Предобработка и генерация признаков с учетом особенностей моделей

План

Главные темы:

1. Мотивация и введение
2. Числовые признаки
3. Категориальные признаки
4. Признаки с особенной структурой
5. Группы признаков
6. Отбор признаков??
7. Пропущенные значения признаков

1. Мотивация и введение

Мотивация и введение

Примеры признаков (данные из Титаника)

Целевая переменная

Числовая переменная

Переменные-счетчики

Категориальные переменные

id

Текстовая переменная

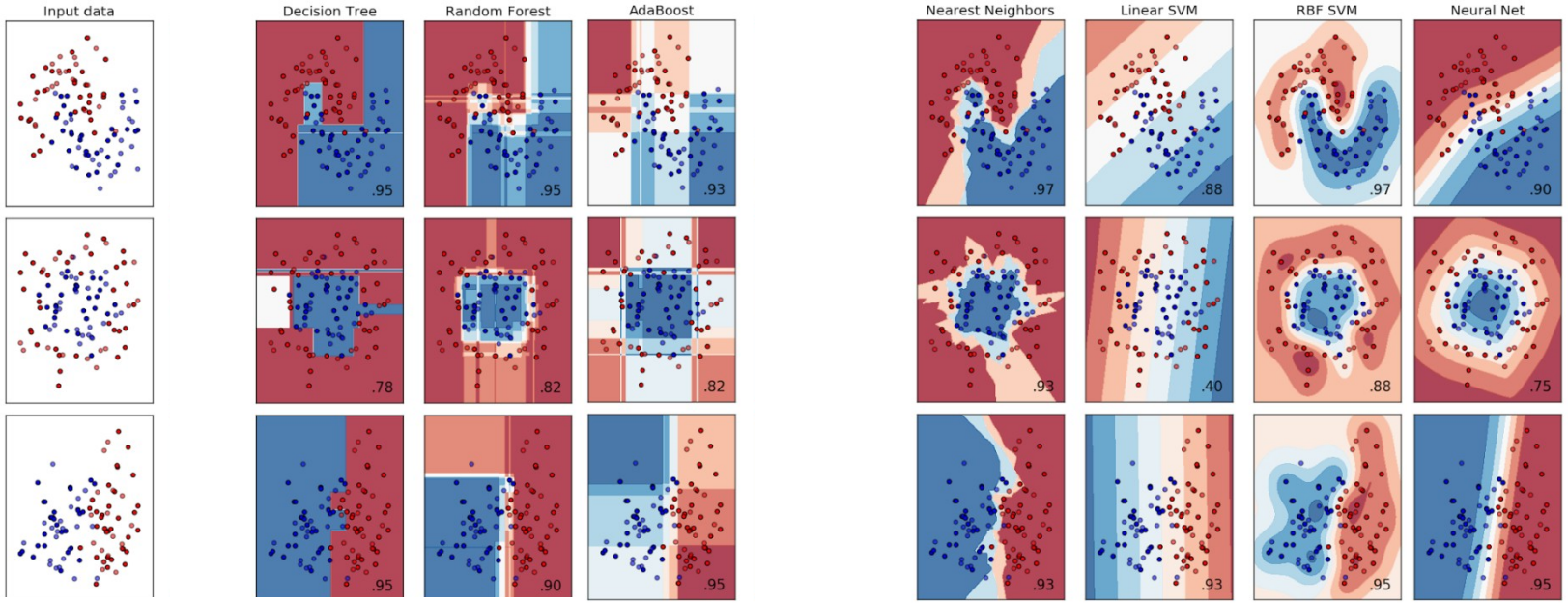
	PassengerId	Survived	Pclass	Name
0	1	0	3	Braund, Mr. Owen Harris
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...
2	3	1	3	Heikkinen, Miss. Laina
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)
4	5	0	3	Allen, Mr. William Henry
5	6	0	3	Moran, Mr. James
6	7	0	1	McCarthy, Mr. Timothy J
7	8	0	3	Palsson, Master. Gosta Leonard

	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	male	22.000000	1	0	A/5 21171	7.2500	NaN	S
1	female	38.000000	1	0	PC 17599	71.2833	C85	C
2	female	26.000000	0	0	STON/O2. 3101282	7.9250	NaN	S
3	female	35.000000	1	0	113803	53.1000	C123	S
4	male	35.000000	0	0	373450	8.0500	NaN	S
5	male	29.699118	0	0	330877	8.4583	NaN	Q
6	male	54.000000	0	0	17463	51.8625	E46	S
7	male	2.000000	3	1	349909	21.0750	NaN	S

Мотивация и введение
Предобработка признаков

Модели на деревьях

Остальные модели



«Деревянные» модели, в отличие от остальных, инварианты относительно масштабирования признаков

Classifier comparison¶, http://scikit -learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Мотивация и введение

Предобработка признаков

pclass	1	2	3
target	1	0	1

Мотивация и введение

Предобработка признаков

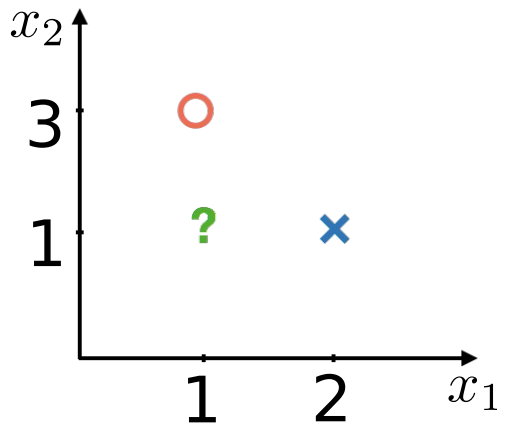
pclass	1	2	3
target	1	0	1

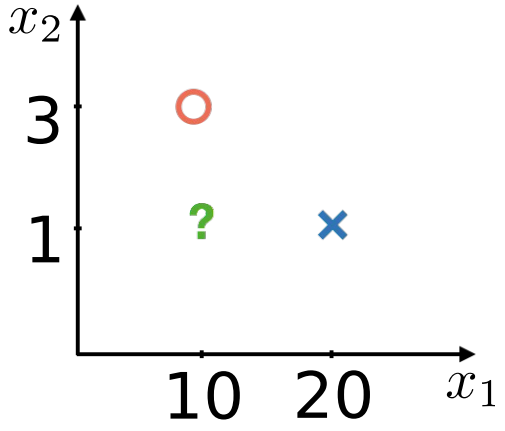
pclass
1
2
1
3

pclass==1	pclass==2	pclass==3
1		
	1	
1		
		1

Мотивация и введение

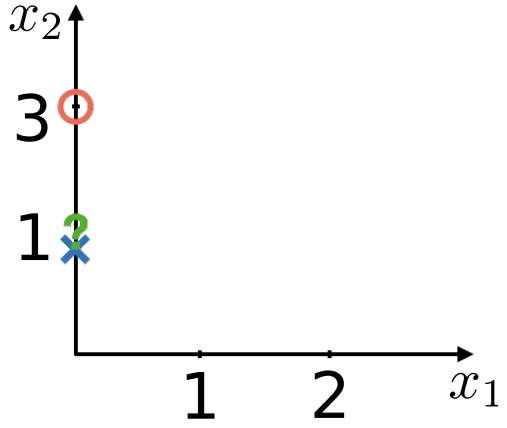
Предобработка признаков

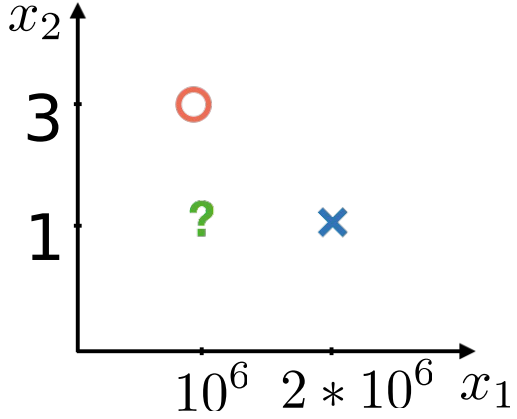




$$x_1 = x_1 * 0$$

$$x_1 = x_1 * 10^6$$





Генерация признаков

Чем руководствоваться:

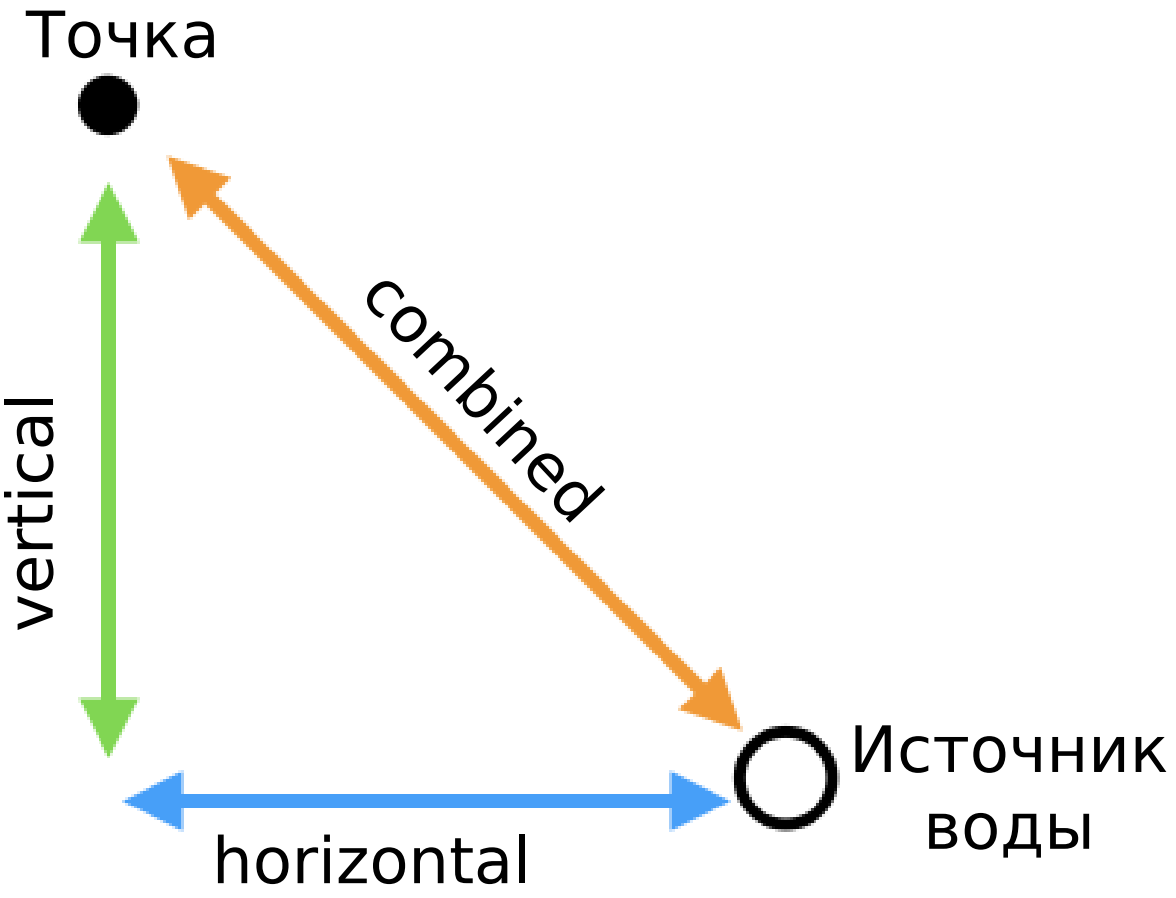
- Знание предметной области
- Исследование данных (EDA)

Мотивация и введение
Генерация признаков

Площадь: 55 м2
Цена: 107000$
Цена за 1м2: 107000$/55 м2

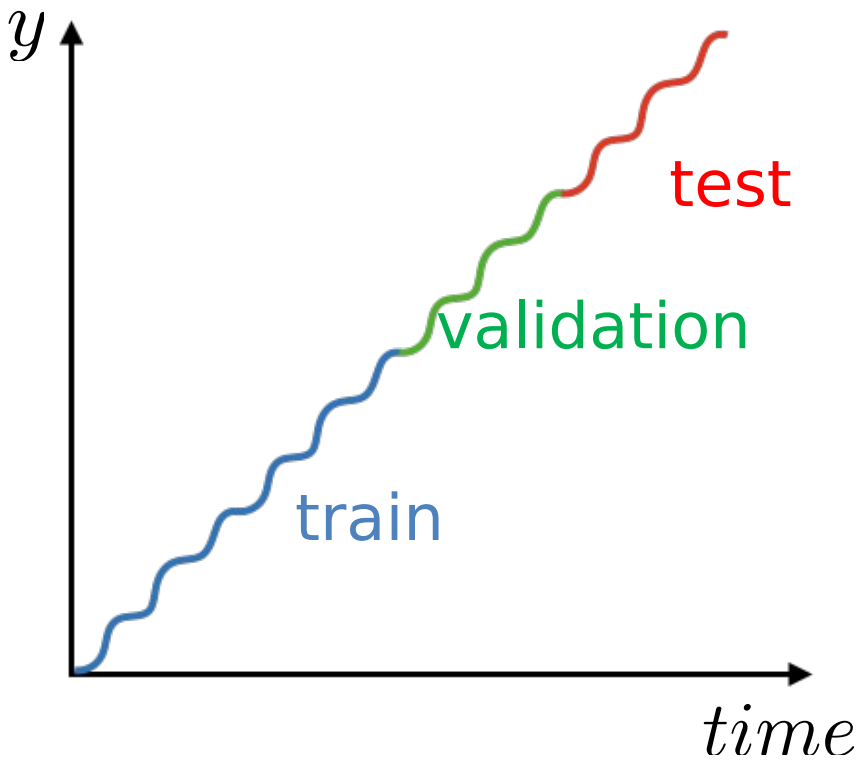
Мотивация и введение

Генерация признаков



Combined = (Horizontal** 2 + Vertical ** 2) **0.5

Мотивация и введение
Генерация признаков



Мотивация и введение

Генерация признаков

	День	Неделя	Число яблок
Train	33	5	45
	34	5	72
	35	5	81
Test	36	6	?
	37	6	?

Мотивация и введение

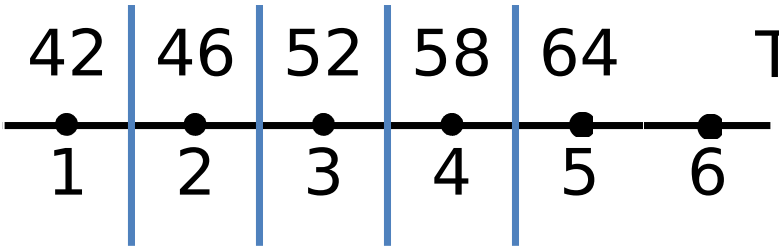
Генерация признаков

Неделя	Число яблок
1	42
2	46
3	52
4	58
5	64

	День	Неделя	Число яблок
Train	33	5	45
	34	5	72
	35	5	81
Test	36	6	?
	37	6	?

Мотивация и введение

Генерация признаков



	День	Неделя	Число яблок
Train	33	5	45
	34	5	72
	35	5	81
Test	36	6	?
	37	6	?

- В данном случае дерево решений не сможет учесть линейный тренд в зависимости целевой переменной от времени
- В то же время, можно сгенерировать признаки, которые окажутся полезным для дерева решений, но не для линейной модели

Мотивация и введение
Выводы

- Предобработка признаков зачастую необходима
- Генерация признаков - мощный метод
- Методы предобработки и генерации признаков зависят от типа модели

2. Числовые признаки

Числовые признаки

План

- Масштабирование (нормализация)
- Работа с выбросами
- Немонотонные преобразования

Числовые признаки

Масштабирование (нормализация)

1. К отрезку [0,1]

$$X = (X - X.min())/(X.max() - X.min())$$

`sklearn.preprocessing.MinMaxScaler`

2. К распределению с мат.ожиданием = 0 и дисперсией = 1

$$X = (X - X.mean())/X.std()$$

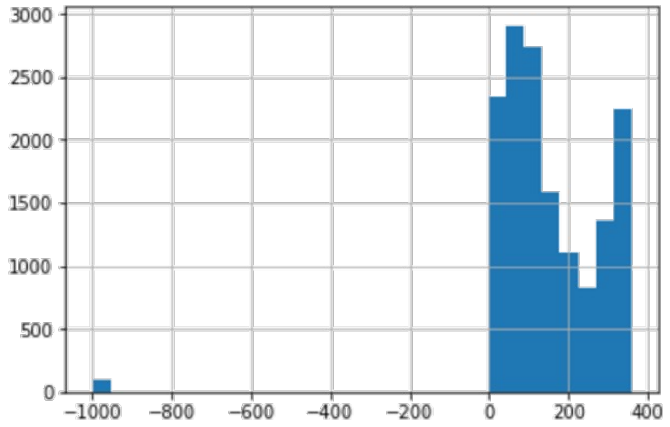
`sklearn.preprocessing.StandardScaler`

Числовые признаки
Работа с выбросами

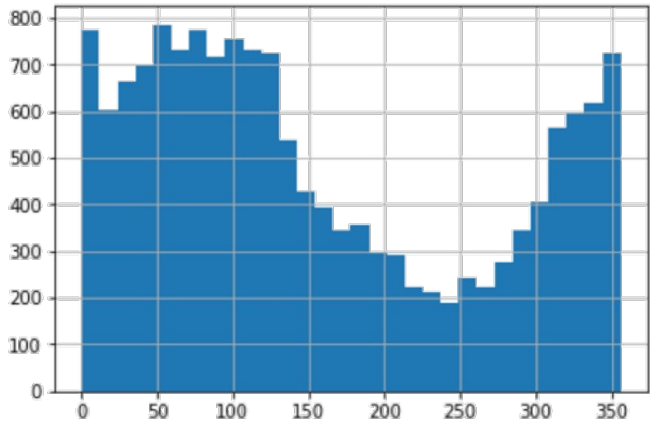
- Винсоризация

  Ограничение значений признаков по некоторому правилу (например, по значению перцентилей)

In [17]:
```
pd.Series(x).hist(bins=30) ;
```



In [18]:
```
UPPERBOUND, LOWERBOUND = np.percentile(x, [1, 99])
y = np.clip(x, UPPERBOUND, LOWERBOUND)
pd.Series(y).hist(bins=30) ;
```



Числовые признаки

Немонотонные преобразования

1. Rank

   ```
   rank([-100, 0, 1e5]) == [0,1,2]
   rank([1000,1,10]) = [2,0,1]
   ```

2. Логарифмирование:
   np.log(1 + x)

3. Возведение в степень <1:
   np.sqrt(x + 2/3)

Числовые признаки

Генерация признаков

Цена	Дробная часть
0.99	0.99
2.49	0.49
1.0	0.0
9.99	0.99

Числовые признаки

Выводы

1. Обработка числовых признаков различна для моделей на деревьях и остальных моделей:
   a. Модели на деревьях не зависят от масштабирования
   b. Другие модели очень зависят от масштабирования

Числовые признаки

Выводы

1. Предобработка числовых признаков различна для моделей на деревьях и остальных моделей:
   a. «Деревянные» модели не зависят от масштабирования
   b. Другие модели зависят от масштабирования
2. Часто используемые методы предобработки:
   a. MinMaxScaler – к отрезку [0,1]
   b. StandardScaler - к mean==0, std==1
   c. Ранг устанавливает одинаковые расстояния между отсортированными значениями
   d. np.log(1+x) and np.sqrt(1+x)

Числовые признаки

Выводы

3. Генерация признаков основана на:
    a. Знания предметной области
    b. Исследования данных (EDA)

3. Категориальные и порядковые признаки

Категориальные признаки

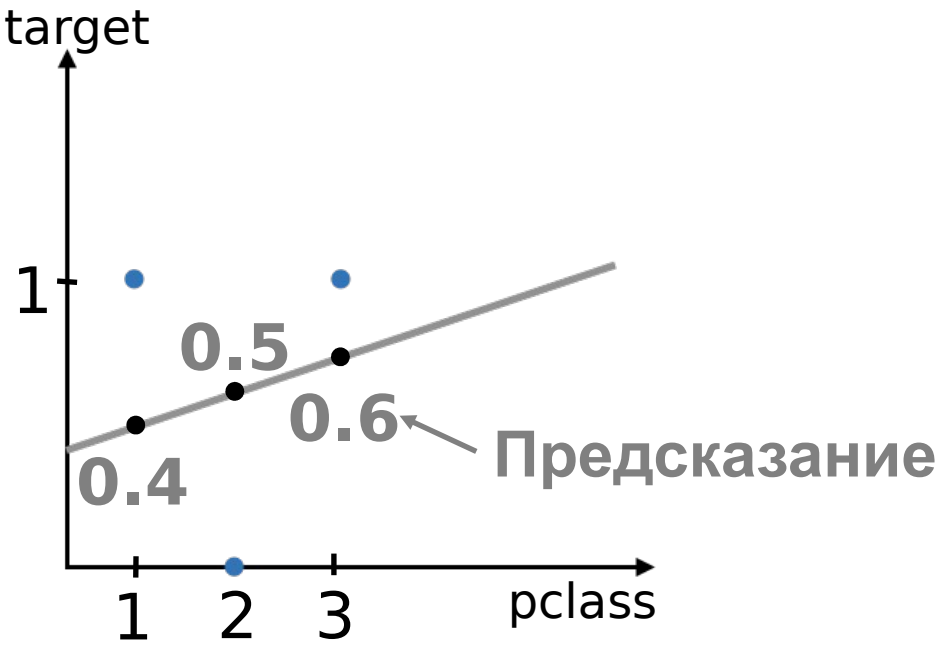
Примеры кат. признаков (датасет «Титаник»)

	PassengerId	Survived	Pclass	Name
0	1	0	3	Braund, Mr. Owen Harris
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...
2	3	1	3	Heikkinen, Miss. Laina
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)
4	5	0	3	Allen, Mr. William Henry
5	6	0	3	Moran, Mr. James
6	7	0	1	McCarthy, Mr. Timothy J
7	8	0	3	Palsson, Master. Gosta Leonard

	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	male	22.000000	1	0	A/5 21171	7.2500	NaN	S
1	female	38.000000	1	0	PC 17599	71.2833	C85	C
2	female	26.000000	0	0	STON/O2. 3101282	7.9250	NaN	S
3	female	35.000000	1	0	113803	53.1000	C123	S
4	male	35.000000	0	0	373450	8.0500	NaN	S
5	male	29.699118	0	0	330877	8.4583	NaN	Q
6	male	54.000000	0	0	17463	51.8625	E46	S
7	male	2.000000	3	1	349909	21.0750	NaN	S

Категориальные признаки
One-Hot Encoding

pclass	1	2	3
target	1	0	1



Категориальные признаки
One-Hot Encoding

pclass
1
2
1
3

pclass==1	pclass==2	pclass==3
1		
	1	
1		
		1

`pandas.get_dummies`, `sklearn.preprocessing.OneHotEncoder`

Категориальные признаки

Порядковые признаки

Класс билетов: 1,2,3

Категории водительский удостоверений: A, B, C, D

Образование: детский сад, школа, бакалавриат, магистратура, аспирантура

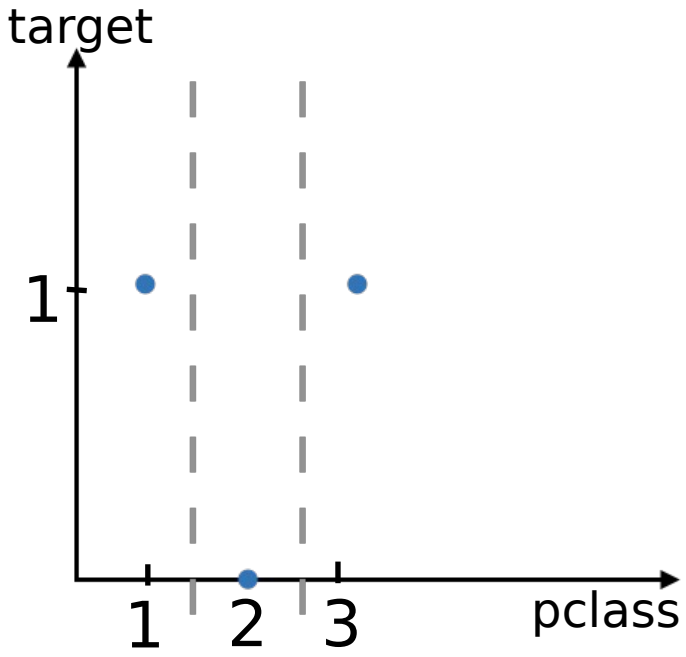
Категориальные признаки

Label Encoding

pclass	1	2	3
target	1	0	1

Категориальные признаки
Label Encoding

pclass	1	2	3
target	1	0	1



Категориальные признаки
Кодирование частотами

	K
	embarked
	S
	C
	S
	S
	S
	Q
	S
	S
	S
	C
	S
	S

[S,C,Q] -> [0.5, 0.3, 0.2]

```
encoding = titanic.groupby('Embarked').size()
encoding = encoding/len(titanic)
titanic['enc'] = titanic.Embarked.map(encoding)
```

```
from scipy.stats import rankdata
```

Категориальные признаки

Комбинации признаков

pclass	sex	pclass_sex
3	male	3male
1	female	1female
3	female	3female
1	female	1female

Pclass_sex==					
1male	1female	2male	2female	3male	3female
				1	
	1				
					1
	1				

Категориальные признаки

Генерация признаков

Места в самолете: 1A, 1B, ..., 1G, 2A, ..., 32F

seat
1A
1C
19B
27F

seat
1
3
110
162

Категориальные признаки

Выводы

1. Значения в порядковых признаках отсортированы в некотором осмысленном порядке
2. Label encoding и кодирование частотами часто используется для деревьев:
   - Label encoding переводит категории в числа
   - Кодирование частотами переводит категории в их частоту
3. One-hot encoding часто используется для остальных моделей
4. Взаимодействия категориальных признаков могут быть полезны и для «деревьев», и для остальных

4. Признаки с особенной структурой

Признаки с особенной структурой

Примеры

- Признаки с периодической структурой
  - Время
  - Датчики измерения с периодического процесса

- Признаки с пространственной структурой
  - Географические координаты
  - Координаты в признаковом пространстве (например, возраст/доход)

Признаки с особенной структурой

Периодическая структура

1. Периодичность
2. Сколько времени прошло/осталось
3. Разница между моментами времени

Признаки с особенной структурой

Периодическая структура

1. Периодичность
   Номер дня в неделе, месяц, сезон, год, секунда, минута, час.

2. Сколько времени прошло/осталось

3. Разница между моментами времени

Признаки с особенной структурой

Периодическая структура

1. Периодичность
   Номер дня в неделе, месяц, сезон, год, секунда, минута, час.
2. Сколько времени прошло/осталось
   a. С одного момента для всех данных
      Например: с 00:00:00 UTC, 1 January 1970;
   b. Момент зависит от выбора объекта
      Например: число оставшихся дней до выходных, число прошедших дней с последней поставки товара
3. Разница между моментами времени

Признаки с особенной структурой

Периодическая структура

1. Периодичность
   Номер дня в неделе, месяц, сезон, год, секунда, минута, час.
2. Сколько времени прошло/осталось
   a. С одного момента для всех данных
      Например: с 00:00:00 UTC, 1 January 1970;
   b. Момент зависит от выбора объекта
      Например: число оставшихся дней до выходных, число прошедших дней с последней поставки товара
3. Разница между моментами времени
   datetime_feature_1 - datetime_feature_2

Признаки с особенной структурой

Периодическая структура - «с тех пор»

TWO SIGMA

График цен закрытия акций на бирже

Two Sigma: Using News to Predict Stock Movements, https://www.kaggle.com/c/two-sigma-financial-news

Признаки с особенной структурой

Периодическая структура - «с тех пор»

Date	weekday	daynumber_since_first_date	is_holiday*	days_till_holidays	*returns*
20.11.2012	0	2119	False	2	*-0.101*
21.11.2012	1	2120	False	1**	*-0.064*
23.11.2012	2	2122	False	1	*-0.103*
26.11.2012	3	2125	False	1	*-0.104*
27.11.2012	4	2126	False	5	*-0.093*
28.11.2012	0	2127	False	4	*-0.125*

* Торговля на бирже происходит по рабочим дням (признак приведен в качестве иллюстрации)

** Один день до выходного в рамках рабочей недели — День Благодарения

Признаки с особенной структурой

Периодическая структура

Date	Volume, 10e6	close	open	*returns*
20.11.2012	23.0	560.91	571.56	*-0.101*
21.11.2012	13.3	561.70	564.25	*-0.064*
23.11.2012	9.7	571.50	567.39	*-0.103*
26.11.2012	22.5	589.53	575.90	*-0.104*
27.11.2012	19.0	584.78	589.60	*-0.093*

- Также в случае периодического вида признаков могут быть полезны статистики признаков и целевой переменной по скользящим окнам: mean, max, min, etc

Признаки с особенной структурой

Периодическая структура — разница между датами

Предсказание оттока клиентов

user_id	registration_date	*last_purchase_date*	*last_call_date*	date_diff	churn
14	10.02.2016	21.04.2016	26.04.2016	5	0
15	10.02.2016	03.06.2016	01.06.2016	-2	1
16	11.02.2016	11.01.2017	11.01.2017	1	1
20	12.02.2016	06.11.2016	08.02.2017	94	0

- В случае, когда наблюдаемые события происходят нерегулярно, бывают полезны величины временных промежутков между рассматриваемыми датами

Признаки с особенной структурой

Периодическая структура — разница между датами

Для конкретного пользователя:

Признаки с особенной структурой

Пространственная структура

1. Соседние («интересные») объекты
   - из обучающей выборки
   - из дополнительных данных
2. Агрегаты и статистики
3. Преобразования базиса (повороты / растяжения)

Признаки с особенной структурой

Пространственная структура

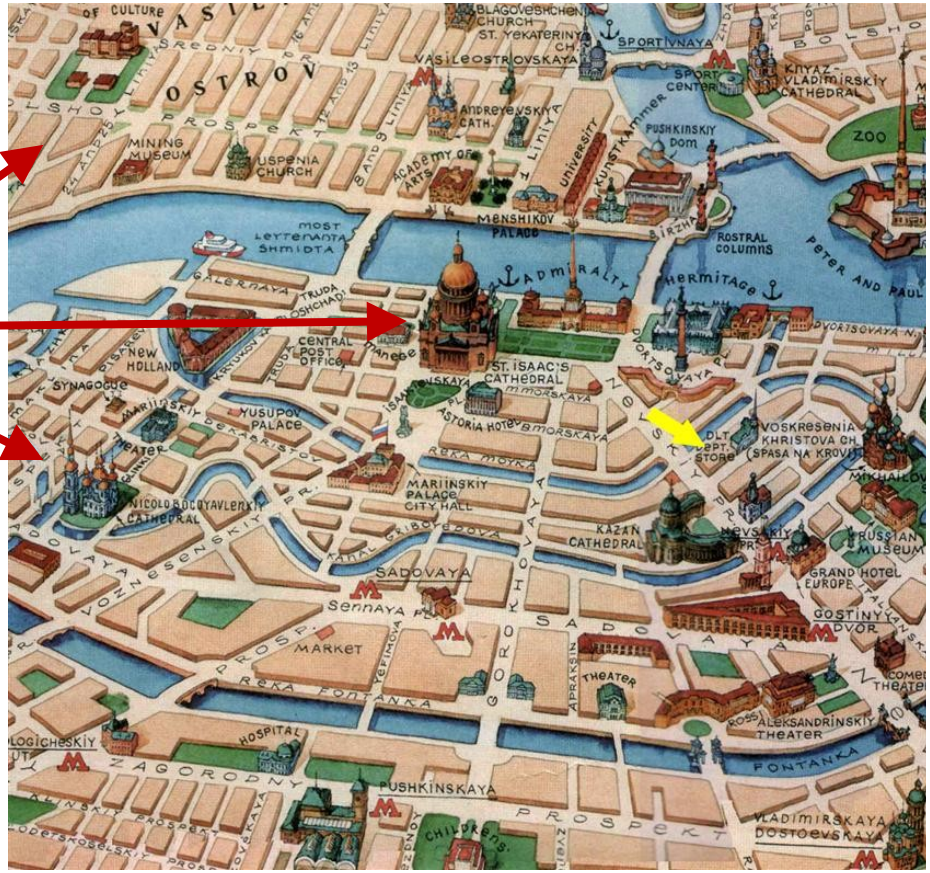
Zillow Prize: Zillow’s Home Value Prediction (Zestimate)

New York City Taxi Fare Prediction

Признаки с особенной структурой

Пространственная структура

Дополнительные данные

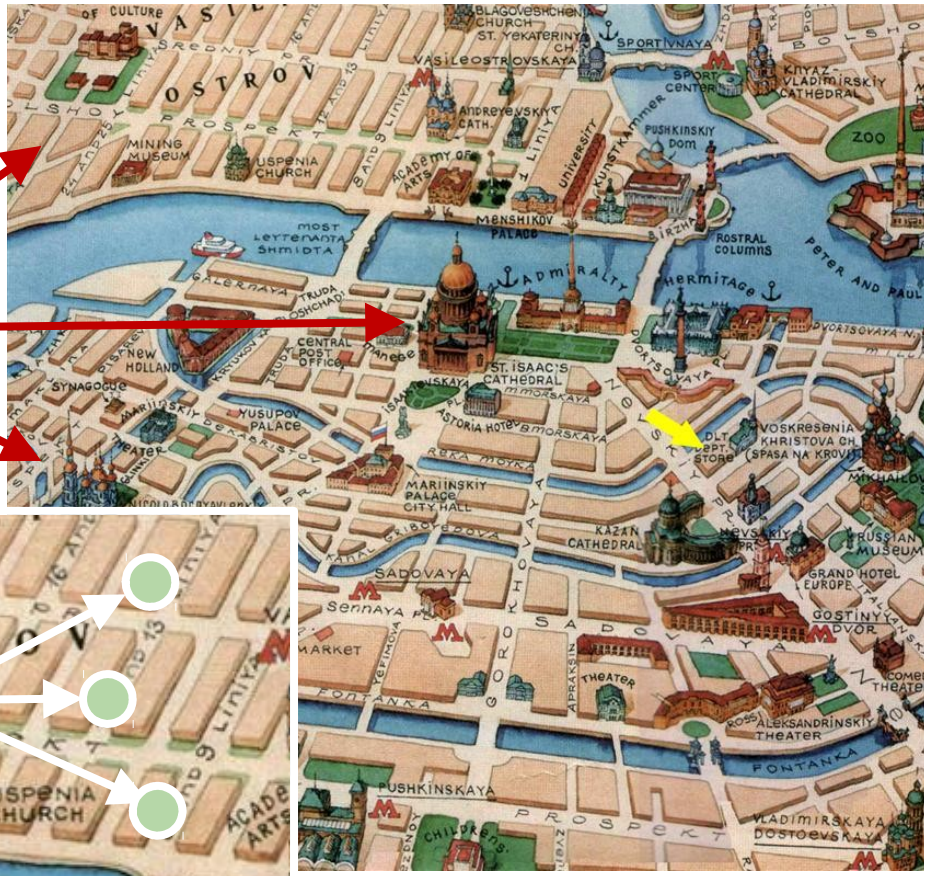


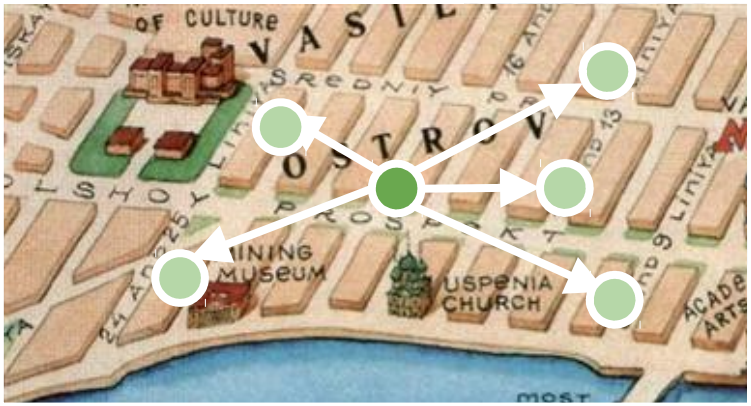
St. Petersburg

Признаки с особенной структурой

Пространственная структура

Доп. данные





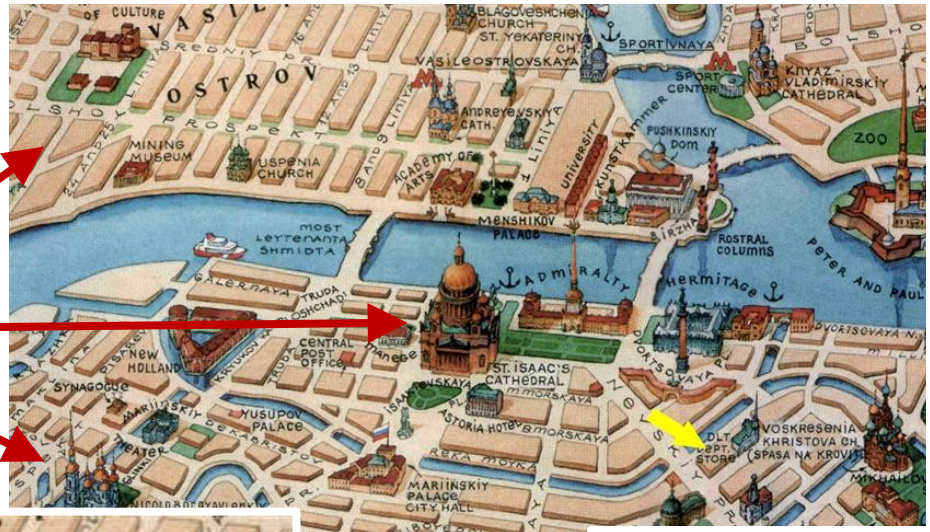
Другие объекты из выборки и центры кластеров

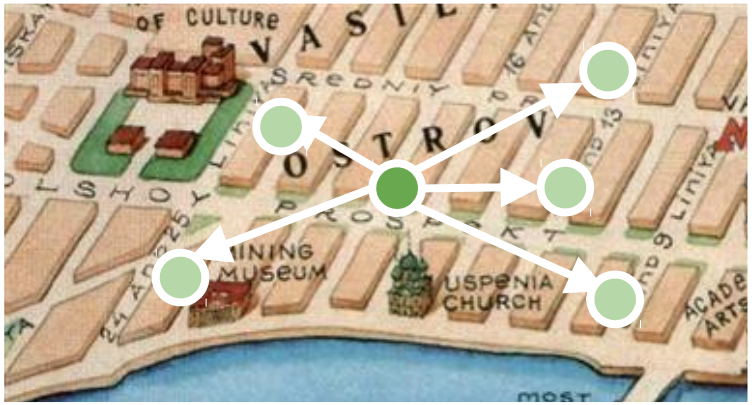
St. Petersburg

Признаки с особенной структурой

Пространственная структура

Доп. данные



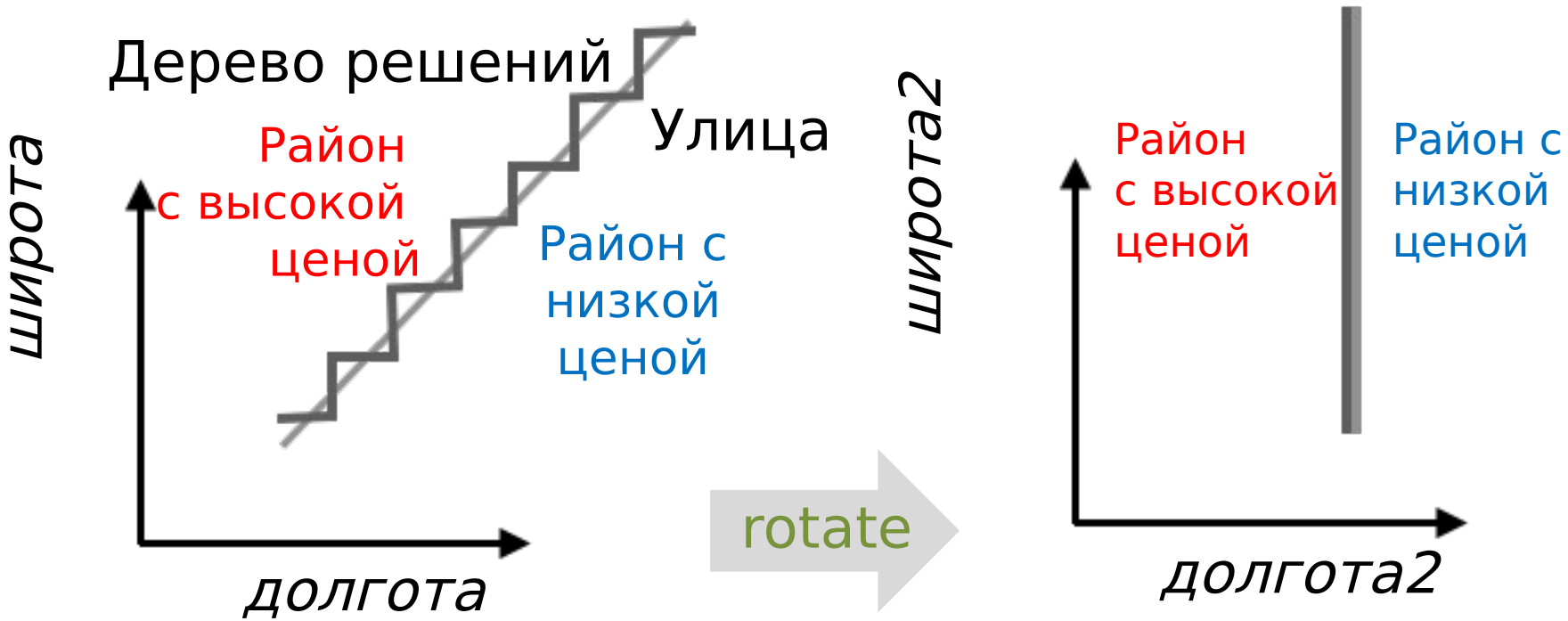


Другие объекты из выборки и центры кластеров

Агрегаты и статистики

Признаки с особенной структурой

Пространственная структура



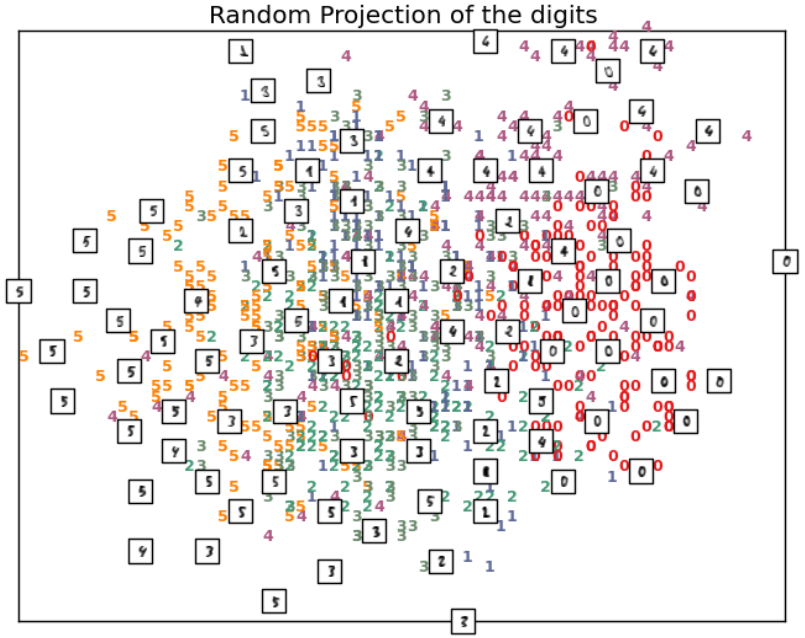
- Повороты координатных осей могут увеличить качество деревьев решений
- Возможно, например,применить повороты на углы, кратные 45° / 22.5°

Признаки с особенной структурой

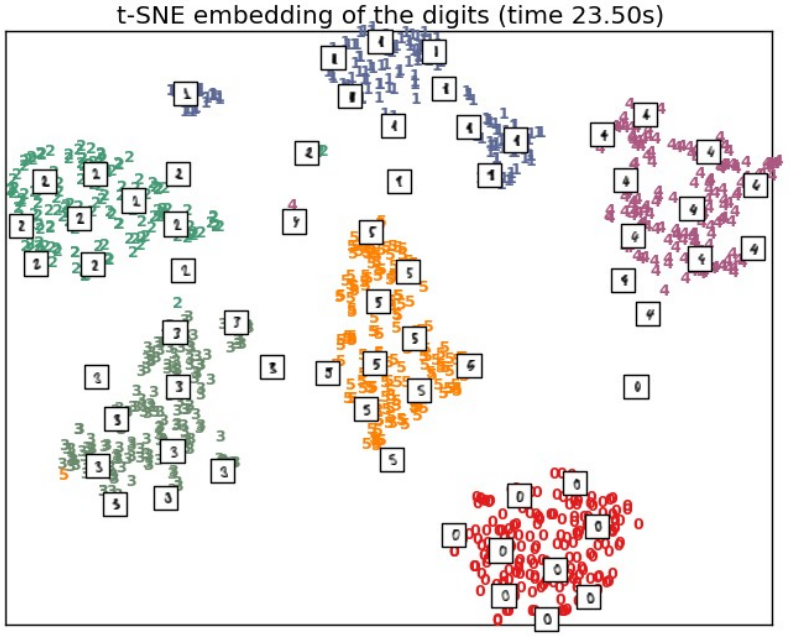
Пространственная структура

Проекции рукописных цифр (MNIST)

Случайная:



t-SNE:



- В целом, могут быть полезны методы снижения размерности признакового пространства путем проекции выборки на линейные (PCA, ...) и нелинейные подпространства (Isomap, t-SNE, ...)

Признаки с особенной структурой

Выводы

1. Время
    a. Периодичность
    b. Прошло/осталось с некоторого момента времени
    c. Разница между моментами времени
2. Координаты
    a. Интересные объекты из train/test или дополнительных данных
    b. Центры кластеров
    c. Агрегаты и статистики

Признаки с особенной структурой

Ссылки

Прогнозирование тренда доходностей акций
https://www.kaggle.com/c/two-sigma-financial-news

Прогнозирование цен на недвижимость
https://www.kaggle.com/c/zillow-prize-1

Предсказание цен поездок на такси в Нью-Йорке
https://www.kaggle.com/c/new-york-city-taxi-fare-prediction

5. Группы признаков

Группы признаков
Примеры

- Статистики баскетбольного матча для обеих команд (победившая/проигравшая)
  - Количество ассистов
  - Количество подборов
  - Количество трехочковых бросков
  - Количество штрафных бросков

- Признаки текстовых блоков в документе
  - Содержимое блока
  - Тип переменной с содержимым блока
  - Пространственные признаки блока (позиция/размер)

Группы признаков

Примеры

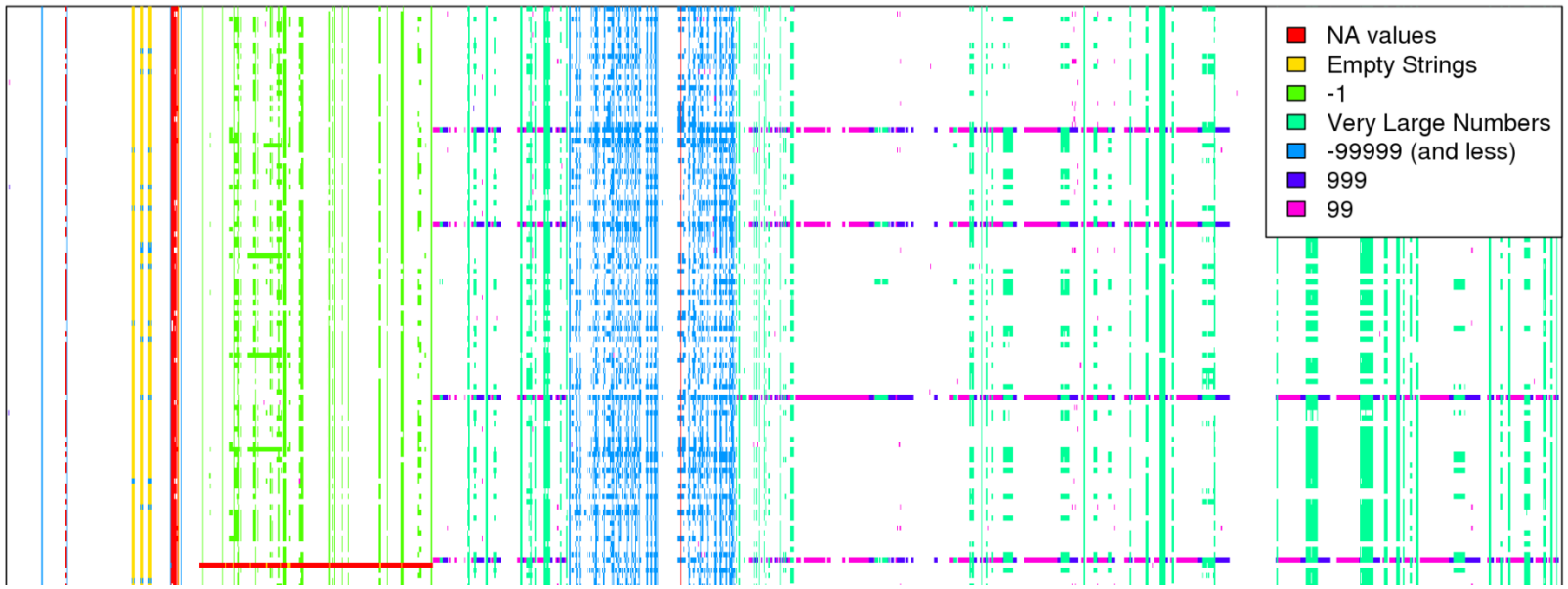
1. Внутри групп
   - Статистики (среднее, квантили)

2. Между группами
   - Суммы/разности/отношения между попарными элементами групп

7. Пропущенные значения признаков

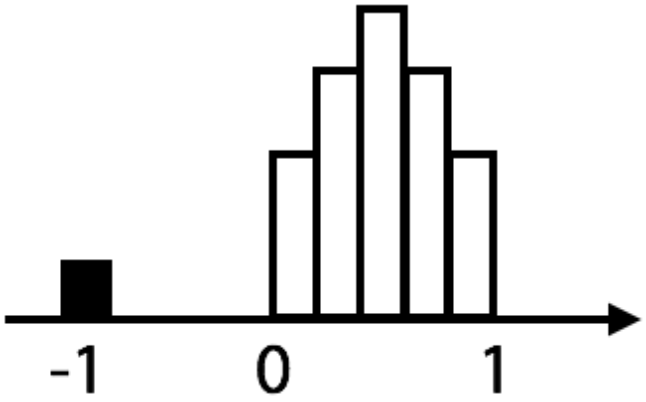
Пропущенные значения признаков

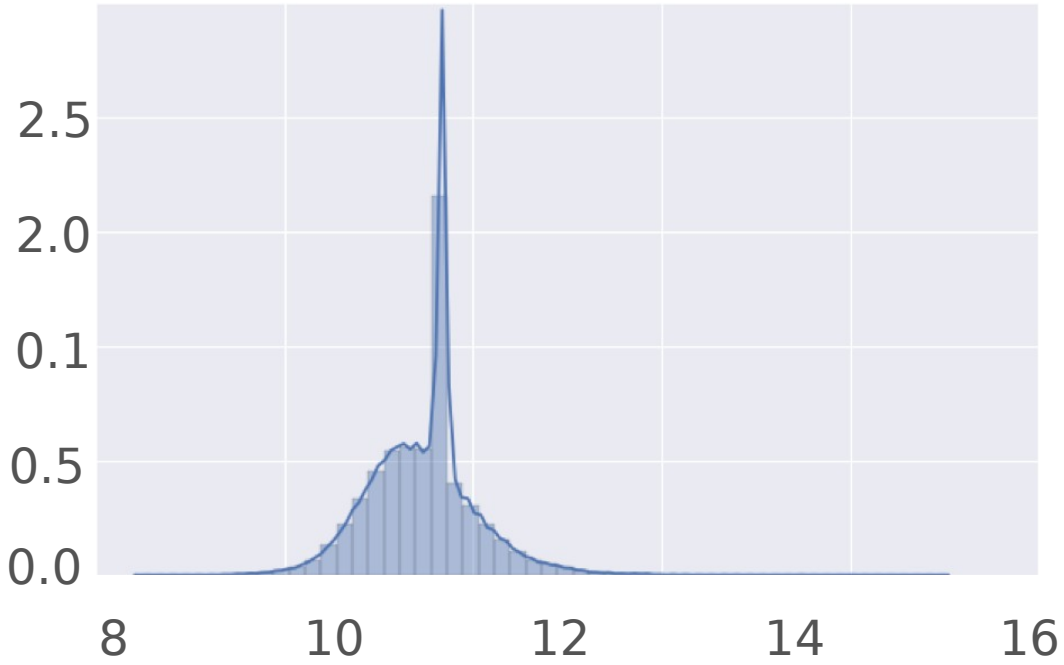
Как выглядят «пропуски» в данных



Пропущенные значения признаков

Скрытые NaN-ы





Организаторы могли заменить пропущенные значения, например, на -1 (слева) или на среднее (справа)

Пропущенные значения признаков

Заполнение пропусков

1. -999, -1, etc
2. Среднее, медиана
3. Реконструированные значения

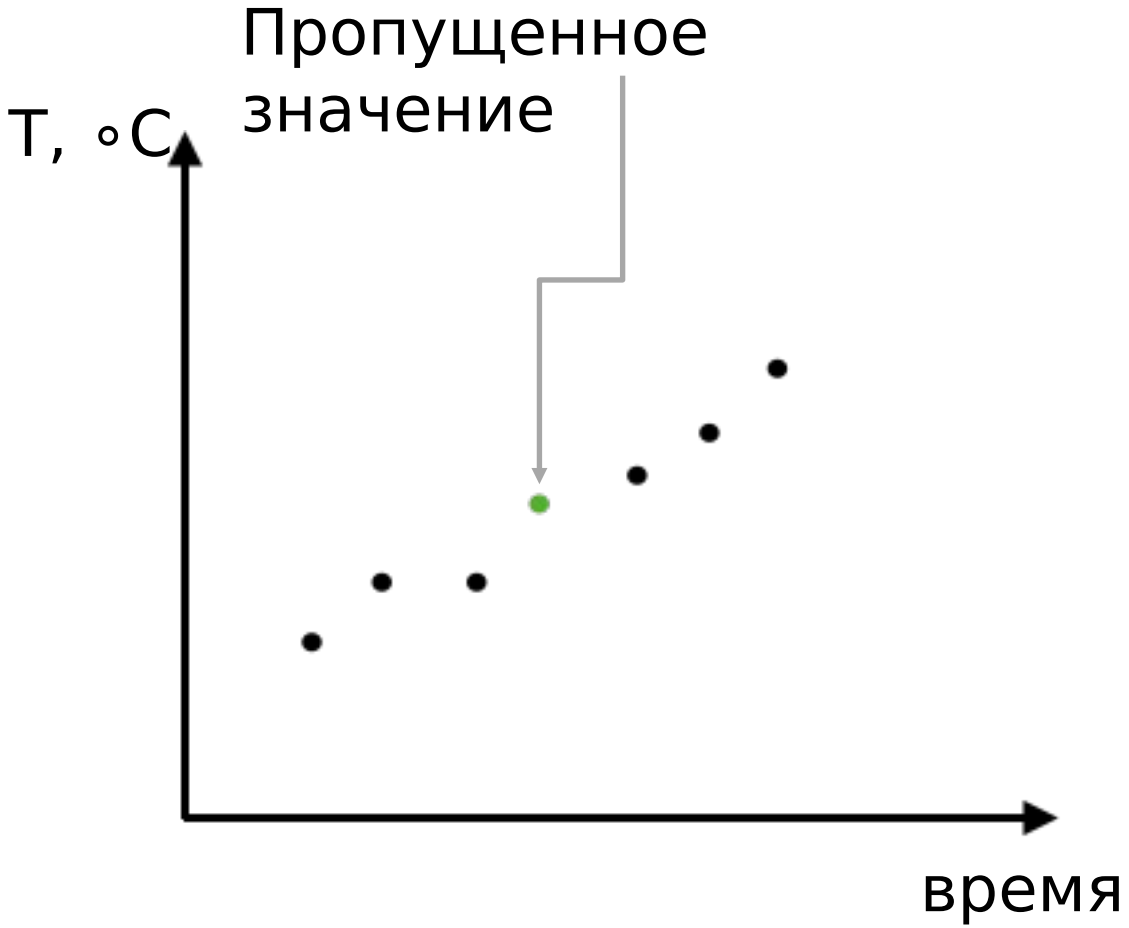
Пропущенные значения признаков

dummy-признак *isnull*

feature	isnull
0.1	False
0.95	False
NaN	True
-3	False
NaN	True

Пропущенные значения признаков

Реконструкция пропущенных значений



Пропущенные значения

Генерация признаков с учетом пропущенных значений

categorical_feature	numeric_feature	numeric_feature_filled	categorical_encoded
A	1	1	1.5
A	4	4	1.5
A	2	2	1.5
A	-1	-1	1.5
B	9	9	-495
B	NaN	-999	-495

Пропущенные значения

Категории, отсутствующие в обучающей выборке

Train:

categorical_feature	target
A	0
A	1
A	1
A	1
B	0
B	0
D	1

Test:

categorical_feature	target
A	?
A	?
B	?
C	?

Пропущенные значения

Категории, отсутствующие в обучающей выборке

Train:

categorical_feature	categorical_encoded	target
A	6	0
A	6	1
A	6	1
A	6	1
B	3	0
B	3	0
D	1	1

Test:

categorical_feature	categorical_encoded	target
A	6	?
A	6	?
B	3	?
C	1	?

Пропущенные значения

Выводы

1. Выбор метода заполнения пропусков зависит от ситуации
2. Обычно пропуски заполняются -999, средним или медианой
3. Пропущенные значения могут быть заранее заполненными организаторами
4. Бинарный признак «isnull» может быть полезным
5. Следует избегать заполнения «NaN»-ов до генерации признаков
6. Xgboost может учитывать NaN

Дополнительно

Категориальные признаки

Хэширование (n_features = 2)

feature
a
b
c
b

feature == a or feature == c	feature == b
1	
	1
1	
	1

`sklearn.feature_extraction.FeatureHasher`

Категориальные признаки

Используйте разделения по группам для добавления статистики

Country	Revenue
2	2000
2	20000
1	10000
3	12000
2	2000

Country	Revenue	Average Revenue by Country
2	2000	8000
2	20000	8000
1	10000	10000
3	12000	12000
2	2000	8000

```
groupby('feature1')['feature2'].mean()
```

Агрегация значений

Train

Customer_id	target
1	1
2	0
3	1

Данные транзакций

Customer_id	datetime	amount
1	2016-09-01	4000
1	2016-09-02	7000
2	2016-09-01	2500

```
transactions.groupby('customer_id').amount.sum()
```