北京航空航天大學
BEIHANG UNIVERSITY

自然语言处理

人工智能研究院

主讲教师 沙磊

Contents

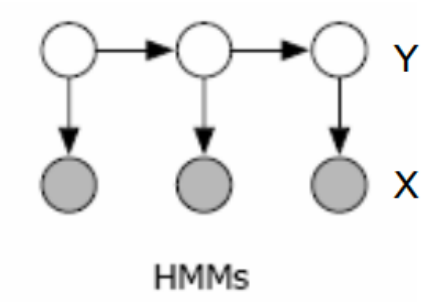
- 条件随机场模型 CRF

CRF

序列模型

- Hidden Markov Model (HMM)

$$p(\mathbf{y},\mathbf{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t)$$

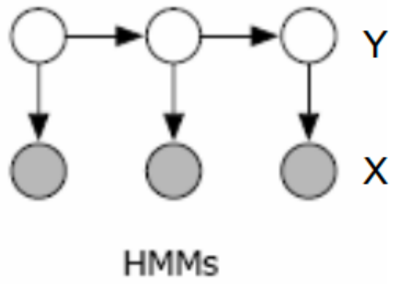


- 独立性假设：
  - 每个状态只直接依赖于其前一个
  - 每个观察变量只依赖于当前状态
- 局限性：
  - 观察变量X之间存在强独立性假设。
  - 建模联合概率p（y，x）引入大量参数，这需要建模分布p（x）

序列模型

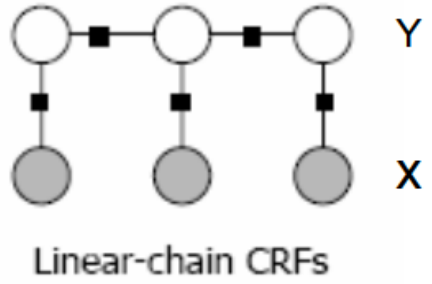
- Hidden Markov Model (HMM)

$$p(\mathbf{y},\mathbf{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t)$$



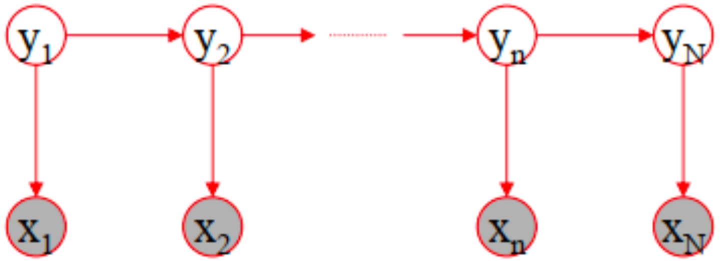
- Conditional Random Fields (CRF)

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$



- 条件随机场（CRF）的一个重要优势是它们具有很大的灵活性，可以包含各种任意的、非独立的观测特征。

Generative Model: HMM



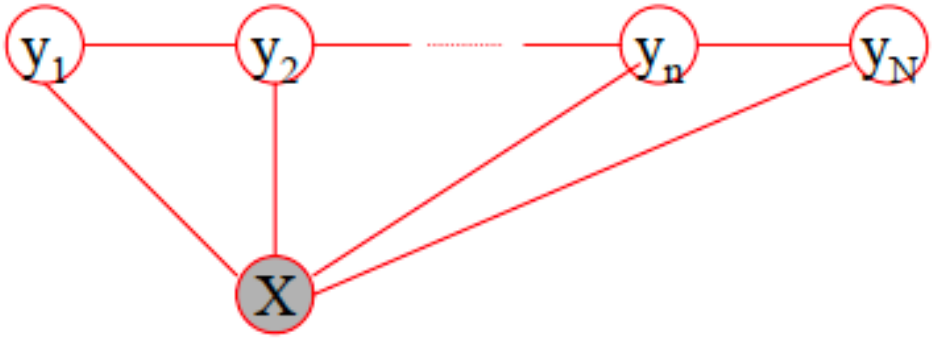
- X is observed data sequence to be labeled,
- Y is the random variable over the label sequences
- HMM is a distribution that models p(Y, X)
- Joint distribution is

$$\boxed{p(\mathbf{Y},\mathbf{X}) = \prod_{n=1}^{N} p(y_n \mid y_{n-1})\, p(\mathbf{x}_n \mid y_n)}$$

- Highly structured network indicates conditional independences,
  - Past states independent of future states
  - Conditional independence of observed given its state.

针对序列模型的判别模型

- CRF模型建模了给定观测值X条件下的条件分布p(Y|X)
- CRF是一个随机场，全局地以观测值X为条件
- 从联合分布p(Y,X)中得出的条件分布p(Y|X)可以被重写为一个马尔可夫随机场。



Markov Random Field (MRF)

- 也称为无向图模型
- 变量集x的联合分布由一个无向图定义

$$p(\mathrm{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathrm{x}_C)$$

- 其中C是最大团（clique）（每个节点与每个其他节点相连）
  - x_C是该团中的变量集，ψ_C是潜在函数potential function（或局部函数(local function)或兼容函数(compatibility function)），满足 $\psi_C(x_C)>0$，通常$\psi_C(x_C)$ = exp{−E(x_C)}，而Z是用于归一化的配分函数。
  - 模型是指一组分布，而场则指一个具体的分布。

$$Z = \sum_{\mathrm{x}}\prod_{C}\psi_C(\mathrm{x}_C)$$

MRF with Input-Output Variables

- X是一组被观测到的输入变量
  - X的元素用x表示
- Y是一组我们要预测的输出变量
  - Y的元素用y表示
- A是$X \cup Y$的子集
  - A中属于$A \cap X$的元素用x_A表示
  - A中属于$A \cap Y$的元素用y_A表示
- 那么无向图模型的形式为

$$\boxed{p(\mathrm{x,y}) = \frac{1}{Z}\prod_{A}\Psi_A(\mathrm{x}_A, \mathrm{y}_A)} \text{ where } \boxed{Z = \sum_{\mathrm{x,y}}\prod_{A}\Psi_A(\mathrm{x}_A, \mathrm{y}_A)}$$

MRF Local Function

- 假设每个局部函数的形式为

$$\Psi_A(x_A, y_A) = \exp\left\{\sum_m \theta_{Am} f_{Am}(x_A, y_A)\right\}$$

- 其中θ_A是一个参数向量， f_A是特征函数，m=1,..M 是特征下标。

From HMM to CRF

- HMM 中

$$p(Y,X)=\prod_{n=1}^{N}p(y_n|y_{n-1})p(x_n|y_n)$$

分布参数：$\theta=\{\lambda_{ij},\mu_{oi}\}$

- 可以被写作：

$$p(Y,X)=\frac{1}{Z}\exp\left\{\sum_{n}\sum_{i,j\in S}\lambda_{ij}\mathbb{I}_{y_n=i}\mathbb{I}_{y_{n-1}=j}+\sum_{n}\sum_{i\in S}\sum_{i\in O}\mu_{oi}\mathbb{I}_{y_n=i}\mathbb{I}_{x_n=o}\right\}$$

- 进一步写作：

$$p(Y,X)=\frac{1}{Z}\exp\left\{\sum_{m=1}^{M}\lambda_m f_m(y_n,y_{n-1},x_n)\right\}$$

特征函数有如下形式：$\mathrm{f}_m(y_n, y_{\mathrm{n-1}}, x_n)$
对每个状态转换 i->j 需要一个特征

$$f_{ij}(y,y',x)=\mathbb{I}(\mathrm{y}=\mathrm{i})\mathbb{I}(\mathrm{y}'=\mathrm{j})$$

对每个状态-观察对也需要一个特征

$$f_{io}(y,y',x)=\mathbb{I}(\mathrm{y}=\mathrm{i})\mathbb{I}(\mathrm{x}=\mathrm{o})$$

From HMM to CRF

$$p(Y, X) = \frac{1}{Z} \exp\left\{ \sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- 进一步

$$p(Y|X) = \frac{p(y, x)}{\sum_{y'} p(y', x)} = \frac{\exp\left\{ \sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}}{\sum_{y'} \exp\left\{ \sum_{m=1}^{M} \lambda_m f_m(y'_n, y'_{n-1}, x_n) \right\}}$$

CRF definition

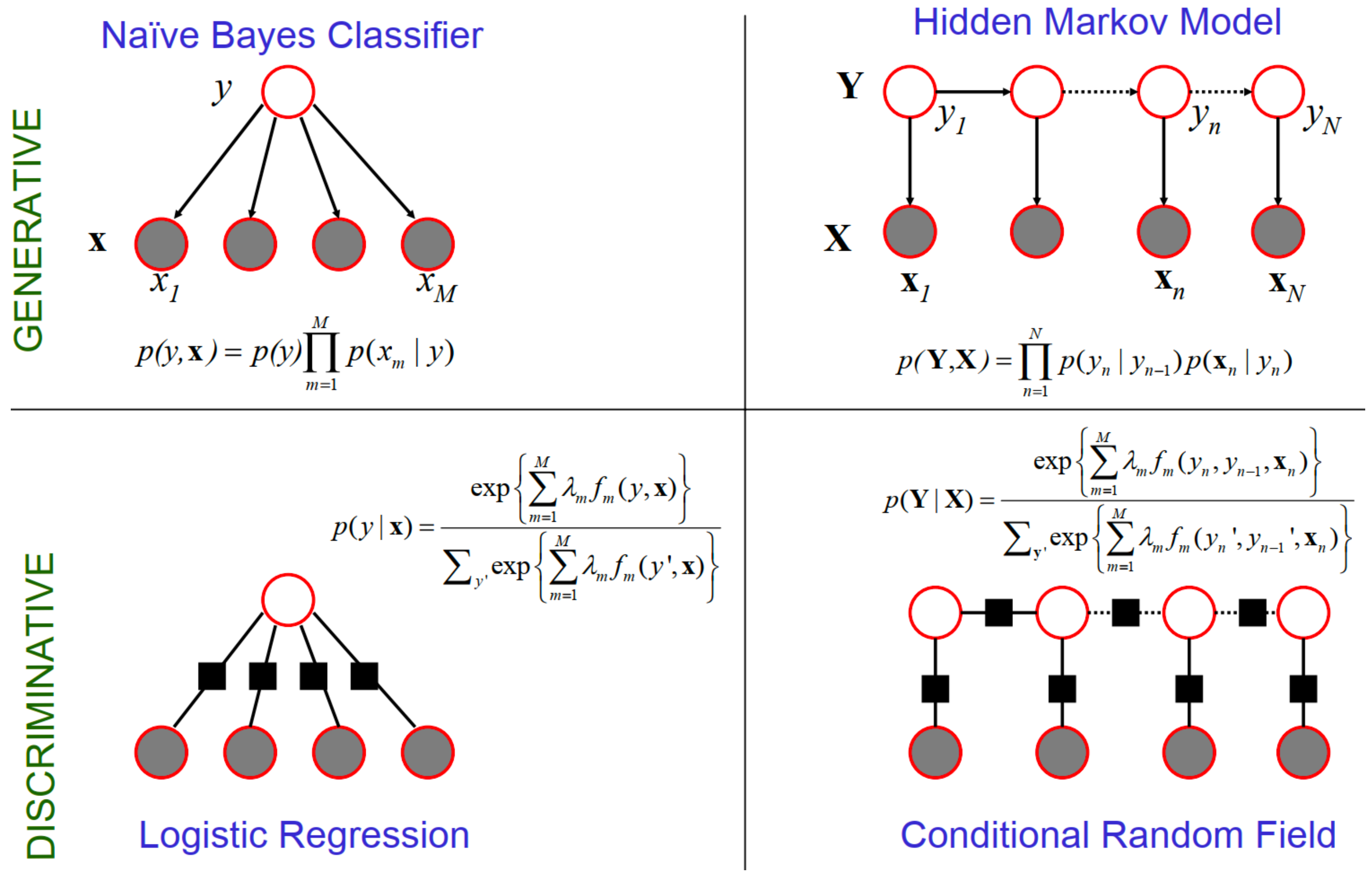
- 线性链CRF定义为分布p($Y|X$) ,其形式如下：

$$p(Y|X) = \frac{1}{Z(X)} \exp\left\{ \sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- 其中Z(X)是一个实例特定的归一化函数。

$$Z(X) = \sum_{y'} \exp\left\{ \sum_{m=1}^{M} \lambda_m f_m(y'_n, y'_{n-1}, x_n) \right\}$$

功能模型


Naïve Bayes Classifier
Hidden Markov Model
GENERATIVE
DISCRIMINATIVE
y
x
x_1
x_M
$p(y,\mathbf{x}) = p(y)\prod_{m=1}^{M} p(x_m \mid y)$
Y
X
y_1
y_n
y_N
$\mathbf{x}_1$
$\mathbf{x}_n$
$\mathbf{x}_N$
$p(\mathbf{Y},\mathbf{X}) = \prod_{n=1}^{N} p(y_n \mid y_{n-1}) p(\mathbf{x}_n \mid y_n)$
$p(y \mid \mathbf{x}) = \frac{\exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y,\mathbf{x})\right\}}{\sum_{y'} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y',\mathbf{x})\right\}}$
$p(\mathbf{Y} \mid \mathbf{X}) = \frac{\exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n)\right\}}{\sum_{\mathbf{y}'} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y_n', y_{n-1}', \mathbf{x}_n)\right\}}$
Logistic Regression
Conditional Random Field

CRF的优势

- CRF放松了对于给定标签的观测数据的条件独立性的假设
- CRF可以包含任意的特征函数
    - 每个特征函数可以使用整个输入数据序列。 观测数据片段的标签概率可能取决于任何过去或未来的数据片段。
- CRF可以避免其他具有偏向后继状态较少的状态的判别性马尔可夫模型的限制

CRF 优化

$$p(Y|X) = \frac{1}{Z(X)} \exp\left\{ \sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- **目标函数（极大似然法）**

$$\max_{\lambda_m \in \mathbb{R}^+} \prod_{x,y} p(y|x;\lambda)^{\tilde{P}(x,y)}$$

- **Why not set the loss as this?**

$$\boxed{\tilde{P}(x,y) = \tilde{P}(y|x)\tilde{P}(x)}$$

$$\max_{\lambda_m \in \mathbb{R}^+} \prod_{x,y} p(y|x;\lambda)^{\tilde{P}(y|x)}$$

CRF 优化

- 实际依然采用对数值进行优化

$$\min_{\lambda\in\mathbb{R}^+} f(\lambda) = -\sum_{x,y} \tilde{P}(x,y)\log p(y|x;\lambda)$$

- 优化方法：
  - 梯度下降
  - 拟牛顿法
  - L-BFGS

CRF 解码-- Viterbi

$$p(Y|X) = \frac{1}{Z(X)} \exp\left\{ \sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- 韦特比变量$\delta_t(i)$
- $\delta_t(i)$的含义是，给定模型λ，在时刻t处于状态i，观察到$o_1\, o_2\, o_3 \ldots o_t$的最佳状态转换序列为$q_1 q_2 \ldots q_t$的概率。

$$\delta_t = \prod_{i=1}^{t} \exp[\sum_k \lambda_k f_k(y_{i-1}, y_i, x)]$$

- 递推公式：

$$\delta_{t+1} = \delta_t \cdot \exp[\sum_k \lambda_k f_k(y_t, y_{t+11}, x)]$$

CRF 词性标注

- w = The quick brown fox jumped over the lazy dog
- s = DET VERB ADJ NOUN-S VERB-P PREP DET ADJ NOUN-S

- Baseline is already 90%
  - Tag every word with its most frequent tag
  - Tag unknown words as nouns

Model	Error
HMM	5.69%
CRF	5.55%

Shallow Parsing

Model	F score
CRF	94.38%
Generalized winnow	93.89%
Voted perceptron	94.09%
MEMM	93.70%

- 完整的parsing或信息提取的前身。
  - 识别文本中各种短语类型的非递归核心。
- 输入：带有POS tag单词的句子
- 任务: 为每个单词打上标签，指示单词是否在短语块(chunk)之外（O），是否开始一个短语块（B），或者是否继续一个短语块（I）。
- CRF 在标准评估数据集上击败了所有单一模型的 NP 分块结果。

NP chunks

Rockwell International Corp. | 's Tulsa unit | said | it | signed | a tentative agreement | extending
its contract | with | Boeing Co. | to provide | structural parts | for | Boeing | 's 747 jetliners | .

Thank you!