Lab5实验报告

**一、实验目的：**

综合本课程所学习知识，用线性回归模型、决策树模型、神经网络模型、支持向量机以及XGBoost等分类模型来完成标签预测任务，并分析实验结果

**二、实验原理：**

1.线性回归模型：对于分类任务，在广义线性回归模型中，用sigmoid函数将线性回归模型的预测值与分类标签关联起来，即可得到对数几率回归（logistic regression）。

对于二分类任务，预测样本为正例的可能性为

当时将其划分为正例，否则为反例。

通过最大化对数似然

来确定模型中的参数。为了减少过拟合风险，常在上式中添加正则项，常用的为L1和L2正则项（为超参数，控制正则项强度）

2.决策树模型：对于分类任务，在决策树模型中，从根节点出发，在内部节点对特定属性进行测试，导出判定结果或进一步的判定问题（进入新的节点）。从根结点出发到每个叶结点的路径对于了一个判定测试序列。决策树的训练方式遵循简单的“分而治之”策略，每次划分的方式（属性和阈值）用信息熵（information entropy）增益或者划分后基尼系数（gini）来决定。该模型防止过拟合的主要手段是剪枝。

3.神经网络模型：多层感知机（MLP）是神经网络模型的一种，由多层“神经单元”组成，具有非常强大的拟合能力。神经网络常遭遇过拟合问题，通常用“早停”策略或添加正则项等方式解决。可调整参数：感知机激活函数（一般为：单位映射、sigmoid函数、Relu函数等）、网络层数以及每层神经元数、正则项大小、是否早停等

4.支持向量机：在分类任务中，该模型试图在数据空间中寻找一个将不同类型划分开的超平面，且使得两类数据距超平面的距离最大。训练该模型，即求解以下最优化问题：

可以转化为对偶问题用SMO、梯度下降等方法求解。

5.XGBoost：XGBoost是由多个基模型组成的一个加法模型，假设第个基本模型是 , 那么前 个模型组成的模型的输出为

在学习第 个基模型时，XGBoost 要优化的目标函数为:

其中loss为基学习器要最小化的误差函数，penalty为XGBoost模型的复杂度惩罚项。对比基学习器的优化公式，只需给第t个基决策树投放如下数据进行训练：

和神经网络一样，XGBoost模型的拟合能力很强，需要注意防止过拟合。

6.用二分类学习器完成多分类任务：常用手段为一对一（OvO）、一对多（OvR）、多对多（MvM）等。本次采用OvO模型，为了避免接口出错可能导致的问题，直接调用sklearn库的实现。

7.填补缺失数据：常用方法有：重采样法、用平均值（中位数）填补、K近邻方法填补（KNN）、直接过滤等

**三、实验步骤：**

0. 文件功能划分

main.ipynb：实验记录

Preprocess.ipynb：预处理记录，记录了数据预处理的草稿

    funs.py：将自己实现的学习算法的代码整理到其中

1. 数据预处理

  数据读取

去噪（有大量outliers，必须进行）：将离群值设置为缺失，交由第四步补全数据

处理缺失数据：将缺失特征填补为数据该特征上的平均值、KNN方法、过滤法（分别采用）

降维：Relieff方法、嵌入式选择

标准化：通过缩放，将每个特征的均值调整为0，方差调整为1

2. 对每个模型获取训练算法的实现（整理到funs.py中）或调包

    线性回归模型（Lab1）：由于数据集较大出现指数下溢，因此调用其他的实现

    决策树模型（Lab3，调包）

    神经网络模型（调包）

    支持向量机（Lab2）

    XGBoost模型（Lab3，调包）

3. 对每个模型都走一遍**调参**流程，选取最优的结果

    划分训练集和验证集

    设定参数进行训练

    根据训练好的模型进行验证，获得验证指标（采用**准确率**以及**weighted-accuracy**）

    调整参数，重复测试（采用交叉验证的方法）

    保存模型

5. 根据最好的模型做预测，保存结果

**四、实验结果**

**数据预处理**部分：

训练集标签分布：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | 0 | 1 | 2 | 3 |
| 数量 | 2523 | 2521 | 2500 | 2456 |

不难看出训练集的数据共含4种标签，且分布均匀

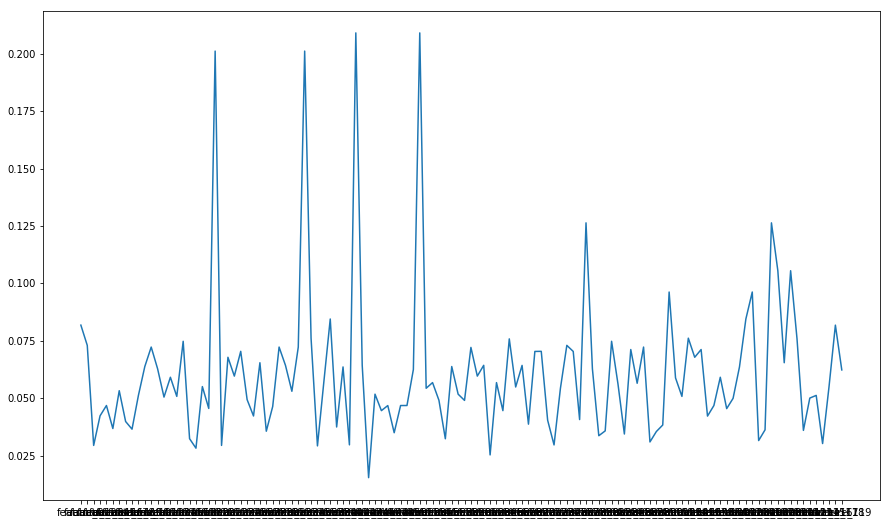
训练集特征分布：

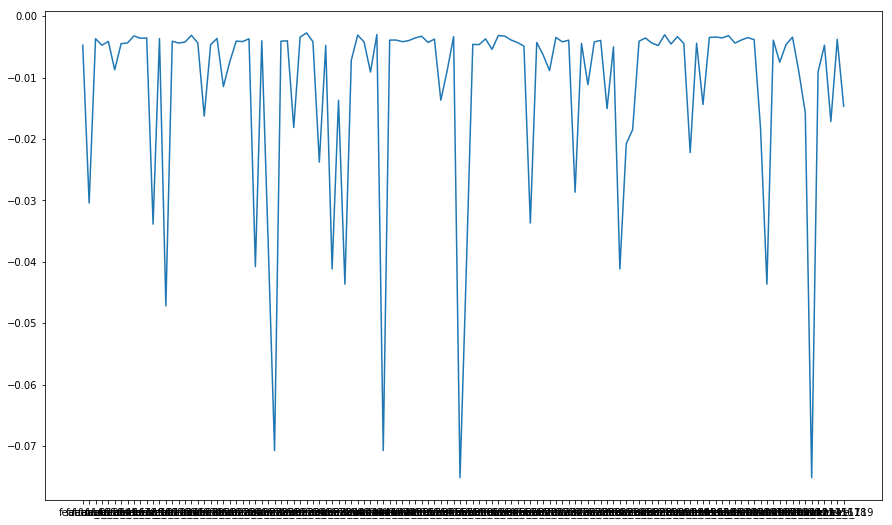
图略，详见附件main.ipynb，各个特征单独绘制

大部分特征取值接近正态分布和均匀分布；少部分特征取值较少，可能是离散型特征。所有特征都有少量的缺失值以及离群点，且往往离群点与正常点相差几个数量级。若将所有带有缺失值和离群点的数据删去，只剩下2415个样本，这会浪费大量数据，对学习过程不利。

特征相关性：

下图分别是是各个特征与其他特征之间**相关系数**的最大值和最小值（绘图形式选取不是恰当，但仍能看出问题），可以看出特征之间没有明显的相关性





特征选择：采用Relieff方法对特征进行筛选，得到特征权重如下：

（平均值法填充）

可以看出部分特征具有较大的权重，将它们筛选出来进行学习，可以有效提高训练速度

根据观察，选取前二十个特征进行学习

由于本次实验正确率非常不理想（采用了所提到的所有预处理方式，特征选取采用过滤式，嵌入式等方法，但泛化准确率总是在0.24到0.27之间浮动且不稳定，只有少数情况有所进展但是后来未能复现，如图所示），因此调参的标准只选准确率，且具体展开在此省略，只记录最优参数，详细过程在main.ipynb文件中。



线性回归模型：

Penalty：L1 tol： C：1 时得到最优准确率0.2535

神经网络模型：（选择特征20个）

单隐藏层：学习率：，tol：，α： 时得到最优准确率0.261

双隐藏层：学习率：，tol：，α： 时得到最优准确率0.264

决策树：

划分判据: entropy

max\_depth: 7

ccp\_alpha: 1

最优准确率：0.2523

支持向量机：（甚至无法收敛，只能手动指定最大迭代次数）

kernel：线性核（与之相对的是rbf）

Tol：0.001

C：0.5

最优准确率：0.2524

XGBoost：（15个特征）

learning\_rate: 1

max\_depth: 4

colsample\_bytree: 1

subsample: 0.8

最优准确率：0.2675

**五、实验分析**

本次实验准确率都不理想，主要有几个方面的问题：

1.多分类问题中，准确率本身就是较为严苛的指标

2.数据预处理时，缺失值和outliers可能含有重要信息被忽略

3.knn方法和relieff算法本身对冗余特征效果不理想

对于实验结果，主要的改进方向还是数据预处理和调参，以及补上缺失的假设检验步骤