

ПОВЫШЕНИЕ ТОЧНОСТИ ЭКСТРАПОЛЯЦИИ РЕГРЕССИИ ДЛЯ ДАННЫХ, СОДЕРЖАЩИХ КОМПЛЕКСНЫЕ ЧИСЛА, ПУТЕМ УЧЕТА ПОГРЕШНОСТИ ИЗМЕРЕНИЯ ПЕРЕМЕННЫХ И ИСПОЛЬЗОВАНИЯ СТАНДАРТНЫХ РЕГРЕССОРОВ SCIKIT-LEARN

Постановка задачи

Имеется набор данных `dataset_0`, который содержит количественные предикторы `A_1`, `A_2`, `A_3`, `A_4`, `A_5` и количественную целевую переменную `b`. Набор данных `dataset_0` представляет собой точную зависимость целевой переменной `b` от предикторов.

Даны два набора данных `dataset_1` и `dataset_2`, каждый из которых содержит количественные предикторы `A_h_1`, `A_h_2`, `A_h_3`, `A_h_4`, `A_h_5` и количественную целевую переменную `b_d`. Наборы данных `dataset_1` и `dataset_2` моделируют результаты измерений `b` от предикторов и сгенерированы по данным `dataset_0` путем случайных отклонений предикторов в пределах 1% и целевой переменной `b` в пределах 5% от соответствующих точных значений из `dataset_0`.

Необходимо по данным `dataset_1` и `dataset_2` спрогнозировать значение целевой переменной `b_max` при значениях предикторов ($1800 + j1000$, $1500 + j1000$, $1800 + j1400$, $2000 + j1700$, $1500 + j1200$). Датасет `dataset_0` использовать только для валидации обученных моделей.

РЕШЕНИЕ ЗАДАЧИ

1. Характеристика датасетов

Значения предикторов в каждом наборе данных являются положительными комплексными числами, максимальное значение которых не превышает $1230 + j513$. Значения целевой переменной в каждом наборе данных также являются положительными комплексными числами, максимальное значение которых не превышает $83 + j109$.

Для `dataset_0` выявлена функциональная линейная зависимость целевой переменной от всех предикторов. При этом имеется практически абсолютная мультиколлинеарность всех предикторов.

Для dataset_1 и dataset_2 коэффициенты корреляции для всех предикторов между собой практически равны 1, а для целевой переменной с предикторами практически равны 0.99. Таким образом, для dataset_1 и dataset_2 также имеются практически функциональная линейная зависимость целевой переменной от предикторов и практически абсолютная мультиколлинеарность всех предикторов.

2. Обучение различных моделей

Данная задача относится к классу задач экстраполяции регрессии.

Обучение моделей выполнено на языке Python 3.

Обучение проводилось как при разбиении на обучающую и тестовую выборки (70/30), так и на полных наборах данных.

Модели строились на всех предикторах, которые масштабировались с помощью MaxAbsScaler.

Для обучения моделей применялись алгоритмы:

1) LinearRegression, Ridge, Lasso, ElasticNet, TheilSenRegressor из библиотеки scikit-learn с поиском оптимальных значений их гиперпараметров по сетке с кросс-валидацией.

2) регуляризация Тихонова с поиском коэффициента регуляризации α методом обобщенной невязки [1].

Алгоритмы scikit-learn не поддерживают работу с комплексными числами, что не позволяет построить непосредственные модели регрессии для данных, содержащих комплексные числа, с помощью стандартных регрессоров scikit-learn. Для устранения этого препятствия автором разработан метод преобразования исходного датасета размером $M \times N$ с комплексными числами в датасет размером $2M \times (2N - 1)$ с вещественными числами, что позволило использовать для построения достоверных моделей регрессии комплексных чисел стандартные регрессоры scikit-learn (подробнее смотри репозиторий автора https://github.com/GorshkovAndrey/Regression_of_complex_numbers_using_sklearn).

Отличие алгоритма регуляризации Тихонова от модели Ridge из scikit-learn состоит в том, что коэффициент регуляризации α определяется по условию оптимизации нестандартной метрики – обобщенной невязки, учитывающей информацию о погрешности измерения предикторов и целевой переменной.

Решение задачи регрессии – системы линейных уравнений $Az = b$ для модели на основе алгоритма регуляризации Тихонова имеет следующий вид

$$z = (A^T A + \alpha E)^{-1} A^T b \quad (1)$$

Уравнение обобщенной невязки для определения коэффициента регуляризации α имеет следующий вид

$$\|Az(\alpha) - b\|^2 - (h\|z(\alpha)\| + d)^2 = 0 \quad (2)$$

$$h = \delta_A \|A\| \quad (3)$$

$$d = \delta_b \|b\| \quad (4)$$

где $\| \cdot \|$ – обозначение нормы матрицы или вектора;

δ_A – относительная погрешность измерения элементов матрицы A ;

δ_b – относительная погрешность измерения элементов вектора b .

По условию данной задачи $\delta_A = 0,01$, $\delta_b = 0,05$.

Метрики, результаты прогноза целевой переменной b_max и их осредненные значения Mean для различных моделей регрессоров из библиотеки scikit-learn приведены в таблицах 1 – 7. Там же приведены результаты прогноза Regularized для модели на основе алгоритма регуляризации Тихонова и **точный прогноз** Accurate.

Таблица 1 – Метрики и результаты прогноза целевой переменной для dataset_1 (random_state = 1902 для train_test_split)

| | LinRegr | Ridge | Lasso | ElastNet | TheilSen | Mean | Regularized | Accurate |
|---------------|--------------------|--------------------|-------------------|-------------------|--------------------|------------------|--------------------|----------------|
| R2 train | 0.988 | 0.988 | 0.987 | 0.987 | 0.987 | 0.988 | 0.931 | |
| R2 test | 0.981 | 0.982 | 0.985 | 0.984 | 0.981 | 0.983 | 0.928 | |
| b_max_complex | (305.464+440.399j) | (272.488+411.849j) | (192.317+320.62j) | (222.348+343.56j) | (271.979+427.573j) | (252.919+388.8j) | (159.659+324.404j) | (147.1+303.7j) |
| b_max_modul | 535.966 | 493.831 | 373.876 | 409.234 | 506.746 | 463.825 | 361.565 | 337.449 |

Таблица 2 – Метрики и результаты прогноза целевой переменной для dataset_1 (random_state = 1903 для train_test_split)

| | LinRegr | Ridge | Lasso | ElastNet | TheilSen | Mean | Regularized | Accurate |
|---------------|--------------------|-------------------|--------------------|------------------|-------------------|--------------------|--------------------|----------------|
| R2 train | 0.986 | 0.986 | 0.986 | 0.986 | 0.986 | 0.986 | 0.932 | |
| R2 test | 0.985 | 0.984 | 0.985 | 0.985 | 0.985 | 0.985 | 0.929 | |
| b_max_complex | (208.258+343.867j) | (166.59+329.601j) | (220.384+327.825j) | (217.115+327.2j) | (194.87+341.702j) | (201.443+334.039j) | (160.012+325.055j) | (147.1+303.7j) |
| b_max_modul | 402.015 | 369.309 | 395.017 | 392.682 | 393.363 | 390.079 | 362.305 | 337.449 |

Таблица 3 – Метрики и результаты прогноза целевой переменной для dataset_1 (обучение на полном наборе данных)

| | LinRegr | Ridge | Lasso | ElastNet | TheilSen | Mean | Regularized | Accurate |
|---------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|----------------|
| R2 | 0.987 | 0.986 | 0.986 | 0.986 | 0.986 | 0.986 | 0.933 | |
| b_max_complex | (210.911+350.847j) | (173.188+330.798j) | (185.271+314.552j) | (192.028+321.218j) | (201.778+364.422j) | (192.635+336.367j) | (160.015+325.052j) | (147.1+303.7j) |
| b_max_modul | 409.362 | 373.392 | 365.059 | 374.24 | 416.555 | 387.622 | 362.303 | 337.449 |

Таблица 4 – Метрики и результаты прогноза целевой переменной для dataset_2 (random_state = 1994 для train_test_split)

| | LinRegr | Ridge | Lasso | ElastNet | TheilSen | Mean | Regularized | Accurate |
|---------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|----------------|
| R2 train | 0.99 | 0.99 | 0.989 | 0.989 | 0.989 | 0.989 | 0.936 | |
| R2 test | 0.977 | 0.978 | 0.98 | 0.98 | 0.977 | 0.979 | 0.907 | |
| b_max_complex | (288.46+435.473j) | (272.626+420.048j) | (231.557+337.462j) | (231.365+337.574j) | (271.936+439.099j) | (259.189+393.931j) | (159.907+324.525j) | (147.1+303.7j) |
| b_max_modul | 522.347 | 500.765 | 409.267 | 409.25 | 516.485 | 471.551 | 361.783 | 337.449 |

Таблица 5 – Метрики и результаты прогноза целевой переменной для dataset_2 (random_state = 1961 для train_test_split)

| | LinRegr | Ridge | Lasso | ElastNet | TheilSen | Mean | Regularized | Accurate |
|---------------|-------------------|-------------------|--------------------|--------------------|------------------|--------------------|--------------------|----------------|
| R2 train | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.93 | |
| R2 test | 0.985 | 0.986 | 0.987 | 0.986 | 0.983 | 0.986 | 0.939 | |
| b_max_complex | (102.463+215.89j) | (152.992+296.72j) | (155.754+281.007j) | (158.619+292.201j) | (91.844+216.47j) | (132.334+260.458j) | (160.358+325.441j) | (147.1+303.7j) |
| b_max_modul | 238.971 | 333.84 | 321.285 | 332.478 | 235.148 | 292.148 | 362.804 | 337.449 |

Таблица 6 – Метрики и результаты прогноза целевой переменной для dataset_2 (обучение на полном наборе данных)

| | LinRegr | Ridge | Lasso | ElastNet | TheilSen | Mean | Regularized | Accurate |
|---------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|----------------|
| R2 | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.987 | 0.933 | |
| b_max_complex | (189.993+321.392j) | (170.353+329.699j) | (188.332+299.377j) | (207.623+310.257j) | (175.489+323.464j) | (186.358+316.838j) | (160.354+325.421j) | (147.1+303.7j) |
| b_max_modul | 373.35 | 371.109 | 353.688 | 373.319 | 368.002 | 367.581 | 362.784 | 337.449 |

Таблица 7 – Метрики и результаты прогноза целевой переменной для dataset_0 (обучение на полном наборе данных)

| | LinRegr | Ridge | TheilSen | Mean | Regularized | Accurate |
|---------------|----------------|----------------|------------------|----------------|------------------|----------------|
| R2 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | |
| b_max_complex | (147.1+303.7j) | (147.1+303.7j) | (147.1+303.701j) | (147.1+303.7j) | (147.1+303.701j) | (147.1+303.7j) |
| b_max_modul | 337.449 | 337.449 | 337.45 | 337.449 | 337.45 | 337.449 |

Примечание. dataset_0 сгенерирован при коэффициентах $z = (0.04 + j0.02, 0.035 + j0.015, 0.03 + j0.02, 0.02 + j0.012, 0.016 + j0.01)$

3. Анализ прогнозов различных моделей

Для dataset_1 диапазон модуля осредненного прогноза Mean различных моделей регрессоров scikit-learn составил от 387,6 до 463,8. Диапазон прогноза регуляризованного решения, учитывающего информацию о погрешности измерения предикторов и целевой переменной, составил от 361,6 до 362,3.

Для dataset_2 диапазон модуля осредненного прогноза Mean различных моделей регрессоров из библиотеки scikit-learn составил от 292,1 до 471,6. Диапазон прогноза регуляризованного решения, учитывающего информацию о погрешности измерения предикторов и целевой переменной, составил от 361,8 до 362,8.

Для dataset_0 значение модуля прогноза составляет 337,5. Это же значение получается при вычислении **точного прогноза** по выражению $b_{\max} = A_{\max} \cdot z$, где $z = (0.04 + j0.02, 0.035 + j0.015, 0.03 + j0.02, 0.02 + j0.012, 0.016 + j0.01)$ – коэффициенты регрессии для генерации dataset_0. Значение модуля прогноза регуляризованного решения с учетом того, что погрешности измерения переменных для dataset_0 равны нулю, также составило 337,5.

Таким образом, несмотря на отличные метрики всех моделей регрессоров из библиотеки `scikit-learn`, результаты их прогнозов не только имеют неприемлемую точность, но и являются неустойчивыми – небольшие погрешности измерений данных приводят к недопустимым погрешностям прогноза целевой переменной (в терминологии ML это означает, что модели переобучаются). При этом модель регуляризации Тихонова, учитывающая информацию о погрешности измерения предикторов и целевой переменной, выдает устойчивый прогноз с приемлемой точностью. Так в данной задаче разброс (variance) осредненного прогноза различных моделей из библиотеки `scikit-learn` составил $\pm 23\%$ при смещении (bias) 13% , тогда как разброс (variance) прогноза модели регуляризации Тихонова составил $\pm 0,2\%$ при смещении (bias) 7% .

4. Выводы

1. Алгоритмы `scikit-learn` не поддерживают работу с комплексными числами, что не позволяет построить непосредственные модели регрессии для данных, содержащих комплексные числа, с помощью стандартных регрессоров `scikit-learn`.

2. Разработан метод преобразования исходного датасета размером $M \times N$ с комплексными числами в датасет размером $2M \times (2N - 1)$ с вещественными числами, что позволяет использовать для построения достоверных моделей регрессии комплексных чисел стандартные регрессоры `scikit-learn`.

3. Игнорирование даже незначительной погрешности измерения предикторов и целевой переменной может привести к крайне большой погрешности прогноза экстраполяции регрессии (как для наборов данных, содержащих комплексные числа, так и для наборов данных, содержащих только вещественные числа).

4. Модели, обученные на наборах данных с мультиколлинеарными предикторами по условию оптимизации стандартных метрик из библиотеки `scikit-learn`, могут привести к неустойчивым прогнозам экстраполяции с недопустимой погрешностью.

5. Для моделей, обучаемых на наборах данных с мультиколлинеарными предикторами, для получения устойчивого прогноза экстраполяции с допустимой погрешностью следует использовать регуляризацию с учетом погрешности измерения предикторов и целевой переменной.

Список литературы

1. Тихонов А. Н., Гончарский А. В., Степанов В.В., Ягола А. Г. Регуляризирующие алгоритмы и априорная информация. – М.: Наука, 1983. – 200 с.