文档

张苏月

1. 自定义函数说明

accuracy\_rate(predict,true\_value)

功能：计算模型在训练集上的正确率

参数：预测类别向量theta，真实类别向量y

返回值：该模型正确率

1. 流程说明
2. 数据处理。使用pandas.read\_csv引入数据集和测试集，统计缺失值小于2的行留下，再查看列缺失值数量为以后选择属性作参考（F列缺失值较多，训练时去除。）。
3. 缺失值填补。利用replace替换?为NAN，再用均值代替NAN。
4. 可视化处理。利用pandas.plt直接作出箱式图，观察异常值，I列数据异常，训练时不采用。
5. train\_test\_split分割数据集，实现计算正确率的函数。类别标签一定要减一，给的数据类别为1~9，但分类模型类别从零开始。
6. 利用基于决策树的lightlgb模型训练分割好的数据集，学习率为0.01，λ1惩罚项常数为0.1，λ2惩罚项常数为0.1，，最大深度为12，目标为多分类，类别数量为9。建立分类器并拟合。
7. 预测划分好的测试集，并利用classification\_report给出分值报告，利用accuracy\_rate查看最直接的正确率。
8. 用test集进行预测并把标签加一，结果转化为DataFrame形式作为submit\_test.csv输出并显示。
9. 遇到的问题及解决办法

|  |  |
| --- | --- |
| 问题 | 解决方法 |
| 用不做处理的数据做模型预测的时候报错。 | 数据类别标签减一 |
| KNN模型正确率在百分之二十 | 换模型为决策树 |
| 决策树正确率在百分之三十左右 | 换模型为lightlgb |
| 模型正确率还是不高 | 去掉缺失值过多的行和列，剔除异常值多的列。 |