Работа с выбросами и пропущенными значениями

Кантонистова Елена

elena.kantonistova@yandex.ru

27 октября 2017

План лекции

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

Виды признаков

Признаки бывают трех видов:

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

Виды признаков

Числовые признаки

Порядковые признаки

Категориальные признаки

Типы пропусков

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

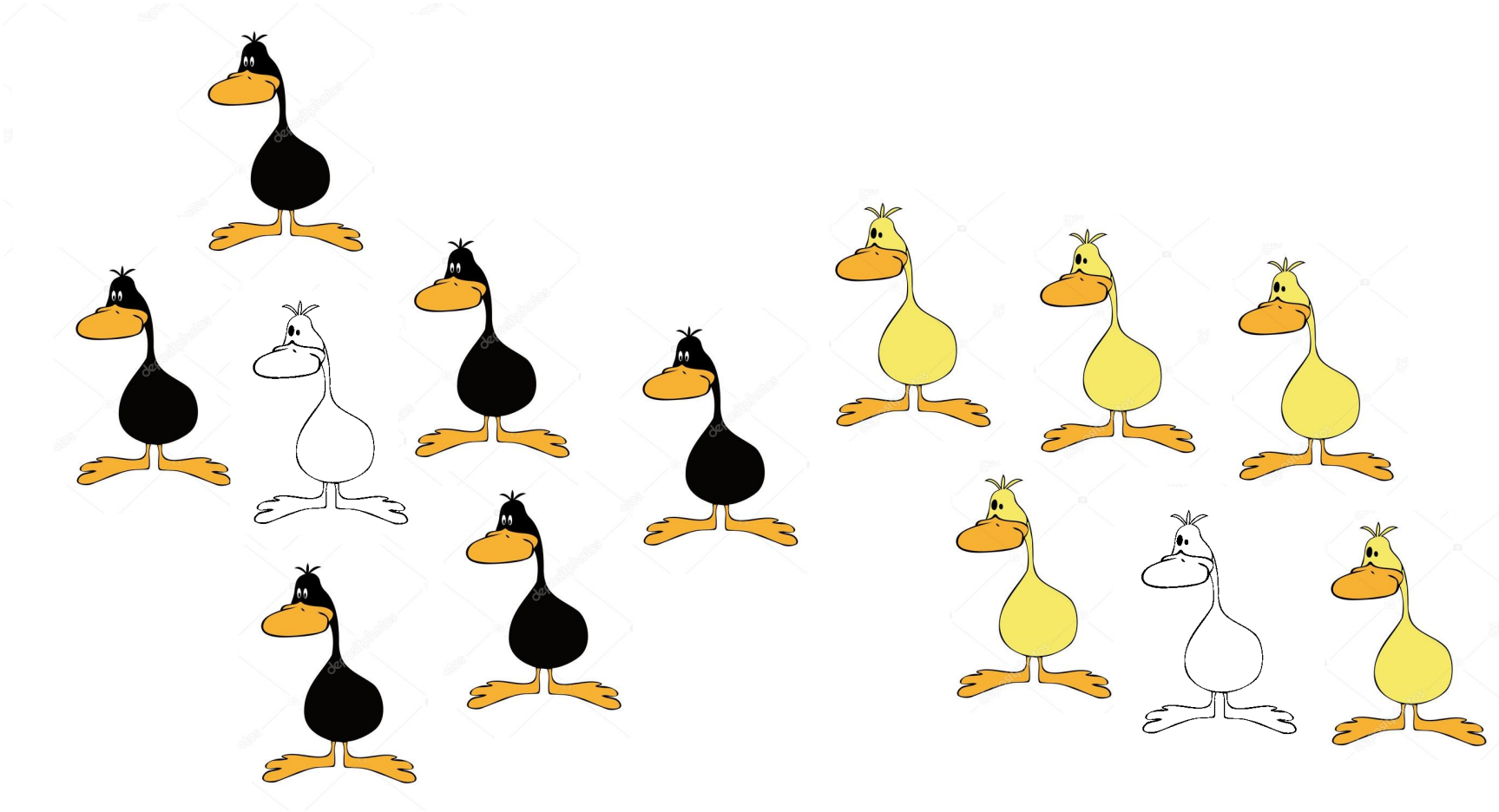
Простые методы работы с пропусками

- Удалить все строки в данных, содержащие пропуски
- Заменить пропущенные значения:
  a) самым частотным значением
  b) средним или медианой
  c) нулем
  d) 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

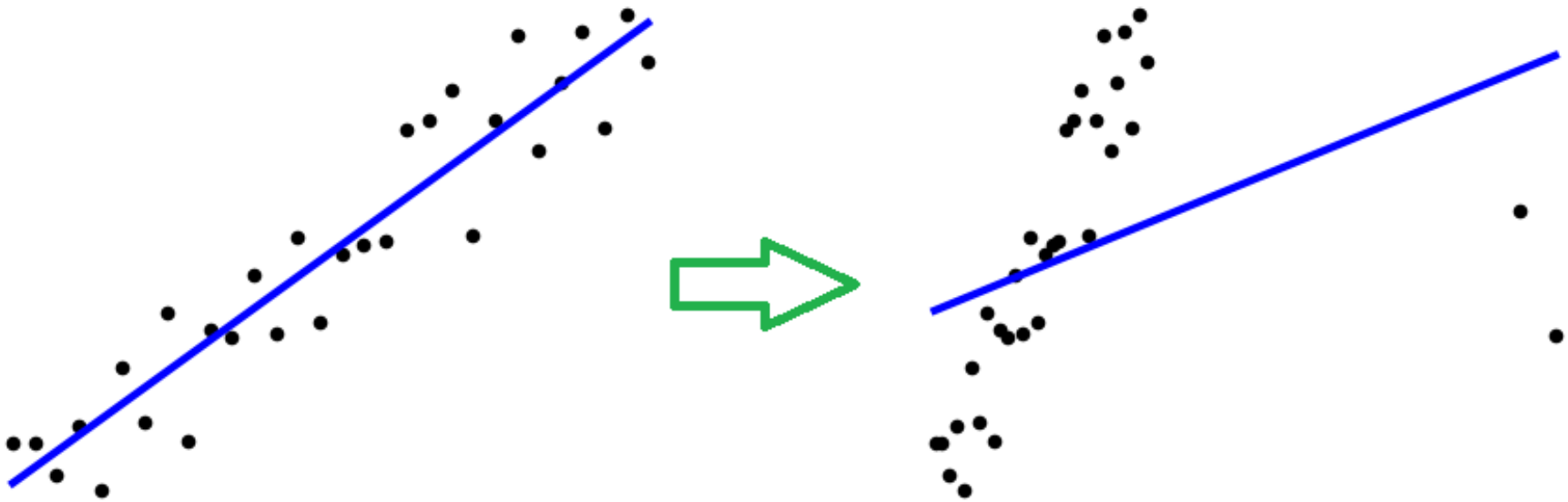
Идея: возьмем строки данных, не содержащие пропусков, и обучимся на них с целью предсказать пропуски. То есть столбцы с пропусками - это целевые переменные.

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 и другие методы, оптимизирующие ошибку R^2



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

Поиск выбросов: методы sklearn

- Robust covariance: этот метод используется в предположении, что данные имеют нормальное распределение. Основан на аппроксимации данных эллипсоидом нормального распределения.
- One-class SVM: этот метод предпочтительнее использовать в случае, если данные не распределены нормально (например, если в данных есть два хорошо разделенных кластера). Данный метод - это специальный случай применения метода опорных векторов.
- Isolation forest: метод основан на концепции random forests, поэтому он хорошо работает на данных с большим количеством признаков.
- и другие (см. sklearn)

Масштабирование признаков

Почти всегда стоит масштабировать данные. Распространенные варианты масштабирования:

- MinMaxScaler:

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

- StandardScaler:

$$x \to \frac{x - mean(x)}{std(x)}$$