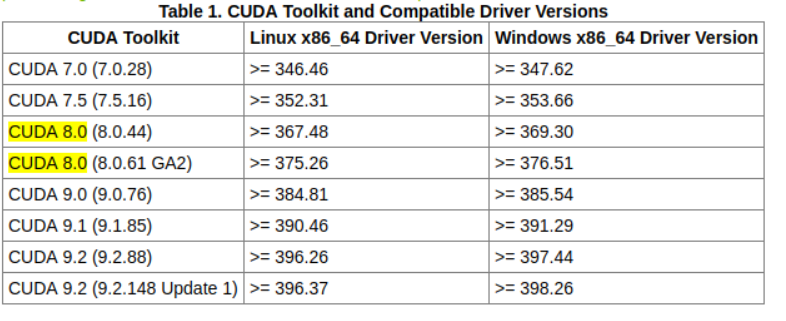
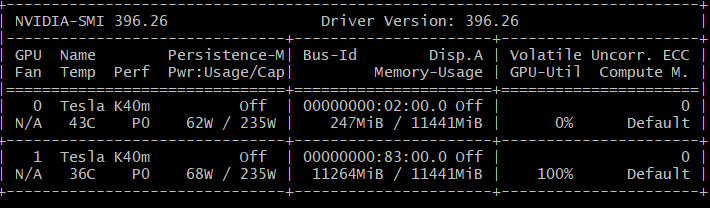
**氧结合蛋白的分类预测**

**之前几个模型的整体指标没有使用宏平均，重跑评价指标：**

遇到的问题：可能因为tensorflow版本的更新，原来在-1-5节点可以跑的代码现在显示cuda版本和tf冲突











回退tf版本，依然报错。

重新安装，版本变成2.1，依然不行。

解决方法：使用gpu-1-4节点，是cuda10，重新装tf，版本为2.1，回退为1.14.0，成功。

现在在重新跑之前几个模型的指标，由于服务器被占用，还没有跑完。

**整合：**

数据集：在uniprot上获得9类氧结合蛋白，长度分布在0-800之间，数量2751条。（cytoglobin：107条，erythrocruorin：13条，flavohemoprotein：943条，hemerythrin：35条，hemocyanin：31条，hemoglobin：1374条，leghemoglobin：28条，myoglobin：120条，neuroglobin：100条）

训练集2200条，测试集551条。，每一类都按随机分成五折再拼接成一个数据集。

使用模型：使用2层cnn连接双向gru。

超参数调优：两层cnn的filters大小设置为choice([32,64, 128])，kernel\_size设置为choice([3,5,7, 11, 15, 19, 21])，dropout层均设置为uniform(0, 1)，池化层的poolsize设置为choice([4,8,16, 32])，步长设置为choice([4,8,16, 32, 64])，两层双向gru输出维度设置为choice([100,200, 300, 400])，dropout层均设置为uniform(0, 1)，Regularizer（权重上的正则项）设置为choice([0.001, 0.0001, 0.00001])，学习率设置为choice（[1e-03,1e-04]）, epochs（轮次）设置为choice([100, 150, 200]), batch\_size（批次）设置为choice([32, 64, 128，256])，最后得出在五折上不同的最优组合

输入数据：onehot编码后的序列和pssm特征（把序列长度补全为800，pssm同理）

输出：9类的预测概率

过程：首先获得数据集中序列的pssm特征，进行归一化获得pssm的npy文件，其次按照序列-标签-id的方式生成pickle文件。再把数据分为五折，放入模型中训练超参数，把获得的超参数带入模型，训练权重因子，最后得到测试集的各项指标如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | loss | precision | recall | f1score | acc |
| 第一折 | 0.27383 | 0.859 | 0.876 | 0.858 | 0.981 |
| 第二折 | 0.27424 | 0.792 | 0.789 | 0.788 | 0.985 |
| 第三折 | 0.32414 | 0.900 | 0.808 | 0.842 | 0.982 |
| 第四折 | 0.29305 | 0.918 | 0.858 | 0.861 | 0.983 |
| 第五折 | 0.23949 | 0.798 | 0.614 | 0.671 | 0.973 |
| 平均值 | 0.28095 | 0.853 | 0.789 | 0.804 | 0.981 |

（之前记录的数据，未使用宏平均，需要修改）

每一类的分类性能的评价：

Cytoglobin（107）细胞珠蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 0.560 | 0.952 | 0.721 |
| 第二折 | 0.833 | 0.714 | 0.769 |
| 第三折 | 0.667 | 0.636 | 0.651 |
| 第四折 | 0.833 | 0.238 | 0.370 |
| 第五折 | 0.5 | 0.5 | 0.5 |
| 平均值 | 0.679 | 0.608 | 0.602 |

Erythrocruorin（13）红细胞凝血素

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 0.5 | 0.5 | 0.5 |
| 第二折 | 0 | 0 | 0 |
| 第三折 | 1.0 | 0.5 | 0.67 |
| 第四折 | 1.0 | 1.0 | 1.0 |
| 第五折 | 1.0 | 0.333 | 0.5 |
| 平均值 | 0.7 | 0.467 | 0.53 |

Flavohemoprotein（943）黄素血红蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 0.995 | 0.973 | 0.984 |
| 第二折 | 0.973 | 0.952 | 0.963 |
| 第三折 | 0.978 | 0.941 | 0.959 |
| 第四折 | 0.958 | 0.958 | 0.958 |
| 第五折 | 0.953 | 0.963 | 0.958 |
| 平均值 | 0.971 | 0.957 | 0.964 |

Hemerythrin（35）蚯蚓血红蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 1.0 | 1.0 | 1.0 |
| 第二折 | 1.0 | 1.0 | 1.0 |
| 第三折 | 1.0 | 1.0 | 1.0 |
| 第四折 | 1.0 | 1.0 | 1.0 |
| 第五折 | 1.0 | 1.0 | 1.0 |
| 平均值 | 1.0 | 1.0 | 1.0 |

Hemocyanin（31）血蓝蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 1.0 | 0.833 | 0.909 |
| 第二折 | 1.0 | 1.0 | 1.0 |
| 第三折 | 1.0 | 1.0 | 1.0 |
| 第四折 | 1.0 | 1.0 | 1.0 |
| 第五折 | 1.0 | 0.714 | 0.833 |
| 平均值 | 1.0 | 0.909 | 0.948 |

Hemoglobin（1374）血红蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 0.968 | 0.876 | 0.920 |
| 第二折 | 0.930 | 0.967 | 0.948 |
| 第三折 | 0.913 | 0.956 | 0.934 |
| 第四折 | 0.939 | 0.953 | 0.946 |
| 第五折 | 0.870 | 0.953 | 0.910 |
| 平均值 | 0.924 | 0.941 | 0.932 |

Leghemoglobin（28）豆血红蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 1.0 | 1.0 | 1.0 |
| 第二折 | 0.857 | 1.0 | 0.923 |
| 第三折 | 1.0 | 0.8 | 0.889 |
| 第四折 | 1.0 | 0.667 | 0.8 |
| 第五折 | 0.5 | 0.333 | 0.4 |
| 平均值 | 0.871 | 0.76 | 0.802 |

Myoglobin（120）肌红蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 0.75 | 1.0 | 0.857 |
| 第二折 | 0.846 | 0.917 | 0.880 |
| 第三折 | 0.833 | 0.833 | 0.833 |
| 第四折 | 0.958 | 0.958 | 0.958 |
| 第五折 | 0.933 | 0.583 | 0.718 |
| 平均值 | 0.864 | 0.858 | 0.849 |

Neuroglobin（100）神经珠蛋白

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1score |
| 第一折 | 0.938 | 0.75 | 0.833 |
| 第二折 | 0.688 | 0.55 | 0.611 |
| 第三折 | 0.705 | 0.6 | 0.649 |
| 第四折 | 0.576 | 0.95 | 0.717 |
| 第五折 | 0.429 | 0.15 | 0.222 |
| 平均值 | 0.667 | 0.6 | 0.606 |

对比试验1：

使用模型：使用3层cnn.

超参数调优：两层cnn的filters大小设置为choice([32,64, 128])，kernel\_size设置为choice([3,5,7, 11, 15, 19, 21])，dropout层均设置为uniform(0, 1)，池化层的poolsize设置为choice([4,8,16, 32])，步长设置为choice([4,8,16, 32, 64])，Regularizer（权重上的正则项）设置为choice([0.001, 0.0001, 0.00001])，学习率设置为choice（[1e-03,1e-04]）, epochs（轮次）设置为choice([100, 150, 200]), batch\_size（批次）设置为choice([32, 64, 128，256])，最后得出在五折上不同的最优组合

输入数据：onehot编码后的序列和pssm特征（把序列长度补全为800，pssm同理）

输出：9类的预测概率

过程：首先获得数据集中序列的pssm特征，进行归一化获得pssm的npy文件，其次按照序列-标签-id的方式生成pickle文件。再把数据分为五折，放入模型中训练超参数，把获得的超参数带入模型，训练权重因子，最后得到测试集的各项指标如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | loss | precision | recall | f1score | acc |
| 第一折 | 0. | 0. | 0. | 0. | 0. |
| 第二折 | 0. | 0. | 0. | 0. | 0. |
| 第三折 | 0. | 0. | 0. | 0. | 0. |
| 第四折 | 0. | 0. | 0. | 0. | 0. |
| 第五折 | 0. | 0. | 0. | 0. | 0. |
| 平均值 | 0. | 0. | 0. | 0. | 0. |

（未使用宏平均，之前的结果需要改进，暂时不记录）

对比试验2：

使用模型：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

超参数调优：（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）

输入数据：20维的protr转换后的氨基酸序列组成占比

输出：9类的预测概率

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。

最后得到测试集的各项指标如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | loss | precision | recall | f1score | acc |
| 第一折 | 0.2965 | 0.675608 | 0.720236 | 0.692783 | 0.934222 |
| 第二折 | 0.3012 | 0.513572 | 0.511846 | 0.511866 | 0.920283 |
| 第三折 | 0.3985 | 0.575980 | 0.619194 | 0.595692 | 0.911583 |
| 第四折 | 0.4489 | 0.581598 | 0.486962 | 0.518431 | 0.877768 |
| 第五折 | 0.4969 | 0.423939 | 0.405770 | 0.412635 | 0.864841 |
| 平均值 | 0.3884 | 0.5541394 | 0.5488016 | 0.5462814 | 0.9017394 |

（结果正确，使用宏平均）

超参数：

第一折：

batchsize:064,lr: 0.01000,weightdecay: 0.000100

第二折：

batchsize:064,lr: 0.01000,weightdecay: 0.000100

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:128,lr: 0.01000,weightdecay: 0.000100