最大熵和条件随机场模型

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

概要

- **导引**
- 最大熵模型
- 条件随机场模型

最大熵模型导引

- 机器学习：通过具体的输入输出实例$(\boldsymbol{x}_i, y_i)$学习从输入到输出的映射函数$f(\boldsymbol{x})$

  $\boldsymbol{x} = (x_1, x_2, \dots, x_m)$是输入的特征向量表示

  考虑K类分类问题，即$y \in \{1,2,\dots,K\}$
- 假定函数为线性函数，K个线性函数

  $$\phi_i(\boldsymbol{x}) = w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{im}x_m, \qquad 1 \le i \le K$$
- 分类决策规则

  $$\hat{y} = \operatorname*{argmax}_{1 \le i \le K} \phi_i(\boldsymbol{x})$$
- 模型是特征的线性组合

最大熵模型导引

- 概率型模型

$$\hat{y} = \underset{1 \le i \le k}{\operatorname{argmax}}\, p(y = i|\boldsymbol{x})$$

- 将$\phi_i(\boldsymbol{x})$转换成概率分布
  - 概率需要非负

$$\psi_i\,(\boldsymbol{x}) = \exp(\phi_i(\boldsymbol{x}))$$

  - 概率需要归一

$$p(y = i|\boldsymbol{x}) = \frac{\psi_i\,(\boldsymbol{x})}{\sum_{j=1}^{k} \psi_j(\boldsymbol{x})}$$

- 也就是

$$p(y = i|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\big(\phi_i(\boldsymbol{x})\big), \qquad Z(\boldsymbol{x}) = \sum_{j=1}^{k} \exp\Big(\phi_j(\boldsymbol{x})\Big)$$

can的词类

- can是兼类词(MD/VB/NN)

  Gabriella **can** speak French fluently (MD)

  Two large **cans** of paint ought to be enough (NN)

  Fruits and vegetables that will be **canned**, skinned, diced or otherwise processed (VB)
- 构建模型，给定句子，判别can的词类

  Did you hear that they **canned** Linda (MD/**VB**/NN)

$$p(y = \mathrm{MD}|\boldsymbol{x})$$
$$p(y = VB|\boldsymbol{x})$$
$$p(y = \mathrm{NN}|\boldsymbol{x})$$

can的词类

- 设置一个窗口

  Two large **cans** of paint ought to be enough

  w_{-1} w_0 w_{+1}

- 设计特征(三类特征)

$$x_i = \begin{cases} 1 & w_{-1} = w_i, w_i \in V \\ 0 & otherwise \end{cases}$$

$$x_{|V|+i} = \begin{cases} 1 & w_{+1} = w_i, w_i \in V \\ 0 & otherwise \end{cases}$$

$$x_{2|V|+1} = \begin{cases} 1 & ending(w_0) = s \\ 0 & otherwise \end{cases}$$

$$x_{2|V|+2} = \begin{cases} 1 & ending(w_0) = ed \\ 0 & otherwise \end{cases}$$

$$\boldsymbol{x} = (\overbrace{0, \cdots, 0, \underset{\substack{\uparrow \\ \text{large}}}{1}, 0, \cdots,}^{w_{-1}} \overbrace{0, \cdots, 0, \underset{\substack{\uparrow \\ \text{of}}}{1}, 0, \cdots,}^{w_{+1}} \overset{w_0}{\underset{\substack{\uparrow \\ -\text{s}}}{1}}, 0)$$

做点形式变化…

- 特征**表示**将标签y考虑在内，视作$(\boldsymbol{x}, y)$上的指标函数

$$f_i(\boldsymbol{x}, y) = \begin{cases} 1 & w_{-1} = w_i, \& y = k \\ 0 & otherwise \end{cases}$$

- 线性判别函数可以写成

$$\phi(\boldsymbol{x}, y) = \lambda_1 f_1(\boldsymbol{x}, y) + \lambda_2 f_2(\boldsymbol{x}, y) + \cdots + \lambda_N f_N(\boldsymbol{x}, y) = \phi_y(\boldsymbol{x})$$

$$x_i = \begin{cases} 1 & w_{-1} = w_i, w_i \in V \\ 0 & otherwise \end{cases}$$

$\lambda_{i_1} = w_{1i}$

$$f_{i_1}(\boldsymbol{x}, MD) = \begin{cases} 1 & w_{-1} = w_i \;\&\; y = MD \\ 0 & otherwise \end{cases}$$

$\lambda_{i_2} = w_{2i}$

$\lambda_{i_3} = w_{3i}$

$$f_{i_2}(\boldsymbol{x}, NN) = \begin{cases} 1 & w_{-1} = w_i \;\&\; y = NN \\ 0 & otherwise \end{cases}$$

$$f_{i_3}(\boldsymbol{x}, VB) = \begin{cases} 1 & w_{-1} = w_i \;\&\; y = VB \\ 0 & otherwise \end{cases}$$

最大熵模型导引

- 特征$f_i(\boldsymbol{x}, y)$代表特征x与类y共现，取值为0和1

$$f_i(\boldsymbol{x}, y) = \begin{cases} 1 & if\ x \text{ co}-\text{occur with } y \\ 0 & \text{otherwise} \end{cases}$$

- 模型分布形式

$$p(y|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_i \lambda_i f_i(\boldsymbol{x}, y)\right)$$

$$Z(\boldsymbol{x}) = \sum_y \exp\left(\sum_i \lambda_i f_i(\boldsymbol{x}, y)\right)$$

- 采用最大似然估计方法估计参数$\lambda_1, \lambda_2, \dots.$

如果我们的问题是序列标注任务，怎么办?

输入对象$\boldsymbol{x}$是序列，输出也是序列$\boldsymbol{y}$，长度相等

$$\boldsymbol{x} = x_1 x_2 \cdots x_n \rightarrow \quad \boldsymbol{y} = y_1 y_2 \cdots y_n$$

从最大熵模型到条件随机场

- 序列$\boldsymbol{y}$随着长度指数增长
- 针对所有可能的序列计算得分
- 分解问题(在子序列上定义特征并计算子序列得分)

$$
\begin{array}{ccccccccc}
\cdots & x_{i-1} & & x_i & & x_{i+1} & \cdots \\
 & \updownarrow & & \updownarrow & & \updownarrow & \cdots \\
\cdots & y_{i-1} & \leftrightarrow & \boxed{y_i} & \leftrightarrow & y_{i+1} & \cdots
\end{array}
\quad \Longrightarrow f_j(\boldsymbol{x}, y_i)
$$

$$
\begin{array}{ccccccccc}
\cdots & x_{i-1} & & x_i & & x_{i+1} & \cdots \\
 & \updownarrow & & \updownarrow & & \updownarrow & \cdots \\
\cdots & y_{i-1} & \leftrightarrow & y_i & \leftrightarrow & y_{i+1} & \cdots
\end{array}
\quad \Longrightarrow f_j(\boldsymbol{x}, y_{i-1}y_i)
$$

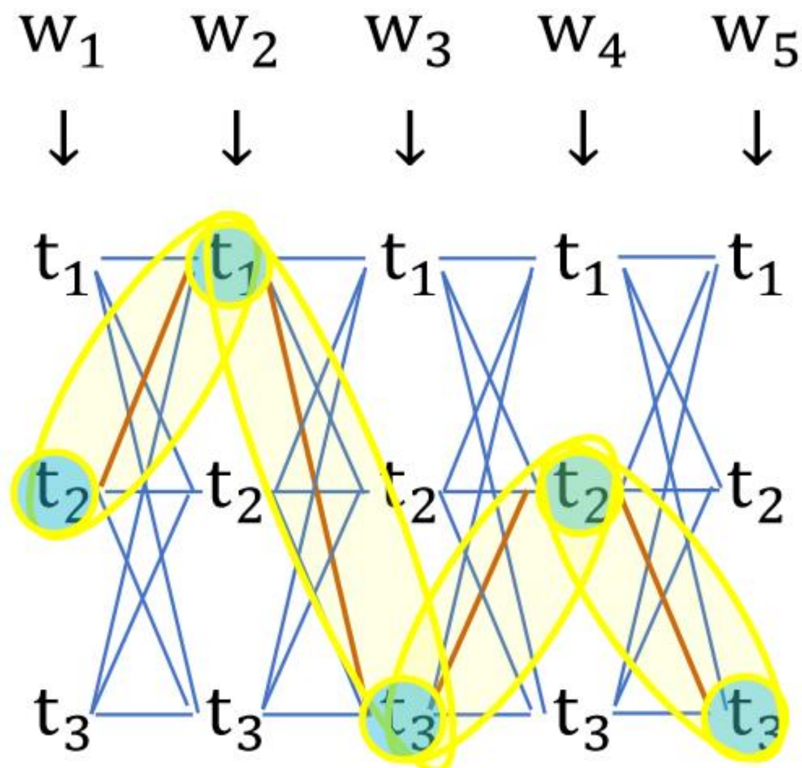
$$\Downarrow$$

$$f_j(\boldsymbol{x}, \boldsymbol{y}_c)$$

从最大熵模型到条件随机场

two	large	cans	of	paint
↓	↓	↓	↓	↓
CD	JJ	NN	IN	NN

two	large	cans	of	paint
↓	↓	↓	↓	↓
CD	JJ	NN	IN	NN



从最大熵模型到条件随机场

- 子序列$\boldsymbol{y}_c$得分仍然是特征加权组合

$$\sum_i \lambda_i f_i(\boldsymbol{x}, \boldsymbol{y}_c)$$

- 序列得分为子序列得分之和

$$\sum_{\boldsymbol{y}_c} \sum_i \lambda_i f_i(\boldsymbol{x}, \boldsymbol{y}_c)$$

- 考虑所有的特征，得到分布形式

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{\boldsymbol{y}_c} \sum_i \lambda_i f_i(\boldsymbol{x}, \boldsymbol{y}_c)\right)$$

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp\left(\sum_{\boldsymbol{y}_c} \sum_i \lambda_i f_i(\boldsymbol{x}, \boldsymbol{y}_c)\right)$$

可是，为什么叫最大熵模型呢…

概要

- 导引
- **最大熵模型**
- 条件随机场模型

例子

- 令$x \in \{a, b\}$且$y \in \{0,1\}$
  已知: $p(a, 0) + p(b, 0) = 0.6$
  求解: $p(x, y)$

$P(x,y)$	**0**	**1**	
a	?	?	
b	?	?	
	0.6		**1.0**

- 满足条件的概率分布有无数多个

例子

- 在众多的概率分布中如何做出选择?

$P(x,y)$	**0**	**1**	
a	0.5	0.1	
b	0.1	0.3	
	0.6		**1.0**

$P(x,y)$	**0**	**1**	
a	0.3	0.2	
b	0.3	0.2	
	0.6		**1.0**

- 不增加任何未知的约束信息，在符合已知约束条件的前提下，尽可能选择均匀分布

最大熵原则(Principle of Maximum Entropy)

- 熵描述了随机变量的不确定性，熵越大表明随机变量的不确定性越大，该随机变量也就越接近均匀分布。

- 在只掌握关于未知分布的部分信息时，应该选取满足这些信息约束但熵最大的概率分布。这就是最大熵原则。

- 按最大熵原则所做的选择，是人们可作出的风险最小的选择，任何其它选择都意味着增加了额外的约束和假设，这些约束和假设根据人们掌握的信息无法作出。

最大熵原则(Principle of Maximum Entropy)

- 基于最大熵原则构建的统计模型称为最大熵模型，利用最大熵原则进行统计建模的方法称为最大熵方法。
- 按照最大熵原则，对于前面的例子进行建模，即为求解下面的问题

$$
\begin{aligned}
& p^* = \underset{p \in P}{\operatorname{argmax}}\, H(p) \\
& p(a,0) + p(b,0) = 0.6 \\
& p(a,0) + p(a,1) + p(b,0) + p(b,1) = 1
\end{aligned}
$$

最大熵原则(Principle of Maximum Entropy)

- 给定样本数据
  $O = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\},\quad x_i \in X, y_i \in Y$
  求解概率分布$p(x, y)$

  例子

- 最大熵分布需满足

  (1) $p^* = \underset{p \in P}{\operatorname{argmax}}\, H(p)$

  (2) $p(x, y)$服从样本中已知的统计证据

- 针对实际问题，一般不存在解析方法

最大熵方法中的特征表示

例子

- 利用特征 表示和提取样本中的已知信息

$$f: X \times Y \rightarrow \{0,1\}$$

  刻画x和y之间的某种共现关系

  对特定样本而言，若共现关系成立，特征值是1，否则特征值是0

- 特征在样本中出现的期望

$$E_{\tilde{p}} f = \sum_{x \in X, y \in Y} \tilde{p}(x,y) f(x,y)$$

  其中：$\tilde{p}(x,y) = \frac{c(x,y)}{N}$

最大熵方法中约束表达

- 特征$f(x,y)$的模型期望可表示为：

$$E_p f = \sum_{x \in X, y \in Y} p(x,y) f(x,y)$$

- 模型分布需符合样本中的统计证据，特征的模型期望应该与特征的观察期望值一致

$$E_p f = E_{\tilde{p}} f$$

- 共有k个特征，则

$$E_p f_j = E_{\tilde{p}} f_j, \qquad 1 \le j \le k$$

最大熵方法中约束表达

$\boldsymbol{x}$	y
1: the soil **can** be from ,	**MD**
2: anything I **can** to end ,	**MD**
3: <bos> Gabriella **can** speak French ,	**MD**
4: all they **can** to find ,	**MD**
5: <bos> <bos> **Can** you help me ,	**MD**
6: <bos> It **can** be quite ,	**MD**
7: bought a **can** of hairspray,	**NN**
8: three large **cans** of paint,	**NN**
9: of cat-food **cans** and smirking,	**NN**
10:that they **canned** Linda ?,	**VB**
11:any provider **can** be built,	**MD**
12:invention **can** be used,	**MD**

$$f_i(\boldsymbol{x}, MD) = \begin{cases} 1 & w_{+1} = be \ \& \ y = MD \\ 0 & otherwise \end{cases}$$

4次 $\Leftrightarrow$ 期望:4/12

$$f_i(\boldsymbol{x}, MD) = \begin{cases} 1 & w_{+1} = to \ \& \ y = MD \\ 0 & otherwise \end{cases}$$

2次 $\Leftrightarrow$ 期望:2/12

返回

求解最大熵分布

- 令P表示所有满足特征约束的分布，则：

$$P = \{p | E_p f_j = E_{\tilde{p}} f_j, 1 \le j \le k\}$$

- 求解最大熵模型(约束最优化)

$$p^* = \underset{p \in P}{\operatorname{argmax}} H(p)$$
$$E_p f_j = E_{\tilde{p}} f_j, 1 \le j \le k$$
$$\textstyle\sum_{x,y} p(x,y) = 1$$

- 拉格朗日乘数法

$$L(p, \mathbf{\Lambda}, \alpha) =$$
$$H(p) + \textstyle\sum_j \lambda_j \left(E_p f_j - E_{\tilde{p}} f_j\right) + \alpha\left(\sum_{x,y} p(x,y) - 1\right)$$
$$\mathbf{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$$

求解最大熵分布

- 利用变分法求解$p(x,y)$

$$p(x,y)=\frac{1}{Z}\exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

$$Z=\sum_{x,y}\exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

- 对数线性模型

  指数部分是特征的线性加权组合，特征f_i对分布的影响通过拉格朗日乘数λ_i来体现。

条件最大熵分布

- 自然语言处理中常用判别模型$p(y|x)$
- 最大化条件熵

$$H(p) = -\sum_{x,y} p(x,y) \log p(y|x)$$

- $p(x,y)$未知，近似表示条件熵

$$H(p) = -\sum_{x,y} p(x)p(y|x) \log p(y|x)$$

$$\approx -\sum_{x,y} \tilde{p}(x)p(y|x) \log p(y|x)$$

条件最大熵分布

- 特征的样本期望

$$E_{\tilde{p}}f = \sum_{x\in X, y\in Y} \tilde{p}(x,y)f(x,y)$$

- 特征的模型期望

$$E_p f = \sum_{x\in X, y\in Y} p(x,y)f(x,y) \approx \sum_{x\in X, y\in Y} \tilde{p}(x)p(y|x)f(x,y)$$

- 模型求解(约束最优化)

$$p^* = \underset{p\in P}{\operatorname{argmax}}\, H(p)$$
$$E_p f_j = E_{\tilde{p}} f_j, 1 \le j \le k$$
$$\sum_y p(y|x) = 1$$

条件最大熵分布

- 利用变分法求解$p(y|x)$

$$p(y|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\exp\left(\sum_i \lambda_i f_i(\boldsymbol{x}, y)\right)$$

$$Z(\boldsymbol{x}) = \sum_y \exp\left(\sum_i \lambda_i f_i(\boldsymbol{x}, y)\right)$$

- 对数线性模型

  指数部分是特征的线性加权组合，特征f_i对分布的影响通过λ_i来体现

  模型参数是$\lambda_1, \lambda_2, \dots, \lambda_k$

最大熵方法的模型训练

- 采用数值最优化算法。设定优化目标：
  - 样本熵值最大化
  - 样本似然值最大化(最大似然估计)
- 二者训练结果一致

$$p^* = \underset{p \in P}{\operatorname{argmax}}\, H(p) = \underset{p \in P}{\operatorname{argmax}}\, L(p)$$

- 最大熵模型训练算法(迭代)

  GIS算法(Generalized Iterative Scaling)

  IIS算法(Improved Iterative Scaling)
- 梯度下降法、拟牛顿法等最优化算法也可用于训练最大熵模型，例如L-BFGS算法

最大熵方法的优点

- 只需针对具体任务，集中精力选择特征
- 特征选择灵活，特征之间无需独立
- 特征的类型与数量都可随时调整
- 无需专门考虑平滑问题

特征选择

- 给定样本数据，可设计出成千上万的特征，并非所有特征的样本期望都是可靠的，很多特征的样本期望带有偶然性，与特征的真实期望并不一致，引入这样的特征无益于统计建模工作。

- 特征选择

  (1) 截止频率

  要求特征在样本中出现的频率大于截止频率

  (2) 特征选择算法

  从所有特征集合F中选择对建模有益的特征S

词类标注中最大熵方法的应用

- 词性标注是一个分类问题
- 当前标记词的环境描述

$$h_i = \{w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, t_{i-2}, t_{i-1}\}$$

- 特征举例

$$f_j(h_i, t_i) = \begin{cases} 1 & if\ t_{i-1} = verb \,\&\, t_i = noun \\ 0 & \text{otherwise} \end{cases}$$

$$f_k(h_i, t_i) = \begin{cases} 1 & if\ t_{i-1} = adverb \,\&\, t_i = verb \\ 0 & \text{otherwise} \end{cases}$$

…一般 都 要 写 **报告**，反映 工作 中…
…当事人 应当 迅速 **报告** 公安 机关…

概要

- 导引
- 最大熵模型
- **条件随机场模型**

图模型(Graphical Model)

- 条件随机场模型是图模型
- 图模型用来为若干随机变量的联合分布进行统计建模，用以将联合分布进行适当的分解
- 据链式规则，n个随机变量的联合分布可分解为

$$p(x_1 x_2 \dots x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1} x_{i-2} \dots x_1)$$

- 若已知随机变量之间的依赖关系，上述分解式中条件分布可略去和变量x_i独立的变量。

图模型(Graphical Model)

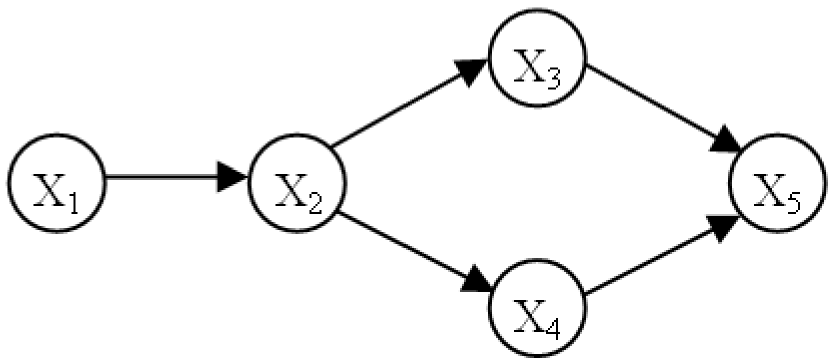
- 用图描述随机变量及其关系
  结点 — 随机变量
  边 — 随机变量间的关系
- 图可是有向图和无向图，分别针对
  (1) 有向图模型
  (2) 无向图模型
- 有向图模型：无环有向图 $G = (X, E)$
  (1) $X = \{X_1, X_2, \dots, X_n\}$是结点集，代表随机变量
  (2) $E = \{(X_i, X_j)\}$是有向边的集合
  X_i是父结点，X_j是子结点，代表X_j依赖X_i

有向图模型举例

- 有向图模型对应分解形式

$$p(x_1 x_2 \dots x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})$$

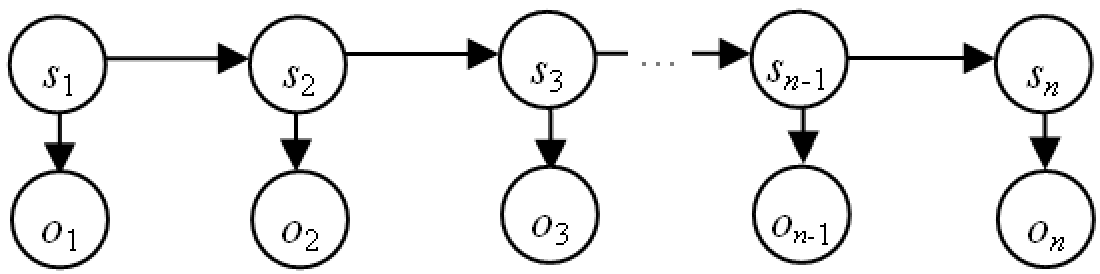
- 示例



$$p(x_1 x_2 x_3 x_4 x_5) = p(x_1) p(x_2|x_1) p(x_3|x_2) p(x_4|x_2) p(x_5|x_3 x_4)$$

有向图模型举例

- HMM模型是有向图模型

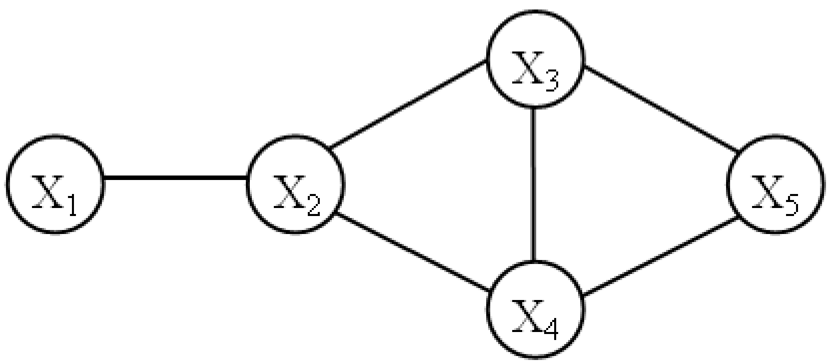


- 对应的分解式

$$p(\mathbf{s}, \mathbf{o}) = p(s_1) \prod_{t=2}^{n} p(s_t | s_{t-1}) \prod_{t=1}^{n} p(o_t | s_t)$$

无向图模型

- 无向图模型：无向图$G = (X, E)$
  (1) $X = \{X_1, X_2, \ldots, X_n\}$是结点集，代表随机变量
  (2) $E = \{(X_i, X_j): i \neq j\}$是无向边的集合
  代表随机变量间的关系
- 团(clique)：无向图的全连通子图
- 极大团(maximal clique)：不能被其它团所包含的团



$\{X_1\}, \{X_2\}, \{X_3\}, \{X_4\}, \{X_5\}$

$\{X_1, X_2\}, \{X_2, X_3\}, \{X_2, X_4\}$

$\{X_3, X_4\}, \{X_3, X_5\}, \{X_4, X_5\}$

$\{X_2, X_3, X_4\}, \{X_3, X_4, X_5\}$

无向图模型

- 势函数(potential function)

$$\psi: X_c \to R^+$$

- 以团为单位将联合概率分布分解为势函数的乘积

$$p(x_1 x_2 \dots x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

- 无向图模型需要(全局)归一化

$$Z = \sum_{X_1} \sum_{X_2} \dots \sum_{X_n} \prod_{c \in C} \psi_c(x_c)$$

- 指数势函数

$$\psi_c(x_c) = \exp\big(\phi_c(x_c)\big)$$

- 无向图模型

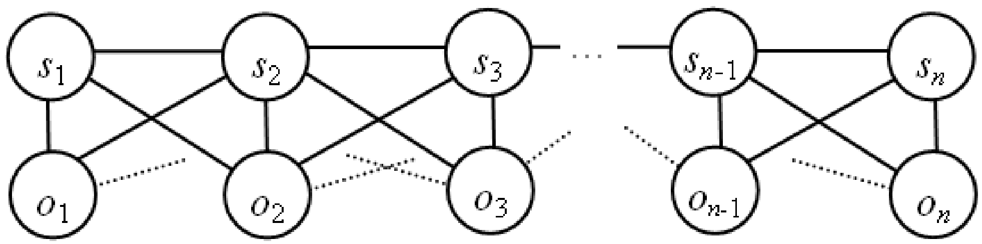
$$p(x_1 x_2 \dots x_n) = \frac{1}{Z} \prod_{c \in C} \exp\big(\phi_c(x_c)\big) = \frac{1}{Z} \exp\left(\sum_{c \in C} \phi_c(x_c)\right)$$

有向图模型与无向图模型的对比

- 共同之处
  将联合分布分解为多个因子
- 不同之处
  有向图模型: 因子是概率分布、无需全局归一
  无向图模型: 因子是势函数，需要全局归一
- 优缺点
  无向图模型中势函数设计不受概率分布约束，设计灵活，但全局归一代价高
  有向图模型无需全局归一、训练相对高效

条件随机场模型(Conditional Random Fields)

- 2001年Lafferty提出，在自然语言处理中应用广泛
- 无向图模型，特征设计灵活，但需全局归一
- 链式条件随机场模型的图结构



- 链式条件随机场模型图结构中的团

$$C = \{\{s_t\}|t = 1,2,\dots,n\} \cup \{\{(s_{t-1}, s_t)|t = 2,3,\dots,n\}$$

- 条件随机场模型的分解式

$$p(\boldsymbol{s}|\boldsymbol{o}) = \frac{1}{Z(\boldsymbol{o})} \prod_{c \in C} \exp\big(\phi_c(\boldsymbol{s}_c, \boldsymbol{o})\big) = \frac{1}{Z(\boldsymbol{o})} \exp\left(\sum_{c \in C} \phi_c(\boldsymbol{s}_c, \boldsymbol{o})\right)$$

条件随机场模型

- 以团为单位定义特征

$$f_i(\boldsymbol{s}_c, \boldsymbol{o}), \qquad i = 1, 2, \dots, k$$

- 指数势函数

$$\psi_c(\boldsymbol{s}_c, \boldsymbol{o}) = \exp\big(\phi_c(\boldsymbol{s}_c, \boldsymbol{o})\big)$$

- 指数部分

$$\phi_c(\boldsymbol{s}_c, \boldsymbol{o}) = \sum_i \lambda_i f_i(\boldsymbol{s}_c, \boldsymbol{o})$$

- 分布形式

$$p(\boldsymbol{s}|\boldsymbol{o}) = \frac{1}{Z(\boldsymbol{o})} \exp\left(\sum_{c \in C} \sum_i \lambda_i f_i(\boldsymbol{s}_c, \boldsymbol{o})\right)$$

条件随机场模型

- 以团为单位定义特征

$$f_i(\boldsymbol{s}_c, \boldsymbol{o}), \qquad i = 1, 2, \dots, k$$

- 按照最大熵原则求解$p(\boldsymbol{s}|\boldsymbol{o})$

$$p^* = \underset{p \in P}{\operatorname{argmax}} H(p)$$

$$E_p f_j = E_{\tilde{p}} f_j, 1 \le j \le k$$

$$\sum_{\boldsymbol{o}} p(\boldsymbol{s}|\boldsymbol{o}) = 1$$

- 优化目标：条件熵

$$H(p) \approx -\sum_{\boldsymbol{o}} \tilde{p}(\boldsymbol{o}) \sum_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{o}) \log p(\boldsymbol{s}|\boldsymbol{o})$$

条件随机场模型

- 特征f_i在$(\boldsymbol{s},\boldsymbol{o})$上出现的次数

$$\sum_{c\in C} f_i(\boldsymbol{s}_c,\boldsymbol{o})$$

- 约束特征的样本期望与模型期望相同

$$\sum_{(\boldsymbol{s},\boldsymbol{o})} \tilde{p}(\boldsymbol{s},\boldsymbol{o}) \sum_{c\in C} f_i(\boldsymbol{s}_c,\boldsymbol{o}) = \sum_{\boldsymbol{o}} \tilde{p}(\boldsymbol{o}) \sum_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{o}) \sum_{c\in C} f_i(\boldsymbol{s}_c,\boldsymbol{o})$$

- 运用变分法，求解$p(\boldsymbol{s}|\boldsymbol{o})$

$$p(\boldsymbol{s}|\boldsymbol{o}) = \frac{1}{Z(\boldsymbol{o})} \exp\left(\sum_{c\in C}\sum_{i} \lambda_i f_i(\boldsymbol{s}_c,\boldsymbol{o})\right)$$

- 对数线性模型

条件随机场

- 训练：最大似然估计
- 对数似然函数$L(\boldsymbol{\Lambda})$

$$L(\boldsymbol{\Lambda}) = \frac{1}{N}\log\prod_{i=1}^{N} p(\boldsymbol{s}_i|\boldsymbol{o}_i)$$

- 参数求解

$$\boldsymbol{\Lambda}^* = \underset{\boldsymbol{\Lambda}}{\operatorname{argmax}}\, L(\boldsymbol{\Lambda})$$

- 优化算法
  - 梯度下降、拟牛顿法等
  - GIS和IIS算法

条件随机场模型

- 解码

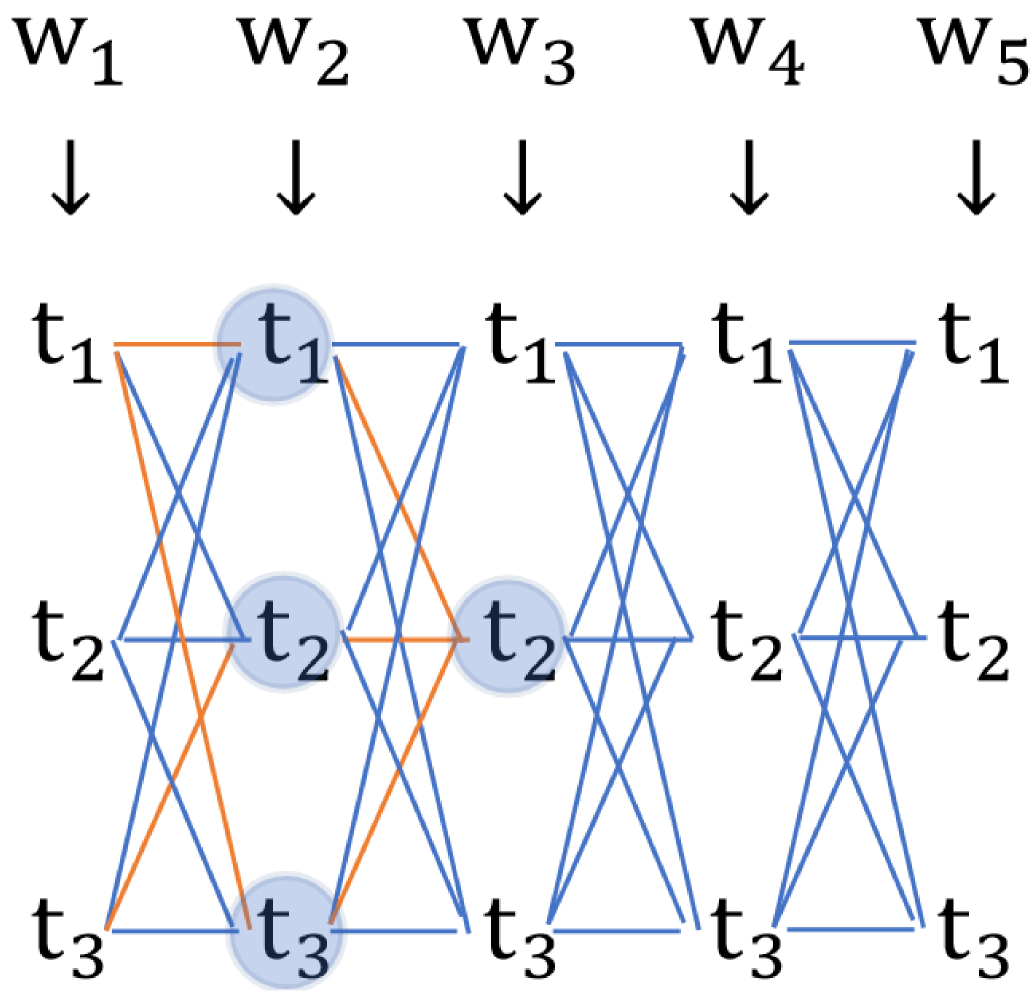
$$\boldsymbol{s}^* = \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{o})$$
$$= \operatorname*{argmax}_{\boldsymbol{s}} \sum_i \phi_{c_i}(s_{i-1}, s_i, \boldsymbol{o})$$

- Viterbi变量

$$\delta_i(s) \triangleq \max_{s_1,\dots,s_{i-1}} \left[\sum_{k=1}^{i-1} \phi_{c_k}(s_{k-1}, s_k) + \phi_{c_i}(s_{i-1}, s) \right]$$

时刻i到达结点s最佳路径得分

条件随机场模型



返回

条件随机场模型

- 解码算法—Viterbi算法
  - 初始化

$$\delta_1(s) = 0, \qquad \forall s \in \mathcal{S}$$
$$\psi_1(s) = 0, \qquad \forall s \in \mathcal{S}$$

  - 递归计算

$$\delta_i(s) = \max_{s' \in \mathcal{S}} \left[\delta_{i-1}(s') + \phi_{c_i}(s', s)\right], \forall s \in \mathcal{S}, 1 < i \leq n$$
$$\psi_i(s) = \operatorname{argmax}_{s' \in \mathcal{S}} \left[\delta_{i-1}(s') + \phi_{c_i}(s', s)\right], \forall s \in \mathcal{S}, 1 < i \leq n$$

  - 终止

$$\delta_n(s_n^*) = \max_{s' \in \mathcal{S}} \delta_n(s')$$
$$s_n^* = \operatorname{argmax}_{s' \in \mathcal{S}} \delta_n(s')$$

  - 路径回溯

$$s_i^* = \psi_{i+1}(s_{i+1}^*), \qquad i = n-1, n-2, \dots, 1$$

图示

条件随机场模型

- 理论上较为完善的序列标记模型
  无标记偏执问题
- 兼具判别模型和无向图模型的优点
  特征设计灵活、无需要求特征独立
- 条件随机场模型训练代价大、复杂度高
- 除链式图结构，亦可设计其他图结构
  如：网页的链接结构
- 在自然语言处理领域应用广泛

判别模型和生成模型

- 序列标注问题：给观察序列标注标记序列
- 令$\boldsymbol{o}$和$\boldsymbol{s}$代表观察序列和标记序列

$$\boldsymbol{o} = o_1 o_2 \dots o_N$$

$$\boldsymbol{s} = s_1 s_2 \dots . s_N$$

则给定$\boldsymbol{o}$标注$\boldsymbol{s}$的过程

$$\boldsymbol{s}^* = \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{o})$$

- 为此需要对**o**和**s**进行统计建模

判别模型和生成模型

- 对$\boldsymbol{o}$和$\boldsymbol{s}$进行统计建模，通常有两种方式

  (1) 生成模型

  构建$\boldsymbol{o}$和$\boldsymbol{s}$的联合分布 $p(\boldsymbol{s},\boldsymbol{o})$

  (2) 判别模型

  构建$\boldsymbol{o}$和$\boldsymbol{s}$的条件分布 $p(\boldsymbol{s}|\boldsymbol{o})$
- 判别模型与序列标记问题有较好的对应性
- 在利用生成模型进行序列标注时，理论上需要

$$p(\boldsymbol{s}|\boldsymbol{o}) = \frac{p(\boldsymbol{s},\boldsymbol{o})}{\sum_{\boldsymbol{s}'} p(\boldsymbol{s}',\boldsymbol{o})}$$

判别模型和生成模型对比

- 生成模型 —联合分布
  判别模型 —条件分布
- 如何对待观察序列
  生成模型中，观察序列作为模型的一部分
  判别模型中，观察序列只作为条件，可以针对观察序列灵活提取特征
- 训练复杂度不同
  判别模型训练复杂度通常高
- 是否支持无指导训练
  生成模型支持无指导训练，判别模型无指导训练代价高

判别模型和生成模型

- HMM模型是生成模型

$$p(\mathbf{s},\mathbf{o}) = p(s_1)\prod_{t=2}^{N} p(s_t|s_{t-1})\prod_{t=1}^{N} p(o_t|s_t)$$

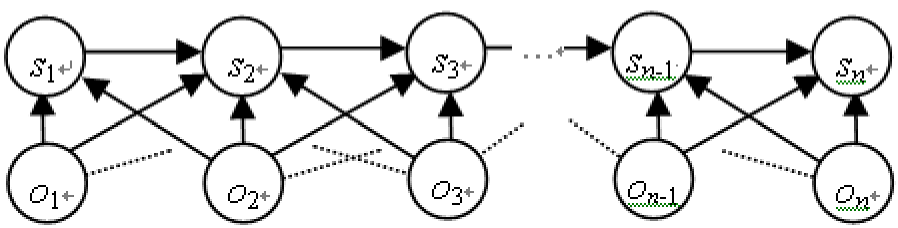
- 针对联合概率$p(\boldsymbol{s},\boldsymbol{o})$建模
- 支持无指导训练：Baum-Welch算法
- 过强的独立性假设限制了模型的改进，无法使用关于观察值的多重特征以及不相互独立的特征
- 训练使用联合分布、标注使用条件分布，对性能造成负面影响

介绍一个有缺陷的模型

条件马尔科夫模型

条件马尔可夫模型

- HMM的缺陷，特征设计受限
- 条件马尔可夫模型
  判别模型、有向图模型
- 条件马尔可夫模型的图结构



- 条件马尔可夫模型的分解式

$$p(\boldsymbol{s}|\boldsymbol{o}) = \prod_{i=1}^{N} p(s_i|s_{i-1}, \boldsymbol{o})$$

条件马尔可夫模型

- 因子是条件分布(标记转移分布)，无需全局归一
- 观察序列在因子中作为条件出现，特征设计灵活
- 对标记转移分布，采用最大熵模型

$$p(s_i|s_{i-1},\boldsymbol{o}) = \frac{1}{Z(s_{i-1},\boldsymbol{o})}\exp\left(\sum_k \lambda_k f_k(s_{i-1},s_i,\boldsymbol{o})\right)$$

$$Z(s_{i-1},\boldsymbol{o}) = \sum_{s_j}\exp\left(\sum_k \lambda_k f_k(s_{i-1},s_i,\boldsymbol{o})\right)$$

条件马尔可夫模型

- 最大熵马尔可夫模型(MEMM)是一种简化了的条件马尔可夫模型

$$p(s_i|s_{i-1},\boldsymbol{o}) = p(s_i|s_{i-1},o_i)$$

- 模型训练：最大似然估计+梯度下降
- 模型解码：Viterbi算法
- 模型缺陷 — 标记偏执问题 (Label Bias problem)
- 标记偏执的例子

All	**the**	**indexes**	**dove**
PDT	*DT*	…	
DT			

the只有一个词类标记，故 $p(DT|the,s_{t-1}) = 1$

$$p(DT|the,PDT) = p(DT|the,DT) = 1$$

PDT-*DT*转移概率远大于*DT*-*DT*，被忽略

条件马尔可夫模型

- 若标记转移分布熵值低，则会有标记偏执问题
- 标记偏执的原因在于局部归一

$$\sum_{s_t} p(s_t|s_{t-1}, \boldsymbol{o}) = 1$$

- 标记偏执—条件马尔可夫模型失败
- 解决策略—取消局部归一，代之以全局归一