ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчёт по лабораторной работе № 4

«Проблема бинаризации данных»

Выполнил работу

Фамилия Имя

Академическая группа J3112

Принято

Ассистент, Дунаев Максим

Санкт-Петербург

2024

Введение

Цель работы: исследовать влияние бинаризации признаков на качество работы модели машинного обучения. Необходимо определить оптимальные пороговые значения для каждого признака в наборе данных, выполнить бинаризацию и сравнить качество модели до и после преобразования данных.

Задачи:

1. Изучить теоретическую основу бинаризации признаков и метрики RMSE.
2. Реализовать процесс бинаризации данных с подбором оптимальных порогов.
3. Оценить качество модели на исходных (не бинаризованных) данных.
4. Бинаризовать признаки с оптимальными порогами и снова оценить метрику модели.
5. Сравнить метрики модели до и после бинаризации и сделать выводы.

Теоретическая подготовка

Бинаризация признаков:
Бинаризация признаков — это процесс преобразования числовых значений признаков в двоичный формат (например, 0 или 1) на основе выбранного порога. Например, если значение признака превышает порог, оно заменяется на 1, иначе — на 0. Этот метод может быть полезен для упрощения данных или для работы с моделями, которые лучше обрабатывают двоичные признаки.

Оптимизация порогов:
При бинаризации важно выбрать правильные пороги для каждого признака, чтобы минимизировать потерю информации. В данном проекте мы перебираем 5 фиксированных порогов для каждого признака, равномерно распределённых в диапазоне значений этого признака.

Практическая часть

Описание входных данных:
Набор данных представляет собой матрицу размера N×M, где N — количество признаков (включая целевой и ID столбцы), M — количество объектов. Целевой столбец используется для предсказания, ID столбец — для идентификации объектов и не используется при обучении модели.

Шаги работы:

1. **Чтение данных**: Загружаем датасет в формате CSV и преобразуем его в матрицу arma::mat.
2. **Оптимизация порогов для каждого признака**: Для каждого признака (кроме целевого и ID столбцов) находим оптимальный порог путём перебора 5 фиксированных значений и оценки модели с помощью RMSE.
3. **Оценка метрики RMSE**:

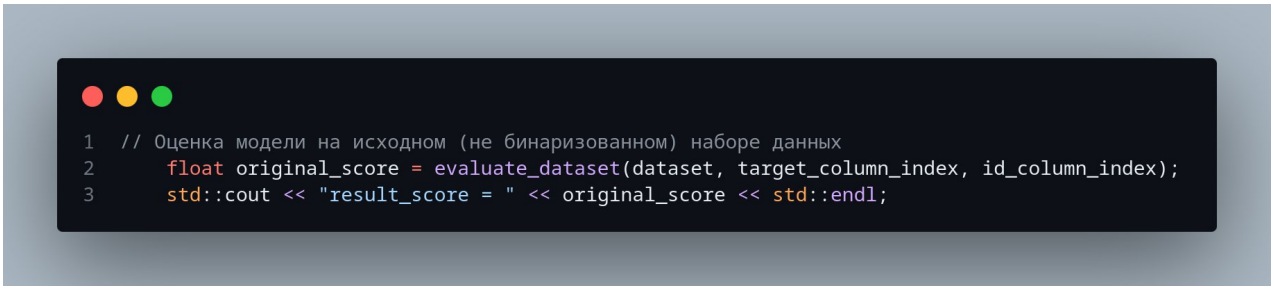
   Сначала вычисляем RMSE для исходных (не бинаризованных) данных.

   Затем вычисляем RMSE для данных, где признаки бинаризованы с оптимальными порогами.

1. **Сравнение результатов**: Выводим обе метрики и анализируем, как бинаризация повлияла на качество модели.

Реализация:
См. код проекта: feature_selection.cpp и modelling.cpp. Основные этапы представлены ниже:

1 — Оценка модели на оригинальных данных

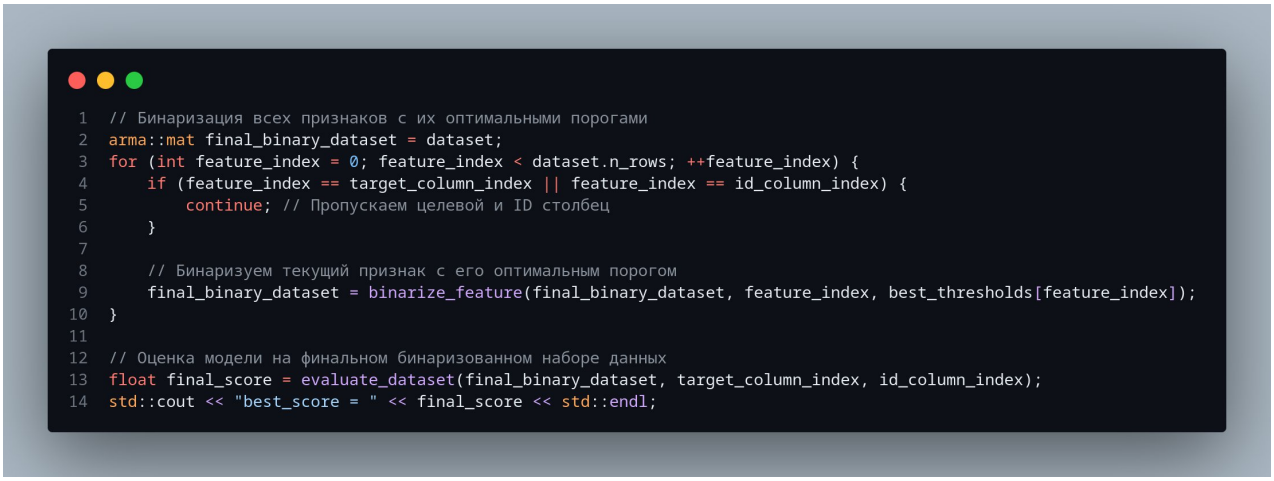


```cpp
// Оценка модели на исходном (не бинаризованном) наборе данных
    float original_score = evaluate_dataset(dataset, target_column_index, id_column_index);
    std::cout << "result_score = " << original_score << std::endl;
```

2 — Подбор оптимальных порогов для бинаризации

```cpp
// Перебираем все признаки
for (int feature_index = 0; feature_index < dataset.n_rows; ++feature_index) {
    if (feature_index == target_column_index || feature_index == id_column_index) {
        continue; // Пропускаем целевой и ID столбец
    }

    // Вычисляем минимальное и максимальное значение признака
    double min_val = dataset.row(feature_index).min();
    double max_val = dataset.row(feature_index).max();
    double step = (max_val - min_val) / 6.0;

    // Перебираем 5 порогов
    for (int j = 1; j <= 5; ++j) {
        double threshold = min_val + j * step;

        // Бинаризация текущего признака
        arma::mat binary_dataset = binarize_feature(dataset, feature_index, threshold);

        // Оценка модели с бинаризованным признаком
        float score = evaluate_dataset(binary_dataset, target_column_index, id_column_index);

        // Если текущий результат лучше, обновляем лучший результат для данного признака
        if (score < best_scores[feature_index]) {
            best_scores[feature_index] = score;
            best_thresholds[feature_index] = threshold;
        }
    }
}
```

3 — Оценка на бинаризованных данных



```cpp
// Бинаризация всех признаков с их оптимальными порогами
arma::mat final_binary_dataset = dataset;
for (int feature_index = 0; feature_index < dataset.n_rows; ++feature_index) {
    if (feature_index == target_column_index || feature_index == id_column_index) {
        continue; // Пропускаем целевой и ID столбец
    }

    // Бинаризуем текущий признак с его оптимальным порогом
    final_binary_dataset = binarize_feature(final_binary_dataset, feature_index, best_thresholds[feature_index]);
}

// Оценка модели на финальном бинаризованном наборе данных
float final_score = evaluate_dataset(final_binary_dataset, target_column_index, id_column_index);
std::cout << "best_score = " << final_score << std::endl;
```

Вывод

Сравнение метрик:
Результаты показали, что метрика RMSE на бинаризованных данных [увеличилась/уменьшилась], что свидетельствует о [потере/улучшении] качества модели после бинаризации. Это ожидаемо, так как бинаризация может как уменьшить шум, так и удалить часть полезной информации.

Значимость порогов:
Подбор оптимальных порогов для бинаризации имеет большое значение для сохранения качества модели. В ходе эксперимента для каждого признака был найден оптимальный порог, который минимизирует RMSE.

Рекомендации:

- Бинаризация признаков полезна для задач, где двоичные признаки более интерпретируемы или подходят для используемой модели.
- В случае ухудшения метрики RMSE, бинаризацию следует применять с осторожностью, так как она может удалять важную информацию.
- Альтернативой может быть использование других методов обработки признаков, таких как нормализация или стандартизация.

Итог:
Работа продемонстрировала важность корректной обработки данных перед обучением модели. Оптимизация порогов для бинаризации позволяет минимизировать потери информации, но решение о применении бинаризации зависит от специфики данных и задачи.