机器学习大作业

2025 年 4 月 20 日

1 任务目标

1. 模型选择与分析：比较不同回归/分类模型（如线性模型与非线性模型）的性能差异，分析重要模型参数的对预测的影响。

2. 特征学习技术：特征学习是提升机器学习性能的重要方法，比较各种不同的特征学习技术对实验结果的影响，包括但不限于：主成分分析 (PCA)、线性判别分析 (LDA) 等降维方法，以及特征选择与构造特征等技术。

3. 训练优化技术：探索数据预处理和模型正则化对训练过程的影响，研究超参数搜索策略对模型超参进行调优。

4. 模型集成策略：研究不同集成方法对预测性能的提升效果，分析集成规模与计算成本的平衡关系。

2 数据集

2.1 基本信息

请到 Canvas 大作业页面或文件页面中下载数据集 “TripDataset.zip”。

数据集详细信息请参考压缩包文件 “FeatureDescription.xlsx” 的各个工作表。

2.2 任务介绍

本研究包含以下两项预测任务：

任务 1（回归预测） 基于旅客的旅游特征数据，预测连续型目标变量：旅客停留天数（*Number of nights in CITY*）。

任务 2（分类预测） 根据旅客的旅游特征数据，预测类别型目标变量：旅客旅游目的（*Purpose of visit to CITY*）。

2.3 数据选择与划分

为保证数据的时效性，本研究选取 2015–2019 年间的样本进行建模。要求采用滚动预测策略，使用 `sklearn.model_selection.TimeSeriesSplit`（$n\_splits = 5$）对时间序列数据进行多折划分。在每一折中，分别计算回归任务与分类任务的性能指标详见5，最终对所有折的指标取平均值，以全面评估模型的平均性能与稳定性。

3 工具集推荐

实验推荐使用以下工具链：

表 1: 推荐工具清单

工具类型	推荐方案
编程语言	Python 3.10+
机器学习框架	scikit-learn
数据处理库	pandas
开发环境	Jupyter Notebook/Lab
可视化库	matplotlib/seaborn

4 评分标准

实验报告将根据以下维度进行评分：

表 2: 评分细则

评分维度	权重
实践探索完整度	30%
模型性能指标	30%
可视化与分析的逻辑性	20%
实验报告书写质量	20%

提交要求

- 抄袭必究，禁止使用 AI 生成报告
- 截止时间：第 16 周的周末（2025 年 6 月 8 日）晚上 12 点整。
- 命名规范：
  - 报告文件：`StuNum_Name_report.pdf`
  - 代码压缩包：`StuNum_Name_code.zip`

5 性能指标

5.1 回归任务

在回归任务中，常用的性能评估指标包括：

1. 均方误差 MSE：衡量预测值与真实值之间差异的平方的平均值。

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

2. 平均绝对误差 MAE：平均绝对误差计算预测值与真实值之间差异的绝对值的平均值。

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

3. 决定系数 R^2:，决定系数表示自变量对因变量变异的解释比例。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

请同学们分析具体任务要求，根据需求选择合适的评价指标（不仅限于上述提及的指标），需分析选择各个指标的理由。

5.2 分类任务

在多分类任务中，混淆矩阵不仅是一个简单的模型性能评估方法，更是深度理解模型行为、发现潜在问题并指导改进的重要手段。

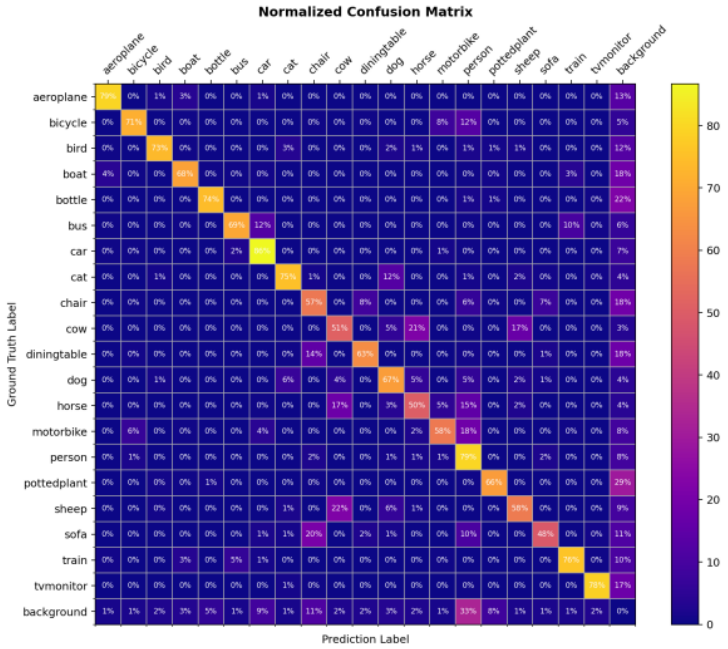

图 1: 多分类问题混淆矩阵

混淆矩阵通常是一个 $N \times N$ 的方阵，其中 N 是类别的数量。每一行代表实际类别，每一列代表预测类别。矩阵中的每个元素 M_{ij} 表示实际为类别 i，但被预测为类别 j 的样本数。

对于每个类别 C_i ，基于混淆矩阵可以计算以下指标：

- **TP**: 模型正确预测为类别 C_i 的样本数，即混淆矩阵中第 i 行第 i 列的值。
- **FP**: 模型错误预测为类别 C_i 的样本数，即混淆矩阵中第 i 列（去除对角线）所有值的总和。
- **FN**: 模型未能预测为类别 C_i 的样本数，即混淆矩阵中第 i 行（去除对角线）所有值的总和。
- **TN**: 模型正确预测为非类别 C_i 的样本数，即混淆矩阵中去除第 i 行和第 i 列的所有值的总和。

对应多分类问题的计算指标为：

$$\text{Accuracy}_i = \frac{\sum_{i=1}^{N} M_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}} \tag{4}$$

$$\text{Precision}_i = \frac{M_{ii}}{\sum_{j=1}^{N} M_{ji}} \tag{5}$$

$$\text{Recall}_i = \frac{M_{ii}}{\sum_{j=1}^{N} M_{ij}} \tag{6}$$

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{7}$$

在多分类任务中，评估模型性能时，常用的平均策略包括宏平均（Macro Average）和加权平均（Weighted Average）。

宏平均对所有类别的指标取平均，不考虑类别的不平衡。

$$\begin{aligned} \text{Macro Precision} &= \frac{1}{N}\sum_{i=1}^{N} \text{Precision}_i \\ \text{Macro Recall} &= \frac{1}{N}\sum_{i=1}^{N} \text{Recall}_i \\ \text{Macro F1} &= \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i \end{aligned} \tag{8}$$

加权平均对所有类别的指标按每个类别的样本数加权平均，考虑类别的不平衡。其中 $Support_i$ 表示类别 C_i 的样本数。

$$\begin{aligned} \text{Weighted Precision} &= \frac{\sum_{i=1}^{N}(\text{Precision}_i \times \text{Support}_i)}{\sum_{i=1}^{N} \text{Support}_i} \\ \text{Weighted Recall} &= \frac{\sum_{i=1}^{N}(\text{Recall}_i \times \text{Support}_i)}{\sum_{i=1}^{N} \text{Support}_i} \\ \text{Weighted F1} &= \frac{\sum_{i=1}^{N}(\text{F1}_i \times \text{Support}_i)}{\sum_{i=1}^{N} \text{Support}_i} \end{aligned} \tag{9}$$

对于多分类任务，AUC 和 ROC 通常是通过以下几种方式计算的，分别适用于不同场景：

1.（One-vs-Rest, OvR）一对多策略。将每个类别视为正类，其他类别视为负类，分别计算每个类别的 ROC 曲线和 AUC 值，取宏平均或加权平均。

2.（One-vs-One, OvO）一对一策略。将每一对类别视为一个二分类任务，单独计算其 ROC 曲线和 AUC 值，取宏平均或加权平均。

请同学们分析具体任务要求，根据需求选择合适的评价指标（不仅限于上述提及的指标），需分析选择各个指标的理由。