N} \left[\sum_{t=1}^{T} f(y_{t-1}, y_t, x^{(i)}) - \nabla_\lambda \underbrace{\log Z(x^{(i)}, \lambda, \eta)}_{\text{log-partition function}} \right]$$

$$E\left[\sum_{t=1}^{T} f(y_{t-1}, y_t, x^{(i)})\right]$$

$$= \sum_{y} P(y|X^{(i)}) \cdot \sum_{t=1}^{T} f(y_{t-1}, y_t, X^{(i)})$$

$$= \sum_{t=1}^{T} \left[\sum_{y} P(y|X^{(i)}) \cdot f(y_{t-1}, y_t, X^{(i)}) \right]$$

$$= \sum_{t=1}^{T} \sum_{y_{<1:t-2>}} \sum_{y_{t-1}} \sum_{y_t} \sum_{y_{<t+1,T>}} P(y|X^{(i)}) \cdot f(0)$$

$$= \sum_{t=1}^{T} \sum_{y_{t-1}} \sum_{y_t} \left[\sum_{y_{<1:t-2>}} \sum_{y_{<t+1,T>}} P(y|X^{(i)}) \cdot f(0) \right]$$

$$= \sum_{t=1}^{T} \sum_{y_{t-1}} \sum_{y_t} \underbrace{P(y_{t-1}, y_t, X^{(i)})}_{A(y_{t-1}, y_t)} f(y_{t-1}, y_t, X^{(i)})$$

$$P(y_{t-1}, y_t | X) = \frac{1}{Z} \alpha_{t-1}(i) \psi(y_{t-1}, y_t) \beta_t(j)$$

$$\nabla_\lambda L = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[f(y_{t-1}, y_t, X^{(i)}) - \sum_{y_{t-1}} \sum_{y_t} A(y_{t-1}, y_t) \cdot f(y_{t-1}, y_t, X^{(i)}) \right]$$

梯度上升算法

$$\begin{cases} \lambda^{(t+1)} = \lambda^{(t)} + step \cdot \nabla_\lambda L(\lambda^{(t)}, \eta^{(t)}) \\ \eta^{(t+1)} = \eta^{(t)} + step \cdot \nabla_\eta L(\lambda^{(t)}, \eta^{(t)}) \end{cases}$$