《数据科学导引 B》期中项目

1. 选题一

名称：

Predict Health Outcomes of Horses

问题简介：

本次比赛的数据集（包括训练和测试）是从一个深度学习模型中生成的，该模型是在 马生存数据集 的一部分上训练的。特征分布与原始分布接近，但不完全相同。您可以随意使用原始数据集作为本次比赛的一部分，既可以探索差异，也可以看看在训练中加入原始数据集是否会提高模型性能。

这是一个三分类任务，目标变量是 outcome。

问题链接：

https://www.kaggle.com/competitions/playground-series-s3e22/overview

联系助教邮箱：

2301213107@stu.pku.edu.cn

2. 选题二

名称：

Sentiment Analysis on Movie Reviews

问题简介：

烂番茄电影评论数据集是一个用于情感分析的电影评论语料库。该项目是一个五分类问题，句子否定、讽刺、简洁、语言歧义等障碍使得这个任务非常具有挑战性，同学们需要对短语进行 5 个等级的标记：

- 0: negative 负面
- 1: somewhat negative 稍微负面
- 2: neutral 中立
- 3: somewhat positive 稍微正面
- 4: positive 正面

问题链接：

https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/overview

联系助教邮箱：

keer.lu@stu.pku.edu.cn

3. 选题三

名称：

ML Olympiad -Sustainable Urban Living

问题简介：

利用机器学习技术构建一个模型，根据详细信息预测房产的宜居性分数。

共有 13 个自变量，包括离散和连续取值，训练集中包含 31599 条数据。

使用测试集上的 RMSE 作为评价指标。

该竞赛已关闭，但仍可提交获取在测试集上的分数。

问题链接：

https://www.kaggle.com/competitions/ml-olympiad-sustainable-urban-living/over...

联系助教邮箱：

libozhou@pku.edu.cn

4. 选题四

名称：

Digit Recognizer

问题简介：

一个入门级的计算机视觉图像识别竞赛，使用了著名的 MNIST 数据集,参赛者的目标是从数万张手写测试图像中正确识别数字。

比赛的数据集包含训练集和测试集，其中训练集有 42000 条数据，测试集有 28000 条数据，每条数据包含 784 个像素点，即原始图像的像素为 28x28。训练集中的 Label 列表示手写数字的类别（共 10 个类别，0-9）。参赛者需要使用这些数据来训练模型，并预测测试集中手写数字图像的数字。

问题链接：

https://www.kaggle.com/competitions/digit-recognizer

联系助教邮箱：

zhaoks@stu.pku.edu.cn

5. 选题五

名称：

Tabular Playground Series

问题简介：

这是一个 Kaggle 聚类预测项目，比赛的目标是对一个无标签的数据集进行聚类，找到数据中的潜在模式，将数据中的每一行分配到对应的簇中。由于这是一个无监督聚类任务，因此没有提供训练数据和真实标签，也不会告知簇的数量。参赛者需要探索数据，并自行决定聚类的方法和簇的数量。

评价标准（比赛最终得分）为 Adjusted Rand Index（调整兰德指数），用来衡量预测簇标签与真实簇标签的相似程度。该指标考虑了随机簇分配的影响，能够更加客观地评价聚类效果。

推荐同学们按照课上讲述的探索性数据分析-建模评估的过程去完成比赛，并按照要求准备输出文件。比赛得分只是最终作业评分的一部分，书面报告的完整度同样重要。

问题链接：

https://www.kaggle.com/competitions/tabular-playground-series-jul-2022/overview

联系助教邮箱：

ytgeng@pku.edu.cn

要求

时间和提交方式：

选 5 道题中的一道，在 11 月 27 日前，发送课程报告（word 或 PDF 格式）到对应选题的助教邮箱。

邮件标题：姓名+学号+学院+题目 x

注：如果是团队，需要在邮件正文和课程报告里，均注明团队成员姓名+学号，并在课程报告里明确指出团队成员的分工。

报告内容：

无固定模板，可以包含问题描述、数据处理、算法建模、实验和分析、心得体会等

加分项：

1. 比现有最好的 kernel 得分最高

2. 独特的数据处理和建模方法
3. 课程报告清晰认真
4. 课程 Presentation（邀请部分同学）

组队形式：

6 人以内/个人均可