ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчёт по лабораторной работе № 4

«Выбор признаков в задачах машинного обучения»

Выполнил работу

Ежов Дмитрий

Академическая группа №J3113

Принято

Должность, звание Фамилия Имя преподавателя

Санкт-Петербург

2024

**Структура отчёта:**

1. Введение

Цель данной лабораторной работы заключалась в реализации и тестировании алгоритма отбора признаков для задачи машинного обучения. В рамках работы необходимо было проверить различные комбинации признаков и определить минимальное их количество, при котором качество модели ухудшается не сильно

1. Теоретическая подготовка

Теоретической основой метода отбора признаков является информация о признаках, оценка их важности для модели и эффект на метрику качества модели (например, RMSE). Расчет информационного выигрыша применен для оценки значимости признаков. Алгоритм полным перебором (exhaustive selection) позволяет исследовать все возможные комбинации подмножеств признаков.

1. Реализация

Алгоритм реализован на языке C++ с использованием библиотеки mlpack. Основные этапы выполнения работы включали:

Релизация функции расчета информационного выигрыша get\_information\_gain.

Отбор признаков на основе информационного выигрыша с помощью функции filter\_information\_gain.

Реализация функции exhaustive\_selection, выполняющей полный перебор возможных комбинаций признаков.

Чтение датасета из файла, вызов функции feature\_selection для выполнения отбора и обучения модели.

Код программы представлен в приложении А.

1. Экспериментальная часть

Для проверки работоспособности алгоритма использовался датасет из файла WineQT.csv.

Из-за маленького количества входных данных (11 фич) время выполнение даже на медленном алгоритме (2^N) не белее 2ух секунд.

1. Заключение

В результате выполнения лабораторной работы был успешно реализован алгоритм отбора признаков на основе информационного выигрыша. Цель работы достигнута: удалось протестировать алгоритм на датасете и подобрать оптимальный набор признаков с учетом заданных ограничений. В будущем возможно исследование параллелизации данного подхода с целью уменьшения времени выполнения.

1. Приложения

ПРИЛОЖЕНИЕ А

<https://github.com/ITMO-ML-algorithms-and-data-structures/polygon/pull/581>