Разбор задач анализа данных

Надежда Чиркова
nchirkova@hse.ru, @nadiinchi (Telegram)

Что такое машинное обучение?

Машинное обучение — «обучение с помощью машины»?
Machine learning — «обучение машины»

Определение с сайта machinelearning.ru:
Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Зачем нужно машинное обучение?

- Заменить человека при решении задач (автоматизация);
- Поиск закономерностей в данных, которые человек не находит.

Постановка задачи машинного обучения

Задача машинного обучения:

Постановка задачи машинного обучения

Задача машинного обучения:

- данные (что такое объект, какие признаки + типы признаков);
- что предсказывать;
- оценивание качества (критерий качества + способ валидации).

Матрица объекты–признаки

Числовая матрица:

	Признак 1	Признак 2	…	Признак K
Объект 1				
Объект 2				
Объект 3				
…				
Объект N				

Виды признаков

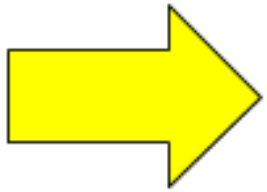
Объект — вектор в [конечномерном] пространстве признаков.
Виды признаков:

1. вещественные;
2. бинарные;
3. категориальные;
4. порядковые (упорядоченные категориальные);
5. подмножество супермножества;
6. строковые.

One-hot кодирование

Бинарное кодирование категориальных признаков:

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

«Мешок слов» для текстов

Бинарное кодирование текстов:

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Что предсказываем?

Два типа обучения:

- Обучение с учителем (пытаемся понять, как зависят ответы, известные на объектах обучающей выборки, от входных данных):
  - Классификация (бинарная, multiclass, multilabel)
  - Регрессия
  - Прогнозирование временных рядов
  - Рекомендации
  - ...
- Обучение без учителя (как можем формализуем, что хотим найти в данных, и ищем).
  - Кластеризация
  - Понижение размерности
  - Визуализация
  - ...

Критерии качества

y_i — правильный ответ на i-м объекте
$a(x_i)$ — предсказанный ответ на i-м объекте
ℓ — число объектов в выборке

$$Q(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i))$$

Критерии качества для регрессии

$y_i \in \mathbb{R}$ — правильный ответ на i-м объекте
$a(x_i) \in \mathbb{R}$ — предсказанный ответ на i-м объекте
ℓ — число объектов в выборке

$$Q(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i))$$

Варианты $L(y_i, a(x_i))$:

- $L(y_i, a(x_i)) = (y_i - a(x_i))^2$ (mean squared error)
- $L(y_i, a(x_i)) = |y_i - a(x_i)|$ (mean absolute error)
- $L(y_i, a(x_i)) = \begin{cases} 1, |y_i - a(x_i)| < \varepsilon \\ 0, |y_i - a(x_i)| \geqslant \varepsilon \end{cases}$

Критерии качества для классификации с непересекающимися классами

$y_i \in \{c_1, \ldots, c_K\}$ — правильный ответ на i-м объекте
$a(x_i) \in \{c_1, \ldots, c_K\}$ — предсказанный ответ на i-м объекте
ℓ — число объектов в выборке

$$Q(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i))$$

Варианты $L(y_i, a(x_i))$:

- $L(y_i, a(x_i)) = \begin{cases} 1, y_i = a(x_i) \\ 0, y_i \neq a(x_i) \end{cases}$ (accuracy)

Критерии качества для классификации с пересекающимися классами

$y_i \in \{0,1\}^K$ — правильный ответ на i-м объекте

$a(x_i) \in \{0,1\}^K$ — предсказанный ответ на i-м объекте

ℓ — число объектов в выборке

$$Q(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i))$$

Варианты $L(y_i, a(x_i))$:

- $L(y_i, a(x_i))$ — среднее accuracy по классам

Критерий качества

- фантазия не ограничена :)
- определяется заказчиком исходя из цели решения задачи
- должен легко вычисляться по имеющимся данным в offline

Критерий качества

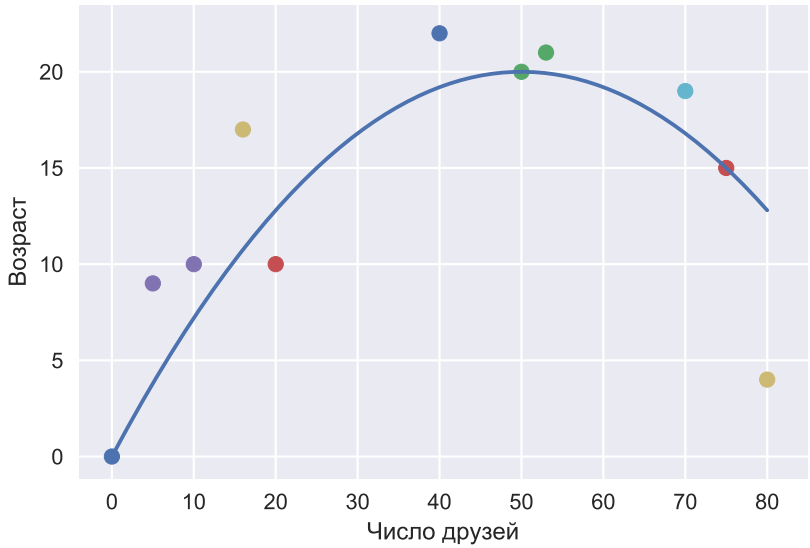
- фантазия не ограничена :)
- определяется заказчиком исходя из цели решения задачи
- должен легко вычисляться по имеющимся данным в offline

По какой выборке измеряется качество?

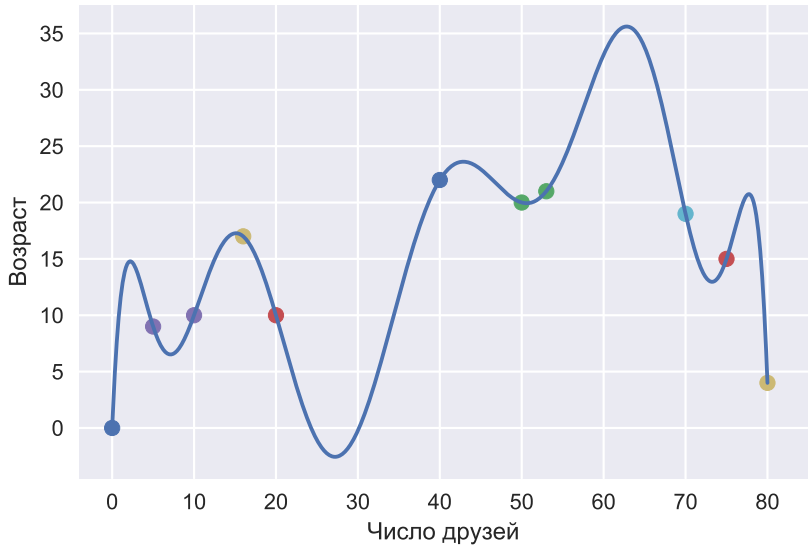
Качество на обучающей выборке

По оси абсцисс — признак, по оси ординат — целевая переменная.

Хорошая модель



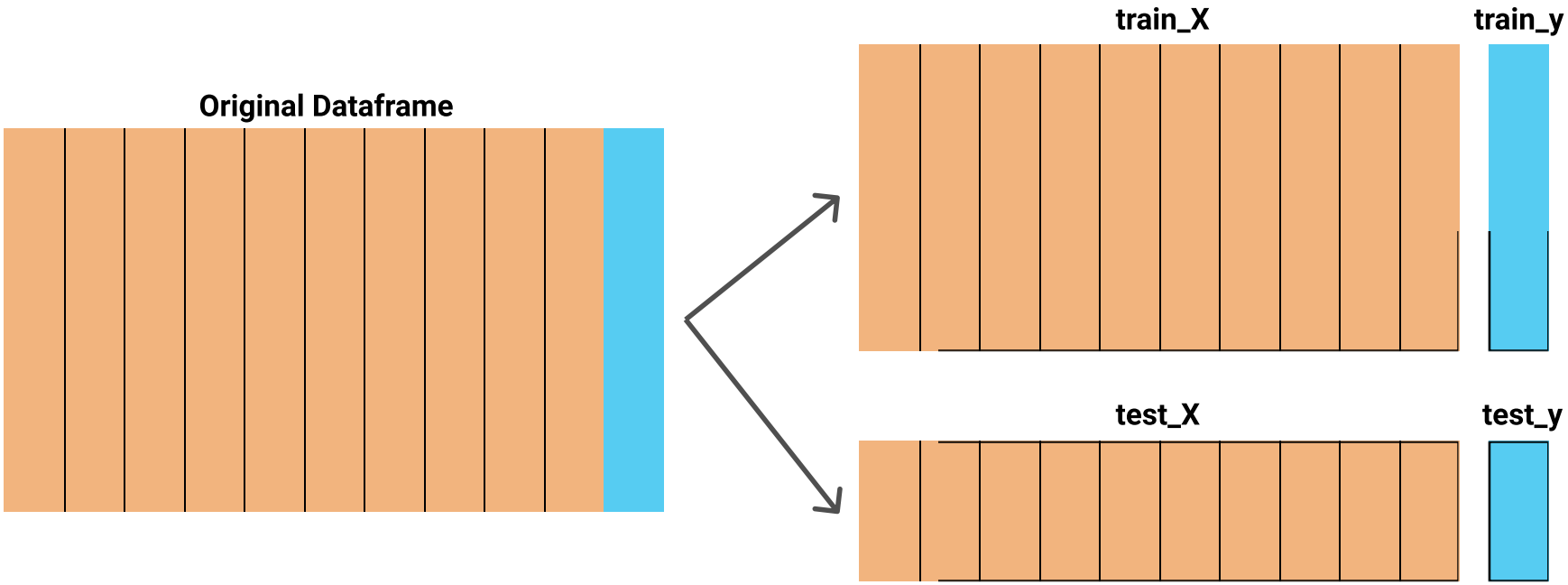
Переобучение



Качество нужно оценивать по отдельной выборке!

Разделение выборки

Качество нужно оценивать по отдельной выборке!



Кейсы

Что нужно сделать: формализовать пожелания заказчика в виде задачи машинного обучения:

- данные (что такое объект, какие признаки + виды признаков);
- что предсказывать (с типом задачи);
- оценивание качества (критерий качества + способ валидации);

Дополнительный вопрос: где взять данные?

Простой пример

Задача кредитного скоринга: вернет ли заемщик кредит
Постановка задачи?

Простой пример

Задача кредитного скоринга: вернет ли заемщик кредит

Постановка задачи?

- данные:
  - объект?
  - признаки (с видом признака)?

Простой пример

Задача кредитного скоринга: вернет ли заемщик кредит
Постановка задачи?

- данные:
  - объект?
  - признаки (с видом признака)?
- что предсказывать (с типом задачи)?
- критерий качества?
- метод валидации решения?
- где взять данные?

Self-driving cars

Задачи машинного обучения:

- распознавание знаков, сигналов светофора
- распознавание текстов на изображениях
- предсказание следующих действий пешеходов и других объектов
- планирование маневров
- прогнозирование времени прибытия в пункт назначения

Self-driving cars

Задачи программирования:

- поиск оптимального маршрута
- физические расчеты, моделирование движения

Выводы

- Формальная постановка задачи — важный процесс, перевод задачи с языка прикладной области на математический язык методов решения.
- Не всегда очевидно, что является объектом, где взять данные, какой критерий качества выбрать...
- Хорошо поставленную задачу проще решать :)