个人作业

1. 波士顿房价预测
2. 列车载客量预测
3. 心跳信号分类预测
4. 跨学科视角下的大模型应用探索

1. 波士顿房价预测

本数据集来源于美国人口普查局，包含1970年波士顿标准大都市统计区（SMSA）的住房价格相关信息。数据集规模较小，每个案例有14个属性，其中13个特征变量与房价密切相关，目标是通过分析这些特征与房价的关系，建立模型预测波士顿地区的房价。

数据集中各属性的具体描述及其含义：

1.CRIM - 城镇人均犯罪率【城镇人均犯罪率】
2.ZN - 占地面积超过25,000平方英尺的住宅用地比例【住宅用地所占比例】
3.INDUS - 每个城镇非零售业务的比例【城镇中非商业用地占比例】
4.CHAS - 查尔斯河虚拟变量（如果是河道，则为1；否则为0）【查尔斯河虚拟变量，用于回归分析】
5.NOX - 一氧化氮浓度（每千万份）【环保指标】
6.RM - 每间住宅的平均房间数【每栋住宅房间数】
7.AGE - 1940年以前建造的自住单位比例【1940年以前建造的自住单位比例】
8.DIS - 到波士顿五个就业中心的加权距离【与波士顿的五个就业中心加权距离】
9.RAD - 径向高速公路的可达性指数【距离高速公路的便利指数】
10.TAX - 每10,000美元的全额物业税率【每一万美元的不动产税率】
11.PTRATIO - 城镇的学生与教师比例【城镇中教师学生比例】
12.B - 1000 × (Bk - 0.63)²，其中Bk是城镇黑人的比例【城镇中黑人比例】
13.LSTAT - 低社会阶层人口比例（%）【房东属于低等收入阶层比例】
14.MEDV - 自有住房的中位数报价，单位：1000美元【自住房屋房价中位数】

评估指标：均方误差（Mean Squared Error, MSE）

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

文件说明：

- **train.csv - the training set**
- **test.csv - the test set**
- **submission.csv - a sample submission file in the correctformat**

PPT和报告提交至邮箱*ymlei@buaa.edu.cn*

https://bhpan.buaa.edu.cn/link/AA8741473E4C9F42F8A3798C5CD955E00C

提示：这是一个典型的回归问题。

2. 列车占用率预测

火车是许多人日常通勤和旅行的主要交通工具，但在高峰时段或热门线路上，列车可能会非常拥挤。为了帮助乘客更好地规划行程，可以通过机器学习技术预测未来列车的占用率。这不仅有助于乘客选择更合适的车次，还能为铁路部门优化调度提供支持。你需要需要利用训练数据构建模型，预测测试集中每趟列车的占用率，并生成符合要求的提交文件。

提交类别说明：

- 0 = low
- 1 = medium
- 2 = high

提示：这是一个典型的分类问题。

评估指标：准确率（accuracy）

文件说明：

- **train.csv - the training set**
- **test.csv - the test set**
- **submission.csv - a sample submission file in the correctformat**

PPT和报告提交至邮箱*ymlei@buaa.edu.cn*

https://bhpan.buaa.edu.cn/link/AA5CCA1CAF69DC48E1BFB4D03D4CE05701

3. 心跳信号分类预测

通过对心电图信号的分析，可以识别不同类型的异常心跳信号，从而帮助医生判断患者的心脏健康状况。然而，手动分析心电图数据耗时且容易受到主观因素的影响。本题以心电图信号数据为背景，希望你根据心电图感应数据预测心跳信号所属类别，这是一个多分类问题。每条数据为1列心跳信号序列，其中每个样本的信号序列采样频次一致，长度相等。你需要预测4种不同心跳信号的概率。

数据集包含：
训练集：80000条信号序列
测试集：20000条信号序列

评估指标：abs-sum

若真实值为$[y_1, y_2, y_3, y_4]$，模型预测概率为$[a_1, a_2, a_3, a_4]$
那么指标**abs-sum**的计算方式为：

$$abs - sum = \sum_{j=1}^{n} \sum_{i=1}^{4} |y_i - a_i|$$

文件说明：

- **train.csv - the training set**
- **test.csv - the test set**
- **submission.csv - a sample submission file in the correctformat**

PPT和报告提交至邮箱*ymlei@buaa.edu.cn*

https://bhpan.buaa.edu.cn/link/AA6B2B521834B44750A04553AA84E83A08

4. 跨学科视角下的大模型应用探索

随着大语言模型（LLM）如 ChatGPT、Claude、Gemini 等的快速发展，机器学习在各学科中的应用边界不断拓展。相比传统机器学习方法，大模型具备更强的泛化能力、推理能力和跨领域适应性。请结合你的专业背景或研究兴趣，自行选取一个曾经使用传统机器学习方法解决的问题，调研该问题在现阶段是否可以借助大语言模型进行改进或替代，可以围绕自己的科研、课程项目、行业兴趣等方向自由选择题目，给出使用大模型的解决方案。

作业要求

1.简要介绍选题背景，包括该问题的专业场景、重要性及相关已有的机器学习方法；
2.使用代码实现一个该问题中传统机器学习方法的简化示例；
3.展示如何使用大语言模型（如 GPT-4、Claude、Gemini、开源大模型等）解决该问题，可以通过 API、网页工具或本地部署实现；
4.对比分析传统方法与大模型方案在本问题中的表现与特点；

样例数据集（供参考）： 医疗搜索检索词意图分类

在医学搜索中，对搜索问题的意图分类可以极大提升搜索结果的相关性，特别是医学知识具备极强的专业性，对问题意图进行分类也有助于融入医学知识来做增强搜索结果的性能。

https://bhpan.buaa.edu.cn/link/AA3E00AAB12396481B84C53CC37A1B60C6

评估指标：PPT展示+提交报告

PPT和报告提交至邮箱*ymlei@buaa.edu.cn*

评估和提交

1、Evaluate_kit使用方法

在每个作业目录下，提供了 evaluate_kit 工具包，用于对模型的预测输出进行评估。以下是具体使用步骤：

步骤 1：准备预测结果

- 使用训练好的模型对测试数据集（test.csv）中的样本进行预测。
- 确保预测结果的格式与示例文件 submission.csv 完全一致，包括：
  - 列名需与 submission.csv 中的列名一致。
  - 数据的顺序需与 test.csv 中的样本顺序一致。
  - 预测值的数据类型需符合要求（如整数或浮点数）。

步骤 2：保存预测结果

- 将整理好的预测结果保存为一个 CSV 文件，命名为 submission.csv。
- 将该文件放置在 evaluate_kit 目录下，确保路径正确。

步骤 3：运行评估脚本

- 打开终端或命令行工具，进入 evaluate_kit 目录，运行以下命令以执行评估脚本：python evaluate.py
- 脚本会自动加载预测结果文件 submission.csv 和真实标签文件（answer.csv），并计算模型的效果指标。

2、提交

- 命名：学号_姓名_作业名称_展示.pptx、学号_姓名_作业名称_报告.pdf、学号_姓名_作业名称_代码
- 提交内容：PPT 演示文稿、实验报告、代码、模型最终结果
- 内容要求：包含对项目背景、数据集描述、模型设计、实验结果及分析的说明，使用图表（如混淆矩阵、损失曲线等）直观展示关键结果。
- 提交方式：请将上述文件压缩为一个ZIP文件，命名为学号_姓名_作业名称.zip，发送至***ymlei@buaa.edu.cn***。

团队作业

1. 地址相关性判断
2. 金融风险预测
3. 宠物年龄辨识

1. 地址相关性判断

地址文本相关性任务在现实世界中存在着广泛的应用场景，如：基于地理信息搜索的地理位置服务、对于突发事件位置信息的快速搜索定位、不同地址信息系统的对齐等等。

日常生活中输入的地址文本可以为以下几种形式：

- 包含四级行政区划及路名路号POI的规范地址文本；
- 地址要素缺省的规范地址文本，例：只有路名+路号、只有POI；
- 非规范的地址文本、口语化的地址信息描述，例：阿里西溪园区东门旁亲橙里；

地址文本相关性主要是衡量地址间的相似程度，地址要素解析与地址相关性共同构成了中文地址处理两大核心任务，具有很大的商业价值。

输入：

Query：江苏省南京市清水亭东路**9**号金域蓝湾**15**幢

Doc：江宁区万科金域蓝湾**15**栋

……

输出：

完全匹配

……

输入：

Query：江苏省南京市栖霞区西岗街道学森路**199**号保利罗兰春天**13**幢二单元

Doc：仙林湖学森路**199**号保利罗兰春天**9**号

输出：

部分匹配

1. 地址相关性判断

标注数据集中每条数据的格式为：


```
{
    "text_id":"1",
    "query":"华侨村西堤1巷12栋",
    " candidate":[{
        "text":"华侨新村西堤一巷12号", "label":"部分匹配"
        },
        {
        "text":"宝安区华侨新村西堤一巷", "label":"部分匹配"
        },
        {
        "text":"海丰县米巷西12幢", "label":"不匹配"
        },
        {
        "text":"余姚市大施巷村西片12号楼", "label":"不匹配"
        },
        {
        "text":"中山市西堤路一巷", "label":"不匹配"
        }]
}
```


评估指标：macro F1

g标签的召回率定义为：
Rg=|intersection(PREDg,GOLDg)|/|GOLDg|

g标签的准确率定义为：
Pg=|intersection(PREDg,GOLDg)|/|PREDg|

g标签的F1定义为：
F1g=2*Pg*Rg/(Pg+Rg)

```
macro F1 = (F1g1 + F1g2 + F1g3) / 3
```

文件说明：

- **train.jsonl - the training set**
- **test.jsonl - the test set**
- **submission.jsonl - a sample submission file in the correctformat**

PPT和报告提交至邮箱*ymlei@buaa.edu.cn*

https://bhpan.buaa.edu.cn/link/AAE7189040DBAC4526BC6B7535584C345C

2. 金融风险预测

以预测用户贷款是否违约为任务，该数据来自某信贷平台的贷款记录，总数据量超过80w，包含47列变量信息，其中15列为匿名变量。提交结果为每个测试样本是1的概率，也就是y为1的概率。评价方法为AUC评估模型效果（越大越好）。提交前请确保预测结果的格式与submission.csv中的格式一致，以及提交文件后缀名为csv。

部分字段说明如下（完整字段说明见数据集下载）：

Field	Description
id	为贷款清单分配的唯一信用证标识
loanAmnt	贷款金额
term	贷款期限（year）
interestRate	贷款利率
installment	分期付款金额
grade	贷款等级
subGrade	贷款等级之子级
employmentTitle	就业职称
employmentLength	就业年限（年）
homeOwnership	借款人在登记时提供的房屋所有权状况
annualIncome	年收入
verificationStatus	验证状态

评估指标：AUC-ROC

文件说明：

- **train.csv - the training set**
- **test.csv - the test set**
- **submission.csv - a sample submission file in the correctformat**

PPT和报告提交至邮箱*ymlei@buaa.edu.cn*

https://bhpan.buaa.edu.cn/link/AA84356E2AE19B4CFCB2DDC0C3C06F64C8

3. 宠物年龄辨识

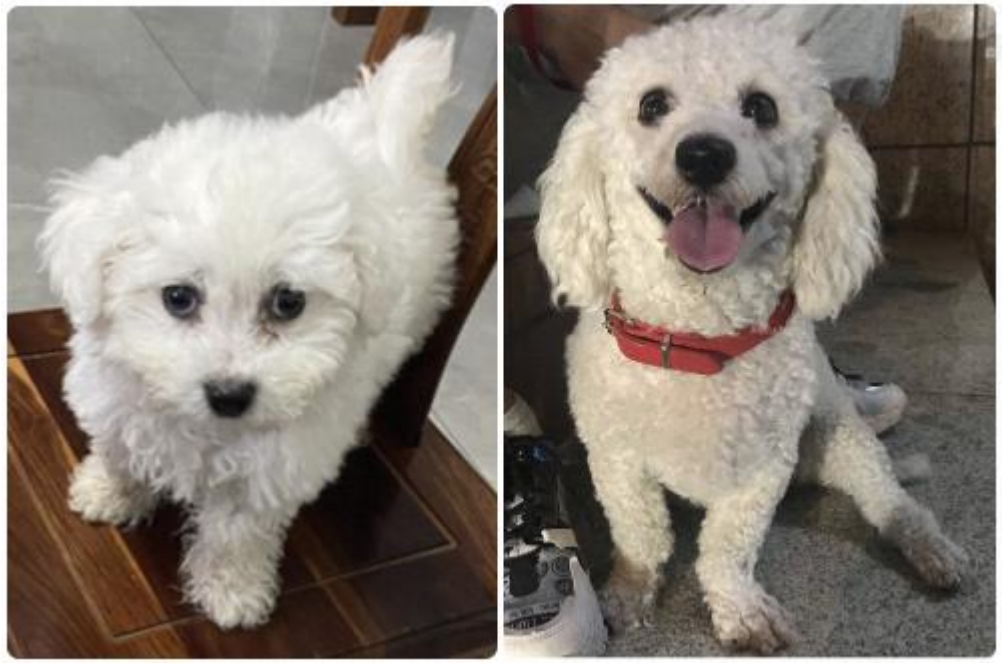
宠物医疗保险则是宠物经济中重要的组成部分，其保费规模也逐年攀升。随着宠物平均年龄的逐渐上升，宠物的出险率也在爬升。当务之急，需要在用户投保时，识别出投保宠物的实际年龄，有效拦截高龄宠物风险，将赔付率维持在健康水位。本次任务希望参赛者基于计算机视觉技术，通过宠物正脸图片，预估其年龄值。

数据集说明：

（1）本次比赛以宠物犬作为数据集，分为训练集（20000张）和测试集（3000张）。

（2）训练集和验证集为带噪数据集，噪声含量约6%，即存在**约6%**的数据标签可能不正确。测试集中不含噪声。

（3）由于宠物犬的寿命一般在15年左右（不超过16年），本次年龄标签范围为[0, 192)，年龄按月龄方式给出



评估指标：MAE

文件说明：

- **trainset.zip - the training set**
- **testset.zip - the test set**
- **annotations – the annotations of train set**
- **submission.txt - a sample submission file in the correct format**

PPT和报告提交至邮箱*ymlei@buaa.edu.cn*

https://bhpan.buaa.edu.cn/link/AAD0810C88A3994FE096D21038A7C2F581

评估和提交

1、Evaluate_kit使用方法

在每个作业目录下，提供了 evaluate_kit 工具包，用于对模型的预测输出进行评估。以下是具体使用步骤：

步骤 1：准备预测结果

- 使用训练好的模型对测试数据集（test.csv）中的样本进行预测。
- 确保预测结果的格式与示例文件 submission.csv 完全一致，包括：
  - 列名需与 submission.csv 中的列名一致。
  - 数据的顺序需与 test.csv 中的样本顺序一致。
  - 预测值的数据类型需符合要求（如整数或浮点数）。

步骤 2：保存预测结果

- 将整理好的预测结果保存为一个 CSV 文件，命名为 submission.csv。
- 将该文件放置在 evaluate_kit 目录下，确保路径正确。

步骤 3：运行评估脚本

- 打开终端或命令行工具，进入 evaluate_kit 目录，运行以下命令以执行评估脚本：python evaluate.py
- 脚本会自动加载预测结果文件 submission.csv 和真实标签文件（answer.csv），并计算模型的效果指标。

2、提交（仅组长提交）

- 命名：队长学号_姓名_作业名称_团队展示.pptx、队长学号_姓名_作业名称_团队报告.pdf、队长学号_姓名_作业名称_团队代码
- 提交内容：PPT 演示文稿、实验报告、代码、模型最终结果
- 内容要求：包含对项目背景、数据集描述、模型设计、实验结果及分析的说明，使用图表（如混淆矩阵、损失曲线等）直观展示关键结果。
- 提交方式：请将上述文件压缩为ZIP文件，命名为学号_姓名_作业名称.zip，发送至***ymlei@buaa.edu.cn***。