**CRF词性标注学习笔记**

# 概述

CRF：Conditional Random Fields判别式模型，在观测序列的基础上对目标序列进行建模，重点解决序列标注问题。以词性标注为例，观测序列X就是分割好的语料，输出序列Y就是语料中的词对应的词性的组合,通过定义给定输入X条件下Y的概率，也就是条件概率P(Y|X)来描述模型，训练出一个由权重w和特征函数f组成的模型，将此模型用于解决语料的词性标注问题。

应用场景：主要自然语言处理中序列标注、数据分割、组块分析等任务。在中文分词、中文人名识别、歧义消除等任务中表现较好。

贝叶斯公式：

马尔科夫过程：状态、初始向量、转移过程；每一个状态的转移只依赖于其之前的那一个状态。

隐马模型：评价、解码、学习



下标t表示当前所在节点(token)的位置

下标k表示第k个特征函数，并且每个特征函数fk都有一个权重wk。

# 算法流程

1. 确定特征模板，产生特征函数fk，fk∈(0,1)，并且给每个特征函数fk随机赋一个初始权重w。w是由状态特征的权重λ和转移特征的权重η组成的列向量。在CRF模型中，每个节点的特征函数仅依赖当前节点和前一节点。
2. 在给定的观测数据上训练模型，优化条件概率p(y|x)，训练模型的特征函数的权重wk ，wk可以为正数（加分，靠谱）、负数（减分，不靠谱）或0（没有贡献）。
3. 用确定的模型做序列标注/序列求概率问题。在新的观测序列上找出一条概率最大的隐状态序列。

# 实例

**训练集：**假设n条有分割和标注的语料。语料长度为T。共5个标签。整个词典有D个不重复的词。如：他们(n)/是(v)/超级(adv)/勇敢的(adj)/中国人(n)。标签labels={n, v, adj, prep, adv}。

**输出结果：**一个概率最大的隐状态序列。上面的主语“他们”的词性有5种可能取值，我们取概率最大的哪一种。

**模型（训练和预测）：**对于一个样本{x,y}，总共有T个节点，每个节点的参数包括权重w和特征函数f。训练的时候，当训练到第i节点，我们要计算其前向向量和后向向量，从而得到当前位置的非规范化概率，再将对结果进行归一化就得到了当前位置的条件概率，这个结果是一个5维的向量，每一维表示一种词性的可能性。对这一个样本，我们的特征总共有5\*T+5\*5\*T个，T表示节点数目。对应的权值参数w是大小为5\*D+5\*5\*D维的列向量。

**训练过程**

得到权值参数w之后要对它进行训练，因为w可正可负可为0。权重w为正时，特征函数就很靠谱，非常有可能得到正确的词性，权重为负时，特征函数就不靠谱，容易导致结果错入，权重为0，那就表示该特征函数对结果没有影响。

我们并不能预知最佳的w取值，所以通过优化条件概率来训练w。前面说到每个节点都学习到一个5维的向量也就是条件概率值，那么一个长度为T的样本的可能输出就有5^T种，一个样本集合有N条语料，如果直接计算的话复杂度非常高。所以，我们可以通过极大似然估计获取训练集 条件概率的最优值，在对参数的求导过程中，可以采用梯度下降法或是牛顿法进行计算，从而更新权重w。

以拟牛顿法的BFGS算法为例

算法流程：

输入：特征函数和经验分布(X,Y)

输出：最优参数值，最优模型

(1)选定初始点,取正定对称矩阵,k=0

(2)计算的梯度，若为0，则停止计算；否则转(3)

(3)由=- 求出

(4)一维搜索，求使得

(5)更新权值w,

(6)计算新的权值的梯度： ，若=0，则停止计算；否则，求出

其中，，

(7)令k=k+1，转3

**预测过程：**输入分割好的语料，输出一个概率最大的隐藏序列。如输入：坐在/树上的/猫猫/特别/可爱，经过训练，输出概率值最大的序列：{v, adj, n, adv, adj}。在本例中，每个位置都包括5个状态特征值和一个5\*5的转移特征矩阵。通过前向和后向向量计算每个样本我们得到5T种可能取值，一般用viterbi算法进行求解，这种算法运用动态规划的思想，从局部最优得到全局最优。



标注问题本质是在观测序列上求非规范化概率最大的隐状态路径问题

**参考资料**

<https://zhuanlan.zhihu.com/p/68874312>