

《人工智能导论》Project-2

一、银行精准营销解决方案

任务描述

根据提供的银行营销数据，构建合适的分类模型，预测客户是否会购买该银行的产品。

数据描述

数据见“data/classification/train_set.csv”

字段名称	数据类型	字段描述
ID	Int	客户唯一标识
Age	Int	客户年龄
Job	String	客户的职业
Marital	String	婚姻状况
Education	String	受教育水平
Default	String	是否有违约记录
Balance	Int	每年账户的平均余额
Housing	String	是否有住房贷款
Loan	String	是否有个人贷款
Contact	String	与客户联系的沟通方式
Day	Int	最后一次联系的时间（几号）
Month	String	最后一次联系的时间（月份）
Duration	Int	最后一次联系的交流时长
Campaign	Int	在本次活动中，与该客户交流过的次数
Pdays	Int	距离上次活动最后一次联系该客户，过去了多

		久（999 表示没有联系过）
Previous	Int	在本次活动之前，与该客户交流过的次数
Poutcome	String	上一次活动的结果
Y	Int	客户是否会订购定期存款业务

作业要求

● 模型

- 使用**不少于三种**分类器完成该分类任务，可以是单个分类器，也可以是通过 bagging, boost 等方式组合的分类器
- 其中自己实现的分类器**不少于一种**，并比较自己的实现效果和所使用的库（如 scikit-learn 或 matlab 等）中提供的实现之间的差异
 - ◆ 注：自己的实现不可以调用第三方的库，全部代码自己完成
- 对比各个算法在该任务上的效果差异并分析原因

● 数据

- 选择合适的字段作为输入特征
 - ◆ 可以手工选择，也可以使用特征选择算法进行选择
 - ◆ 请选择**不少于两种**的特征组合，说明选择的原因，并比较不同特征组合的实际效果

● 评价

- 由于提供的数据集全部为有标数据，因此建议采用 K-折交叉验证来验证模型
- 使用合理的评价指标，并说明使用该指标的原因

二、青蛙叫声聚类分析

任务描述

本任务根据青蛙的叫声所提取的 MFCC 特征，对不同科属的青蛙进行聚类分析。

数据描述

数据见“data/clustering”

- 数据集：青蛙叫声的 MFCC 特征数据
 - Frogs_MFCCs.csv: 数据集全集
 - Readme.txt: 数据集的说明文件
 - 数据集每一行代表一个青蛙叫声的特征样例。其中第 1 列到第 22 列表示 MFCC 的 22 维特征向量，后面 3 列分别表示这个青蛙属于哪个科 (Family)、属 (Genus)、种 (Species)，最后一列表示音频数据的记录来源。
- 默认对青蛙所属的科 (Family) 进行聚类结果的比较，感兴趣的同学也可以将属 (Genus) 和种 (Species) 作为聚类的目标。
 - 注：数据集中的青蛙共有 4 个科，8 个属，10 个种。

作业要求

- 数据
 - 选择合适的字段作为输入特征
 - ◆ 可以手工选择，也可以使用特征选择算法进行选择
 - ◆ 请选择**不少于两种**的特征组合，说明选择的原因，并比较不同特征组合的实际效果
- 模型与算法
 - 选择**至少两种**聚类模型或算法
 - 其中**至少一种**模型或算法为自己实现，并比较自己的实现效果和所使用

的库（如 `scikit-learn` 或 `matlab` 等）中提供的实现之间的差异

◆ 注：自己的实现不可以调用第三方的库，全部代码自己完成

- 了解音频特征提取的过程与背景知识，了解 MFCC 的算法，思考并设计距离度量方式，在实验报告中说明你的设计原因
- 使用合适的方法对聚类算法中的超参数进行选择，如 K-means 中的 K 等
- 评价
 - 使用**不少于两种**指标对聚类结果进行评估
 - 使用可视化的方法对聚类结果进行展示，如 t-SNE (https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding)

三、提交要求

- 本次作业总分 20 分，**个人独立**完成
- 内容
 - 所有实现作业要求所需的代码，语言不限
 - 文档需覆盖作业要求中的所有点，**使用模板撰写，中英文不限，不超过 5 页，提交 PDF 格式的文件**
- 时间：**2020 年 05 月 26 日 23:59 前**