模式识别与机器学习

条件随机场CRF

主讲：图像处理与模式识别研究所
赵群飞　教授

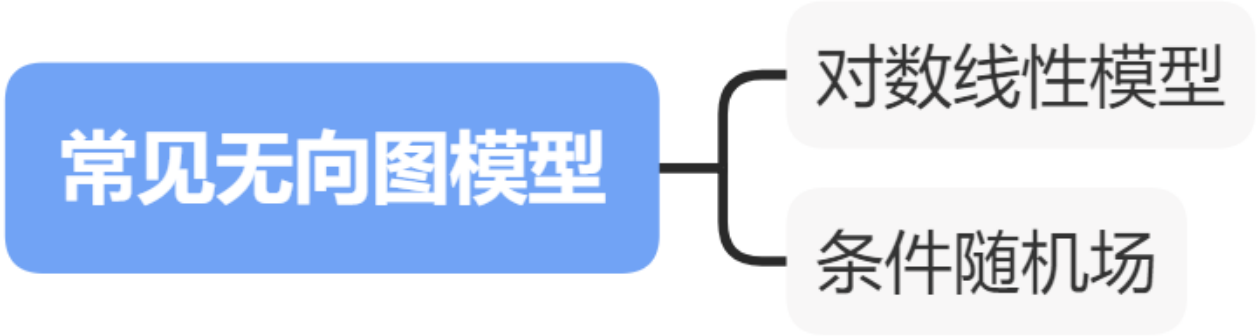
邮　箱：zhaoqf@sjtu.edu.cn
办公室：电院2-441
电　话：13918191860

学习目标

- 掌握条件随机场的模型表示
- 了解线性链条随机场与HMM的关系
- 掌握用于模型推理的前向-后向算法、韦特比算法
- 掌握参数学习的最大后验估计算法

常见的无向图模型

很多机器学习模型可以使用无向图模型来描述，比如对数线性模型（也叫最大熵模型）和条件随机场等。



本节以对数线性模型，介绍其模型假设及其概率图模型表示。

对数线性模型

- 势能函数一般定义为：

$$\phi_c(\mathbf{x}_c|\theta_c) = \exp\left(\theta_c^{\mathsf{T}} f_c(\mathbf{x}_c)\right),$$

- 其中　函数$f_c(\mathbf{x}_c)$为定义在$\mathbf{x}_c$上的特征向量，参数θ_c为权重向量。

□ 这样联合概率分布的对数形式为：

$$\ln p(\mathbf{x}|\theta) = \sum_{c \in C} \theta_c^{\mathsf{T}} f_c(\mathbf{x}_c) - \ln Z(\theta),$$

其中θ代表所有势能函数中的参数$\{\theta_c\}$。

- 这种形式的无向图模型也称为对数线性模型或最大熵模型。如果用对数线性模型来建模条件概率分布$p(\mathbf{y}|\mathbf{x})$，那么带有参数的条件概率分布表示$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$为：

$$p(\mathbf{y}|\mathbf{x},\theta)=\frac{1}{Z(\mathbf{x},\theta)}\exp(\theta^{\top}f(\mathbf{x},\mathbf{y}))\,,$$

其中

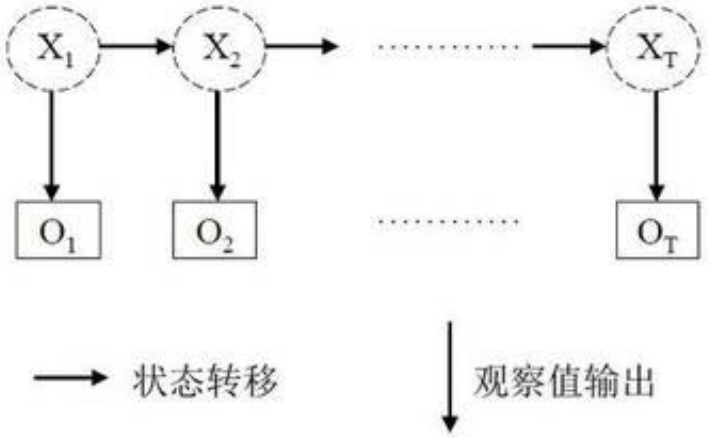
$$Z(\mathbf{x},\theta)=\sum_{\mathbf{y}}\exp(\theta^{\top}f(\mathbf{x},\mathbf{y}))$$

一、条件随机场CRF

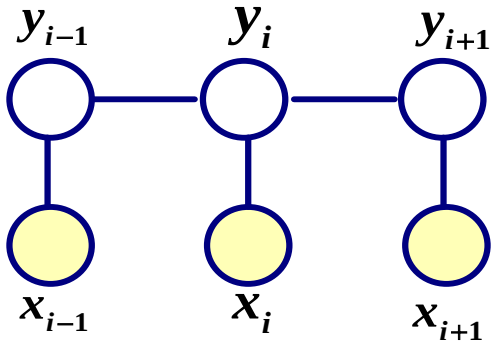
- 在NLP领域，分词、词性标注、命名实体识别、语法解析等基础任务都归结为结构化预测。当处理系列数据时，结构化预测也可以称为序列标注，其中序列中的每一个元素对应一个标签，标签之间，以及标签与元素之间存在着依赖关系。
- *条件随机场*(*CRF:*Conditional Random Fields)是解决序列标注问题的概率模型，是一种判别式无向图模型，无向图的边可以捕获必要的依赖关系，直接建模结构化输出与输入之间的关系。

CRF与HMM

- CRF与HMM对样本标签的识别过程类似，都常用来做序列标注的建模然，而两者的训练模式及模型假设具有较大差异。
  - HMM最大的缺点就是由于其输出独立性假设，导致其不能考虑上下文的特征，限制了特征的选择；在每一节点都要进行归一化，所以只能找到局部的最优值，同时也带来了标记偏见的问题（label bias）；
  - CRF：选择上下文相关特性；不在每一个节点进行归一化，而是所有特征进行全局归一化，可以求得全局的最优值。



HMM



CRF

- 以命名实体识别为例：
  - **CRF**是监督模型：训练通常用带标签的语料数据库，输入序列是有汉字组成的句子，输出序列是每个汉字对应的标签。
  - **CRF**不对观测文本进行概率分布假设，而是通过特征函数的方式构建观测与标签、标签与标签之间的关系，并建立标签序列在给定观测文本情况下的全局条件分布。

标注：人名 地名 组织名

观察序列：毛泽东

实体命名识别

标注：名词 动词 助词 形容词 副词 ……

观测序列：今天天气非常好！

汉语词性标注

二、一般条件随机场模型表示

- CRF是在判别式框架下对成对的观测序列和标签构建一个条件模型。

定义：设 G=(V,E) 是一个无向图，$Y=\{Y_v \mid v \in V\}$ 是以G中节点 v 为索引的随机变量 Y_v 构成的集合。在给定 X 的条件下，如果每个随机变量 Y_v 服从马尔可夫属性，即

$$p(Y_v \mid X, Y_u, u \neq v) = p(Y_v \mid X, Y_u, u \sim v),$$

则 (X,Y) 就构成一个条件随机场。 $u \sim v$ 表示 u 和 v 在图G中是相邻的节点。

- 从定义可以看出一般的条件随机场是在给定观测X 的条件下关于随机变量Y 的马尔可夫随机场。它对数据的结构没有链图的约束，应用范围很广。

- 用于图像分割时，每个像素特征可以看成输入节点，每个像素的标注类别是对应的输出节点，这些节点就构成网状结构。

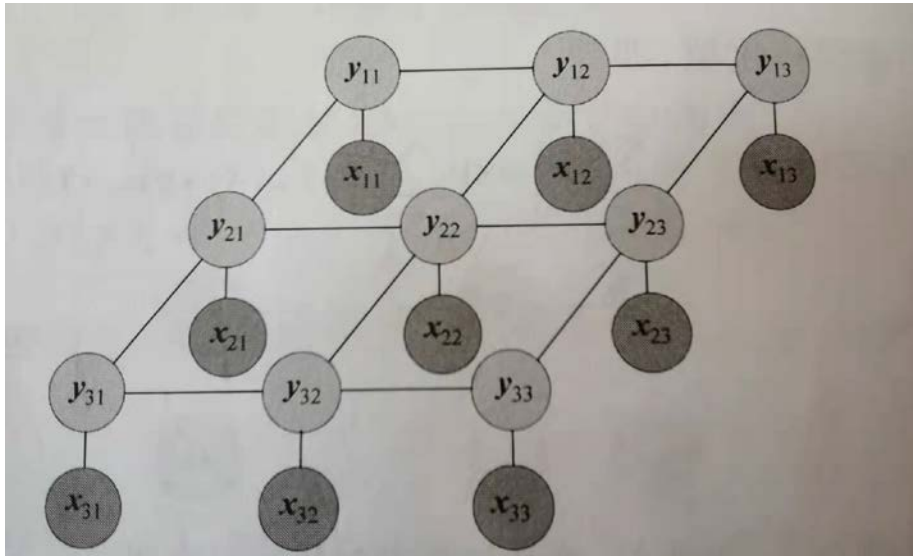
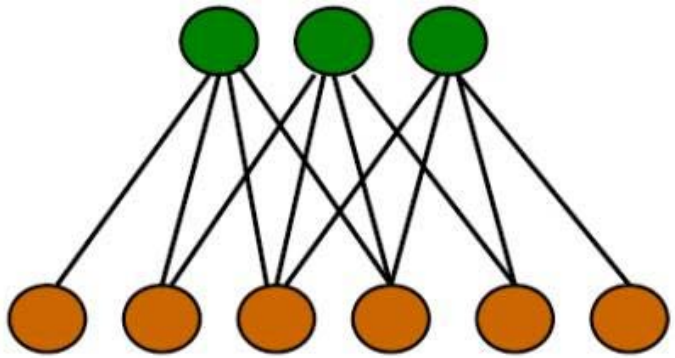


图 6-4　网状结构条件随机场示意图

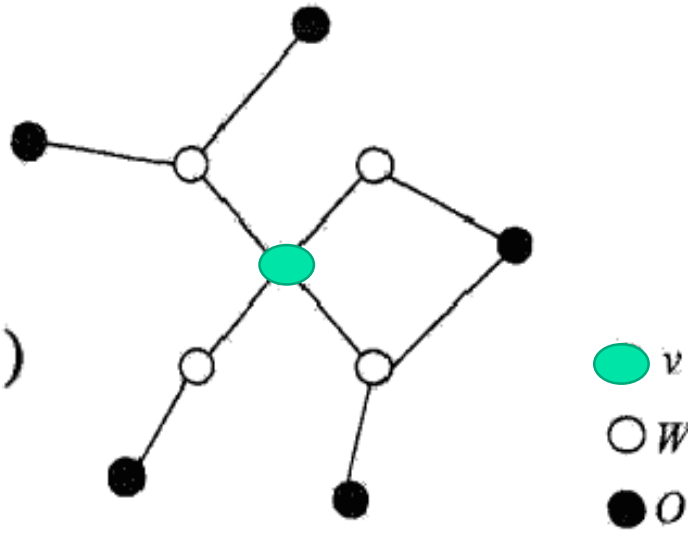
- 成对马尔可夫性 (Pairwise Markov property)
  - 设u和v是无向图G中任意两个没有边连接的结点，结点u和v分别对应随机变量Y_u和Y_v，
  - 其他所有结点为 O，对应的随机变量组是Y_O
  - 给定随机变量组Y_O的条件下随机变量Y_u和Y_v是条件独立的

$$P(Y_u, Y_v \mid Y_O) = P(Y_u \mid Y_O)P(Y_v \mid Y_O)$$



- 局部马尔可夫性(Local Markov properly)
  - v 任意结点
  - W与v有边相连
  - O 其它

$$P(Y_v, Y_O \mid Y_W) = P(Y_v \mid Y_W) P(Y_O \mid Y_W)$$



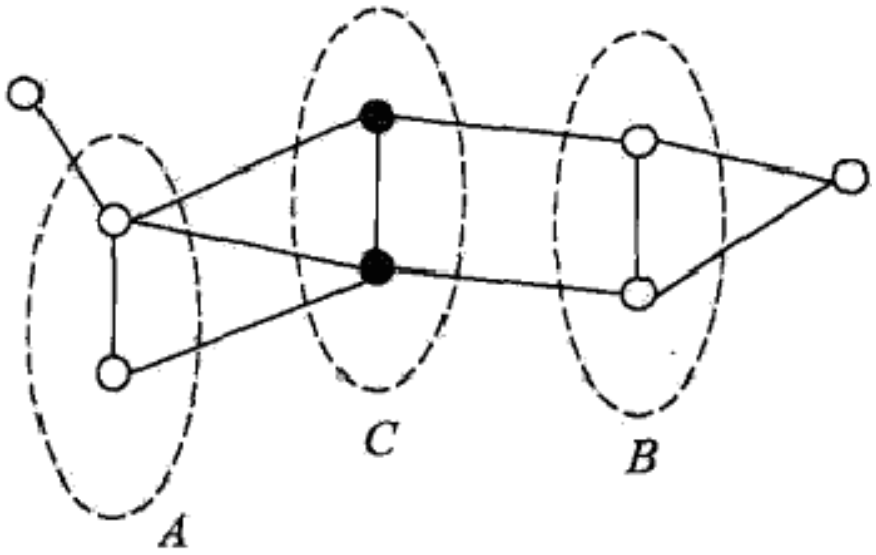
在 $P(Y_O \mid Y_W) > 0$ 时，等价于

$$P(Y_v \mid Y_W) = P(Y_v \mid Y_W, Y_O)$$

模型定义

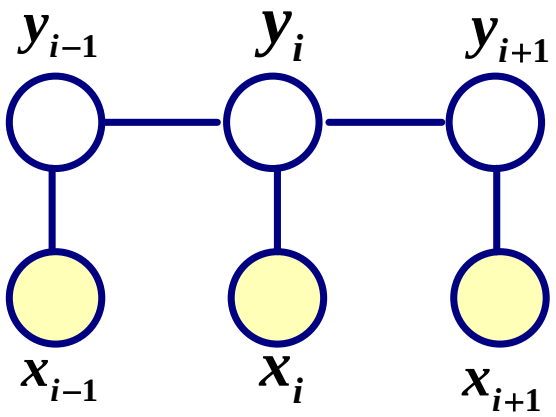
- 全局马尔可夫性(Global Markov property)
  - **结点集合A, B是在无向图G中被结点集合C分开的任意结点集合，**

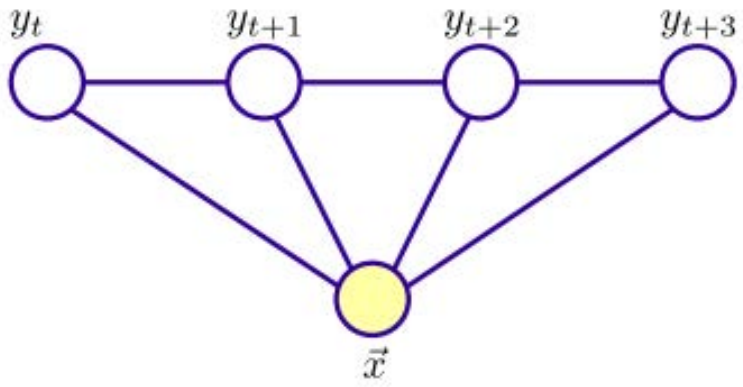
$$P(Y_A, Y_B \mid Y_C) = P(Y_A \mid Y_C) P(Y_B \mid Y_C)$$



三、Linear-chain CRFs 模型

最简单且最常用的是一阶链式结构，即线性链结构
(Linear-chain CRFs)





Linear-chain CRFs 模型

令 $x=\{x_1,x_2,\cdots,x_n\}$ 表示观察序列，$y=\{y_1,y_2,\cdots,y_n\}$ 是有限状态的集合，

λ 为模型参数，根据随机场的基本理论：

$$p(y|x,\lambda)\propto \exp\left(\sum_j \lambda_j t_j(y_{i-1},y_i,x,i)+\sum_k \mu_k s_k(y_i,x,i)\right)$$

$t_j(y_{i-1},y_i,x,i)$：对于观察序列的标记位置 *i-1* 与 *i* 之间的转移特征函数

$s_k(y_i,x,i)$：观察序列的 *i* 位置的状态特征函数

将两个特征函数统一为：$f_j(y_{i-1},y_i,x,i)$

$$p(y|x,\lambda)=\frac{1}{Z(x)}\exp\left(\sum_{i=1}^{n}\sum_j \lambda_j f_j(y_{i-1},y_i,x,i)\right)$$

$$Z(x)=\sum_j \exp\left(\sum_{i=1}^{n}\sum_j \lambda_j f_j(y_{i-1},y_i,x,i)\right)$$

四、CRF关键问题

1.特征函数的选择

特征函数的选取直接关系模型的性能。

2.参数估计

从已经标注好的训练数据集学习条件随机场模型的参数，即各特征函数的权重向量 λ 。

3.模型推断

在给定条件随机场模型参数 λ 下，预测出最可能的状态序列。

1.特征函数的选择

- CRF模型中特征函数的形式定义：$f_j(y_{j-1}, y_i, x, i)$
- 它是状态特征函数和转移特征函数的统一形式表示。特征函数通常是二值函数，取值要么为1要么为0。
- 在定义特征函数的时候，首先构建观察值上的真实特征 $b(x,i)$ 的集合，即所有 i 时刻的观察值 x 的真实特征，结合其对应的标注结果，就可以获得模型的特征函数集。

$$b(x,i)=\begin{cases}1 & \text{如果时刻 i 观察值x是大写开头}\\ 0 & \text{否则}\end{cases}$$

$$f(y_{i-1}, y_i, x, i)=\begin{cases}b(x,i) & if \quad y_{i-1}=<title>, y_i=<author>\\ 0 & otherwise\end{cases}$$

2.参数估计

极大似然估计（MLE：Maximum Likelihood Estimation）

假定对于训练数据有一组样本集合D=$\{(x_j, y_j)|j = 1,2,\ldots,n\}$，

样本是相互独立的。

那么，CRF模型中的条件概率 $p(y|x,\lambda)$ 的对数极大似然函数为：

$$L(\lambda)=\sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n}\left(\sum_{j}\lambda_j f_j((y_{i-1},y_i,x,i))\right)-\sum_{x}\tilde{p}(x)\log Z(x)$$

其中，$\tilde{p}(x,y)$ 为训练样本中(x,y)的经验概率

$$\tilde{P}(x,y)=\frac{(x,y)\text{在样本中同时出现的次数}}{\text{样本空间的容量}}$$

$\tilde{P}(x)$ 是随机变量x在训练样本中的经验分布

$$\tilde{P}(x)=\frac{x\text{在样本中出现的次数}}{\text{样本空间的容量}}$$

分别对 λ_j 求导:

$$\frac{\partial L(\lambda)}{\lambda_j} = \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x) - \sum_{x,y} \tilde{p}(x) p(y|x,\lambda) \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x)$$

$$= E_{\tilde{p}(x,y)}\left[f_j(x,y)\right] - \sum_{k} E_{p(y|x^{(k)},\lambda)}\left[f_j(x^{(k)},y)\right]$$

令上式等于0，求λ

模型分布中特征的期望等于经验分布中的期望值——最大熵原理

- 上述方法直接使用对数最大似然估计，可能会发生过度学习问题，通常引入惩罚函数的方法解决这一问题。

使用惩罚项 $\frac{\sum_j \lambda_j^2}{2\sigma^2}$ 对数似然函数公式变为：

$$L(\lambda) = \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^{n} \left(\sum_j \lambda_j f_j((y_{i-1}, y_i, x, i)) \right) - \sum_x \tilde{p}(x) \log Z(x) - \frac{\sum_j \lambda_j^2}{2\sigma^2}$$

对上式中每个 λ_j 求偏导，并令结果为0，求 λ_j

- 由于极大似然估计并不一定能得倒一个近似解，因而需要利用一些迭代技术来选择参数，使对数似然函数最大化。

迭代缩放

Lafferty 提出迭代缩放的算法用于估计条件随机场的极大似然参数

迭代缩放是一种通过更新规则以更新模型中的参数，通过迭代改善联合或条件模型分布的方法。更新规则如下：

$$\lambda_j \leftarrow \lambda_j + \delta\lambda_j$$

其中更新值 $\delta\lambda_j$ 使得新的值 λ_j 比原来的值 λ_j 更接近极大似然值。

迭代缩放的基本原理

假定我们有一个以 $\lambda=\{\lambda_1,\lambda_2,\cdots\}$ 为参数的模型 $p(y|x,\lambda)$， 并且要找到一组新的参数： $\lambda+\Delta=\{\lambda_1+\delta\lambda_1,\lambda_2+\delta\lambda_2,\cdots\}$ 使得在该参数条件下的模型具有更高的对数似然值。 通过迭代，使之最终达到收敛。

对于条件随机场对数似然值的变化可以表示为：

$$L(\lambda+\Delta)-L(\lambda)=\sum_{x,y}\tilde{p}(x,y)\log p(y|x,\lambda+\Delta)-\sum_{x,y}\tilde{p}(x,y)\log p(y|x,\lambda)$$

$$=\sum_{x,y}\tilde{p}(x,y)\left[\sum_{i=1}^{n}\sum_{j}\delta\lambda_j f_j(y_{i-1},y_i,x)\right]-\sum_{x}\tilde{p}(x)\log\frac{Z_{\lambda+\Delta}(x)}{Z_{\lambda}(x)}$$

引入辅助函数：

$$A(\lambda,\Delta)\square\sum_{x,y}\tilde{p}(x,y)\left[\sum_{i=1}^{n}\sum_{j}\delta\lambda_j f_j(y_{i-1},y_i,x)\right]+1$$

$$-\sum_{x}\tilde{p}(x)p(y|x,\lambda)\left[\sum_{i=1}^{n}\sum_{j}\left(\frac{f_j(y_{i-1},y_i,x)}{T(x,y)}\right)\exp(\delta\lambda_j T(x,y))\right]$$

$$T(x,y)=\sum_{i=1}^{n}\sum_{j}f_j(y_{i-1},y_i,x)$$

定义为在观察序列和标记序列为(x,y)的条件下，特征值为1的特征的个数。

根据 $L(\lambda+\Delta)-L(\lambda)\geq A(\lambda,\ \Delta)$，寻找使 $A(\lambda,\ \Delta)$ 最大化的Δ，使用迭代算法计算最大似然参数集。

迭代过程：（A）将每个 λ_j 设初始值；

（B）对于每个 λ_j，计算 $\frac{\partial A(\lambda,\Delta)}{\partial \delta\lambda_j}=0$，即

$$\frac{\partial A(\lambda,\Delta)}{\partial \delta\lambda_j}=\sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n}f_j(y_{i-1},y_i,x)$$

$$-\sum_{x}\tilde{p}(x)\sum_{y}p(y|x,\lambda)\sum_{i=1}^{n}f_j(y_{i-1},y_i,x)\exp(\delta\lambda_j T(x,y))=0$$

应用更新规则 $\lambda_j \leftarrow \lambda_j+\delta\lambda_j$，更新每个参数，直到收敛。

Lafferty 提出两个迭代缩放的算法用于估计条件随机场的极大似然参数

- GIS算法（Generalised Iterative Scaling）
- IIS算法（Improved Iterative Scaling）

3.模型推断

对于一个给定观察序列 $x=\{x_1,x_2,\cdots,x_n\}$，求使得该观察序列出现概率最大的标记序列(状态序列)$y=\{y_1,y_2,\cdots,y_n\}$。

$$p(y|x,\lambda)=\frac{1}{Z(x)}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right)$$

$$Z(x)=\sum_{j}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right)$$

常见的两个问题：

1. 在模型训练中，需要边际分布 $p(y_t,y_{t-1}|x)$ 和 $Z(x)$；
2. 对于未标记的序列，求其最可能的标记。

第1个问题采用前向后向法解决；第2个问题通过Viterbi算法解决：Viterbi算法是一种动态规划算法，其思想精髓在于将全局最佳解的计算过程分解为阶段最佳解的计算。

条件随机场模型举例——中文命名实体识别

在中文信息处理领域，命名实体识别是各种自然语言处理技术的重要基础。

命名实体：人名、地名、组织名三类

模型形式

$$p(y|x,\lambda)=\frac{1}{Z(x)}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right)$$

$$Z(x)=\sum_{j}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right)$$

关键：特征函数的确定

适用于人名的特征模板

“上下文”，指的是包括当前词w_0及其前后若干个词的一个“观察窗口”（w_{-n}, w_{-n+1}，…w_0…，w_n）。理论上来说，窗口越大，可利用的上下文信息越多，但窗口开得过大除了会严重降低运行效率，还会产生过拟合现象；而窗口过小，特征利用的就不够充分，会由于过于简单而丢失重要信息。

通过一些模板来筛选特征。模板是对上下文的特定位置和特定信息的考虑。

“人名的指界词”：主要包括称谓词、动词和副词等，句首位置和标点符号也可。

根据指界词与人名同现的概率的大小，将人名的左右指界词各分为两级，生成4个人名指界词列表：

类型	级别	列表名称	举例
左指界词	1 级	PBW1	记者、纪念
	2 级	PBW2	称赞、叮咛
右指界词	1 级	PAW1	报道、会见
	2 级	PAW2	供认、坚决

还建立了若干个资源列表，包括：中国人名姓氏用表、中国人名名字用表、欧美俄人名常用字表、日本人名常用字表。

定义了用于人名识别特征的原子模板，每个模板都只考虑了一种因素：

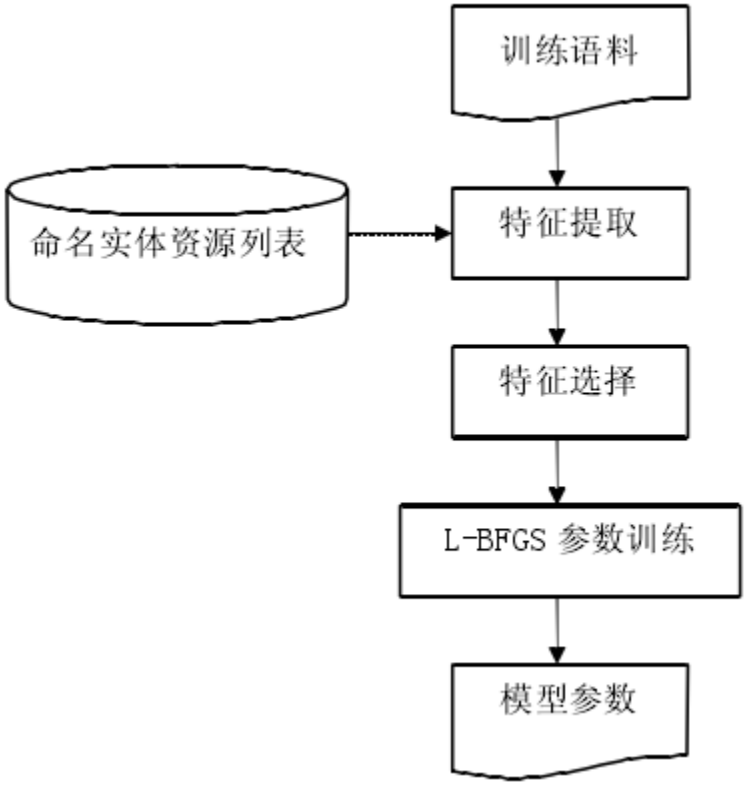
序号	原子模板	意义
P1	ChSurName	当前词是否为中国人名姓氏用字
P2	ChLastName	当前词是否为中国人名名字用字
P3	EurName	当前词是否为欧美俄人名常用字
P4	JapName	当前词是否为日本人名常用字
P5	PerFirRightBoundary	当前词后面第一个词是否为右指界词（1、2 级）
P6	PerSecRightBoundary	当前词后面第二个词是否为右指界词（1、2 级）
P7	PerFirLeftBoundary	当前词前面第一个词是否为左指界词（1、2 级）
P8	PerSecLeftBoundary	当前词前面第二个词是否为左指界词（1、2 级）

当特征函数取特定值时，特征模板被实例化就可以得到具体的特征。

“当前词的前一个词w_{-1}在人名1级左指界词列表中出现”

$$f_i(x,y)=\begin{cases}1 & \text{If } PBW1(w_{-1})=ture \text{ and } y=person\\ 0 & \text{else}\end{cases}$$

类似的，做地名、组织名的特征提取和选择，并将其实例化，得到所有的特征函数。



模型训练流程图

评测指标

$$\text{召回率（Recall）} = \frac{\text{正确识别的命名实体首部（尾部）的个数}}{\text{标准结果中命名实体首部（尾部）的的总数}} \times 100\%$$

$$\text{精确率（Precision）} = \frac{\text{正确识别的命名实体首部（尾部）的个数}}{\text{识别出的命名实体首部（尾部）的总数}} \times 100\%$$

$$\text{F-值} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}}$$

条件随机场CRF

- 如果在给定某些条件的前提下，马尔可夫随机场就变成*条件随机场 (CRF:*Conditional Random Fields*)*。如果使用条件随机场解决标注问题，并且进一步将条件随机场中的网络拓扑变成线性的，则得到*线性链条件随机场*。
- CRF思想的主要来源是最大熵模型，模型的三个基本问题的解都用到了HMM中提到的前向-后向法和韦特比算法。
- CRF是判别式无向图模型，可直接构建结构化输入输出特征之间的关系，是一种用来标记和切分序列化数据的统计模型，在序列数据和图像数据相关的结构化预测任务中表现优异，广泛应用于NLP中的词性标注和命名实体识别，图像分割，目标识别和动作识别等人工智能领域。

整体评价：

优点：条件随机场模型既具有判别式模型的优点，又具有产生式模型考虑到上下文标记间的转移概率，以序列化形式进行全局参数优化和解码的特点，解决了其他判别式模型(如最大熵马尔科夫模型)难以避免的标记偏见问题。

缺点：模型训练时收敛速度比较慢

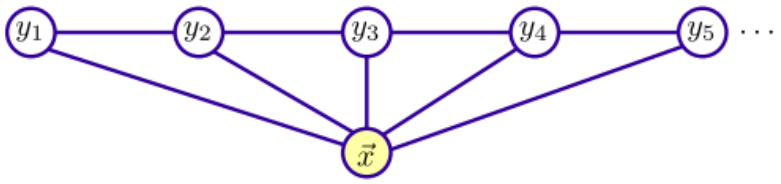
CRF研究方向：

1. 复杂拓扑结构的CRF

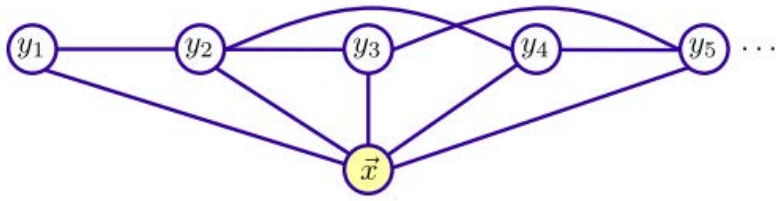
（skip-CRFs ，层叠CRFs）

2. 模型训练和推断的快速算法

3. CRF模型特征的选择和归纳



(a) Linear Chain

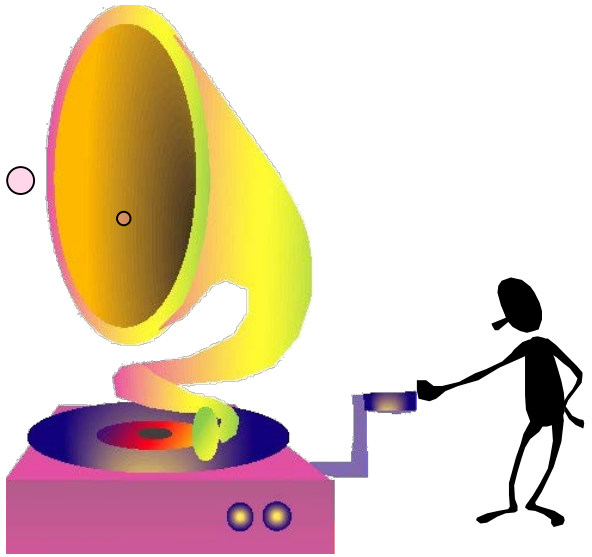


(b) Skip Chain

参考文献：

- An Introduction to Conditional Random Fields for Relational Learning
- Conditional Random Fields An Introduction
- Conditional Random Fields for Activity Recognition
- Conditional random fields Probabilistic models for segmenting and labeling sequence data
- 条件随机场综述
- 基于条件随机场的古文自动断句与标点方法

谢谢大家！



本课件制作过程中，多处引用了国内外同行的网页、教材、以及课件PPT的内容或图片，没有随处标注，特此说明，并在此向各位作者表示感谢！

比谁更牛