**基于CRF实现中文分词和词性标注**

目录

[一、 预备知识 1](#_Toc46325662)

[1. 中文分词和词性标注 1](#_Toc46325663)

[2. CRF模型原理 1](#_Toc46325664)

[二、 实验过程和结果 4](#_Toc46325665)

[1. 实验环境 4](#_Toc46325666)

[2. 模型训练 4](#_Toc46325667)

[3. 预测算法和实验结果 5](#_Toc46325668)

[4. 实验结果 5](#_Toc46325669)

[三、 HMM和CRF原理对比 5](#_Toc46325670)

[四、 参考资料和数据集 6](#_Toc46325671)

[小结 6](#_Toc46325672)

[练习 6](#_Toc46325673)

# 预备知识

## 中文分词和词性标注

中文分词就是对汉字以字为单位进行处理。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 输入X | 我 | 去 | 北 | 京 |
| 分词Y | S | S | B | E |

S：single，单个词；B：begin，起始词；E：end，结束词。

词性标注是对汉字以词为单位进行标注。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 输入X | I | can | open | this | can | . |
| 词性标注Y | PN | MD | VV | PN | NN | PU |

PN：代词；MD：情态动词；VV：动词；NN：名词；PU：标点符号。

分词和词性标注两个任务同时进行可以减少错误传播。

## CRF模型原理

条件随机场（conditional random field）是一种用来标记和切分序列化数据的统计模型。在NLP领域可以用来做序列标注任务。

t

s

图中输入x为句子，输出y为句子中每个词的标注。

条件随机场模型引入状态特征和转移特征来表达语法；引入权值来表达语法重要性。引入指数函数来简化计算。



转移特征

状态特征

权值w=(λ,μ)

特征函数f

1. CRF模型输入和输出

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 输入 | 我 | 去 | 北 | 京 |
| 分词结果 | S | S | B | E |
| 词性标注结果 | PN | VV | NN | |

1. CRF模型训练过程
2. CRF模型训练参数

转移概率矩阵T：给定标签下另一标签出现的概率

观测概率矩阵S：某一标签下生成特定词的概率

初始概率分布π：初始时刻单词分布情况

1. 构建特征函数

构建特征函数所需特征模板如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 单个词 | 单个词性 | 相邻两个词 | 相邻两个词性 |
| word[-2,0] word[-1,0]  word[0,0] word[1,0] | pos[-2,1]pos[-1,1]  pos[0,1]pos[1,1] | word[-1,0]/word[0,0]  word[0,0]/word[1,0] | pos[-2,1]/pos[-1,1]  pos[-1,1]/pos[0,1] |

特征函数初始权重为函数值为1的概率。

1. 计算条件概率

引入前向-后向算法，计算条件概率和特征函数的数学期望。

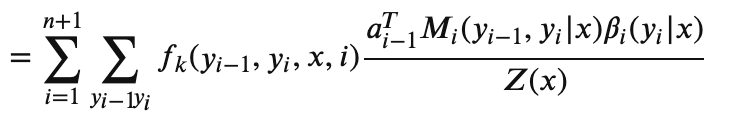
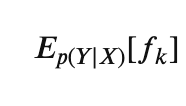
由此可以简化条件概率的计算。

前向算法：

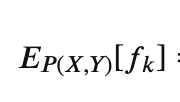
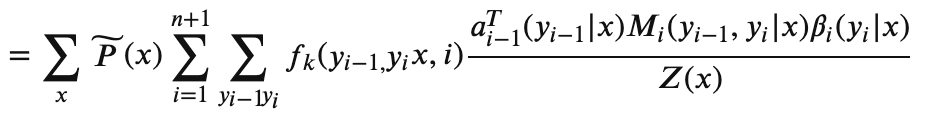
后向算法：

故P(X|λ=)=, t=1,2,…,T-1

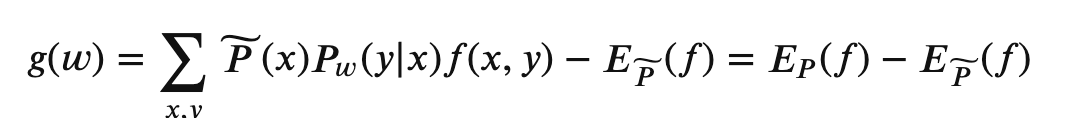
引入前后向算法之后，复杂度降为O(TN2)

特征函数的数学期望为：

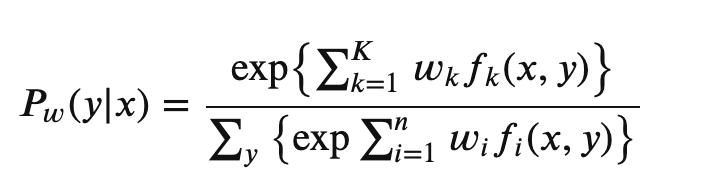
1. 梯度下降训练CRF模型



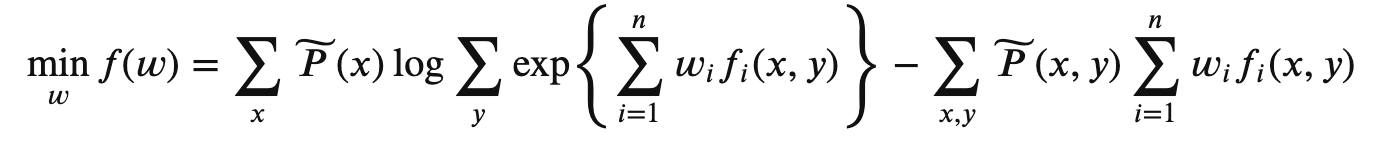
梯度函数



条件随机场模型



目标函数



在训练阶段使用拟牛顿法的l-bfgs算法，提升训练速度，降低时间和空间复杂度。

1. Viterbi算法解码



T=1也就是第一个单词的时候，用δ来记录这4个状态生成单词v1的概率，然后用ψ记录这4个δ中的最大值；

求解下一个单词的输出的时候，不需要管前一个单词的输出了，第一个单词的输出已经是最佳，所以这个问题就可以推广到求解t+1时的最佳输出。

在t+1时刻，依然用δ来记录从某个状态生成输入的概率值，用ψ来记录的概率δ的最大值。

# 实验过程和结果

## 实验环境

Windows: 训练阶段使用.bat文件

MacOS：训练阶段使用.sh文件

## 模型训练

本次实验以MacOS环境下的操作为例。

1. 用人民日报1998语料进行训练：train.py中的trainfile变量指定语料

sh run\_train.sh

1. 测试：使用一部分剩余料进行测试，test\_eva.py中的testfile变量指定语料

先进行中文分词，再进行词性标注

sh run\_test\_raw.sh

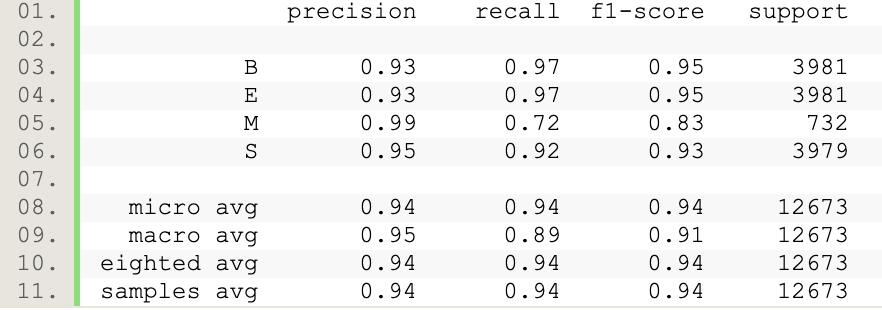
## 预测算法和实验结果

针对哈工大信息检索研究中心语料，使用Viterbi算法进行预测。

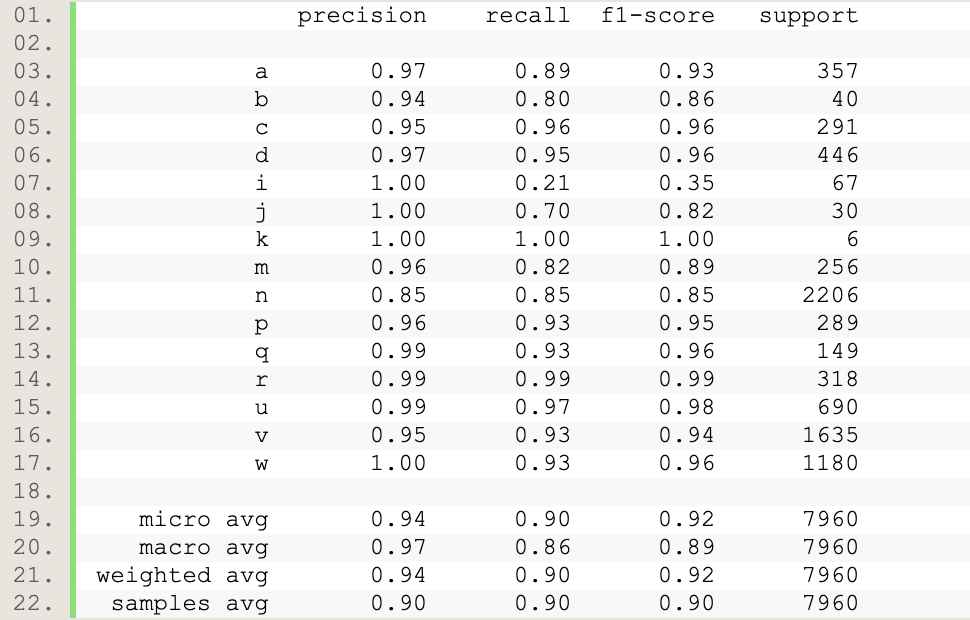
sh run\_test\_eva.sh

## 实验结果

分词结果



词性标注结果



# HMM和CRF原理对比

|  |  |  |
| --- | --- | --- |
| 模型 | CRF | HMM |
| 本质 | 生成式模型 | 判别式模型 |
| 图 | 概率无向图 | 概率有向图 |
| 求解过程 | 全局最优 | 局部最优 |
| 归一化 | 合理 | 标签偏移 |
| 特征 | 全局特征 | 局部特征 |

# 参考资料和实验源代码

[1] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[2] L. Rabiner and B. Juang, "An introduction to hidden Markov models," in IEEE ASSP Magazine, vol. 3, no. 1, pp. 4-16, Jan 1986, doi: 10.1109/MASSP.1986.1165342.

[3] <https://github.com/xjtuerz0/crf_seg_pos>

# 小结

条件随机场（conditional random field）是一种用来标记和切分序列化数据的统计模型。在NLP领域可以用来做序列标注任务。模型训练过程中涉及特征函数的构建、梯度下降法和viterbi算法的使用。本次实验得到的结果是被分割好的单词和已标注的中文词汇。

# 练习

1. CRF模型是什么图模型？
2. 概率有向图模型
3. 概率无向图模型
4. 有向图模型
5. 无向图模型

正确答案：B

1. CRF模型解码阶段一般使用什么算法？
2. Viterbi
3. Softmax
4. ReLu
5. SVM

正确答案：A

1. CRF模型训练阶段一般使用什么算法?

A. 拟牛顿法

B. 拟牛顿法的l-bfgs算法

C. Gradient Ascent

D. Xgboost

正确答案：B