Подготовка данных

Дьяконов А.Г.

Московский государственный университет имени М.В. Ломоносова (Москва, Россия)

Данные

More data beats clever algorithms,
but better data beats more data
(P. Norvig)

На что смотреть:

- **размеры, размерность, число элементарных порций (объектов), разреженность, разрешение**

- **семантика данных**

- **структура данных, режим доступа к данным (online / offline), способ доступа**

Виды данных

- признаковые описания (матрица объект-признак)
- измерения
  - одномерные сигналы (ряды, звук и т.п.), последовательности, тексты
  - изображения
  - видео
- метрические данные
- данные в специальных форматах
  - графы
  - XML-файлы
  - пространственно-временные
  - сырые логи
  - и т.п.

Признаки (Features)

Признак – функция на множестве объектов

$$f : X \to A$$

Типы признаков

- вещественными (+ временнные)
  - интервальные (Interval)
  - относительные (Ration)
- категориальными
  - неупорядоченные категориальные (номинальные – Nominal, факторные)
  - порядковые (Ordinal) или упорядоченные категориальные
- текстовыми

+ Дискретные

Типы признаков

Тип признака	Операции	Трансформации	Примеры
номинальные	== перестановки mode, entropy, contingency, correlation, X2-test	перестановка	ID, пол, цвет, профессия
порядковые	> median, percentiles, rank correlation, run tests, sign tests	Монотонное преобразование	оценка, рейтинг, место в соревновании
интервальные	+, – mean, standard deviation, Pearson's correlation, t and F tests	A*X + B	дата, температура по Цельсию
относительные	*, / geometric mean, harmonic mean, percent variation	A*X	возраст, масса, длина, цена, температура по Кельвину

Контекстный признак

смысл явно прописан в постановке задачи

или понятен из контекста

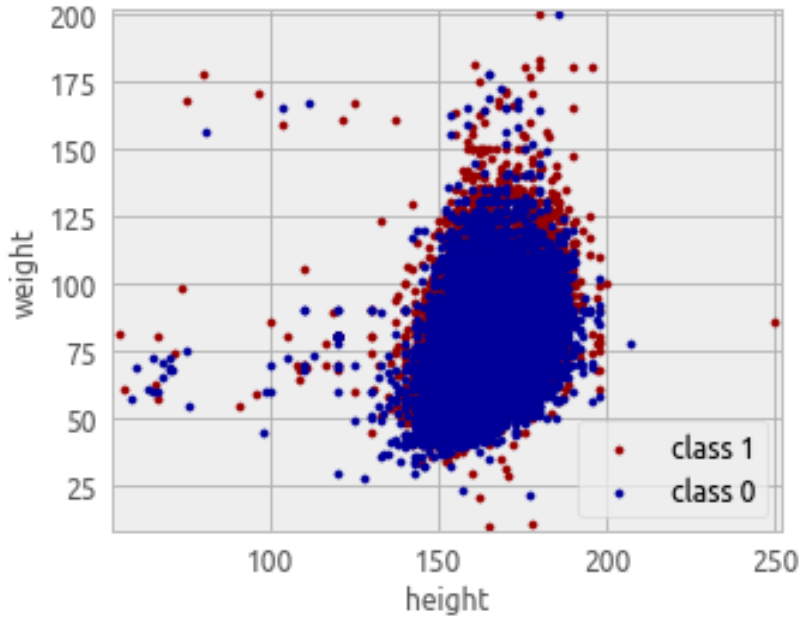
Предполагаем

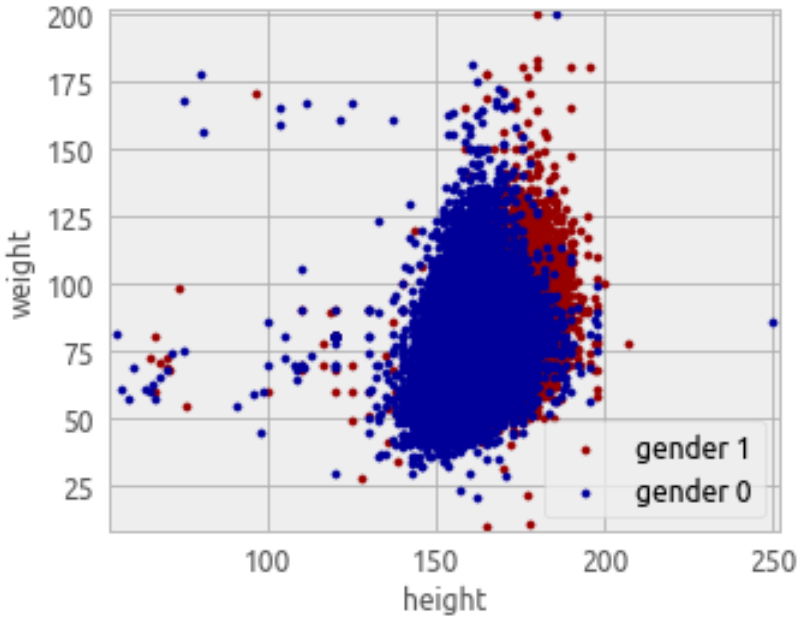
- **область значений**
- **примерное распределение значений в этой области**

Признак	Гипотеза
Число кликов	Максимальна в рабочие дни, в дневные часы
Уровень дохода	Унимодальное распределение, значения положительные
Температура	Лежит на отрезке [36, 42]

Контекстный признак

ap_hi	ap_lo	ap_hi_new	ap_lo_new
150	1100	150	110
11	70	110	70
12	80	120	80
11	570	115	70
1	2080	120	80





Dummy-признаки и утечки в данных (Leakages)

Dummy-признаки – признаки, которые могут не входить в явном виде в признаковую матрицу, но их значения определяются из способа организации данных.

- **номер строки** (а также производные признаки, например, чётность номера строки)

- **номер объекта в какой-то внутренней нумерации** (например, id объекта), **производные признаки от этого номера**

- **константный признак**

- **номер порции данных (если изначально датасет разбит на несколько частей)**

В идеале должны быть бесполезны!

Dummy-признаки, которые могут быть созданы

- имена и характеристики записей (в задачах, где объекты хранятся отдельно, например как файлы изображений в задаче классификации изображений)

- характеристический признак, есть ли в строке какие-то особенности (например, пропуски, аномальные значения и т.п.)

На практике могут быть полезны

Вес 60, 50, 40 кг (наверняка неточный)
Вес 62.5, 63, 67.2 (наверное, точнее)

Зачем нужны Dummy-признаки

для правильной организации работы

Пример: области значений целевого (маленькие, средние, большие) + StratifiedKFold

Утечка в данных

информация, которая повышает качество решения задачи машинного обучения, но теряет эти свойства при тестировании на независимом и правильно организованном контроле

Как правило

1. Зависимость целевого признака от dummy-признаков

2. Содержание ответа в исходных данных

(пример про номер страницы и число страниц в сессии)

Свойства данных

Свойства данных	Что мешает этому свойству	Причины нарушения свойства	Средство борьбы
Корректность (точность)	Выбросы, аномалии Шумовые значения	Погрешность приборов, ошибки при заполнении	Очистка данных (Data Cleaning)
Полнота	Пропуски Разреженность	Недоступность данных, ошибки при заполнении, сбои при записи	Очистка данных (Data Cleaning)
Непротиворечивость (согласованность)		Различные источники данных	Data Integration
Безызбыточность	Дубликаты Шумовые признаки Излишняя дискретизация		Data Reduction Data Transformation
Ясность			Data Transformation
Доступность			
Актуальность			

Предобработка данных (Data Preprocessing / Preparation)

– замена, модификация или удаление частей набора данных с целью повышения непротиворечивости, полноты и корректности набора данных, а также уменьшения избыточности.

На полном наборе данных (и на контрольных объектах тоже).

РАЗДЕЛЫ Предобработки данных

Очистка данных (Data Cleaning)

- Обнаружение (и удаление / замена) аномалий / выбросов 2do
- Обнаружение (и удаление / замена) пропусков (Missing Data Imputation) next
- Обнаружение (и удаление / замена) шумов (Noise Identification)
- Обнаружение (и удаление / исправление) некорректных значений (correct bad data / filter incorrect data) next

Трансформация данных (Data Transformation)

- Переименование признаков, объектов, значений признаков
- Кодирование значений категориальных переменных 2do
- Дискретизация (Discretization / Binning) next
- Нормализация (Normalization) next
- Сглаживание (Smoothing)
- Создание признаков (Feature creation) 2do
- Агрегирование (Aggregation) next

Интеграция данных (Data Integration)

- Перевод в нужный формат (в том числе, объединение таблиц…)
  next

Сокращение данных (Data Reduction) next

- Сэмплирование (Sampling)
- Сокращение размерности (Dimensionality reduction)
- Отбор признаков (Feature subset selection)
- Отбор объектов (Instance Selection)
  - удаление дубликатов

Пропуски (NA, NaN, Impute missing variables)

- **оставляем** (но не все модели могут работать с пропусками)
- **удаляем описания объектов с пропусками** (радикальная мера, которая редко используется)
- **заменяем на фиксированное** значение (например, если признак бинарный, то на 0.5)
- **заменяем на легковычислимое** значение (среднее, медиана, мода)
- **восстановление значение** (построение специальной модели для восстановления)
- **экспертная замена** (см. ниже)
- **+ добавление характеристического признака пропусков**

Пропуски (NA, NaN, Impute missing variables)

Важно понимать природу пропуска:

- **значение может не быть доступно**

клиент банка не указал в анкете свой возраст

- **значение может не существовать**

«Доход» для детей моложе 18 (=0)

- **значение не является числом**

0/0 = NaN

средняя покупка в категории товаров

Обучение и тест – одинаковые распределения.
Тоже самое для пропусков!

Корректировка значений

время	давление	температура
'23:10'	120/80	36.6C
'10 часов'	120/70	37.1C
'7:40'	110/70	37C

время	в. давл	н. давл	темп
23:10.00	120	80	36.6
10:00.00	120	70	37.1
07:40.00	110	70	37

Агрегация (Aggregation)

PAY1	PAY2	PAY3			
100	100	100			
100	100	90			
70	60	60			
70	50	50			
10	0	0			

Составляющие суммы, замеры разными датчиками и т.п.

Интеграция данных (Data Integration)

ID	BKI	DATE	SUM	CITY	<30
101	1	03/12/16	20000	Москва	1
101	2	03/12/16	20000	Г. Москва	NA
101	3	03/12/16	20000	Москва	1
101	1	01/10/14	15000	Москва	0
101	2	01/10/14	15000	Г. Москва	0

Нормировки (Data Normalization)

Для большинства алгоритмов машинного обучения необходимо, чтобы все признаки были вещественными и «в одной шкале».

- **Стандартизация (Z-score Normalization)**
- **Нормировка на отрезок (Min-Max Normalization)**
- **Нормировка по максимуму**
- **Decimal Scaling Normalization**

$$\mathrm{N}_{\mathrm{ds}}(x) = \frac{x}{10^{\min\{i : 10^i > x\}}}$$

- **Ранговая нормировка (tiedrank, rankdata)**

Трансформация

Box-Cox Transformation положительного признака

$$y = \begin{cases} x^{\lambda-1}/\lambda, & \lambda \neq 0 \\ log(x), & \lambda = 0 \end{cases}$$

Как правило применяют, чтобы распределение признака стало похожим на нормальное

Дискретизация (биннинг, Binning)

переход от вещественного признака к порядковому за счёт кодирования интервалов одним значением.

доход от 0 до 10000, от 10000 до 25000, от 25000 до 50000 и т.д.

Способы:

Equal-width (distance) partitioning

Делим область значения признаков на области-интервалы равной длины.

Equal-depth (frequency) partitioning

Делим область значения признаков на области-интервалы: в каждую попало одинаковое число точек.

Экспертно

Сокращение данных (Data Reduction)

– уменьшение объёма исходных данных, сохраняя полезную информацию

- **отбор признаков (Feature Selection)**

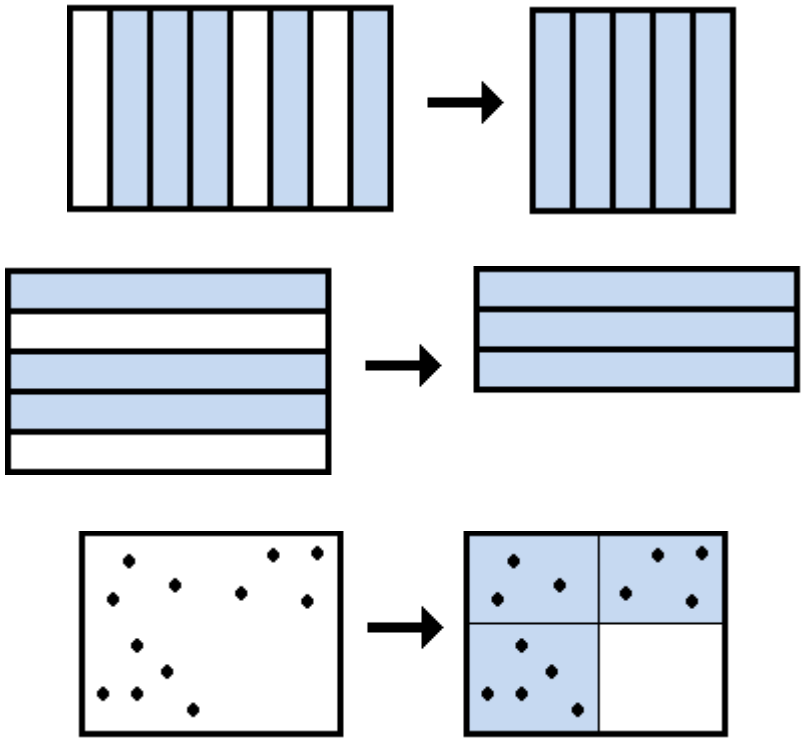
  отдельная тема **next**

- **отбор объектов (Instance Selection)**

  редко используется,
  как правило, по анализу или экспертами

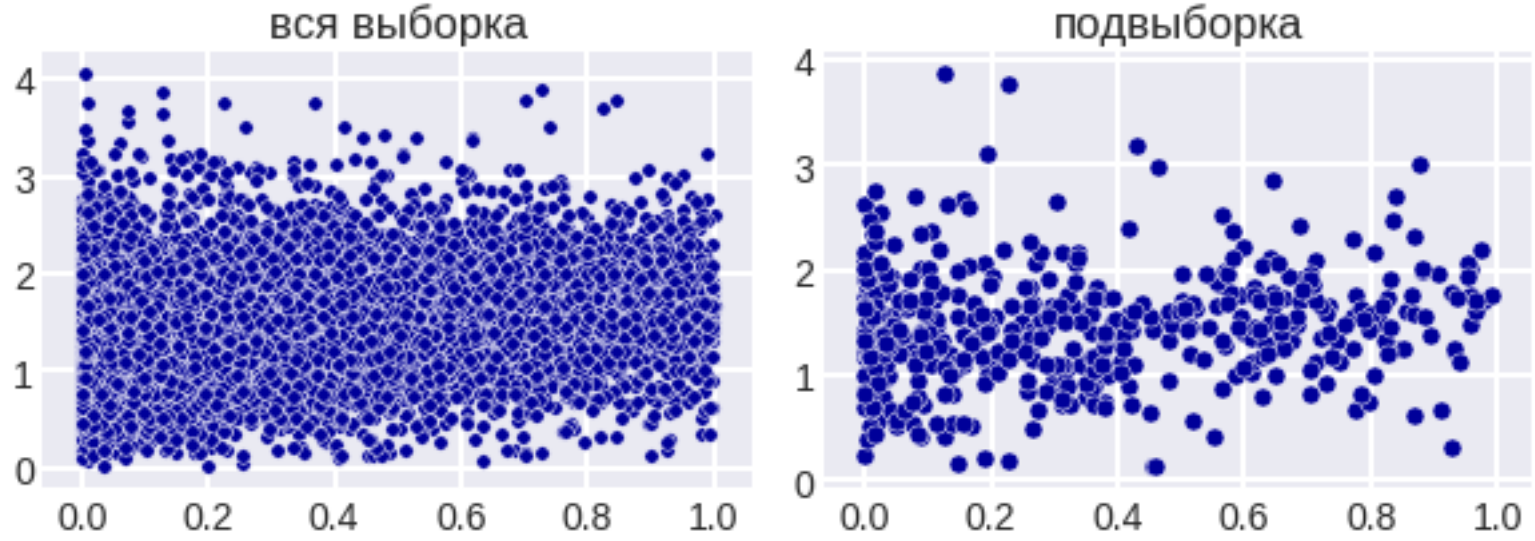
- **дискретизацию, огрубление информации (Discretization)**

  увеличение шага дискретизации
  перевод вещественных признаков в дискретные



Сокращение данных (Data Reduction)

- **сэмплированое (Sampling)**



- **сокращение размерности (Dimensionality reduction)**
  - **факторный анализ (factor analysis)**
  - **метод главных компонент (PCA), SVD**
  - **нелинейные модели: LLE, ISOMAP**
  - **многомерное шкалирование (MDS)**

Сокращение данных (Data Reduction)

цели

- удаление лишних (нерелавантых) данных
- повышение качества решения задачи
- уменьшение стоимости данных
- увеличение скорости последующего анализа

(в частности, настройки моделей)

- повышение интерпретируемости моделей

Сэмплирование

- **Без возвратов** (Simple random sampling without replacement)
- **С возвратами** (Simple random sampling with replacement)
- **Балансированное** (Balanced sampling) – сэмплирование при котором подвыборка будет удовлетворять некоторому заранее заданному условию (например, 90% описаний будет соответствовать пациентам старше 60 лет)
- **Кластерное** (Cluster sampling) – предварительно данные разбиваются на кластеры и выбирается поднабор кластеров.
- **Стратифицированное** (Stratified sampling) – предварительно данные разбиваются на кластеры, в каждом кластере отдельно осуществляется сэмплирование, таким образом в подвыборку попадают представители всех кластеров.

Для более быстрого поиска оптимальных параметров.
Составляющая часть алгоритма (RF)
Для получения выборки, обладающей специальными свойствами.