**氧结合蛋白的分类预测**

与其他论文的对比试验

特征：20维的protr转换后的氨基酸序列组成占比

模型：3CNN

处理过程：提取csv格式的序列成分占比数据，打标签，分成5折，转成pickle文件，按照label：pdb\_feature格式，方便特征fit模型。

后来发现csv格式可以作为向量直接使用pandas读取，不需要转成pickle再进一步处理

由于服务器一段时间内被人占用，没法调试，改用colab调试程序，也方便一步步查看问题所在。

Colab上遇到的问题：

1586850198(1)

调试查看测试训练集的shape都是条数，显示list的维度为1，20维的数组shape为1导致出错，但是后来在服务器上调试并没有这个问题。

遇到的问题：输入层的序列维度不匹配。

解决办法：修改embedded层

遇到的问题：显示kernel过大

解决方法：调整超参数的范围（缩小）

遇到的问题：



不适用adabound报错：  
ValueError: Could not interpret optimizer identifier: <module 'hyperas.optim' from '/home/jiajunh/miniconda3/envs/py36/lib/python3.6/site-packages/hyperas/optim.py'>

说明是无法使用优化器

小插曲：在碰到问题的时候咨询王雷学长，发现论文使用的特征没问题，但是不是3层cnn，而是全连接层，所以是line而不是3cnn模型。在此进行改正。

学习并安装测试使用pytorch，便于后期工作的开展。

使用pytorch遇到的问题：数据格式不同。由于提取的label是str类型，而feature也是一个个list，要转成矩阵处理。（需要构造成为tensor（张量））

问题：打标的时候从1开始，在转onehot的时候classes=9出错。

重新打标记，改成从0开始。

对比论文中的模型细节：输入层+隐藏层1（使用relu激活函数）+dropout层（设置为0.2）+隐藏层2（使用relu激活函数）+隐藏层3（使用relu激活函数）+dropout层（设置为0.2）+输出层

初步来看效果较好，由于涉及的模型较简单，参数较少，手动调参（batchsize，learning\_rate,weight\_decay)设置为（32，64，128）（0.01，0.001，0.0001）（0.001，0.0001，0.00001）设置300个epoch（测试效果较好）

结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | loss | precision | recall | f1score | acc |
| 第一折 | 0.3388 | 0.923556 | 0.923556 | 0.923556 | 0.923556 |
| 第二折 | 0.2245 | 0.944198 | 0.944198 | 0.944198 | 0.944198 |
| 第三折 | 0.3245 | 0.922193 | 0.922193 | 0.922193 | 0.922193 |
| 第四折 | 0.4202 | 0.892826 | 0.892826 | 0.892826 | 0.892826 |
| 第五折 | 0.4913 | 0.877208 | 0.877208 | 0.877208 | 0.877208 |
| 平均值 | 0.35986 | 0.9119962 | 0.9119962 | 0.9119962 | 0.9119962 |

第一折：

batchsize:064,lr: 0.01000,weightdecay: 0.000100

第二折：

batchsize:032,lr: 0.00100,weightdecay: 0.000010

第三折：

batchsize:064,lr: 0.01000,weightdecay: 0.000100

第四折：

batchsize:032,lr: 0.01000,weightdecay: 0.000100

第五折：

batchsize:032,lr: 0.00100,weightdecay: 0.000100

问题：precision ,recall, fscroe和acc率都相同

原因：多分类的precision，recall和f1需要制定average=“macro”用宏平均。

修改后的指标：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | loss | precision | recall | f1score | acc |
| 第一折 | 0.2965 | 0.675608 | 0.720236 | 0.692783 | 0.934222 |
| 第二折 | 0.2245 | 0.944198 | 0.944198 | 0.944198 | 0.944198 |
| 第三折 | 0.3245 | 0.922193 | 0.922193 | 0.922193 | 0.922193 |
| 第四折 | 0.4202 | 0.892826 | 0.892826 | 0.892826 | 0.892826 |
| 第五折 | 0.4913 | 0.877208 | 0.877208 | 0.877208 | 0.877208 |
| 平均值 | 0.35986 | 0.9119962 | 0.9119962 | 0.9119962 | 0.9119962 |

第一折：

batchsize:064,lr: 0.01000,weightdecay: 0.000100

第二折：

第三折：

第四折：

第五折：