Машинное обучение

Лекция 1

Основные понятия, стандартные задачи и простые модели

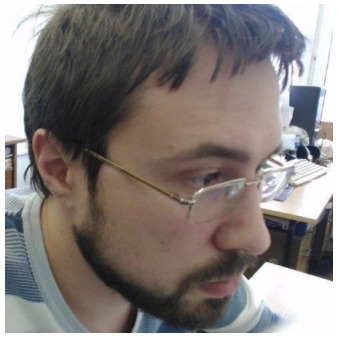
Материалы: Виктор Кантор

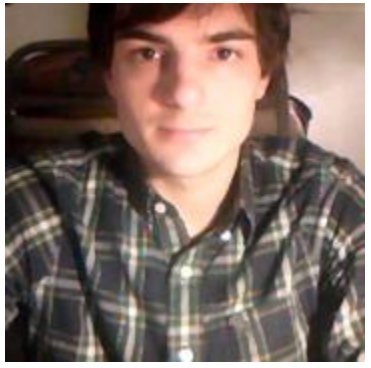
Корректировки: Анастасия Зухба

Содержание курса

1. Введение: основные понятия и задачи, простые методы
2. Решающие деревья и ансамбли решающих деревьев
3. Линейные модели в задачах классификации и регрессии
4. Нейронные сети
5. Обучение без учителя

Команда курса





На этой лекции

I. Примеры применения машинного обучения
II. Стандартные задачи и простые методы
III. Идеи часто используемых моделей
IV. Оптимизационные задачи в машинном обучении
V. Переобучение и недообучение
VI. Инструменты

I. Примеры применения

Кредитный скоринг

Выдача кредита

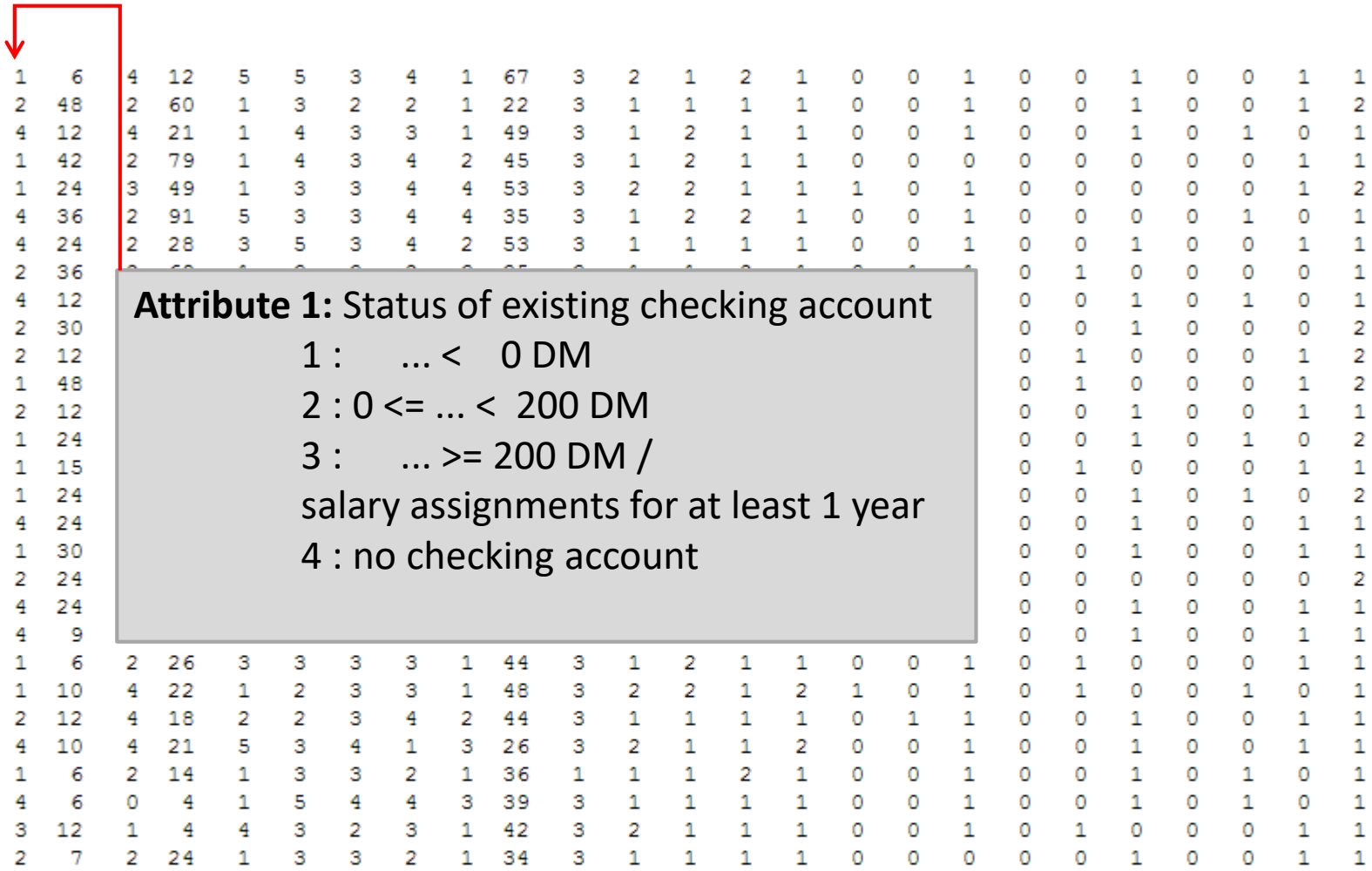
German credit data set (UCI репозиторий)

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	0	0	1	0	0	1	0	0	1	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
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	0	0	0	1	0	0	1	1

Выдача кредита

German credit data set (UCI репозиторий)



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 0 1 0 0 0 0 1
4 12 0 0 1 0 1 0 1
2 30 0 0 1 0 0 0 2
2 12 0 1 0 0 0 1 2
1 48 0 1 0 0 0 1 2
2 12 0 0 1 0 0 1 1
1 24 0 0 1 0 1 0 2
1 15 0 1 0 0 0 1 1
1 24 0 0 1 0 1 0 2
4 24 0 0 1 0 0 1 1
1 30 0 0 1 0 0 1 1
2 24 0 0 0 0 0 0 2
4 24 0 0 1 0 0 1 1
4 9 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
2 7 2 24 1 3 3 2 1 34 3 1 1 1 1 0 0 0 0 0 1 0 0 1 1

Attribute 1: Status of existing checking account

1 : ... < 0 DM

2 : 0 <= ... < 200 DM

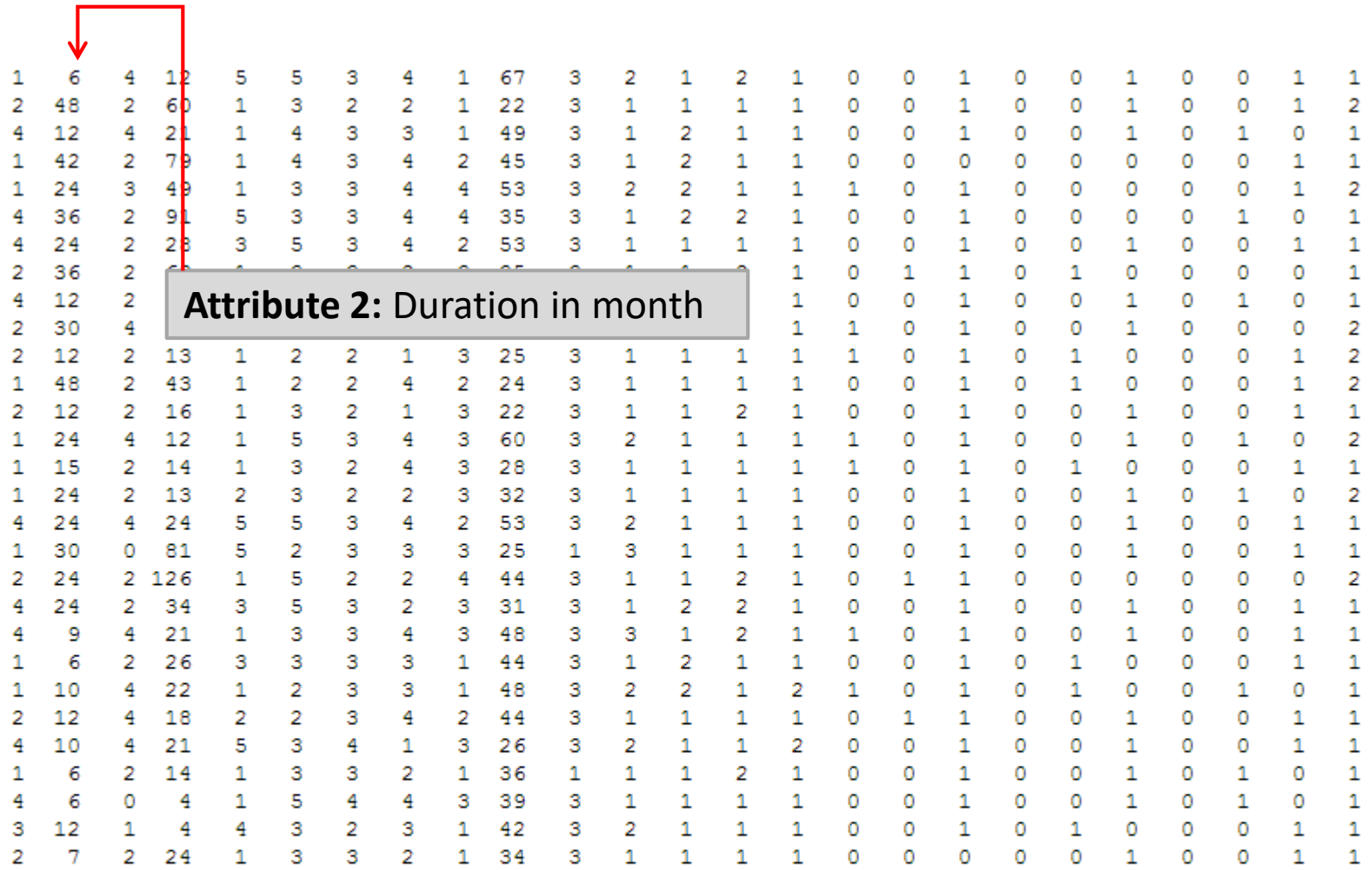
3 : ... >= 200 DM /

salary assignments for at least 1 year

4 : no checking account

Выдача кредита

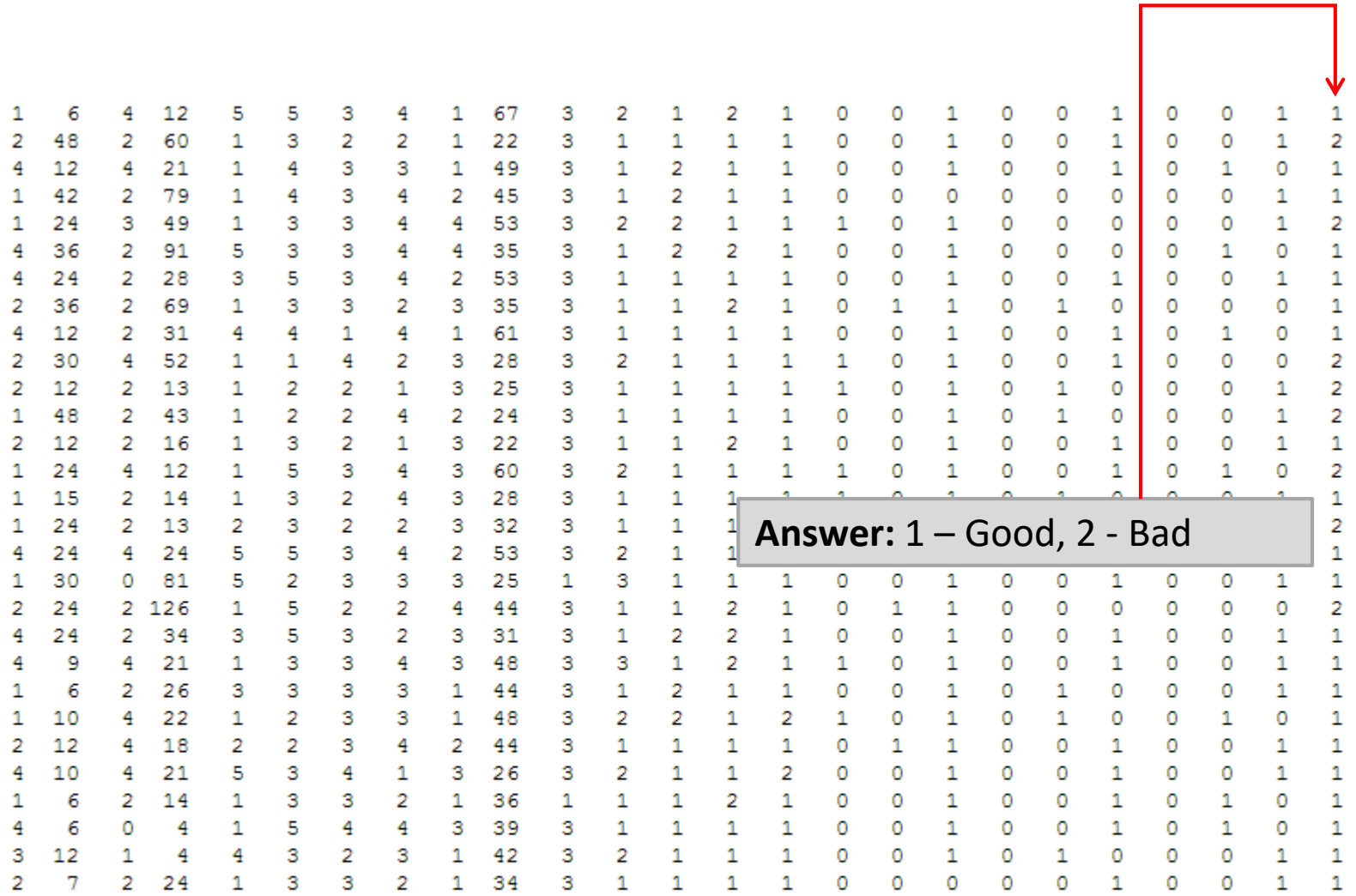
German credit data set (UCI репозиторий)



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												1	0	1	1	0	1	0	0	0	0	1
4	12	2												1	0	0	1	0	0	1	0	1	0	1
2	30	4												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	0	0	1	0	0	1	0	0	1	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
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	0	0	0	1	0	0	1	1

Выдача кредита

German credit data set (UCI репозиторий)



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	24	2	13	2	3	2	2	3	32	3	1	1	1											2
4	24	4	24	5	5	3	4	2	53	3	2	1	1											1
1	30	0	81	5	2	3	3	3	25	1	3	1	1	1	0	0	1	0	0	1	0	0	1	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
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	0	0	0	1	0	0	1	1

Answer: 1 – Good, 2 - Bad

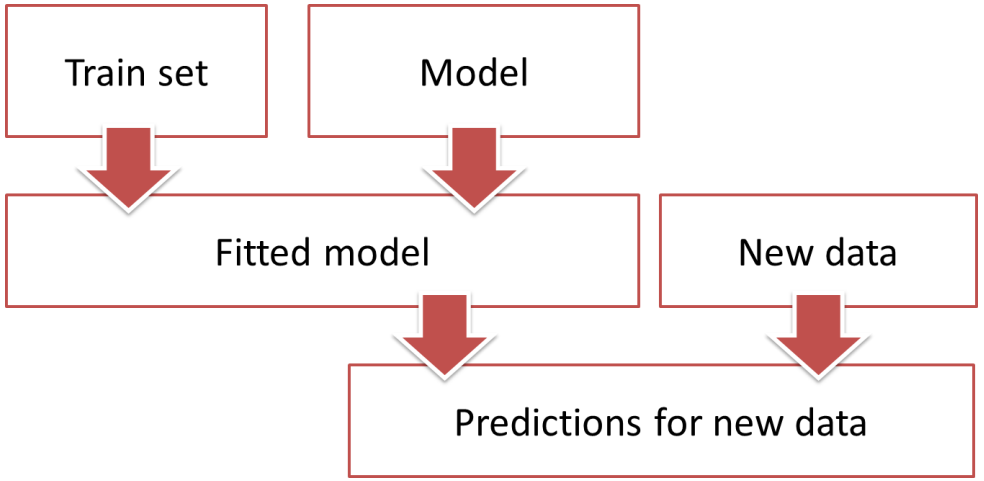
Выдача кредита

Задача (supervised classification): предсказать класс (1 или 2)

1	60	3	68	1	5	3	4	4	63	3	2	1	2	1	0	0	1	0	0	1	0	0	1	?
2	18	2	19	4	2	4	3	1	36	1	1	1	2	1	0	0	1	0	0	1	0	0	1	?
1	24	2	40	1	3	3	2	3	27	2	1	1	1	1	0	0	1	0	0	1	0	0	1	?
2	18	2	59	2	3	3	2	3	30	3	2	1	2	1	1	0	1	0	0	1	0	0	1	?
4	12	4	13	5	5	3	4	4	57	3	1	1	1	1	0	0	1	0	1	0	0	1	0	?
3	12	2	15	1	2	2	1	2	33	1	1	1	2	1	0	0	1	0	0	1	0	0	0	?
2	45	4	47	1	2	3	2	2	25	3	2	1	1	1	0	0	1	0	0	1	0	1	0	?

Test set

Более глобальная задача:
Придумать алгоритм, генерирующий алгоритм классификации (“обученную модель”) на данной выборке



Кредитный скоринг: вопросы

1. Какой экономический эффект может дать модель в этой задаче? Как он связан с качеством модели? (как его измерять)
2. Будет ли оценка ожидаемого экономического эффекта на исторических данных совпадать с реальным экономическим эффектом? Как можно измерить его?
3. Какие данные нужны для построения модели?

Рекомендации товаров

Блок рекомендаций

Товар 1	**Товар 2**	**Товар 3**	**Товар 4**

Возможный вариант заполнения

Puma
Ветровка
3 490 руб.

Crocs
Сланцы
1 990 руб.

Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

История про одинаковое качество

- Интегрировали чужое решение, чтобы сравнить качество со своим
- Оценили качество у обоих
- Совпало до тысячных долей
- Не стали использовать чужое решение
- Позже – выяснили, в чем дело

История про статзначимость

История про статзначимость

Одна кривая отличается от других на 10%

Но разбиение на самом деле – случайное

Рекомендации товаров: вопросы

1. Какой экономический эффект может дать модель в этой задаче? Как он связан с качеством модели? (и как его измерять)
2. Будет ли оценка ожидаемого экономического эффекта на исторических данных совпадать с реальным экономическим эффектом? Как можно измерить его?
3. Какие данные нужны для построения модели?

Ещё примеры

Яндекс

Data Mining in Action

Найти

Войти

ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ ДИСК МУЗЫКА ЕЩЁ

Нашлось 8 млн результатов

Дать объявление Показать все

Data Mining in Action | ВКонтакте

vk.com › **data_mining_in_action**

Москва, Россия Денис Семененко. Администратор сообщества. Data Mining in Action. So it begins. Местоположение: Москва, Россия. . Data Mining in Action запись закреплена. 6 мая в 23:04.

Process **Mining**: знакомство / Хабрахабр

habrahabr.ru › post/244879/

Статья подготовлена на основе материалов онлайн курса Process **Mining**: **Data** Science **in Action**, являющихся собственностью Технического университета Эйндховена.

Process **Mining**: **Data** science **in Action**... | Coursera

coursera.org › learn/process-**mining**


MegaFon 3G
16:06
92 %
8
дтп
1.4 км


Google
self-driving car

II. Стандартные задачи и простые методы их решения

Классификация

Iris setosa

Iris versicolor

Iris virginica

<u>Вход (обучающая выборка):</u>

Признаки N объектов с известными классами

<u>Выход:</u>

Классификатор (алгоритм, прогнозирующий классы новых объектов по их признакам)

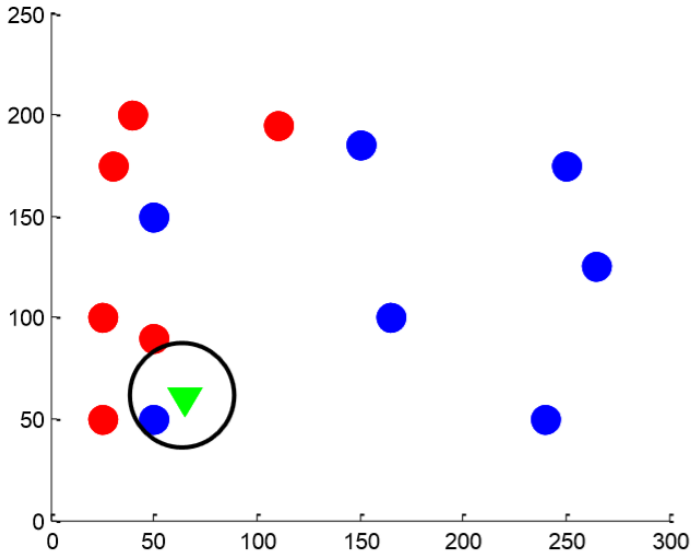
Классификация: обучающая выборка

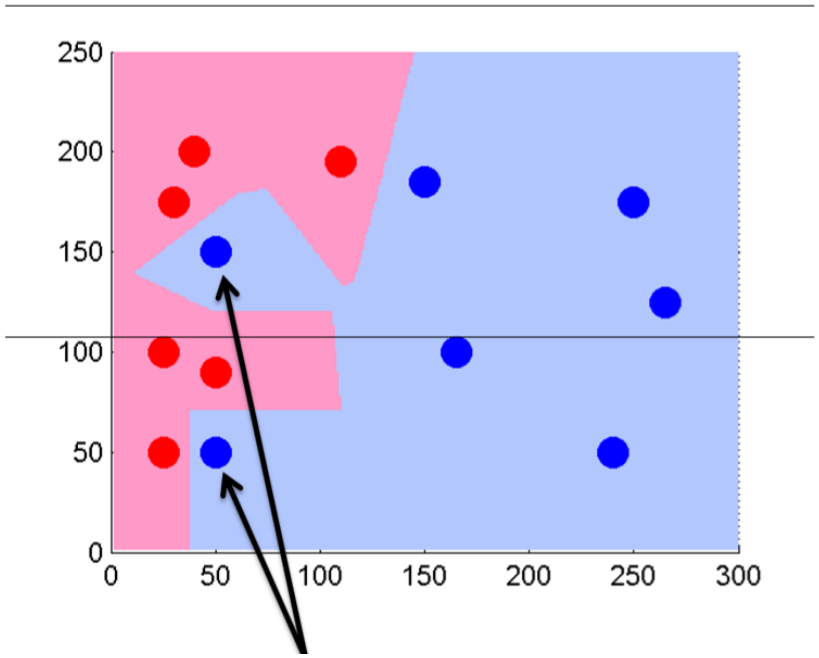
Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
5.7	2.5	5.0	2.0	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*
6.3	2.5	4.9	1.5	*I. versicolor*
4.9	2.5	4.5	1.7	*I. virginica*

Простой классификатор: kNN

k nearest neighbours



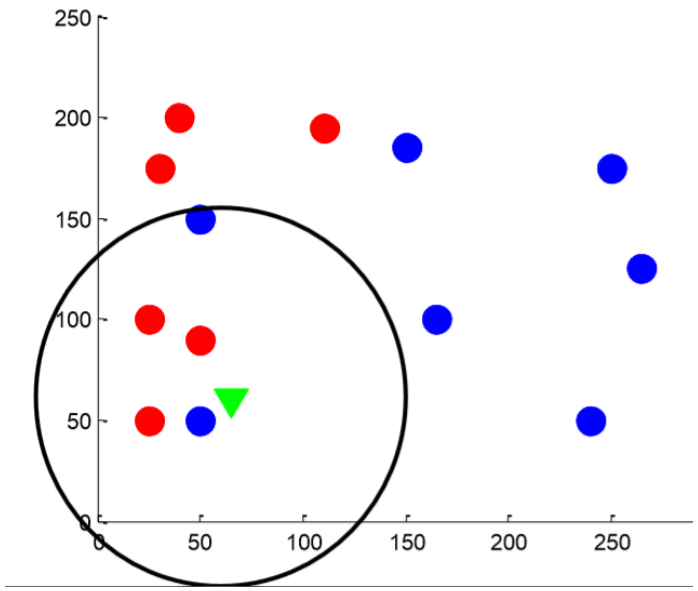


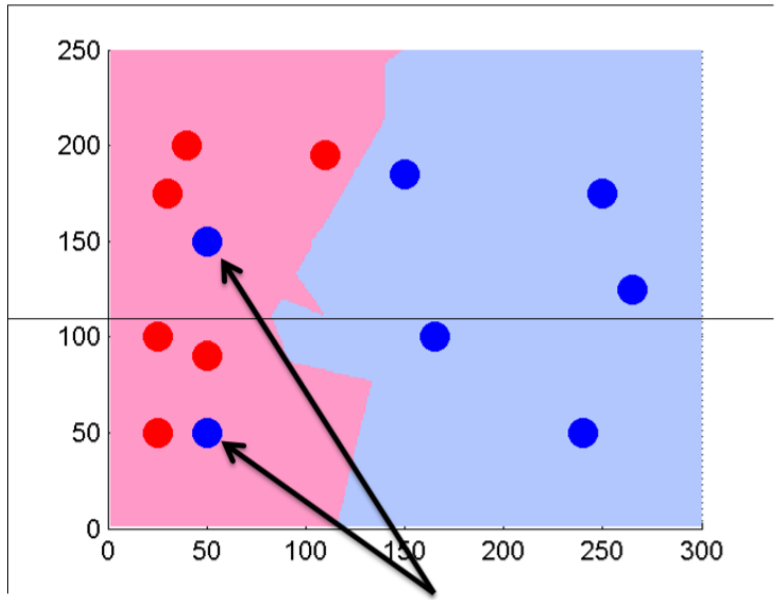
Шумы? (outliers)

k = 1

Простой классификатор: kNN

k nearest neighbours



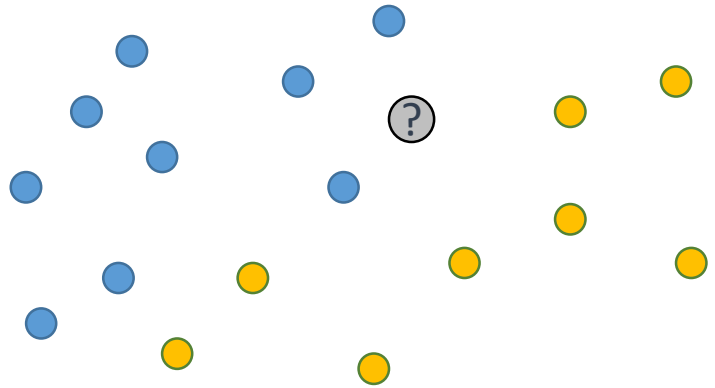


Ошибки?

k = 5

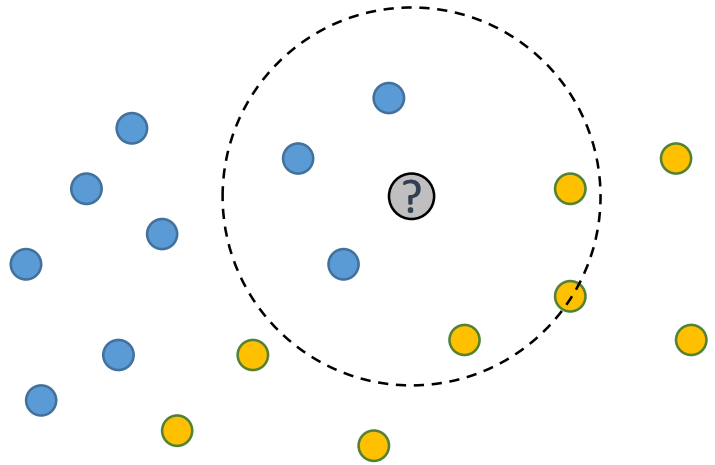
Взвешенный kNN

Пример классификации (k = 6):



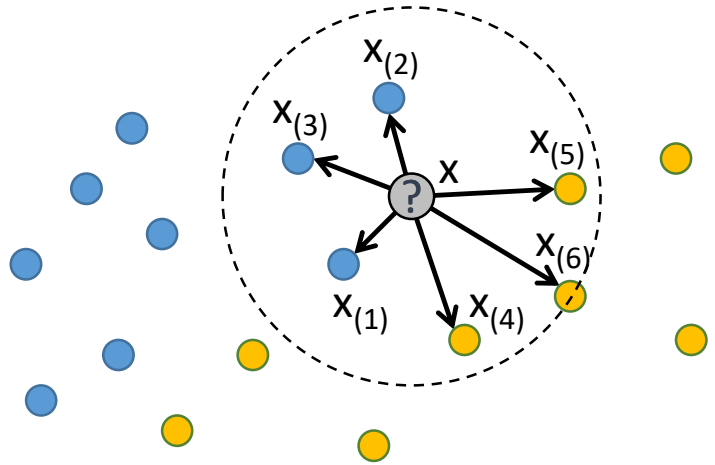
Взвешенный kNN

Пример классификации (k = 6):



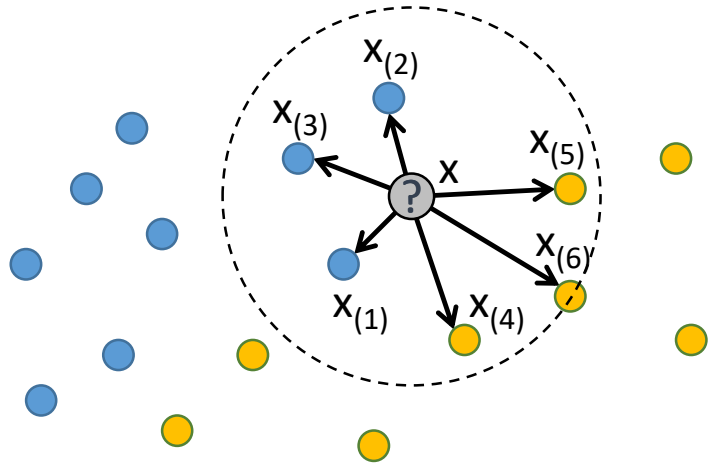
Взвешенный kNN

Пример классификации (k = 6):



Взвешенный kNN

Пример классификации (k = 6):

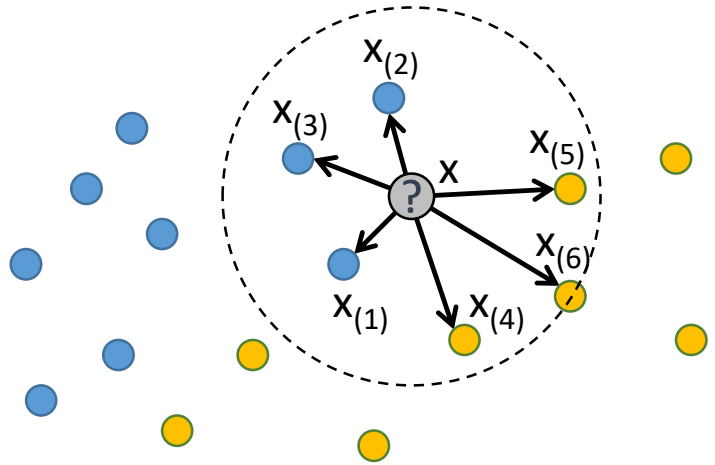


Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

Взвешенный kNN

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

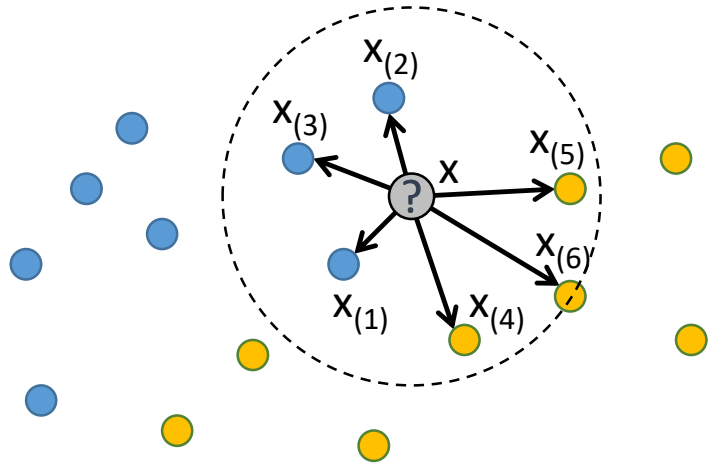
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

Взвешенный kNN

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

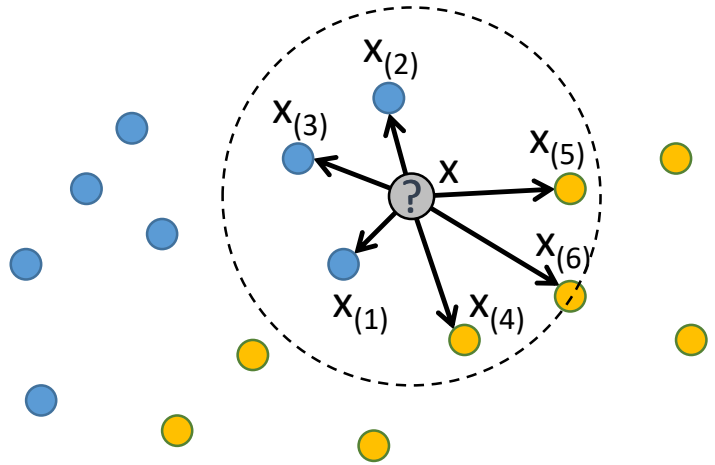
или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

$$Z = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

Взвешенный kNN

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

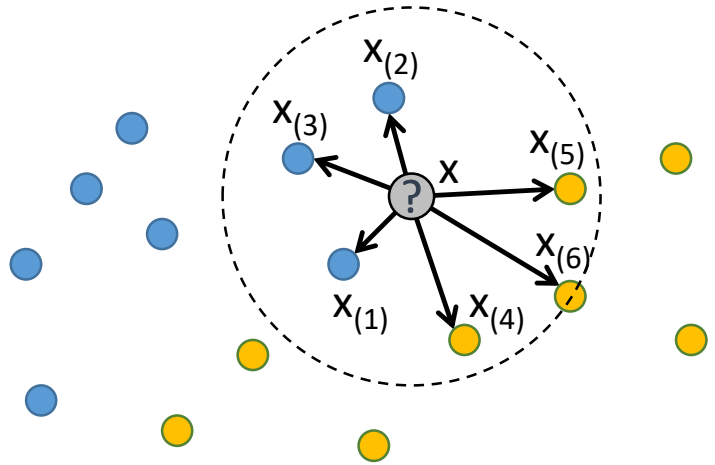
$$w(x(i)) = w(d(x, x_{(i)}))$$

$$Z_{\bullet} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$Z_{\bullet} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

Взвешенный kNN

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

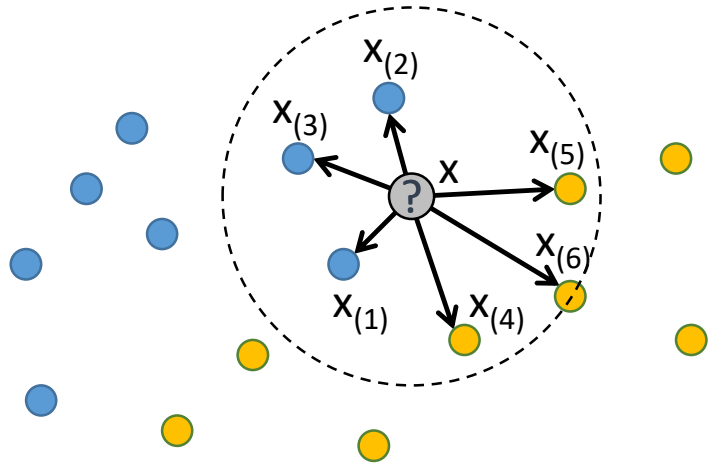
$$Z_{\bullet} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$Z_{\bullet} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$? = \underset{\bullet}{\operatorname{argmax}}\, Z_{\bullet}$$

Взвешенный kNN

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

$$Z_{\bullet} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

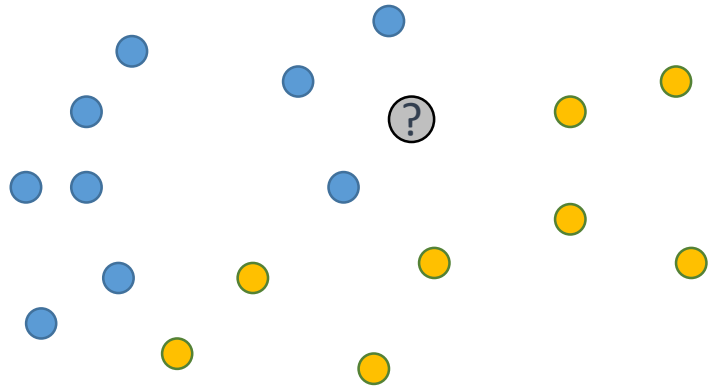
$$Z_{\bullet} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

? = argmax Z

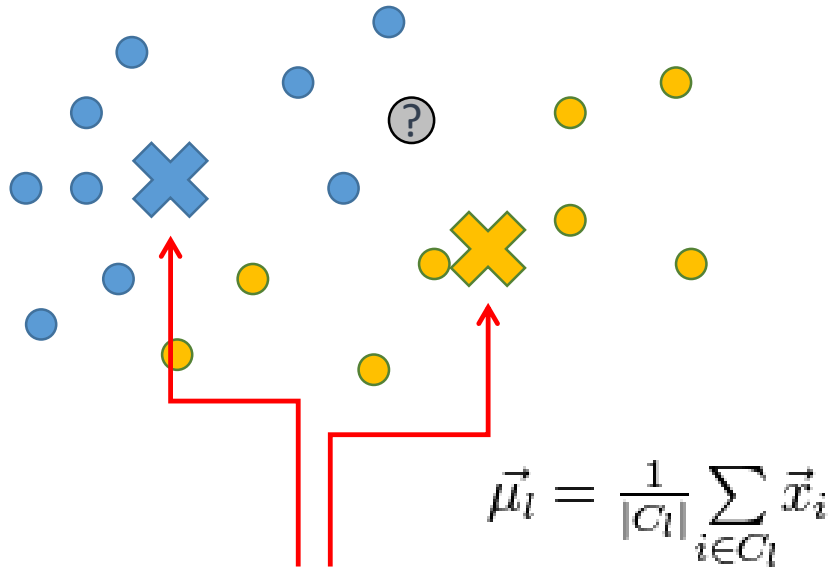
if $Z_{\bullet} > Z_{\bullet}$: ? = ●

if $Z_{\bullet} < Z_{\bullet}$: ? = ●

Центроидный классификатор

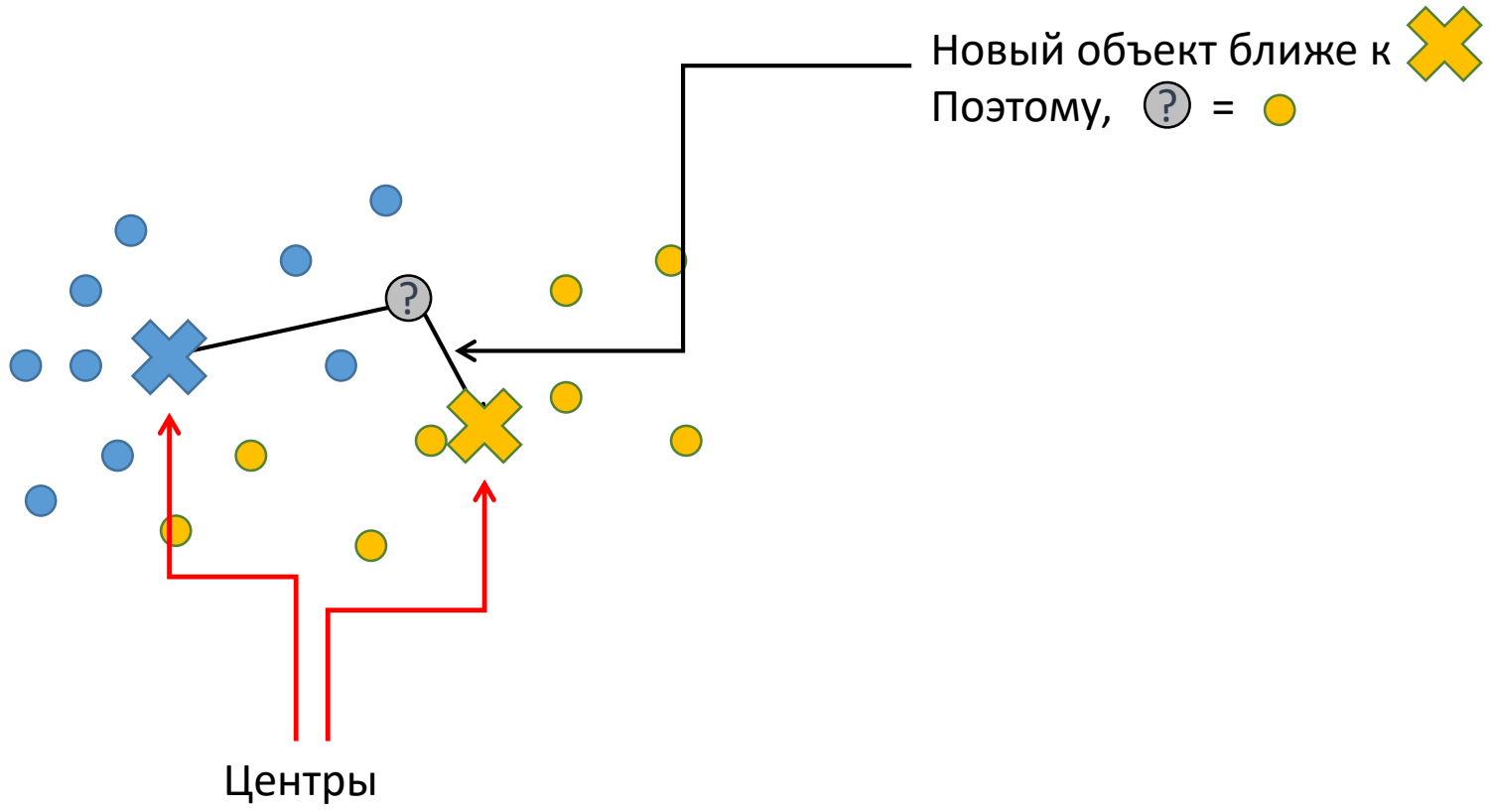


Центроидный классификатор



$$\vec{\mu}_l = \frac{1}{|C_l|} \sum_{i \in C_l} \vec{x}_i$$

Центроидный классификатор



Байесовский классификатор

По известному вектору признаков x
алгоритм относит объект к классу a(x)
по правилу:

$$a(x) = \operatorname{argmax}_y P(y|x)$$

Байесовский классификатор

$$a(x) = \operatorname{argmax}_y P(y|x)$$

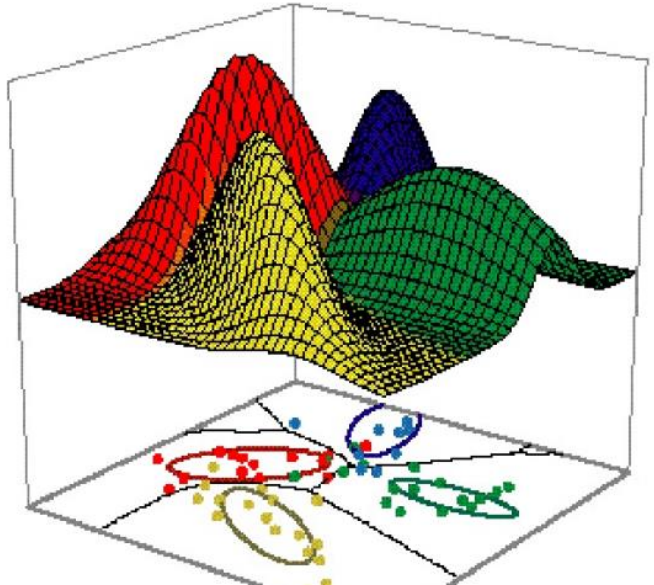
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$a(x) = \operatorname{argmax}_y P(x|y)\, P(y)$$

Байесовский классификатор

$$a(x) = \operatorname{argmax}_y P(x|y)\, P(y)$$

Если P(y) одинаковы для всех классов – мы просто выбираем класс, плотность которого больше в точке x



Зачем нам понадобилась теорема Байеса

- P(y|x) – вероятность класса y при признаках x
- X часто из вещественных чисел и признаков часто очень много
- Всевозможных значений признаков так много, что скорее всего каждый вектор x встретится только один или несколько раз
- Этого недостаточно для оценки P(y|x)

Что оценивается по обучающей выборке

- $P(x|y)$ – вероятность увидеть набор признаков x в классе y, если x дискретный
- Если координаты вектора x – вещественные, $P(x|y)$ – плотность распределения x
- Именно эту величину и можно оценивать по обучающей выборке
- А затем подставлять в классификатор:

$$a(x) = \operatorname{argmax}_y P(x|y)\, P(y)$$

Наивный байесовский классификатор

$$a(x) = argmax_y\, P(y) \prod_{k=1}^{n} P(x^{(k)}|y)$$

$P(x^{(k)}|y)$ оцениваем, предположив, что это распределение из какого-то стандартного семейства: нормальное, Бернулли, мультиномиальное

Немного истории: фильтрации спама

- Задача построения спам-фильтра – в чистом виде задача классификации писем на 2 класса: спам и не спам
- Первые спам-фильтры использовали наивный байесовский классификатор

Примеры спама

- Hi! :) **Purchase Exclusive** Tabs Online **http://**…
- We **Offer** Loan At A **Very Low Rate** Of 3%. If **Interested**, Kindly Contact Us, Reply by this email **….@hotmail.com**
- **Купите** специализацию Машинное обучение и анализ данных от МФТИ и Яндекса с супер-**скидкой** 0,99%! Станьте Data Scientist **за 5 месяцев**!

Фильтруем спам: обучение

1. Посчитать для каждого слова **w** из коллекции текстов количество писем с ним $\mathbf{n_{ws}}$ в спаме (spam) и количество писем с ним $\mathbf{n_{wh}}$ в «не спаме» (ham)

2. Оценить вероятность появления каждого слова **w** в спамном и в неспамном тексте:

   $\mathbf{P(w|spam) = n_{ws} / n_s}$

   $\mathbf{P(w|ham) = n_{wh} / n_h}$

Фильтруем спам: применение

Получив текст письма, для которого нужно определить, относится оно к спаму или нет, мы можем:

1. Оценить вероятность появления всего текста в классе «спам» и в классе «не спам» просто произведением вероятностей слов:

$$P(\text{new text}|\text{spam}) = P(w_1|\text{spam})\,P(w_2|\text{spam})\,\ldots\,P(w_N|\text{spam})$$

$$P(\text{new text}|\text{ham}) = P(w_1|\text{ham})\,P(w_2|\text{ham})\,\ldots\,P(w_N|\text{ham})$$

Фильтруем спам: применение

Получив текст письма, для которого нужно определить, относится оно к спаму или нет, мы можем:

1. Оценить вероятность появления всего текста в классе «спам» и в классе «не спам» просто произведением вероятностей слов:

   $$P(\text{new text}|\text{spam}) = P(w_1|\text{spam})\,P(w_2|\text{spam})\,\ldots\,P(w_N|\text{spam})$$

   $$P(\text{new text}|\text{ham}) = P(w_1|\text{ham})\,P(w_2|\text{ham})\,\ldots\,P(w_N|\text{ham})$$

2. Выбрать тот класс, в котором вероятность возникновения этого текста больше:

   $$a(\text{new text}) = \text{argmax}_y\, P(\text{text}|y)$$

Это почти правильный алгоритм, не хватает только **P(y)**

Фильтрация спама: вопросы

- Как рассмотренный классификатор спама связан с рассмотренным до этого наивным байесовским?
- Что является признаками?
- Распределения из какого семейства восстанавливаются?

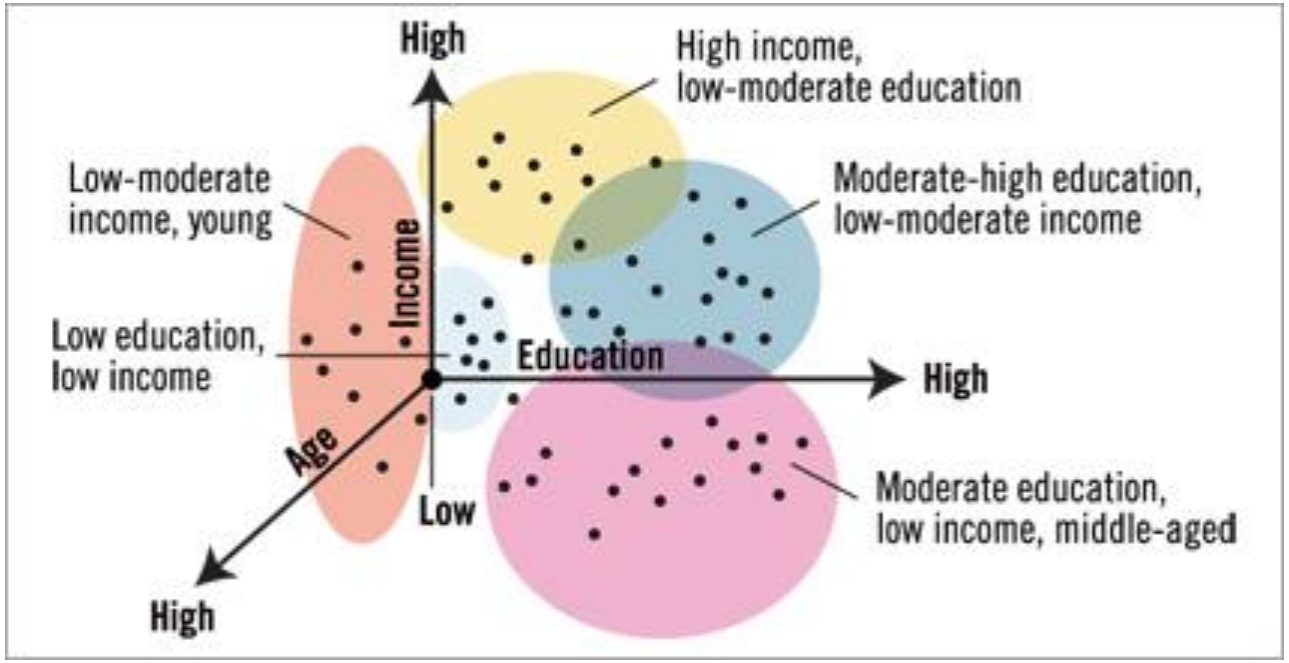
Кластеризация

Вход (обучающая выборка):

Признаки N объектов

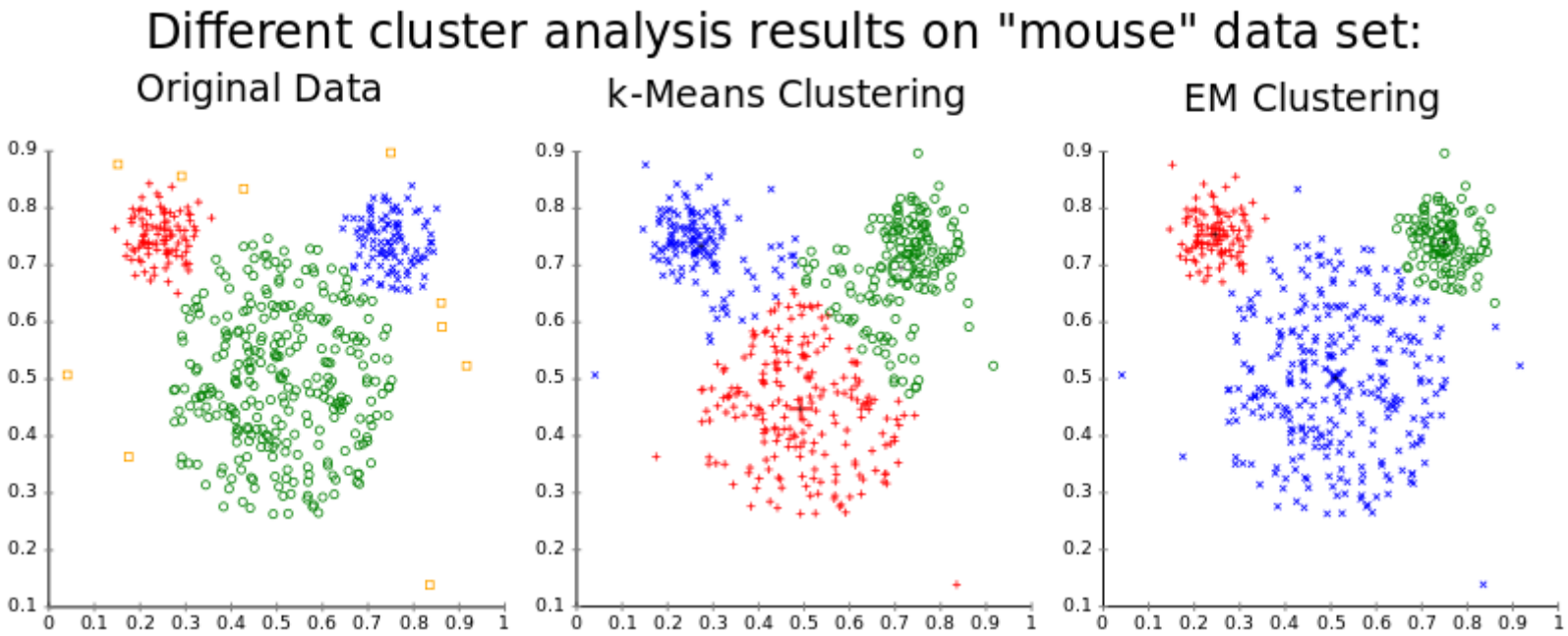
Выход:

Найденные в выборке классы (кластеры), метки кластеров для объектов из обучающей выборки и алгоритм отнесения новых объектов к кластеру

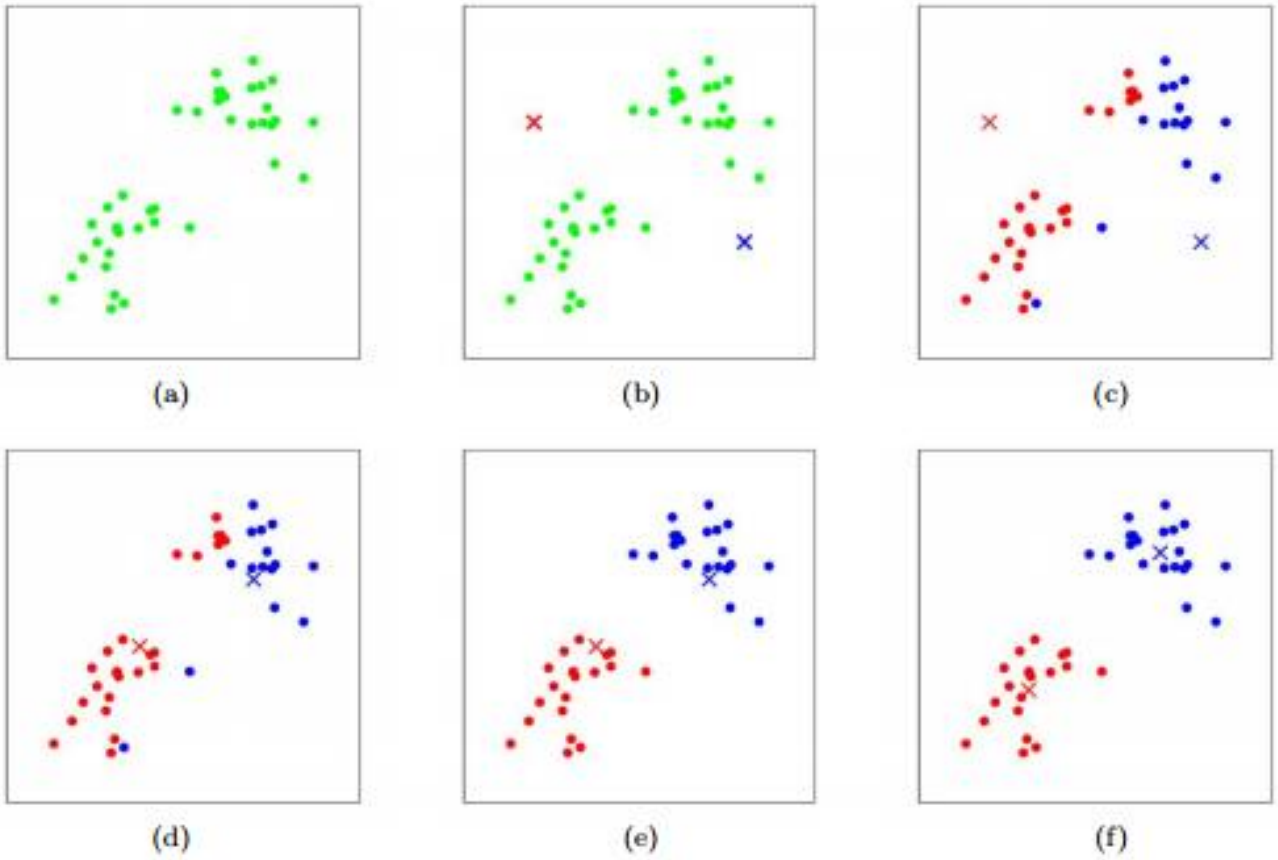


Пример: сегментация рынка

Кластеризация



Простой алгоритм кластеризации: kMeans



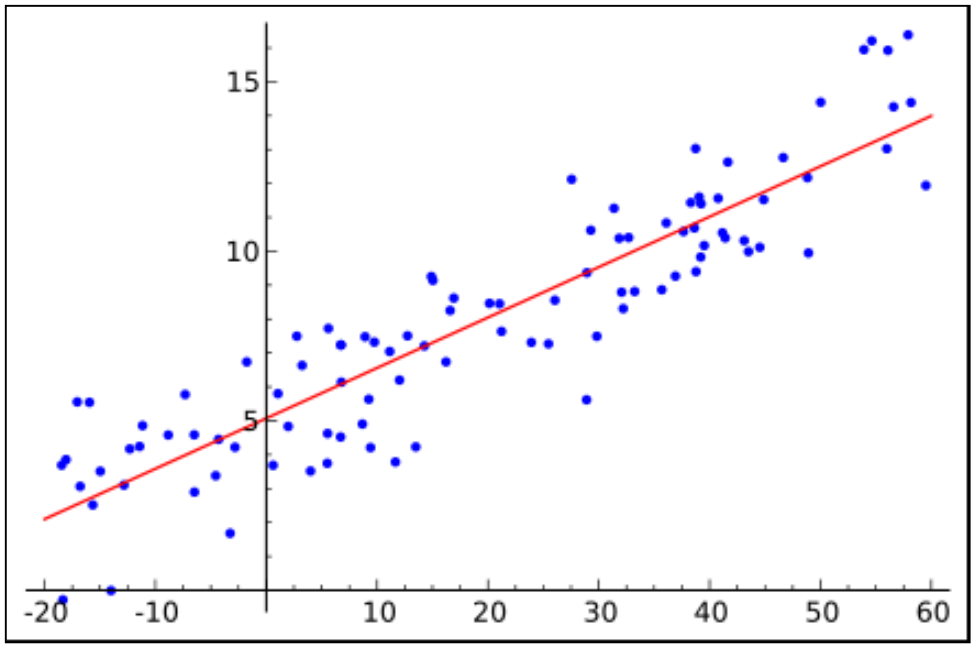
Регрессия

Вход (обучающая выборка):

Признаки N объектов с известными значениями прогнозируемого вещественного параметра объекта

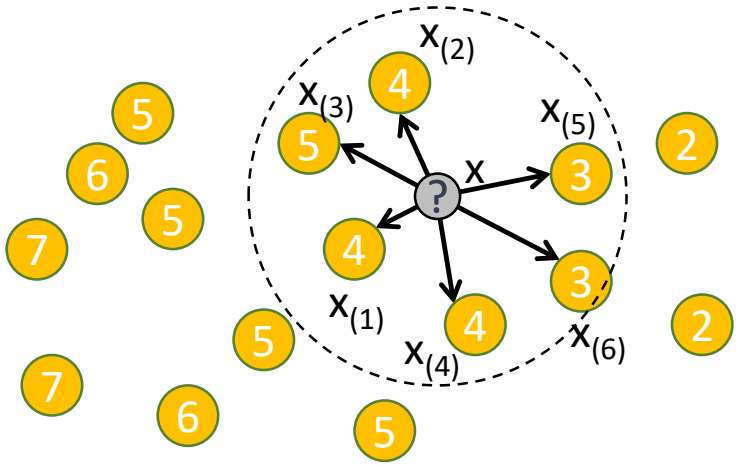
Выход:

Алгоритм, прогнозирующий значение вещественной величины по признакам объекта



Взвешенный kNN для регрессии

Пример (k = 6):



Веса можно определить как функцию от соседа или его номера:

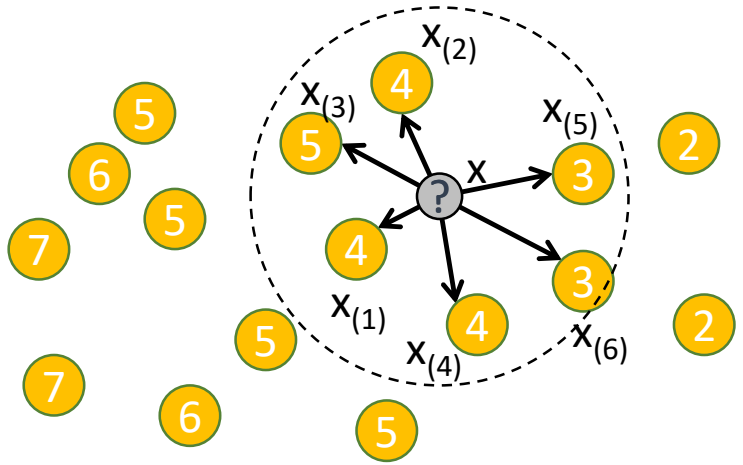
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

Взвешенный kNN для регрессии

Пример (k = 6):



Веса можно определить как функцию от соседа или его номера:

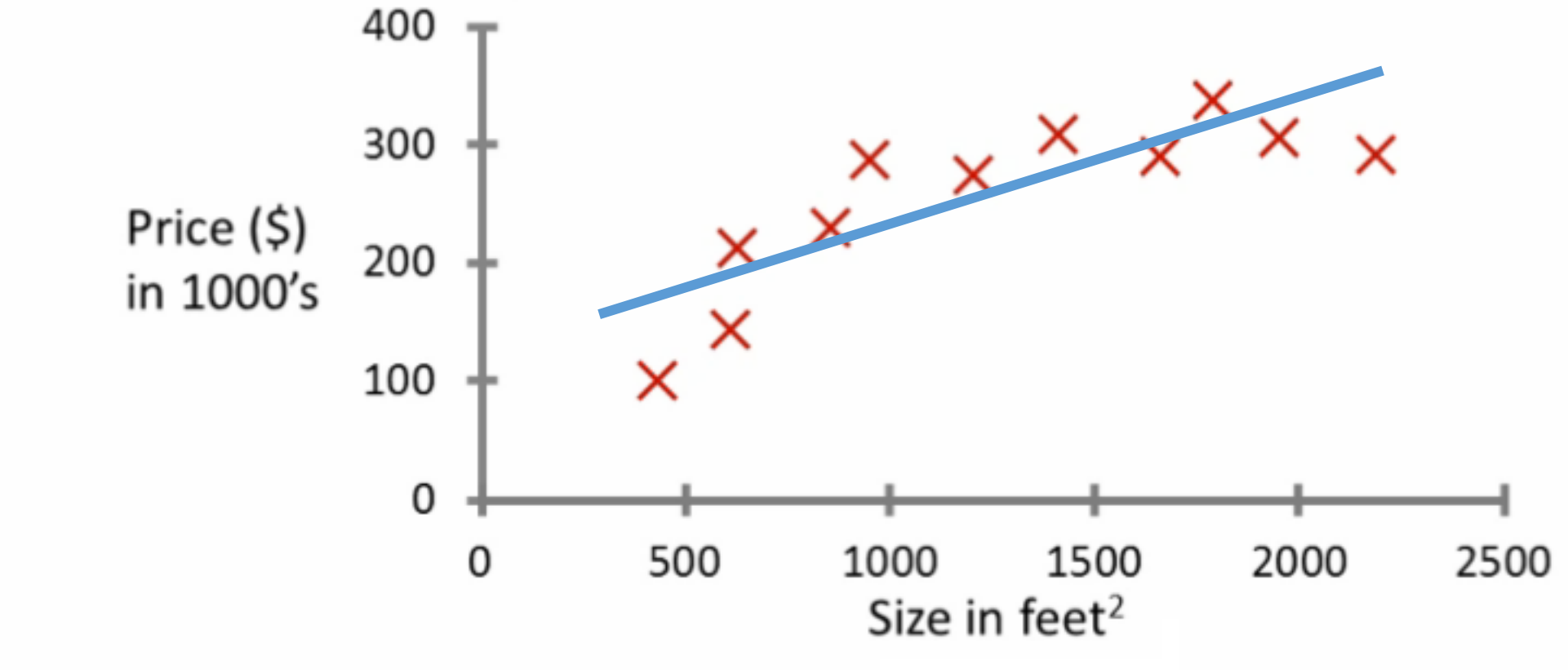
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

$$? = \frac{4 \cdot w(x_{(1)}) + 4 \cdot w(x_{(2)}) + 5 \cdot w(x_{(3)}) + 4 \cdot w(x_{(4)}) + 3 \cdot w(x_{(5)}) + 3 \cdot w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

Линейная регрессия



Линейная регрессия

Модель: $y_i \approx \hat{y}_i = < w, x_i > + w_0$

Линейная регрессия

Модель: $y_i \approx \hat{y}_i = < w, x_i > + w_0$

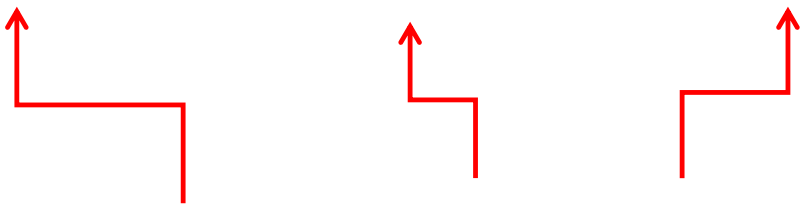
Если добавить $x_{i0} = 1$:

$$y_i \approx \hat{y}_i = < w, x_i >$$

$$y_1 \approx \hat{y}_1 = x_1{}^T w$$
$$\dots$$
$$y_i \approx \hat{y}_i = x_i{}^T w$$
$$\dots$$
$$y_l \approx \hat{y}_l = x_l{}^T w$$

Матричная запись

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_l \end{pmatrix} \approx \begin{pmatrix} \widehat{y_1} \\ \widehat{y_2} \\ \cdots \\ \widehat{y_l} \end{pmatrix} = \begin{pmatrix} x_1{}^T \\ x_2{}^T \\ \cdots \\ x_l{}^T \end{pmatrix} w$$



$$y \approx \hat{y} = Fw$$

$$w = argminw\|y - \hat{y}\|^2$$

Веса признаков

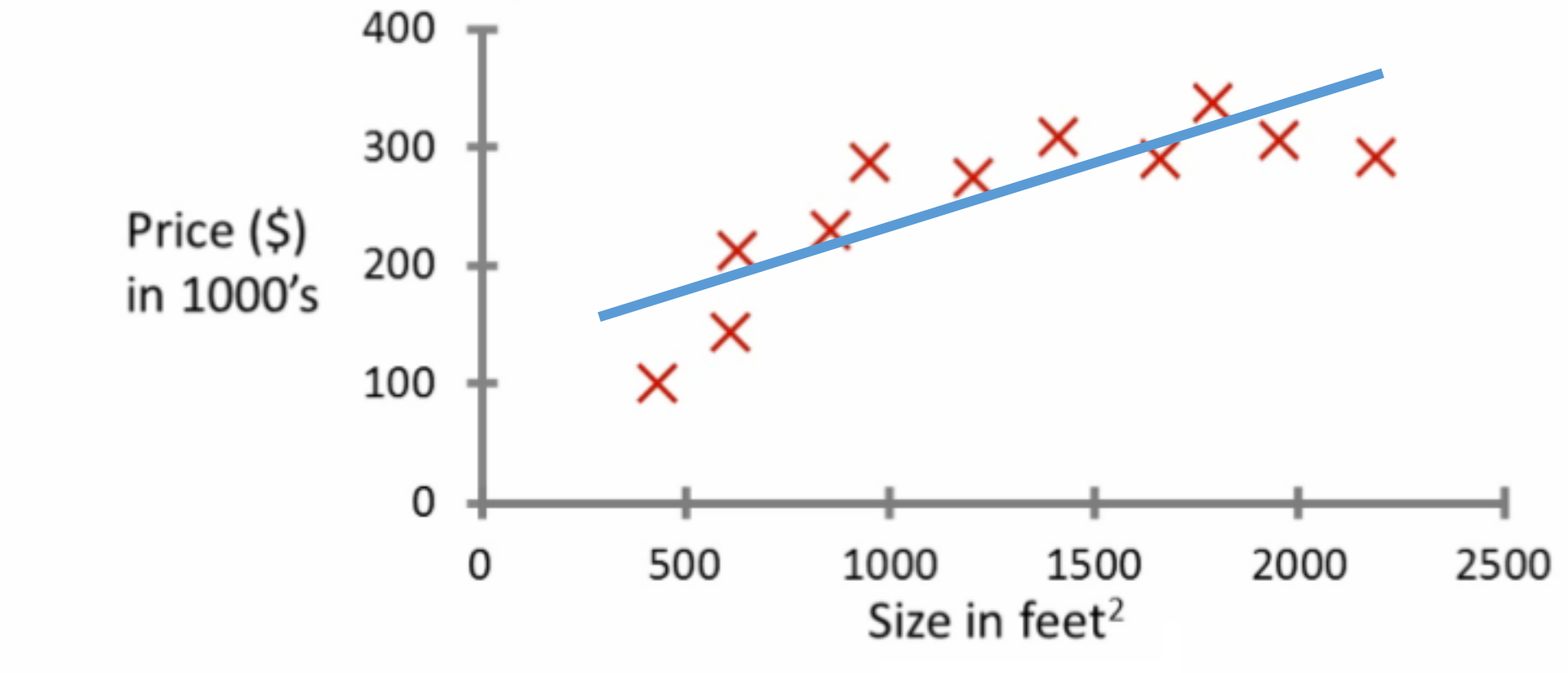
$$\frac{\partial (y - Fw)^2}{\partial w} = 2F^T(y - Fw) = 0$$

$$F^T F w = F^T y$$

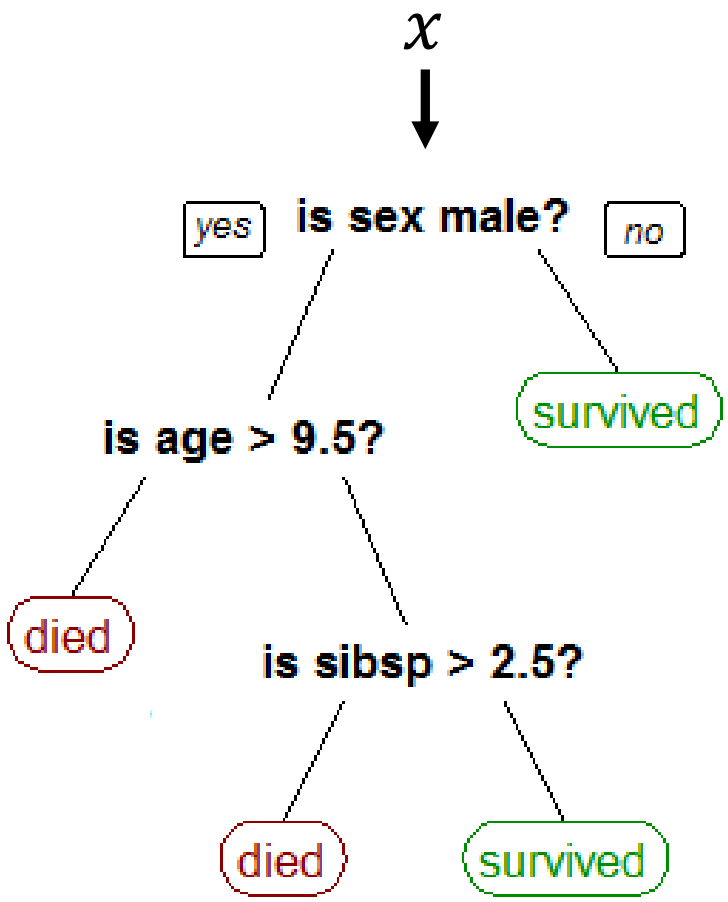
$$\boxed{w = (F^T F)^{-1} F^T y}$$

III. Идеи часто используемых методов

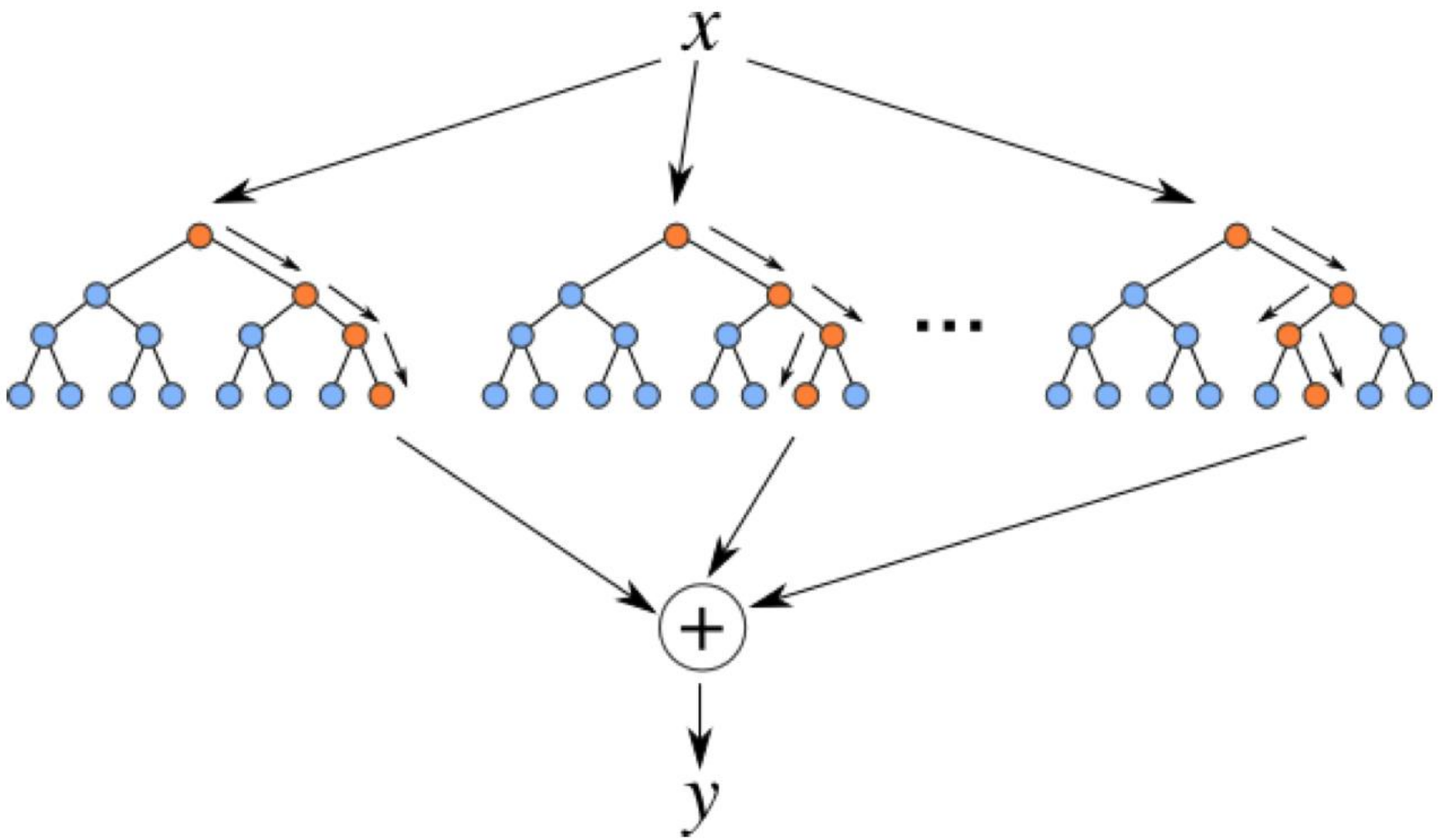
Линейные модели



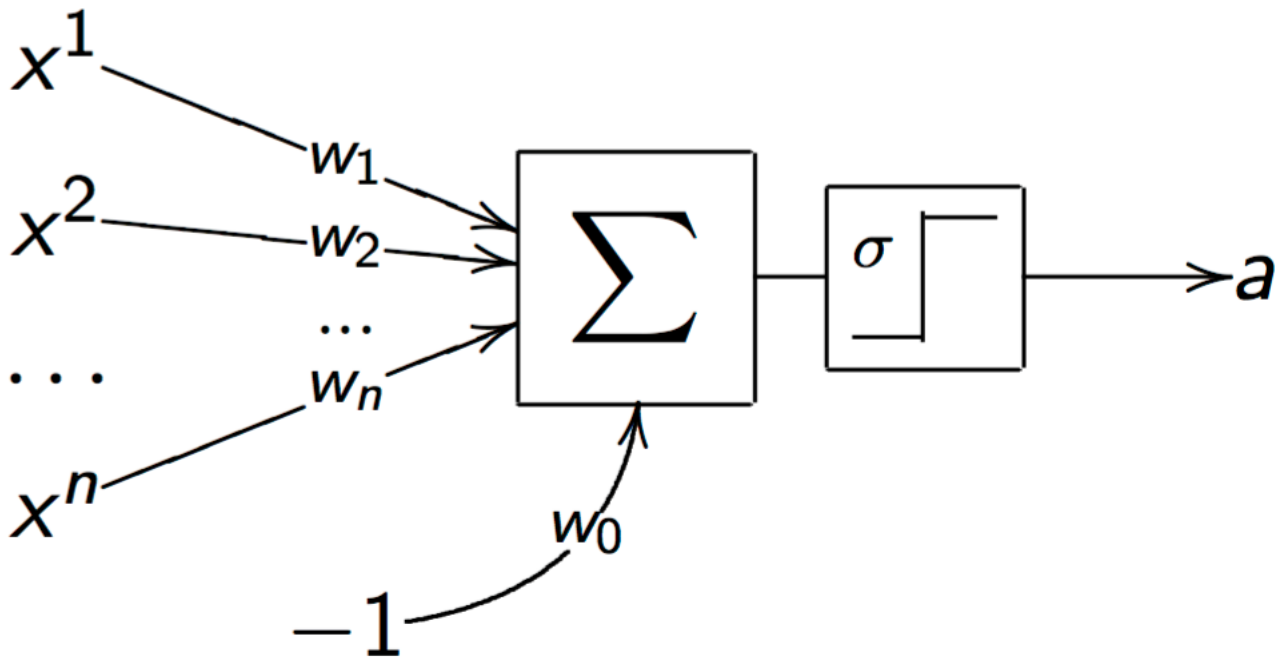
Решающие деревья



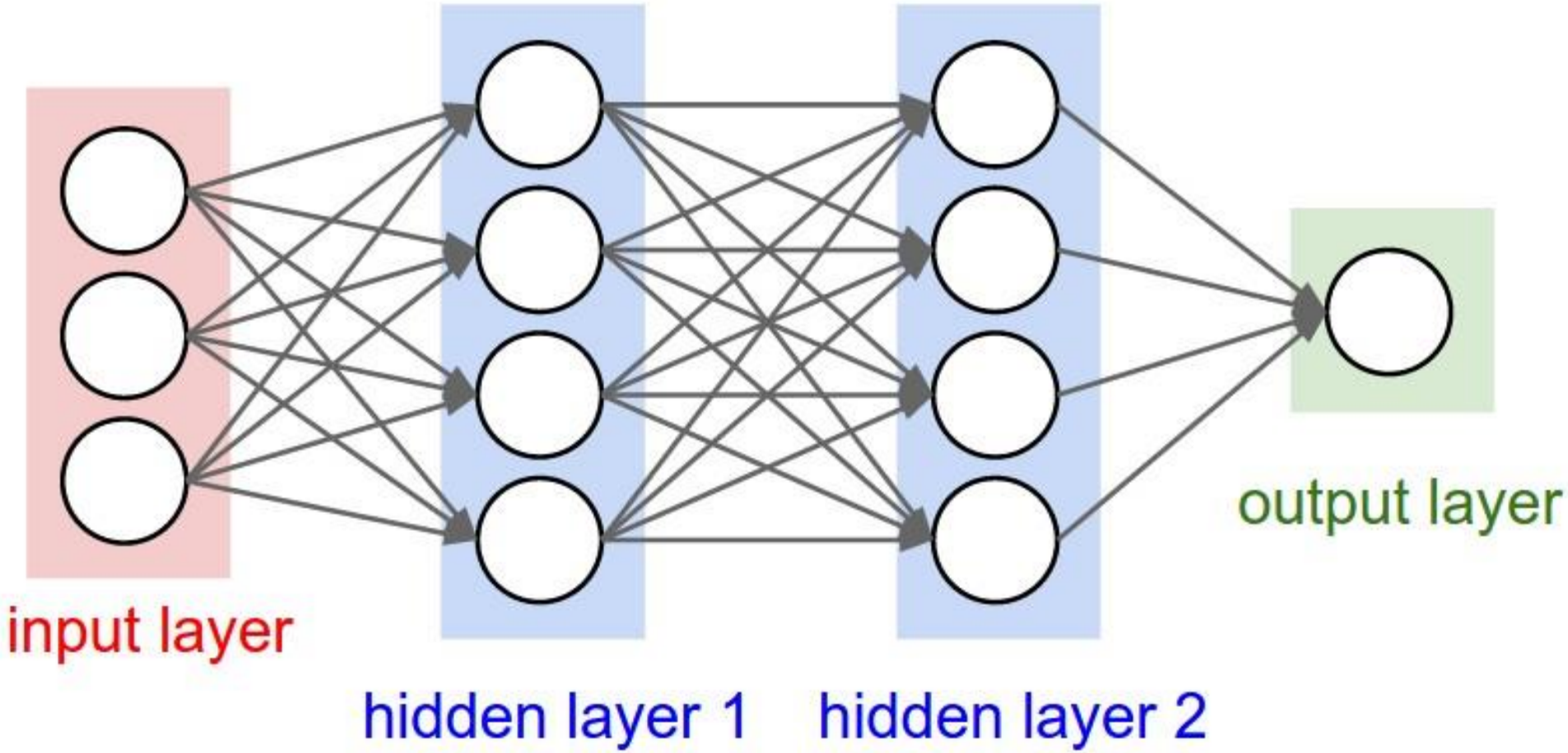
Ансамбли решающих деревьев



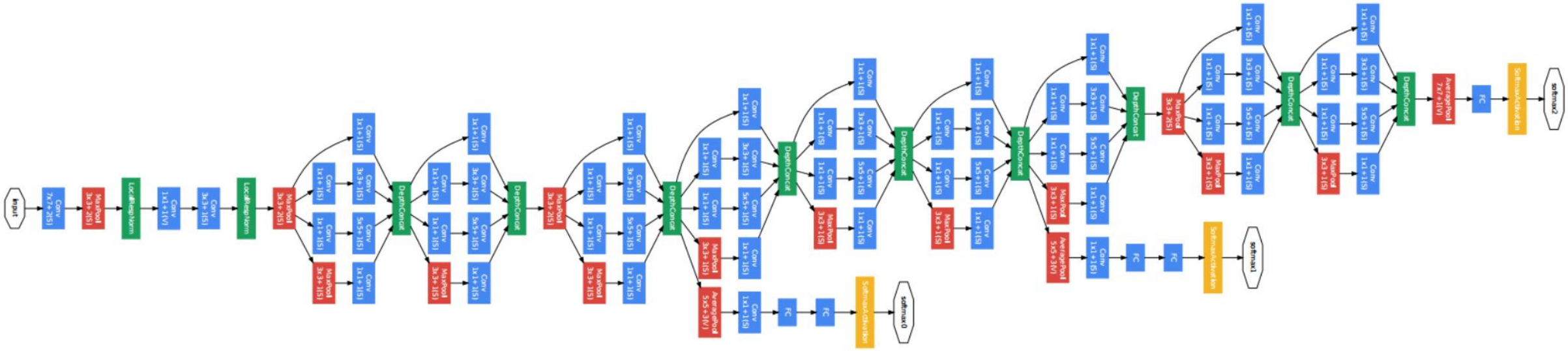
Нейронные сети



Нейронные сети



Нейронные сети



GoogLeNet

IV. Задачи оптимизации в машинном обучении

Задача регрессии

$x_1, x_2, \dots, x_l$ - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

Задача регрессии

$x_1, x_2, \dots, x_l$ - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

Например, это:

$$\sum_{i=1}^{l} \left(y_i - a(x_i)\right)^2 \to min$$

Задача регрессии

$x_1, x_2, \dots, x_l$ - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

В общем случае:

$$\sum_{i=1}^{l} L\big(y_i, a(x_i)\big) \quad \to min$$

Задача классификации

$x_1, x_2, \ldots, x_l$ - объекты, для которых известны их классы:

$$y_1, y_2, \ldots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i = a(x_i)$$

Как выразить то, что он должен угадывать класс как можно чаще?

Задача классификации

$x_1, x_2, \dots, x_l$ - объекты, для которых известны их классы:

$$y_1, y_2, \dots, y_l$$

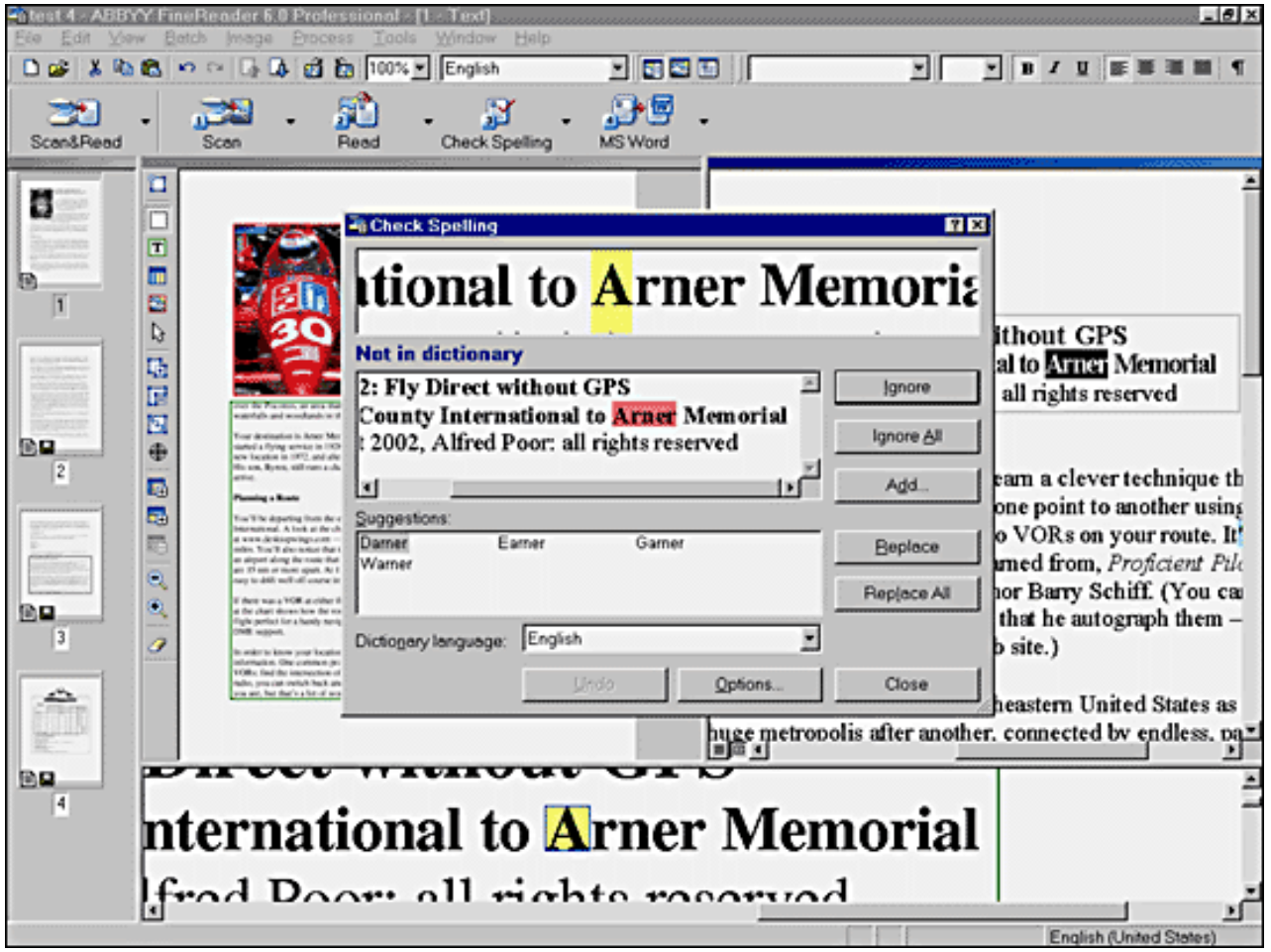
Мы строим прогнозирующий алгоритм:

$$y_i = a(x_i)$$

Как выразить то, что он должен угадывать класс как можно чаще?

$$\sum_{i=1}^{l} [y_i \neq a(x_i)] \to min$$

Сложный пример: исправление опечаток



Сложный пример: исправление опечаток

$$Suggest(w) = [w_1, w_2, \dots, w_k]$$

В алгоритме есть параметры, которые когда-то были заданы «вручную». Хочется настроить их так, чтобы suggest был как можно «адекватней».

Есть выборка:
w (слово с опечаткой), cw(правильное написание)

Как сформулировать «адекватность» suggest’а, как настроить параметры?

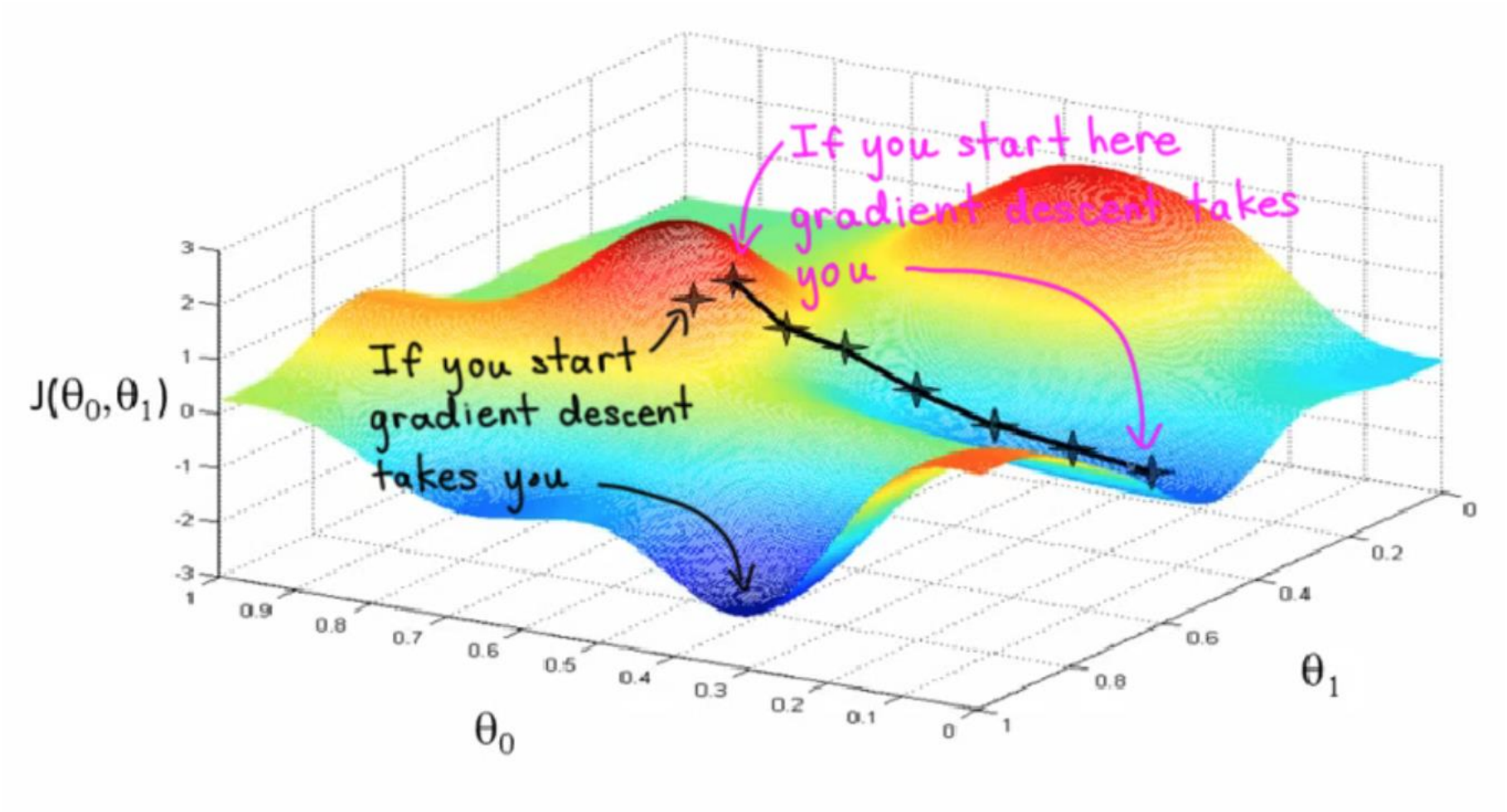
Сложный пример: исправление опечаток

Возможное решение:

$$Suggest(w) = [w_1, w_2, \dots, w_k]$$
$$Pos\big(w_j, [w_1, w_2, \dots, w_k]\big) = j$$

$$\sum_{i=1}^{l} Pos(cw_i, Suggest(w_i)) \to min$$

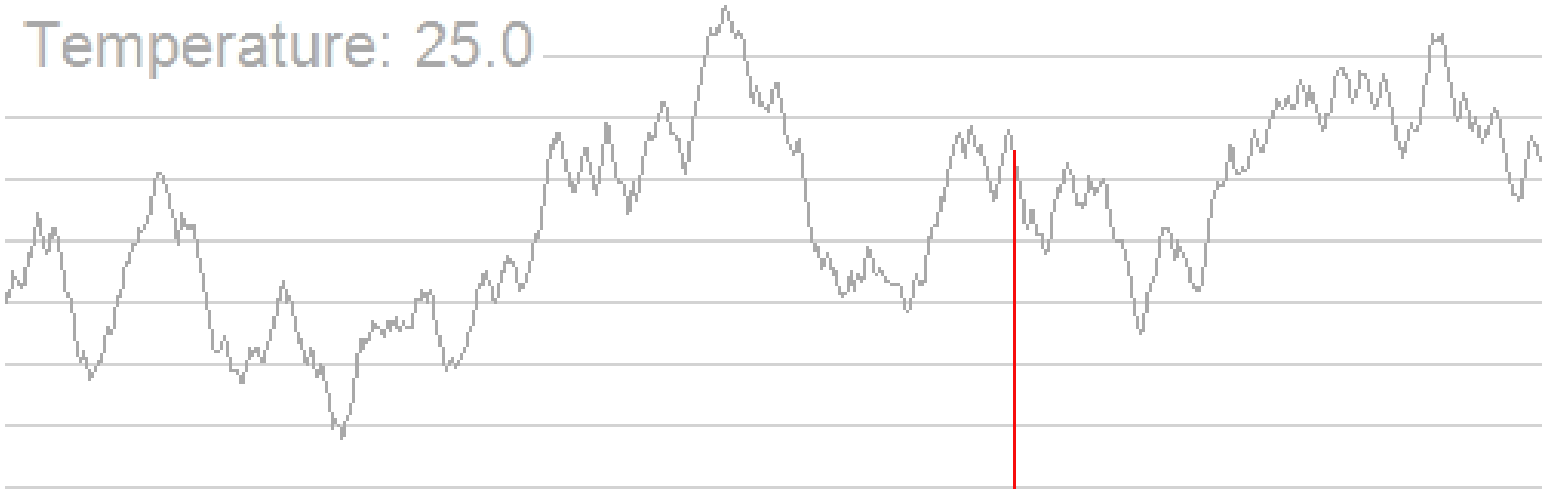
Градиентные методы оптимизации



Методы глобальной оптимизации

$$P(\overline{x^*} \to \overline{x_{i+1}} \mid \overline{x_i}) = \left\{ \begin{array}{ll} 1, & F(\overline{x^*}) - F(\overline{x_i}) < 0 \\ \exp\left(-\dfrac{F(\overline{x^*}) - F(\overline{x_i})}{Q_i}\right), & F(\overline{x^*}) - F(\overline{x_i}) \geqslant 0 \end{array} \right\}.$$

Методы глобальной оптимизации

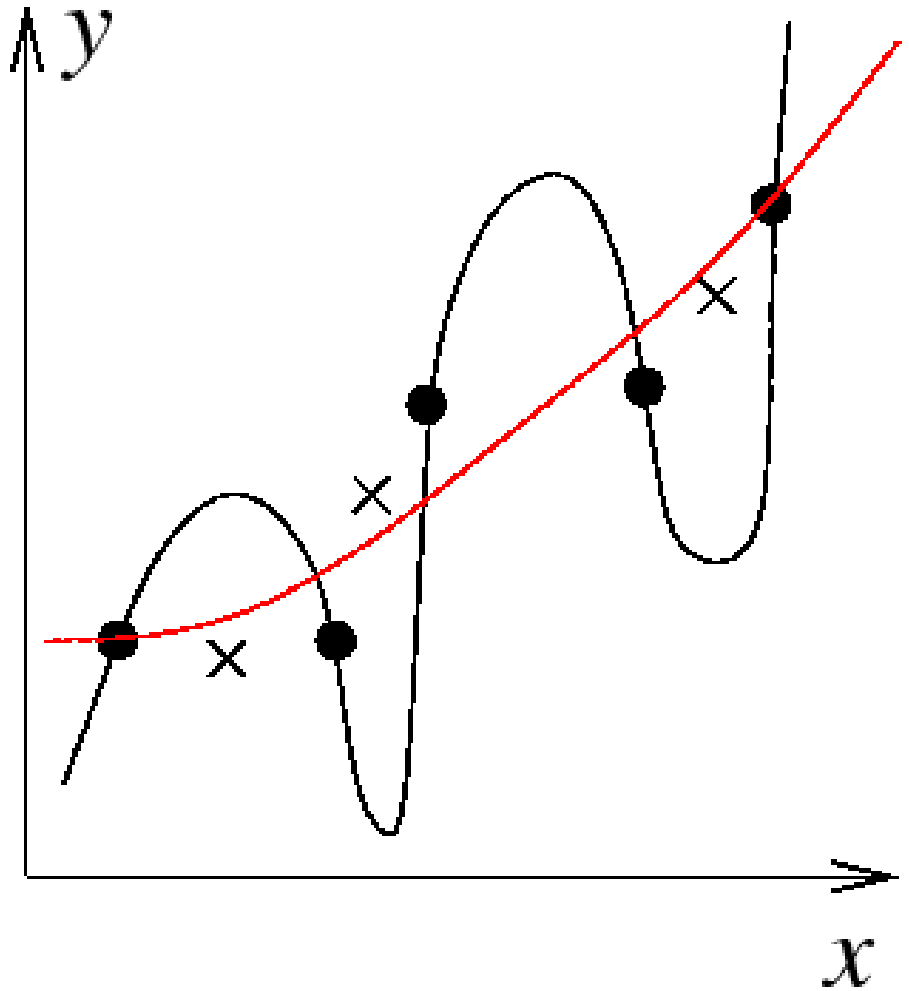


$$P(\overline{x^*} \to \overline{x_{i+1}} \mid \overline{x_i}) = \begin{cases} 1, & F(\overline{x^*}) - F(\overline{x_i}) < 0 \\ \exp\left(-\dfrac{F(\overline{x^*}) - F(\overline{x_i})}{Q_i}\right), & F(\overline{x^*}) - F(\overline{x_i}) \geqslant 0 \end{cases}.$$

https://ru.wikipedia.org/wiki/Алгоритм_имитации_отжига

V. Переобучение и недообучение

Переобучение на примере регрессии



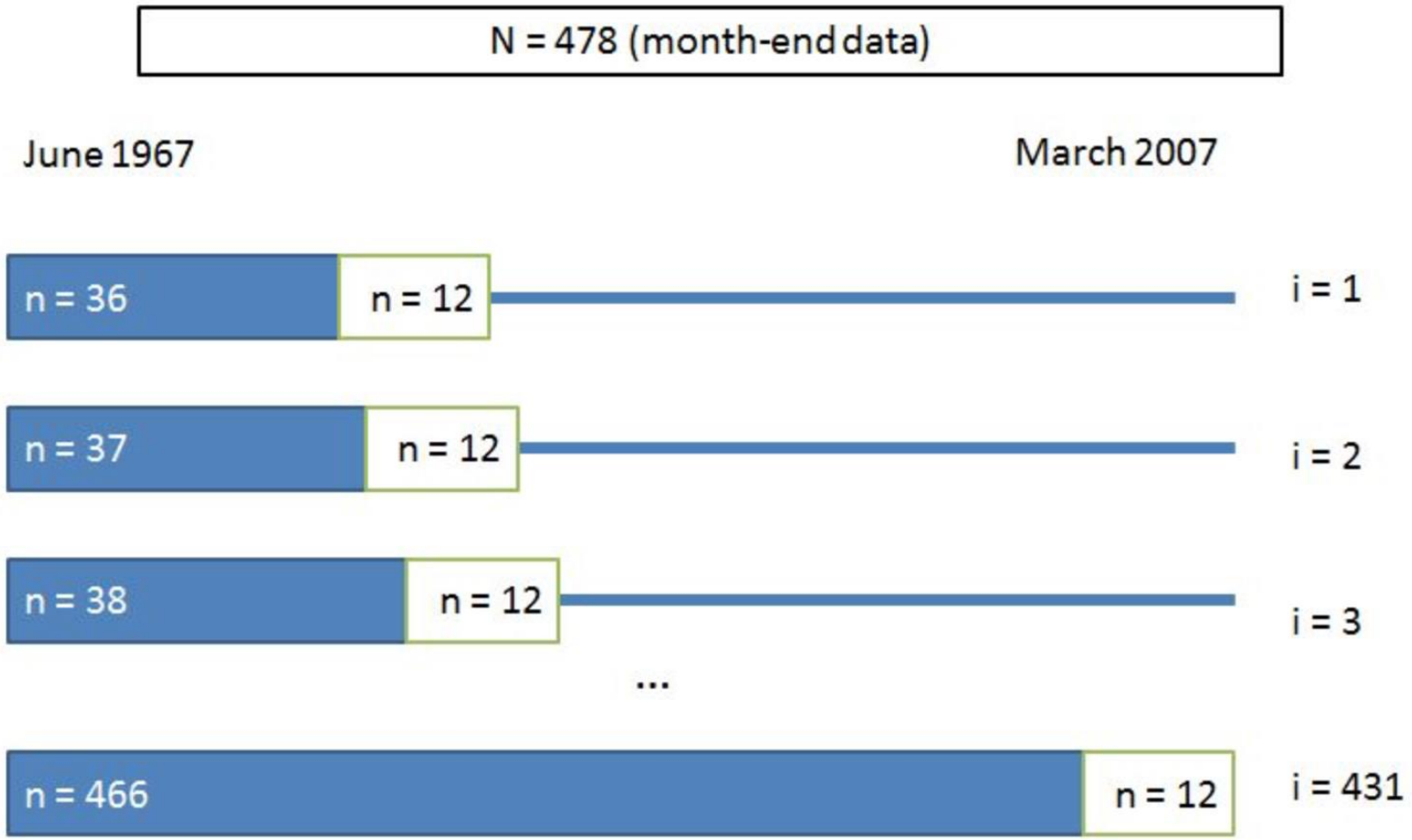
Оценка качества

Кросс-валидация

K-Fold cross validation:

На картинке k = 5, обычно такое k и используют. Другие частые варианты – 3 и 10.

Кросс-валидация и данные «из будущего»



История про танки

Классификатор: есть танки на снимке или нет

История про танки

Классификатор: есть танки на снимке или нет

Задача

Для некоторой задачи построили алгоритм обучения с учителем и он работает очень плохо

А) Как понять, проблема в недостаточном размере обучающей выборки или в чем-то еще?

Б) В чем еще может быть проблема?