Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

基于LSTM-CRF的序列标注算法

硕士研究生　　尹继泽

2018年01月28日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

- 机器学习的任务
  - 分类：将数据划分为不同的类别，通常以“数据-标签”的形式展现。
  - 序列标注：为序列数据中的每一个值标注标签。
- 背景案例
  - 命名实体识别：从非结构化文本中抽取出特定类别的命名实体，比如人名、地名和组织机构名。
  - 命名实体：所有以名称为标识的实体。
  - I am Xiao Ming.
  - I-O, am-O, Xiao-PER.B, Ming-PER.I, .-O

背景简介

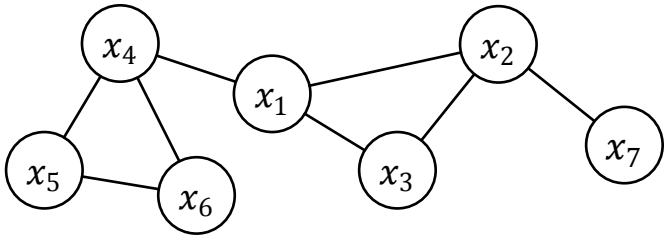
- LSTM-CRF模型
  - 自然语言处理中常用的序列标注模型
  - 深度神经网络和概率图模型的结合
  - 输入的序列数据经过深度神经网络LSTM的处理获得特征向量序列输出
  - 条件随机场CRF对LSTM的输出进行判别，确定每一个特征向量对应的标签值
  - LSTM 参见 胡雅娴《长短期记忆网络（LSTM）》

基本概念

- 概率模型
  - 提供一种描述框架，将学习任务归结于计算变量的概率分布。
- 推断
  - 在概率模型中，利用已知变量推测未知变量的[条件]分布。
- 生成模型
  - 能够随机生成观测数据的模型，尤其是在给定某些隐含参数的条件下，通常为一个联合概率分布。
  - 例如：$P(\text{回答}, \text{问题})$
- 判别模型
  - 根据不可观测数据对可观测数据的依赖而建立的模型，通常为一个条件概率分布。
  - 例如：$P(\text{标签}|\text{单词})$

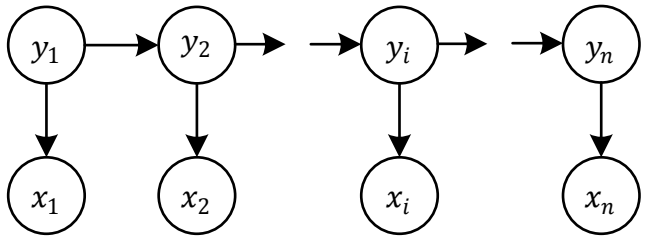
基本概念

- 概率图模型
  - 一类用图来表达变量相关关系的概率模型。
  - 以图为表示工具。
  - 一个结点表示一个或一组随机变量。
  - 结点之间的边表示变量间的概率相关关系。
  - 有向图模型/无向图模型



算法原理：隐马尔可夫模型HMM

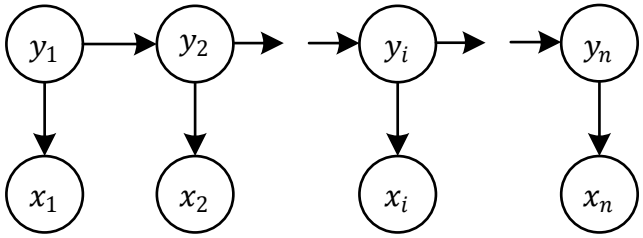
- 变量
  - 状态变量/隐变量
  - 观测变量
- 系统在多个状态间转换
- 马尔可夫链
  - 系统下一时刻的状态仅由当前状态决定，不依赖以往的任何状态。
- 参数
  - 状态转移概率
  - 输出观测概率
  - 初始状态概率
- 参见 王文浩《隐马尔可夫模型的三个基本问题》



$$P(x_1, y_1, \dots, x_n, y_n) = P(y_1)P(x_1|y_1)\prod_{i=2}^{n} P(y_i|y_{i-1})P(x_i|y_i)$$

算法原理：隐马尔可夫模型HMM

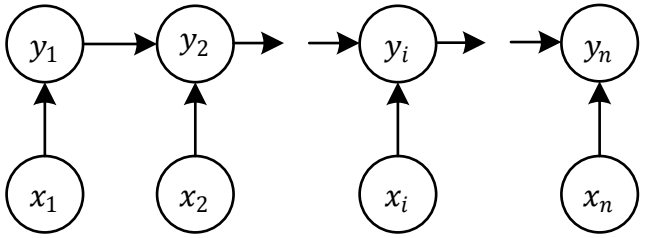
- 问题
  - 特征表示有限，且不能表示交叉的特征（输出独立性假设）
  - 常常通过生成式的联合概率模型解决条件概率问题



$$P(x_1, y_1, \ldots, x_n, y_n) = P(y_1)P(x_1|y_1)\prod_{i=2}^{n} P(y_i|y_{i-1})P(x_i|y_i)$$

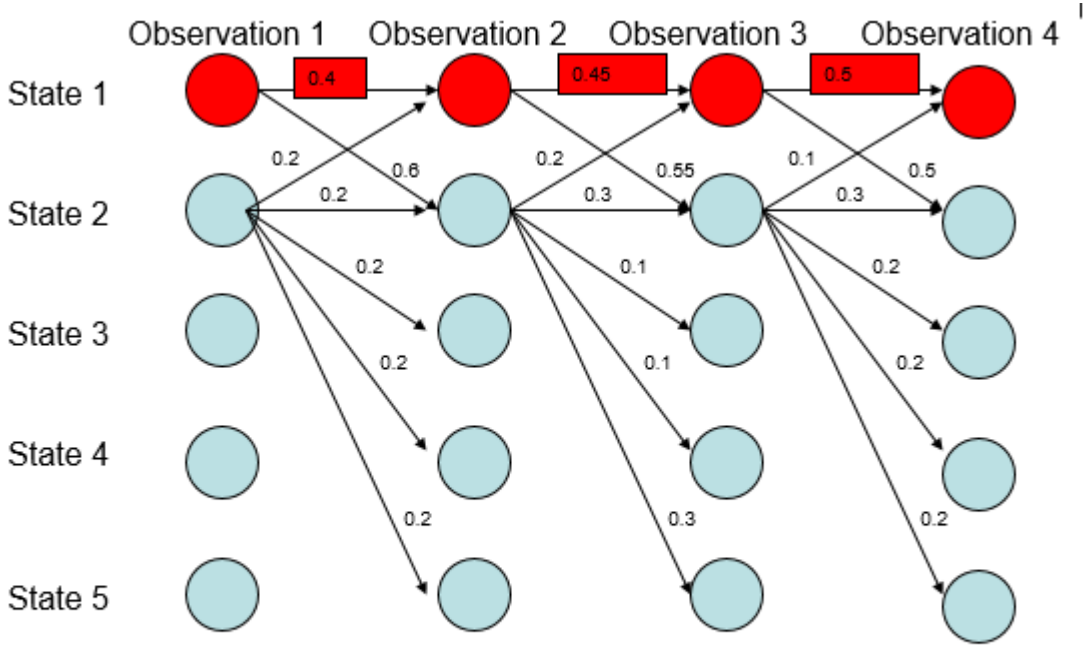
算法原理：最大熵马尔可夫模型MEMM

- 特点
  - $P(S_1 \cdots S_T | O_1 \cdots O_T) = \prod_{t=1}^{T} P(S_t | S_{t-1}, O_t)$
  - $P(s|s', o) = \frac{\exp(\sum_a \lambda_a f_a(o,s))}{Z(o,s')}$
    - $Z(o, s')$为归一化因子
    - $f_a(o, s)$为特征函数
    - $f_{\langle b,s \rangle}(o_t, s_t) = \begin{cases} 1 & \text{if } b(o_t) \text{ is true and } s = s_t \\ 0 & \text{otherwise} \end{cases}$
      - b表示特征，s表示目标状态
    - λ_a为特征函数的权重



算法原理：最大熵马尔可夫模型MEMM

- 标记偏置问题
  - 由于分支数不同，概率分布不均衡，导致状态的转移不公平

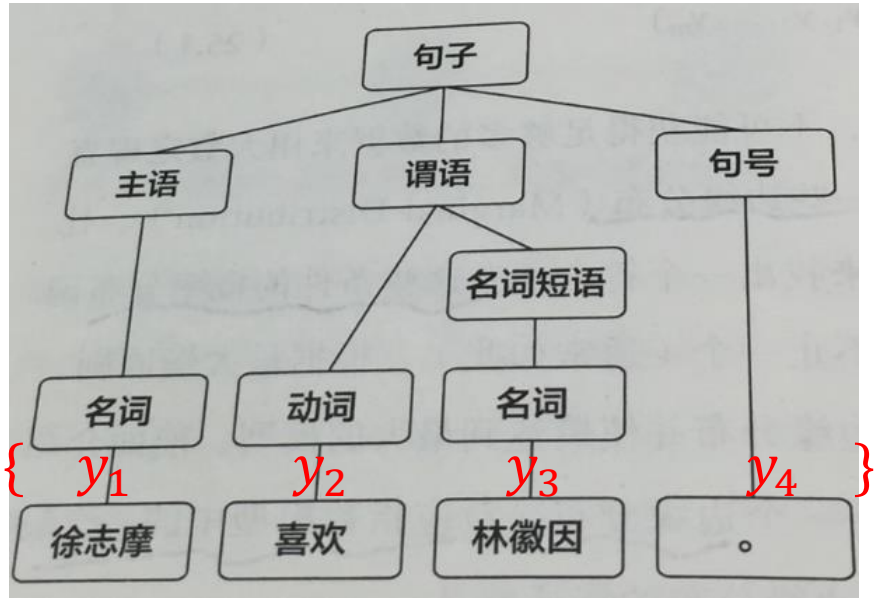


Most Likely Path: 1-> 1-> 1-> 1

- State 1 has only two transitions but state 2 has 5:
  - Average transition probability from state 2 is lower

算法原理：条件随机场CRF

- 变量
  - 观测变量$\boldsymbol{x}$
  - 标记变量$\boldsymbol{y}$
  - 标记变量可以是结构型变量，即其分量之间有某种相关性



$\boldsymbol{y}$ $\{ y_1 \quad y_2 \quad y_3 \quad y_4 \}$

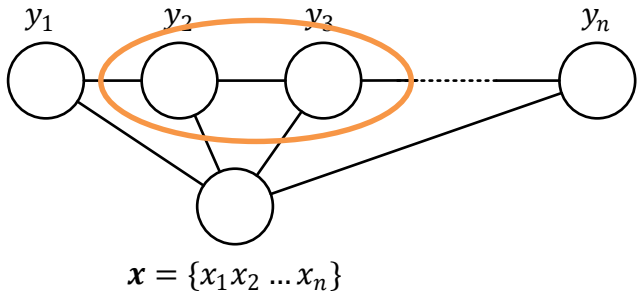
$\boldsymbol{x}$ $\{ x_1 \quad x_2 \quad x_3 \quad x_4 \}$

语法分析

词性标注

算法原理：条件随机场CRF

- 定义
  - 令$G = \langle V, E \rangle$表示结点与标记变量$\boldsymbol{y}$中元素一一对应的无向图，y_v表示与结点v对应的标记变量，$n(v)$表示结点v的邻接结点，若图G的每个变量y_v都满足马尔可夫性，即$P(y_v|\boldsymbol{x}, \boldsymbol{y}_{V\setminus\{v\}}) = P(y_v|\boldsymbol{x}, \boldsymbol{y}_{n(v)})$，则$(\boldsymbol{y}, \boldsymbol{x})$构成一个条件随机场。
  - 理论上图G可具有任意结构
  - 团：任意两结点间都有边连接的结点子集
  - 势函数：定义在变量子集上的非负实函数
- 链式条件随机场（chain-structured CRF）



算法原理：条件随机场CRF

- 构造条件概率
  - 选用指数势函数并引入特征函数
  - $P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_j \sum_{i=1}^{n-1} \lambda_j \, t_j(y_{i+1}, y_i, \boldsymbol{x}, i) + \right.$

算法原理：条件随机场CRF

- 矩阵形式
  - $\boldsymbol{M}_i(\boldsymbol{x}) = [\boldsymbol{M}_i(y_{i-1}, y_i|\boldsymbol{x})] = \begin{matrix} A & B & C \\ D & E & F \\ G & H & I \end{matrix}$
  - $\boldsymbol{M}_i(y_{i-1}, y_i|\boldsymbol{x}) = \exp\left(\boldsymbol{W}_i(y_{i-1}, y_i|\boldsymbol{x})\right)$
  - $\boldsymbol{W}_i(y_{i-1}, y_i|\boldsymbol{x}) = \sum_{j=1}^{J} \lambda_j f_j(y_{i-1}, y_i, \boldsymbol{x}, i)$
  - $P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\prod_{i=1}^{n} \boldsymbol{M}_i(y_{i-1}, y_i|\boldsymbol{x})$
  - $Z = \left(\boldsymbol{M}_1(\boldsymbol{x})\boldsymbol{M}_2(\boldsymbol{x})\cdots\boldsymbol{M}_{\boldsymbol{n}}(\boldsymbol{x})\right)_{start,stop}$
  - Z是以start为起点、stop为终点通过状态的所有路径的非规范化概率之和。

算法原理：条件随机场CRF

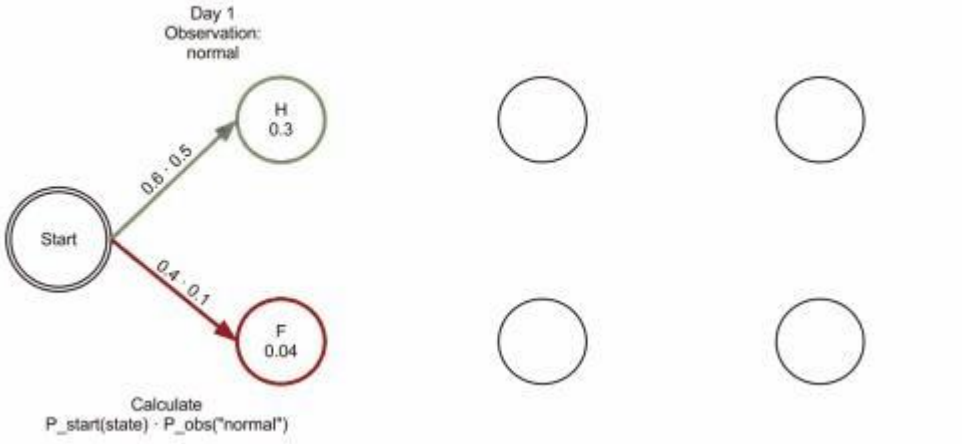
- 参数学习
  - 已知训练数据集，可知经验概率分布$\tilde{p}(\boldsymbol{x},\boldsymbol{y})$，通过最大化对数似然函数来求模型参数。
  - $L(w) = \log \prod_{x,y} P(\boldsymbol{y}|\boldsymbol{x})^{\tilde{p}(x,y)}$

    $= \sum_{x,y} \tilde{p}(\boldsymbol{x},\boldsymbol{y}) \log P(\boldsymbol{y}|\boldsymbol{x})$

    $= \sum_{k=1}^{K} \sum_{j} \sum_{i=1}^{n} \lambda_j f_j(y_{i-1}, y_i, \boldsymbol{x}, i) - \sum_{k=1}^{K} \log Z(\boldsymbol{x}_k)$
  - 参见 董思佳《深度学习优化算法概述》
  - 改进的迭代尺度法

算法原理：条件随机场CRF

- 推断
  - 给定条件随机场的条件概率计算方法和输入序列，求条件概率最大的输出序列，即对观测序列进行标注。
  - 使用维特比算法。
  - $\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \left(\sum_j \sum_{i=1}^{n} \lambda_j f_j(y_{i-1}, y_i, \boldsymbol{x}, i)\right)$
  - 推断问题转变为最优路径问题。



优劣分析

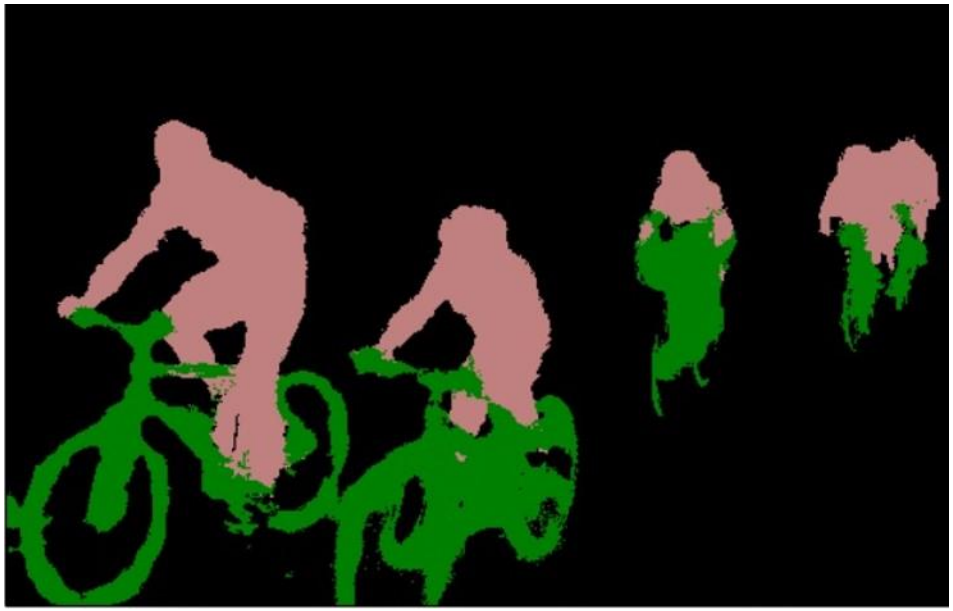
- 优势
  - 避免了HMM的输出独立性假设
  - 避免了MEMM的标记偏置问题
  - 可以结合多种特征
- 劣势
  - 受限于局部特征
  - 训练模型的时间更长
  - 模型规模大
  - 只能捕获一个或几个时间步的联系

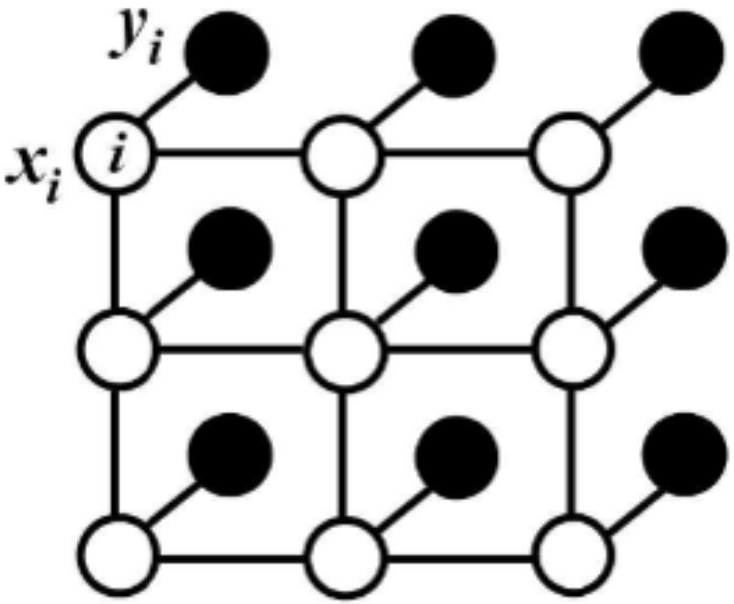
应用总结

- 算法的应用
  - 命名实体识别
  - 词性标注
  - 句法分析
- 未来的发展
  - 针对算法的不足之处进行改进
  - semi-Markov CRFs
  - higher-order CRFs
  - latent-dynamic CRFs
  - grSemi-CRFs

应用总结

- 全连接条件随机场（图像语义分割）





参考文献

[1] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016: 319-327.

[2] 宗成庆. 统计自然语言处理(第二版) [M]. 北京: 清华大学出版社, 2016: 125-128.

[3] Zhuo et al. Segment-Level Sequence Modeling using Gated Recursive Semi-Markov Conditional Random Fields [C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 1413-1423.

[4] 博客.最大熵马尔可夫模型MEMM [CP/OL]. [2016-12-20]. https://www.cnblogs.com/en-heng/p/6201893.html.

[5] Wikipedia. Conditional random field [CP/OL]. [2017-10-08]. https://en.wikipedia.org/wiki/Conditional_random_field.

参考文献

[6] 博客. 条件随机场 [CP/OL]. [2016-08-22].
https://www.hankcs.com/ml/conditional-random-field.html.
[7] 博客.图像语义分割之FCN和CRF [CP/OL]. [2016-09-04].
http://blog.csdn.net/u012759136/article/details/52434826.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。