当前课题：膜蛋白药物作用位点预测

1问题定义：药物小分子名字信息从Drugbank中提取。蛋白质序列中筛选膜蛋白，膜蛋白序列中有交互作用位点，交互的分子是药物分子的位点是正样本，其余是负样本。

2数据来源PDB和Drugbank

3

数据总量：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 全部 | 0.9cd-hit | 0.3cd-hit |
| 蛋白质条数 | 3232 | 653 | 446 |
| 正样本数 | 22141 | 4235 | 2729 |
| 负样本数 | 770702 | 142621 | 97590 |
| 正负样本比 | 1：34.8 | 1：33.6 | 1：35.7 |

4处理步骤

4.1数据处理：

读取数据

分别读取3个数据文件：【seqdicts.pickle】、【labeldicts.pickle】、【seq\_3cutoff.pickle】

【seqdicts.pickle】中存放着的是所有序列的氨基酸信息，由蛋白质名字和氨基酸序列组成

【labeldicts.pickle】中存放着的是所有序列的标签信息，由蛋白质名字和标签序列组成

【seq\_3cutoff.pickle】中存放着的是经过去冗余后的蛋白质名字。

**标识正负样本**

窗口大小取值【31】，左右各15位氨基酸。

正样本是中间位置为作用点的序列，其余为负样本。

负样本中，去除掉左右有空格的样本，去除包含正样本位点的样本。

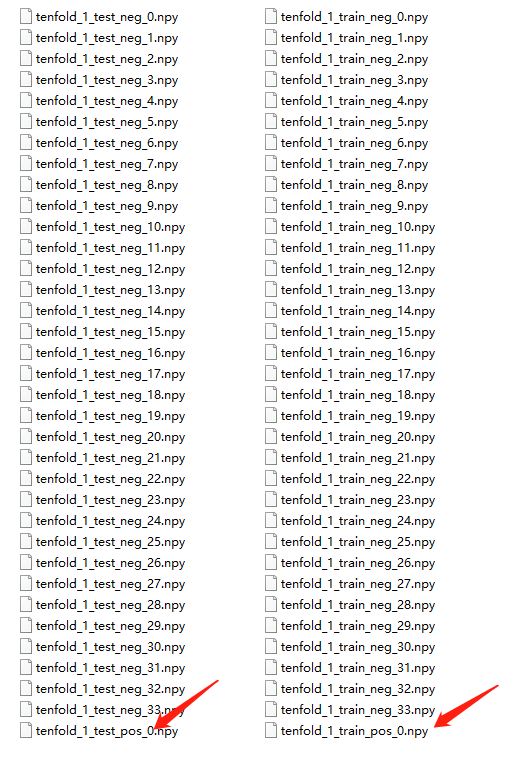
**准备数据（十折交叉验证）**

随机种子设置为【6】

分别对正负样本进行十折分配。10折交叉验证，其中9份用作训练，1份用作验证。

**Bootstrapping后保存**

正负样本比为1：34，有1份正样本标记为【0】，有34份正样本标记为【0~33】，分别存储成train和test集。其中train:test的数据量比是9:1。



**5训练网络：**

**十折交叉分次训练**

分别对每一折中的样本进行训练：一个for循环内分别对某个文件夹目录下的内容进行读取训练。

**对其中一折进行训练**

依次读取数据文件内容，存入对应变量中

根据负样本的数量迭代。第一次训练时，网络新建。之后使用旧网络对数据进行学习。

1号正样本---1号负样本

1号正样本---2号负样本

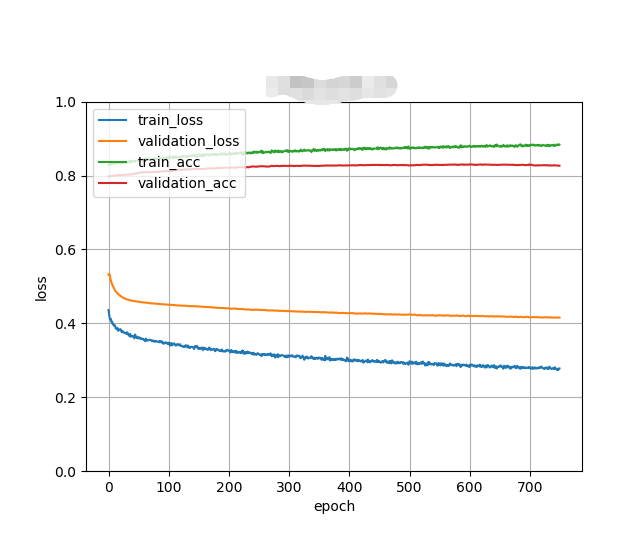
1号正样本---3号负样本

。。。 。。。

1号正样本---33号负样本

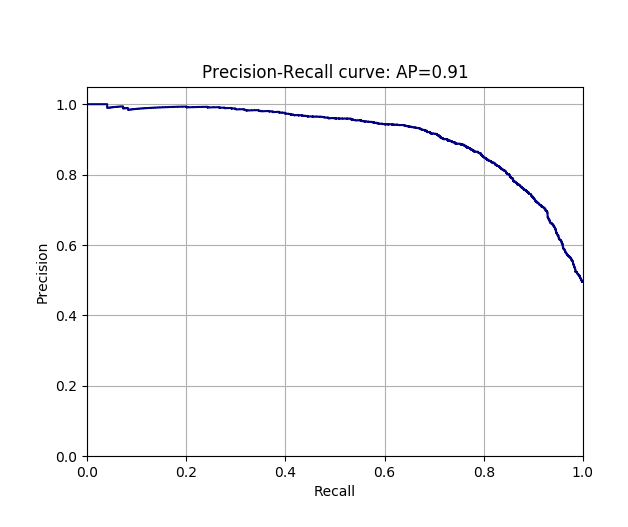
每一组数据都进行一次模型评估。（使用1号验证集中的正样本--1号验证集中的负样本）

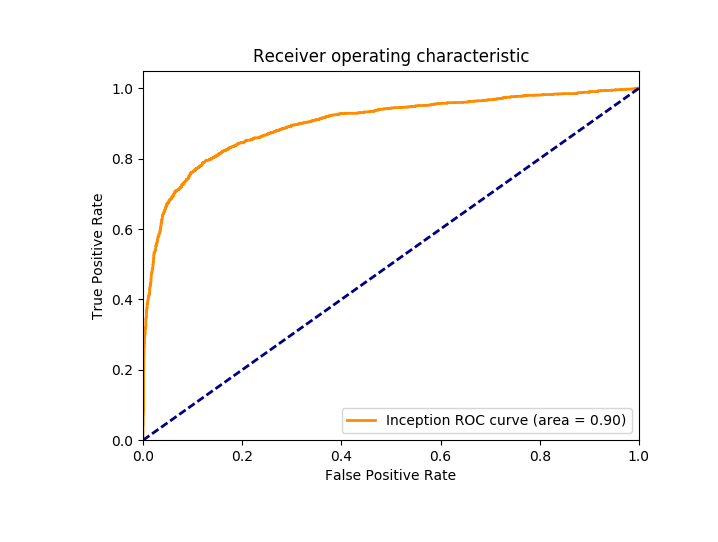
查看其acc变化趋势、loss变化趋势。

（示例图，为了不用一次看2张图）

最终的mcc值等相关数据。

每一组数据画对应的图线。画出Precision-Recall1曲线、ROC曲线。

（示例图）

（示例图）

**5当前结果**

在epoch = 1250时，第三组训练中，达到了val\_acc = 0.85（不是最高，是较高）

同次实验中val\_loss= 0.36

Mcc在0.64-0.68之间。

其他设置如下：

Filters = 64, filtersize = 2