Занятие № 7

Работа с переменными

Сапрыкин Артур
Data Scientist

Содержание

1. Масштабирование
2. Основные способы преобразования пространства
3. Основные способы преобразования категориальных переменных
4. Практика.

Масштабирование

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

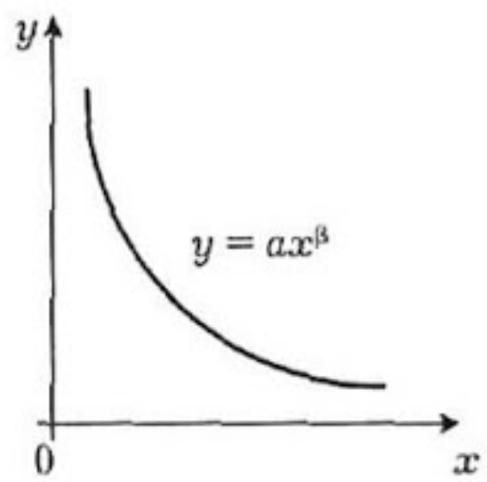
нормализация по методу **минимакс**

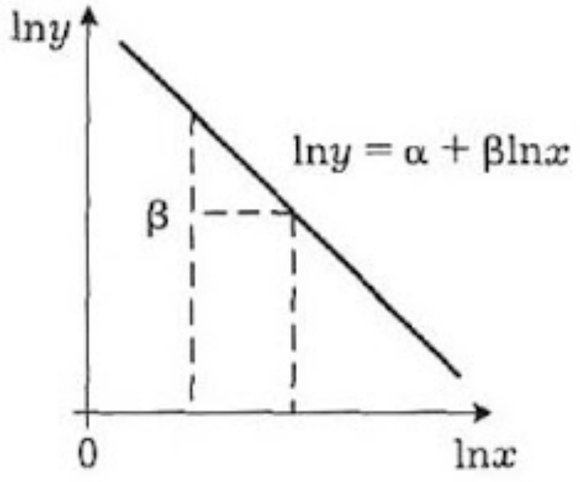
$$z = \frac{x - \mu}{\sigma}$$

Z-масштабирование

Логарифмирование

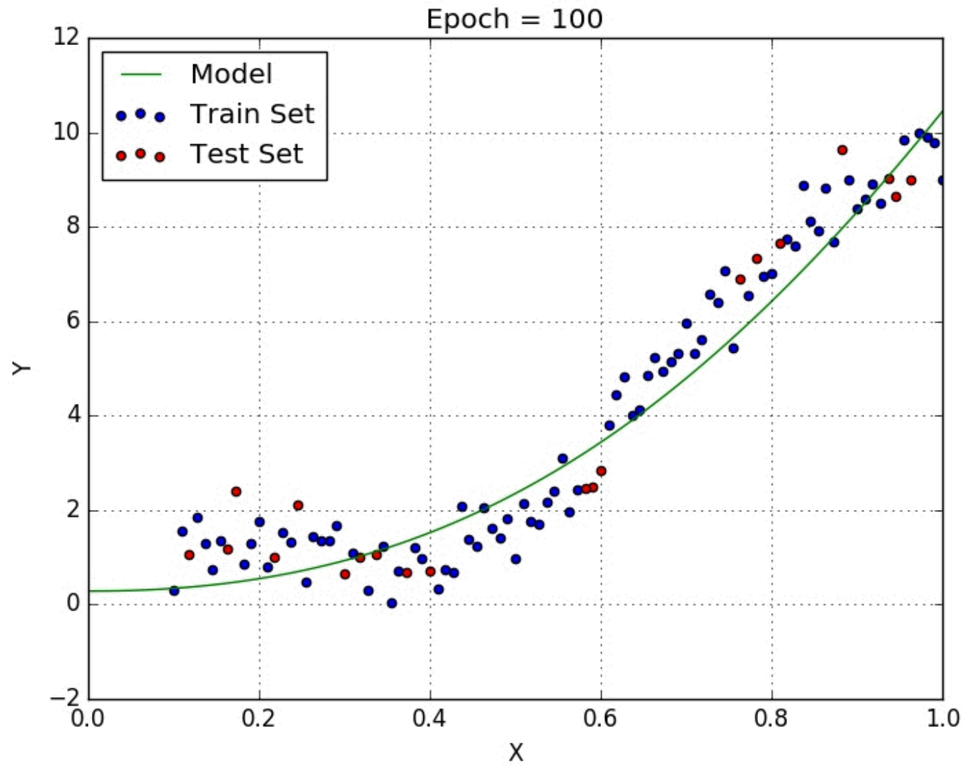
$$\ln(y) = \ln(a) + bx + u$$





Полиномиальная регрессия

$$Y = a\_1*X\_1 + (a\_2)^{2}*X\_2 + (a\_3)^{4}*X\_3 \ldots\ldots a\_n*X\_n + b$$



Обработка категориальных переменных. One-hot/ Label encoding

Company Name	Categorical value	Price
VW	1	20.000
Acura	2	10.011
Honda	3	50.000
Honda	3	10.000

VW	Acura	Honda	Price
1	0	0	20.000
0	1	0	10.011
0	0	1	50.000
0	0	1	10.000

Обработка категориальных переменных. Bins to number

User_ID	Product_ID	Gender	Age	C
1000001	P00069042	F	0-17	
1000001	P00248942	F	0-17	
1000001	P00087842	F	0-17	
1000001	P00085442	F	0-17	
1000002	P00285442	M	55+	
1000003	P00193542	M	26-35	
1000004	P00184942	M	46-50	
1000004	P00346142	M	46-50	
1000004	P0097242	M	46-50	
1000005	P00274942	M	26-35	
1000005	P00251242	M	26-35	

Средняя или медиана

User_ID	Product_ID	Gender	Age	New_Age	Occ
1000001	P00069042	F	0-17	14	
1000001	P00248942	F	0-17	14	
1000001	P00087842	F	0-17	14	
1000001	P00085442	F	0-17	14	
1000002	P00285442	M	55+	60	
1000003	P00193542	M	26-35	30	
1000004	P00184942	M	46-50	47	
1000004	P00346142	M	46-50	47	
1000004	P0097242	M	46-50	47	
1000005	P00274942	M	26-35	30	
1000005	P00251242	M	26-35	30	

Верхняя или нижняя граница

User_ID	Product_ID	Gender	Age	Lower_Age	Upper_Age	C
1000001	P00069042	F	0-17	0	17	
1000001	P00248942	F	0-17	0	17	
1000001	P00087842	F	0-17	0	17	
1000001	P00085442	F	0-17	0	17	
1000002	P00285442	M	55+	55	80	
1000003	P00193542	M	26-35	26	35	
1000004	P00184942	M	46-50	46	50	
1000004	P00346142	M	46-50	46	50	
1000004	P0097242	M	46-50	46	50	
1000005	P00274942	M	26-35	26	35	
1000005	P00251242	M	26-35	26	35	

Обработка категориальных переменных. По бизнес-логике

По бизнес логике

Zip Code	District
110044	South Delhi
110048	South Delhi
110049	South Delhi
110006	North Delhi
110007	North Delhi
110058	West Delhi
110059	West Delhi
110063	West Delhi
110064	West Delhi

По доле таргетинга

Based on Response Rate

Levels	Response_Rate	New_Level
HA014	98%	1
HA001	97%	1
HA003	93%	1
HA009	81%	2
HA015	75%	3
HA010	73%	3
HA006	66%	4
HA017	60%	4
HA007	49%	5
HA004	36%	6
HA005	31%	6
HA012	28%	7

ПРАКТИКА

Спасибо за внимание!

Сапрыкин Артур
Data Scientist

fb.com/asaprykin92

asaprykin92@gmail.com