**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра Кафедра МОЭВМ**

ОТЧЕТ

по учебной практике НИР

Тема: Оптимизация использования вычислительных ресурсов при решении задач автоматизированного машинного обучения.

|  |  |  |
| --- | --- | --- |
| Студент гр. 8303 |  | Гусев К. А. |
| Руководитель |  | Жукова Н. А. |

Санкт-Петербург

2023

**ЗАДАНИЕ**

**НА НАУЧНО-ИССЛЕДОВАТЕЛЬСКУЮ РАБОТУ**

|  |  |  |
| --- | --- | --- |
| Студент Гусев К. А. | | |
| Группа 8303 | | |
| Тема практики:Разработка метода настройки алгоритмов автоматического машинного обучения для решения задач классификации временных рядов | | |
| Задание на практику:  Выбрать библиотеку для работы с временными рядами, способную решать задачи классификации. Ознакомится с ее функционалом, провести набор экспериментов с различными конфигурациями параметров и сделать выводы на основе полученных данных. | | |
| Сроки прохождения практики: : 01.09.2023 – 20.12.2023 | | |
| Дата сдачи отчета: 20.12.2023 | | |
| Дата защиты отчета: 20.12.2023 | | |
|  | | |
| Студент |  | Гусев К. А. |
| Руководитель |  | Жукова Н. А. |

**Аннотация**

В работе приведен обзор на AutoML библиотеку AutoGluon. Выполнен ряд экспериментов, выявлены зависимости базовых параметров на итоговые результаты. Сделаны предположения по поводу итоговых технических решений на весенний семместр 2024 года.

**Summary**

The work provides an overview of the AutoML library AutoGluon. A number of experiments were performed and the dependences of the basic parameters on the final results were identified. Assumptions are made regarding the final technical solutions for the spring semester of 2024.

**Содержание**

|  |  |  |
| --- | --- | --- |
|  | Определения и обозначения | 5 |
|  | Введение | 6 |
| 1. | Результаты работы в осеннем семестре | 7 |
| 1.1. | Ход работы | 7 |
| 2. | Описание предполагаемого метода решения | 13 |
| 2.1. | Предполагаемые решения | 13 |
| 3. | План работы на весенний семестр | 14 |
| 3.1. | Ход работы | 14 |
|  | Заключение | 15 |
|  | Список использованных источников | 16 |

**Определения и обозначения**

ML – машинное обучение – сложный, многоэтапный процесс, который поддается автоматизации: данные загружают в модель, обученная реализация обрабатывает данные и выдает предсказание, которое можно использовать для аналитики.

AutoML – автоматическое машинное обучение – инструмент, который позволяет полностью автоматизировать применение методов ML к реальным задачам

HPO – оптимизация гиперпараметров

Гиперпараметр – В машинном обучении гиперпараметрами называют параметры алгоритмов, значения которых устанавливаются перед запуском процесса обучения. В этом смысле они и отличаются от обычных параметров, вычисляемых в процессе обучения. Гиперпараметры используются для управления процессом обучения.

Модель – подход к решению поставленной задачи

MAC – число операций умножения-накопления

FLOP – число операций с плавающей точкой

Low fidelity – Низкая точность

Surrogate method – Суррогатный метод

Early stopping – Ранняя остановка

Resource-aware – Оптимизация ресурсов

Аугментация данных — методика создания дополнительных данных из имеющихся данных. Чаще всего, проблема ограниченного набора данных возникает при решении задач, связанных с обработкой изображений.

Миноритарным класс – часть набора данных, которая скудно представлена

Мажоритарным класс – часть набора данных, которая представлена изобильно

Pipeline – пайплайн – конвейер

**введение**

Автоматизированное машинное обучение, которое также называется автоматизированным ML или AutoML, представляет собой процесс автоматизации трудоемких и многократно повторяющихся задач разработки моделей машинного обучения.

Машинное обучение (ML) добилось значительных успехов в последние годы, и на него опирается постоянно растущее число дисциплин. Однако этот успех в решающей степени зависит от специалистов по машинному обучению, которые выполняют следующие задачи:

* Предварительная обработка и очистка данных.
* Выбор и построение соответствующих функций.
* Выбор подходящего семейства моделей.
* Оптимизация гиперпараметров модели.
* Разработка топологии нейронных сетей (если используется глубокое обучение).
* Анализ полученных результатов.

Поскольку сложность этих задач в большинстве случаев не под силу специалистам, не специализирующимся на машинном обучении, быстрый рост приложений машинного обучения создал спрос на готовые методы машинного обучения, которые можно легко использовать без специальных знаний. Область исследования, нацеленная на прогрессивную автоматизацию машинного обучения назвали – AutoML.

**1. ПЛАН РАБОТЫ НА осенний СЕМЕСТР**

В результате обсуждения темы с научным руководителем, было принято решение изменить тему работы с «Оптимизация использования вычислительных ресурсов при решении задач автоматизированного машинного обучения.» на «Разработка метода настройки алгоритмов автоматического машинного обучения для решения задач классификации временных рядов »

1.1 Ход работы

В результате исследования нового набора библиотек, я выбрал библиотеку AutoGluon [1]

Причины выбора данной библиотеки следующие:

* Наличие инструментов для работы непосредственно с временными рядами (по данному критерию не проходит большинство библиотек и решений, рассматриваемых мной в предыдущем семестре)
* Полное соответствие технических характеристик моей машины и требований для установки (исключены библиотеки AutoKeras[2], AutoPyTorch-TS[3], AutoTS[4])
* Удобство и возможность установки (исключены библиотеки AutoKeras, AutoPyTorch-TS, AutoTS)
* Объем базового инструментария (исключены библиотеки Azure[5])
* Возможность работы с комплексом временных рядов (исключены библиотеки AutoKeras)

Для исследования была выбрана группа временных рядов, приведенная к следующей структуре:

Таб. 1 – Структура исследуемых временных рядов

|  |  |  |
| --- | --- | --- |
| item\_id | timestamp | target |
| ID временного ряда | Временная точка (каждый час, для данной группы временных рядов) | Значение в конкретной временной точке |

Всего существует 157660 временных точек, что доводит количество элементов временного ряда до 472980

Полный временной ряд можно найти в репозитории проекта на github [6]

Для начала рассмотрим особенность загрузки необработанных входных данных

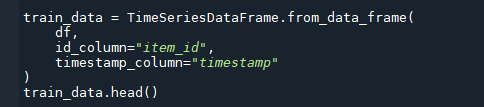
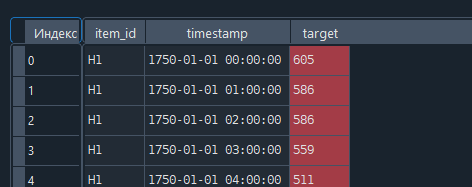


Рис. 1 — Минимально возможное описание входных данных

Набор необработанных данных всегда должен соответствовать этому формату и содержать как минимум три столбца для уникального идентификатора, метки времени и целевого значения, но имена этих столбцов могут быть произвольными. [7]

Собственно как можно заметить по Таб. 1 исследуемые временные ряды, был приведен к подобному формату.

Рис. 2 — Хранение необработанных входных данных

Таб. 2 — Описание аргументов

|  |  |
| --- | --- |
| Название аргумента | Описание |
| eval\_metric | Выбор метрики оценки |
| hyperparameters | Выбор моделей для предиктора |
| known\_covariates\_names | это изменяющиеся во времени характеристики, которые могут влиять на целевой временной ряд. Иногда их также называют динамическими признаками, экзогенными регрессорами или связанными временными рядами. AutoGluon поддерживает два типа ковариат:  известные ковариаты, известные для всего горизонта прогноза  прошлые ковариаты, которые известны только до начала горизонта прогнозирования |
| num\_val\_windows | Количество окон проверки, чем их больше, тем меньше вероятность переобучения, но прямо пропорционально количеству окон, увеличивается время работы |
| prediction\_length | «дальность» предсказывания возможных значений |
| quantile\_levels | представляет квантиль распределения прогноза. Например, если квантиль 0.1 (также известный как P10 или 10-й процентиль) равен x, это означает, что прогнозируемое значение временного ряда будет ниже x 10 % времени. |
| target | Название исследуемого параметра временного ряда |
| time\_limit | Лимит времени работы |

В AutoGluon сущесвует 4 пресета качества работы, более высокая степень качества, увеличивает как время работы, так и качество работы, подробный обзор степеней качества представлен в Таб.3

Таб. 3 — Описание пресетов качеств работы

|  |  |  |  |
| --- | --- | --- | --- |
| Название | Описание | Рекомендуемые случаи использования | Подходящее время (идеальное) |
| fast\_training | Подбор простых статистических и базовых моделей + быстрые древовидные модели | Быстро обучается, но может быть не очень точным | 0.5x |
| medium\_quality | Те же модели, что и в fast\_training + модель глубокого обучения TemporalFusionTransforme | Хорошие прогнозы с разумным временем обучения | 1x |
| high\_quality | Более мощные модели глубокого обучения, машинного обучения и статистического прогнозирования. | Гораздо точнее, чем medium\_quality, но обучение занимает больше времени | 3x |
| best\_quality | Те же модели, что и в high\_quality, больше окон перекрестной проверки | Обычно более точный, чем high\_quality, особенно для наборов данных с небольшим количеством (<50) временных рядов | 6x |

Таб. 4 — Описание метрик оценки

|  |  |
| --- | --- |
| Название | Описание |
| SQL | Масштабированные квантильные потери. |
| WQL | Взвешенная квантильная потеря. |
| MAE | Средняя абсолютная ошибка. |
| MAPE | Средняя абсолютная процентная ошибка. |
| MASE | Средняя абсолютная масштабированная ошибка. |
| MSE | Среднеквадратическая ошибка. |
| RMSE | Среднеквадратическая ошибка. |
| RMSSE | Среднеквадратическая масштабированная ошибка. |
| SMAPE | Симметричная средняя абсолютная процентная ошибка. |
| WAPE | Взвешенная абсолютная процентная ошибка. |

Метрика оценки выбирается под конкретную задачу. По умолчанию рекомендованная — MASE.

Мной была проведена серия экспериментов, на выбранной группе временных рядов.

Изменяемыми параметрами стали:

* prediction\_length
* time\_limit
* пресеты качества
* метрика оценки

Данные параметры были выбраны, т. к. они влияют на качество работы с любой задачей, именно они не могут быть указанны как параметры по умолчанию

Таб. 5 — Результаты экспериментальных запусков

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Номер эксперимента | prediction\_length (временные точки) | time\_limit (секунды) | пресет качества | метрика оценки | Фактическое время работы  (секунды) | Счет модели (качество работы) |
| 0 | 48 | 600 | medium\_quality | MASE | 460.98 | -0.9247 |
| 1 | 48 | 1500 | medium\_quality | MASE | 937.26 | -0.9240 |
| 2 | 60 | 600 | medium\_quality | MASE | 447.27 | -1.0234 |
| 3 | 32 | 600 | medium\_quality | MASE | 426.47 | -0.8192 |
| 4 | 60 | 1500 | medium\_quality | MASE | 902.82 | -1.3182 |
| 5 | 48 | 600 | high\_quality | MASE | 549.92 | -0.9210 |
| 6 | 48 | 1500 | high\_quality | MASE | 1284.25 | -0.9029 |
| 7 | 48 | 600 | fast\_training | MASE | 252.51 | -0.9055 |
| 8 | 48 | 600 | medium\_quality | MAE | 415.43 | -552.2477 |
| 9 | 48 | 600 | medium\_quality | WAPE | 398.03 | -0.0375 |

В AutoGluon более высокие баллы всегда соответствуют лучшей эффективности прогнозирования. Поэтому оценки умножаются на -1, так что более высокие «отрицательные значения» соответствуют более точным прогнозам. [7]

Для используемых оценок наиболее подходят данные со следующими характеристиками [8]:

* MAE — зависит от масштаба (временные ряды с большим абсолютным значением вносят больший вклад в потери); предпочтительны модели, которые точно оценивают медиану [8]
* MASE — следует использовать только в том случае, если все временные ряды имеют положительные значения; не определено для временных рядов, содержащих нулевые значения; наказывает за завышение прогноза более строго, чем за недооценку [8]
* WAPE — масштабированная метрика (нормализует ошибку для каждого временного ряда по шкале этого временного ряда); не определено для постоянного временного ряда; не чувствителен к выбросам; предпочитает модели, которые точно оценивают медиану [8]

Полные логи работы, графики прогнозов, код программы приведены в репозитории проекта.

Выводы:

Наибольшее значение на итоговый счет модели и точность прогнозов влияет метрика оценки

Пресет качества может улучшать работу незначительно, но при этом можно значительно снизить время работы

Чем больше «дальность» предсказывание, тем выше счет модели и меньше время, но это верно не для всех пресетов

Увеличение общего лимита времени работы, может негативно влиять на счет модели

**2. Описание предполагаемого метода решения.**

2.1 Предполагаемые решения:

Из выводов полученных в предыдущем пункте, можно сделать общие предположения на счет технических решений

* Библиотека AutoGluon, показала достойные результаты с высокой вероятностью, именно она будет использоваться в итоге
* С большей вероятностью будет выбран средний или небольшой лимит времени (до 10 мин)
* Для большого объема данных и прогнозирования медианы больше всего подходит метрика оценки MAE, но может быть использована другая метрика, в зависимости от типа задачи и объема данных

**3. ПЛАН РАБОТЫ НА весенний СЕМЕСТР**

3.1 Ход работы

* Получить задачу/задачи для исследования
* Собрать или получить набор/наборы временных рядов, необходимый для реализации итоговой задачи
* Выполнить точную настройку параметров для решения конкретной поставленной задачи

**заключение**

В ходе данной работы был проведен обзор на Auto ML библиотеку AutoGluon для работы с временными рядами, исследовалась зависимость настраиваемых параметров на длительность и качество работы AutoML системы . Продолжена работа над магистерской диссертацией.

**список использованных источников**

1. Описание AutoGluon URL: https://auto.gluon.ai/stable/index.html Дата обращения (17.08.2023)

2. Описание AutoKeras URL: https://autokeras.com/tutorial/timeseries\_forecaster/ Дата обращения (13.08.2023)

3. Описание AutoPyTorch-TS URL: https://2022.ecmlpkdd.org/wp-content/uploads/2022/09/sub\_759.pdf Дата обращения (13.08.2023)

4. Описание AutoTS URL: https://winedarksea.github.io/AutoTS/build/html/source/tutorial.html Дата обращения (13.08.2023)

5. Описание Azure URL: https://learn.microsoft.com/ru-ru/AZURE/machine-learning/v1/how-to-auto-train-forecast?view=azureml-api-1 Дата обращения (13.08.2023)

6. Репозиторий проекта URL: https://github.com/KirillGusev0/Gusev\_2023\_NIR Дата обращения (18.12.2023)

7. Описание базовых параметров AutoGluon URL: https://auto.gluon.ai/stable/tutorials/timeseries/forecasting-quick-start.html# Дата обращения (19.08.2023)

8. Описание оценочных метрик AutoGluon URL: https://auto.gluon.ai/stable/tutorials/timeseries/forecasting-metrics.html#autogluon.timeseries.metrics.MAPE (1.11.2023)