ПОВЫШЕНИЕ ТОЧНОСТИ ЭКСТРАПОЛЯЦИИ РЕГРЕССИИ 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. 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, 1500, 1800, 2000, 1500). Датасет dataset_0 использовать только для валидации обученных моделей.

РЕШЕНИЕ ЗАДАЧИ

1. Характеристика датасетов

Значения предикторов в каждом наборе данных являются положительными вещественными числами, максимальное значение которых не превышает 1269. Значения целевой переменной в каждом наборе данных также являются положительными вещественными числами, максимальное значение которых не превышает 115.

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].

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 $\mathrm{Az} = \mathrm{b}$ для модели на основе 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 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 = 1937 для train_test_split)

	LinRegr	Ridge	Lasso	ElastNet	TheilSen	Mean	Regularized	Accurate
R2 train	0.981	0.981	0.981	0.981	0.981	0.981	0.923	
R2 test	0.979	0.979	0.979	0.979	0.979	0.979	0.916	
b_max	373.321	353.358	340.636	348.531	377.779	358.725	262.794	242.5

Таблица 2 – Метрики и результаты прогноза целевой переменной для dataset_1
(random_state = 1941 для train_test_split)

	LinRegr	Ridge	Lasso	ElastNet	TheilSen	Mean	Regularized	Accurate
R2 train	0.981	0.981	0.981	0.981	0.981	0.981	0.922	
R2 test	0.979	0.979	0.979	0.979	0.979	0.979	0.916	
b_max	277.869	276.078	280.392	277.265	276.307	277.582	262.513	242.5

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	0.981	0.981	0.980	0.981	0.981	0.981	0.921	
b_max	316.934	303.941	291.151	307.120	323.745	308.578	262.533	242.5

Таблица 4 – Метрики и результаты прогноза целевой переменной для dataset_2
(random_state = 1968 для train_test_split)

	LinRegr	Ridge	Lasso	ElastNet	TheilSen	Mean	Regularized	Accurate
R2 train	0.981	0.981	0.981	0.981	0.981	0.981	0.922	
R2 test	0.980	0.980	0.980	0.980	0.979	0.980	0.924	
b_max	114.390	149.757	208.054	215.942	119.067	161.442	262.611	242.5

Таблица 5 – Метрики и результаты прогноза целевой переменной для dataset_2
(random_state = 1901 для train_test_split)

	LinRegr	Ridge	Lasso	ElastNet	TheilSen	Mean	Regularized	Accurate
R2 train	0.980	0.980	0.980	0.980	0.980	0.980	0.921	
R2 test	0.982	0.982	0.981	0.981	0.981	0.981	0.911	
b_max	229.391	267.201	255.334	267.729	232.893	250.510	262.074	242.5

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	0.981	0.981	0.981	0.981	0.981	0.981	0.923	
b_max	183.793	185.529	241.599	219.564	201.463	206.390	262.489	242.5

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	242.5	242.5	242.5	242.5	242.5	242.5

Примечание. dataset_0 сгенерирован при коэффициентах z = (0.04, 0.035, 0.03, 0.02, 0.016)

3. Анализ прогнозов различных моделей

Для dataset_1 диапазон осредненного прогноза Mean различных моделей регрессоров scikit-learn составил от 277,6 до 358,7. Диапазон прогноза регуляризованного решения, учитывающего информацию о погрешности измерения предикторов и целевой переменной, составил от 262,5 до 262,8.

Для dataset_2 диапазон осредненного прогноза Mean различных моделей регрессоров из библиотеки scikit-learn составил от 161,4 до 250,5. Диапазон прогноза регуляризованного решения, учитывающего информацию о погрешности измерения предикторов и целевой переменной, составил от 262,1 до 262,6.

Для dataset_0 значение прогноза составляет 242,5. Это же значение получается при вычислении **точного прогноза** по выражению $b\_\max = \mathrm{A}\_\max \cdot \mathrm{z}$, где z = (0.04, 0.035, 0.03, 0.02, 0.016) – коэффициенты регрессии для генерации dataset_0. Значение прогноза регуляризованного решения с учетом того, что погрешности измерения переменных для dataset_0 равны нулю, также составило 242,5.

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 ±38% при смещении (bias) 7%, тогда как разброс (variance) прогноза модели регуляризации Тихонова составил ±0,1% при смещении (bias) 8%.

4. Выводы

1. Игнорирование даже незначительной погрешности измерения переменных может привести к крайне большой погрешности прогноза экстраполяции.

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

3. 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.