

Машинное обучение

Лекция 1. Основные понятия и примеры



Команда курса



Эмели Драль
Chief DS YDF



Виктор Кантор
Chief DS Я.Такси



Александр Гущин
Senior DS Я.Такси



Илья Ирхин
Lead DS Я.Такси

+ приглашенные преподаватели

Лектор



Работа:

Chief Data Scientist Яндекс.Такси, до этого руководил подразделениями в Yandex Data Factory и ABBYY



Преподавание:

Coursera, МФТИ, ШАД, Яндекс, Яндекс.Такси, ABBYY, Мегафон, МТС, СберТех, РайффайзенБанк и др.

Основатель Applied Data Science Center

Содержание курса (часть 1)

1. Введение: основные понятия и задачи, простые методы
2. Линейные модели в задачах классификации и регрессии
3. Вероятностный взгляд на обучение с учителем
4. Решающие деревья и ансамбли
5. Оценка качества на исторических данных
6. Оценка качества в онлайне
7. Тренинг по постановке задач
8. Резерв на повторение материала

Содержание курса (часть 2)

9. Кластеризация, semi-supervised learning и look-alike модели
10. Работа с признаками: извлечение, отбор, преобразование
11. Основы анализа текстов
12. Рекомендательные системы
13. Предиктивная аналитика
14. Анализ социальных сетей
15. Нейронные сети и глубокое обучение
16. Резерв

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	0	1	1
1	42	2	79	1	4	3	4	2	45	3	1	2	1	1	0	0	0	0	0	0	0	0	0	1
1	24	3	49	1	3	3	4	4	53	3	2	2	1	1	1	0	1	0	0	0	0	0	0	2
4	36	2	91	5	3	3	4	4	35	3	1	2	2	1	0	0	1	0	0	0	0	0	1	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	0	1	0	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	0	1	0	0	0	2
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	1	0	0	1	0	0	0	2
2	12	2	16	1	3	2	1	3	22	3	1	1	2	1	0	0	1	0	0	0	1	0	0	1
1	24	4	12	1	5	3	4	3	60	3	2	1	1	1	1	0	1	0	0	1	0	0	1	2
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	1	0	0	1	0	0	0	1
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	0	0	1	0	0	0	1	0	0	2
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	0	0	1	0	0	0	1	0	0	1
1	30	0	81	5	2	3	3	3	25	1	3	1	1	1	1	0	1	0	0	1	0	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	0	1	0	0	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	0	1	0	0	1	1
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	1	0	0	1	0	0	1	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	0	1	0	0	1
1	6	2	14	1	3	3	2	1	36	1	1	1	2	1	0	0	1	0	0	0	1	0	0	1
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	1	0	0	1	0	0	0	1
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	0	0	0	0	1	0	0	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	60	1	3	2	2	1	22	3	1	1	1	1	0	0	1	0	0	0	0	0	1	1	
4	12	2	30	2	12	1	48	2	12	1	24	1	15	1	24	4	24	1	30	2	24	4	24	4	9
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	0	1	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	0	1	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	0	1	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	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	
4	12	2	Attribute 2: Duration in month	5	3	25	3	1	1	1	1	1	1	1	1	0	0	1	0	0	1	0	0	1	1
2	30	4		1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	2	
2	12	2	13	1	2	2	1	3	25	3	1	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	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	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	0	1	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	1	0	0	1	0	0	1	0	0	2	
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	0	0	1	0	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	1
2	24	2	126	1	5	2	2	4	44	3	1	1	2	1	0	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	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	0	1	0	0	1	1	
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	1	1	0	0	1	0	0	1	
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	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	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	0	1	0	0	1	
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1	
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	1	0	0	1	0	0	0	1	
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	0	0	0	0	1	0	0	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	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	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	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	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	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	0	1	0	1
4	24	2	28	3	5	3	4	2	53	3	1	1	1	1	0	0	1	0	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	0	1
4	12	2	31	4	4	1	4	1	61	3	1	1	1	1	0	0	1	0	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	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	0	1	2
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	1	0	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	0	1	0	0	0	1
1	24	4	12	1	5	3	4	3	60	3	2	1	1	1	1	0	1	0	0	0	1	0	1	0	
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	1	0	0	0	1	0	0	1	
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	1	0	0	1	0	0	0	1	0	2	
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	1	1	1	0	0	0	1	0	1	1	
1	30	0	81	5	2	3	3	3	25	1	3	1	1	1	0	0	1	0	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	0	2
4	24	2	34	3	5	3	2	3	31	3	1	2	2	1	0	0	1	0	0	0	1	0	0	0	1
4	9	4	21	1	3	3	4	3	48	3	3	1	2	1	1	0	1	0	0	0	1	0	0	1	
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	0	0	1	0	0	1	
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	1	0	0	1	0	0	0	1	
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	1	1	0	0	0	1	0	0	1	
4	10	4	21	5	3	4	1	3	26	3	2	1	1	2	0	0	1	0	0	0	1	0	0	1	
1	6	2	14	1	3	3	2	1	36	1	1	1	2	1	0	0	1	0	0	0	1	0	0	1	
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	0	0	0	1	0	1	0	
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	1	0	0	1	0	0	0	1	
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	0	0	0	0	1	0	0	1	

Answer: 1 – Good, 2 - Bad

Выдача кредита

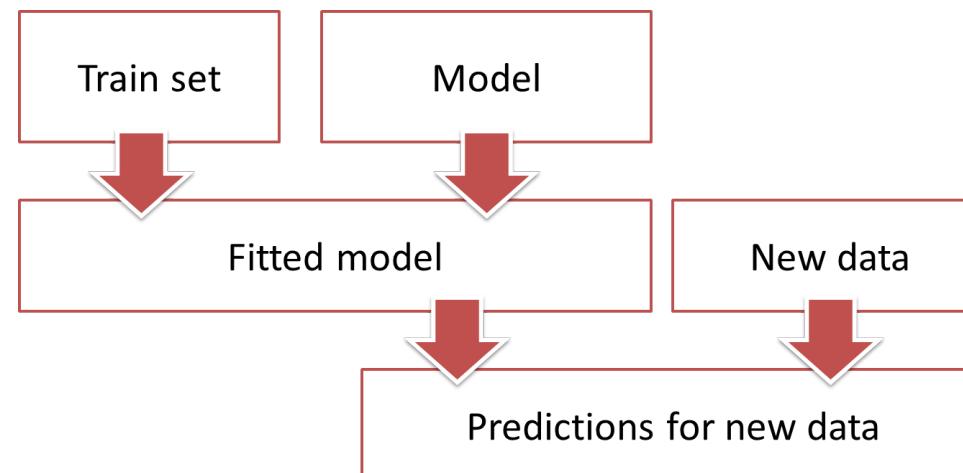
Задача (supervised classification): предсказать класс (1 или 2)

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-r
Слипоны
~~1 999 руб.~~ 1 590 руб.



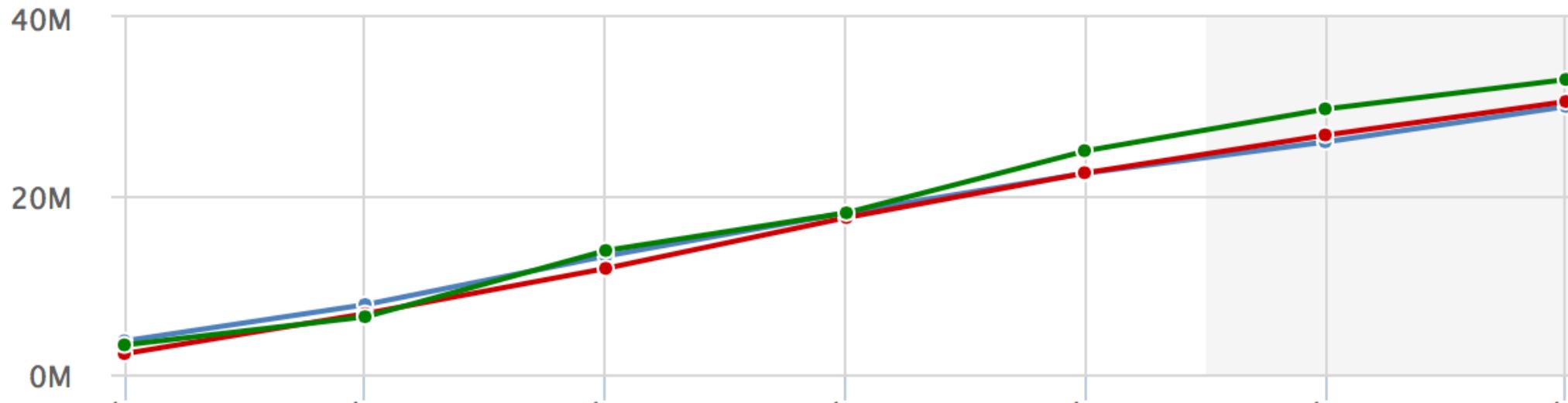
Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

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

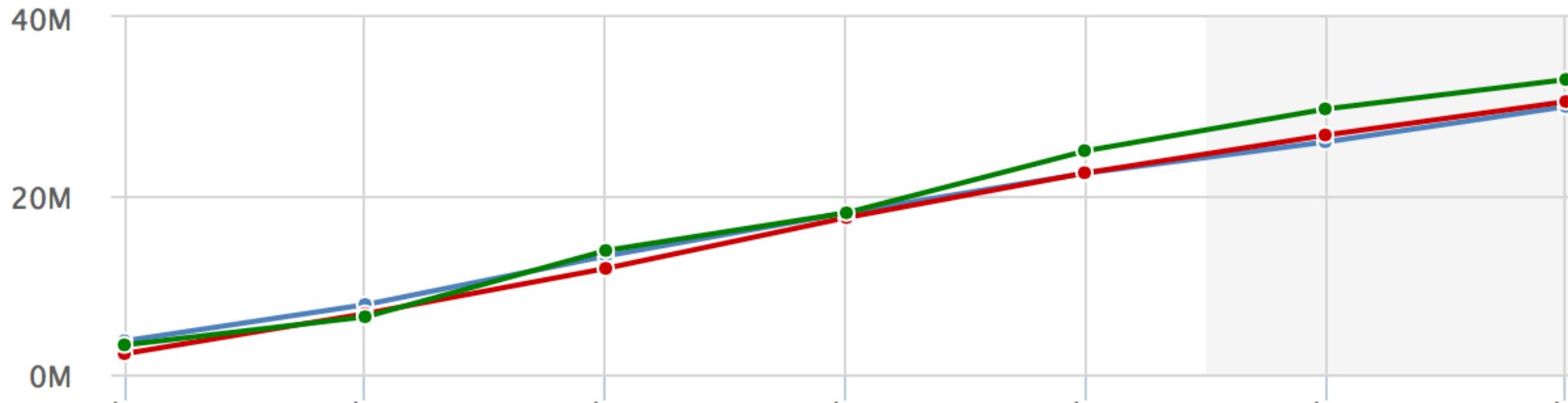
История про статзначимость

Суммарная выручка



История про статзначимость

Суммарная выручка



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?

Еще примеры

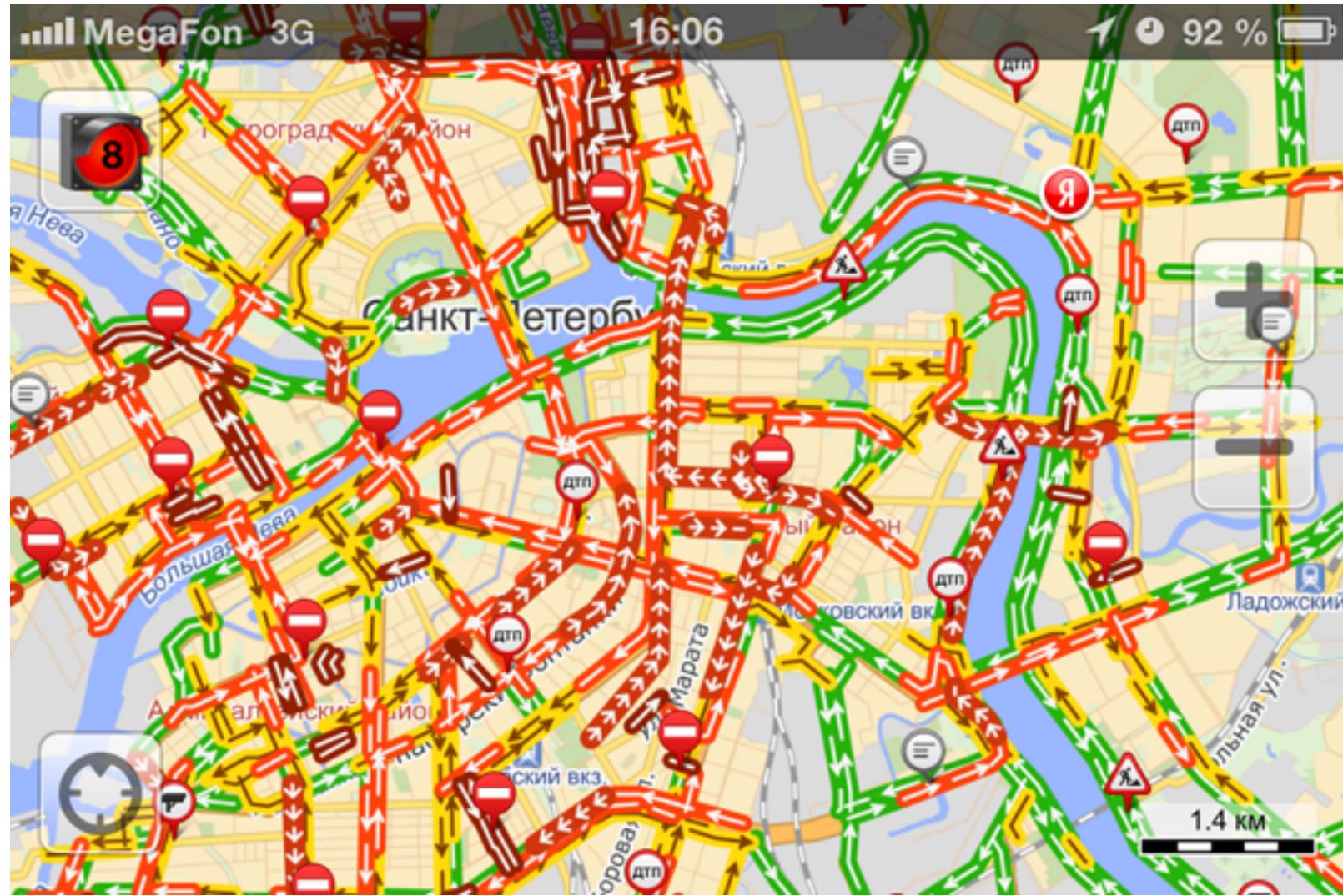
VK Data Mining in Action | ВКонтакте[vk.com > data_mining_in_action](#) ▾

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.

Нашлось 8 млн результатов[Добавить объявление](#) [Показать все](#)**H 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.

Coursera Process Mining: Data science in Action... | Coursera[coursera.org > learn/process-mining](#) ▾





II. Стандартные задачи и простые методы их решения

Классификация



Iris setosa



Iris versicolor



Iris virginica

Вход (обучающая выборка):

Признаки N объектов с известными классами

Выход:

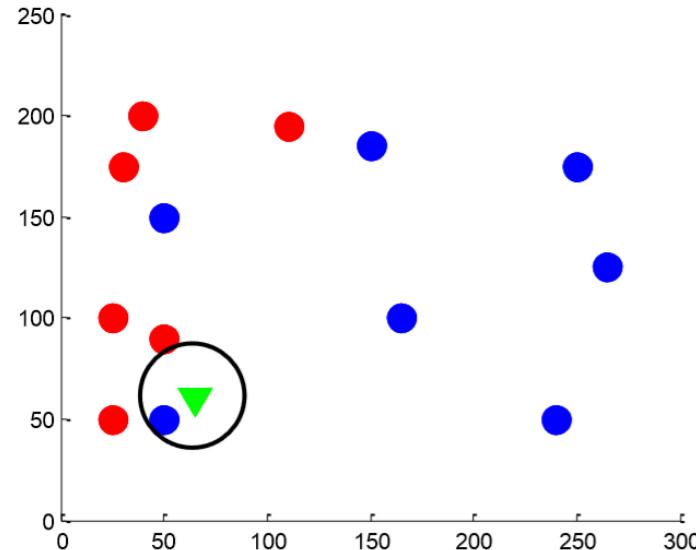
Классификатор (алгоритм, прогнозирующий классы новых объектов по их признакам)

Классификация: обучающая выборка

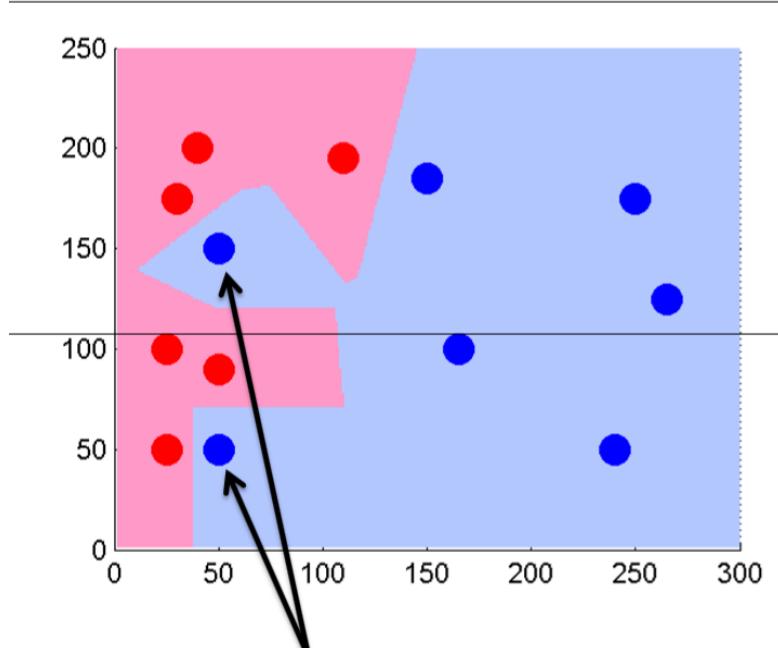
Fisher's Iris Data

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

Простой классификатор: kNN

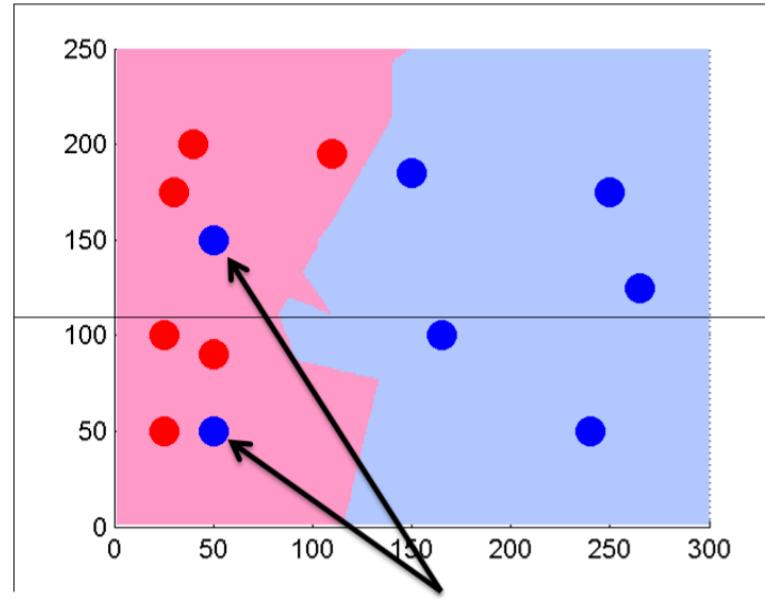
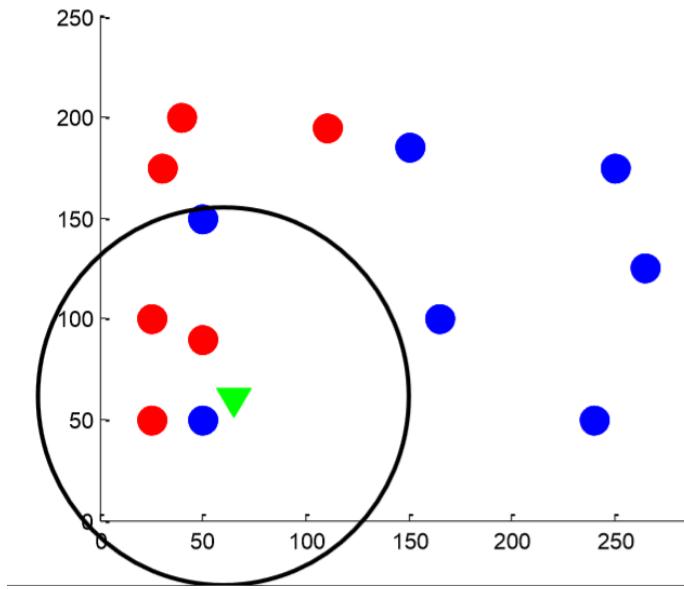


$k = 1$



Шумы? (outliers)

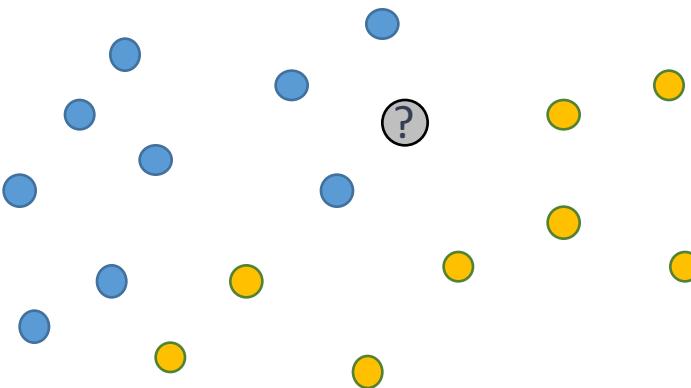
Простой классификатор: kNN



$$k = 5$$

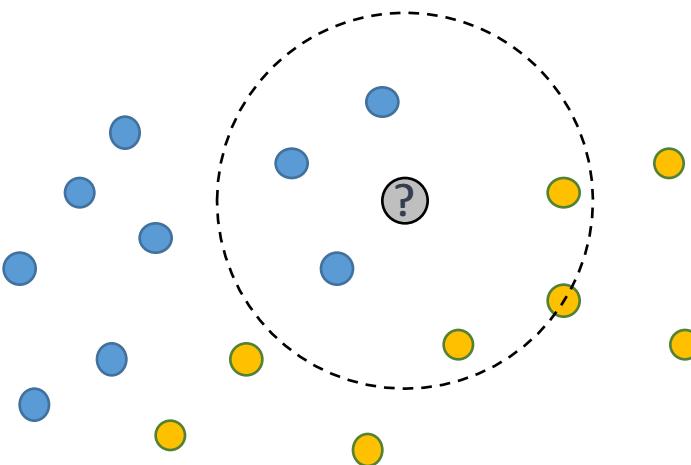
Взвешенный kNN

Пример классификации ($k = 6$):



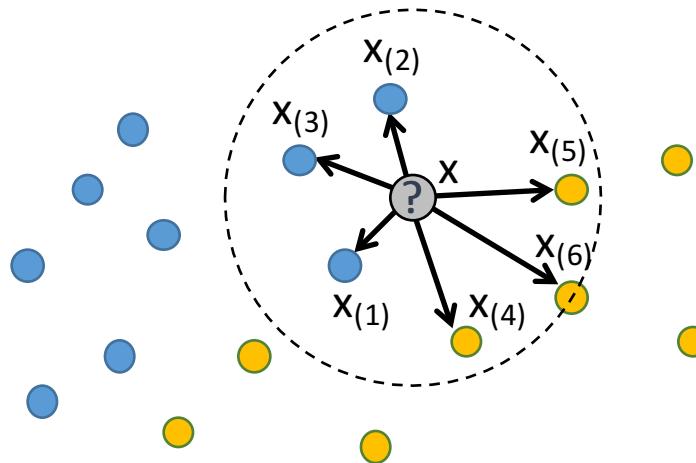
Взвешенный kNN

Пример классификации ($k = 6$):



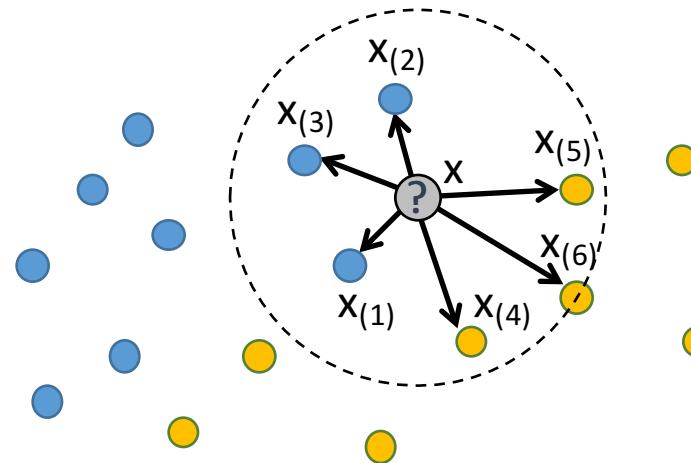
Взвешенный kNN

Пример классификации ($k = 6$):



Взвешенный kNN

Пример классификации ($k = 6$):

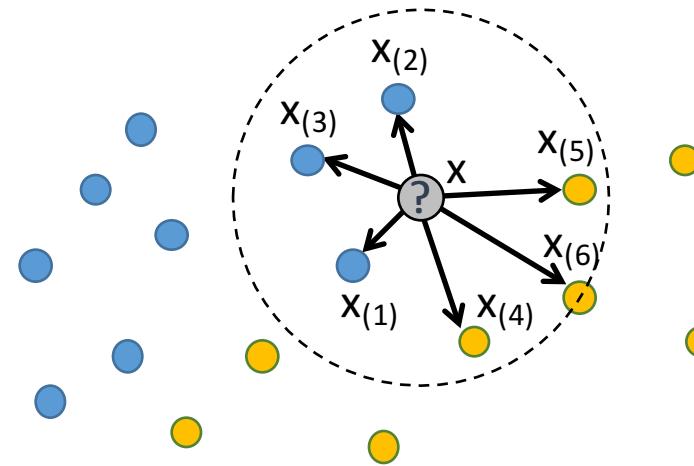


Веса можно определить как функцию от соседа или его номера:

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

Взвешенный kNN

Пример классификации ($k = 6$):



Веса можно определить как функцию от соседа или его номера:

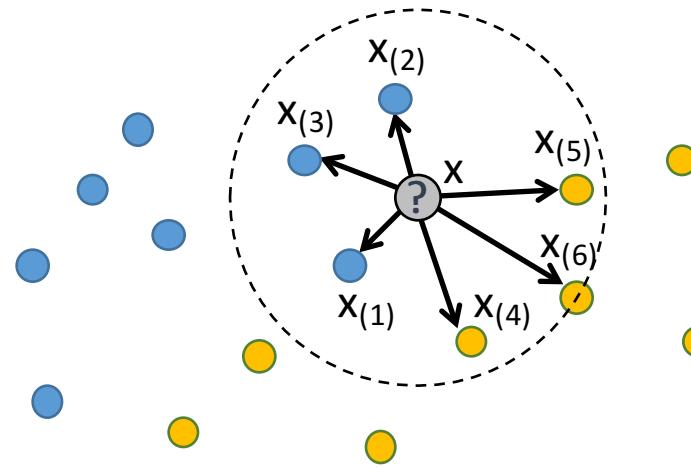
$$w(x_{(i)}) = w(i)$$

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:

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

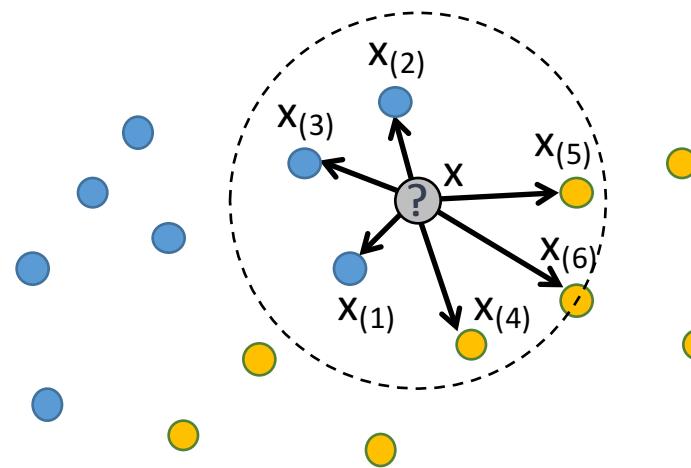
или как функцию расстояния:

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

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

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:

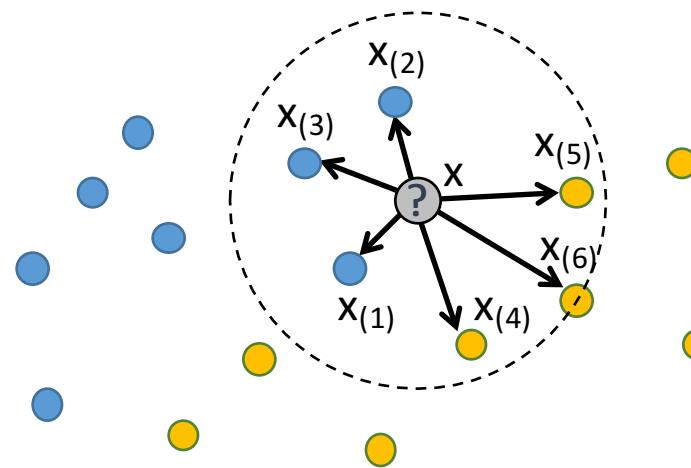
$$w(x(i)) = w(d(x, x_{(i)}))$$

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

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

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:

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

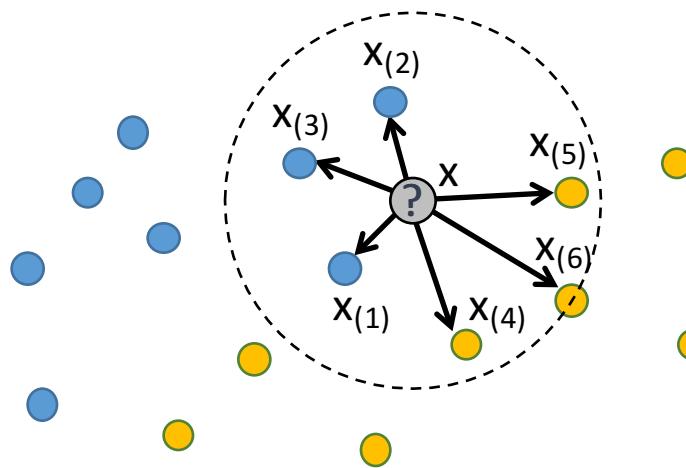
$$Z_{\text{blue}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$\text{?} = \operatorname{argmax}_{\text{grey}} Z_{\text{grey}}$$

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

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:

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

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

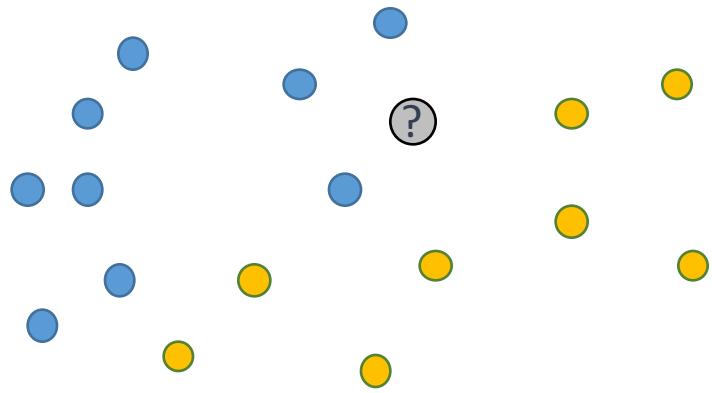
$$Z_{\text{yellow}} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$\text{?} = \operatorname{argmax}_{\circlearrowleft} Z_{\circlearrowleft}$$

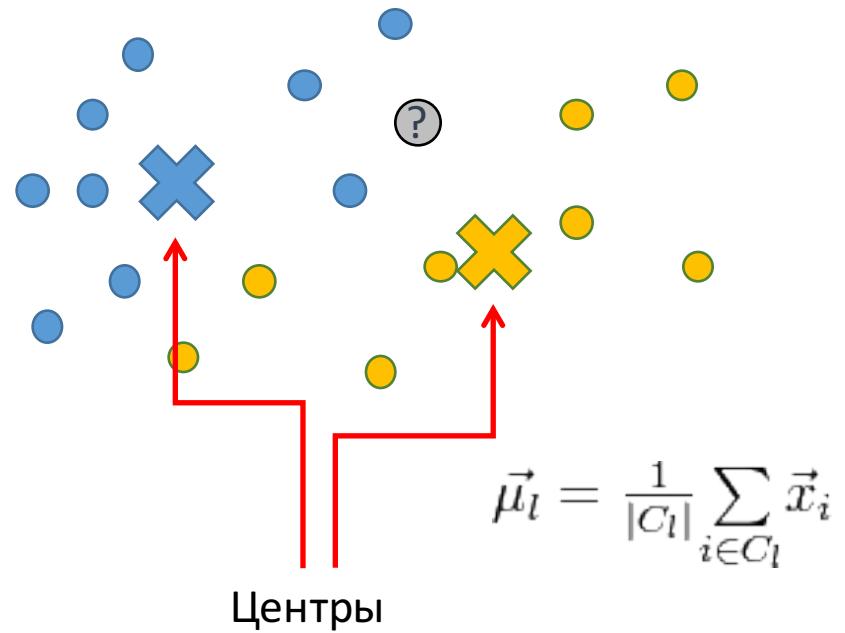
$$\text{if } Z_{\text{yellow}} > Z_{\text{blue}} : \quad \text{?} = \text{yellow}$$

$$\text{if } Z_{\text{yellow}} < Z_{\text{blue}} : \quad \text{?} = \text{blue}$$

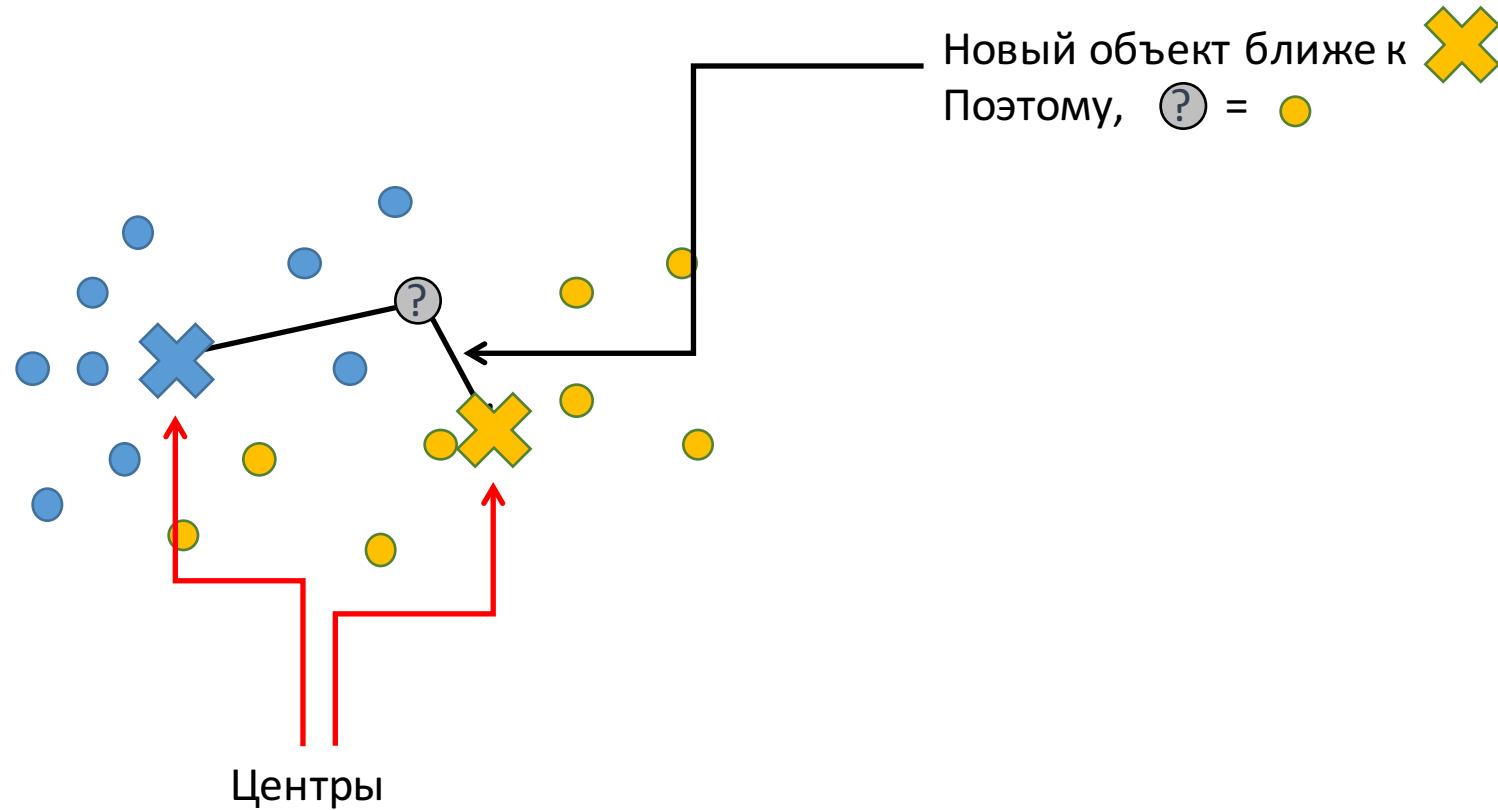
Центроидный классификатор



Центроидный классификатор



Центроидный классификатор



Кластеризация

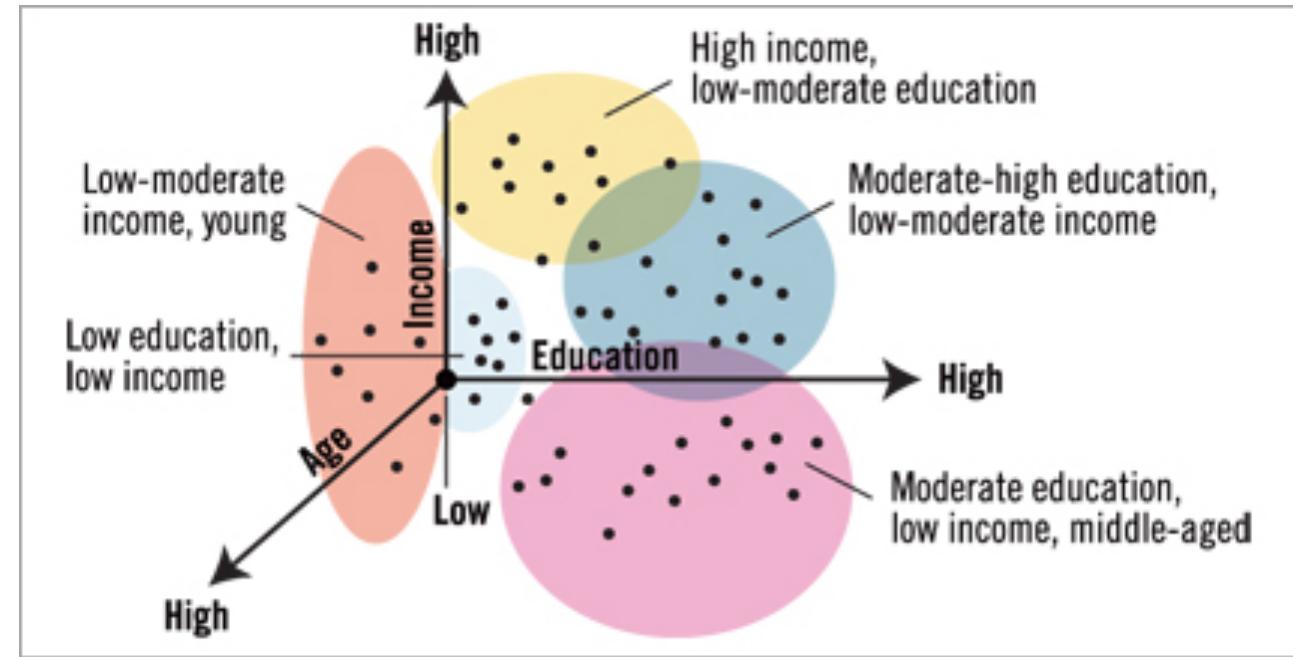
Вход (обучающая выборка):

Признаки N объектов

Выход:

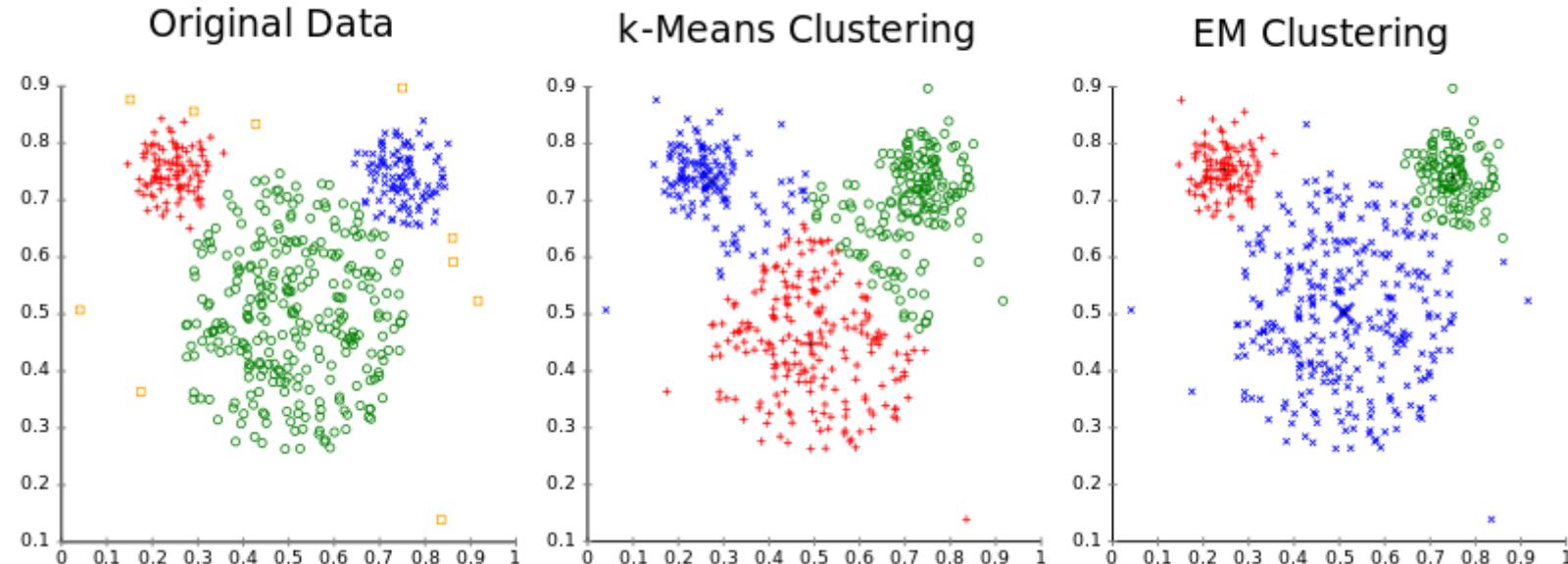
Найденные в выборке классы (кластеры), метки кластеров для объектов из обучающей выборки и алгоритм отнесения новых объектов к кластеру

Пример: сегментация рынка

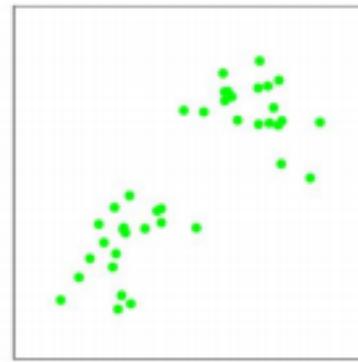


Кластеризация

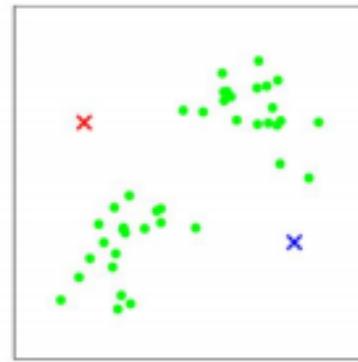
Different cluster analysis results on "mouse" data set:



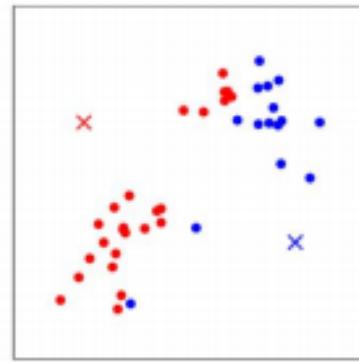
Простой алгоритм кластеризации: kMeans



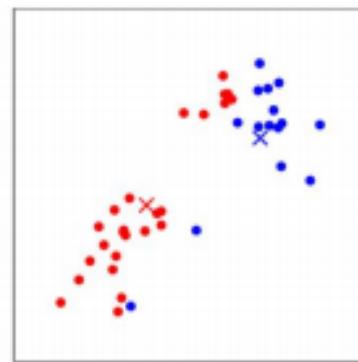
(a)



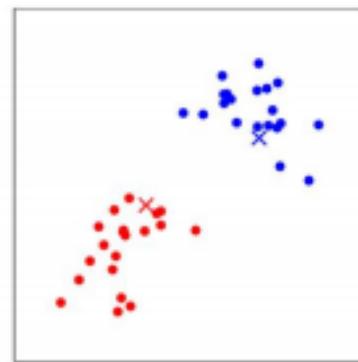
(b)



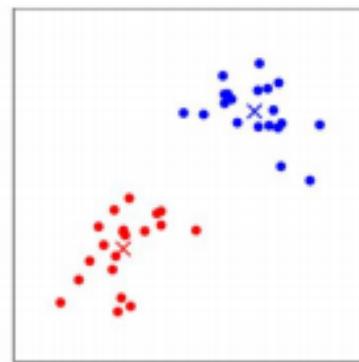
(c)



(d)



(e)



(f)

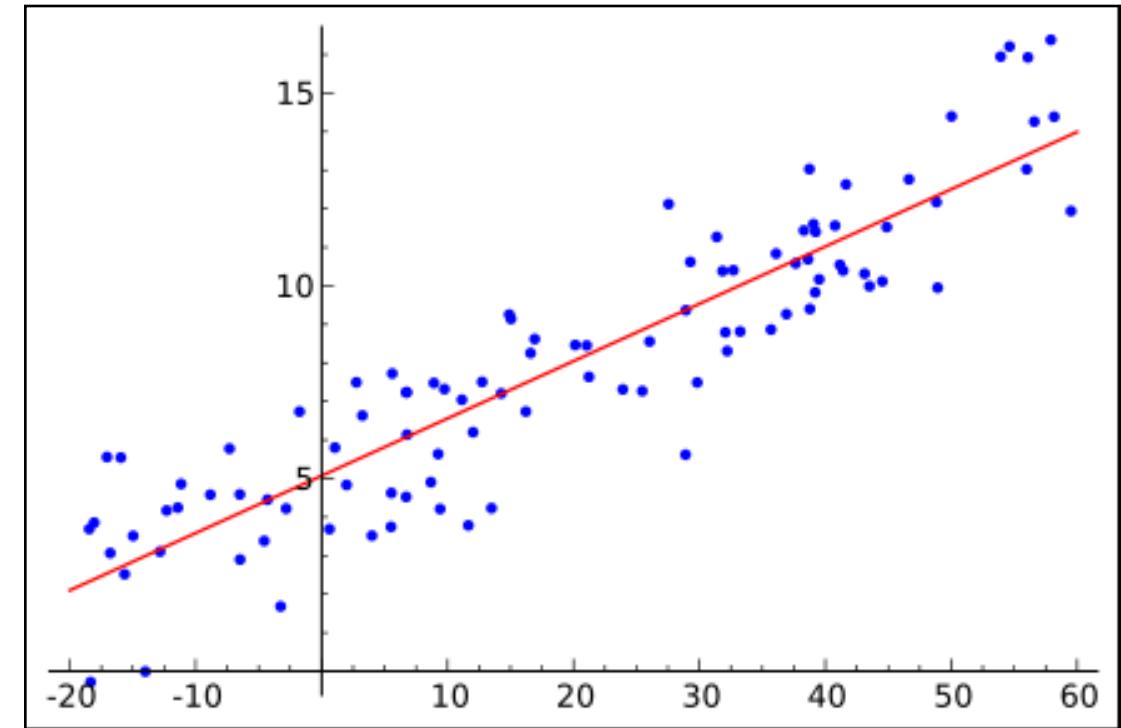
Регрессия

Вход (обучающая выборка):

Признаки N объектов с известными значениями прогнозируемого вещественного параметра объекта

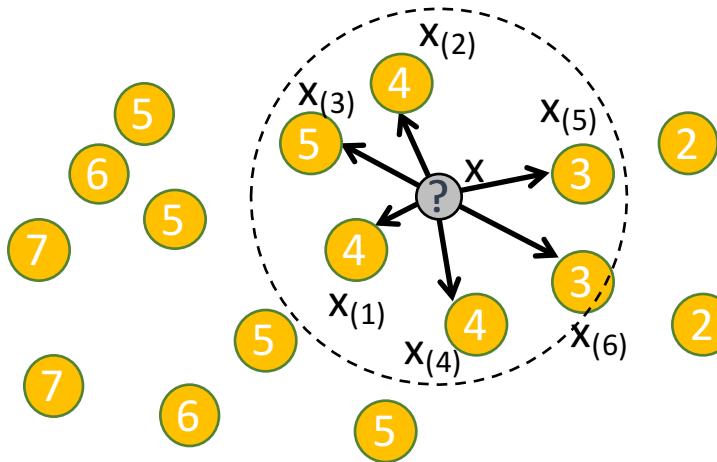
Выход:

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:

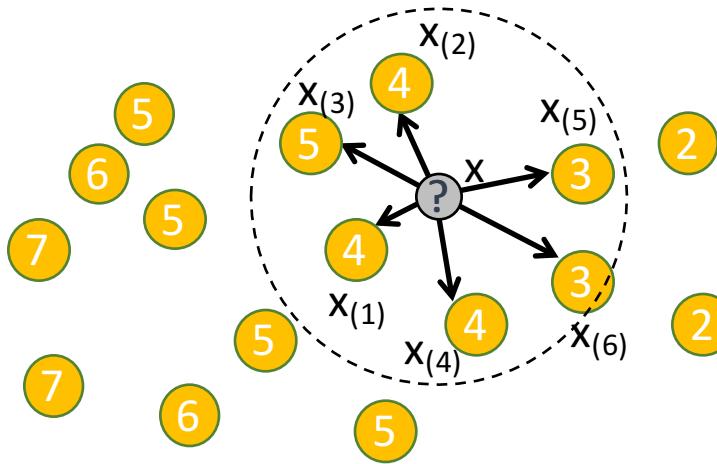
$$w(x_{(i)}) = w(i)$$

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:

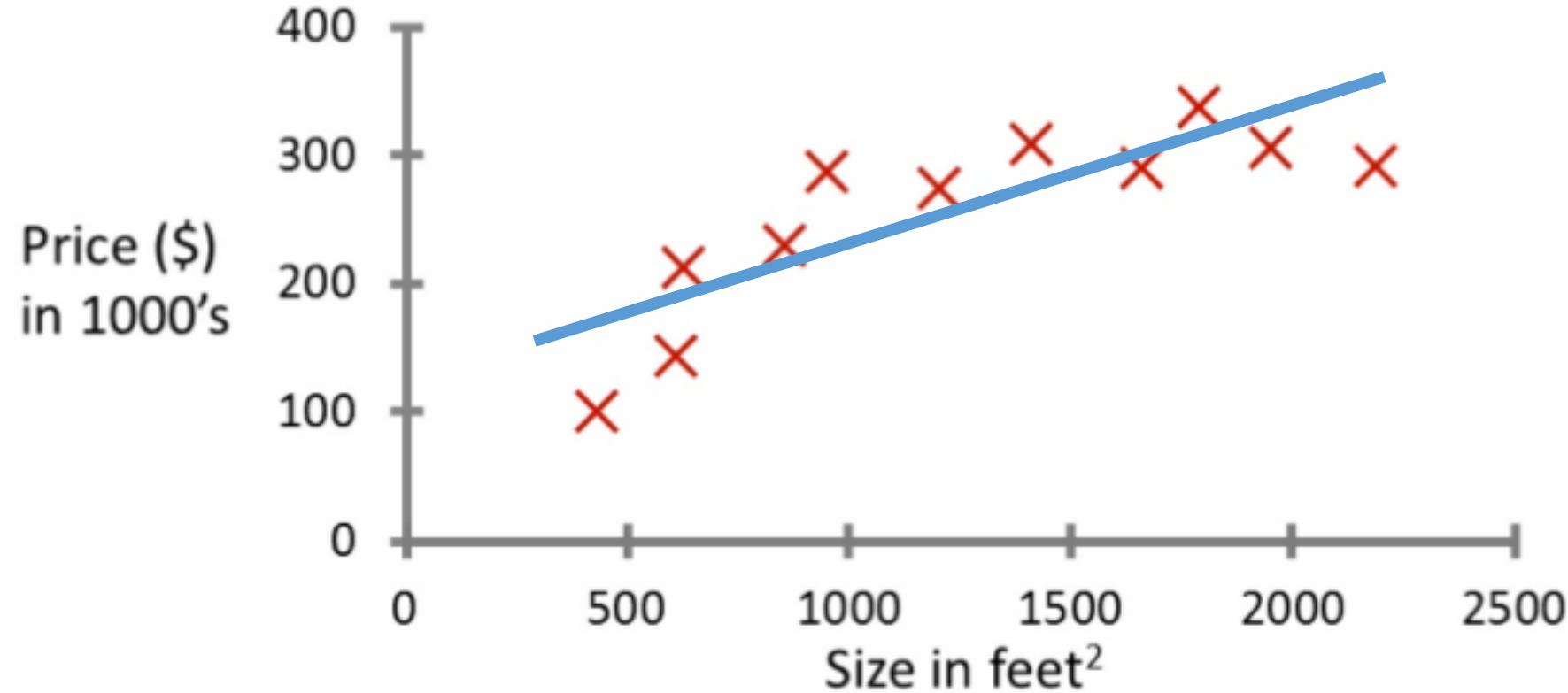
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

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

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

Линейная регрессия



Линейная регрессия

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

Линейная регрессия

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

Если добавить $x_{i0} = 1$:

Линейная регрессия

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

Если добавить $x_{i0} = 1$:

$$y_i \approx \hat{y}_i = \langle w, x_i \rangle$$

$$y_1 \approx \hat{y}_1 = x_1^T w$$

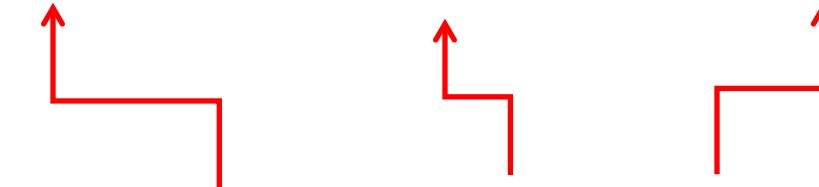
...

$$y_i \approx \hat{y}_i = x_i^T w$$

...

$$y_l \approx \hat{y}_l = x_l^T w$$

Матричная запись

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_l \end{pmatrix} \approx \begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \dots \\ \widehat{y}_l \end{pmatrix} = \begin{pmatrix} {x_1}^T \\ {x_2}^T \\ \dots \\ {x_l}^T \end{pmatrix} w$$

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

$$w = \operatorname{argmin}_w \|y - \hat{y}\|^2$$

Веса признаков

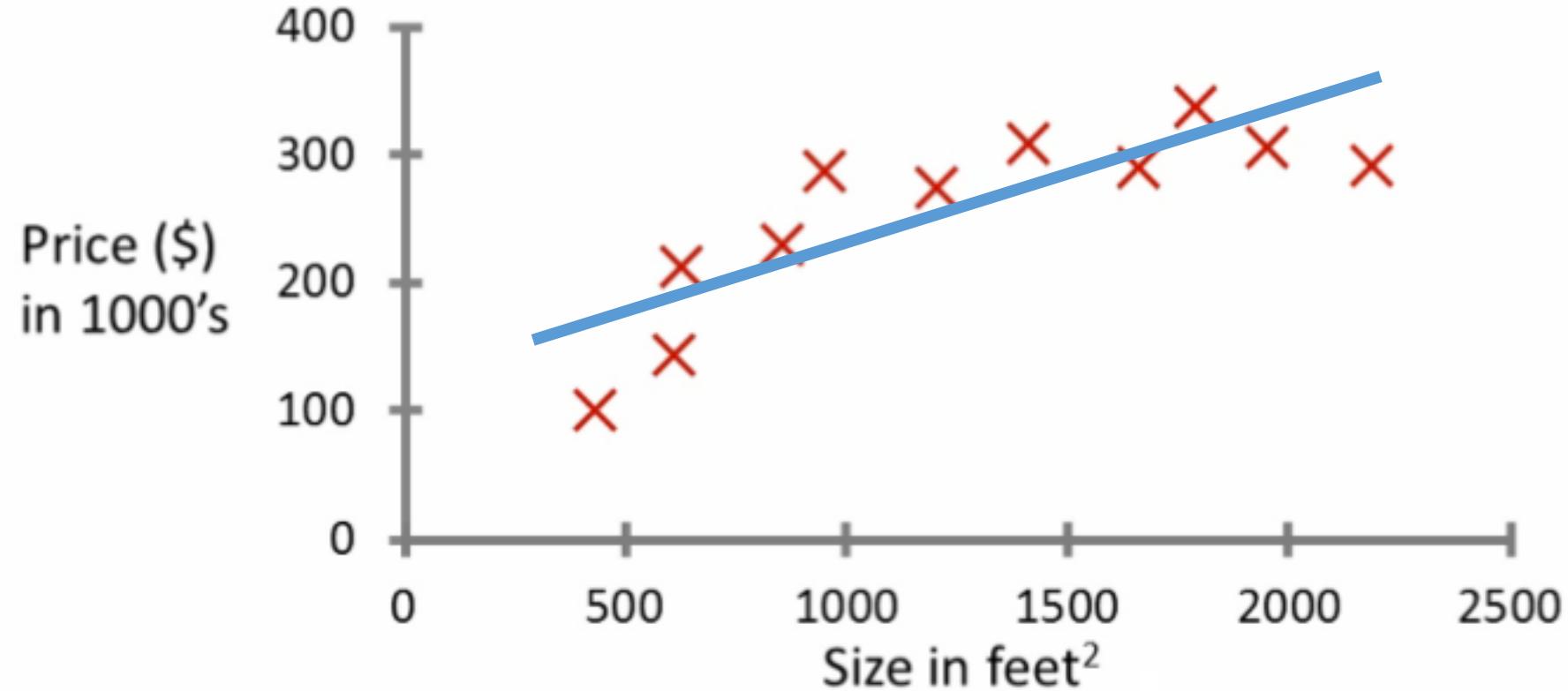
$$\frac{\partial(y - Fw)^2}{\partial w} = 2F^T(y - Fw) = 0$$

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

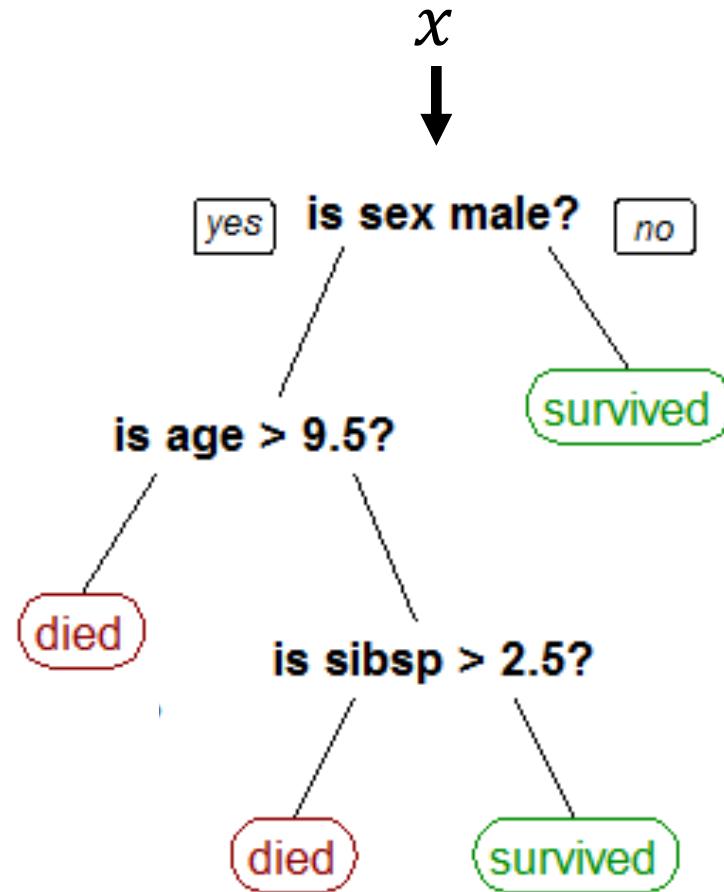
$$w = (F^T F)^{-1} F^T y$$

III. Идеи часто используемых методов

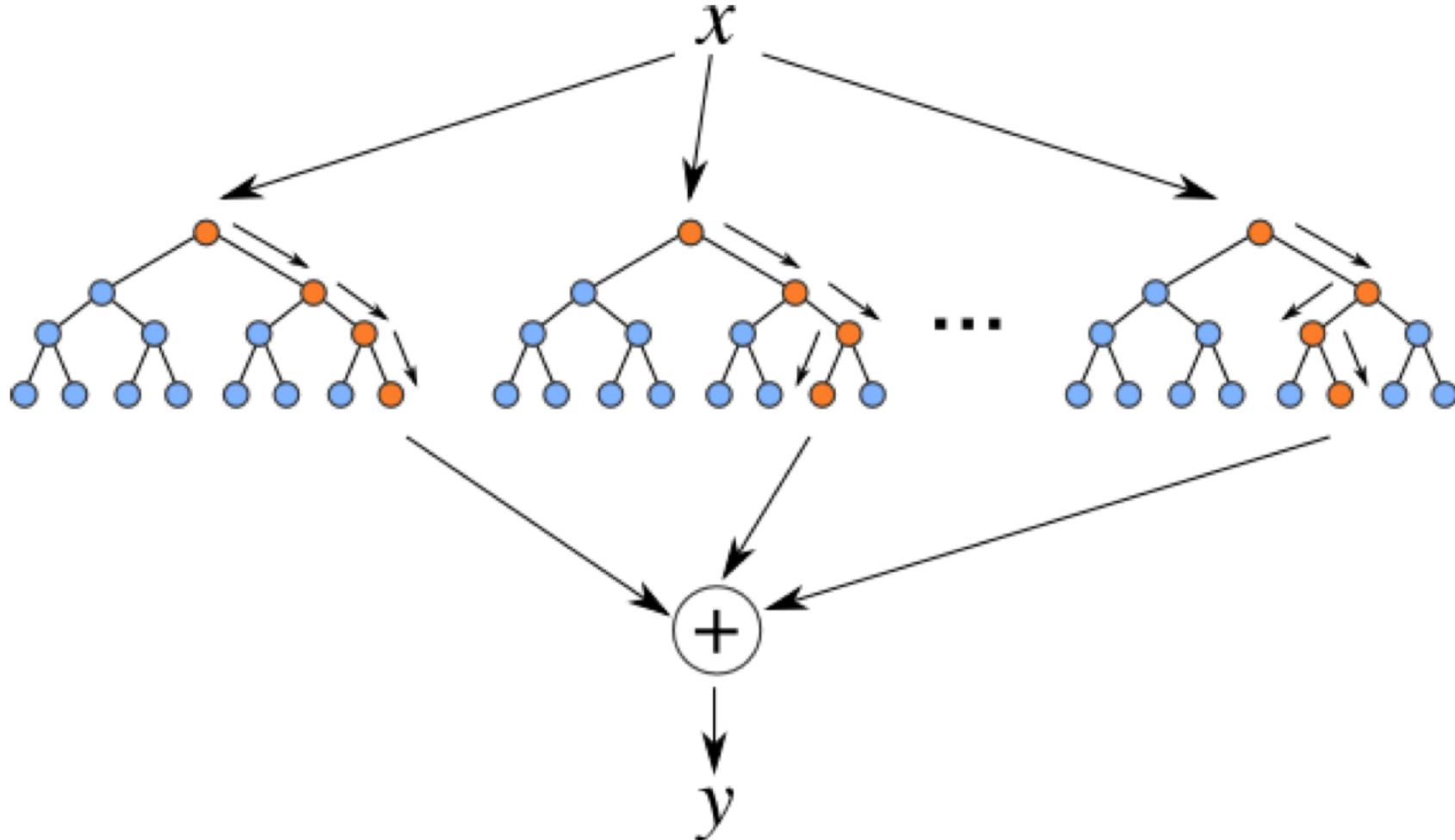
Линейные модели



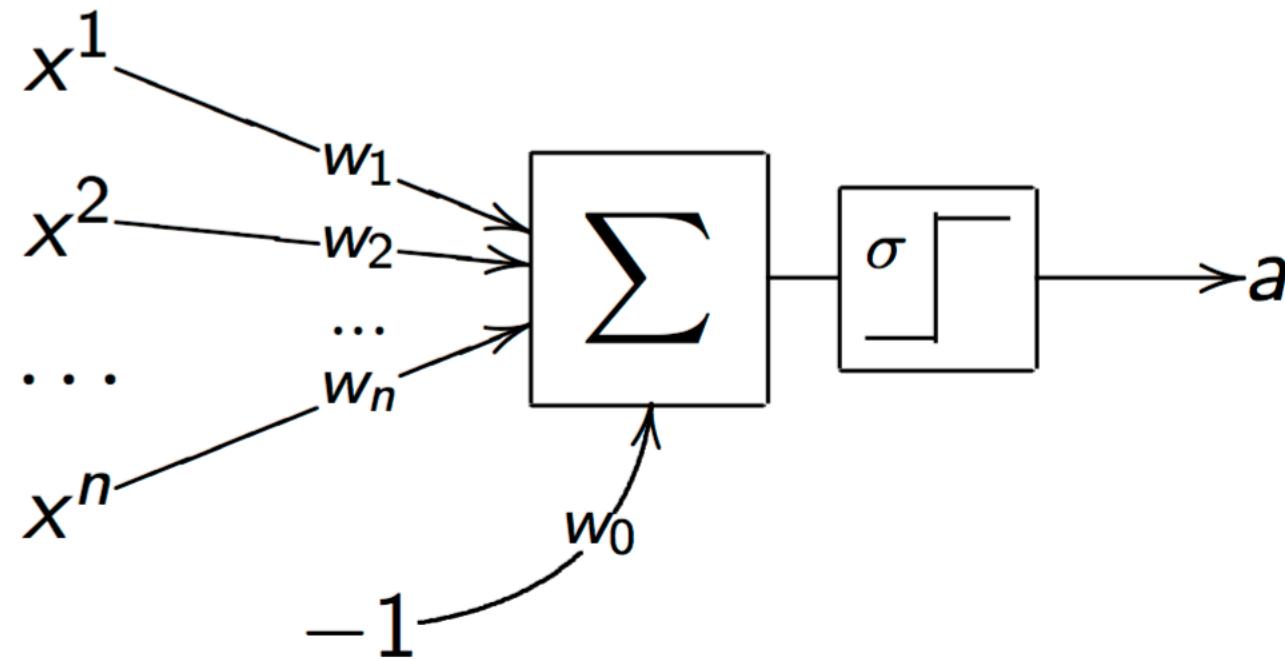
Решающие деревья



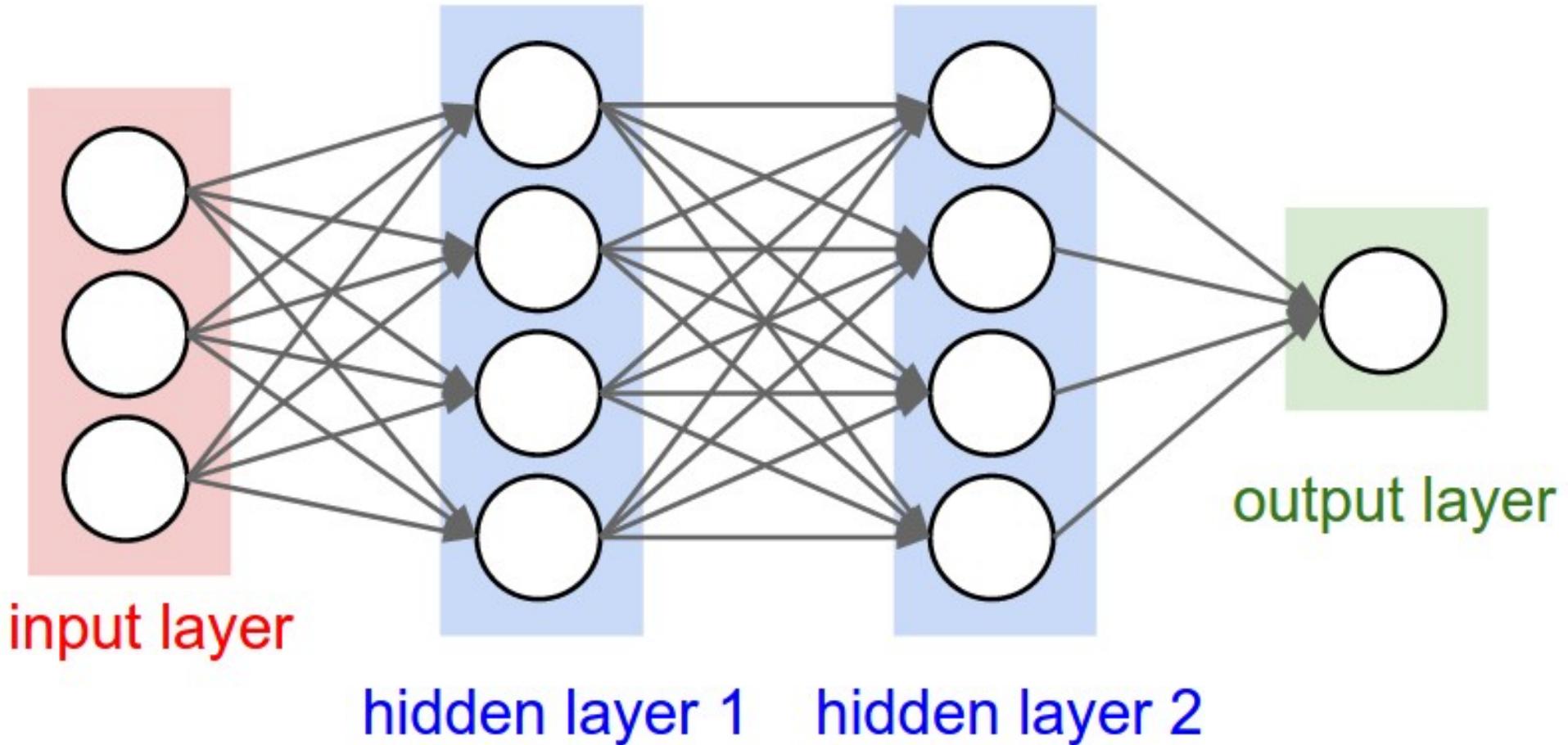
Ансамбли решающих деревьев



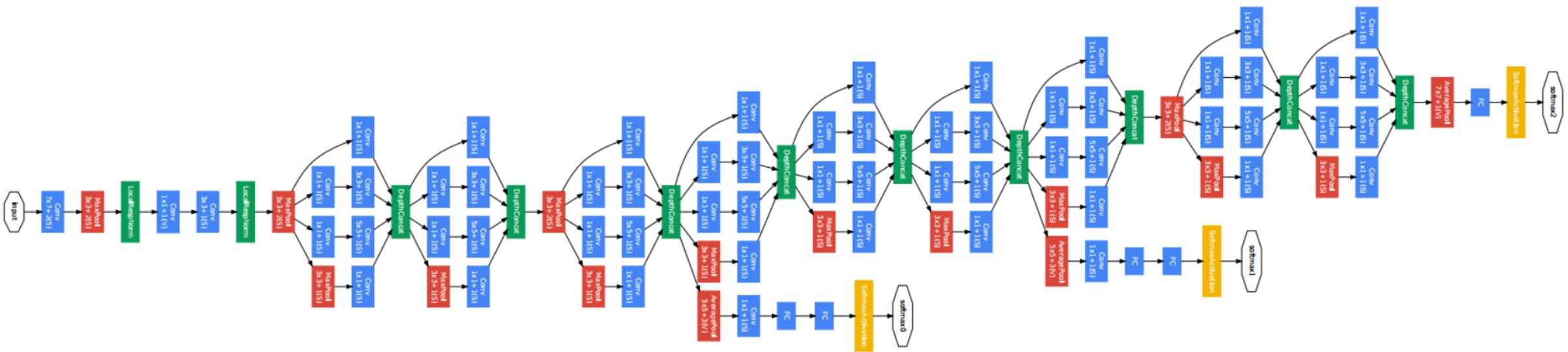
Нейронные сети



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:

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

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»?

Например, это:

$$\sum_{i=1}^l |y_i - a(x_i)| \rightarrow \min$$

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:

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

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?

Задача классификации

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

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

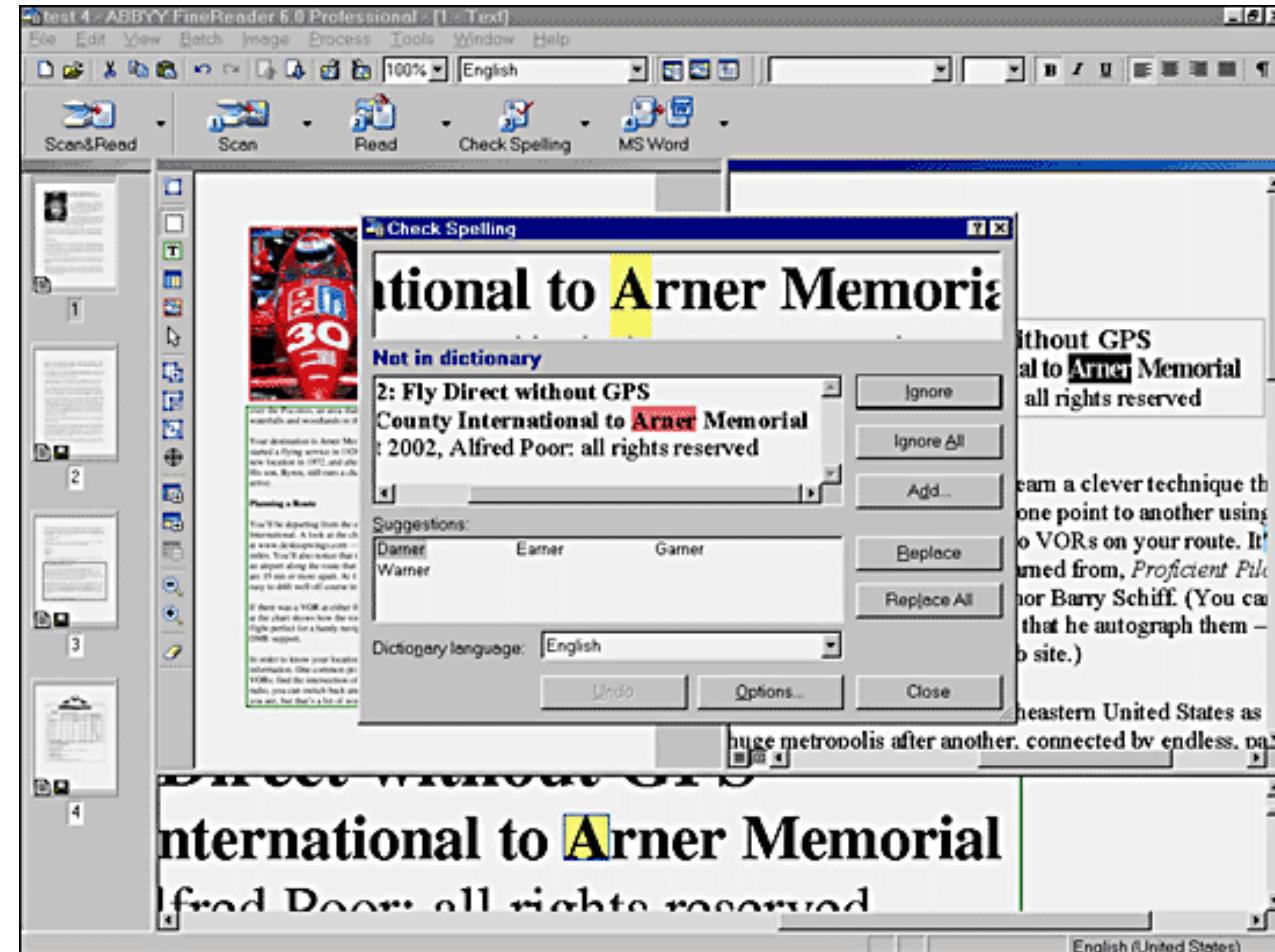
Мы строим прогнозирующий алгоритм:

$$y_i = a(x_i)$$

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

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

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'a, как настроить параметры?

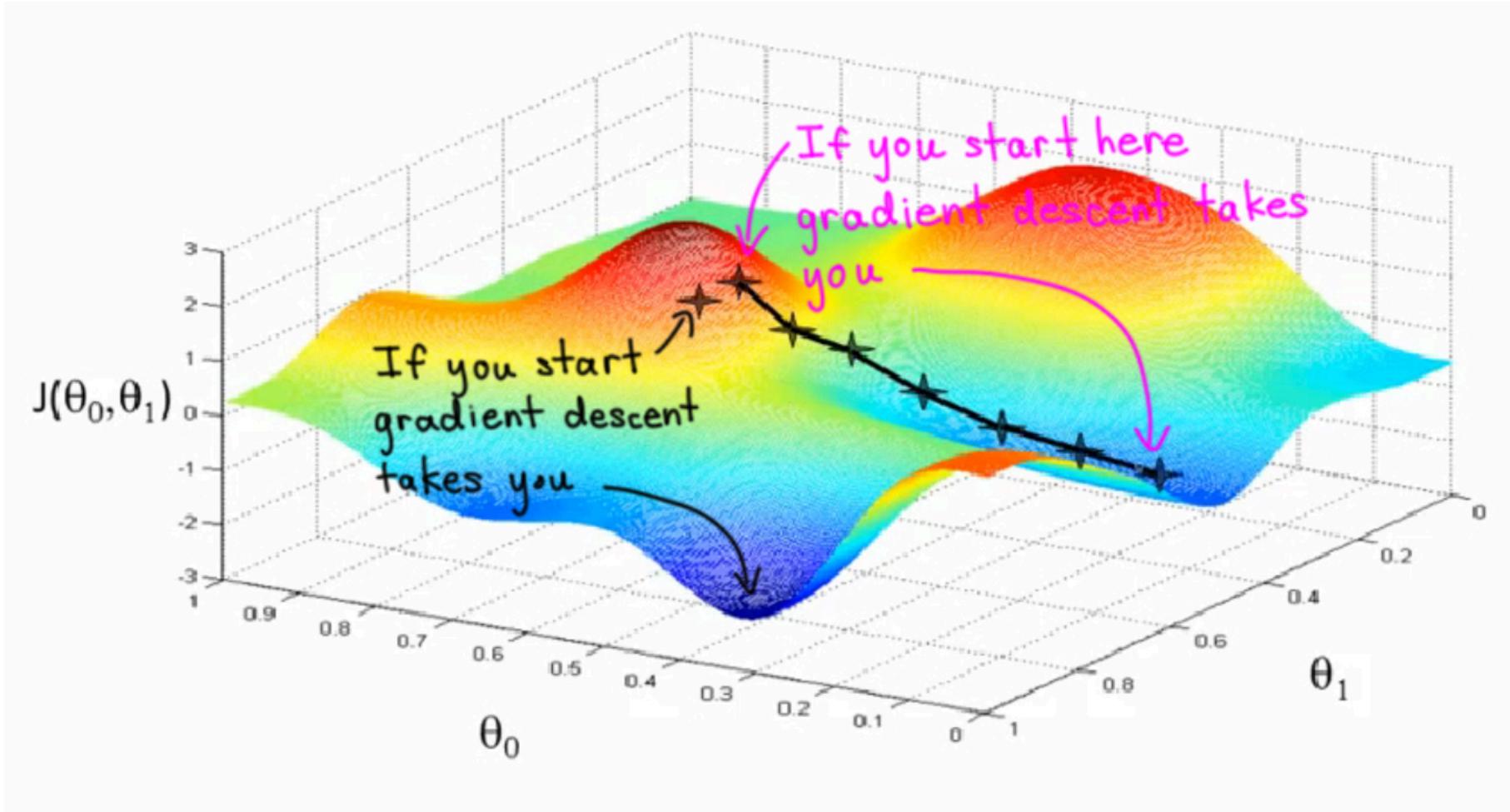
Сложный пример: исправление опечаток

Возможное решение:

$$\begin{aligned}Suggest(w) &= [w_1, w_2, \dots, w_k] \\Pos(w_j, [w_1, w_2, \dots, w_k]) &= j\end{aligned}$$

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

Градиентные методы оптимизации



Методы глобальной оптимизации

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

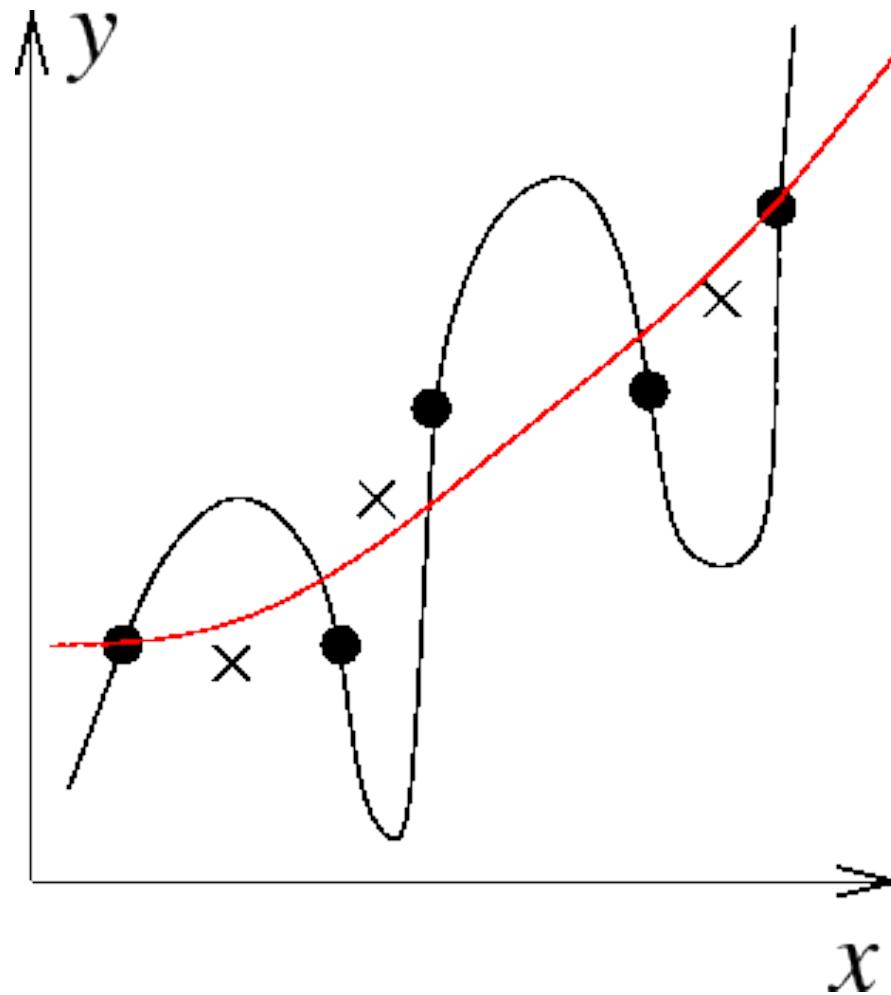
Методы глобальной оптимизации



$$P(\bar{x}^* \rightarrow \bar{x}_{i+1} | \bar{x}_i) = \begin{cases} 1, & F(\bar{x}^*) - F(\bar{x}_i) < 0 \\ \exp\left(-\frac{F(\bar{x}^*) - F(\bar{x}_i)}{Q_i}\right), & F(\bar{x}^*) - F(\bar{x}_i) \geq 0 \end{cases}.$$

v. Переобучение и недообучение

Переобучение на примере регрессии

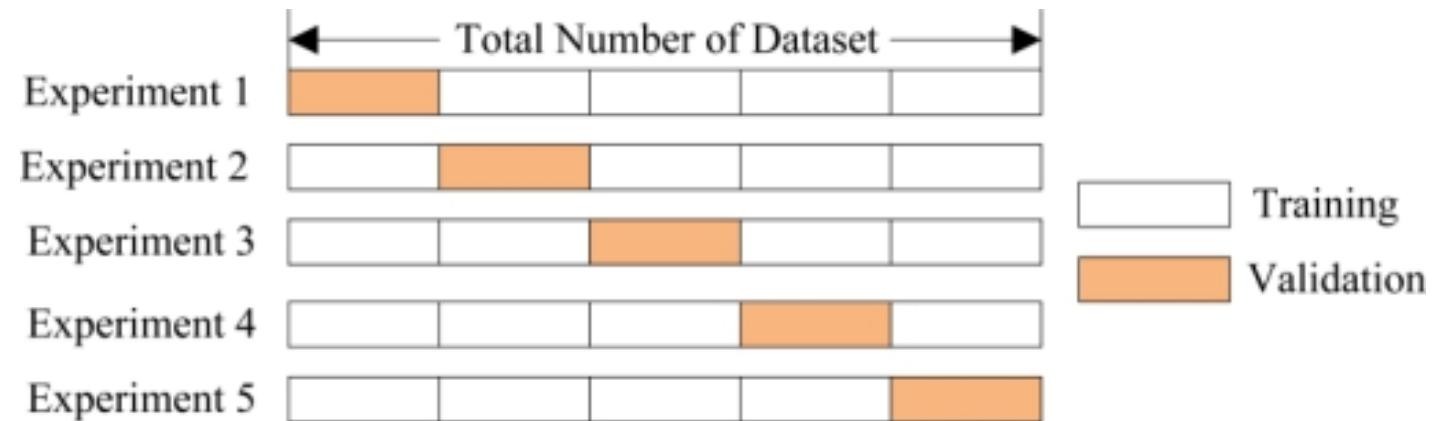


Оценка качества



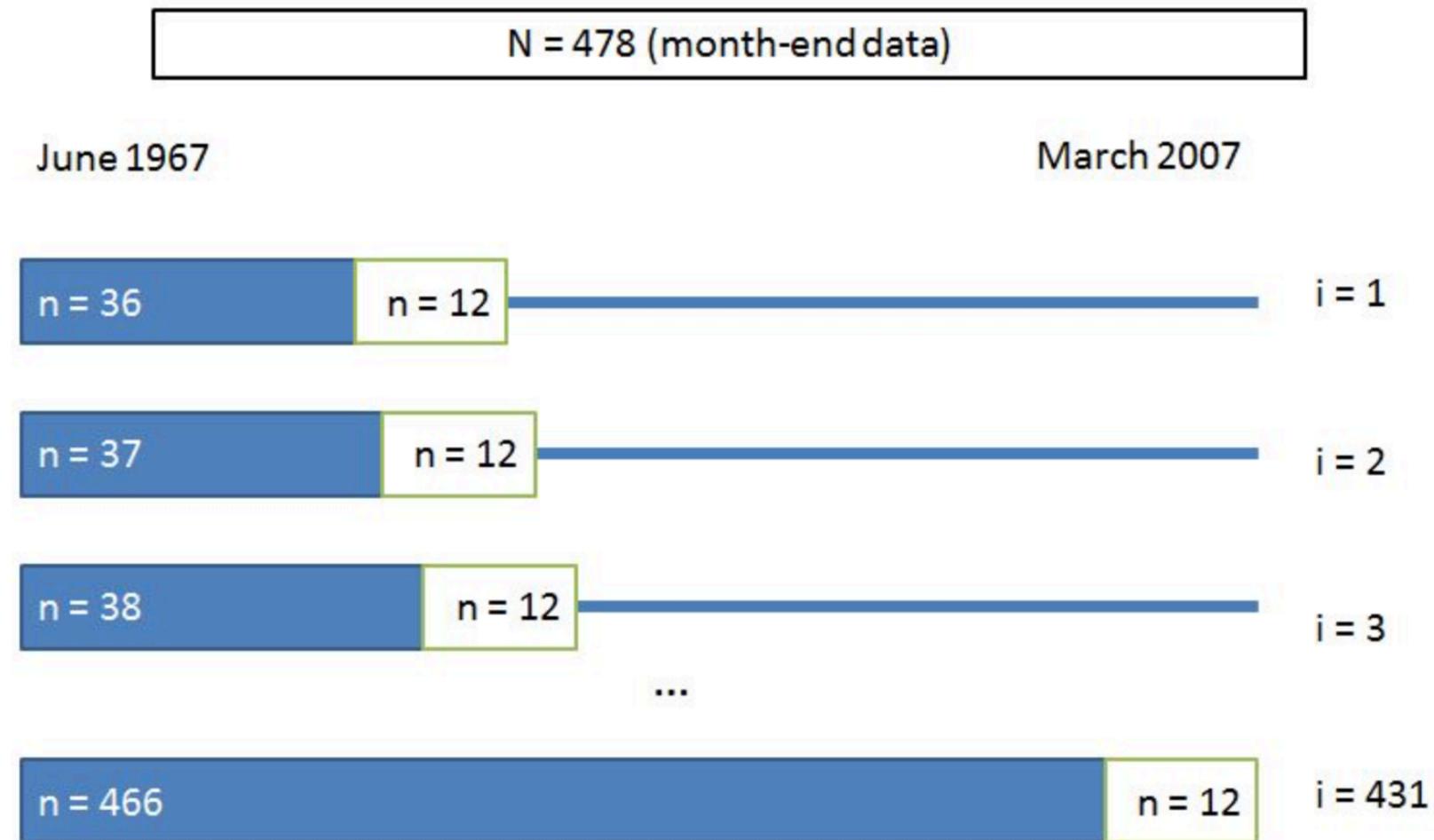
Кросс-валидация

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?

VI. Инструменты

Python

На чем будут примеры

- Python 3.x, библиотеки: numpy, scipy, sklearn, matplotlib
- Почему Python? Потому что можно всего в 5 - 30 строк очень простого кода продемонстрировать интересные явления.
- Что использовать на практике – ваш выбор
- Под Windows проще всего установить Anaconda Python



PyCharm



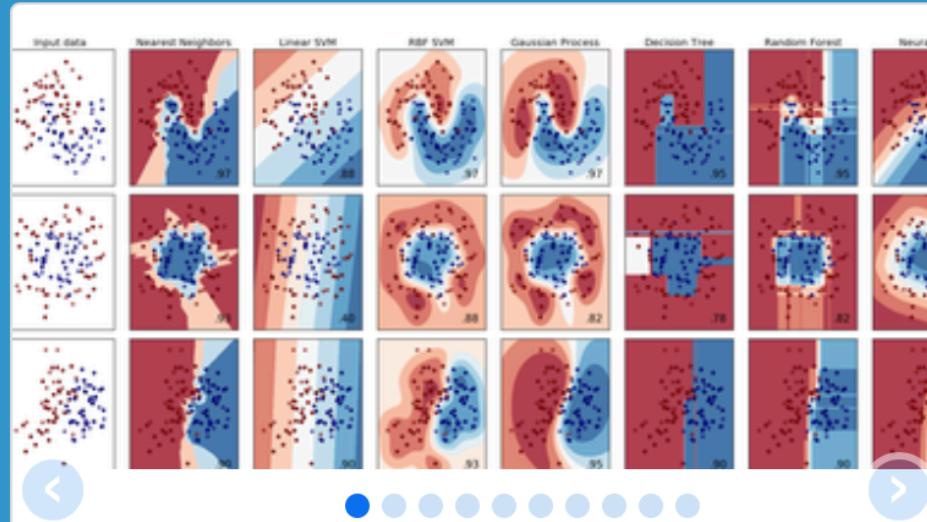
python

Scikit-learn

Scikit-learn

[Home](#)[Installation](#)[Documentation](#)[Examples](#)

Google™ Custom Search

Search x

scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Машинное обучение в несколько строк

```
from sklearn.linear_model import LogisticRegression
```

```
model = LogisticRegression()  
model.fit(X_train, y_train)  
predictions = model.predict(X_test)
```

Спасибо за внимание



info@applieddatascience.ru



https://t.me/joinchat/B10lThC96v0BQCvs_joNew



https://github.com/vkantor/ml2018jan_feb



<https://goo.gl/forms/11uVHzPFeOBGw6u2>