Урок 13. Регрессия (часть 2)

Хакимов Р.И. + ChatGPT

Модель нелинейной регрессии

Модель нелинейной регрессии может быть представлена в виде:

$$Y = f(X_1, X_2, \ldots, X_p; \beta) + \epsilon$$

где:
- Y — зависимая переменная.
- $X_1, X_2, \ldots, X_p$ — независимые переменные.
- $f(\cdot)$ — нелинейная функция, определяющая зависимость между переменными.
- β — вектор параметров модели.
- ϵ — случайная ошибка.

Примеры нелинейных функций

1. **Экспоненциальная функция:**

$$Y = \beta_0 e^{\beta_1 X}$$

2. **Логарифмическая функция:**

$$Y = \beta_0 + \beta_1 \log(X)$$

3. **Полиномиальная функция (высшего порядка):**

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p$$

Примеры нелинейных функций

4. Гиперболическая функция:

$$Y = \frac{\beta_0}{\beta_1 + X}$$

5. Сигмойдальная функция (логистическая регрессия):

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Модель нелинейной регрессии

Преимущества:

Гибкость: Нелинейные модели могут лучше описывать сложные зависимости между переменными.

Точность: Может предоставить более точные предсказания, если связь действительно нелинейная.

Недостатки:

Сложность: Нелинейные модели могут быть сложными для интерпретации и требуют больше вычислительных ресурсов.

Перегрузка: Модели могут переобучаться на данных, если слишком сложные функции используются.

Итеративные методы: Оценка параметров требует итеративных методов, которые могут быть чувствительны к начальным условиям и локальным минимумам.

Методы выбора признаков на основе корреляции

Когда мы работаем с множеством признаков, некоторые из них могут быть слабо коррелированы с целевой переменной или сильно коррелированы друг с другом. Это может негативно сказаться на качестве модели, так как избыточная информация (высокая корреляция между признаками) может привести к многоколлинеарности. Чтобы избежать этого, часто применяют методы выбора признаков на основе корреляции.

Методы выбора признаков на основе корреляции

1. Корреляция признаков с целевой переменной: Признаки, которые слабо коррелированы с целевой переменной, могут быть исключены из модели, так как они не предоставляют полезной информации для предсказания целевой переменной. С другой стороны, признаки с высокой корреляцией (в положительном или отрицательном направлении) более важны и предпочтительны.

2. Корреляция между признаками: Признаки, которые сильно коррелированы между собой, могут дублировать информацию. В таких случаях можно оставить только один из сильно коррелированных признаков, чтобы снизить избыточность в данных и избежать многоколлинеарности.

Многоколлинеарность

Многоколлинеарность (или мультиколлинеарность) в математической статистике и эконометрике — это ситуация, когда в множественной линейной регрессии одна из независимых переменных может быть представлена как линейная комбинация других независимых переменных с высокой точностью. Это приводит к тому, что переменные оказываются сильно взаимосвязанными, что может создавать проблемы при оценке коэффициентов регрессии.

Многоколлинеарность

Основные проблемы, связанные с многоколлинеарностью:

1. Неустойчивость оценок: коэффициенты регрессии могут быть сильно чувствительны к малейшим изменениям в данных, что делает их оценки менее надежными.

2. Проблемы с интерпретацией: если независимые переменные сильно коррелируют, то сложно понять, какой именно переменной приписывать влияние на зависимую переменную.

3. Высокие стандартные ошибки: это снижает статистическую значимость коэффициентов, что может приводить к неверным выводам о значимости переменных.

Многоколлинеарность

Для диагностики многоколлинеарности часто используют:
- **Коэффициент детерминации** R^2 между переменными.
- **Фактор инфляции дисперсии (VIF)**: если *VIF* для переменной превышает 10, это может указывать на многоколлинеарность.

Методы борьбы с многоколлинеарностью:
1. Исключение одной из взаимосвязанных переменных.
2. Применение методов регуляризации, таких как ридж-регрессия (Ridge regression), которые штрафуют большие коэффициенты и снижают влияние многоколлинеарности.

Матрица корреляции

Матрица корреляции — это квадратная таблица, которая показывает коэффициенты корреляции между всеми парами признаков в наборе данных. Каждый элемент матрицы представляет собой коэффициент корреляции Пирсона для двух признаков. Значения коэффициента корреляции находятся в диапазоне от -1 до 1:

- Значение близкое к 1 означает сильную положительную корреляцию (признаки изменяются в одном направлении).
- Значение близкое к -1 означает сильную отрицательную корреляцию (признаки изменяются в противоположных направлениях).
- Значение близкое к 0 означает отсутствие корреляции.

Матрица корреляции

Матрица корреляции помогает:
- Определить взаимосвязи между признаками.
- Найти признаки с сильной взаимной корреляцией, чтобы исключить лишние признаки.
- Выявить важные признаки для целевой переменной.

Пример матрицы корреляции для набора признаков может выглядеть так:

	Признак 1	Признак 2	Признак 3	Целевая переменная
Признак 1	1.0	0.8	-0.5	0.7
Признак 2	0.8	1.0	-0.4	0.6
Признак 3	-0.5	-0.4	1.0	-0.3
Целевая переменная	0.7	0.6	-0.3	1.0

Здесь можно видеть, что Признак 1 и Признак 2 сильно коррелируют между собой (0.8), что может указывать на необходимость исключения одного из них из модели.

Матрица корреляции

Основные шаги выбора признаков на основе корреляции:

1. Построение матрицы корреляции.
2. Оценка корреляции каждого признака с целевой переменной.
3. Исключение признаков с низкой корреляцией с целевой переменной.
4. Исключение признаков с высокой взаимной корреляцией.

Эти шаги помогают упростить модель, сократить количество признаков и улучшить интерпретируемость результатов.