Отчет по Заданию №3

Юй Чанбай

14 май 2025г.

1 Цель работы

обработка кардиологического дата-сета для решения задач бинарной классификации

Загрузить 5000 строк из датасета: (см. файлы с названием «модуль 3….»): https://github.com/AI-is-out-there/data2lab.git. Сформировать обучающую выборку из загруженного датасета, состоящую из столбцов: ['Count_subj', 'rr_interval', 'p_end', 'qrs_onset', 'qrs_end', 'p_axis', 'qrs_axis', 't_axis', 'Healthy_Status'].

Датасет состоит из числовых параметров ЭКГ и классификационного признака Healthy_Status. Исследуй фреймворки AutoML, обоснуй выбор лучшего. Используй его для для бинарной классификации. Постройте матрицу ошибок (confusion matrix) и рассчитайте F1-метрику для оценки обученного классификатора по признаку Healthy_Status на основе данных параметров ЭКГ.

1 目标

处理心脏病学数据集以解决二分类问题

从数据集下载 5000 行数据（参见名为“模块 3”的文件）：https://github.com/AI-is-out-there/data2lab.git。从下载的数据集中生成训练样本，包含以下列:['Count_subj', 'rr_interval','p_end','qrs_onset','qrs_end','p_axis', 'qrs_axis', 't_axis', 'Healthy_Status']。

该数据集包含数值型心电图参数和“健康状态”分类特征。探索 AutoML 框架，论证最佳框架的选择。将其用于二分类。构建混淆矩阵并计算 F1 指标，以评估基于心电图参数数据训练的“健康状态”特征分类器。

2 Реализовано с использованием алгоритма AutoML (H2O)

2.1 Конфигурация среды H2O

Во-первых, вам нужно с помощью Anaconda создать новую виртуальную среду и

2 使用 AutoML（H2O）的算法实现

2.1 H2O 环境配置

首先需要使用 Anaconda 创建一个新的虚拟环境将其命名为 autoML H20，在这个环境

назвать ее AutoML H20, а также установить необходимые файлы в эту среду.

pip install h2o scikit-learn

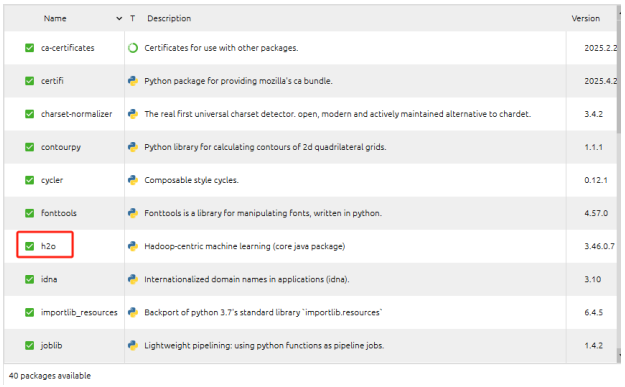

Рисунок 1.Сконфигурированная среда

2.2 Загрузка и подготовка данных

Скачайте файл модуль 3 и загрузите данные с помощью библиотеки pandas.В соответствии с требованиями задачи были выбраны только необходимые столбцы и первые 5000 записей.

2.3 Выберите подходящий фреймворк AutoML

Для выполнения этой задачи я выбрал фреймворк AutoML H2O.H2O AutoML - это масштабируемый,полностью автоматический и контролируемый алгоритм обучения, который может автоматически обучать большое количество моделей-кандидатов и объединять процессы интеграции в единую функцию.

2.4 Автоматическое обучение и результаты

Используйте автоматическое обучение H2O, установите максимальное количество моделей равным 10, тренируйтесь, чтобы получить оптимальную модель, получить оценку F1 и сгенерировать матрицу путаницы.

Все экспериментальные результаты показаны на рисунке ниже.

下安装所需要的文件 pip install h2o scikit-learn。

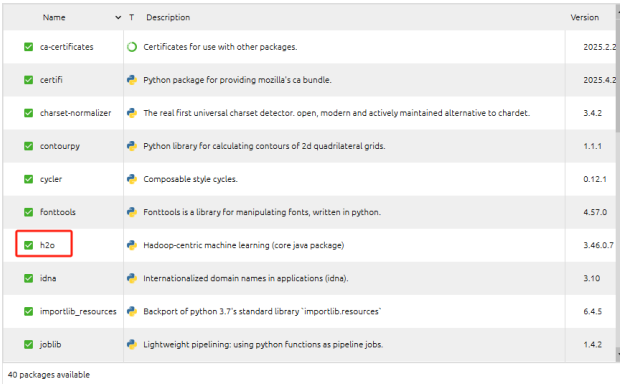

图 1. 配置好的环境

2.2 数据加载与准备

下载 модуль 3 文件，使用 pandas 库加载了数据。根据任务要求，仅选择了必需的列和前 5000 条记录。

2.3 选择合适的 AutoML 框架

为了完成此任务，我选择了 AutoML H2O 框架。H2O AutoML 是一个高度可扩展、全自动、有监督的学习算法，可在单个函数中自动训练大量候选模型和堆叠集成的过程。

2.4 AutoML 训练与结果

使用 H2OAutoML 训练，设置最大模型数为 10，训练得到最优模型，得到 F1 分数，生成混淆矩阵。所有实验结果如下图所示。

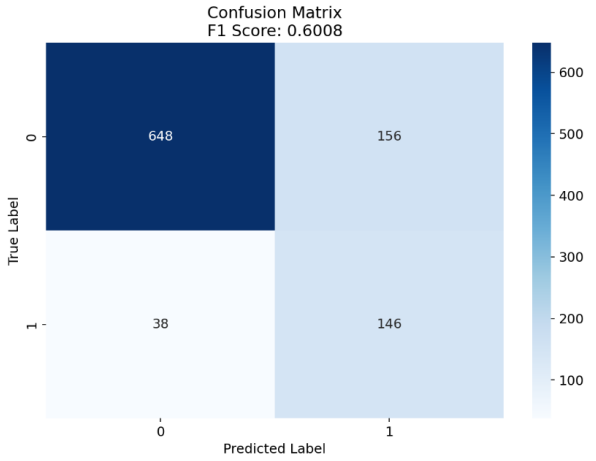

Рисунок 2.Матрица путаницы

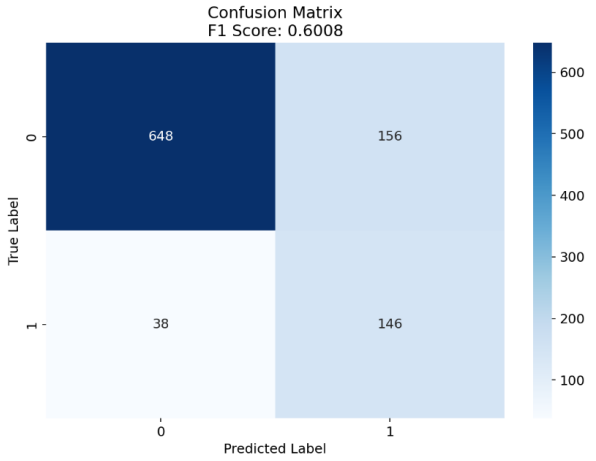

图 2. 混淆矩阵

```
==================================================
最佳模型：StackedEnsemble_AllModels_1_AutoML_1_20250514_103459
模型参数摘要：
dict_keys(['model_id', 'training_frame', 'response_column', 'validation_frame', 'blending_frame',
==================================================
```

Рисунок 3.Лучшая модель

```
==================================================
最佳模型：StackedEnsemble_AllModels_1_AutoML_1_20250514_103459
模型参数摘要：
dict_keys(['model_id', 'training_frame', 'response_column', 'validation_frame', 'blending_frame',
==================================================
```

图 3. 最佳模型

模型排行榜：

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_1_AutoML_1_20250514_103459	0.879161	0.317688	0.558766	0.193956	0.328236	0.107739
StackedEnsemble_BestOfFamily_1_AutoML_1_20250514_103459	0.878488	0.317967	0.55981	0.2027	0.328505	0.107916
GBM_2_AutoML_1_20250514_103459	0.878234	0.320203	0.561678	0.191842	0.328832	0.10813
GBM_grid_1_AutoML_1_20250514_103459_model_1	0.878157	0.323298	0.55313	0.202158	0.329819	0.10878
GBM_3_AutoML_1_20250514_103459	0.87533	0.325036	0.539255	0.205975	0.331185	0.109684
GBM_1_AutoML_1_20250514_103459	0.874938	0.322599	0.555649	0.204219	0.330502	0.109232
GBM_5_AutoML_1_20250514_103459	0.874523	0.324155	0.526855	0.199831	0.331457	0.109864
GBM_4_AutoML_1_20250514_103459	0.871259	0.331564	0.532887	0.195186	0.334726	0.112041
XRT_1_AutoML_1_20250514_103459	0.867723	0.331311	0.534685	0.222392	0.334353	0.111792
DRF_1_AutoML_1_20250514_103459	0.867382	0.331026	0.534927	0.205367	0.334417	0.111835

[10 rows x 7 columns]

Рисунок 4.Данные модели

模型排行榜：

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_1_AutoML_1_20250514_103459	0.879161	0.317688	0.558766	0.193956	0.328236	0.107739
StackedEnsemble_BestOfFamily_1_AutoML_1_20250514_103459	0.878488	0.317967	0.55981	0.2027	0.328505	0.107916
GBM_2_AutoML_1_20250514_103459	0.878234	0.320203	0.561678	0.191842	0.328832	0.10813
GBM_grid_1_AutoML_1_20250514_103459_model_1	0.878157	0.323298	0.55313	0.202158	0.329819	0.10878
GBM_3_AutoML_1_20250514_103459	0.87533	0.325036	0.539255	0.205975	0.331185	0.109684
GBM_1_AutoML_1_20250514_103459	0.874938	0.322599	0.555649	0.204219	0.330502	0.109232
GBM_5_AutoML_1_20250514_103459	0.874523	0.324155	0.526855	0.199831	0.331457	0.109864
GBM_4_AutoML_1_20250514_103459	0.871259	0.331564	0.532887	0.195186	0.334726	0.112041
XRT_1_AutoML_1_20250514_103459	0.867723	0.331311	0.534685	0.222392	0.334353	0.111792
DRF_1_AutoML_1_20250514_103459	0.867382	0.331026	0.534927	0.205367	0.334417	0.111835

[10 rows x 7 columns]

图 4. 模型数据

3 Заключение

Задача по использованию AutoML для обработки наборов кардиологических данных и построения бинарного классификатора была успешно выполнена.Использование H2OAutoML позволяет эффективно сравнивать несколько моделей с минимальными затратами на кодирование и получать результаты оценки производительности классификатора.

3 结论

使用 AutoML 处理心脏病学数据集并构建二元分类器的任务已成功完成。H2OAutoML 的使用使得能够以最少的编码工作高效地比较多种模型，并获得分类器性能的评估结果。