ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчёт по лабораторной работе № 4

«Подбор признаков датасета WineQT»

Выполнил работу

Фиолетов Эдуард

Академическая группа №3114

Принял

Дунаев М.В.

Санкт-Петербург

2024

Отчёт по лабораторной работе

Введение

Цель данной работы — реализация алгоритма отбора признаков для улучшения качества модели машинного обучения. Задачами лабораторной работы являются:

1. Изучение методов загрузки данных и их предобработки.
2. Реализация алгоритма перебора комбинаций признаков для определения оптимального набора.
3. Оценка производительности модели машинного обучения с различными наборами признаков.
4. Анализ сложности алгоритма и построение графиков зависимости времени работы от количества элементов.

Теоретическая подготовка

Основные понятия

- **Отбор признаков** — процесс выбора подмножества значимых признаков из исходного набора данных для повышения производительности модели.
- **Метрика RMSE (Root Mean Squared Error)** — используется для оценки качества регрессионной модели. Рассчитывается как квадратный корень из среднего значения квадрата ошибок.

Используемые алгоритмы

- **Полный перебор комбинаций признаков**:
    - Алгоритм перебирает все возможные подмножества признаков.
    - Для каждого подмножества оценивается качество модели
    - Выбирается комбинация с наименьшим значением RMSE.
- **Линейная регрессия**:

- Простой метод машинного обучения, который ищет линейную зависимость между признаками и целевой переменной.

Используемые структуры данных

1. **Armadillo Matrix** — структура для представления матриц в памяти, поддерживающая эффективные вычисления.
2. **std::vector** — контейнер для хранения индексов и промежуточных результатов.

Реализация

Этапы выполнения

1. **Загрузка данных**:
   1.1. Используется библиотека mlpack для чтения CSV-файлов.
   1.2. Удаляются незначимые столбцы (ID, целевая переменная) с использованием функции drop_columns.
2. **Реализация алгоритма перебора**:
   2.1. Генерируются все подмножества признаков с помощью битовых масок.
   2.2. Для каждой комбинации вычисляется RMSE.
   2.3. Ведется учет лучшего результата.
3. **Оценка модели**:
   3.1. Функция evaluate_dataset обучает модель линейной регрессии и возвращает метрику RMSE.
4. **Вывод результатов**:
   4.1. На консоль выводятся текущая комбинация признаков, её RMSE, а также лучшая комбинация.

Ключевые фрагменты кода

1. Загрузка данных:

```
arma::mat dataset;
    if (!mlpack::data::Load(path, dataset)) {
        throw std::runtime_error("Could not read *.csv!");
    }

    // get target and drop id
    arma::rowvec target = dataset.row(target_column_index);
    std::vector<int> to_drop = {target_column_index};
    if (id_column_index >= 0) {
        to_drop.push_back(id_column_index);
    }
    dataset = drop_columns(dataset, to_drop);
```

2. Генерация комбинаций признаков:

```
for (int mask = 1; mask < (1 << num_cols); ++mask) {
    std::vector<int> combination;
    for (int j = 0; j < num_cols; ++j) {
        if (mask & (1 << j)) {
            combination.push_back(j);
        }
    }
}
```

3. Выделение из датасета нужных для тестирования признаков

```
arma::uvec arma_combination(combination.size());
           for (size_t i = 0; i < combination.size(); ++i) {
               arma_combination[i] = combination[i];
           }
           arma::mat sub_dataset =
dataset.rows(arma_combination);
```

4. Оценка RMSE:

```
float score = evaluate_dataset(sub_dataset, target);
```

Используемые библиотеки

- **mlpack** — для машинного обучения.
- **Armadillo** — для работы с матрицами.

Экспериментальная часть

Условия эксперимента

Входной набор данных: файл WineQT.csv с признаками характеристик вина и оценкой качества.

Результаты

По результатам работы алгоритма лучшее сочетание признаков - все признаки. На полном наборе данных результат оценки RMSE получился **0.405982**

Заключение

В ходе выполнения работы была реализована программа для отбора признаков с использованием метода полного перебора при помощи битовой маски. Цель работы достигнута: разработан алгоритм, способный определять оптимальный набор признаков для модели машинного обучения. Результаты эксперимента подтвердили теоретическую сложность алгоритма.

Приложения

Полный исходный код программы можно найти по ссылке https://github.com/ITMO-ML-algorithms-and-data-structures/polygon/pull/579