

Работа с выбросами и пропущенными значениями

Кантонистова Елена

elena.kantonistova@yandex.ru

27 октября 2017

1 Виды признаков

2 Работа с пропущенными значениями

- Типы пропусков
- Простые методы работы с пропусками
- Заполнение пропусков методом ближайших соседей
- Модель для предсказания пропусков

3 Работа с выбросами

- Поиск выбросов
- Методы машинного обучения и выбросы
- Поиск выбросов: методы sklearn

4 Масштабирование признаков

Признаки бывают трех видов:

- Числовые (дискретные: оценка за экзамен и непрерывные: вес)
- Порядковые (номер дома)
- Категориальные (адрес)

Виды признаков

ФРУКТЫ ОВОЩИ

СВЕЖИЕ ВИТАМИНЫ КАЖДЫЙ ДЕНЬ!

Числовые признаки

Порядковые признаки

Категориальные признаки

Номер	Наименование	Единица измерения	Цена
1	Яблоки Голден	1 кг	79 ⁹⁰
2	Груши Китайские	1 кг	79 ⁹⁰
3	Яблоки Красные	1 кг	89 ⁹⁰
4	Груши осенние: Конфетница	1 кг	109
5	Помело	1 кг	79 ⁹⁰
6	Мандарины	1 кг	89 ⁹⁰
7	Лимоны	1 кг	99 ⁹⁰
8	Слива круглая	1 кг	129
9	Огурец длинный	1 шт.	24 ⁹⁰
10	Лук, зеленый	100 г, 1 уп.	24 ⁹⁰
11	Томаты Черри, красные	200 г, 1 уп.	54 ⁹⁰
12	Сельдерей стебель	1 уп.	59 ⁹⁰
13	Виноград Киш Мели	1 кг	99 ⁹⁰
14	Морковь мясистая	1 уп.	24 ⁹⁰
15	Лук репчатый, красный	300 г, 1 уп.	24 ⁹⁰
16	Дыня сезонная	1 кг	32 ⁹⁰
17	Картофель мытый	2,5 кг, 1 уп.	59 ⁹⁰

- Пропуск появился случайно
- Пропуск можно объяснить, исходя из смысла переменных и задачи
- Вероятность появления пропуска зависит от наблюдаемых переменных

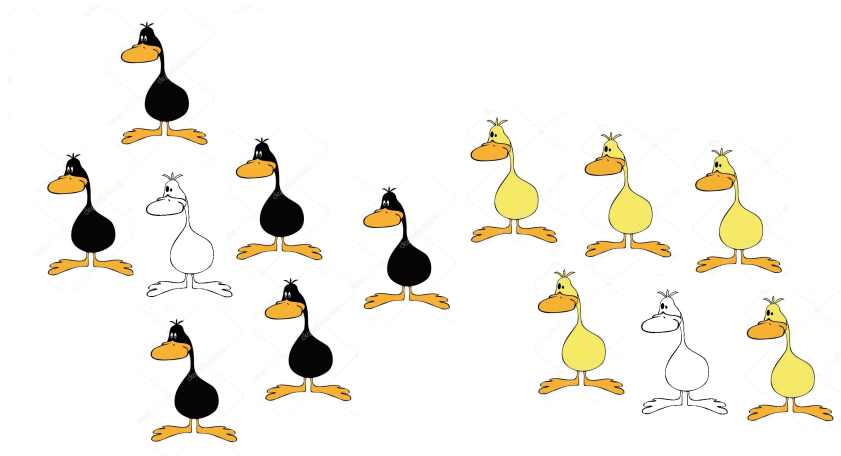
Простые методы работы с пропусками

- Удалить все строки в данных, содержащие пропуски
- Заменить пропущенные значения самым частотным значением
- Заменить пропущенные значения средним или медианой
- Заменить пропущенные значения нулем
- Заменить пропущенные значения некоторым уникальным значением (например, -999999)

Заполнение пропусков методом ближайших соседей

Идея: посмотреть на ближайших соседей и взять среднее

- Для каждого объекта находим k ближайших соседей без пропусков
- Усредняем полученные значения по соседям



Модель для предсказания пропусков

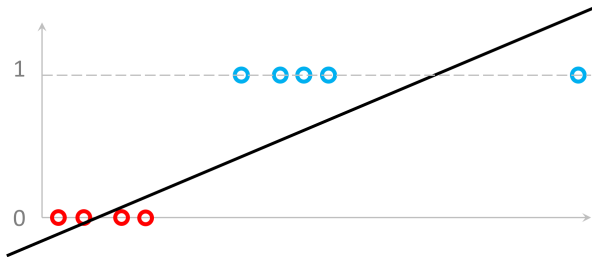
Идея: возьмем строки данных, не содержащие пропусков, и обучимся на них с целью предсказать пропуски. То есть столбцы с пропусками - это целевые переменные.

Поиск выбросов: интуитивный и работающий подход

- визуализация данных (например, объекты на карте)
- построение распределений исходных данных
- поиск редких / ошибочных значений

Методы, чувствительные к выбросам:

- Линейная регрессия



- Все остальные методы (менее чувствительны к выбросам, но выбросы так или иначе влияют на них)

- One-class SVM
- Robust covariance
- Isolation forest
- Local outlier factor

Почти всегда стоит масштабировать данные. Распространенные варианты масштабирования:

- MinMaxScaler:

$$x \rightarrow \frac{x - \min(x)}{\max(x) - \min(x)}$$

- StandardScaler:

$$x \rightarrow \frac{x - \text{mean}(x)}{\text{std}(x)}$$