ПОВЫШЕНИЕ ТОЧНОСТИ ЭКСТРАПОЛЯЦИИ РЕГРЕССИИ ДЛЯ ДАННЫХ, СОДЕРЖАЩИХ КОМПЛЕКСНЫЕ ЧИСЛА, ПУТЕМ УЧЕТА ПОГРЕШНОСТИ ИЗМЕРЕНИЯ ПЕРЕМЕННЫХ

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, значения которых являются комплексными числами. 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, значения которых также являются комплексными числами. 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 подбором оптимальных значений их гиперпараметров по сетке с кросс-валидацией.

2) двухслойная нейронная сеть прямого распространения из библиотеки Keras с подбором оптимального числа нейронов на каждом слое по условию оптимизации коэффициента детерминации R^2.

3) регуляризация Тихонова с определением коэффициента регуляризации α методом обобщенной невязки [1].

Алгоритмы Scikit-learn и TensorFlow не поддерживают работу с комплексными числами, что не позволяет построить непосредственные модели регрессии для данных, содержащих комплексные числа, с помощью стандартных регрессоров Scikit-learn и нейронных сетей TensorFlow. 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 позволяет использовать для построения достоверных моделей регрессии комплексных чисел стандартные регрессоры Scikit-learn и нейронные сети TensorFlow (подробнее смотри репозиторий автора https://github.com/GorshkovAndrey/Regression_of_complex_numbers_using_Sklearn_and_TensorFlow).

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 **вещественных** уравнений $\mathrm{Az} = \mathrm{b}$ для алгоритма регуляризации Тихонова имеет следующий вид

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

Уравнение обобщенной невязки для определения коэффициента регуляризации α имеет следующий вид

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

$$h = \delta_{\mathrm{A}}\|\mathrm{A}\| \quad (3)$$

$$d = \delta_{\mathrm{b}}\|\mathrm{b}\| \quad (4)$$

где $\|\_\|$ – 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 δ_{A} = 0,01, δ_{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 прогнозов для нейронной сети, для модели регуляризации Тихонова и **точный прогноз** Accurate.

Таблица 1 – Метрики и результаты прогноза целевой переменной для dataset_1

(random_state = 1902 для train_test_split)

	LinearRegr	Ridge	Lasso	ElasticNet	TheilSenRegr	Mean	NeuralNet_balance	NeuralNet_best	Regularized	Accurate
R2 train	0.988	0.988	0.987	0.987	0.987	0.988	0.973	0.973	0.931	
R2 test	0.981	0.982	0.985	0.984	0.981	0.983	0.967	0.967	0.928	
b_max_complex	(305.5+440.4j)	(270.1+409.8j)	(192.3+320.6j)	(222.3+343.6j)	(272+427.6j)	(252.4+388.4j)	(172.2+276.5j)	(172.2+276.5j)	(159.7+324.4j)	[(147.1+303.7j)]
b_max_modul	536.0	490.8	373.9	409.2	506.7	463.2	325.7	325.7	361.6	[337.4]
b_max_arg, DEG	55.3	56.6	59.0	57.1	57.5	57.0	58.1	58.1	63.8	[64.2]

Таблица 2 – Метрики и результаты прогноза целевой переменной для dataset_1

(random_state = 1903 для train_test_split)

	LinearRegr	Ridge	Lasso	ElasticNet	TheilSenRegr	Mean	NeuralNet_balance	NeuralNet_best	Regularized	Accurate
R2 train	0.986	0.986	0.986	0.986	0.986	0.986	0.974	0.974	0.932	
R2 test	0.985	0.984	0.985	0.985	0.985	0.985	0.961	0.961	0.929	
b_max_complex	(208.3+343.9j)	(166.9+329.5j)	(220.4+327.8j)	(217.1+327.2j)	(194.9+341.7j)	(201.5+334j)	(173.8+276.7j)	(173.8+276.7j)	(160+325.1j)	[147.1+303.7j]
b_max_modul	402.0	369.4	395.0	392.7	393.4	390.1	326.8	326.8	362.3	[337.4]
b_max_arg, DEG	58.8	63.1	56.1	56.4	60.3	58.9	57.9	57.9	63.8	[64.2]

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

	LinearRegr	Ridge	Lasso	ElasticNet	TheilSenRegr	Mean	NeuralNet_best	Regularized	Accurate
R2	0.987	0.986	0.986	0.986	0.986	0.986	0.969	0.933	
b_max_complex	(210.9+350.8j)	(173.2+330.8j)	(185.3+314.6j)	(192+321.2j)	(201.8+364.4j)	(192.6+336.4j)	(178+272.4j)	(160+325.1j)	(147.1+303.7j)
b_max_modul	409.4	373.4	365.1	374.2	416.6	387.6	325.4	362.3	337.4
b_max_arg, DEG	59.0	62.4	59.5	59.1	61.0	60.2	56.8	63.8	64.2

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

	LinearRegr	Ridge	Lasso	ElasticNet	TheilSenRegr	Mean	NeuralNet_balance	NeuralNet_best	Regularized	Accurate
R2 train	0.99	0.99	0.989	0.989	0.989	0.989	0.97	0.97	0.936	
R2 test	0.977	0.978	0.98	0.98	0.977	0.979	0.966	0.966	0.907	
b_max_complex	(288.5+435.5j)	(273.2+420.6j)	(231.6+337.5j)	(231.4+337.6j)	(271.9+439.1j)	(259.3+394.1j)	(172.1+275.1j)	(172.1+275.1j)	(159.9+324.5j)	[(147.1+303.7j)]
b_max_modul	522.3	501.6	409.3	409.3	516.5	471.7	324.5	324.5	361.8	[337.4]
b_max_arg, DEG	56.5	57.0	55.5	55.6	58.2	56.7	58.0	58.0	63.8	[64.2]

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

	LinearRegr	Ridge	Lasso	ElasticNet	TheilSenRegr	Mean	NeuralNet_balance	NeuralNet_best	Regularized	Accurate
R2 train	0.987	0.987	0.987	0.987	0.987	0.987	0.972	0.972	0.93	
R2 test	0.985	0.986	0.987	0.986	0.983	0.986	0.96	0.96	0.939	
b_max_complex	(102.5+215.9j)	(153.5+297.8j)	(155.8+281j)	(158.6+292.2j)	(91.8+216.5j)	(132.4+260.7j)	(172.4+278.3j)	(172.4+278.3j)	(160.4+325.4j)	[(147.1+303.7j)]
b_max_modul	239.0	335.0	321.3	332.5	235.1	292.4	327.4	327.4	362.8	[337.4]
b_max_arg, DEG	64.6	62.7	61.0	61.5	67.0	63.1	58.2	58.2	63.8	[64.2]

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

	LinearRegr	Ridge	Lasso	ElasticNet	TheilSenRegr	Mean	NeuralNet_best	Regularized	Accurate
R2	0.987	0.987	0.987	0.987	0.987	0.987	0.968	0.933	
b_max_complex	(190+321.4j)	(170.4+329.7j)	(188.3+299.4j)	(207.6+310.3j)	(175.5+323.5j)	(186.4+316.9j)	(178.5+272.6j)	(160.4+325.4j)	(147.1+303.7j)
b_max_modul	373.4	371.1	353.7	373.3	368.0	367.6	325.8	362.8	337.4
b_max_arg, DEG	59.4	62.7	57.8	56.2	61.5	59.5	56.8	63.8	64.2

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

	LinearRegr	Ridge	TheilSenRegr	Mean	NeuralNet_best	Regularized	Accurate
R2	1.0	1.0	1.0	1.0	0.977	1.0	
b_max_complex	(147.1+303.7j)	(147.1+303.7j)	(147.1+303.7j)	(147.1+303.7j)	(179.2+272.1j)	(147.1+303.7j)	(147.1+303.7j)
b_max_modul	337.4	337.4	337.4	337.4	325.8	337.4	337.4
b_max_arg, DEG	64.2	64.2	64.2	64.2	56.6	64.2	64.2

Примечание. dataset_0 сгенерирован при коэффициентах z = (0.04 + *j*0.02, 0.035 + *j*0.015, 0.03 + *j*0.02, 0.02 + *j* 0.012, 0.016 + *j*0.01)

3. Анализ прогнозов различных моделей

Для dataset_1 комплексный диапазон осредненного прогноза Mean различных моделей регрессоров Scikit-learn составил (192,6 ÷ 252,4) + *j*(276,5 ÷ 336,4). Комплексный диапазон прогноза двухслойной нейронной сети составил (172,2 ÷ 178,0) + *j*(272,4 ÷ 276,7). Комплексный диапазон прогноза модели регуляризации Тихонова, учитывающей информацию о погрешности измерения предикторов и целевой переменной, составил (159,7 ÷ 160,0) + *j*(324,4 ÷ 325,1).

Для dataset_2 комплексный диапазон осредненного прогноза Mean различных моделей регрессоров Scikit-learn составил (132,4 ÷ 259,3) + *j*(260,7 ÷ 394,1). Комплексный диапазон прогноза двухслойной нейронной сети составил (172,1 ÷ 178,5) + *j*(272,6 ÷ 278,3). Комплексный диапазон прогноза модели регуляризации Тихонова, учитывающей информацию о погрешности измерения предикторов и целевой переменной, составил (159,9 ÷ 160,4) + *j*(324,5 ÷ 325,4).

Для dataset_0 значение прогноза составляет 147,1 + *j*303,7. Это же значение получается при вычислении **точного прогноза** по выражению $b\_\max = \mathrm{A}\_\max \cdot \mathrm{z}$, где z = (0.04 + *j*0.02, 0.035 + *j*0.015, 0.03 + *j*0.02, 0.02 + *j* 0.012, 0.016 + *j*0.01) – коэффициенты регрессии для генерации dataset_0. Значение прогноза модели регуляризации Тихонова с учетом того, что погрешности измерения переменных для dataset_0 равны нулю, также составило 147,1 + *j*303,7. Значение прогноза двухслойной нейронной сети для dataset_0 составило 179,2 + *j*272,1.

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). Максимальный модуль разброса (variance) осредненного прогноза различных моделей регрессоров достигает 24% при модуле смещения (bias) 18%.

При этом результаты прогнозов двухслойной нейронной сети прямого распространения являются достаточно устойчивыми и имеют приемлемую точность. Максимальный модуль разброса (variance) прогноза достигает 1,5% при модуле смещения (bias) 12%.

Наилучшие устойчивые прогнозы с приемлемой точностью выдает модель регуляризации Тихонова, учитывающая информацию о погрешности измерения предикторов и целевой переменной. Максимальный модуль разброса (variance) прогноза модели регуляризации Тихонова достигает 0,2% при модуле смещения (bias) 7%.

4. Выводы

1. Алгоритмы Scikit-learn и TensorFlow не поддерживают работу с комплексными числами, что не позволяет построить непосредственные модели регрессии для данных, содержащих комплексные числа, с помощью стандартных регрессоров Scikit-learn и нейронных сетей TensorFlow.

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 и нейронные сети TensorFlow.

3. Игнорирование даже незначительной погрешности измерения значений 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 регрессоров Scikit-learn, обученные на наборах данных с мультиколлинеарными предикторами по условию оптимизации стандартных метрик Scikit-learn, могут привести к неустойчивым прогнозам экстраполяции с недопустимой погрешностью.

5. Для наборов данных с мультиколлинеарными предикторами для получения устойчивого прогноза экстраполяции с допустимой погрешностью следует использовать модель регуляризации Тихонова, учитывающую информацию о погрешности измерения предикторов и целевой переменной.

6. Для получения устойчивого прогноза экстраполяции с допустимой погрешностью при отсутствии информации о погрешности измерения предикторов и целевой переменной следует использовать модели нейронных сетей прямого распространения с подбором оптимального числа нейронов на каждом слое по условию оптимизации стандартных метрик Scikit-learn.

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.