Отчет

ЧЕРНОВ АНДРЕЙ

1. Отбор признаков

(1) Я посчитал выборочную дисперсию всех признаков и сразу удалил признак, у которого дисперсия оказалась равной нулю.

(2) Я отделил бинарные признаки от количественных, для количественных признаков и y построил boxplot(рис. 1 по оси абсцисс номер признака, по оси ординат значения, последний - это y). Оказалось, что у очень многих признаков мало ненулевых значений, я удалил признак, у которого только два ненулевых значения(десятый признак), хотя, конечно, в данном месте желательно провести дополнительный анализ и понять какие из разреженных признаков нужны, а какие нет. Также удалил три образца, которые соответствуют трем экстремальным значениям y. Далее я отбирал признаки двумя разными способами.

(3) Первый способ - это корреляционный анализ. С помощью теста Спирмена я построил корреляционную матрицу и посчитал p-value. Я отбирал количественные признаки таким образом, чтобы значение p-value между признаками, присутсвутющими в выборке не было меньше, чем $1e-05$. Бинарные признаки я не трогал. Я оставил 18 признаков.

(4) Второй способ - это посчитать information gain. Собственно, я оставил 20 наиболее информативных признаков.

Author's address: Чернов Андрей, chernov.andrey.98@mail.ru.

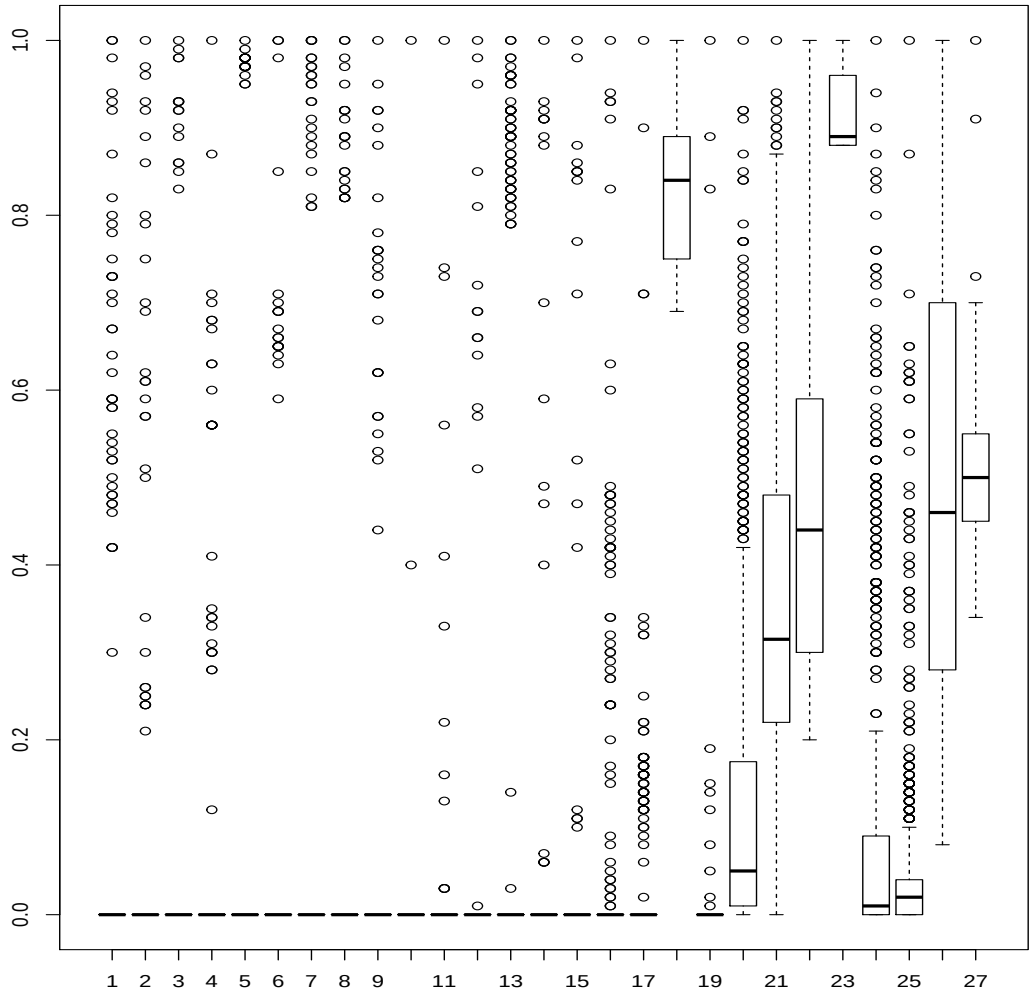

Рис. 1.

2. Выбор модели

Для начала я решил попробовать обычную линеную регрессию, чтобы получить базовые результаты. Несмотря на то, что данных и признаков не очень много я решил попробовать полносвязную нейронную сеть на полном наборе признаков(исключая только константные) с функцией активации: RELU и l1 регуляризацией. Моя идея была в том, что l1 регуляризация и RELU могут занулить веса неинформативных признаков. К сожалению, у меня не получилось заставить работать l1 регуляризацию и я ее убрал с конечной модели. Сама модель состоит из трех полносвязных словев с параметрами 32, 16 и 1 соответсвенно. Также я применил градиентный бустинг над решающими деревьями(использовал библиотеку XGBoost).

3. Результаты

Для оценки результатов, я обучал модели 20 раз, каждый раз перемешивая датасет. В качестве тренировочных данных я брал 500 образцов, а остальные 105 использовал в качестве тестовых данных. Data0 - это датасет со всем признаками, исключая константные. Data1 - это датасет с признаками, отобранными корреляционным анализом. Data2 - это датасет с признаками, отобранными с помощью gain infromation. В качестве метрик качества я использовал:

mae и R^2. На вход полносвязной сети я подавал только Data0. Для линейной регрессии:

	Метрики		
	mae train	mae test	R^2
Data0	0.032	0.032	0.57
Data1	0.035	0.036	0.51
Data2	0.055	0.057	0.91

Для бустинга:

	Метрики		
	mae train	mae test	R^2
Data0	0.032	0.034	0.57
Data1	0.0219	0.0064	1.15e-3
Data2	0.0219	0.0064	1.15e-3

Для полносвязной сети:

	Метрики		
	mae train	mae test	R^2
Data0	0.025	0.031	0.62

Судя по результатам, я не совсем эффективно отобрал признаки. Некоторые из гиперпараметров я брал по умолчанию, какие-то настраивал вручную.