# 课内实验1 音乐流行度预测

指导教师：刘胥影 助教：陈诗泽

2023年3月31日

## 任务描述

构建线性回归模型，对音乐流行度进行预测。

## 数据集简介

人类与歌曲和音乐有着紧密的联系。音乐可以改善情绪，减轻疼痛和焦虑，为情感表达提供机会。研究表明，音乐可以在许多方面有益于我们的身体和心理健康。最近，多项研究已经开展，以了解歌曲和其受欢迎程度之间的关系。本数据集中，歌曲用多种指标度量对进行描述并记录为表格，任务是预测歌曲流行度。数据集包含18835个样本，每个样本包含13个属性与1个实值标记，属性包括连续数据属性和离散属性。要注意数据之间存在多重共线性，这增加了任务的挑战难度。

**目标**：根据歌曲指标预测歌曲流行度。

## 教学要求

1. 掌握线性回归模型的原理与构建
2. 掌握数据集预处理的常用方法，包括属性转换、归一化、共线性（multicollinearity）的检测与处理等。其中关于共线性要求自学，可参考 [1] [2]。
3. 掌握常用评估方法（本实验中为最基本的单次留出法）
4. 掌握实验结果的展示方式

## 实验要求

1. 使用Python编程构建线性回归算法，预测音乐流行度。
2. 数据集预处理：
   1. 离散属性连续化
   2. 连续属性归一化normalization
   3. 共线性的检测与处理
3. 使用80%的数据集进行训练，20%的数据集进行测试。需要报告训练误差和测试误差。请注意，需要保证模型能够在测试集上进行准确的预测。
4. 使用可视化技术来展示你的模型和数据集。可以使用Python的Matplotlib库来完成该任务。
5. 报告模型的性能和准确率。可以使用Python的Scikit-Learn库中的mean\_squared\_error函数来计算均方误差（MSE）。
6. 提交代码与报告，其中报告内容包括以下几个部分：
   1. 预处理
   2. 实验设置，包括实验评估方法与性能度量
   3. 实验结果，包括模型性能，可视化结果展示
   4. 结果分析
   5. 代码附录

## 提交要求

1. 提交的文件为压缩文件，其中包含：.py代码文件（不接受.ipynb或其他后缀的代码文件）、实验报告（doc或.docx），可能包含：模型文件（训练代码和预测代码分开的话需要）、运行脚本（视自己情况而定）
2. 提交的报告为doc或docx，代码需要复制后放在附录中。
3. 提交的代码需要一键运行，或者命令行（脚本）直接运行。批改设备是win10系统，使用pycharm运行。（脚本运行需要注意路径，路径错误扣分）
4. 模型使用pickle库保存。
5. 不要包含太多库导致批改设备运行时报错。

参考文献

|  |  |
| --- | --- |
| [1] | 流病与统计, “一看就懂的多重共线性,” [联机]. Available: https://zhuanlan.zhihu.com/p/355241680. |
| [2] | 喀喀湖, “线性回归的决定系数,” [联机]. Available: https://blog.csdn.net/weixin\_44005185/article/details/121273629. |

