集成学习

## 一、实验内容：

利用AdaBoost和XgBoost模型完成链接<https://archive.ics.uci.edu/ml/datasets/HIV-1+protease+cleavage>的任务

## 二、实验任务：

一共有四个数据集，分别用以其中一个数据集作为训练集，其他三个数据集作为测试集，测试模型性能。下面为四个数据集的相关信息。

- 746 data: 401 cleaved, 345 non-cleaved

- 1625 data: 374 cleaved, 1251 non-cleaved

- Schilling data: 434 cleaved, 2838 non-cleaved

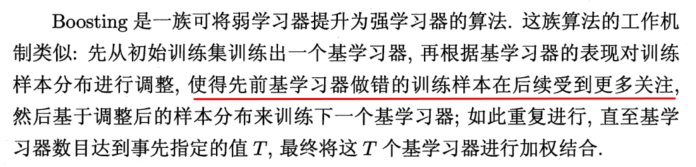
- Impens data: 149 cleaved, 798 non-cleaved

数据集的属性是8个字母的字符串（每个字母代表一个独立的属性）。'ARNDCQEGHILKMFPSTWYV'一共有20个供选择的值，分别代表20个不同的氨基酸。要预测的是此字符串是否代表HIV-1蛋白酶对其进行切割的肽段（或蛋白质）中的位点（如果是，则为+1，否则为-1）。

此时，我们的任务可以用下列式子表示：

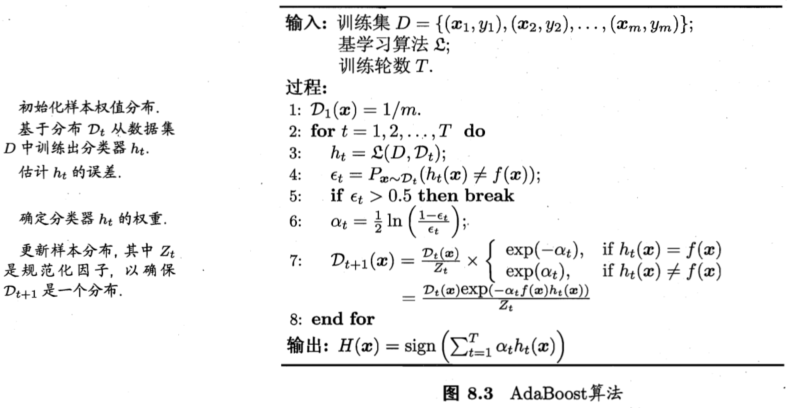
## 三、实验算法：

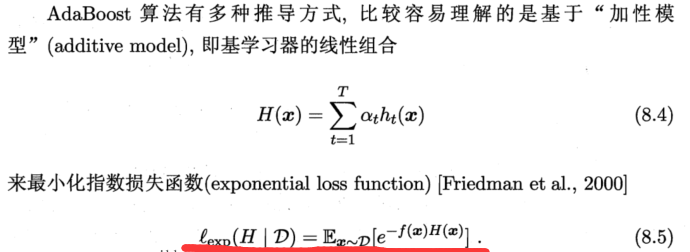
### 1、Boosting



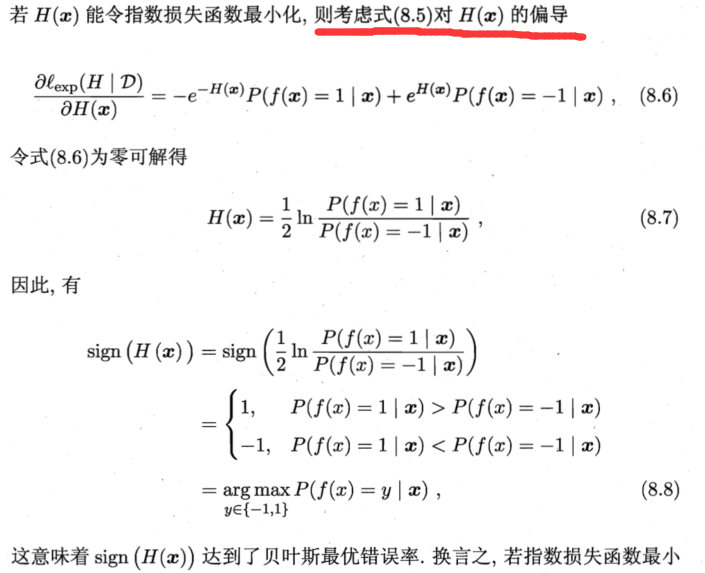
### 2、AdaBoost

基学习器的线性组合

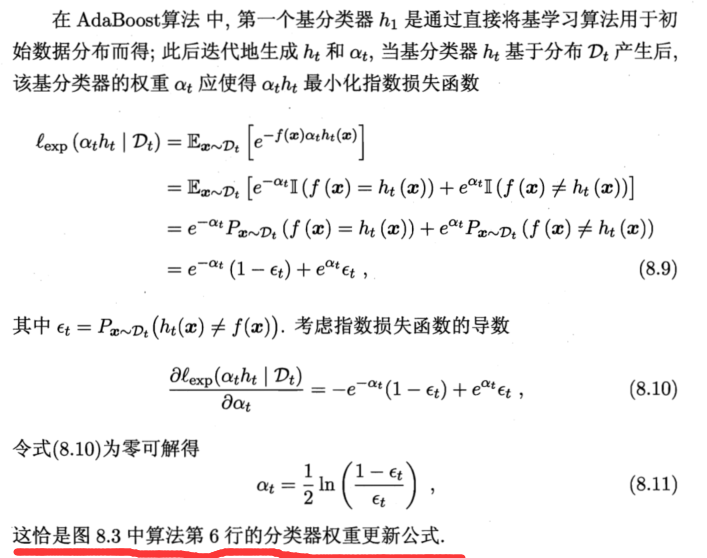




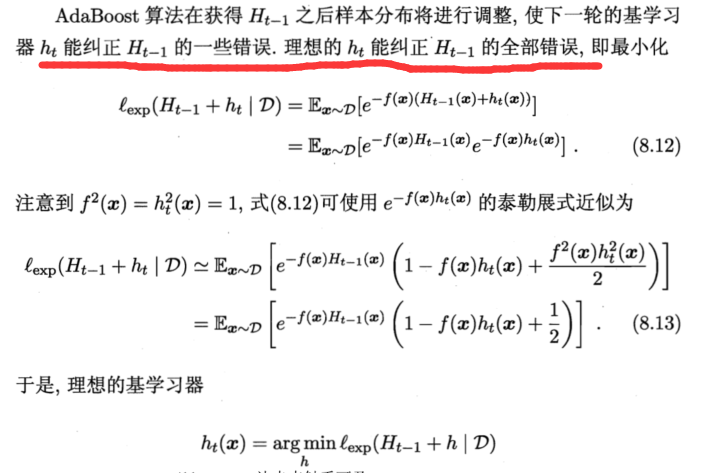
为什么最小化指数损失函数？最小化指数损失函数等同于最小化分类误差

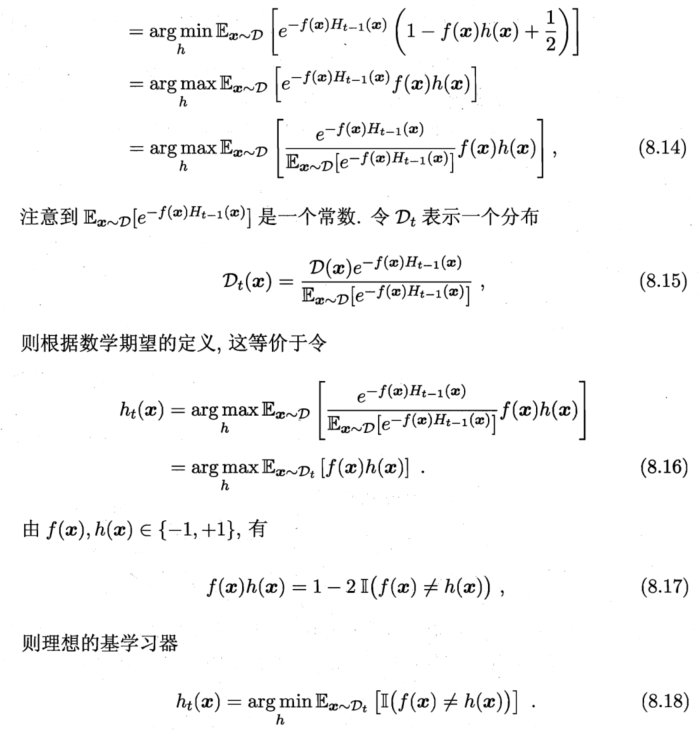


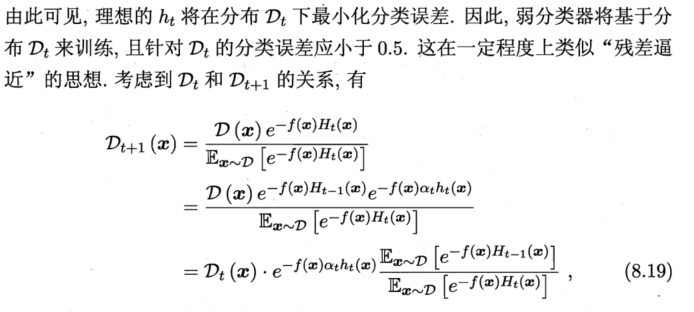
第一个基分类器？



如何获得下一轮的基学习器ht？







每个分类器下最小化分类误差

最小化指数损失函数一定能保证当前地集成学习器效果比前一轮地集成学习器效果好嘛？

### 3、XgBoost

## 四、实验数据处理：

**一共有4个数据集，按照实验要求，以任意一个数据集作为训练集，以其他数据集作为测试集，来对模型性能进行评估。**

**考虑到每条数据的属性都是字符型，需要将其离散数值化。**'ARNDCQEGHILKMFPSTWYV'将其分别进行one-hot编码

**A = [1, 0, 0, …, 0]**

**R = [0, 1, 0, …, 0]**

## 五、实验结果：

### 1、通过sklearn包实现

Sklearn.ensemble.AdaBoostClassifier

模型参数

1. base\_estimators：基分类器，默认是决策树。
2. n\_estimators：基分类器提升（循环）次数，默认是50次。

3、learning\_rate：学习率，表示梯度收敛速度，默认为1。

4、algorithm：boosting算法，也就是模型提升准则，有SAMME和SAMME.R两种，前者是对样本集预测错误的概率进行划分，后者是对样本集预测错的比例进行划分，默认是SAMME.R。

5、random\_state：随机种子设置。

#### 设置参数进行实验

随机森林 使用默认参数

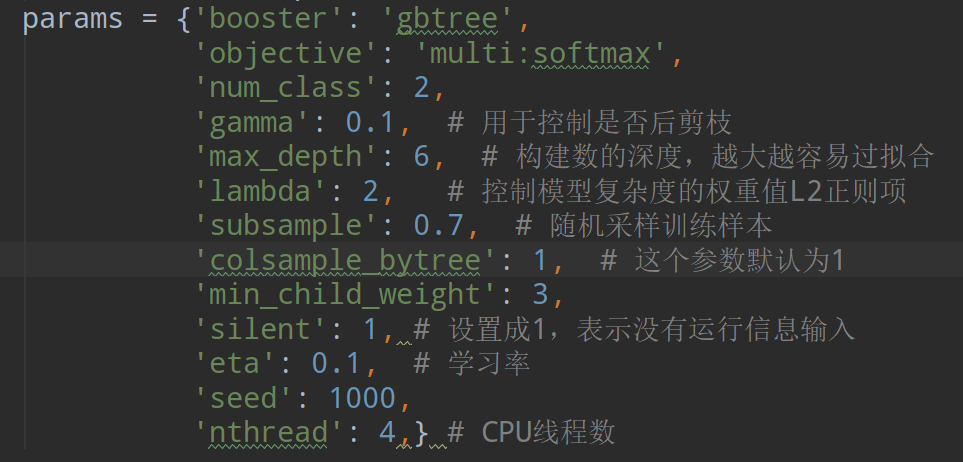
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gini | 训练集 | 验证集 | 测试集 | | |
| 746Data | 2/3的数据 | 1/3的数据 | 1625Data | impensData | schillingData |
|  | 0.9960 | 0.8996 | 0.8917 | 0.8110 | 0.8310 |

AdaBoost 使用默认参数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| 746Data | 2/3的数据 | 1/3的数据 | 1625Data | impensData | schillingData |
|  | 0.9738 | 0.8795 | 0.8911 | 0.8574 | 0.8493 |

XgBoost

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| 746Data | 2/3的数据 | 1/3的数据 | 1625Data | impensData | schillingData |
|  | 0.9779 | 0.8996 | 0.9102 | 0.8448 | 0.8603 |



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| 1625Data | 2/3的数据 | 1/3的数据 | 746Data | impensData | schillingData |
|  | 0.9631 | 0.9594 | 0.9196 | 0.8691 | 0.8829 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| impensData | 2/3的数据 | 1/3的数据 | 746Data | 1625Data | schillingData |
|  | 0.9461 | 0.9146 | 0.7560 | 0.7975 | 0.8658 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| schillingData | 2/3的数据 | 1/3的数据 | 746Data | 1625Data | ImpensData |
|  | 0.9418 | 0.9285 | 0.7064 | 0.84 | 0.8902 |

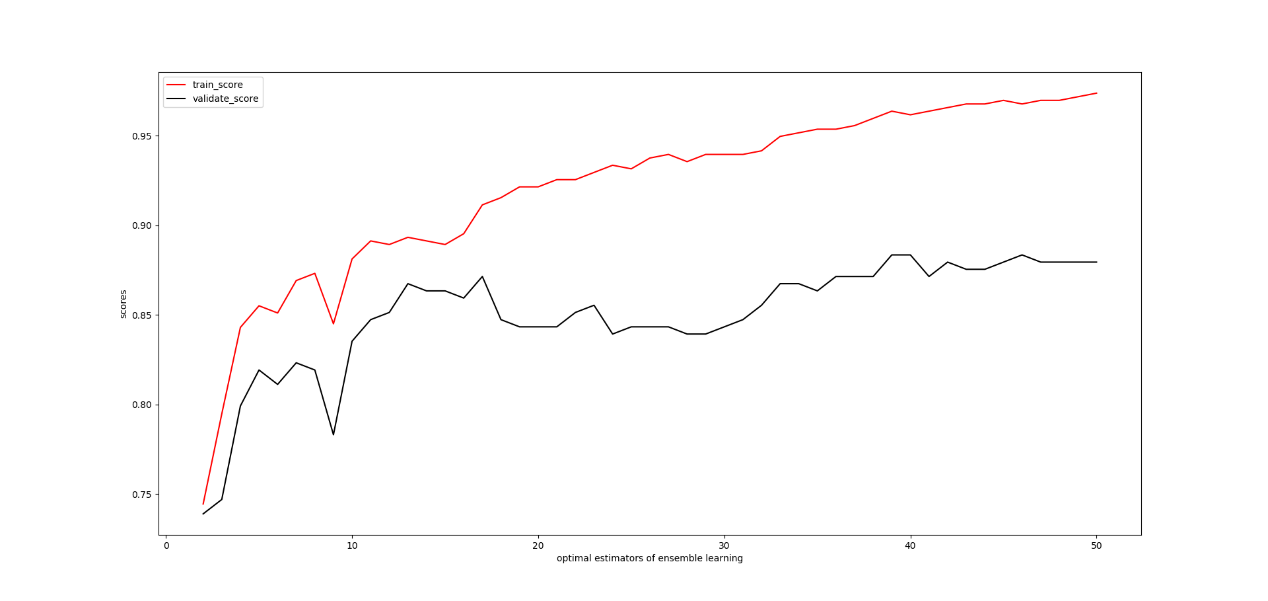
* 1. 利用GridSearchCV求最优参数



有进行了交叉验证，cv=5表示每次计算都把数据集分成5份，其中一份作为交叉验证数据集，其他作为训练集，从而得出最优参数。

### 2、确定模型相关参数

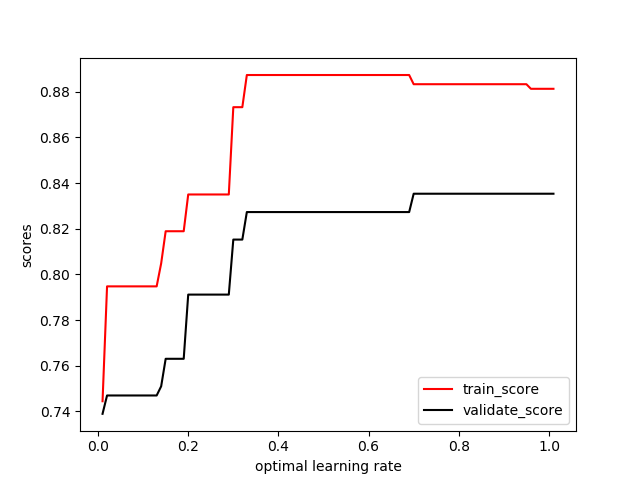
初步探索，基学器的数目

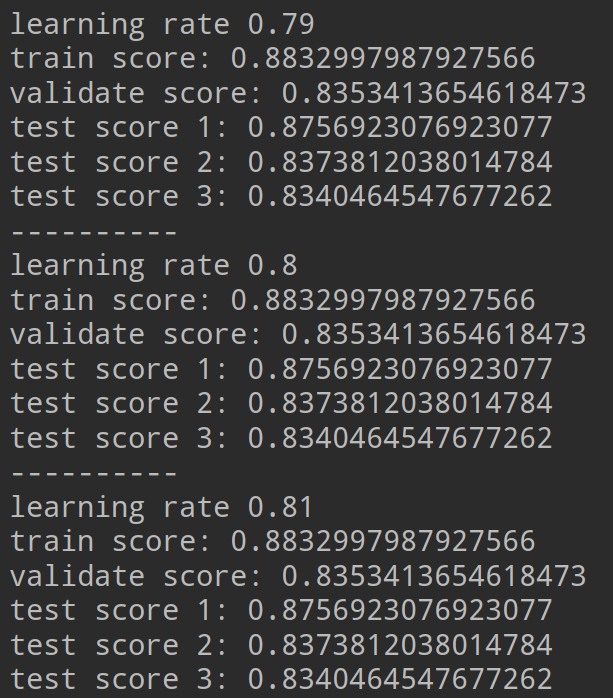


会发现一个明显的趋势，随着基学器数目增加，集成模型容易发生过拟合。所以不是模型数越多就越好。当模型数增多的时候，适当减少学习率，可以防止过拟合。

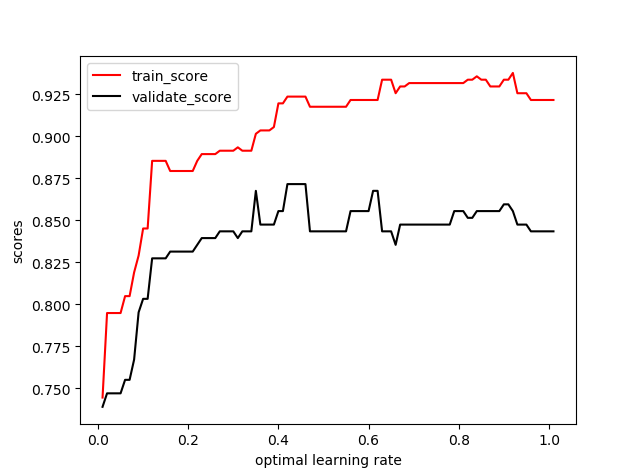
探索基学器数目与学习率之间的关系

（基学器数目n=10）

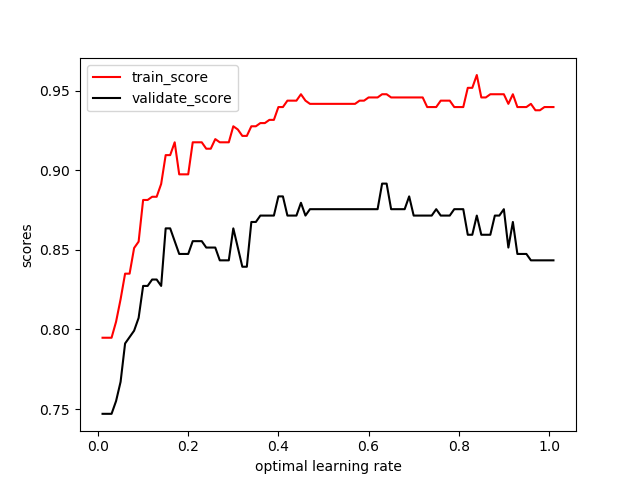


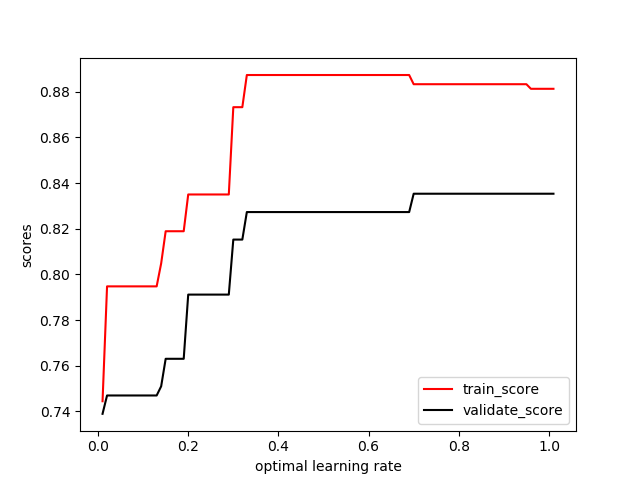


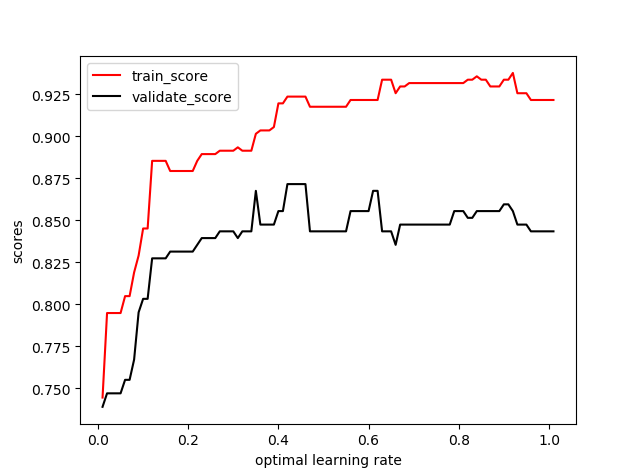
当基学器数目n=20时

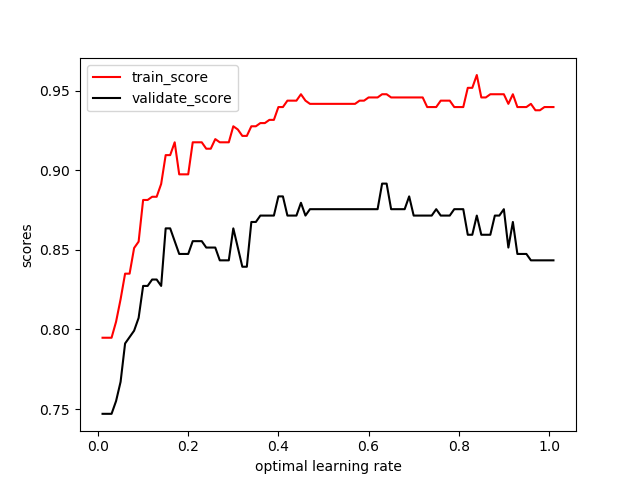


当基学习器数目n=30时

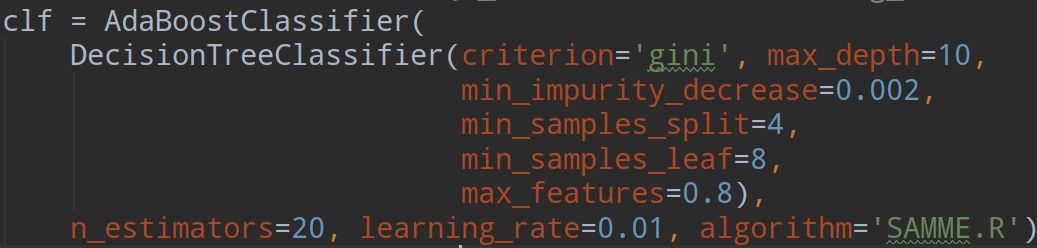








基学器器模型数增多的时候，精确率会增加，但同时会带来过拟合问题，所以要适当地减少学习率。

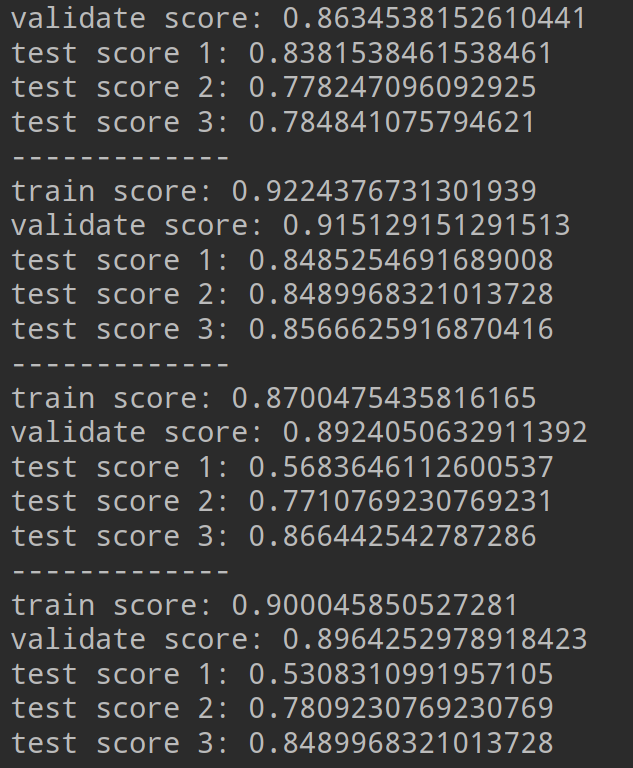


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| 746Data | 2/3的数据 | 1/3的数据 | 1625Data | impensData | schillingData |
|  | 0.9738 | 0.8795 | 0.8911 | 0.8574 | 0.8493 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| 1625Data | 2/3的数据 | 1/3的数据 | 746Data | impensData | schillingData |
|  | 0.9631 | 0.9594 | 0.9196 | 0.8691 | 0.8829 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| impensData | 2/3的数据 | 1/3的数据 | 746Data | 1625Data | schillingData |
|  | 0.9461 | 0.9146 | 0.7560 | 0.7975 | 0.8658 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 训练集 | 验证集 | 测试集 | | |
| schillingData | 2/3的数据 | 1/3的数据 | 746Data | 1625Data | ImpensData |
|  | 0.9418 | 0.9285 | 0.7064 | 0.84 | 0.8902 |



### 3、结论