Практическая работа №3

Решение задачи линейной регрессии

ОГЛАВЛЕНИЕ

ЦЕЛЬ:

1. Изучение основных принципов линейной регрессии: Понять, как строится модель линейной регрессии, и какие методы используются для определения коэффициентов модели.

2. Практическое применение аналитического и численного подходов: Закрепить знания о методе наименьших квадратов (аналитическое решение) и численных методах, таких как градиентный спуск, для решения задач линейной регрессии.

3. Навыки программирования на Python для машинного обучения: Развить навыки написания кода для реализации численных методов решения задач регрессии с использованием библиотек, таких как NumPy и Matplotlib.

4. Сравнение различных моделей регрессии: Научиться различать методы линейной регрессии, такие как стандартная регрессия, Lasso и Ridge, и понять их преимущества и ограничения.

5. Оценка качества моделей с использованием метрик: Ознакомиться с основными метриками качества регрессионных моделей (MSE, MAE, R^2) и научиться интерпретировать их результаты.

6. Применение знаний на практике: Применить теоретические знания на практике, решая задачи с реальными данными и анализируя полученные результаты.

ЗАДАНИЕ

Порядок выполнения работы:

1. Решить задачу регрессии на примере следующих данных:

```
x = [6.1101, 5.5277, 8.5186, 7.0032, 5.8598, 8.3829, 7.4764, 8.5781, 6.4862,
5.0546, 5.7107, 14.164, 5.734, 8.4084, 5.6407, 5.3794,6.3654, 5.1301, 6.4296,
7.0708, 6.1891, 20.27, 5.4901, 6.3261, 5.5649, 18.945, 12.828, 10.957, 13.176,
22.203, 5.2524, 6.5894,9.2482, 5.8918, 8.2111, 7.9334, 8.0959, 5.6063, 12.836,
6.3534, 5.4069, 6.8825, 11.708, 5.7737, 7.8247, 7.0931, 5.0702, 5.8014,11.7,
5.5416, 7.5402, 5.3077, 7.4239, 7.6031, 6.3328, 6.3589, 6.2742, 5.6397, 9.3102,
9.4536, 8.8254, 5.1793, 21.279, 14.908, 18.959,7.2182, 8.2951, 10.236, 5.4994,
20.341, 10.136, 7.3345, 6.0062, 7.2259, 5.0269, 6.5479, 7.5386, 5.0365, 10.274,
5.1077, 5.7292, 5.1884,6.3557, 9.7687, 6.5159, 8.5172, 9.1802, 6.002, 5.5204,
5.0594, 5.7077, 7.6366, 5.8707, 5.3054, 8.2934, 13.394, 5.4369]
y = [17.592, 9.1302, 13.662, 11.854, 6.8233, 11.886, 4.3483, 12, 6.5987,
3.8166, 3.2522, 15.505, 3.1551, 7.2258, 0.71618, 3.5129, 5.3048, 0.56077,3.6518,
5.3893, 3.1386, 21.767, 4.263, 5.1875, 3.0825, 22.638, 13.501, 7.0467, 14.692,
24.147, -1.22, 5.9966, 12.134, 1.8495, 6.5426, 4.5623,4.1164, 3.3928, 10.117,
5.4974, 0.55657, 3.9115, 5.3854, 2.4406, 6.7318, 1.0463, 5.1337, 1.844, 8.0043,
1.0179, 6.7504, 1.8396, 4.2885, 4.9981,1.4233, -1.4211, 2.4756, 4.6042, 3.9624,
5.4141, 5.1694, -0.74279, 17.929, 12.054, 17.054, 4.8852, 5.7442, 7.7754, 1.0173,
20.992, 6.6799, 4.0259,1.2784, 3.3411, -2.6807, 0.29678, 3.8845, 5.7014, 6.7526,
2.0576, 0.47953, 0.20421, 0.67861, 7.5435, 5.3436, 4.2415, 6.7981, 0.92695,
0.152, 2.8214,1.8451, 4.2959, 7.2029, 1.9869, 0.14454, 9.0551, 0.61705]
```

2. Написать функцию, которая реализует численное решение задачи регрессии.
3. Построить график построенной модели.
4. Сравнить результат численного решения с аналитическим.
5. Решить задачу регрессии для структурированных данных. Повторов в группе по выбору данных быть не должно. Данные можно выбрать в следующих источниках: https://www.kaggle.com/, sklearn.datasets,
6. Сравнить работу линейной регрессии с Lasso-регрессией и с Ridge-регрессией.
7. Вывести метрики оценки качества модели для задачи регрессии.

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ:

Линейная регрессия используется для моделирования связи между зависимой переменной и одной или несколькими независимыми переменными. Она основывается на предположении, что эта связь может быть описана прямой линией.

Основное уравнение линейной регрессии:

$y = b_0 + b_1 * x$

где y — зависимая переменная, x — независимая переменная, b_0 — свободный член, b_1 — коэффициент наклона.

Решение задачи линейной регрессии может быть аналитическим (метод наименьших квадратов) или численным (градиентный спуск).

АНАЛИТИЧЕСКОЕ РЕШЕНИЕ ЗАДАЧИ РЕГРЕССИИ

Аналитическое решение — это нахождение точного решения задачи с помощью математических формул. В случае линейной регрессии аналитическое решение основано на методе наименьших квадратов (МНК), который минимизирует сумму квадратов отклонений предсказанных значений от фактических.

Для линейной регрессии уравнение аналитического решения выглядит так:

$$\beta = (X^T \times X)^{-1} \times X^T \times y$$

где:

- β — вектор коэффициентов модели, который мы хотим найти.
- X — матрица признаков.
- X^T — транспонированная матрица признаков.
- $(X^TX)^{-1}$ — обратная матрица произведения X^T и X.
- y — вектор наблюдаемых значений.

Преимущества аналитического решения:

- точность: дает точное решение без приближений.
- скорость: при небольших размерах матрицы вычисляется очень быстро.

Ограничения аналитического решения:

- Требует вычисления обратной матрицы, что возможно не всегда (например, если матрица вырожденная или сильно коррелированы признаки).
- Неподходит для больших и разреженных данных, так как требует больших вычислительных ресурсов.

ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ РЕГРЕССИИ

Численное решение — это приближенный метод нахождения коэффициентов модели с помощью итеративных процедур. Один из самых популярных численных методов — градиентный спуск.

Градиентный спуск минимизирует функцию потерь (например, среднеквадратичную ошибку) путем итеративного обновления коэффициентов модели. Алгоритм работает следующим образом:

1. Инициализируются начальные значения коэффициентов (обычно случайные).
2. Рассчитывается градиент функции потерь относительно коэффициентов.
3. Обновляются коэффициенты, двигаясь в направлении, противоположном градиенту, на определенный шаг (скорость обучения).
4. Процесс повторяется до тех пор, пока функция потерь не станет минимальной или изменения коэффициентов не станут достаточно малыми.

Формула обновления коэффициентов:

$$\beta_{new} = \beta_{old} - \alpha \times \nabla L(\beta)$$

где:

- α — шаг обучения (learning rate).
- ∇L(β) — градиент функции потерь.

Преимущества численного решения:

- Гибкость: подходит для сложных моделей, где аналитическое решение невозможно.
- Применимость к большим данным: может обрабатывать большие и разреженные матрицы.
- Возможность регулирования: можно использовать различные модификации градиентного спуска (например, стохастический, мини-батч), чтобы улучшить скорость и сходимость.

Ограничения численного решения:

- Сходимость: возможна застревание в локальных минимумах, особенно для сложных функций потерь.

- Зависимость от начальных условий и параметров: неправильный выбор шага обучения может привести к плохой сходимости.

МЕТРИКИ ОЦЕНКИ КАЧЕСТВА МОДЕЛИ

Метрики, используемые для оценки качества модели:

1. MSE (среднеквадратичная ошибка): показывает среднее значение квадратов ошибок предсказаний.
2. MAE (средняя абсолютная ошибка): среднее значение абсолютных ошибок предсказаний.
3. R^2 (коэффициент детерминации): показывает, какая доля дисперсии зависимой переменной объясняется моделью.

MSE (Среднеквадратичная ошибка)

Среднеквадратичная ошибка (Mean Squared Error, MSE) измеряет среднее значение квадратов ошибок предсказаний модели. Она вычисляется как среднее арифметическое квадратов разниц между предсказанными и фактическими значениями.

Формула для MSE:

$$MSE = \frac{1}{n} \times \sum (y_i - \hat{y}_i)^2$$

где:

- n — количество наблюдений,
- y_i — фактическое значение,
- $\hat{y}_i$ — предсказанное значение модели.

MSE используется для оценки степени разброса предсказанных значений от реальных. Чем меньше значение MSE, тем точнее модель.

Пример

Допустим, у нас есть следующие данные:

- Фактические значения: [3, -0.5, 2, 7]
- Предсказанные значения: [2.5, 0.0, 2, 8]

Расчет MSE:

- Ошибки: [0.5, -0.5, 0, -1]
- Квадраты ошибок: [0.25, 0.25, 0, 1]

MSE = (0.25 + 0.25 + 0 + 1) / 4 = 0.375

MAE (Средняя абсолютная ошибка)

Средняя абсолютная ошибка (Mean Absolute Error, MAE) измеряет среднее значение абсолютных ошибок предсказаний модели. Она показывает среднюю величину ошибок, не учитывая их направление (положительное или отрицательное).

Формула для MAE:

$$MAE = \frac{1}{n} \times \sum |y_i - \hat{y}_i|$$

где:

- n — количество наблюдений,
- y_i — фактическое значение,
- $\hat{y}_i$ — предсказанное значение модели.

Пример:

Используем те же данные:

- Фактические значения: [3, -0.5, 2, 7]
- Предсказанные значения: [2.5, 0.0, 2, 8]

Расчет MAE:

- Ошибки: [0.5, 0.5, 0, 1]

MAE = (0.5 + 0.5 + 0 + 1) / 4 = 0.5

MAE оценивает среднее отклонение предсказанных значений от фактических и легче интерпретируется, так как выражается в тех же единицах, что и сами данные.

R^2 (Коэффициент детерминации)

Коэффициент детерминации (R^2) показывает, какая доля дисперсии зависимой переменной объясняется моделью. Значение R^2 варьируется от 0 до 1, где 1 означает, что модель идеально объясняет все вариации в данных, а 0 — модель не объясняет их вообще.

Формула для R^2:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

где:

- y_i — фактическое значение,
- $\hat{y}_i$ — предсказанное значение модели,
- $\bar{y}$ — среднее значение фактических значений.

Пример:

Опять используем те же данные:

- Фактические значения: [3, -0.5, 2, 7]
- Предсказанные значения: [2.5, 0.0, 2, 8]
- Среднее фактических значений: (3 + (-0.5) + 2 + 7) / 4 = 2.875

Расчет R^2:

RSS (сумма квадратов ошибок): $(0.5)^2 + (-0.5)^2 + (0)^2 + (-1)^2 = 1.5$

TSS (общая сумма квадратов): $(3 - 2.875)^2 + (-0.5 - 2.875)^2 + (2 - 2.875)^2 + (7 - 2.875)^2 = 23.1875$

$R^2 = 1 - (RSS / TSS) = 1 - (1.5 / 23.1875) \approx 0.935$

R^2 часто используется для оценки общего качества модели: чем ближе значение R^2 к 1, тем лучше модель описывает данные.

РЕГУЛЯРИЗАЦИЯ: LASSO И RIDGE РЕГРЕССИИ

Регуляризация используется для борьбы с переобучением. Lasso и Ridge — методы, которые добавляют штраф к модели, тем самым уменьшая её сложность:

1. Lasso-регрессия (L1-регуляризация): добавляет штраф за сумму модулей коэффициентов. Это приводит к занулению некоторых коэффициентов, что эффективно отбирает признаки.

2. Ridge-регрессия (L2-регуляризация): добавляет штраф за сумму квадратов коэффициентов, что сглаживает модель, не зануляя коэффициенты.

Регуляризация используется для борьбы с переобучением моделей, добавляя штраф к функции потерь за сложность модели. Регуляризация помогает уменьшить величину коэффициентов модели, что приводит к более устойчивым и обобщаемым результатам. Наиболее популярные методы регуляризации в линейной регрессии — это Lasso и Ridge регрессии.

Lasso-регрессия (L1-регуляризация)

Lasso (Least Absolute Shrinkage and Selection Operator) добавляет штраф за сумму модулей коэффициентов модели. Этот штраф заставляет некоторые коэффициенты стать равными нулю, что эффективно выполняет отбор признаков.

Формула функции потерь с Lasso-регуляризацией выглядит так:

$$L = RSS + \lambda \sum |\beta_i|$$

где:

- L — функция потерь модели,
- RSS — сумма квадратов остатков (ошибок предсказания),
- λ — параметр регуляризации, определяющий вес штрафа,
- β_i — коэффициенты модели.

Основное преимущество Lasso-регрессии — возможность автоматического отбора признаков, что упрощает модель и делает её интерпретируемой.

Пример

Допустим, у нас есть данные с несколькими предикторами, некоторые из которых мало влияют на результат.

Фактические данные (X1, X2, X3) и целевая переменная (y):

X1 = [1, 2, 3, 4]

X2 = [0, 0, 0, 0]

X3 = [4, 3, 2, 1]

y = [3, 2.5, 2, 1.5]

При использовании Lasso-регрессии:

Функция потерь учитывает штраф за ненулевые коэффициенты, и коэффициент для X2 может стать нулевым, так как X2 не влияет на результат.

Итоговая модель может выглядеть как: y = 0.5 * X1 + 0.5 * X3, что значительно упрощает интерпретацию.

Ridge-регрессия (L2-регуляризация)

Ridge регрессия добавляет штраф за сумму квадратов коэффициентов модели. Этот штраф предотвращает слишком большие значения коэффициентов, делая модель более устойчивой и сглаженной, особенно в условиях мультиколлинеарности признаков.

Формула функции потерь с Ridge-регуляризацией выглядит так:

$$L = RSS + \lambda \sum \beta_i^2$$

где:

- L — функция потерь модели,
- RSS — сумма квадратов остатков (ошибок предсказания),
- λ — параметр регуляризации, определяющий вес штрафа,
- β_i — коэффициенты модели.

Пример

Используем те же данные:

Фактические данные (X1, X2, X3) и целевая переменная (y):

X1 = [1, 2, 3, 4]

X2 = [4, 3, 2, 1]

X3 = [10, 10, 10, 10]

y = [5, 5, 5, 5]

При использовании Ridge-регрессии:

Функция потерь добавляет штраф за большие коэффициенты. Это сглаживает коэффициенты и уменьшает их значения, особенно для тех переменных, которые сильно коррелированы (например, X1 и X2).

Модель все равно учитывает все признаки, но их вклад будет более сбалансированным, избегая доминирования одного признака.