BiLSTM+CRF算法总结

erazhan

2020 年 3 月 20 日

BiLSTM+CRF算法可用于解决中文分词,词性标注和命名实体识别等序列标注问题.首先序列标注问题描述为:给定观测输入$x = (x_1, \ldots, x_n)$,找到使得$P(y|x)$最大的隐藏状态$y = (y_1, \ldots, y_n)$,其中序列长度n也可以用seq_length表示,用公式表示如下:

$$y^* = \arg\max_y P(y|x) = \arg\max_y \frac{e^{\sum_{k=1}^{K} f_k(y,x)}}{\sum_y e^{\sum_{k=1}^{K} f_k(y,x)}}.$$

其中$f_k(y,x) = \sum_{i=1}^{n+1} f_k(y_{i-1}, y_i, x, i)$,是由同一个特征函数$f_k(y_{i-1}, y_i, x, i)$遍历$n+1$个时间步计算求和得到.特征函数一共有$K$个,包括状态特征函数和转移特征函数.(参考《统计学习方法》)

仅考虑一条数据x,定义损失函数为负对数似然函数:

$$L = -\ln P(y|x) = -(S - \ln Z),$$

其中$S = S(y,x) = \sum_{k=1}^{K} f_k(y,x), Z = \sum_y e^{\sum_{k=1}^{K} f_k(y,x)} = \sum_y e^S$.

BiLSTM+CRF算法模型中考虑分开计算两种特征函数对应的分值.首先BiLSTM模型的输出$logits$维度为$(batch_size, seq_length, num_tags)$, 状态转移矩阵$transition_params$维度为$(num_tags, num_tags)$,数据的真实标签$tag_indices$ 维度为$(batch_size, seq_length)$.

状态特征函数对应的分数$unary_scores$是根据$tag_indices$对每条数据每个seq_length的真实label在$logits$中找到对应$label$的数值,并对所有时间步(一共seq_length)上的数值求和得到.

转移特征函数对应的分数$binary_scores$根据$tag_indices$对每条数据所有$(seq_length - 1)$个相邻状态对(yp, y)在$transition_params$矩阵中找到对应位置上的数值并求和得到.

那么对于每条数据将分别计算状态特征函数值S_1和转移特征函数值S_2求和即可得到$S = S_1 + S_2$,然后计算给定数据x对应的归一化因子Z(考虑所有可能出现的y),最后得到损失函数值.