

Введение в Машинное обучение: примеры и основные методы

Зуева Надежда
ФИВТ МФТИ

March 2018

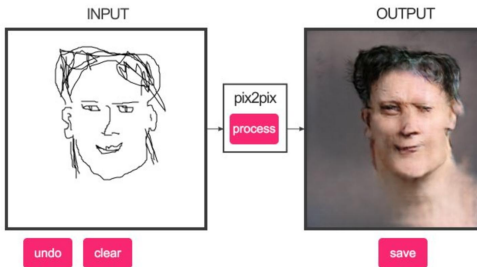
- ❶ Организационные моменты
- ❷ Напоминания
- ❸ Короткая историческая справка
- ❹ Закон Мура
- ❺ Терминология и мотивировка
 - ❶ Основные определения: объекты и признаки
 - ❷ Задача обучения по прецедентам
 - ❸ Мотивационные примеры
- ❻ Виды задач
- ❼ Инструменты

Программа курса

- 1 Математика и Питон для Машинного Обучения
- 2 Введение в машинное обучение и обработку данных. Работа с Git
- 3 Классификация
- 4 Регрессия
- 5 Отбор признаков и снижение размерности
- 6 Кластеризация лекция и семинар
- 7 Работа с текстовыми данными
- 8 Введение в глубокое обучение
- 9 Работа с изображениями

Историческая справка

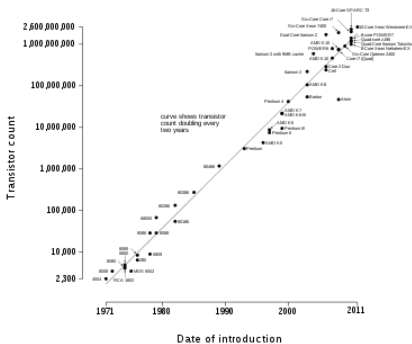
- 1 Артур Сэмюэль, Checkers-playing, 1952 год
- 2 Джозеф Вейцбаум, ELIZA, 1966
- 3 Фрэнк Розенблатт, Perceptron, конец 1950х
- 4 Big Data, MapReduce, Hadoop начало 2000х
- 5 Deep Learning, новые алгоритмы 2010е



Закон Мура

Количество транзисторов на интегральной схеме удваивается каждые 24 месяца, то есть с каждым годом производительность компьютеров увеличивается, открывается простор для изучения **больших данных**.

Microprocessor Transistor Counts 1971-2011 & Moore's Law

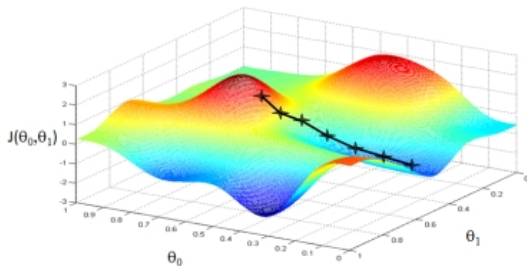
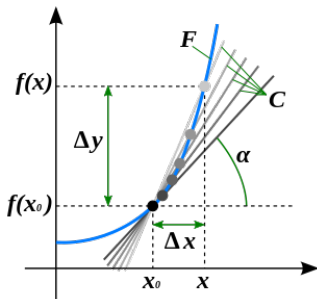


Напоминание. Производная

Производная функция — понятие дифференциального исчисления, характеризующее скорость изменения функции в данной точке.

Определяется как предел отношения приращения функции к приращению её аргумента при стремлении приращения аргумента к нулю, если такой предел существует.

Если производная равна нулю в некоторой точке, то эта точка — **экстремум** (локальный или глобальный максимум или минимум функции). $f'(x_0) = \frac{df}{dx}(x_0)$



Напоминание. Линейная алгебра

Вектор — в линейной алгебре вектором называется элемент линейного пространства. Векторы могут иметь различную природу: направленные отрезки, матрицы, числа, функции и другие, однако все линейные пространства одной размерности изоморфны между собой.

Матрица — математический объект, записываемый в виде прямоугольной таблицы элементов

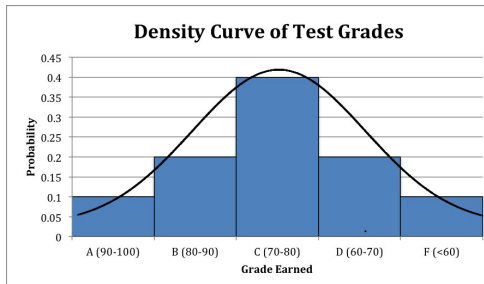
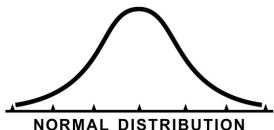
$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

Напоминание. Теория вероятностей

Случайная величина — это переменная, значения которой представляют собой исходы какого-нибудь случайного феномена или эксперимента.

Выборка — набор независимых между собой случайных величин

Распределение вероятностей — это закон, описывающий область значений случайной величины и вероятности их исхода (появления).



X — множество **объектов**

Y — множество **допустимых ответов**

y^* — целевая функция, $y^*: X \rightarrow Y$, $y_i = y^*(x_i)$ известны только на **конечном** подмножестве объектов x_1, \dots, x_m из X

Пары (x_i, y_i) — прецеденты

Совокупность пар таких пар при i из $1, \dots, m$ — **обучающая выборка** (X_{train})

a — **решающая функция** (алгоритм), которая любому объекту из X ставит в соответствие допустимый ответ из Y и приближает целевую функцию y^*

X_{test} — **выборка прецедентов** для тестирования построенного алгоритма a

Для решения задачи обучения по прецедентам в первую очередь фиксируется восстанавливаемой зависимости.

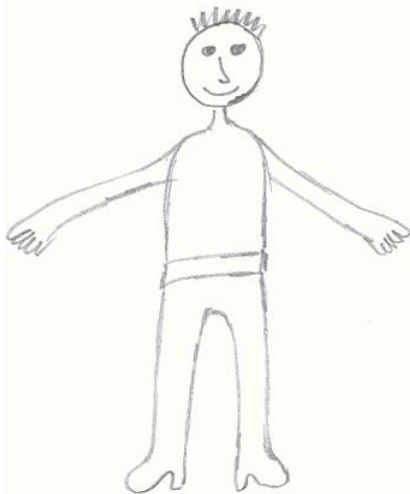
Признак (feature) f объекта x — это результат измерения некоторой характеристики объекта. Формально признаком называется отображение $f : X \rightarrow D_f$, где D_f — множество допустимых значений признака. В частности, любой алгоритм $a : X \rightarrow Y$ также можно рассматривать как признак

Пусть дан набор признаков $f_1(x), \dots, f_n(x)$.

Признаковое описание объекта x — вектор (одномерный массив) (f_1, \dots, f_n) . Совокупность признаковых описаний всех объектов выборки длины m , записанную в виде таблицы размера mn , называют матрицей объектов–признаков.

Основные понятия

Знакомьтесь — это **Вася**. Кем он может быть в нашей терминологии? Какие признаки могут быть у Васи?



Задача обучения по прецедентам

По выборке X_{train} построить решающую функцию (*decisionfunction*) $a : X \rightarrow Y$, которая приближает целевую функцию y^* , причём не только на объектах **обучающей выборки, но и на всём множестве X .**

Решающая функция a должна быть вычислимой.

Кредитный скоринг

Обучающая выборка

1	6	4	12	5	5	3	4	1	67	3	2	1	2	1	0	0	1	0	0	1	0	0	1	1
2	48	2	60	1	3	2	2	1	22	3	1	1	1	1	0	0	1	0	0	1	0	0	1	2
4	12	4	21	1	4	3	3	1	49	3	1	2	1	1	0	0	1	0	0	1	0	1	0	1
1	42	2	79	1	4	3	4	2	45	3	1	2	1	1	0	0	0	0	0	0	0	0	1	1
1	24	3	49	1	3	3	4	4	53	3	2	2	1	1	1	0	1	0	0	0	0	0	1	2
4	36	2	91	5	3	3	4	4	35	3	1	2	2	1	0	0	1	0	0	0	0	1	0	1
4	24	2	28	3	5	3	4	2	53	3	1	1	1	1	0	0	1	0	0	1	0	0	1	1
2	36	2	69	1	3	3	2	3	35	3	1	1	2	1	0	1	1	0	1	0	0	0	0	1
4	12	2	31	4	4	1	4	1	61	3	1	1	1	1	0	0	1	0	0	1	0	1	0	1
2	30	4	52	1	1	4	2	3	28	3	2	1	1	1	1	0	1	0	0	1	0	0	0	2
2	12	2	13	1	2	2	1	3	25	3	1	1	1	1	1	0	1	0	1	0	0	0	1	2
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	1	0	1	0	0	0	1	2
2	12	2	16	1	3	2	1	3	22	3	1	1	2	1	0	0	1	0	0	1	0	0	1	1
1	24	4	12	1	5	3	4	3	60	3	2	1	1	1	1	0	1	0	0	1	0	1	0	2
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	1	0	1	0	0	0	1	1
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	0	0	1	0	0	1	0	1	0	2
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	0	0	1	0	0	1	0	0	1	1
1	30	0	81	5	2	3	3	3	25	1	3	1	1	1	1	0	0	1	0	0	1	0	0	1
2	24	2	126	1	5	2	2	4	44	3	1	1	2	1	0	1	1	0	0	0	0	0	0	2
4	24	2	34	3	5	3	2	3	31	3	1	2	2	1	0	0	1	0	0	1	0	0	1	1
4	9	4	21	1	3	3	4	3	48	3	3	1	2	1	1	0	1	0	0	1	0	0	1	1
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	1	0	0	0	1	1
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	1	0	1	0	0	1	0	1
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	1	1	0	0	1	0	0	1	1
4	10	4	21	5	3	4	1	3	26	3	2	1	1	2	0	0	1	0	0	1	0	0	1	1
1	6	2	14	1	3	3	2	1	36	1	1	1	2	1	0	0	1	0	0	1	0	1	0	1
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	0	0	1	0	1	0	1
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	1	0	1	0	0	0	1	1

Кредитный скоринг

Задание

Данные — информация о выданных кредитах, требуется предсказать вероятность успешного погашения кредита.

X_{train} содержится в файле train.csv, X_{test} — test.csv.

Информация о значениях признаков содержится в файле featureDescr.csv

Целевой признак — $loan_{status}$, бинарный признак. 1 означает, что кредит успешно погашен.

Задача

Предсказать, кому стоит выдавать кредит?

Рекомендательные системы

Популярные товары



4 620 P

Автомобильная
шина MICHELIN...



16.

Форма для кулича
Жостовская фабрика...



4 000 p

Матрас Аскона
Balance Forma...



405 p

Matrix шампунь Total
Results So Long...



9 999 ₪

Диван Hoff Хаген



6 430 P

Кроватка
Mimi 7 в 1



Женская парфюмерия



1 506 P

LACOSTE Lacoste
pour Femme



2745P

Christian Dior J'adore
Eau de Parfum



2 145 P

Dolce &
Gabbana 3...



1 468 P

Versace Bright Crystal




3 620 ₪

Guerlain Mon Guei

[Все товары](#)

Техника для красоты
Фены и многое другое

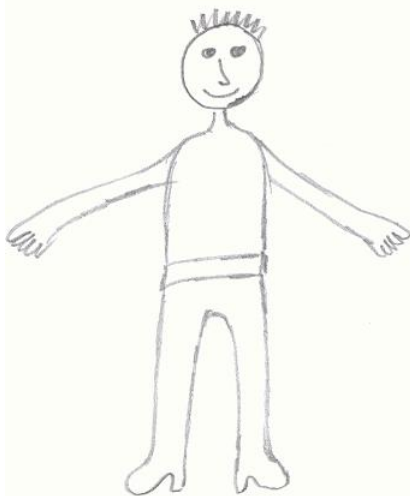


от **300 Р**

Территория детства
Игрушки

Features

Какие важные для конкретных задач признаки могут быть у Васи?



1 Обучение с учителем

Каждый прецедент представляет собой пару «объект, ответ». Требуется найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ. Функционал качества обычно определяется как средняя ошибка ответов, выданных алгоритмом, по всем объектам выборки.

2 Обучение без учителя

В этом случае ответы не задаются, и требуется искать зависимости между объектами.

1 Частичное обучение

Комбинация первых двух вариантов

2 Обучение с подкреплением

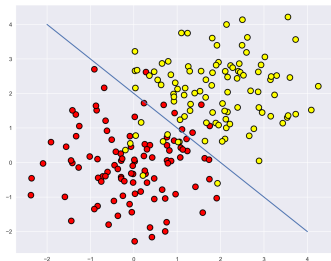
Роль объектов играют пары «ситуация, принятое решение», ответами являются значения функционала качества, характеризующего правильность принятых решений (реакцию среды). Как и в задачах прогнозирования, здесь существенную роль играет фактор времени. Примеры прикладных задач: формирование инвестиционных стратегий, автоматическое управление технологическими процессами, самообучение роботов, и т.д.

3 etc

Трансдуктивное, активное, метаобучение..

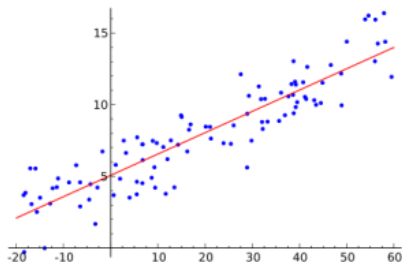
Классификация

Множество допустимых ответов конечно. Их называют метками классов (class label). Класс — это множество всех объектов с данным значением метки.



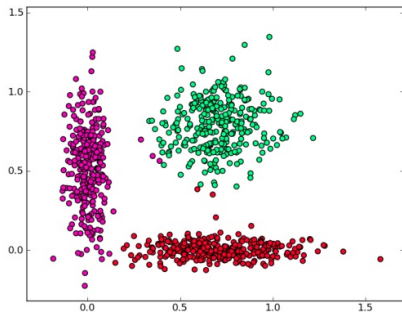
Регрессия

Отличается тем, что допустимым ответом является действительное число или числовой вектор.



Кластеризация

Заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.



- 1 Jupyter Notebook
- 2 Python (NumPy, SciPy, Sklearn, Pandas,...)
- 3 Математический аппарат