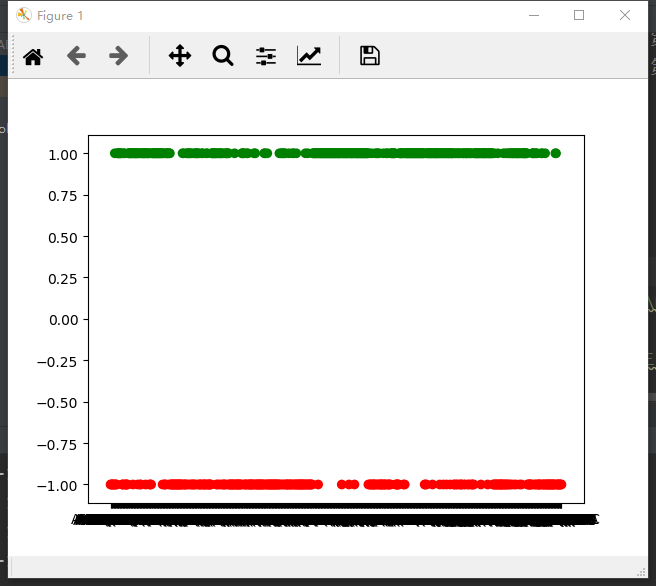
第二次训练——logistic回归

要求：利用逻辑回归完成数据集<https://archive.ics.uci.edu/ml/datasets/HIV-1+protease+cleavage>的二分类任务。

1. data analysis

该数据共包含四个数据集，内容为：八聚体(Octamers)、label，该标签指示该字符串是否代表肽(或蛋白质)中HIV-1蛋白酶切割它的位点(是则为+1，不是为-1)；代表氨基酸的字母有：AR ND CQ EG HI LK MF PS TW YV；

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| dataset | Octamers数量 | 裂解数 | 未裂解数 | 是否线性可分 |
| 746 | 746 | 401 | 345 | Y |
| 1625 | 1625 | 374 | 1251 | Y |
| Schilling | 3272 | 434 | 2838 | Y |
| Impens | 947 | 149 | 798 | Y |



将数据进行可视化，四个数据集的结果图都类似上图所示，其中绿色点为label=1的点，红色为-1的点，这四个数据集都是明显线性可分的。

实验原理：针对二分类问题，logistic回归输出的预测函数数学表达式为：

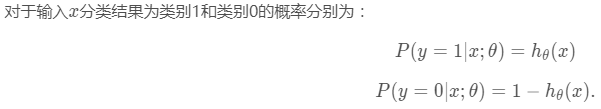
,(=1)

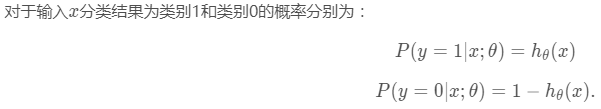
其中θ是参数向量，对于的直观解释为：对于给定的输入x，表示其对应的标签y=1的概率

上式取对数，则有被称为对数几率回归

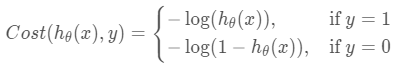
分子分母分别为正例、反例的概率。逻辑回归是对于特征的线性组合来拟合真实标记为正例的概率的对数几率,即为：







故损失函数为：



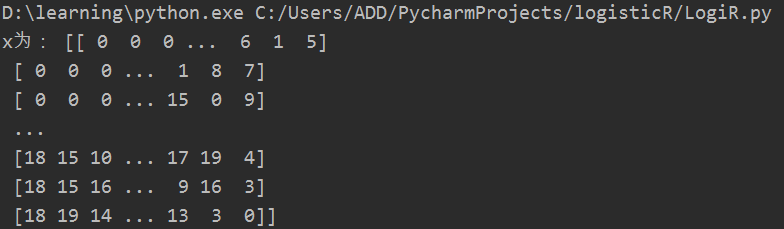
对最优参数θ的获取，除了梯度下降的算法，还可以利用牛顿法、拟牛顿法、共轭梯度法等，这里仍选择梯度下降法来最优化θ。

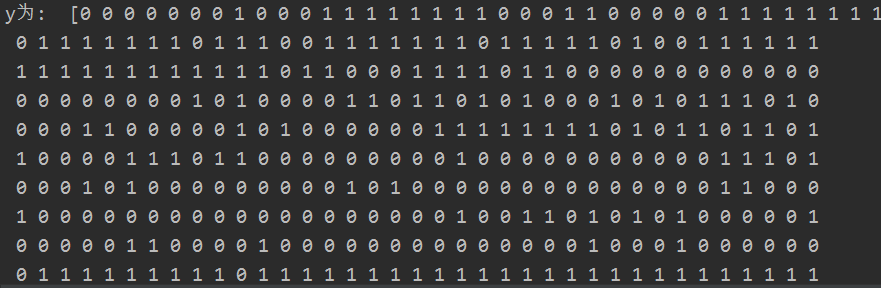
1. 设计实验
2. 划分数据、数据处理

数据的训练组合有很多种方式，此次做法是：选择其中一个数据集进行train，剩下的三个数据集进行test。

此外，由于属性x为字符串，这里要转化为数值量，’ARNDCQEGHILKMFPSTWYV’分别对应’0、1、2、3……19’,每一个字母对应于一个one-hot编码，每个八聚体则为每个字母的数字向量拼接而成的一个8\*20的一个向量矩阵。同理，y这里为{-1, 1}，应转化为{0, 1}，即把-1转化为0。

数据结果为：





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试集  训练集 | 746 | 1625 | impens | schilling |
| 746 | 0.697 | 0.692 | 0.664 | 0.632 |
| 1625 | 0.697 | 0.692 | 0.664 | 0.639 |
| impens | 0.461 | 0.461 | 0.843 | 0.867 |
| schilling | 0.461 | 0.461 | 0.843 | 0.867 |

1. 调包跑数据

如上表可看出，红色标记的组合，模型跑出来的分数比较高，这其中可能存在一个过拟合因素：样本的label分布不均衡。

1. 手写模型

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试集  训练集 | 746(lr=0.00001) | 1625(lr=0.00001) | impens(lr=0.0001) | schilling(lr=0.0001) |
| 746 | 0.694 | 0.694 | 0.699 | 0.689 |
| 1625 | 0.694 | 0.694 | 0.699 | 0.689 |
| impens | 0.461 | 0.461 | 0.843 | 0.865 |
| schilling | 0.461 | 0.461 | 0.843 | 0.867 |

参数设置为：最多迭代次数200次，测试集为746Data、1625Data的数据集学习率lr为0.00001，impens、schilling的lr=0.0001时，模型结果才会收敛，并且分数最高。

1. 总结
2. 对数据的分析：要把握数据集的成分、特点、属性，标签分布的均衡性
3. 手写模型与sklearn中的模型调包结果相差不大，在调节迭代次数与学习率的参数的时候能够控制结果的收敛情况。