Жадные регионы

Бугакова Надежда

Научный руководитель: Кураленок И. Е.

СПбАУ

21 июня 2017 г.

Мотивация

Измерения обычно содержит ошибку. Если на ошибке зависимой случайной величины построен весь регрессионный анализ, то моделирование ошибки в признаках моделируется довольно сложно. Ошибка в зависимой случайной величине:

$$Y = Y^* + N(0, \sigma^2)$$

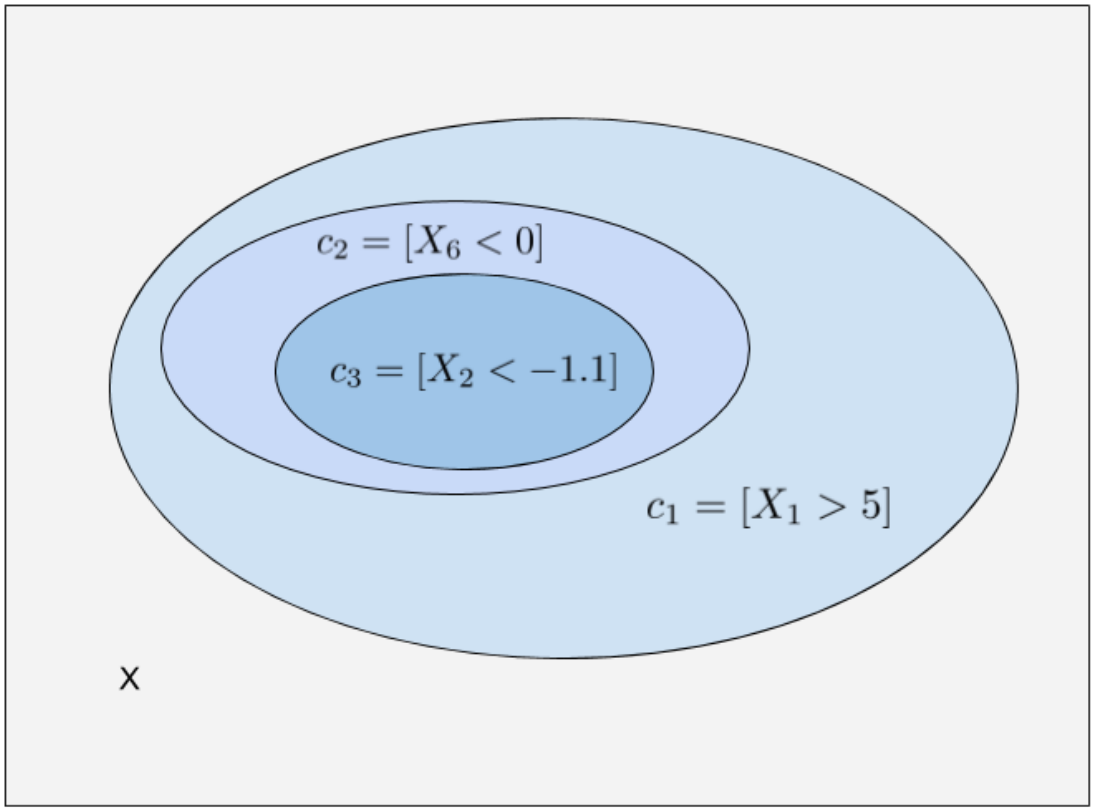
Ошибка в признаках:

$$X = X^* + ?$$

Понятие жадных регионов

Хотим простую структуру. Чем она проще, тем легче ее строить и с ней работать.

$$Y = \alpha_0 + \alpha_1 c_1(X) + \alpha_2 c_1(X) c_2(X) + \dots$$



Линейность

Модель: $Y = \alpha_0 + \alpha_1 c_1(X) + \cdots + \alpha_r c_1(X) \dots c_r(X), X \in M(n, m)$
Если присмотреться - модель линейная. Воспользуемся МНК для $k = 1..r$.

$$X_c^{(k)} = \begin{pmatrix} c_1(X_1) & c_1(X_1)c_2(X_1) & \dots & c_1(X_1)\dots c_k(X_1) \\ c_1(X_2) & c_1(X_2)c_2(X_2) & \dots & c_1(X_2)\dots c_k(X_2) \\ \dots & \dots & \dots & \dots \\ c_1(X_n) & c_1(X_n)c_2(X_n) & \dots & c_1(X_n)\dots c_k(X_n) \end{pmatrix}$$

Если выбрали c_k, то:

$$\alpha^{(k)} = (X_c^{(k)T} X_c^{(k)})^{-1} X_c^{(k)T} Y$$

c_k :

$$c_k = arg \min_{c_k} T(Y - \alpha_0^{(k)} - \alpha_1^{(k)} c_1(X) - \cdots - \alpha_k^{(k)} c_k(X))$$

Жадность

Модель: $Y = \alpha_0 + \alpha_1 c_1(X) + \cdots + \alpha_r c_1(X) \dots c_r(X), X \in M(n, m)$
Тут мы не меняем коэффициенты каждый раз, в отличие от линейного подхода:
Если выбрали c_k, то:

$$\alpha_k = arg \min_{\alpha_k} T(Y - \alpha_0 - \alpha_1 c_1(X) - \cdots - \alpha_k c_1(X) \dots c_k(X))$$

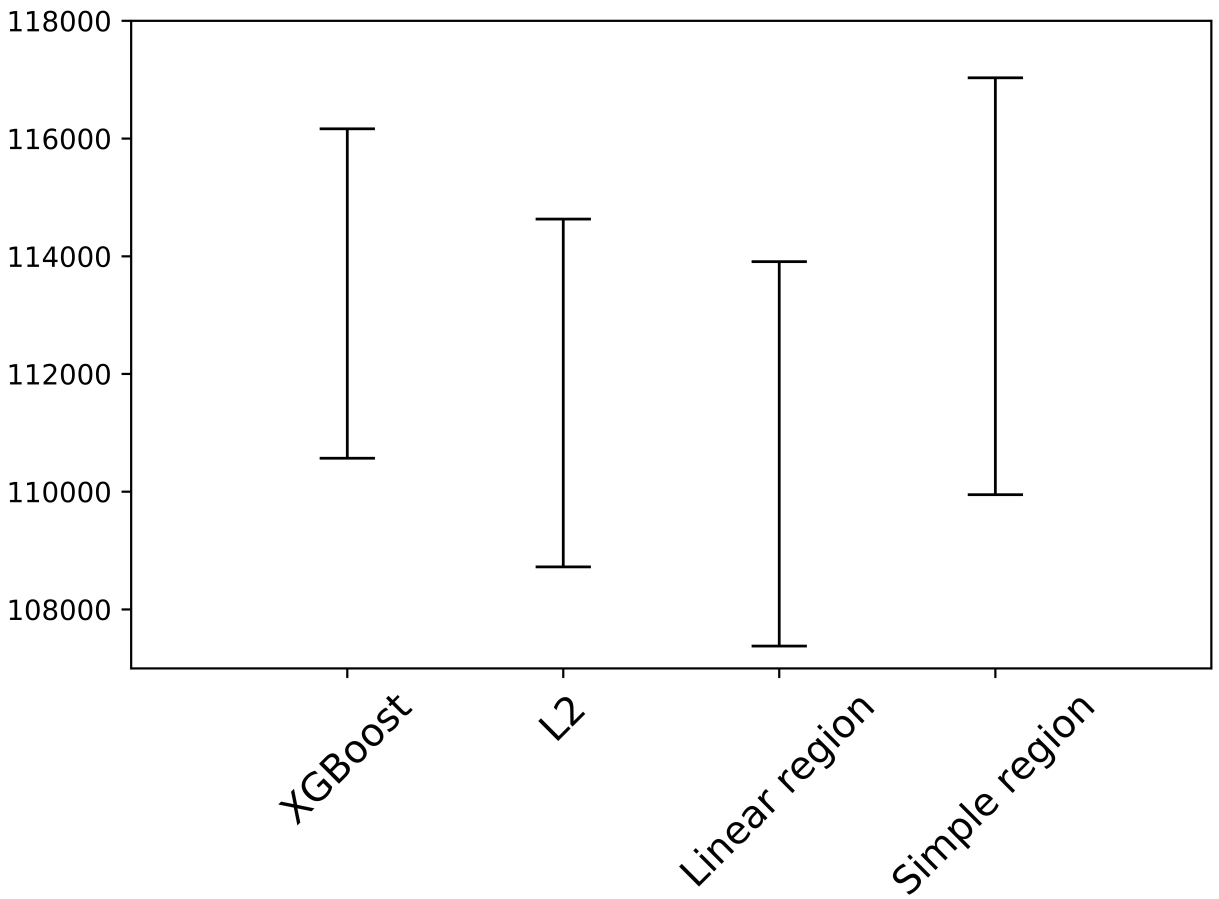
Если T = L2, то:

$$\alpha_k = mean\{Y_i - \alpha_0 - \alpha_1 c_1(X_i) - \cdots - \alpha_{k-1} c_1 \dots c_{k-1}(X_i) | c_k(X_i)\}$$

c_k выбираем из:

$$c_k = arg \min_{c_k} T(Y - \alpha_0 - \alpha_1 c_1(X) - \cdots - \alpha_k c_1(X) \dots c_k(X))$$

Результаты



Доверительные интервалы, RMSE. Датасет KC house pricing: размер - 21613,количество признаков 19.

Результаты

	XGBoost	L2	Linear	Simple
XGBoost	1.0	4.12e-08	8.24e-12	1.87e-02
L2	4.12e-08	1.0	1.4e-02	8.9e-05
Linear	8.24e-12	1.4e-02	1.0	7.98e-09
Simple	1.87e-02	8.9e-05	7.98e-09	1.0

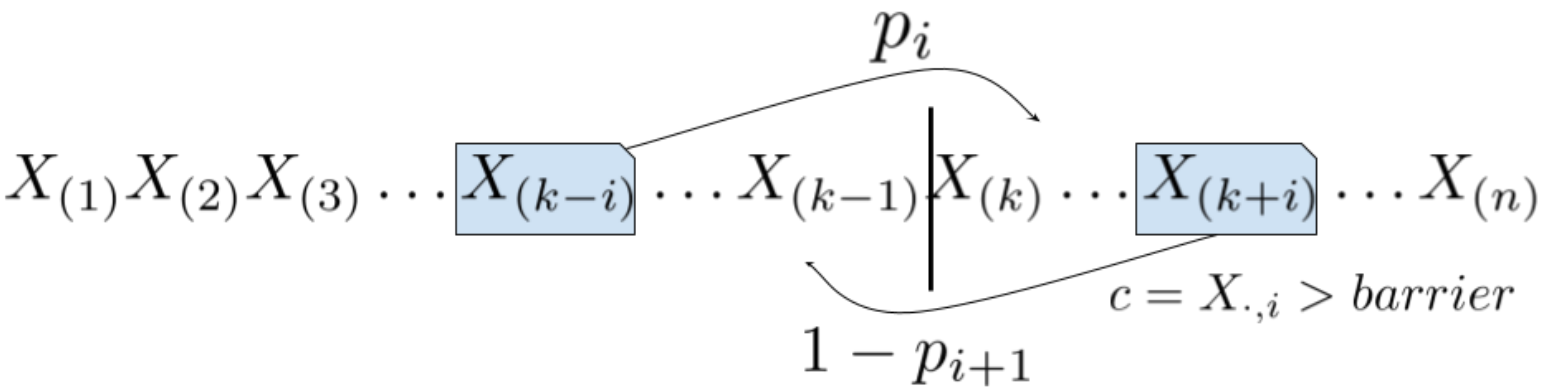
(a) Wilcoxon rank sum test. P-value matrix. Датасет KC house pricing.

	XGBoost	L2	Linear	Simple
XGBoost	0.0	5.49	6.83	2.35
L2	-5.49	0.0	2.45	-3.92
Linear	-6.83	-2.45	0.0	-5.77
Simple	-2.35	3.92	5.77	0.0

(b) Wilcoxon rank sum test. Test statistic matrix. Датасет KC house pricing.

Вероятностные регионы

Идея: Точка принадлежит региону с какой-то вероятностью, так как могут быть неточности в измерении фичей.



$p_i = (p_1)^i$

Что дальше?

- $p_1 = exp(-\alpha)$
- Разная вероятность для разных признаков.