第14讲 条件随机场

余正涛 郭剑毅 李玉惠 李勇 苏磊 范玉刚等

昆明理工大学
信息工程与自动化学院

大纲

- 产生式模型和判别式模型
- 概率图模型
- 条件随机场理论
- 条件随机场模型应用示例
- 条件随机场工具包使用
- 条件随机场研究进展

产生式模型和判别式模型

- o和s分别代表观察序列和标记序列
- 产生式模型：构建o和s的联合分布p(s,o)，因可以根据联合概率来生成样本，如HMM，BNs，MRF。
- 判别式模型：构建o和s的条件分布p(s|o)，因为没有s的知识，无法生成样本，只能判断分类，如SVM，CRF，MEMM 。
- 产生式模型：无穷样本 ==》 概率密度模型 = 产生模型 ==》预测
- 判别式模型：有限样本 ==》 判别函数 = 预测模型 ==》预测

举例

- (1,0), (1,0), (2,0), (2, 1)
- 产生式模型：
  P (x, y)：
  P(1, 0) = 1/2, P(1, 1) = 0, P(2, 0) = 1/4,
  P(2, 1) = 1/4.
- 判别式模型：
  P (y | x)：
  P(0|1) = 1, P(1|1) = 0, P(0|2) = 1/2, P(1|2) = 1/2

两种模型比较

- Generative model ：从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度，不关心判别边界。
- 优点:
  - 实际上带的信息要比判别模型丰富， 研究单类问题比判别模型灵活性强
  - 能更充分的利用先验知识
  - 模型可以通过增量学习得到
- 缺点:
  - 学习过程比较复杂
  - 在目标分类问题中易产生较大的错误率

两种模型比较

- Discriminative model：寻找不同类别之间的最优分类面，反映的是异类数据之间的差异。
- 优点：
  - 分类边界更灵活，比使用纯概率方法或生产模型得到的更高级。
  - 能清晰的分辨出多类或某一类与其他类之间的差异特征
  - 在聚类、viewpoint changes, partial occlusion and scale variations中的效果较好
  - 适用于较多类别的识别
- 缺点：
  - 不能反映训练数据本身的特性。
  - 能力有限，可以告诉你的是1还是2，但没有办法把整个场景描述出来。
- 二者关系：由生成模型可以得到判别模型，但由判别模型得不到生成模型。

概率图模型

- 基本思想
- 朴素贝叶斯分类器
- 隐马尔可夫模型
- 最大熵模型
- 最大熵马尔可夫模型

概率图模型基本思想

- 概率图模型：是一类用图的形式表示随机变量之间条件依赖关系的概率模型，

  是概率论与图论的结合。图中的节点表示随机变量，缺少边表示条件独立假设。

  $G=(V,E)$

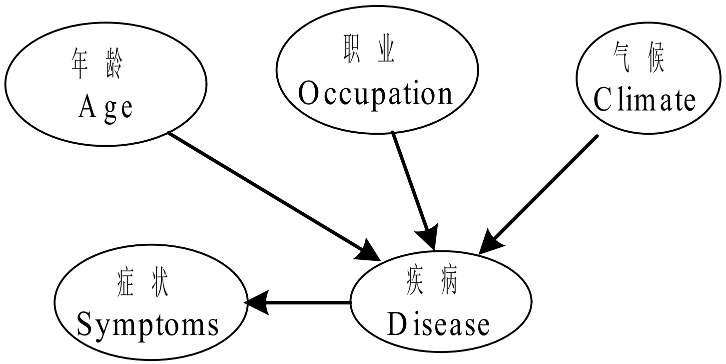
  V：顶点/节点，表示随机变量

  E：边/弧

- 两个节点邻接：两个节点之间存在边，记为 $X_i \sim X_j$，不存在边，表 示条件独立
- 路径：若对每个 I, 都有 $X_{i-1} \sim X_i$ $X_1,\ldots,X_N$ 则称序列为一条路径。

概率图模型基本思想

- 根据图中边有无方向，常用的概率图模型分为两类：
  - 有向图：最基本的是贝叶斯网络(Bayesian Networks ,BNs)
  - 举例

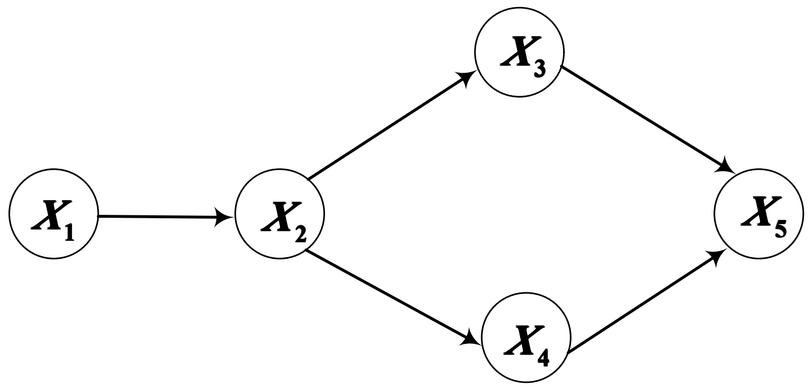


$$P(A,O,C,D,S|M)=$$

$$P(A|M)P(O|M)P(C|M)P(D|A,O,C,M)P(S|D,M)$$

概率图模型基本思想

- 有向图模型的联合概率分解：
  - 每个节点的条件概率分布表示为：
    P(当前节点|它的父节点)
  - 联合分布：



$$P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} p(X_i | \pi(X_i))$$

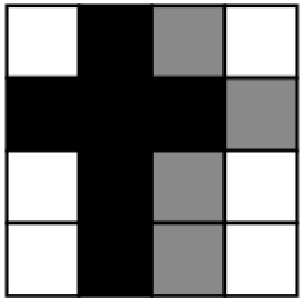
$$P(X_1, X_2, \cdots, X_5) = p(X_1)p(X_2|X_1)p(X_3|X_2)p(X_4|X_2)p(X_5|X_3X_4)$$

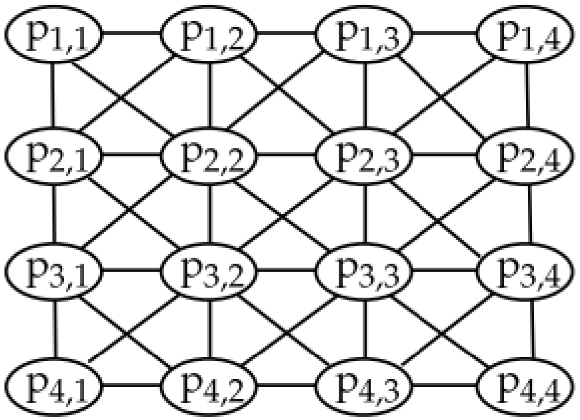
概率图模型基本思想

- 无向图：马尔可夫随机场(*Markov Random Fields*, MRF)

  马尔可夫随机场模型中包含了一组具有马尔可夫性质的随机变量，这些变量之间的关系用无向图来表示

  - 马尔科夫性： $$p(x_i \mid x_j, j \neq i) = p\left(x_i \mid x_j, x_i \sim x_j\right)$$
  - 举例





有向图模型和无向图模型的对比

- 共同之处

将复杂的联合分布分解为多个因子的乘积

- 不同之处

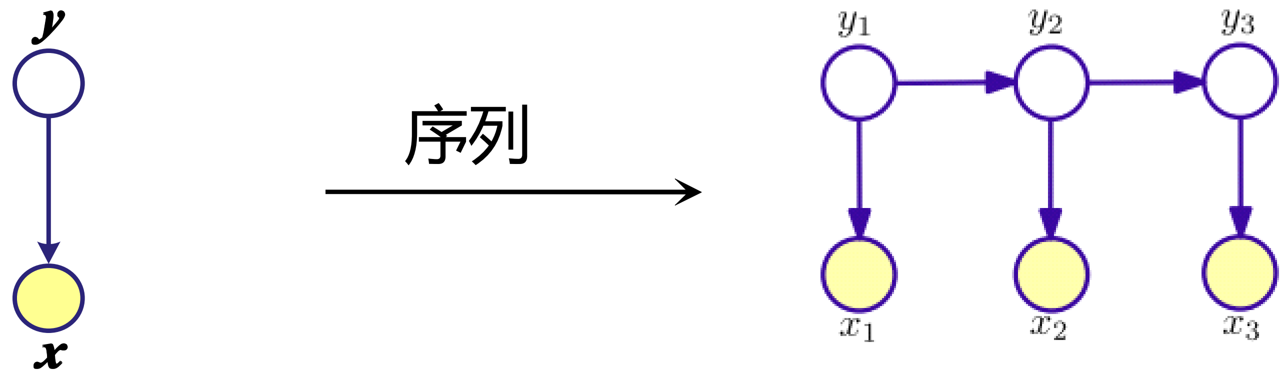
无向图模型因子是势函数，需要全局归一

有向图模型因子是概率分布、无需全局归一

- 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

有向图模型和无向图模型的对比



NBs

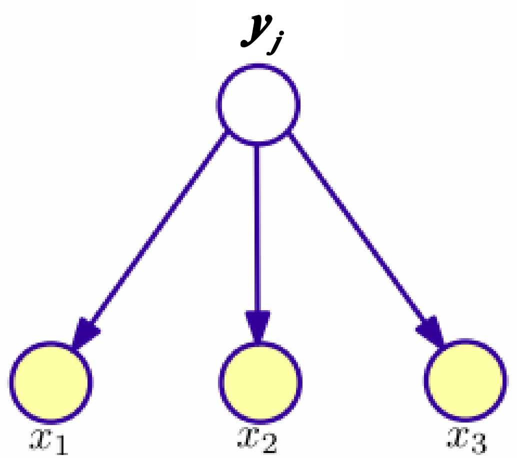
HMMs

MEMs

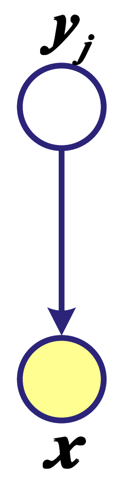
朴素贝叶斯分类器

- 设$x \in \Omega$是一个类别未知的数据样本，Y为类别集合，若数据样本x属于一个特定的类别y_j，那么分类问题就是决定$P(y_j|x)$，即在获得数据样本x时，确定x的最佳分类。所谓最佳分类，一种办法是把它定义为在给定数据集中不同类别y_j先验概率的条件下最可能的分类。贝叶 斯理论提供了计算这种可能性的一种直接方法。

朴素贝叶斯分类器的概率图表示



(a) Independency graph



$$P(x_1,x_2,x_3,y_j)=p(y_j)p(x_1|y_j)p(x_2|y_j)p(x_3|y_j)$$

隐马尔可夫概念(HMM)

- HMM是一种随机过程，它用概率统计的方法来描述语音信号的变化过程。
- HMM与通常的Markov链的不同之处在于其观察结果不是与状态有确定的对应关系，而是系统所处状态的概率函数，所以模型本身是隐藏的，它与观察结果之间还有一层随机的关系。
- HMM是对语音信号的时间序列结构建立统计模型，将之看做一个数学上的双重随机过程：
  - 一个是用具有有限状态的Markov链来模拟语音信号统计特性变化的隐含随机过程，
  - 另一个是与Markov链的每一个状态相关联的观测序列的随机过程。前者通过后者表现出来，但前者的具体参数是不可测的。

隐马尔可夫概念(HMM)

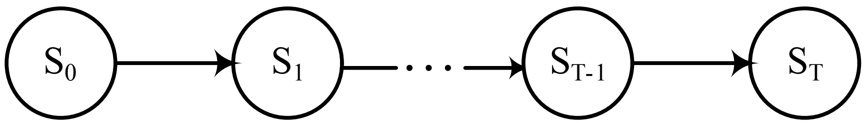
- HMM的状态是不确定或不可见的，只有通过观测序列的随机过程才能表现出来
- 观察到的事件与状态并不是一一对应，而是通过一组概率分布相联系
- HMM是一个双重随机过程，两个组成部分：
  - 马尔可夫链：描述状态的转移，用转移概率描述。
  - 一般随机过程：描述状态与观察序列间的关系，用观察值概率描述。

隐马尔可夫模型

- 马尔可夫模型：是一个三元组 $\lambda=(S, \Pi, A)$

  其中 S是状态的集合，Π是初始状态的概率，A是状态间的转移概率。

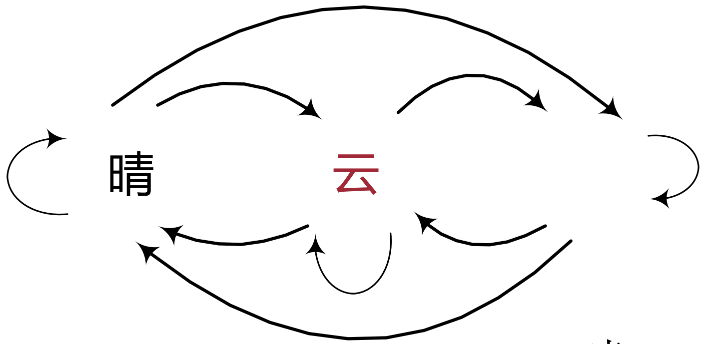
一阶马尔可夫链



$$\boxed{p(S_0, S_1, \ldots, S_T) = \prod_{t=1}^{T} p(S_t|S_{t-1})\ p(S_0)}$$

隐马尔可夫模型

- 一阶马尔可夫模型的例子



$$
\begin{array}{cc}
 & \textit{today} \\
 & \begin{array}{ccc} \textit{sun} & \textit{cloud} & \textit{rain} \end{array} \\
\begin{array}{r} \textit{yesterday}\ \ \textit{sun} \\ \textit{cloud} \\ \textit{rain} \end{array} &
\begin{bmatrix} 0.50 & 0.375 & 0.125 \\ 0.25 & 0.125 & 0.625 \\ 0.25 & 0.375 & 0.375 \end{bmatrix}
\end{array}
$$

晴　云　雨

$$S=\{s_1, s_2, s_3\}$$

$$\pi = (1,\ 0,\ 0)$$

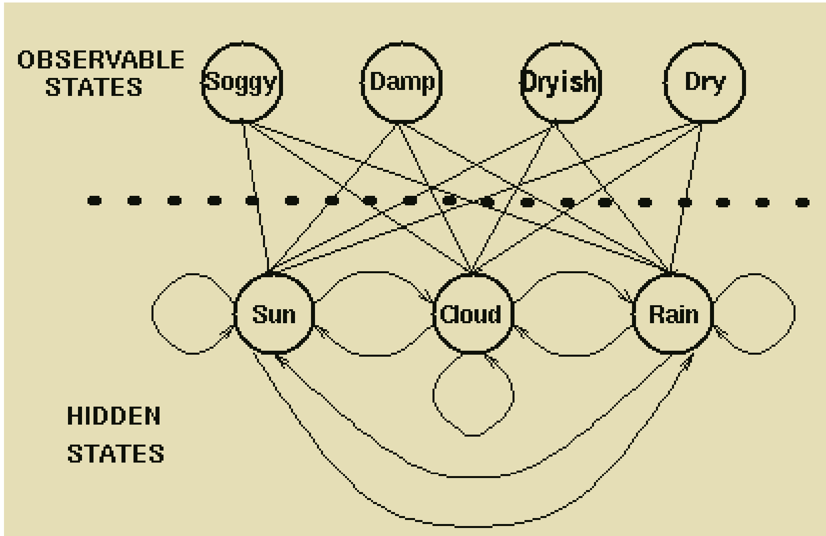
- 问题：假设今天是晴天，请问未来三天的天气呈现云雨晴的概率是多少？

隐马尔可夫模型

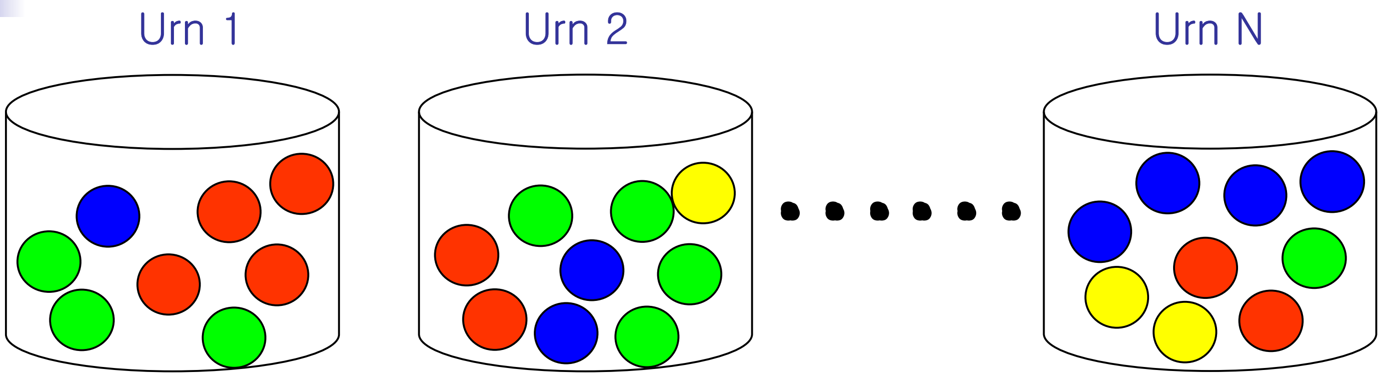
- HMM是一个五元组 λ = （Y, X, Π, A, B），其中 Y是隐状态（输出变量）的集合，）X是观察值（输入）集合， Π是初始状态的概率，A是状态转移概率矩阵，B是输出观察值概率矩阵。

$$
\begin{array}{ll}
 & \textit{today} \\
 & \begin{array}{ccc} \textit{sun} & \textit{cloud} & \textit{rain} \end{array} \\
\textit{yesterday} \begin{array}{r} \textit{sun} \\ \textit{cloud} \\ \textit{rain} \end{array} & \begin{bmatrix} 0.50 & 0.375 & 0.125 \\ 0.25 & 0.125 & 0.625 \\ 0.25 & 0.375 & 0.375 \end{bmatrix}
\end{array}
$$

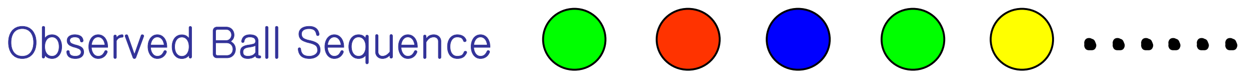
$$
\begin{array}{ll}
 & \begin{array}{cccc} \textit{soggy} & \textit{damp} & \textit{dryish} & \textit{dry} \end{array} \\
\begin{array}{r} \textit{sun} \\ \textit{cloud} \\ \textit{rain} \end{array} & \begin{bmatrix} 0.05 & 0.15 & 0.20 & 0.60 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.35 & 0.10 & 0.05 \end{bmatrix}
\end{array}
$$



HMM实例

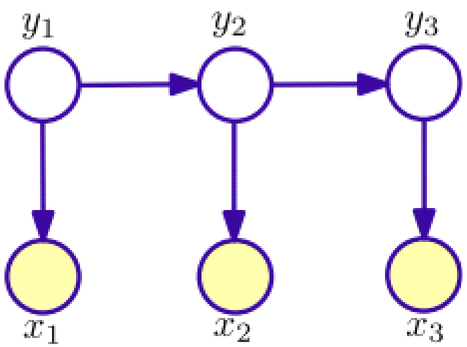


- 实验进行方式如下：
  - 根据初始概率分布，随机选择N个缸中的一个开始实验
  - 根据缸中球颜色的概率分布，随机选择一个球，记球的颜色为x1，并把球放回缸中
  - 根据缸的转移概率分布，随机选择下一口缸，重复以上步骤。
- 最后得到一个描述球的颜色的序列x1, x2, …称为观察值序列X。



- HMMs等生产式模型存在的问题：

$$P(X)=\sum_{\text{所有的Y}}\prod_{i=1}^{T}p(y_i|y_{i-1})p(x_i|y_i)$$



- 由于生成模型定义的是联合概率，必须列举所有观察序列的可能值，这对多数领域来说是比较困难的。

- 基于观察序列中的每个元素都相互条件独立。即在任何时刻观察值仅仅与状态（即要标注的标签）有关。对于简单的数据集，这个假设倒是合理。但大多数现实世界中的真实观察序列是由多个相互作用的特征和观察序列中较长范围内的元素之间的依赖而形成的。

最大熵模型

- 最大熵模型主要是在已有的一些限制条件下估计未知的概率分布。

  熵的计算公式： $H(X) = -\sum_{x \in X} p(x)\log p(x)$

  熵的性质： $0 \le H(X) \le \log|X|$

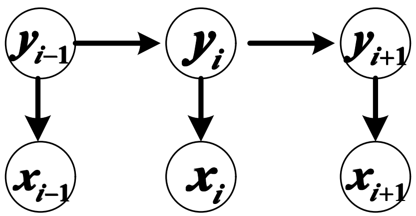
  - 其中X在离散分布时是随机变量的个数；
  - 当X为确定值，即没有变化的可能时，左边等式成立；
  - 可以证明，当X服从均匀分布时，右边等式成立，即均匀分布时熵最大。

- 最大熵的原理认为，从不完整的信息（例如有限数量的训练数据）推导出的唯一合理的概率分布应该在满足这些信息提供的约束条件下拥有最大熵值。求解这样的分布是一个典型的约束优化问题。

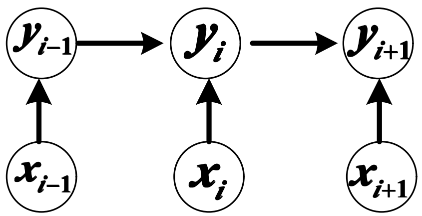
最大熵马尔科夫模型

- HMM：状态集合Y，观察值集合X， 两个状态转移概率：从y_{i-1}到y_i的条件概率分布$P(y_i \mid y_{i-1})$，状态y_i的输出观察值概率$P(x_i \mid y_i)$，初始概率$P_0(y)$.
- MEMM：用一个$P(y_i \mid y_{i-1}, x_i)$分布来替代HMM中的两个条件概率分布，它表示从先前状态，在观察值下得到当前状态的概率，即根据前一状态和当前观察预测当前状态。每个这样的分布函数都是一个服从最大熵的指数模型。

$$p_{y_{i-1}}(y_i \mid x_i) = \frac{1}{Z(x_i, y_{i-1})} \exp\left\{ \sum_a \lambda_a f_a(x_i, y_i) \right\} \quad i = 1, 2, \cdots, T$$



HMM



MEMM

最大熵马尔科夫模型

- 参数学习
  - 目的：通过学习 λ_a 使得MEMM中的每个转换函数达到最大熵。
  - GIS（Generalized Iterative Scaling）算法
- 编码问题
  - Viterbi算法的思想

最大熵马尔科夫模型举例

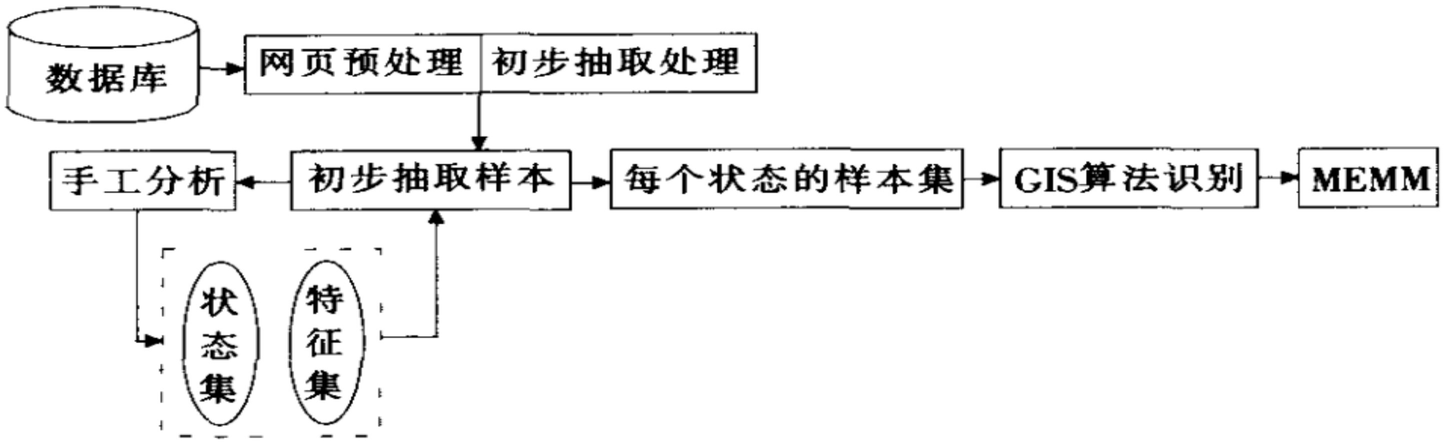
- 基于文本的网络地址信息抽取

<P>公司:青岛银河钢木家具厂</P><P>地址:青岛市重庆南路 247 号</P>
<P>电话:0532-5032359</P><P>传真:0532-5032263</P>
<P>Email:<A HREF="mailto:daiguoli@yhfnt.com">daiguoli@yhfnt.com</FONT></A></P>

<div align="center">地址:胶南市泰薛路中端 邮编:266401
电话:0532-6151696 传真:0532-6151333 手机:13589367999
<div>

- 任务：完成地址，电话，传真，E-mail 等信息的识别和抽取

最大熵马尔科夫模型举例



页面预处理

流程图

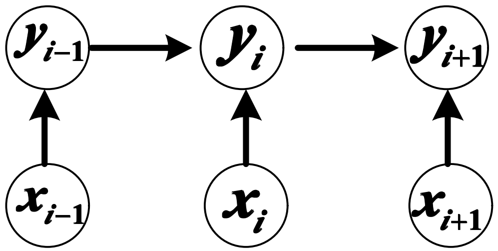
#地址:胶南市泰薛路中端# 邮编:266401#电话:0532-6151696#
传真:0532-6151333# 手机:13589367999#http://www.hengrun-qd.cn#
E-mail:qdjntlx@sohu.com#hengrun@hengrun-qd.cn#

页面文本中加入#用于保留结构信息和页面内容的自然划分，便于对文本页面的进一步处理。

模型建立

- 确定状态集合Y ，观察值（特征）集合 X
- 状态集合包含：邮编、电话、电邮、地址、联系人、账号、手机、网址、传真，对于其他可能出现的状态定义了“other”来代表。
- 特征集合包含： “具有@符号”
  “最大数字串长度为6”
  “最大数字串长度为11”
  “最大数字串长度介于6到11”
  “最大数字长度大于15”
  “最大数字长度小于6，字符串总长度大于30”
  “最大数字长度小于6，字符串总长度介于8到30”
  “最大数字长度小于6， 字符串总长度小于6”，……

模型建立



$$p_{y_{i-1}}(y_i|x_i)=\frac{1}{Z(x_i,y_{i-1})}\exp\left\{\sum_a \lambda_a f_a(x_i,y_i)\right\}$$

- 特征函数$f_a(x,y)$ 表示数据集<X， Y>的特性：

$$f_a(x,y)=\begin{cases}1 & \text{如果x只含有6位数字\&y=邮编}\\ 0 & \text{其他}\end{cases}$$

进一步引入一系列的特征函数 $\{f_1,f_2,\cdots f_n\}$

模型建立

- 参数学习

$$p_{y_{i-1}}(y_i|x_i)=\frac{1}{Z(x_i,y_{i-1})}\exp\left\{\sum_a \lambda_a f_a(x_i,y_i)\right\}$$

  - 用上述的状态和特征集对初步抽取样本进行统计，得到每个状态所对应的样本集，通过对于每个这样的样本集合采用 GIS算法进行参数学习，最终得到 MEMM。

- **说明：**

GIS算法要求对于每一个<x, y>,特征之和达到一个常数C，即有

$$\sum_{i=1}^{n} f_i(x,y)=C$$

如果不满足，则令 $$C=\max_{<x,y>}\sum_{i=1}^{n} f_i(x,y)$$

并加入一个修正函数，使得 $$f_{n+1}(x,y)=C-\sum_{i=1}^{n} f_i(x,y)$$

模型建立

- 评测指标

$$召回率（Recall）= \frac{正确识别出的实体个数}{标准结果中实体的总数} \times 100\%$$

$$精确率（Precision）= \frac{正确识别出的实体个数}{识别出的实体总数} \times 100\%$$

- 关键：特征的选择

标记偏见问题

最大熵马尔科夫模型与 HMMs 不同，它不是一个生产模型，而是一个基于下状态分类器的有限状态模型，但是，它却存在一个缺点就是标记偏见问题(label bias problem)，如图 1 所示。

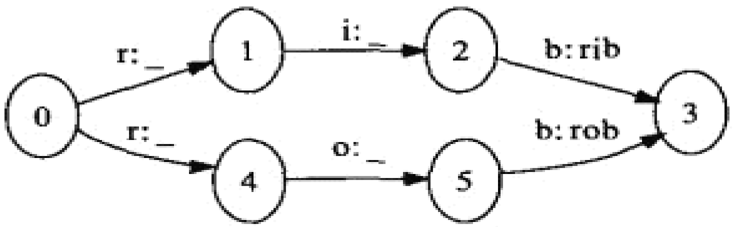
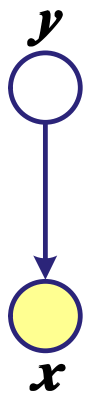


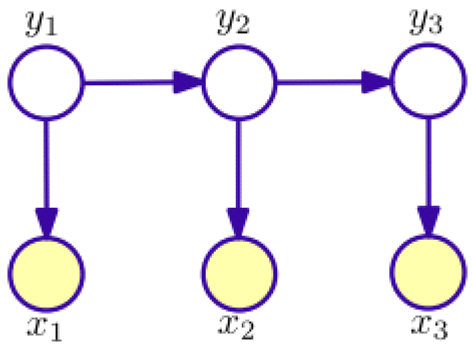
图 1　标记偏见问题

图 1 为一个简单的有限状态机用来区别单词 rob 和 rib。从状态 0 到 1 或 4 的概率相同(他们面对相同的观测 r)从状态 1 到 2 以及从状态 4 到 5 均只有惟一的转换，故而这个有限状态机无法区分 rob 和 rib。产生这种问题的原因就在于 MEMMs 对于状态序列的计算是局部的。而 CRFs 则是对状态序列进行全局的计算，从而克服了标记偏见问题。

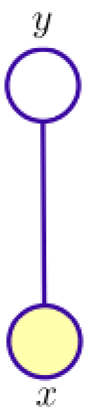


NBs

序列 →

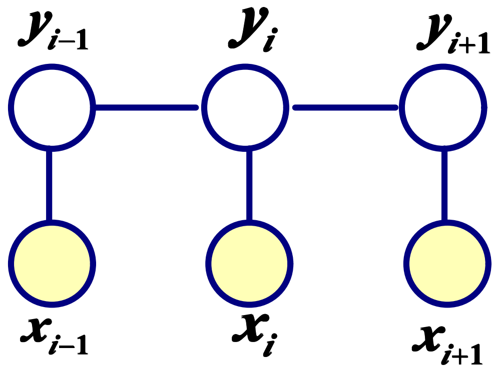


HMMs



MEMs

序列 →



linear-chain CRF

条件随机场理论

- 概述
- 定义
- Linear-chain CRFs 模型
- 特征函数选择
- 参数估计
- 模型推断
- 条件随机场特点

条件随机场概述

- 条件随机场模型是Lafferty于2001年，在最大熵模型和隐马尔科夫模型的基础上，提出的一种判别式概率无向图学习模型，是一种用于标注和切分有序数据的条件概率模型。
- CRF最早是针对序列数据分析提出的，现已成功应用于自然语言处理(Natural Language Processing，NLP）、生物信息学、机器视觉及网络智能等领域。

条件随机场概述

- 序列标注

  - 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　……
    - 观察序列：今天天气非常好！

汉语词性标注

条件随机场概述

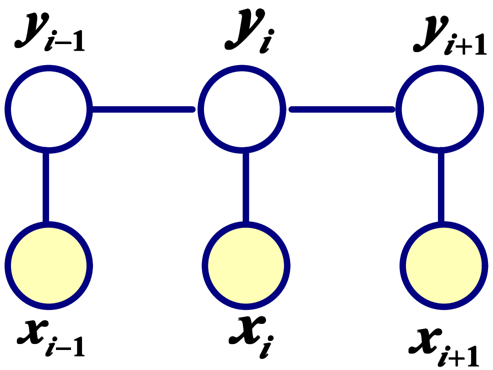
- 简单地讲，随机场可以看成是一组随机变量的集合（这组随机变量对应同一个样本空间）。当给每一个位置按照某种分布随机赋予一个值之后，其全体就叫做随机场。当然，这些随机变量之间可能有依赖关系，一般来说，也只有当这些变量之间有依赖关系的时候，我们将其单独拿出来看成一个随机场才有实际意义。
- 马尔科夫随机场（MRF）对应一个无向图。这个无向图上的每一个节点对应一个随机变量，节点之间的边表示节点对应的随机变量之间有概率依赖关系。因此，MRF的结构本质上反应了我们的先验知识——哪些变量之间有依赖关系需要考虑，而哪些可以忽略。具有马尔科夫性质：离当前因素比较遥远（这个遥远要根据具体情况自己定义）的因素对当前因素的性质影响不大。

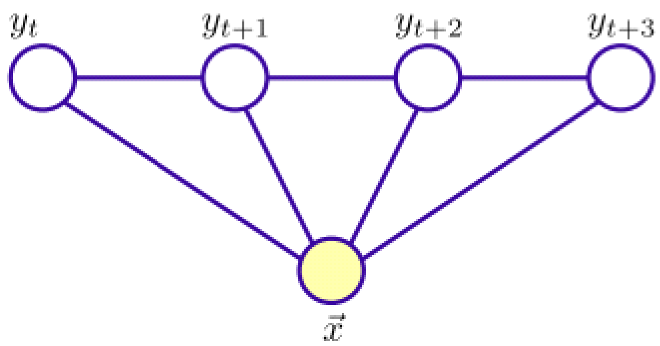
条件随机场概述

- 现在，如果给定的MRF中每个随机变量下面还有观察值，我们要确定的是给定观察集合下，这个MRF的分布，也就是条件分布，那么这个MRF就称为CRF。它的条件分布形式完全类似于MRF的分布形式，只不过多了一个观察集合x。最通用角度来看，CRF本质上是给定了观察值 (observations)集合的MRF

CRF定义：

- 设G=（V，E）是一个无向图，$Y=\{Y_v|v\in V\}$ 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 $p(Y_v|X,Y_u,u\neq v)=p(Y_v|X,Y_u,u\sim v)$， 则 (X,Y) 就构成一个条件随机场。
- 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，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：$f_j(y_{i-1},y_i,x,i)$

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

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

- 关键问题
  - 特征函数的选择

    特征函数的选取直接关系模型的性能。

  - 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。

  - 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。

特征函数的选择

- CRFs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。

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

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

参数估计

- 极大似然估计（Maximum Likelihood Estimation, MLE）

假定对于训练数据有一组样本集合 $D=\left\{x^{(j)},y^{(j)}\right\},\forall j=1,\cdots,N,$

样本是相互独立的， $\tilde{p}(x,y)$ 为训练样本中(x,y)的经验概率，

对于某个条件模型 $p\left(y\middle|x,\Theta\right)$ ，训练数据D的似然函数公式为：

$$L(\Theta)=\prod_{x,y}p(y|x,\Theta)^{\tilde{p}(x,y)}$$

取对数形式：

$$L(\Theta)=\sum_{x,y}\tilde{p}(x,y)\log p(y|x,\Theta)$$

参数估计

- CRFs模型中极大似然函数：

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

$$L(\lambda)=\sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n}\lambda f-\sum_{x}\tilde{p}(x)\log Z(x)$$

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

参数估计

- 迭代缩放

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）

- 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。

参数估计

- 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
- 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

  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：

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

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

参数估计

- 引入辅助函数：

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

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

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

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。

参数估计

- 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。

参数估计

- GIS算法

GIS是迭代缩放的一种，为了确保参数收敛的结果达到全局最优，GIS需要对特征集进行约束，即令每个训练数据中的事件 $T(x,y)=C$ 。

定义了一个全局修正特征S(x, y)：

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

其中C是训练语料中所有的x和y情况下T(x, y)的最大值，即等于最大可能的特征个数，特征S(x, y)的加入确保了T(x, y)=C。

假定对于所有的事件，条件随机场选定的特征的总和是常量C。

更新值按下式计算

$$\delta\lambda_j=\frac{1}{C}\log\left(\frac{E_{\tilde{p}(x,y)}\left[f_k\right]}{E_{p(y|x,\lambda)}\left[f_k\right]}\right)$$

参数估计

$$\delta\lambda_j = \frac{1}{C}\log\left(\frac{E_{\tilde{p}(x,y)}\left[f_k\right]}{E_{p(y|x,\lambda)}\left[f_k\right]}\right)$$

$$E_{\tilde{p}(x,y)}\left[f_k\right] = \sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n} f_j(y_{i-1}, y_i, x)$$

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

- 问题：
  - GIS算法的收敛速度由计算更新值的步长确定。C值越大，步长越小，收敛速度就越慢；反之C值越小，步长越大，收敛的速度也就越快。
  - GIS算法是依赖于一个额外的全局修正特征S(x,y)，以确保对于每个(x,y)对的有效特征的总和是一个常量。但是一旦加入这个新的特征，就认为这个特征和特征集中所有其他的特征之间是相互独立的，并且它的参数也需要使用上式来更新。计算期望需要对所有可能的标记序列求和，这将是一个指数级的计算过程。

参数估计

- 迭代改进规模（IIS）算法

重新定义：$$T(x,y) \approx T(x) \triangleq \max_{y} T(x,y)$$

将每个对观察序列和标记序列对(x, y)起作用的特征值的和近似等于对于观察序列x的最大可能的观察特征的和

$$E_{p(y|x,\lambda)}\left[f_k\right] = \sum_{m=0}^{T_{\max}} a_{k,m} \exp(\delta\lambda_j)^m$$

$$a_{k,m} = \sum_{x} \tilde{p}(x) \sum_{y} p(y|x,\lambda) \sum_{i=1}^{n} f_k(y_{i-1}, y_i, x)\delta(m, T(x))$$

使用牛顿一拉夫森方法求解

参数估计

- 梯度算法
- L-BFGS算法：

$$\frac{\partial L(\lambda)}{\lambda_j} = E_{\tilde{p}(x,y)}\left[f_j(x,y)\right] - \sum_k E_{p(y|x^{(k)},\lambda)}\left[f_j(x^{(k)},y)\right] - \frac{\lambda_k}{\sigma^2}$$

Dong C. Liu and Jorge Nocedal : [On The Limited Memory BFGS Method For Large Scale Optimization]

Jorge Nocedal用Fortran语言实现了L-BFGS工具包来进行条件随机场的参数估计与训练，该数学工具包可从 http://www.ece.northwestern.edu/nocedal/下载。

另外，Taku Kudo实现了L-BFGS算法的c语言版本，该工具集成在了其开发的CRF++工具包中，网址为 http://www.chasen.org/taku/software/CRF++/。

模型推断

- 常见两个问题：一、在模型训练中，需要边际分布 $p(y_t, y_{t-1} | x)$ 和 $Z(x)$ ；二、对于未标记的序列，求其最可能的标记。

- 第一个问题采用前向后向法解决；
- 第二个问题通过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。

条件随机场特点

- 在建模时，实验者只需集中精力选择特征，并不需要花费精力考虑如何使用特征。
- 特征选择灵活，可使用不同类型的特征，特征容易更换，可以将观察到的各种相关或不相关的概率知识拟合到统一的一个框架之中。
- 不需要像其他产生式模型那样做独立性假设条件。
- 克服了判别模型普遍存在的标记偏见问题。
- 无需对每个知识源单独建模，可以直接构造一个联合了多知识源的混合模型。

条件随机场模型应用示例

- 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

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

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

条件随机场模型应用示例

- 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。

条件随机场模型应用示例

- "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：

条件随机场模型应用示例

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）

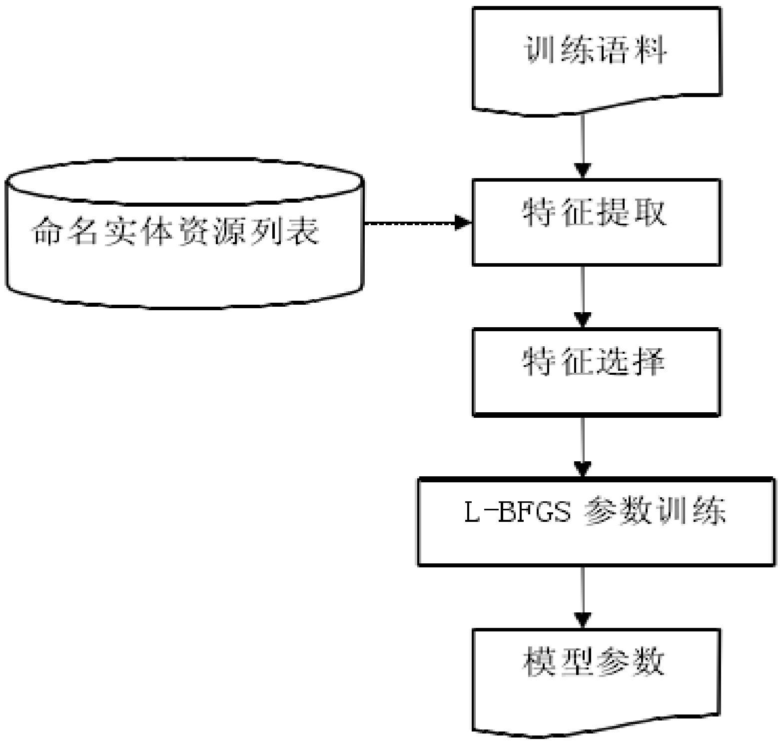
- 当特征函数取特定值时，特征模板被实例化就可以得到具体的特征。

“当前词的前一个词w_{-1}在人名1级左指界词列表中出现”

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

条件随机场模型应用示例

- 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

条件随机场模型应用示例

- 评测指标

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

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

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

条件随机场模型应用示例

- 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

条件随机场工具包使用

- 安装
- 训练语料的格式
- 特征模板的格式
- 训练模型
- 识别

安装

参考网站：

- http://crf.sourceforge.net/doc/
  - java写的CRF工具包，有很详细的说明文档。
- FlexCRFs: Flexible Conditional Random Fields
  - 作者:Xuan-Hieu Phan 和 Le-Minh Nguyen
  - C++语言实现
  - 下载地址：http://www.jaist.ac.jp/~hieuxuan/flexcrfs/flexcrfs.html

训练语料格式

- 训练和测试文件必须包含多个tokens
- 每个token包含多个列
- token的定义可根据具体的任务，如词、词性等
- 每个token必须写在一行，且各列之间用空格或制表格间隔
- 一个token的序列可构成一个sentence，sentence之间用一个空行间隔

特征模板的格式

- 模板的基本格式为*%x[row,col]*，它用于确定输入数据中的一token 其中，row确定与当前的token的相对行数。col用于确定绝对列数。

- 特征模板的例子
  - 训练语料

	col0	col1	col2	
	Input: Data			
r-2	He	PRP	B-NP	
r-1	reckons	VBZ	B-VP	
r0	the	DT	B-NP	<< 当前的token
r1	current	JJ	I-NP	
r2	account	NN	I-NP	

  - 特征模板

template	expanded feature
%x[0,0]	the
%x[0,1]	DT
%x[-1,0]	rokens
%x[-2,1]	PRP
%x[0,0]/%x[0,1]	the/DT
ABC%x[0,1]123	ABCthe123

训练模型

- 使用crf_learn命令
  - **crf_learn template_file train_file model_file**
    其中，template_file和train_file需由使用者事先准备好。crf_learn将生成训练后的模型并存放在model_file中。

识别

- 使用crf_test 命令
  - **crf_test –m model_file test_file**

    其中，model_file是crf_learn创建的。在测试过程中，使用者不需要指定template file，因为，mode file已经有了template的信息。test_file是使用者想要标注序列标记的测试语料。这个文件的书写格式应该与训练文件一致。

条件随机场研究进展

- 2001 年，卡耐基. 梅隆大学的 Lafferty 教授针对序列数据处理提出了 CRF 模型。[Conditional random fields- Probabilistic models for segmenting and labeling sequence data]
- 2003 年，Kumar 博士将 CRF 模型扩展到 2-维格型结构，开始将其引入到图像分析领域，吸引了学术界的高度关注。
  Ariadna Quattoni Michael Collins Trevor Darrell
  [Conditional Random Fields for Object Recognition]
- Asela Gunawardana等人
  [Hidden Conditional Random Fields for Phone Classification]
- 2007年，Charles Sutton，Andrew McCallum
  [Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data]

参考文献

- 基础的参考文献：
  - 概率图

    [An Introduction to Variational Methods for Graphical models]

    [Classical Probabilistic Models and Conditional Random Fields]

  - 经典概率模型与CRF

    [An Introduction to Conditional Random Fields for Relational Learning]

    [Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data]

    [Operations for learning with Graphical models]

- 有用的参考文献：

[Efficient Training of Conditional Random Fields]

[Efficiently Inducing features of random fields]

[A maximum entropy approach to natural language processing]

[Multiscale Conditional Random Fields for Image Labeling]

[Training Conditional Random Fields via Gradient Tree Boosting]

谢谢！