

# 第三次大作业说明

## 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
- 对比各个算法在该任务上的效果差异并分析原因

### ● 数据

- 挑选合适的字段组成输入特征
- 考虑降维操作对结果的影响

### ● 评价

- 由于提供的数据集全部为有标数据，因此建议采用 N-fold cross validation 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、学生调查问卷结果分析

### 任务描述

根据提供的调查问卷的回答结果，对学生进行类型划分，分析每种类型的特点，并结合数据对学生、老师或学校提出建议。

### 数据描述

数据见“data/clustering/data.csv”

字段描述见“data/clustering/description.txt”

### 作业要求

- 数据
  - 选择合适的字段作为模型的输入
  - 比较是否对数据进行降维处理对效果的影响
- 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
  - 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`，或者 `DBSCAN` 中的 `epsilon` 等
- 评价
  - 使用常见的量化的聚类算法指标对结果进行评估，不少于**两种**
  - 使用可视化的方法对聚类结果进行展示
- 应用
  - 根据聚类结果，结合数据本身的特点，对学生、老师或学校提出不少于**两点**建议，如哪个老师的不满意率较高，是否集中在某些班级上，可能的原因是什么，是否需要建议学校重点关注等

### 三、提交要求

- 本次作业分组完成，沿用第二次大作业的分组
- 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：2019 年 12 月 16 日 23:59:59 前