Лекция 5. Генерация новых признаков. Скалирование данных. Устранение проблемы мультиколлинеарности

Практикум по программированию, 5 семестр

Иван Евгеньевич Бугаенко,
ассистент каф. ПМиФИ

Что такое генерация новых признаков?

Генерация признаков – это процесс создания новых признаков (столбцов) на основе уже имеющихся значений

Зачем вообще нужна генерация признаков?

Генерация признаков может увеличить ~~(а может не увеличить)~~ итоговые метрики модели и пофиксить проблемы в данных

1. Повышение точности моделей. Хорошо подобранные признаки помогают алгоритму лучше понять структуру данных и закономерности
2. Учет сложных зависимостей. Иногда исходные данные не содержат явных признаков, которые могут хорошо объяснять целевую переменную
3. Упрощение модели. С помощью создания новых фич можно упростить признаковое пространство

Подходы по генерации признаков

1. Генерация на основе свойств признакового пространства
2. Генерация на основе арифметических операций
3. Генерация на основе агрегации данных
4. Статистические преобразования (скалирование)
5. Функциональные преобразования
6. Формирование интервальных признаков
7. Использование сторонних данных
8. Генерация для временных рядов

Генерация на основе свойств признакового пространства

Для генерации новых признаков можно использовать формулы прикладной области (если они есть, конечно же…)

feature1	feature2
24	12
23	4
43	20
23	23
54	27

feature3 = feature1 / feature2

feature1	feature2	feature3
24	12	2
23	4	5.75
43	20	2.15
23	23	1
54	27	2

Генерация на основе арифметических операций

Новые признаки можно генерировать как всевозможные комбинации признаков и арифметических операций

feature1	feature2
24	12
23	4
43	20
23	23
54	27

feature1	feature2	/	*	+	-
24	12	2			
23	4	5.75			
43	20	2.15			
23	23	1			
54	27	2			

Генерация на основе агрегации данных

Для панельных данных (серия измерений для одного объекта) и данных с категориальными признаками можно создавать агрегации

id	t	feature
24	1	12
24	2	4
24	3	20
55	1	23
55	2	27
55	3	6

id	t	feature	mean
24	1	12	mean24
24	2	4	mean24
24	3	20	mean24
55	1	23	mean55
55	2	27	mean55
55	3	6	mean55

mean(i) – среднее значение признака feature для значения признака id=i

Статистические преобразования (скалирование)

Вычисление нового значения признака для улучшения распределения или снижения влияния выбросов / диспропорции значений

1. Стандартизация
2. MinMax-преобразование
3. Робастное скалирование

$$x_i' = \frac{x - \bar{x}}{\sigma} \qquad x_i' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad x_i' = \frac{x - \bar{m}}{\mathrm{IQR}}$$

Решение о скалировании применяется на основе df.describe()

Функциональные преобразования

Применение некоторых функциональных зависимостей к признакам для построения нелинейных связей

feature1	feature2
24	12
23	4
43	20
23	23
54	27

feature1	feature2	ln(f1)
24	12	3.178
23	4	3.135
43	20	3.761
23	23	3.135
54	27	3.989

Формирование интервальных признаков

Вещественные признаки можно представлять как интервал значений

feature1	feature2
24	12
23	4
43	20
23	23
54	27

feature1	feature2	int_f1
24	12	(23; 43]
23	4	(0; 23]
43	20	(23; 43]
23	23	(0; 23]
54	27	(43; 55]

Использование сторонних данных

Для исходного датасета можно использовать связи с другими датасетами с целью пополнения данных

id	feature1
1	12
2	4
2	20
1	23
3	27

id	feature2
1	5
2	66
3	24

join

id	feature1	feature2
1	12	5
2	4	66
2	20	66
1	23	5
3	27	24

Генерация для временных рядов

Для временных рядов можно использовать структуру с целью выделения паттернов и динамики в данных (лаговые сдвиги, вычисление разностей)

t	feature2
1	12
2	4
3	20
4	23
5	27

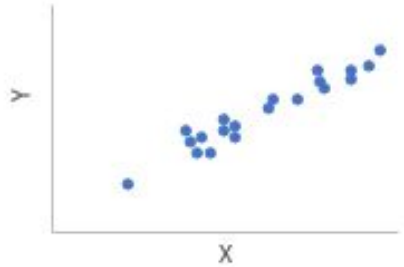
t	feature2	lag1	delta
1	12	NaN	NaN
2	4	12	8
3	20	4	-16
4	23	20	-3
5	27	23	-4

Советы по генерации признаков

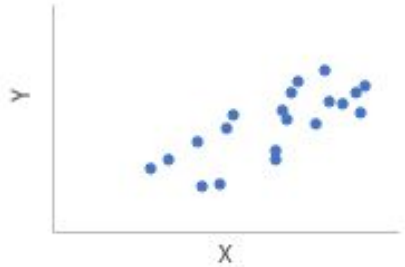
1. Если есть возможность, всегда генерируйте дополнительные признаки! Неинформативные признаки можно отсеять в результате отбора (после обучения модели)
2. Все вспомогательные вычисления осуществляются исключительно на обучающей выборке (в противном случае – утечка данных) без целевых значений. Это касается скалирования, интервальных признаков, формул, агрегация и временных рядов
3. При помощи формул прикладной области можно решать проблему **мультиколлинеарности** в данных. Генерация признаков также решает проблемы слабых зависимостей в данных и диспропорции величин

Что такое мультиколлинеарность в данных?

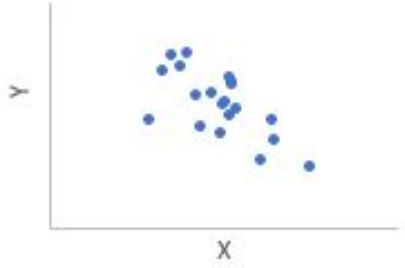
Прямая



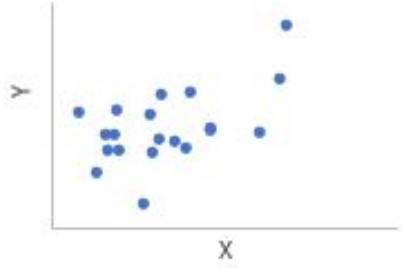
Сильная



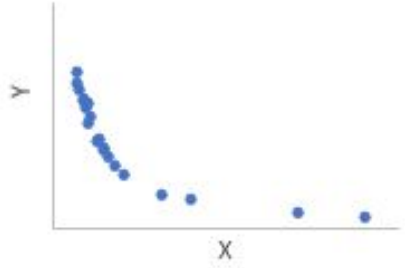
Линейная



Обратная



Слабая

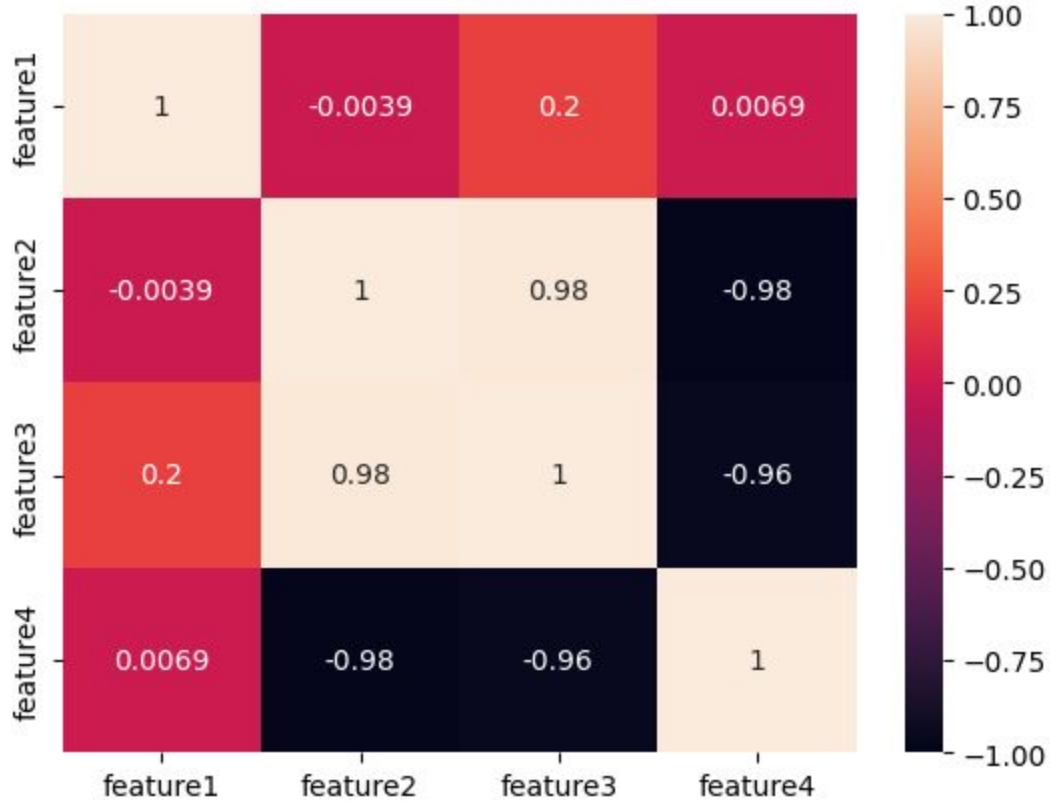


Нелинейная

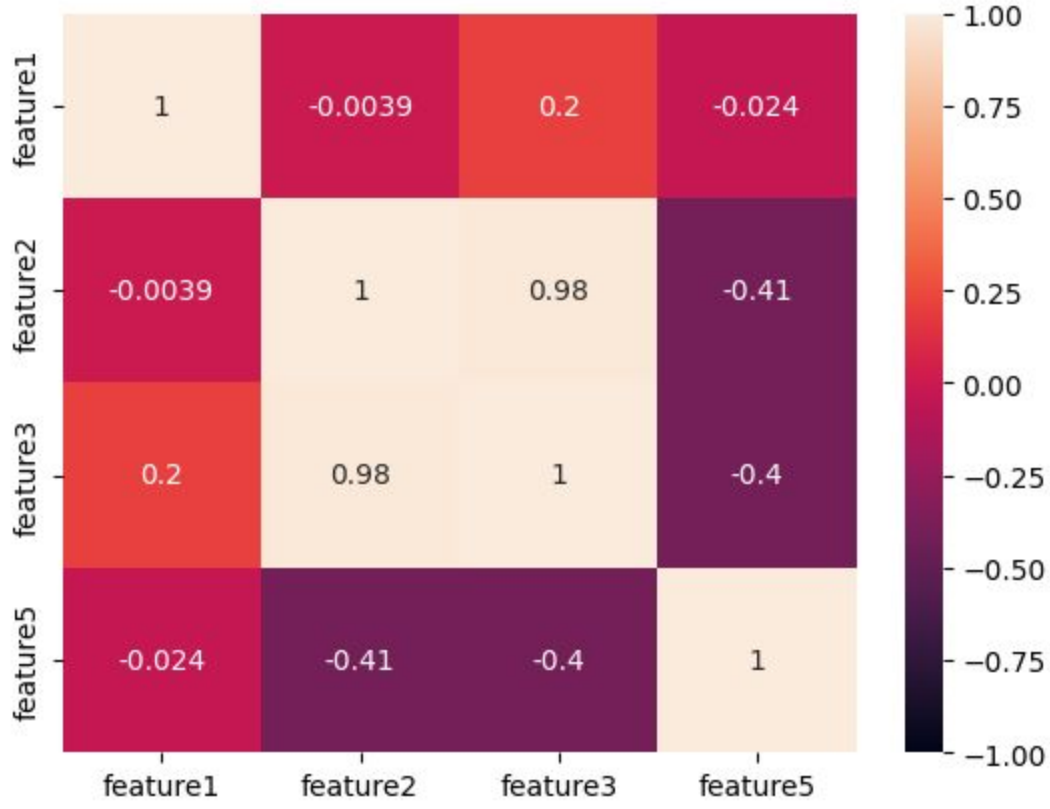
Мультиколлинеарность в данных – это свойство, которое заключается в наличии сильной линейной зависимости между парами признаков

$$r_{xy} = \frac{\sum (x_i - M_x)(y_i - M_y)}{\sqrt{\sum (x_i - M_x)^2 \cdot (y_i - M_y)^2}}$$

Величина коэффициента корреляции	0.1 - 0.3	0.3 - 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 1.0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая
		средняя		сильная	


feature1
feature2
feature3
feature4
1
-0.0039
0.2
0.0069
-0.0039
1
0.98
-0.98
0.2
0.98
1
-0.96
0.0069
-0.98
-0.96
1
feature1
feature2
feature3
feature4
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

feature5 = (feature2 / feature4) * feature3


feature1
feature2
feature3
feature5
1
-0.0039
0.2
-0.024
-0.0039
1
0.98
-0.41
0.2
0.98
1
-0.4
-0.024
-0.41
-0.4
1
feature1
feature2
feature3
feature5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00