Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |  |
| --- | --- |
| Институт | математики и компьютерных наук |
| Кафедра | компьютерной безопасности |

ОТЧЕТ

ЛАБОРАТОРНАЯ РАБОТА 6.

ПОСТРОЕНИЕ ПАЙПЛАЙНА ОДНОМЕРНОЙ РЕГРЕССИИ

Выполнил: Окунев Николай Александрович,

студент 2 курса

группы КМБ-с-о-23-1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Проверено с оценкой:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

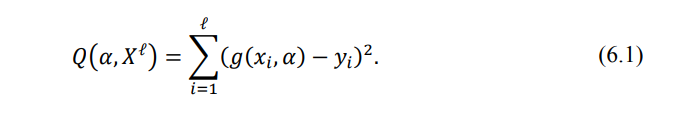
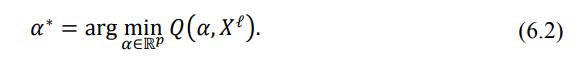
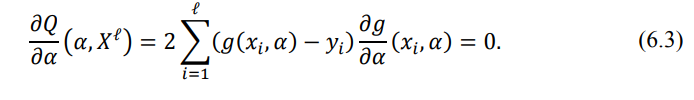
«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Ставрополь, 2025

1. **Цели и задачи**

Цель лабораторной работы: разработка единого пайплайна для решения задачи регрессии. Основные задачи: – реализовать каонвейер для выполнения всех стадий обработки данных при решении задачи одномерной регрессии; – получение теоретических представлений о задаче регресии; – получение навыков использования пайплайна при решении задачи машинного обучения; – получение навыков рефакторинга кода в задачах машинного обучения.

1. **Теоретическое обоснование**

Для решения задачи одномерной регрессии необходимо использовать универсальный пайплайн предварительной обработки данных. К имеющемусяя шаблонному коду необходимо добавить код для обучения модели, интерпретации и визуализации результатов. Линейная регрессия – метод восстановления зависимости между двумя переменными. Пусть задана модель регрессии – параметрическое семейство функций– вектор параметров модели. Определим функционал качества аппроксимации целевой зависимости на выборке Xℓ как сумму квадратов ошибок:  Обучение по методу наименьших квадратов (МНК) состоит в том, чтобы найти вектор параметров Х∗ , при котором достигается минимум среднего квадрата ошибки на заданной обучающей выборке  Стандартный способ решения этой оптимизационной задачи – воспользоваться необходимым условием минимума. Если функция g(x,a) достаточное число раз дифференцируема по X, то в точке минимума выполняется система X уравнений относительно X неизвестных:С использованием библиотек машинного обучения формулы (6.1) – (6.2) можно реализовать автоматически, но следует понимать, что конкретно реализует каждый метод.

1. **Методика и порядок выполнения работы**

Перед выполнением индивидуального задания рекомендуется выполнить все пункты учебной задачи. 3.1 Учебная задача Устовие. Построить пайплайн, реализующий решение задачи линейной одномерной регрессии. Решение. Для решения задачи необходимо написать скрипт на яыке Python (рисунок 6.1)

.

Рисунок 6.1 – Код Python, отражающий общий пайплайн для задачи одномерной регрессии

1. **Контрольные вопросы**

1)Почему при реализации линейной модели регрессии нет необходимости выполнять масштабирование признаков?

Масштабирование признаков в линейной регрессии не является обязательным, но часто желательно.

Без регуляризации: Если в модели нет регуляризации (L1 или L2), то масштабирование признаков не влияет на результат. Коэффициенты модели просто подстроятся под масштаб признаков.

С регуляризацией: Если в модели есть регуляризация, то масштабирование признаков важно. Регуляризация штрафует большие значения коэффициентов. Если признаки имеют разный масштаб, то регуляризация будет сильнее штрафовать коэффициенты для признаков с большим масштабом, что может привести к неоптимальным результатам.

Интерпретация коэффициентов: Масштабирование признаков облегчает интерпретацию коэффициентов модели. Коэффициенты становятся сравнимыми по величине, и можно оценить относительную важность признаков.

2)Почему при реализации модели линейной регрессии в качестве функции потерь используется квадратичное отклонение, а не модуль отклонения?

Квадратичное отклонение (MSE) имеет несколько преимуществ перед модулем отклонения (MAE):

Гладкость: MSE — гладкая функция, что позволяет использовать градиентные методы оптимизации. MAE — негладкая функция, что затрудняет использование градиентных методов.

Чувствительность к выбросам: MSE более чувствительна к выбросам, чем MAE. Это может быть как преимуществом, так и недостатком. Если выбросы являются ошибками в данных, то MSE поможет их выявить и уменьшить их влияние на модель. Если выбросы являются важной частью данных, то MAE может быть более подходящей функцией потерь.

Математические свойства: MSE имеет хорошие математические свойства, что упрощает анализ и разработку алгоритмов.

3)Что именно реализовано в методе fit(X, y) класса LinearRegression?

В методе fit(X, y) класса LinearRegression происходит обучение модели на обучающих данных X и y.

Вычисление коэффициентов: Метод вычисляет коэффициенты линейной регрессии, минимизируя функцию потерь (обычно MSE). Это делается с помощью аналитического решения (метод наименьших квадратов) или с помощью градиентных методов оптимизации.

Сохранение коэффициентов: Вычисленные коэффициенты сохраняются в атрибутах модели (например, coef\_ и intercept\_).

4)Что такое p-значение? Как p-значение используется при оптимизации моделей регрессии?

P-значение (p-value) — это вероятность получить наблюдаемое значение статистики (или еще более экстремальное) при условии, что нулевая гипотеза верна.

Нулевая гипотеза: Гипотеза о том, что нет связи между признаком и целевой переменной.

Интерпретация p-значения:

Маленькое p-значение (обычно меньше 0.05) говорит о том, что нулевая гипотеза, скорее всего, неверна, и есть связь между признаком и целевой переменной.

Большое p-значение говорит о том, что нет достаточных оснований для отклонения нулевой гипотезы, и связь между признаком и целевой переменной может быть случайной.

Использование p-значения при оптимизации моделей регрессии: P-значение используется для отбора признаков. Признаки с большим p-значением (то есть признаки, которые не имеют статистически значимой связи с целевой переменной) могут быть удалены из модели, чтобы улучшить ее обобщающую способность.

5)Поясните назначение метода predict класса LinearRegression.

Метод predict класса LinearRegression используется для предсказания значений целевой переменной на основе новых данных.

Вход: На вход метод принимает матрицу признаков X.

Вычисление предсказаний: Метод умножает матрицу признаков X на коэффициенты модели и добавляет смещение (intercept), чтобы получить предсказания для целевой переменной.

Выход: Метод возвращает вектор предсказанных значений.

6)Поясните назначение метода plot и scatter класса pyplot.

Методы plot и scatter класса pyplot используются для визуализации данных.

plot: Используется для построения графиков функций или зависимостей между двумя переменными. Обычно используется для отрисовки линий.

scatter: Используется для построения диаграмм рассеяния (scatter plots). Каждая точка на диаграмме представляет собой объект из набора данных. Диаграммы рассеяния используются для визуализации связи между двумя переменными.

7)По какой подвыборке необходимо оценивать точность модели машинного обучения: тестовой или тренировочной?

Точность модели необходимо оценивать на тестовой выборке.

Оценка на тренировочной выборке: Оценка на тренировочной выборке показывает, насколько хорошо модель запомнила тренировочные данные. Она не показывает, насколько хорошо модель будет работать на новых данных.

Оценка на тестовой выборке: Оценка на тестовой выборке показывает, насколько хорошо модель обобщает на новые данные. Это более надежная оценка качества модели.

**Вывод:** в ходе лабораторной работы мы разработали единый пайплайн для решения задачи регрессии.