实践2 Python数据分析与识别实践1

1、实验目的

(1) 掌握Python智能数据分析的流程

(2) 掌握数据缺失值的处理办法、掌握数据特征抽取方法

(3) 掌握机器学习模型在实际数据背景下的应用

(4) 学会自己分析给定的数据

2、实验平台

操作系统：Windows或Ubuntu

软件：Python 2.7或Python 3.6搭配Anaconda+Jupyter或Pycharm等

3、实验内容和要求

（一）窃电用户行为识别实践

阅读所给的文档，了解窃电用户识别流程，在给定的数据集上，分别完成如下步骤，并将结果截图：

(1) 通过拉格朗日插值法补全数据缺失（代码清单6-1）

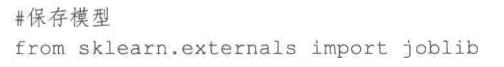
该用法已经过时，修改为：

y = s.reindex(list(range(n-k, n)) + list(range(n+1, n+1+k))) #取数

(2) 编写如下截图给出的混淆矩阵(cm\_plot.py)文件，下一步展示时需import该文件：



(3) 采用提示代码的决策树方法进行窃电用户行为识别（代码清单6-2、代码清单6-4、代码清单6-6），展示决策树模型的混淆矩阵和ROC曲线。



该用法已经过时，修改为（需事先安装好pip install joblib）：

import joblib

(4)采用K近邻模型、多元线性回归模型进行窃电用户行为识别，并与决策树进行对比

(5)（选做）该例子采用随机划分80%作为训练集，20%作为测试集，结果具有随机性。请修改代码采用留一交叉验证法，进行窃电用户行为识别。

（二）航空客户价值分析实践

阅读所给的文档，了解航空客户价值分析的过程和方法。在给定的数据集上，分别完成如下步骤（所给介绍文档中代码有一些缺失，以代码文档的分析过程为准），并将结果截图：

（1）数据预处理、数据规约和数据标准化运行结果：

（2）KMeans聚类运行结果：

（3）（选做）如何确定聚类数目？请尝试不同数目聚类的结果，对比分析，如何选择合适的聚类数目？

思考：给定具有一定背景的数据集后，如何进行数据清洗？缺失值如何处理？异常值如何处理（如机票价格为0属于异常值）？ 如何选择合适的模型进行分析（窃电识别有标签，选择有监督模型，用户行为无标签，选择无监督模型）？如何进行结果展示？

# 4.实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 《人工智能技术应用实践》上机实践2报告 | | | | |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： | | | | |
| 实验内容与完成情况： | | | | |
| 出现的问题： | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | | | |