特征工程作业 (管理类 + 开发类)

关键词

特征工程

数据描述

数据获取：
自己构建学习数据（可以通过Numpy等库进行构建）

学习目的

- 会使用Vscode开发环境，Python依赖安装
- 掌握课上涉及Python基本语法
- 理解特征预处理
- 理解特征选择
- 理解特征扩展
- 熟悉绘图
  实践过程可以参考sklearn， Numpy， Pandas， Matplotlib等相关文档。

环境及要求

- 通过python实现即可。有余力的同学可以通过sklearn加深理解和学习。
- 本项目建议使用python3.x 来完成。
- 参考前置安装文档安装环境。对于仅安装python解释器用户，由于非共性问题较多，建议下载相应编译环境后自行搜索安装和问题解决方法。对于直接安装anaconda的学员，可以直接在shell(cmd)中输入pip install sklearn自动完成sklearn库以及相关依赖的安装。

任务

参考程序：课程使用的代码示例文件夹：

- （1）特征工程：
  - 特征排序：通过递归特征消除，获取鸢尾花数据集特征排名。
  - 参考文件：a3_feature_engineering/feature_selection/wrapper/RFE.py
- （2）可视化：
  - （3.1）通过matlpotlib绘制y=x*xx* + 10函数。
  - 参考文件：a6_visualize/2_simply_plot.py
  - （3.2）绘制散点图，设置X = np.linspace(-2, 2, 20)，Y = 2 * X + 1。绘制散点图。
  - 参考文件：a6_visualize/4_scatter.py
- （3）使用鸢尾花数据集，
  - 分析4个特征统计信息和状况。
  - 通过最近邻（KNN）分类器,设置k = 5。预测花种类，尝试使用特征扩展或特征选择方式提升预测评分
  - 特征工程参考文件：a3_feature_engineering

评估

请确定你已完整的读过了这个任务，提交前对照检查过了你的项目，并按照提交要求完成任务。

有余力同学可以做选做题加深理解。

提交：

- PDF 报告文件，将代码实现过程以及心得。
- 项目相关代码（包括从raw data开始到最终结果以及你过程中所有代码）
- 包含使用的库，机器硬件，机器操作系统等数据的 README 文档（建议使用 Markdown）

泰坦尼克Titanic 求生预测 (开发类作业)

题目描述

本项目中需要利用监督学习算法对于Titanic获救概率进行分析。
通过对于性别、年龄等研究对于获救概率产生的影响。

数据描述

各列信息：

列名	属性
"row.names"	ID
"pclass"	类别
"survived"	是否获救
"name"	名字
"age"	年龄
"embarked"	从事工作
"home.dest"	家乡地址
"room"	房间
"ticket"	船票
"boat"	船号
"sex"	性别

要求

- 本项目要求使用python3.x + sklearn + pyccharm(或jupyter notebook)来完成。
- 鼓励用开源代码

任务

- 通过KNN完成预测（或使用内置的决策树进行预测）
- 进行数据探索分析
- 首先进行基本模型的实验，记录基准模型评测结果
- 绘制学习曲线
- 参考文件：a8_titanic/9_model_train.py