逻辑回归

## 一、实验内容：

利用逻辑回归完成链接<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

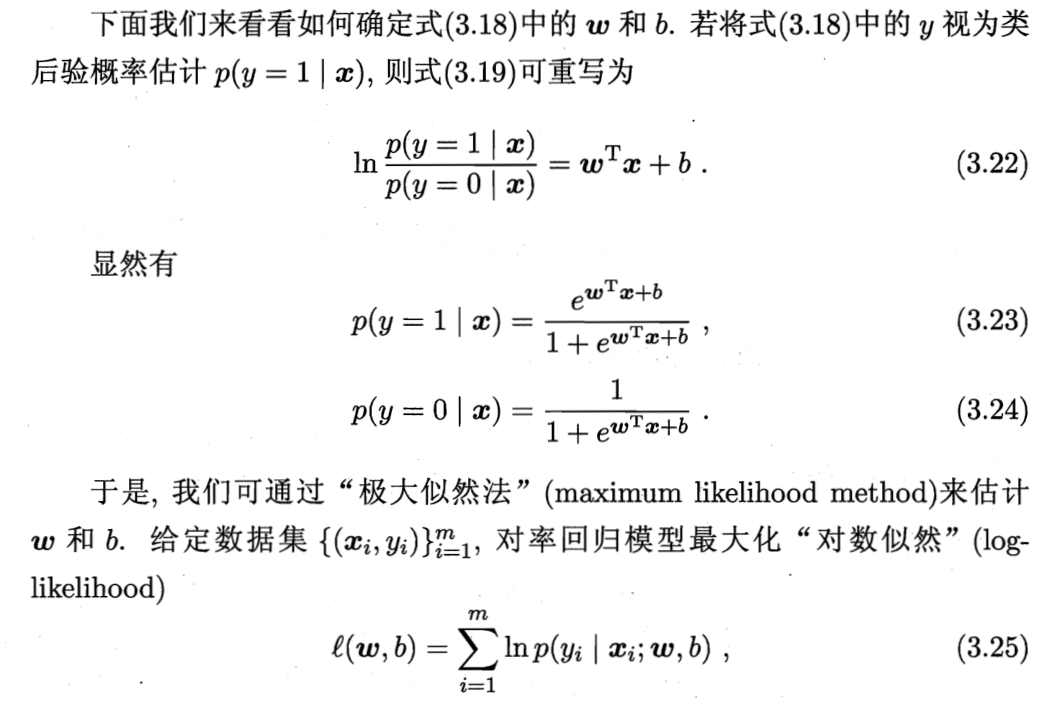
数据集的属性是8个字母的字符串（每个字母代表一个独立的属性）。'ARNDCQEGHILKMFPSTWYV'一共有20个供选择的值，分别代表20个不同的氨基酸。要预测的是此字符串是否代表HIV-1蛋白酶对其进行切割的肽段（或蛋白质）中的位点（如果是，则为+1，否则为-1）。

此时，我们的任务可以用下列式子表示：

## 三、实验算法：

**逻辑回归模型是一个对对数几率进行建模的方法，针对二分类问题，对数几率即为：**

这里的y是**x**作为正例的概率，1-y即为**x**作为负例的概率



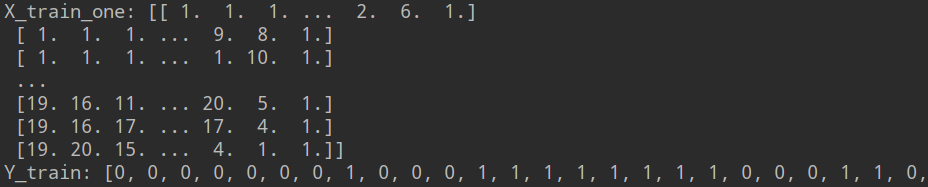
## 四、实验数据处理：

**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离散的特征，需要将其数值化。**'ARNDCQEGHILKMFPSTWYV'将其分别数值化为，1，2，3…，20。是否有更加合适的方法？

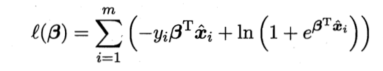
每一个字母进行one-hot编码，最后将每一个数据的各个属性拼接起来，将其作为最后数据的一个表示（每个数据相当于有8\*20=160维， 考虑b的话最后构建出161维的数据）

同时将数据的标签信息改为1，0。



## 五、实验结果：

目标函数为负对数似然函数，最小化负对数似然函数



### 1、通过sklearn包实现

Sklearn.linear\_model.LogisticRegression

Penalty：确定正则项，‘l1’或‘l2’

C=1.0：正则化系数的倒数，值越少，正则化越强

Solver：确定优化算法，（1）‘liblinear’：坐标轴下降法。（2）‘lbfgs’：拟牛顿法。（3）‘netwton-cg’：牛顿法。（4）‘sag’：随机梯度下降。（5）‘saga’:批梯度下降

Class\_weight=None:类型权重参数，选择为‘balanced’自动根据y值计算类型权重。权重与输入数据中的类频率成反比。

#### 损失函数就是负对数似然函数，分别采用上述四种优化算法

1）使用默认参数，penalty=‘l2’，C=1.0，solver=‘lbfgs’未平衡样本

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 训练集  测试集 | 746Data | 1625Data | impensData | schillingData |
| 746Data | 0.6971 | 0.5952 | 0.4611 | 0.4611 |
| 1625Data | 0.6904 | 0.8012 | 0.7692 | 0.7692 |
| impensData | 0.6621 | 0.8163 | 0.8427 | 0.8427 |
| schillingData | 0.6381 | 0.8292 | 0.8674 | 0.8674 |

2）发现可能出现过拟合现象，考虑平衡正负样本后，即class\_weight=‘balanced’，penalty=‘l2’，C=1.0，solver=‘lbfgs’。

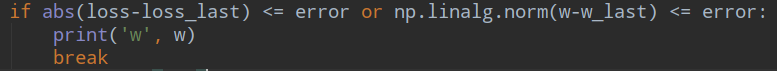
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 训练集  测试集 | 746Data | 1625Data | impensData | schillingData |
| 746Data | 0.6877 | 0.7038 | 0.5536 | 0.4879 |
| 1625Data | 0.7175 | 0.728 | 0.5268 | 0.5415 |
| impensData | 0.6948 | 0.7637 | 0.5913 | 0.5903 |
| schillingData | 0.6770 | 0.6718 | 0.5730 | 0.6021 |

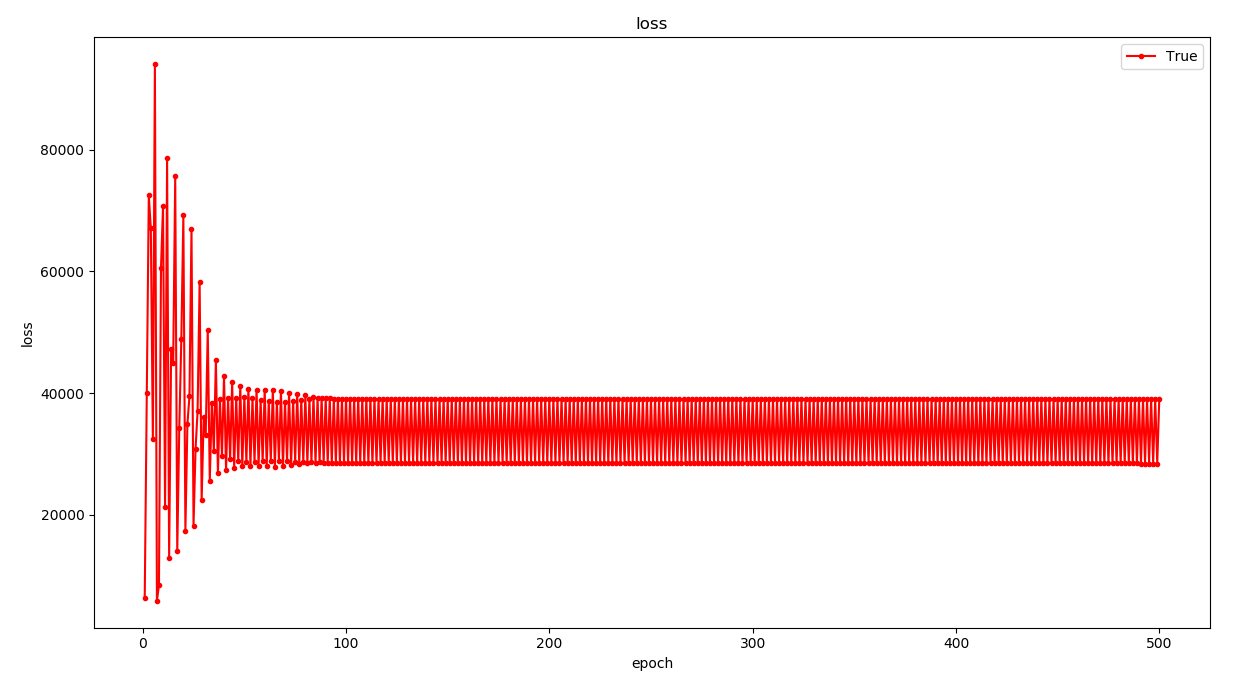
2）在设置正负样本的权重后，发现impensdata和schillingdata数据可能与其他数据集存在较大差异，当以这两个数据集作为训练集时，测试集的准确率都很低。考虑通过重采样来控制正负样本大致相等。

### 2、手动实现

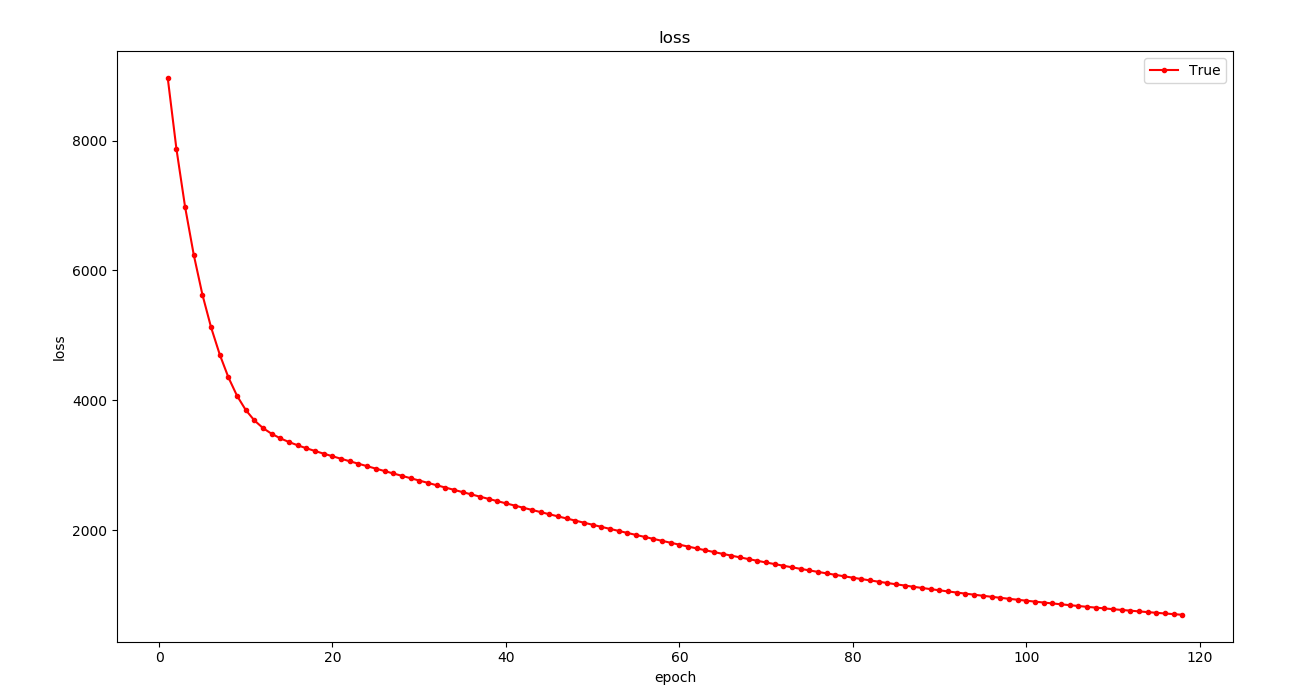
#### 2.1.\_746data为训练集时

使用梯度下降算法，学习率设置为0.001， epoch为500，error=0.1





使用梯度下降算法，学习率设置为0.00001， epoch为200，error=0.01

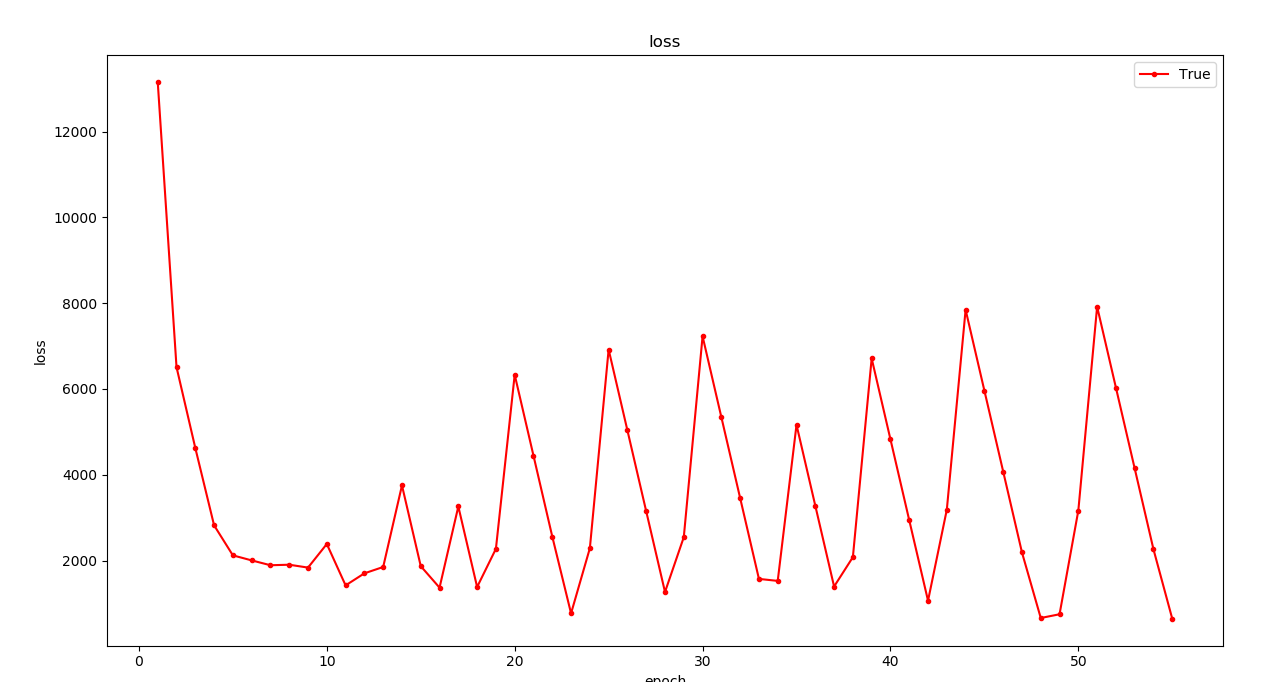


#### 2.2.\_1625data为训练集时

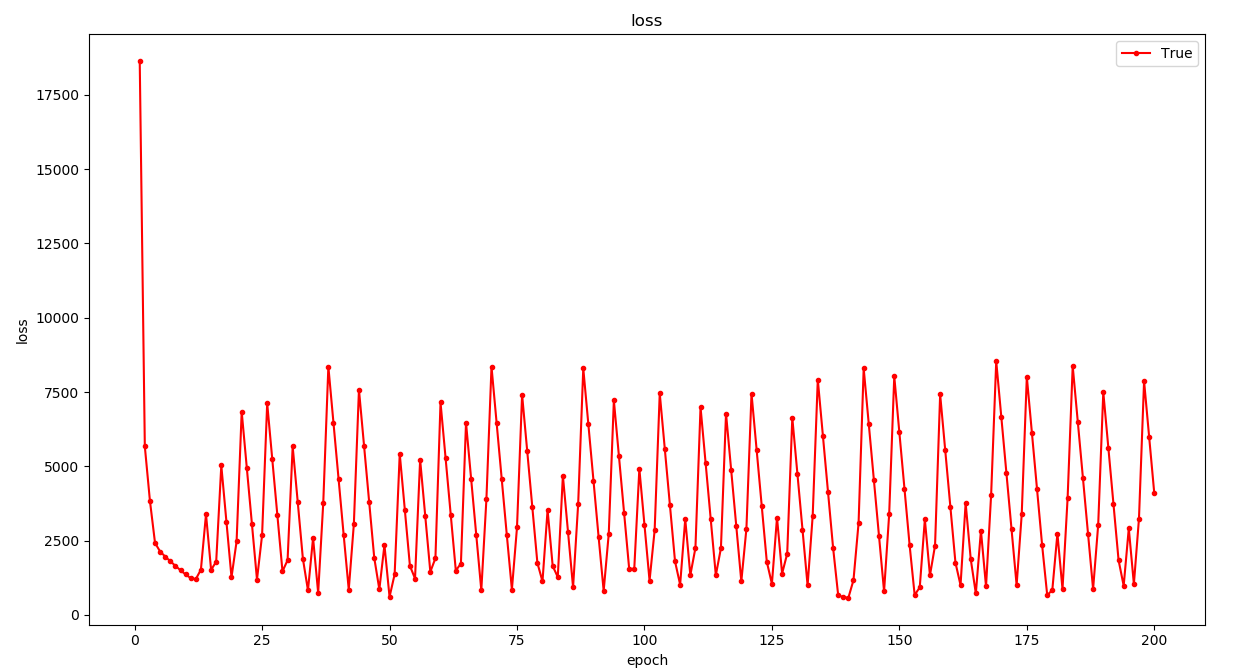
使用梯度下降算法，学习率设置为0.0001， epoch为200，error=0.01

#### 2.3.\_impen sdata为训练集时

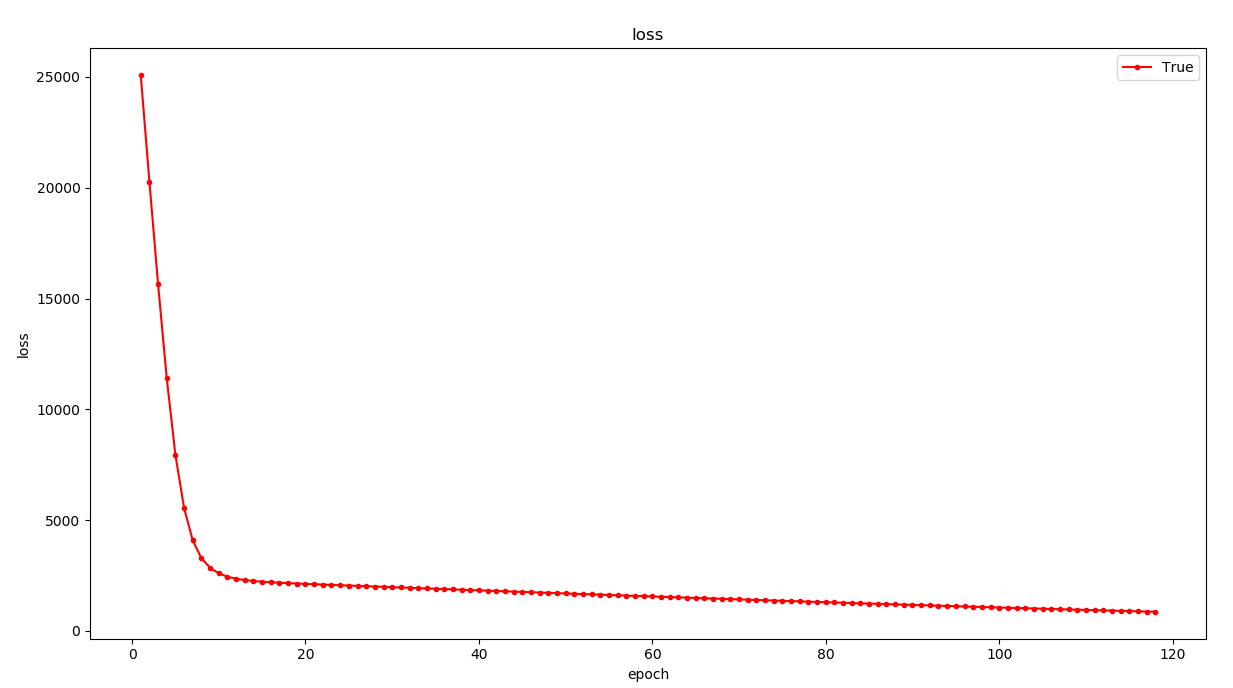
使用梯度下降算法，学习率设置为0.0001， epoch为500，error=0.1



使用梯度下降算法，学习率设置为0.0001， epoch为200，error=0.01

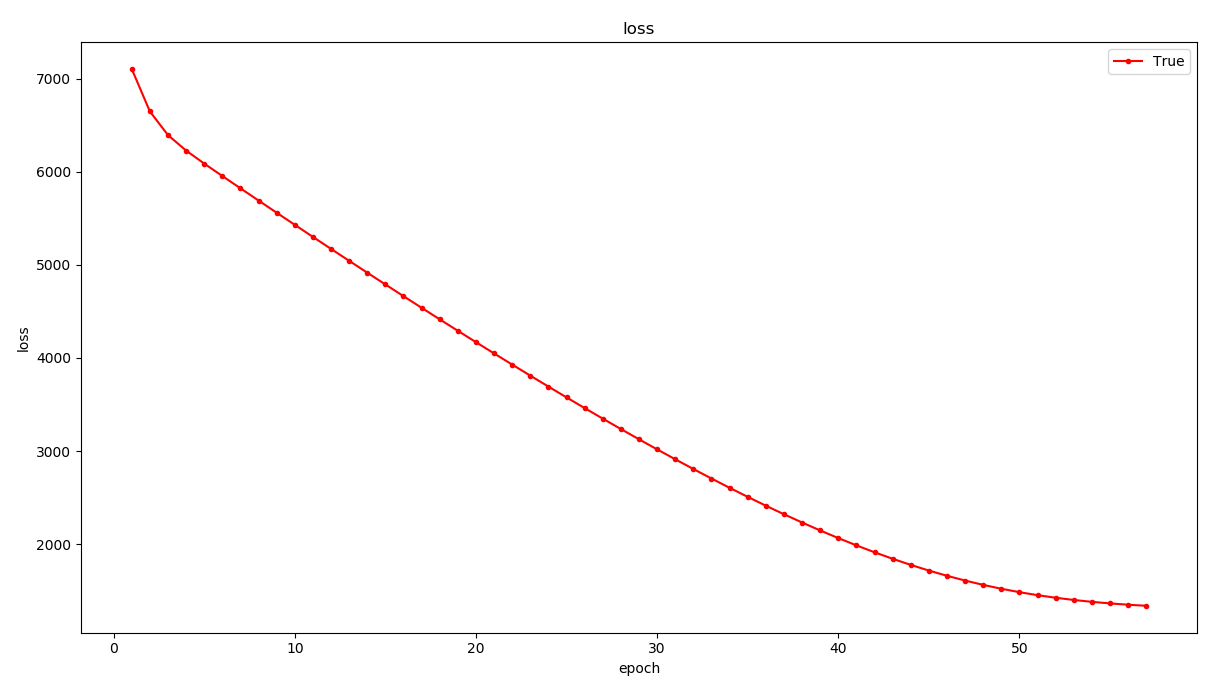


使用梯度下降算法，学习率设置为0.00001， epoch为200，error=0.01



#### 2.3.\_schillingdata为训练集时

使用梯度下降算法，学习率设置为0.00001， epoch为200，error=0.01



迭代停止时的参数作为模型最终参数, 括号里的为500次迭代结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 训练集  测试集 | 746Data | 1625Data | impensData | schillingData |
| 746Data | 0.6434（0.5643） | 0.4611（0.4611） | 0.4424(0.4397) | 0.4584 |
| 1625Data | 0.6794（0.7815） | 0.7618(0.7692) | 0.7003(0.7151) | 0.7668 |
| impensData | 0.6124（0.8036） | 0.8374(0.8426) | 0.7846(0.7972) | 0.8416 |
| schillingData | 0.6201（0.8151） | 0.8554(0.8673) | 0.7919(0.8133) | 0.8649 |

选择一个最优的参数（目标函数取得最小值时的参数）最为模型参数

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 训练集  测试集 | 746Data | 1625Data | impensData | schillingData |
| 746Data | 0.6448（0.6689） | 0.4571（0.6367） | 0.4410 | 0.4584 |
| 1625Data | 0.6775（0.6892） | 0.7575（0.7515） | 0.7145 | 0.7655 |
| impensData | 0.6114（0.6484） | 0.8374（0.7212） | 0.7962 | 0.8405 |
| schillingData | 0.6198（0.6299） | 0.8548（0.7249） | 0.8102 | 0.8652 |

在手动实现过程中，未考虑训练样本中正负样本平衡问题

### 3、结论

针对样本不平衡问题

1、可以对不足的样本进行过采样/重采样，SMOTE（即该算法构造的数据是新样本，原数据集中不存在的。该基于距离度量选择小类别下两个或者更多的相似样本，然后选择其中一个样本，并随机选择一定数量的邻居样本对选择的那个样本的一个属性增加噪声，每次处理一个属性。这样就构造了更多的新生数据）

2、对过多的样本进行欠采样（设小类中有N个样本。将大类聚类成N个簇，然后使用每个簇的中心组成大类中的N个样本，加上小类中所有的样本进行训练。优点是保留了大类在特征空间的分布特性，又降低了大类数据的数目），

3、可以考虑其他的评估指标，P，R，F1值。

其实发现一个问题，就是没有考虑阈值的选取问题，人为地认为正类的概率y>=1-y（负类的概率）时，样本可以判断未正类

746Data可能涵盖更多的数据类型，其他数据集做训练集得到的模型对746Data的分类准确率都很低