

Data Mining in Action

Лекция 1

Примеры и основные понятия

Виктор Кантор

Кто лектор

Работа:

Руководитель группы анализа пользовательского поведения в Yandex Data Factory, ранее – руководитель учебной группы в ABBYY

Преподавание:

- МФТИ: Спецкурс по анализу данных, Машинное обучение, Анализ изображений, Анализ Текстов
- Яндекс: Автоматическая обработка текстов
- СберТех: Машинное обучение, Язык Python и библиотеки для анализа данных
- Coursera: специализация «Машинное обучение и анализ данных» от Яндекса и МФТИ



Кто лектор

Задачи:

- Автоматическое тегирование текстов, аннотирование, анализ тональности, поиск похожих, классификация и кластеризация текстов, выделение полей из распознанных бумажных документов (написанных в свободной форме)
- Построение рекомендательных систем для магазинов
- Оптимизация рекламных кампаний
- Классификация изображений (сканов документов)

На этой лекции

- I. Примеры применения машинного обучения
- II. Стандартные задачи и методы
- III. Инструменты

Часть I: примеры применения

Пример задачи: 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	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	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	
1	24	3	49	1	3	3	4	4	53	3	2	2	1	1	1	0	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	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	0	1	0	0	0	1	0	1	1	
2	36	2	69	1	3	3	2	3	35	3	1	1	2	1	0	0	1	1	0	0	1	0	0	0	1	
4	12	2	31	4	4	1	4	1	61	3	1	1	1	1	0	0	0	1	0	0	0	1	0	1	1	
2	30	4	52	1	1	4	2	3	28	3	2	1	1	1	1	0	0	1	0	0	0	1	0	0	2	
2	12	2	13	1	2	2	1	3	25	3	1	1	1	1	1	0	0	1	0	0	1	0	0	0	1	2
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	0	1	0	0	0	1	0	0	1	2
2	12	2	16	1	3	2	1	3	22	3	1	1	2	1	0	0	0	1	0	0	0	1	0	0	1	1
1	24	4	12	1	5	3	4	3	60	3	2	1	1	1	1	0	0	1	0	0	0	1	0	1	2	
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	1	0	0	0	1	0	0	0	1	0	2	
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	0	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	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	0	2
4	24	2	34	3	5	3	2	3	31	3	1	2	2	1	0	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	0	1	0	0	0	1	0	0	1	1
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	0	1	0	0	0	1	0	0	1	1
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	0	1	0	0	0	1	0	0	1	1
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	0	1	1	0	0	0	1	0	0	1	1
4	10	4	21	5	3	4	1	3	26	3	2	1	1	2	0	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	0	1	0	0	0	1	0	0	1	1
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	0	1	0	0	0	1	0	0	1	1
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	0	1	0	0	0	1	0	0	1	1
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	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	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	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	0	1	1	
1	24	3	49	1	3	3	4	4	53	3	2	2	1	1	1	0	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	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	0	1	0	0	0	0	1	1	1	
2	36	2	60	1	3	2	2	1	22	3	1	1	1	1	0	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	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	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	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	1	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	1	
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	0	1	0	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	0	1	0	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)



Attribute 2: Duration in month																									
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	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	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	1	0	0	1	
2	30	4	1	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	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	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	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	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	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	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	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	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	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	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	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	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	0	1	0	0	0	1	0	0	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	0	1	0	0	0	1	0	0	2	
2	12	2	13	1	2	2	1	3	25	3	1	1	1	1	1	0	0	1	0	0	1	0	0	0	1	
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	0	1	0	0	0	1	0	0	1	
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	0	1	0	0	0	1	0	0	2	
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	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	0	2	
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	1	0	0	1	1	0	0	0	0	0	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	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	0	1	
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	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	0	1	0	0	0	1	0	
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	1	1	0	0	0	1	0	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	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	0	
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	0	
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	0	1	0	0	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	0	1	0	0	1	1	

Answer: 1 – Good, 2 - Bad

Выдача кредита

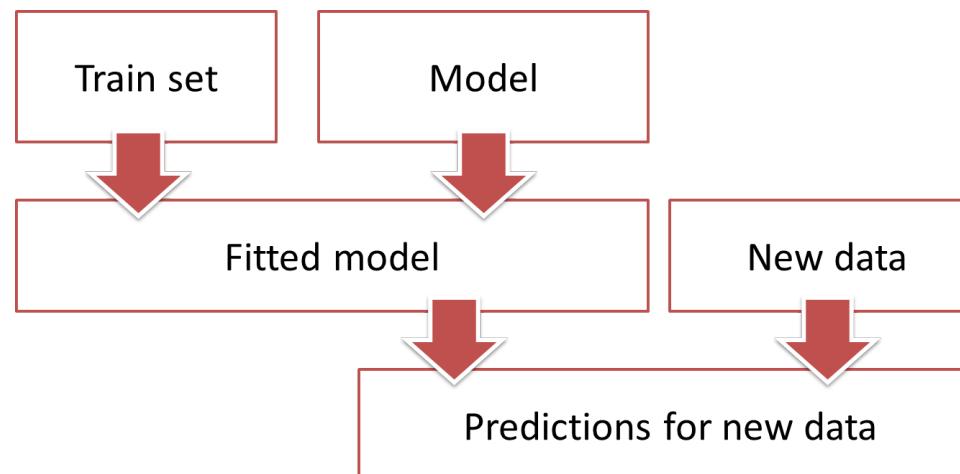
Задача (supervised classification): предсказать класс (1 или 2)

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



Пример проекта: рекомендации товаров

Блок рекомендаций

Товар 1	Товар 2	Товар 3	Товар 4
---------	---------	---------	---------

Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4

Вероятность:	p_1	p_2	p_3	p_4
--------------	-------	-------	-------	-------

Максимизация дохода

	Товар 1	Товар 2	Товар 3	Товар 4
Вероятность:	p_1	p_2	p_3	p_4
Цена:	c_1	c_2	c_3	c_4

Максимизация дохода



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.

Вероятность:	0.05	0.02	0.015	0.009
Цена:	3490	1990	1590	1970

Прогнозирование вероятности

- Объекты: тройки (пользователь, товар, момент времени)
- Классы: 1 - товар будет куплен, 0 – товар не будет куплен
- Признаки: параметры пользователя, товара, момента времени и их «взаимодействие»

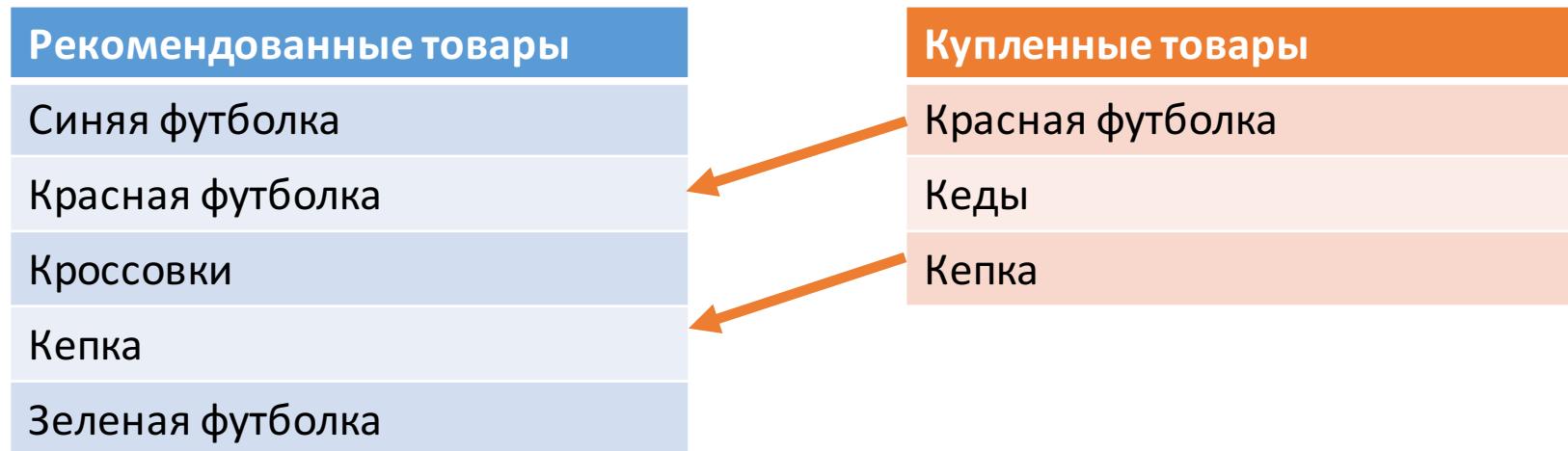
Отбор кандидатов

- Популярные
- Популярные из тех же категорий
- Часто просматриваемые вместе с теми товарами, которые пользователь уже видел
- Из заранее подготовленных списков похожих товаров

Генерация негативных примеров

- Добавить к каждому позитивному примеру весь каталог как негативный (не реально)
- Случайные с равномерным распределением
- Случайные, с вероятностями, пропорциональными популярности объекта
- Самые популярные примеры
- Те объекты, которые рекомендовал бы какой-то алгоритм, но они не были куплены

Точность (Precision@k)

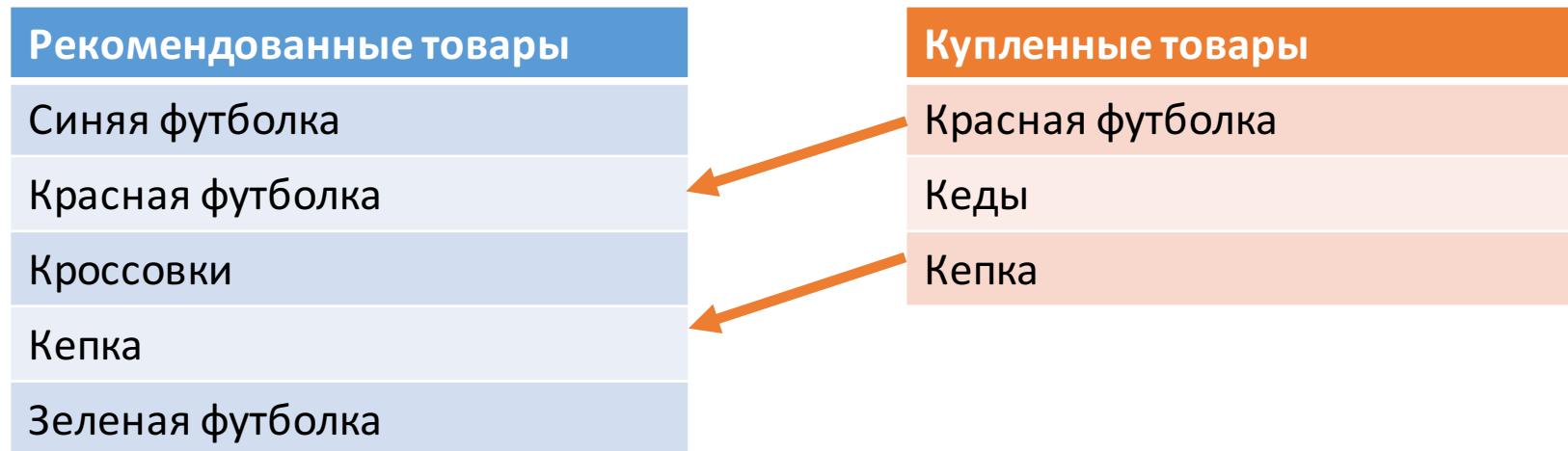


k – количество
рекомендаций

$$\text{Precision}@k = \frac{\text{купленное из рекомендованного}}{k}$$

AveragePrecision@k - усредненный по сессиям Precision@k

Полнота (Recall@k)



k – количество
рекомендаций

$$\text{Recall}@k = \frac{\text{купленное из рекомендованного}}{\text{количество покупок}}$$

AverageRecall@k - усредненный по сессиям Recall@k

Взвешенный ценами recall@k

Рекомендованные товары	Купленные товары
Синяя футболка – 1000р	Красная футболка – 1200р
Красная футболка – 1200р	Кеды – 3000р
Кроссовки – 3500р	Кепка – 900р
Кепка – 900р	
Зеленая футболка – 800р	

Взвешенный ценами Recall@k = $\frac{\text{стоимость купленного из рекомендованного}}{\text{стоимость покупок}}$

AverageRecall@k - усредненный по сессиям Recall@k

Качество классификации против качества рекомендаций

Пример – 2 решения для прогноза купит/не купит товар

	Алгоритм 1	Алгоритм 2
Качество классификации	0.52	0.85
Recall@5	0.72	0.71

История из практики: сравнение методов

- 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 :)

Онлайновая оценка качества

Допустим, на исторических данных качество алгоритма высокое, а будет ли оно высоким в реальности?

Онлайновая оценка качества

Допустим, на исторических данных качество алгоритма высокое, а будет ли оно высоким в реальности?

Идеи:

1. А/В тест
2. Оценка статзначимости результата

A/B тест

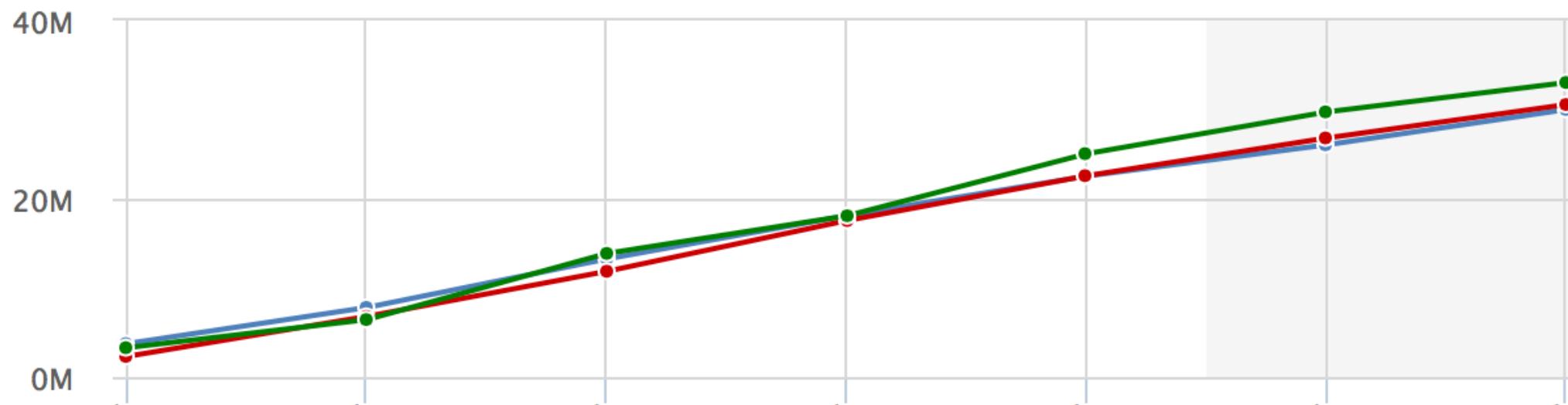
1. Случайным образом делим пользователей на равные группы
2. Измеряем целевые метрики (например, количество заказов или доход) в каждой группе за длительный период времени
3. Получаем какое-то число для каждой группы
4. Что дальше?

Истории из практики: разбиение на группы

- Предложено:
 - Брать hash от user_id
 - Смотреть на остаток от деления на 2
- Сделано:
 - Брать hash от user_id+user_email
 - Смотреть на остаток от деления на 2

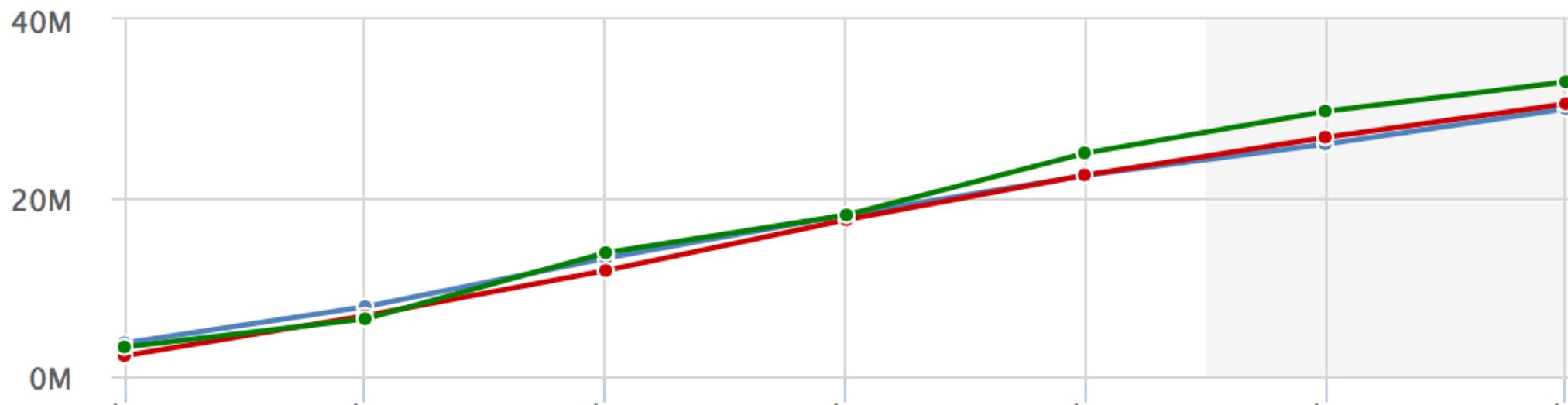
Статистическая значимость: пример

Суммарная выручка



Статистическая значимость: пример

Суммарная выручка



Одна кривая отличается от других на 10%
Но разбиение на самом деле – случайное

На какие метрики смотрят в онлайне

- Доход в группе
- Доход с пользовательской сессии
- Средняя стоимость купленного товара
- Средний чек
- Конверсия в покупку
- Клики
- Различные модели атрибуции: last click, first click

Итог: о чём нужно позаботиться

- Высокоуровневая постановка задачи - от экономического эффекта
- Оценка возможного экономического эффекта
- Оценка реализуемости проекта
- Оффлайновая оценка качества
- Онлайновая оценка качества
- Решение задачи – декомпозиция на подзадачи, выбор признаков, выбор моделей

Пример сервиса: аннотирование текстов

Аннотирование текста: сайт

The screenshot shows a web browser window with the URL 2l2r.ru. The page title is "2Long2Read" and the subtitle is "Мы находим главное в текстах". There are two tabs: "Исходный текст" (highlighted) and "С выделенным главным". A progress bar indicates 29% completion. The main content area contains three paragraphs of text. The first paragraph is highlighted in yellow. The second and third paragraphs are partially visible. Below the text, there is a note: "Выберите пример: [новость](#), [отрывок из Википедии](#), [описание компании](#)". At the bottom, there are two numbered steps: "1. Есть длинный нудный текст..." and "2. Скопируйте его...". Step 1 shows a Microsoft Word interface with a long document. Step 2 shows a context menu with "Вырезать" (Cut) and "Копировать" (Copy) options.

Как все устроено

- Текст разбивается на предложения
- Из предложений извлекаются слова и приводятся в начальную форму
- Для каждого предложения вычисляется его ранг
- Выделяются предложения с самым высоким рангом

Аннотирование: плагин для браузера

2L2R: находим главно L Lenta.ru: Россия: Пол +

lenta.ru > Lenta.ru: Россия: Политика: Законопроект о двойном гражданстве прошел первое чтение

18:22, 14 мая 2014

Главное
Россия
Мир
Бывший СССР
Экономика
Наука и техника
Спорт
Культура
Интернет и СМИ
Из жизни

Lenta.doc
Мотор
Дом
МедНовости

Статьи
Галереи

Поиск

18+

Законопроект о двойном гражданстве прошел первое чтение

Заседание Госдумы
Фото: Владимир Федоренко / РИА Новости

В среду, 14 мая, Госдума в первом чтении одобрила законопроект о двойном гражданстве. Документ обязывает граждан в письменном виде уведомить территориальный орган ФМС о получении иностранного гражданства в течении месяца после его получения.

Помимо этого, граждане России, имеющие второе гражданство до вступления в силу новых правил, должны уведомить об этом местный орган ФМС в течение трех месяцев.

За неисполнение этой обязанности предусматривается уголовная ответственность в виде штрафа в размере до 200 тысяч рублей или в размере заработной платы или иного дохода осужденного за период до 1 года, либо обязательные работы на срок до 400 часов.

Вместе с тем, по мнению депутатов, документ нуждается в существенной переработке ко второму чтению.

ПОСЛЕДНИЕ НОВОСТИ

02:24 Ополченцы Донбасса поставили Киеву ультиматум

00:49 Российские хоккеисты уверенно обыграли сборную Казахстана

00:09 Пост главреда The New York Times впервые занял афроамериканец

23:29 США и монархии Персидского залива усилят сотрудничество по ПРО

01:17 Sony Pictures приобрела права на экранизацию книги о Сноудене

20:28 Пентагон подготовился к зомби-апокалипсису

01:56 Годовалый ребенок упал в США с 11 этажа и выжил

2L 2R 25%

Поделиться главным: B f t

Еще примеры

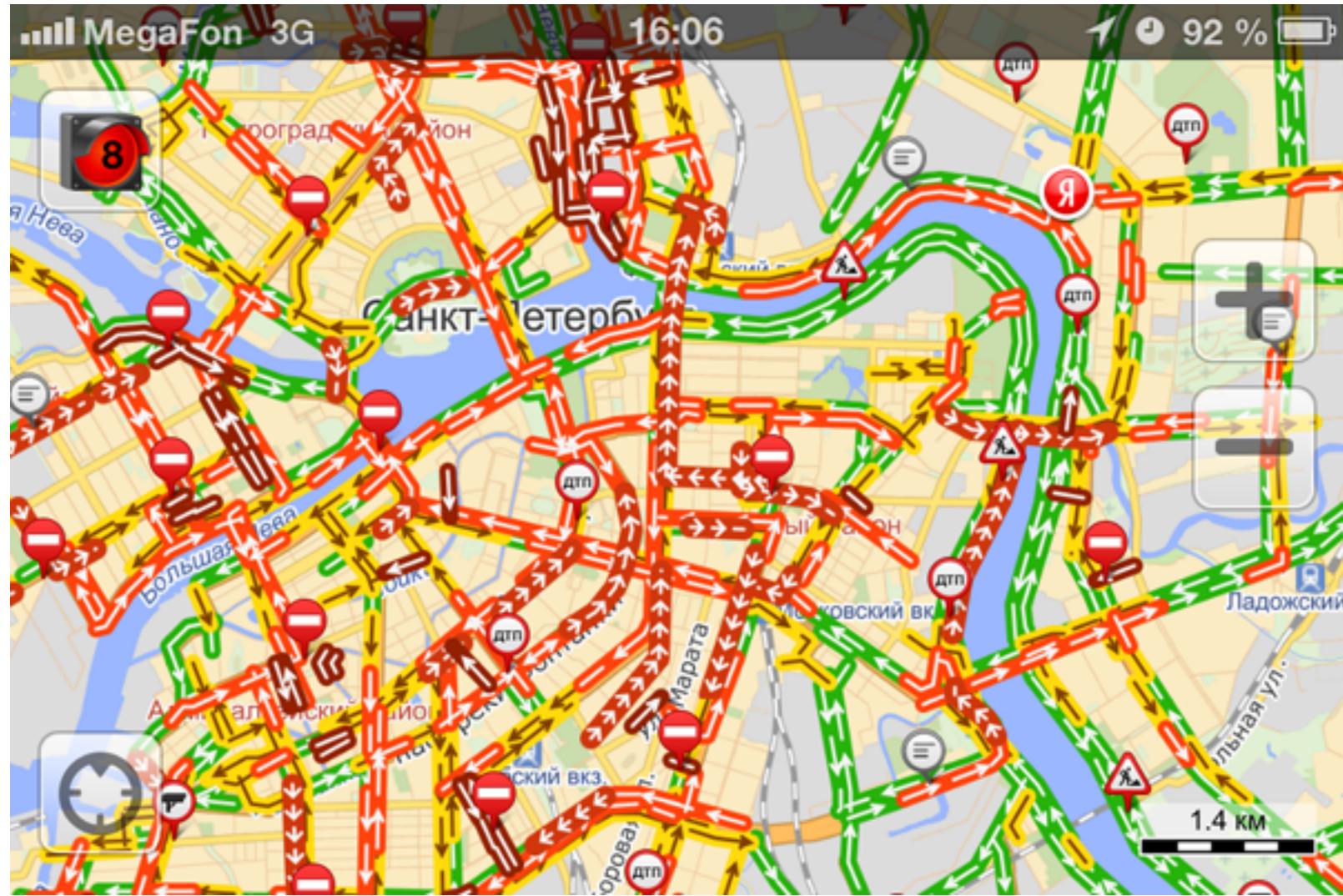
VK Data Mining in Action | ВКонтакте[vk.com > data_mining_in_action](#) ▾

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](#) ▾





Часть II: стандартные задачи и методы

Стандартные постановки задач
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)

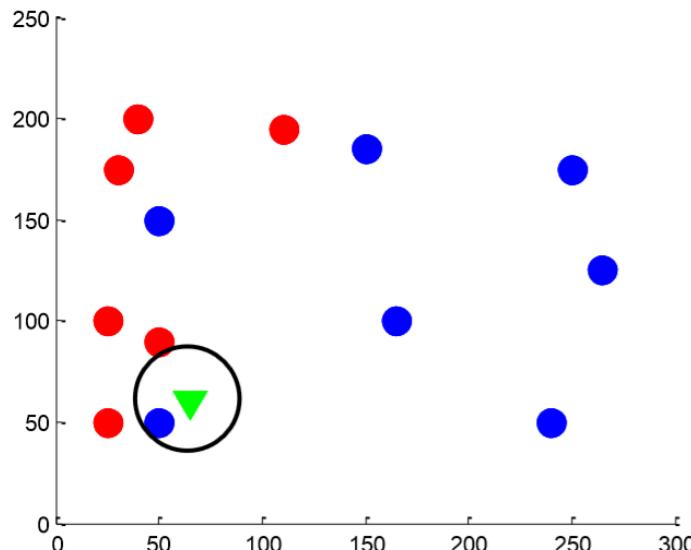
Классификация: обучающая выборка

Fisher's Iris Data

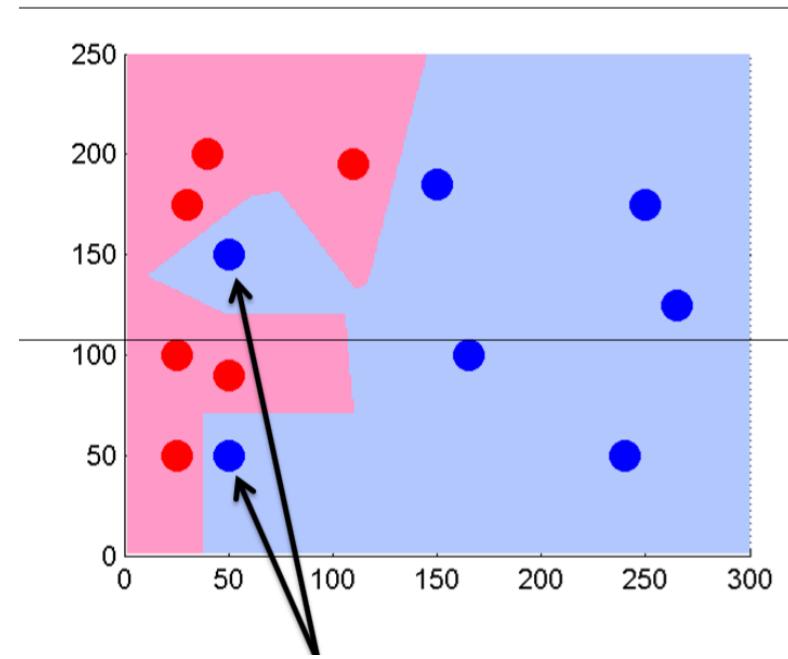
Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	<i>I. versicolor</i>
6.0	2.2	5.0	1.5	<i>I. virginica</i>
6.2	2.2	4.5	1.5	<i>I. versicolor</i>
6.0	2.2	4.0	1.0	<i>I. versicolor</i>
6.3	2.3	4.4	1.3	<i>I. versicolor</i>
5.5	2.3	4.0	1.3	<i>I. versicolor</i>
5.0	2.3	3.3	1.0	<i>I. 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 nearest neighbours



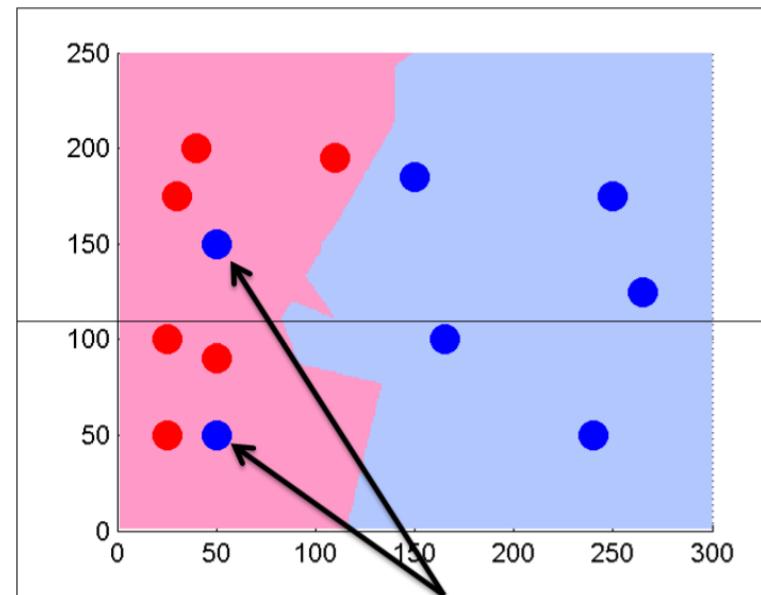
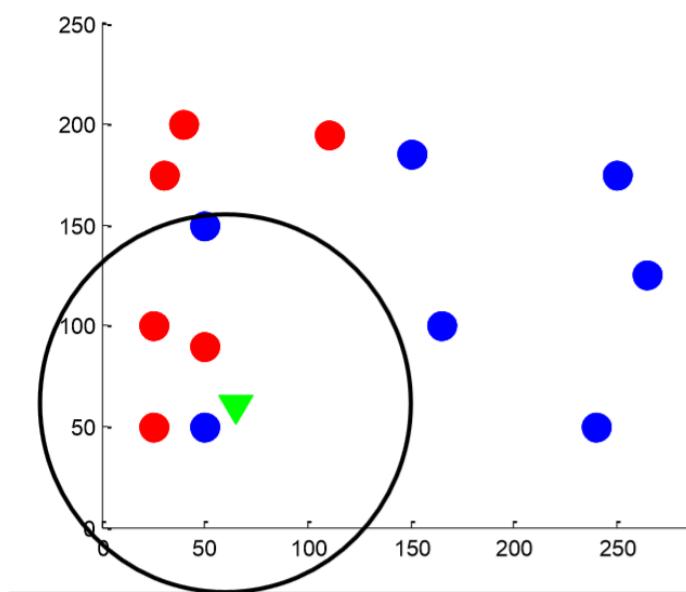
$k = 1$



Шумы? (outliers)

Простой классификатор: kNN

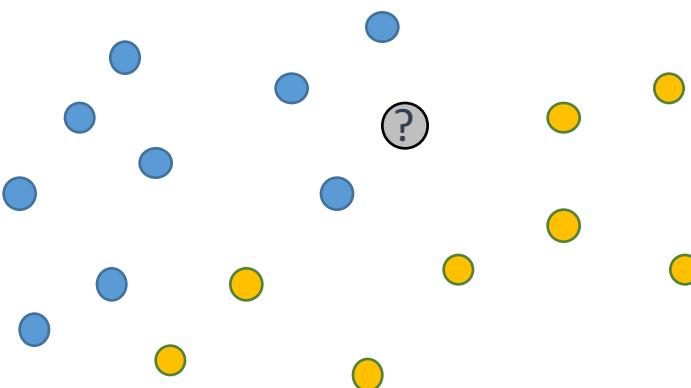
k nearest neighbours



$k = 5$

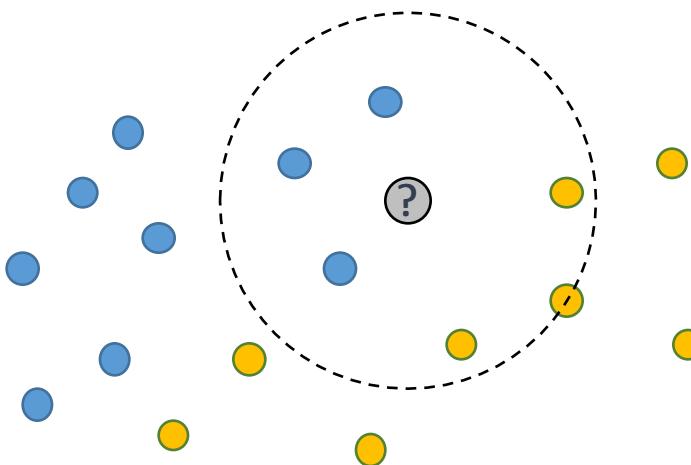
Взвешенный kNN

Пример классификации ($k = 6$):



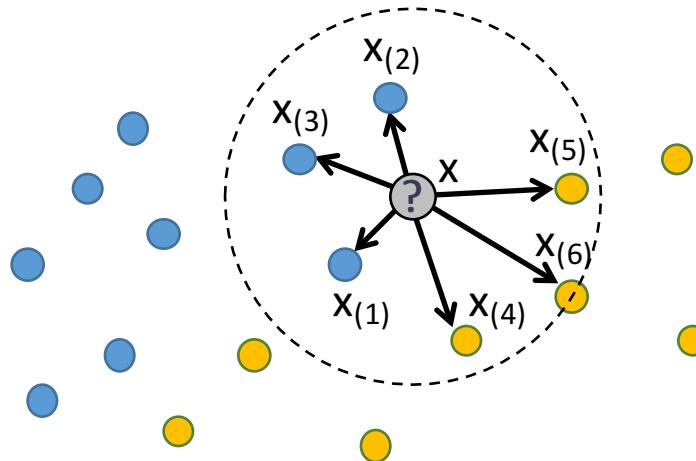
Взвешенный kNN

Пример классификации ($k = 6$):



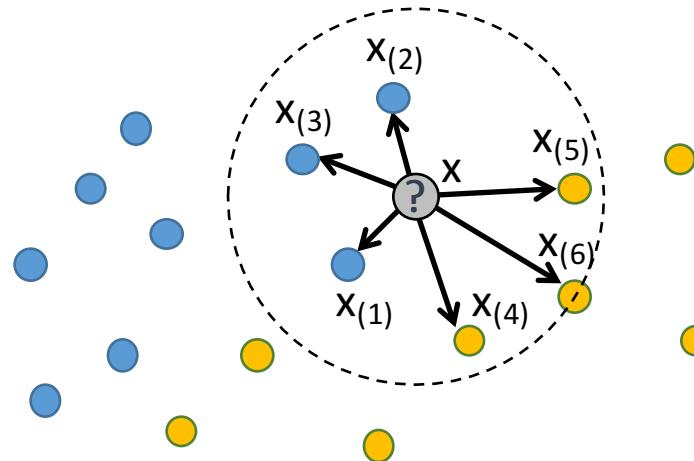
Взвешенный kNN

Пример классификации ($k = 6$):



Взвешенный kNN

Пример классификации ($k = 6$):

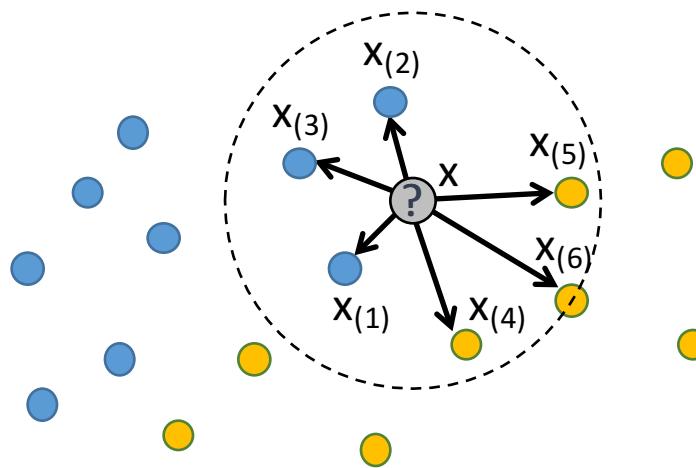


Веса можно определить как функцию от соседа или его номера:

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

Взвешенный kNN

Пример классификации ($k = 6$):



Веса можно определить как функцию от соседа или его номера:

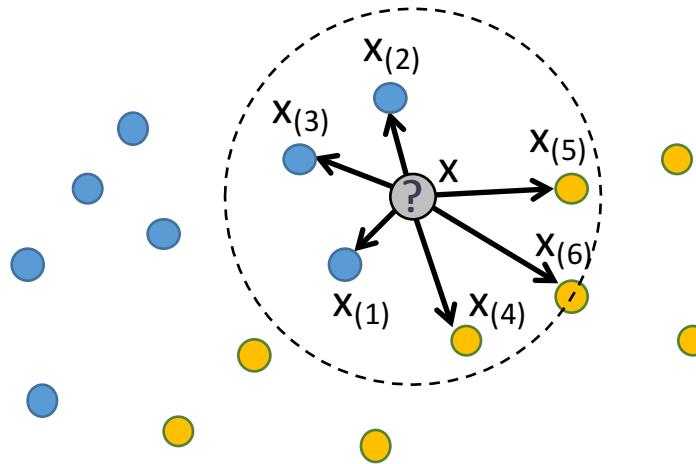
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

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

Взвешенный kNN

Пример классификации ($k = 6$):



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

Веса можно определить как функцию от соседа или его номера:

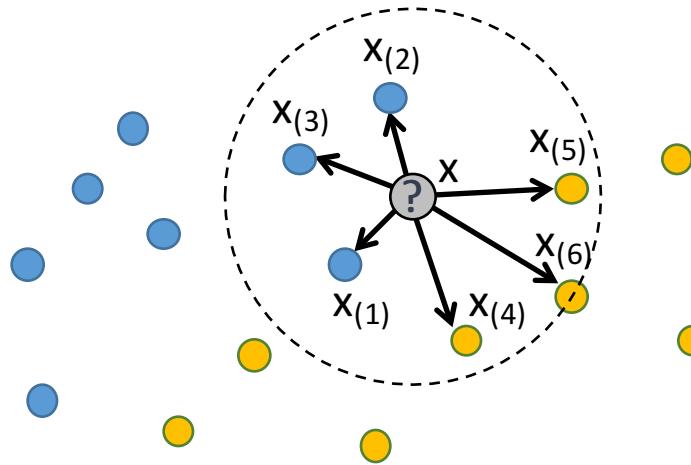
$$w(x_{(i)}) = w(i)$$

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:

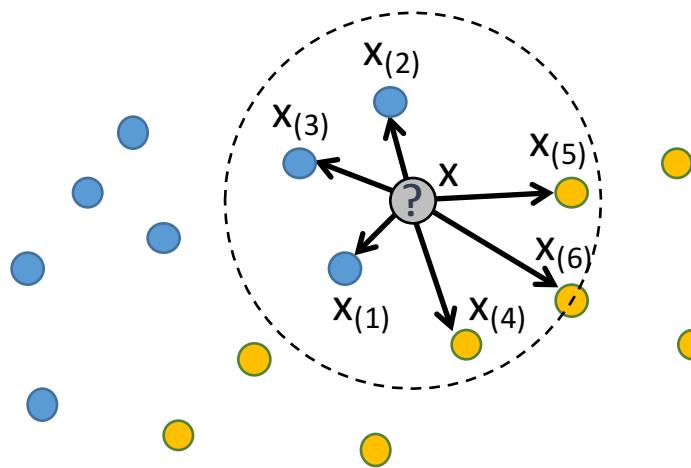
$$w(x(i)) = w(d(x, x_{(i)}))$$

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

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

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:

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

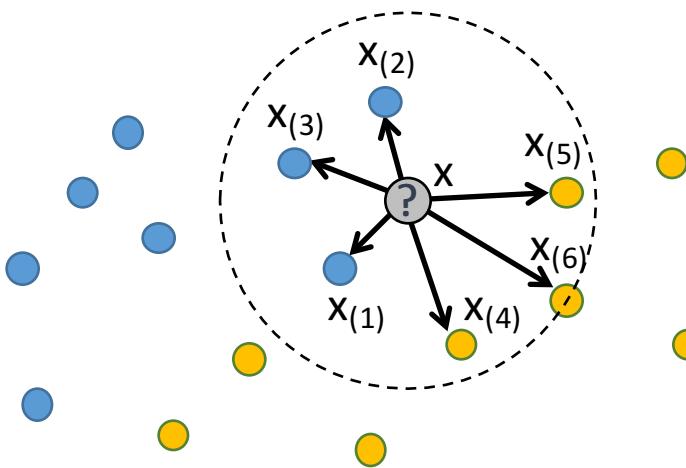
$$Z_{\text{blue}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

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

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

Взвешенный kNN

Пример классификации ($k = 6$):



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

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

Веса можно определить как функцию от соседа или его номера:

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

или как функцию расстояния:

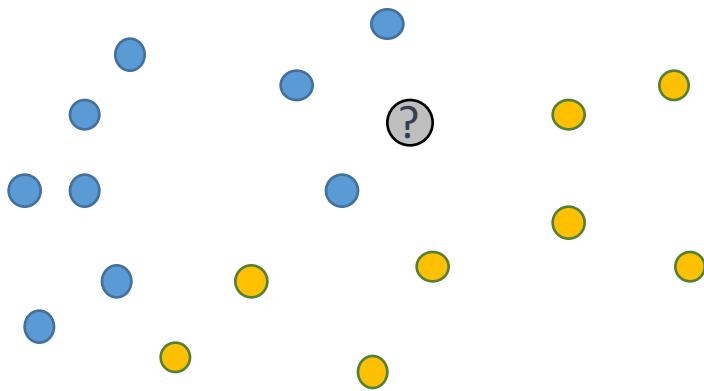
$$w(x(i)) = w(d(x, x_{(i)}))$$

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

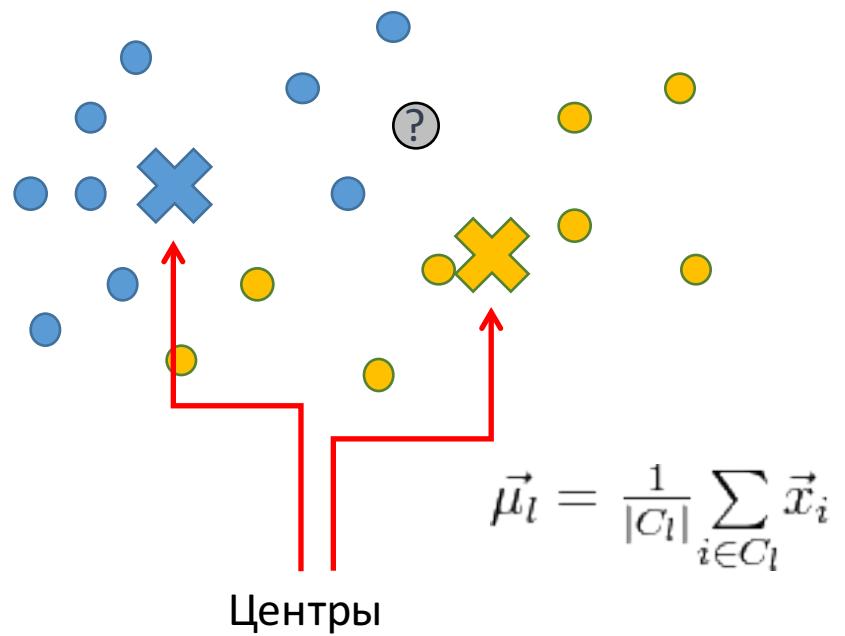
$$\text{if } Z_{\text{yellow}} > Z_{\text{blue}} : \quad \text{?} = \text{yellow}$$

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

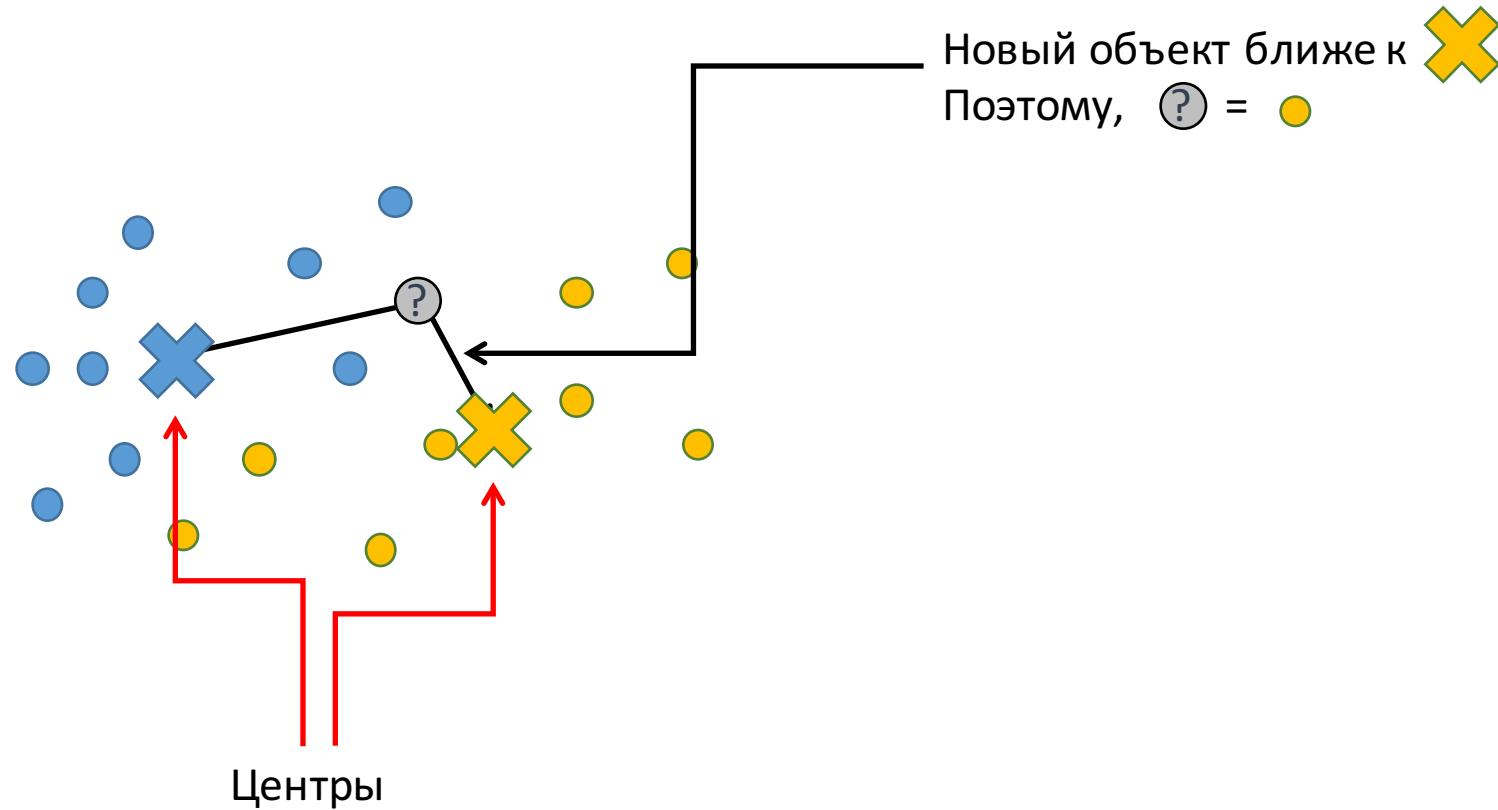
Центроидный классификатор



Центроидный классификатор



Центроидный классификатор



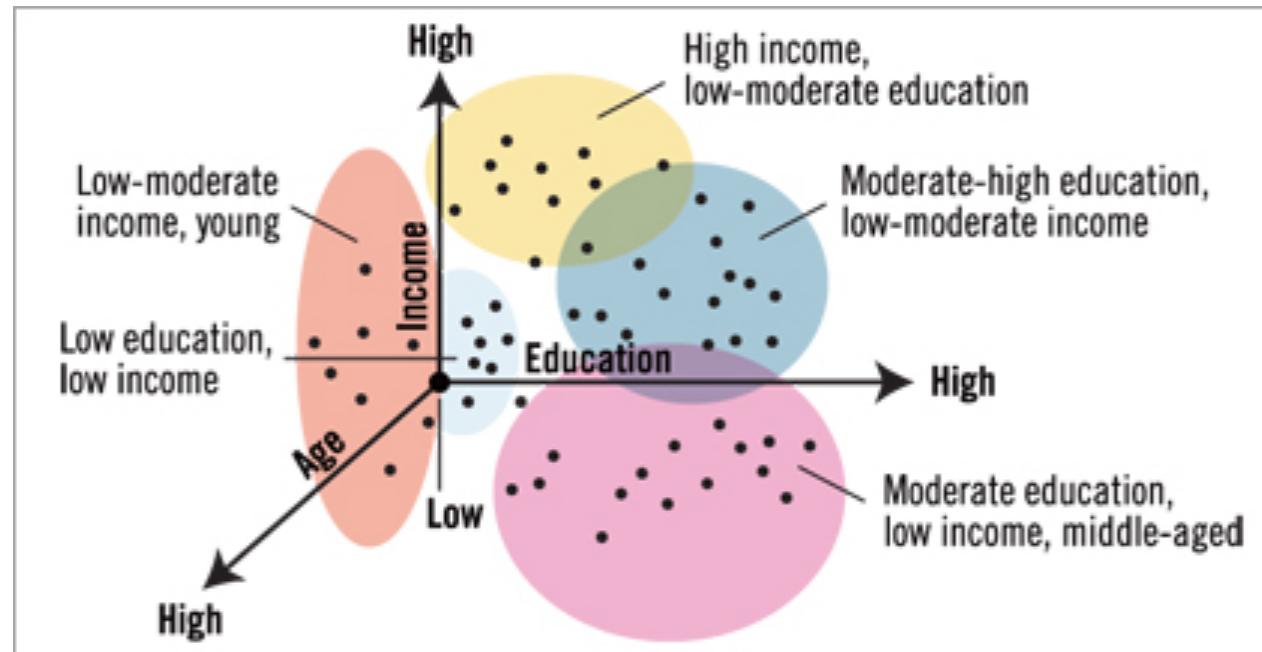
Кластеризация

Вход (обучающая выборка):

Признаки N объектов

Выход:

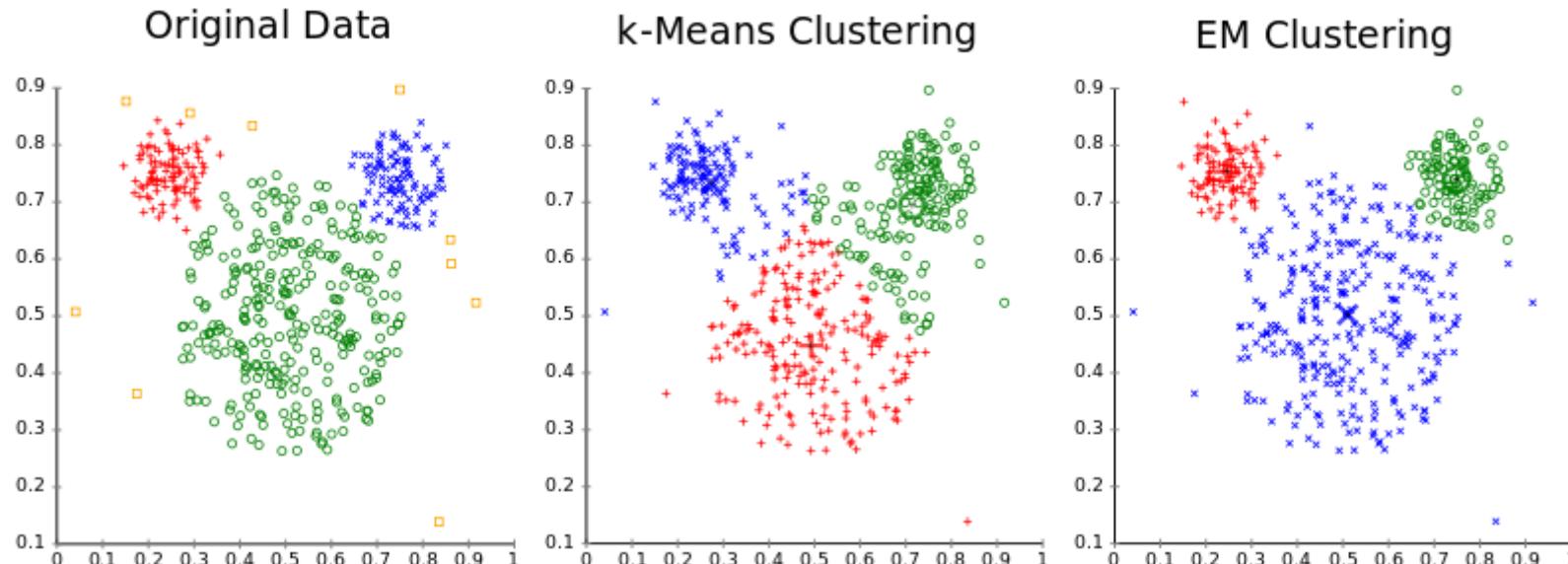
Найденные в выборке классы (кластеры), метки кластеров для объектов из обучающей выборки и алгоритм отнесения новых объектов к кластеру



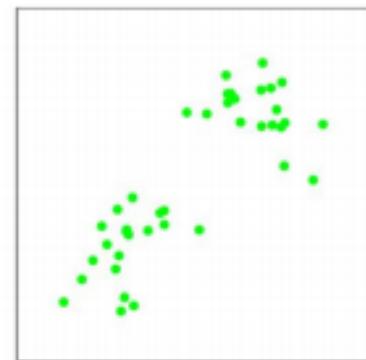
Пример: сегментация рынка

Кластеризация

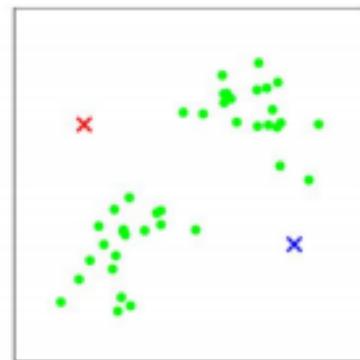
Different cluster analysis results on "mouse" data set:



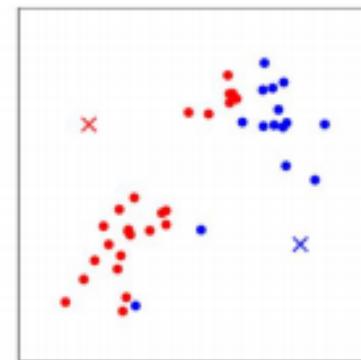
Простой алгоритм кластеризации: kMeans



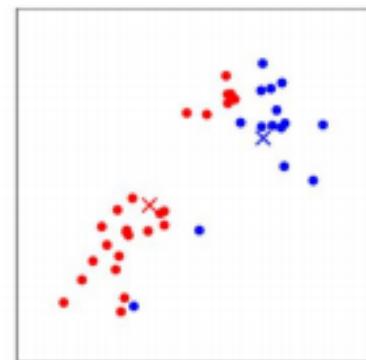
(a)



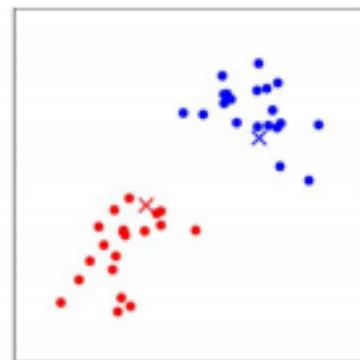
(b)



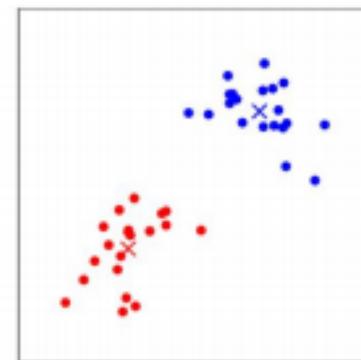
(c)



(d)



(e)



(f)

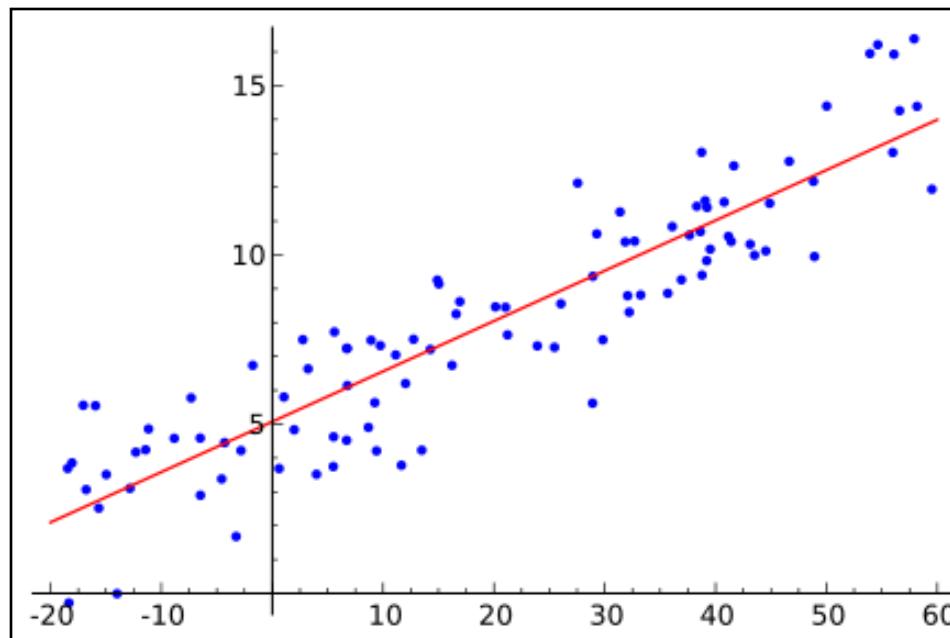
Регрессия

Вход (обучающая выборка):

Признаки N объектов с известными значениями прогнозируемого вещественного параметра объекта

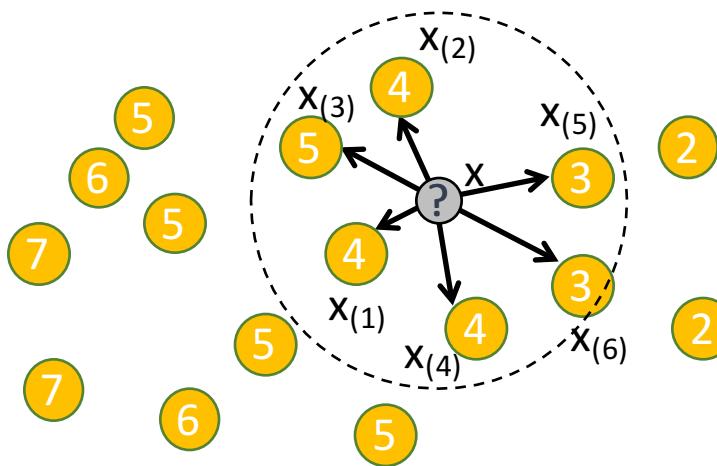
Выход:

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:

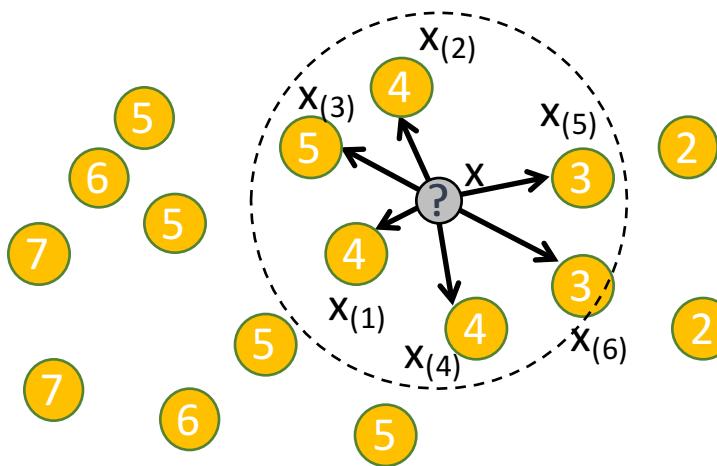
$$w(x_{(i)}) = w(i)$$

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:

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

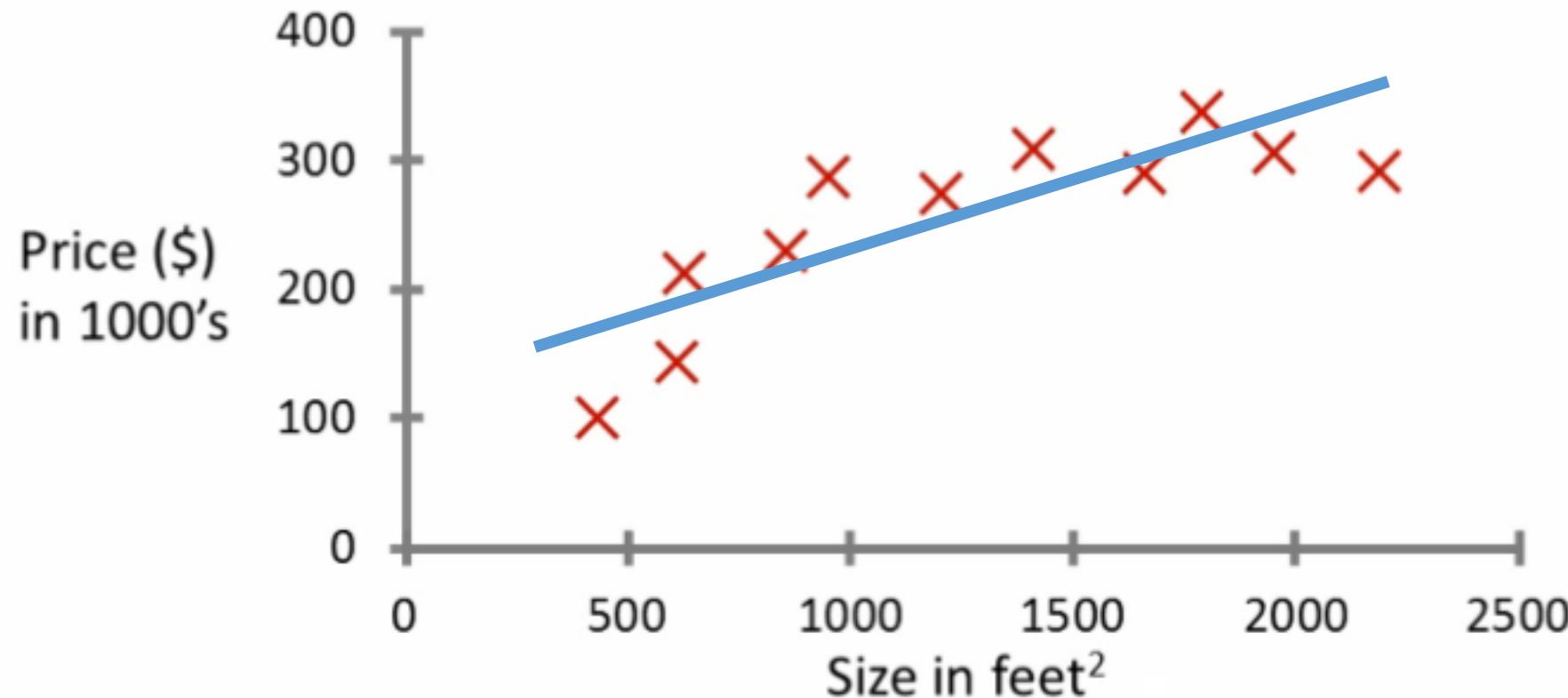
или как функцию расстояния:

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

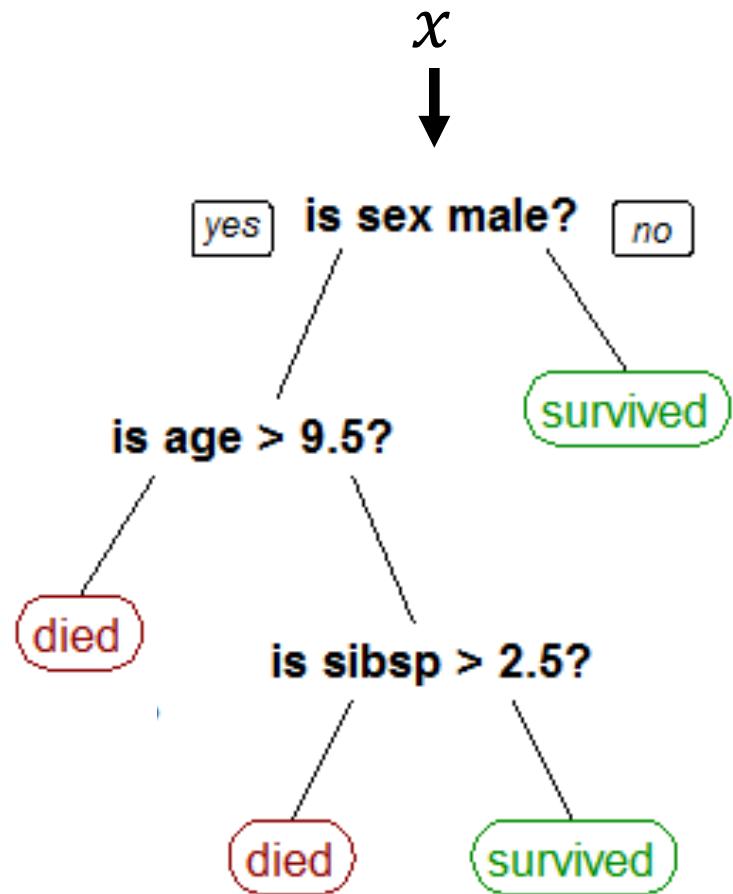
$$\textcircled{?} = \frac{4 \cdot w(x_{(1)}) + 4 \cdot w(x_{(2)}) + 5 \cdot w(x_{(3)}) + 4 \cdot w(x_{(4)}) + 3 \cdot w(x_{(5)}) + 3 \cdot w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

Наиболее часто используемые методы

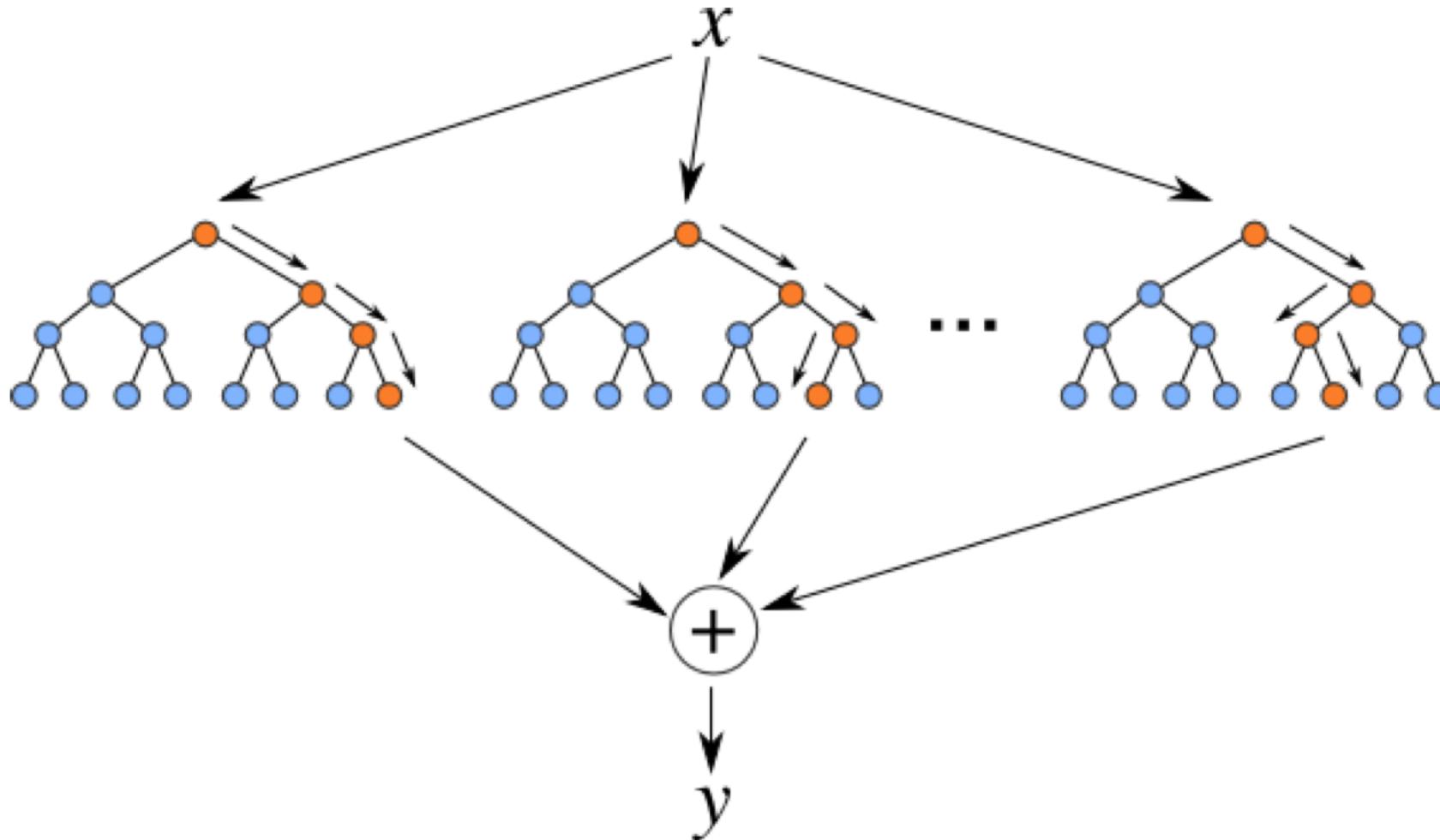
Линейные модели



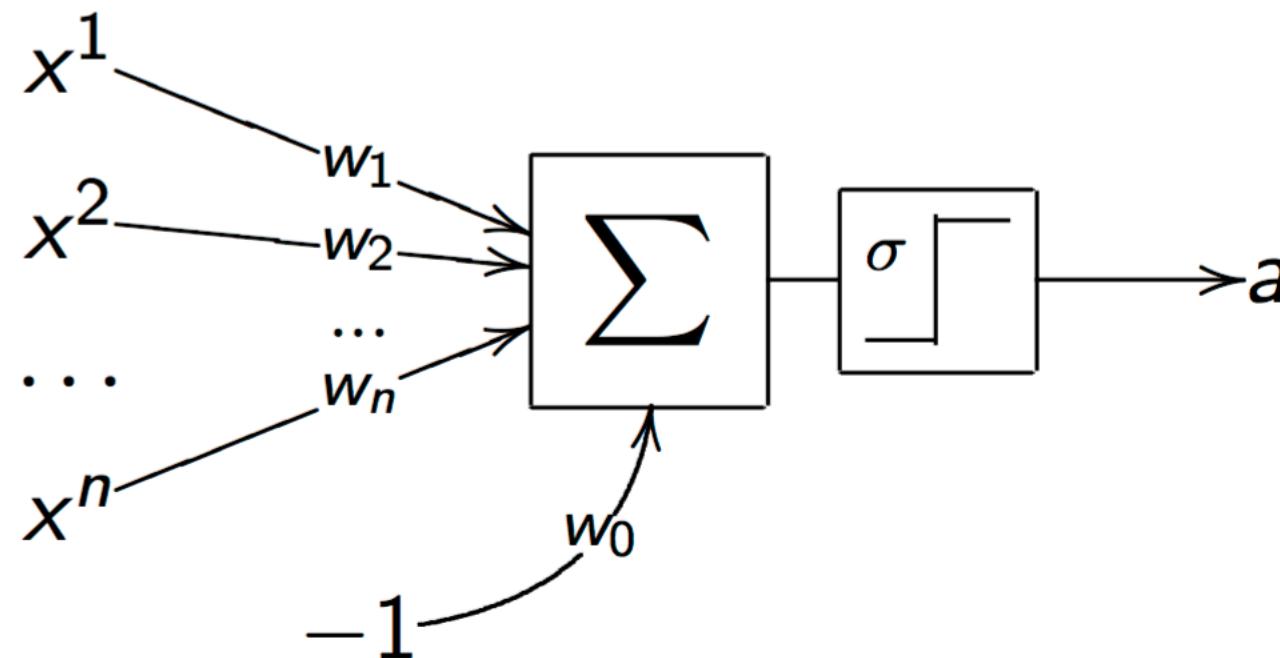
Решающие деревья



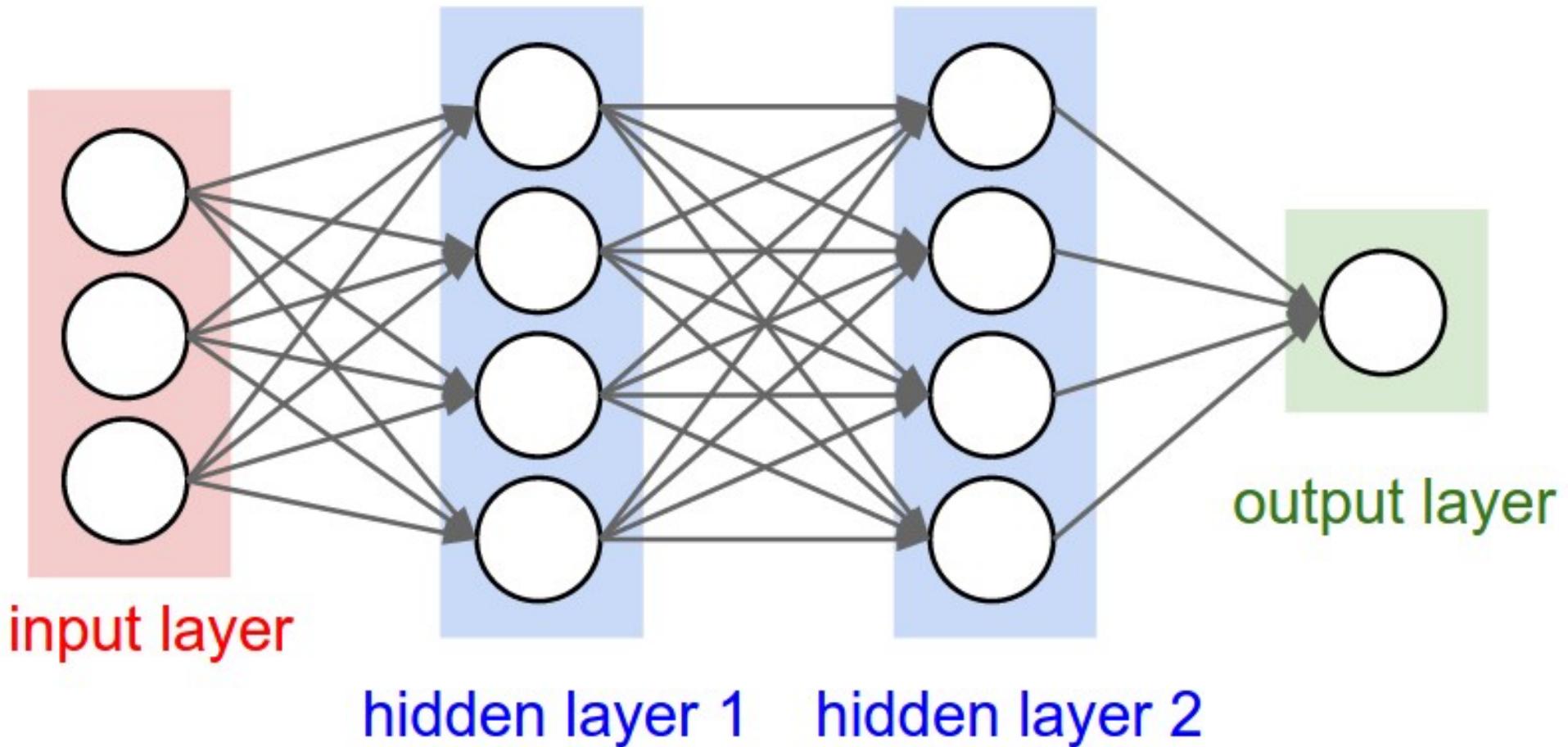
Ансамбли решающих деревьев



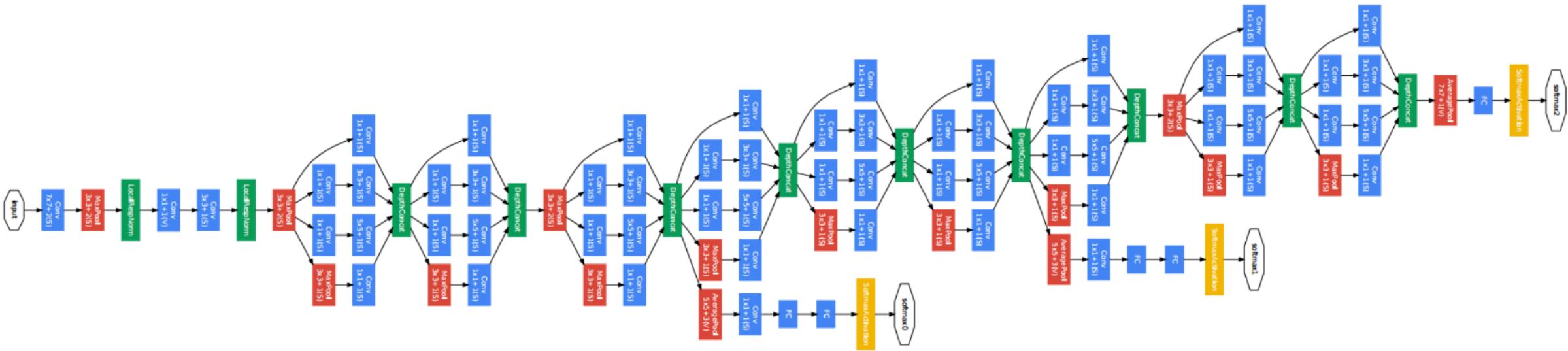
Нейронные сети



Нейронные сети



Нейронные сети



GoogLeNet

Функционалы качества и методы оптимизации

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»?

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:

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

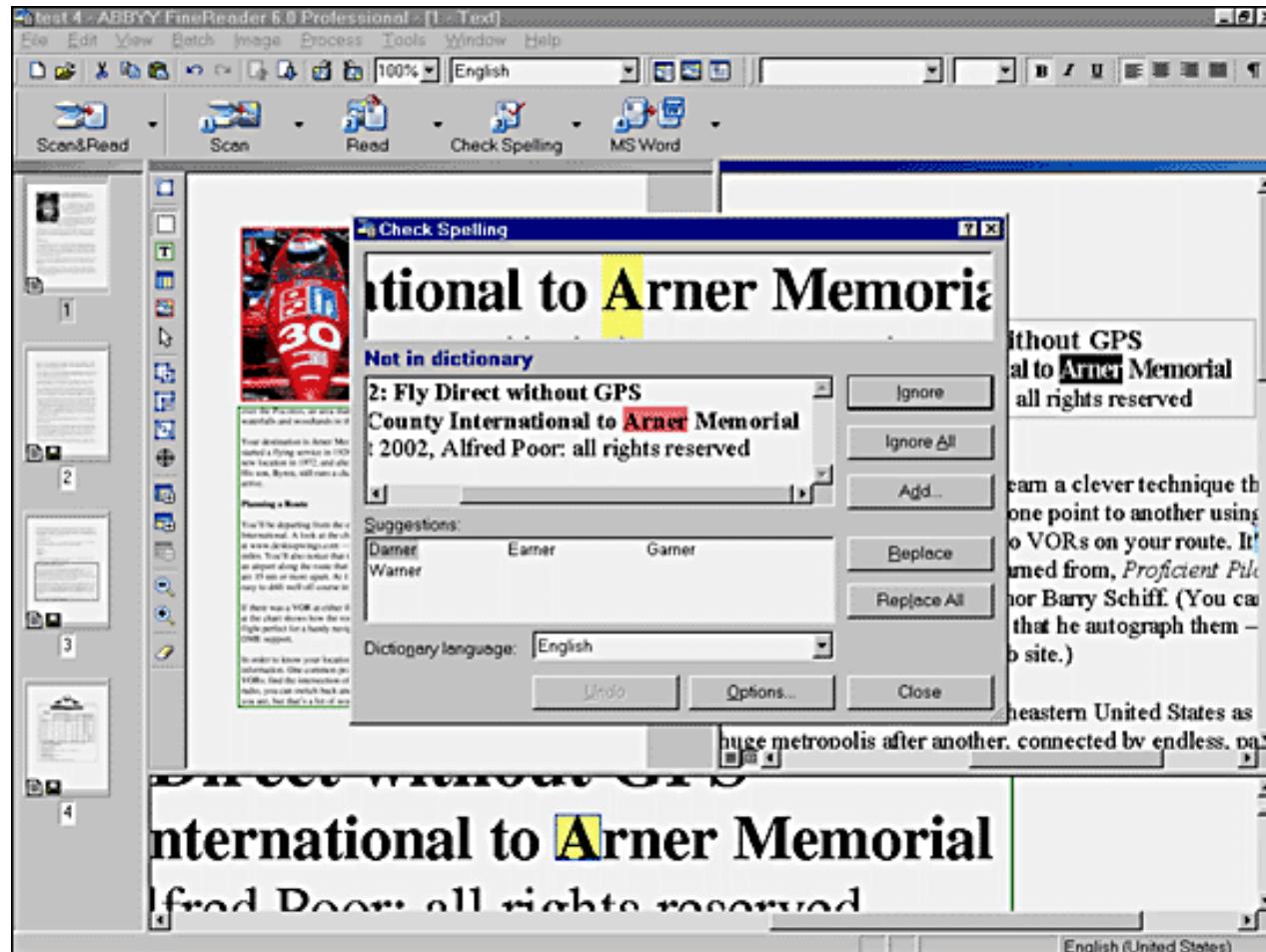
Мы строим прогнозирующий алгоритм:

$$y_i = a(x_i)$$

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?

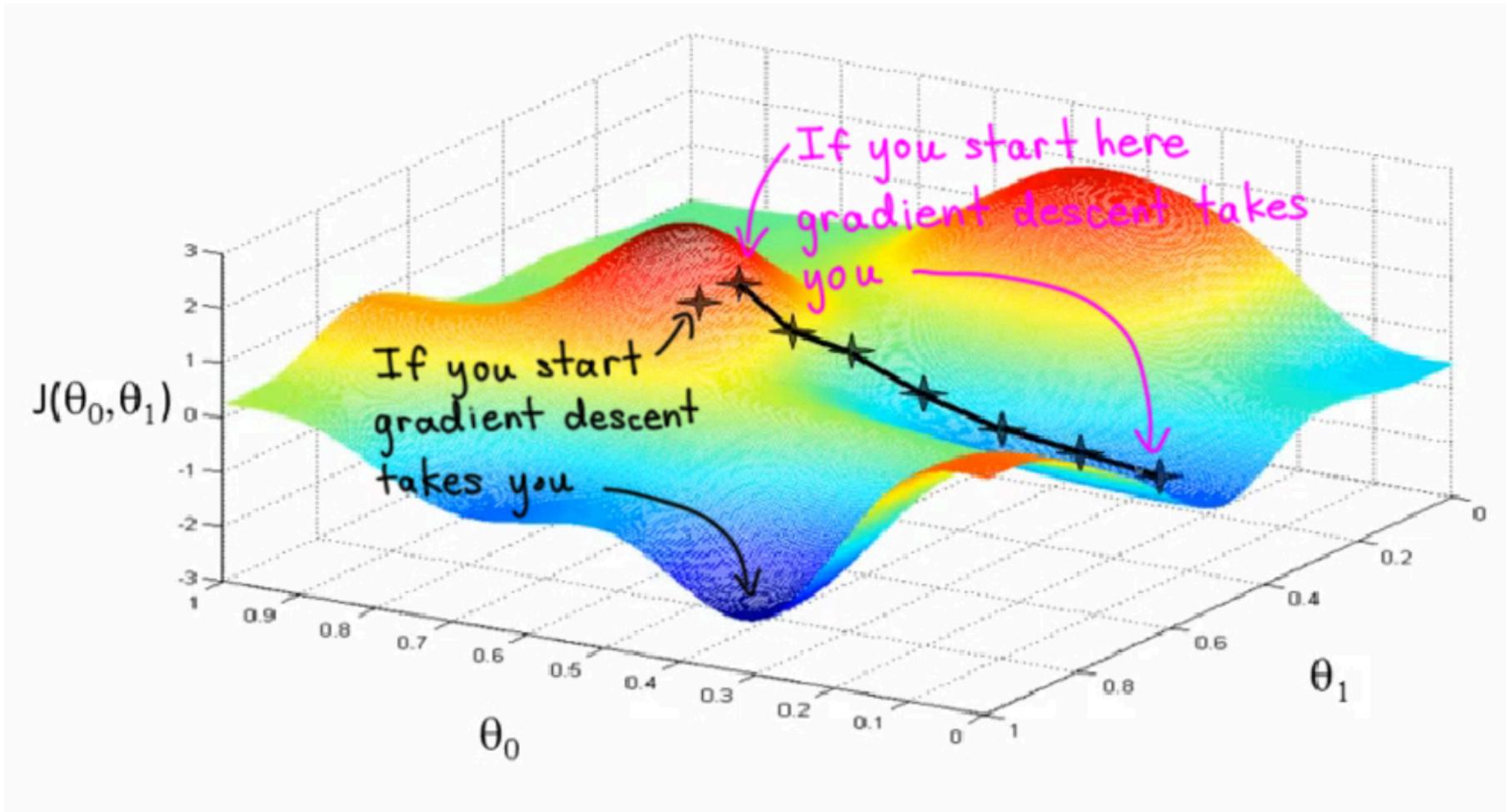
Сложный пример: исправление опечаток

Возможное решение:

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

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

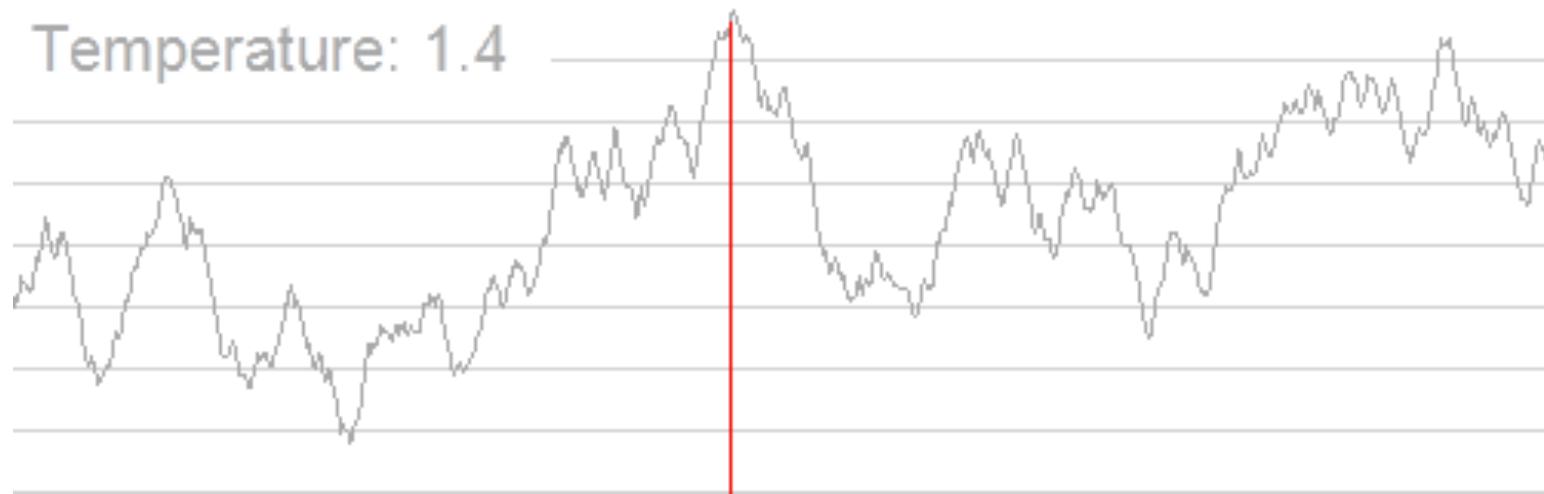
Градиентные методы оптимизации



Методы глобальной оптимизации

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

Методы глобальной оптимизации



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

Признаки

Feature engineering

- Выделение признаков (feature extraction) – генерация признаков по известным данным
- Отбор признаков (feature selection) – ранжирование признаков по «полезности» и выкидывание наименее полезных (или даже наоборот «вредных»)
- Преобразование признаков (feature transform) – создание новых признаков на основе имеющихся

Пример 1: текстовые признаки

- Dataset: 20news_groups
- Электронные письма, разбитые по 20 темам (классам)
- Попробуем придумать классификатор, который различает две темы:
auto и **politics.mideast**

Извлечение текстовых признаков

- Пример письма 1:

From: carl_f_hoffman@cup.portal.com

Newsgroups: rec.autos

Subject: 1993 Infiniti G20

Message-ID: <78834@cup.portal.com>

Date: Mon, 5 Apr 93 07:36:47 PDT

Organization: The Portal System (TM)

Lines: 26

I am thinking about getting an Infiniti G20.

In consumer reports it is ranked high in many
categories including highest in reliability index for compact cars.
Mitsubishi Galant was second followed by Honda Accord.)

Извлечение текстовых признаков

- Пример письма 2:

From: Bob.Waldrop@f418.n104.z1.fidonet.org (Bob Waldrop)

Subject: Celebrate Liberty! 1993

Message-ID: <1993Apr5.201336.16132@dsd.es.com>

Followup-To:talk.politics.misc

Announcing... Announcing... Announcing... Announcing...

CELEBRATE LIBERTY!
1993 LIBERTARIAN PARTY NATIONAL CONVENTION
AND POLITICAL EXPO

THE MARRIOTT HOTEL AND THE SALT PALACE

SALT LAKE CITY, UTAH

Текстовые признаки: bag-of-words



The world of **TOTAL**

all about the company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Basin complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

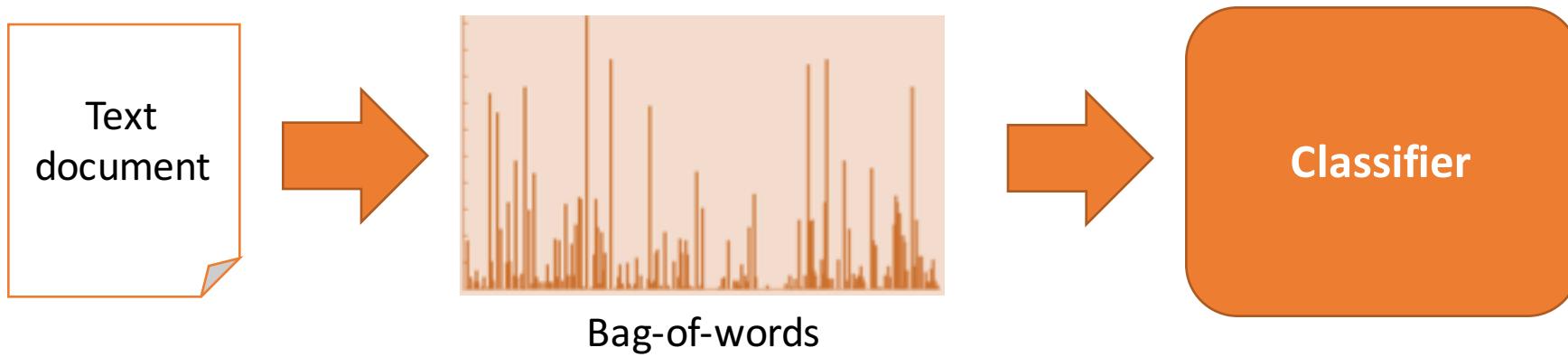
All About The Company

- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage

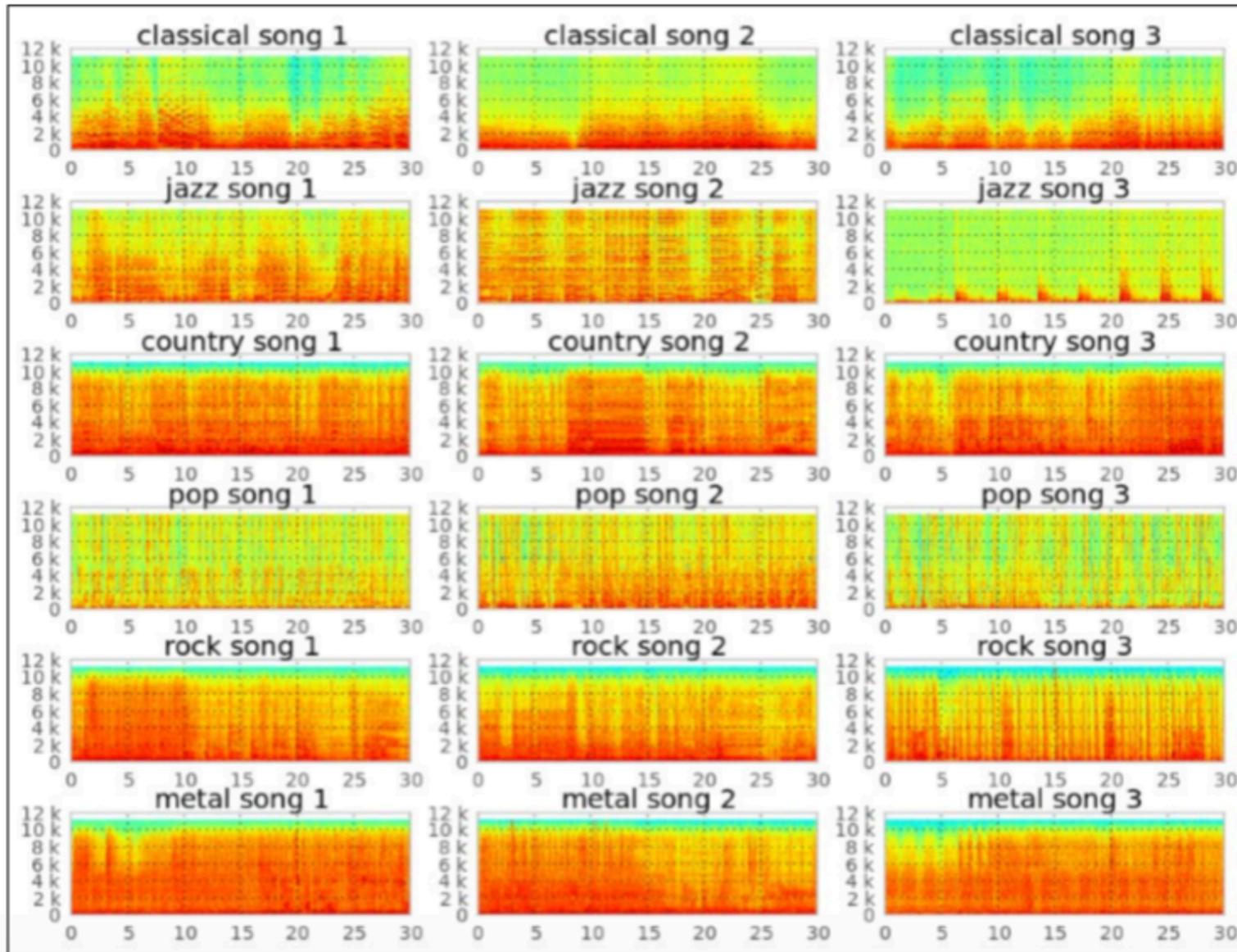


aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

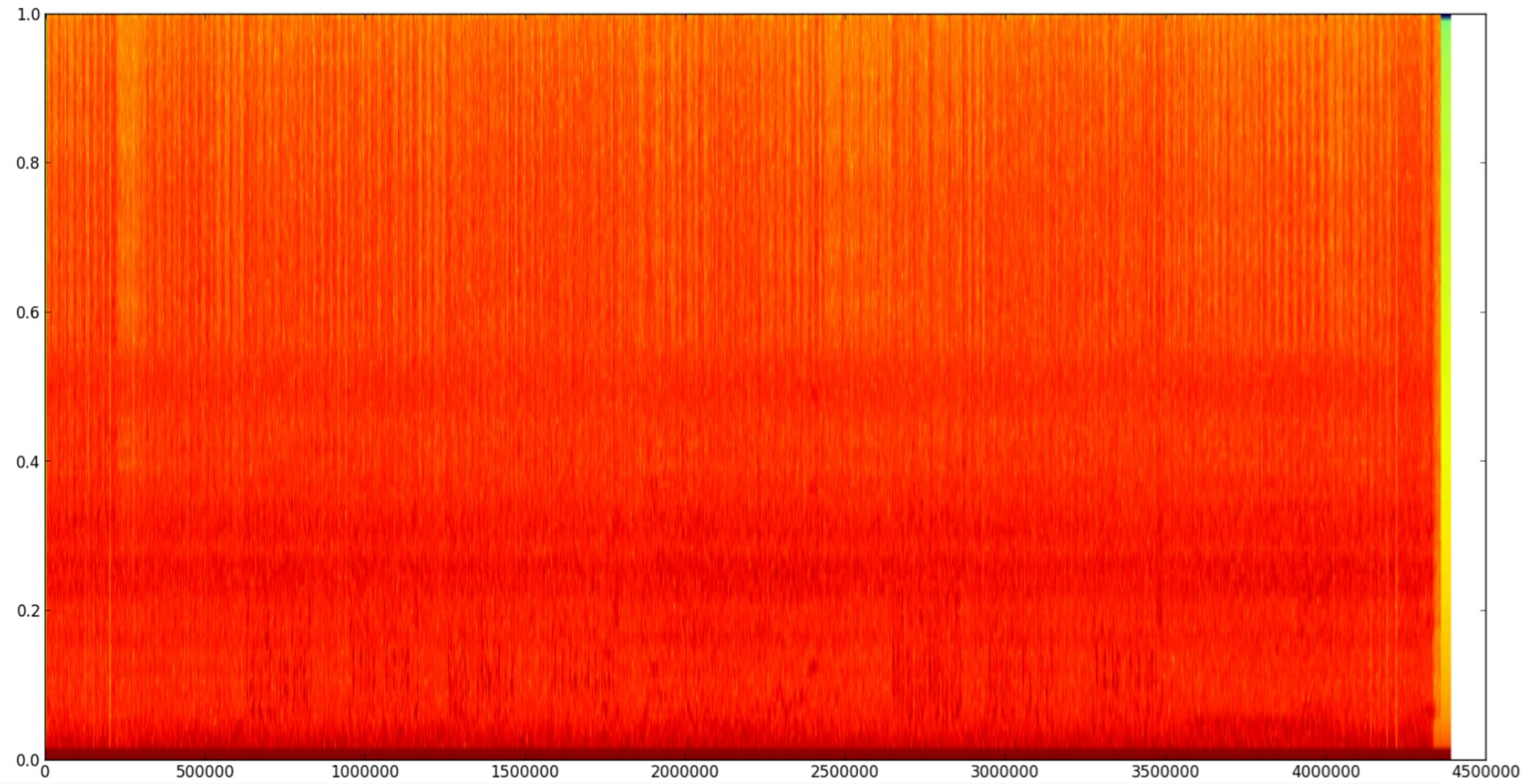
Самый простой классификатор текстов



Пример 2: признаки аудиофайла



Пример 2: признаки аудиофайла

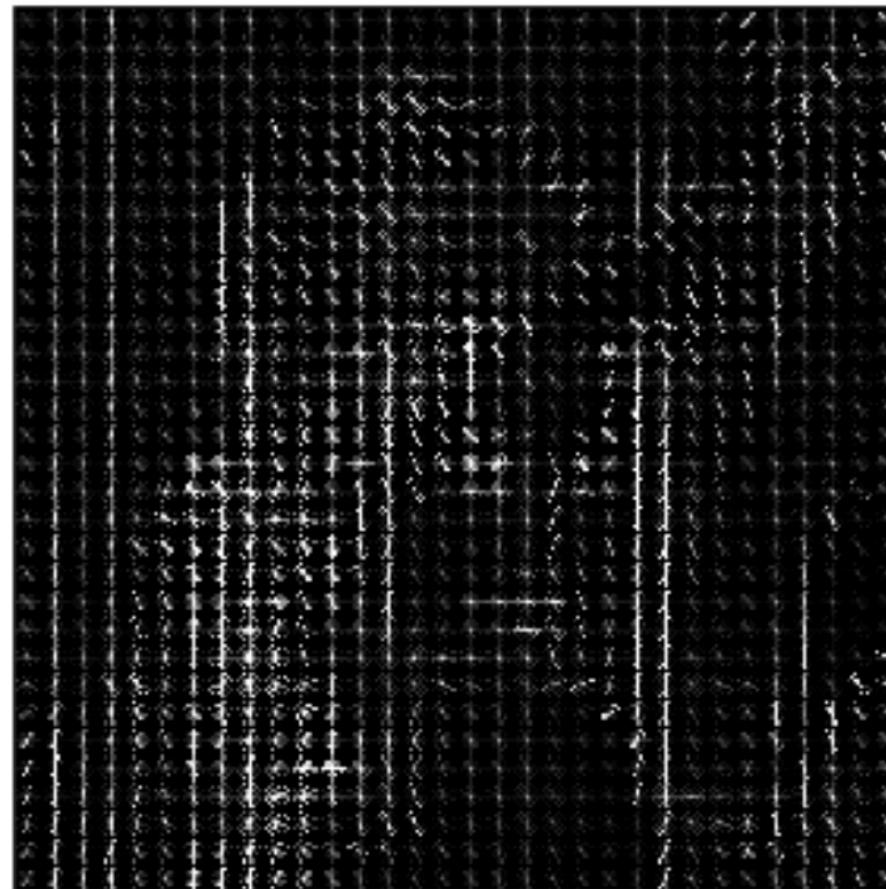


Пример 3: признаки изображения

Input image



Histogram of Oriented Gradients



Пример 4: категориальные признаки

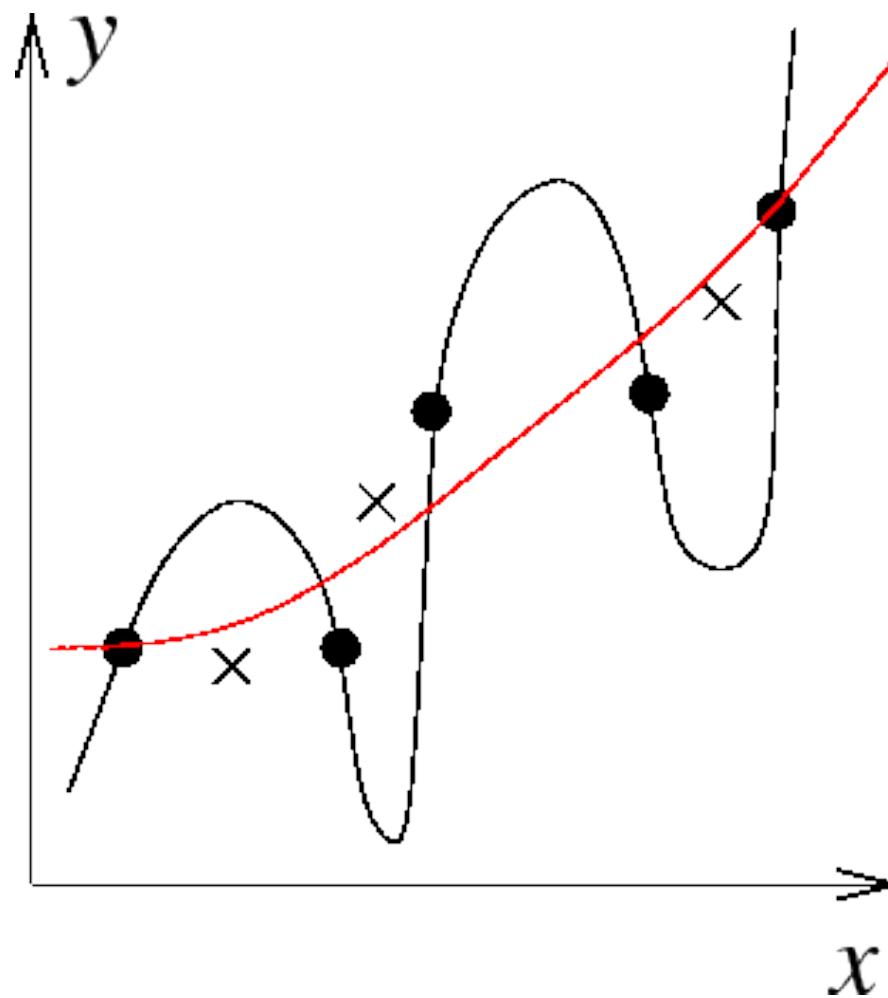
№ склада	Город	...	Продано вина (ящиков)
2343	Москва	...	56000
185	Самара	...	10500
121	Ростов	...	11300
...

Пример 4: категориальные признаки

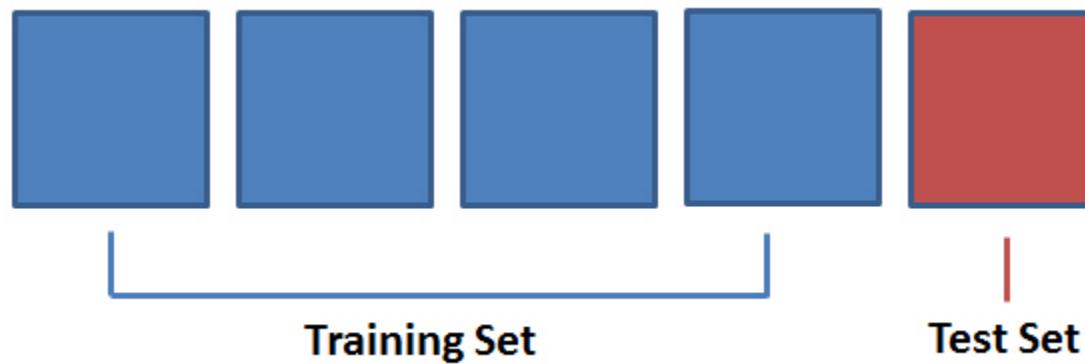
№ склада	В среднем продают в городе	...	Продано вина (ящиков)
2343	59000	...	56000
185	11200	...	10500
121	12100	...	11300
...

Переобучение

Переобучение на примере регрессии

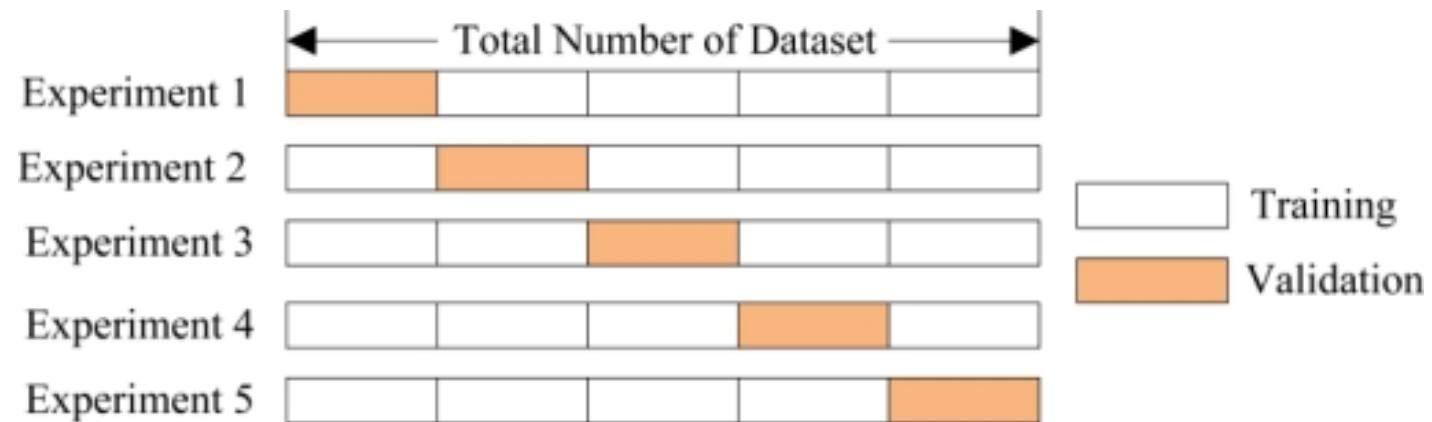


Оценка качества



Кросс-валидация

K-Fold cross validation:



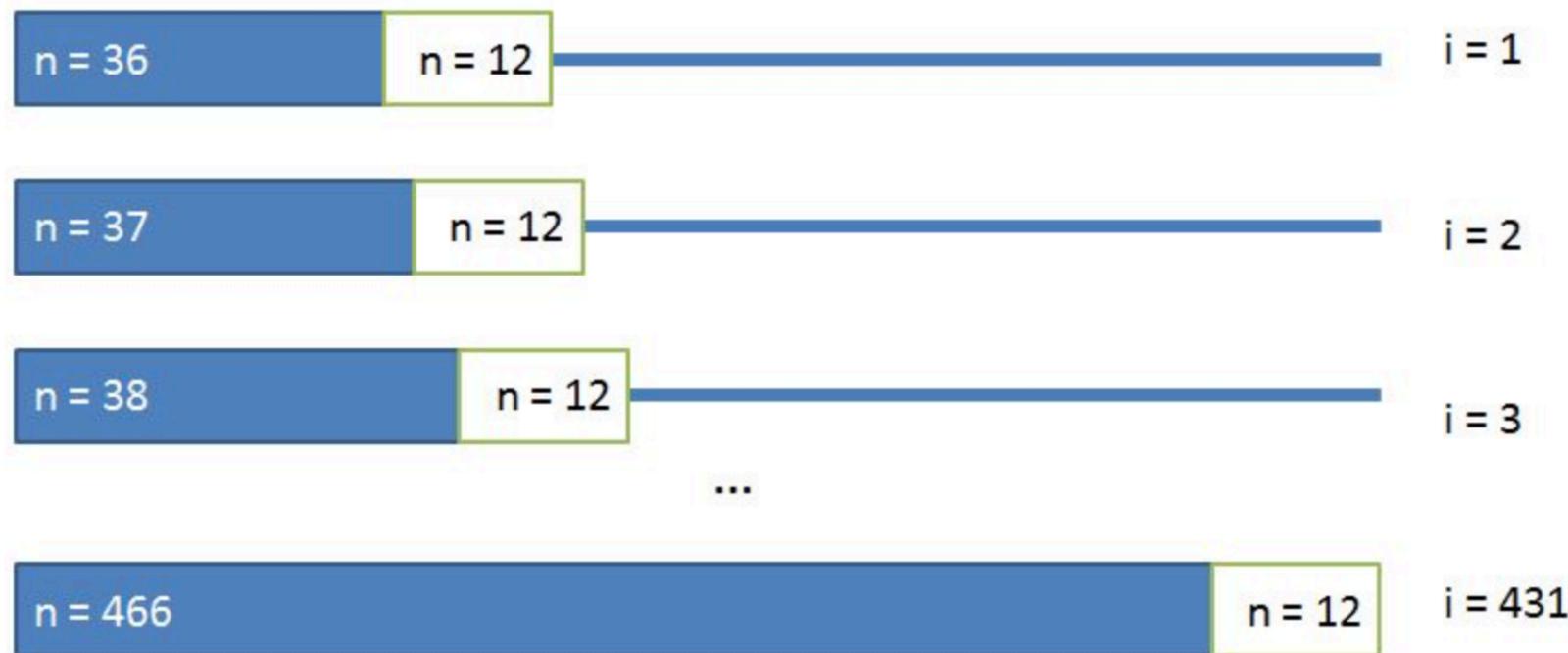
На картинке $k = 5$, обычно такое k и используют. Другие частые варианты – 3 и 10.

Кросс-валидация и данные «из будущего»

N = 478 (month-end data)

June 1967

March 2007



История про танки



Классификатор: есть танки на снимке или нет

История про танки



Классификатор: есть танки на снимке или нет

Часть III: инструменты

Python

На чем будут примеры

- Python 2.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

The screenshot shows the official scikit-learn website. At the top left is the logo, followed by a navigation bar with links for Home, Installation, Documentation (with a dropdown arrow), Examples, a Google Custom Search bar, and a Search button. Below the navigation is a large blue section featuring a grid of nine small plots demonstrating various machine learning models. The first plot is labeled "Input data". The subsequent eight plots show different models: Nearest Neighbors, Linear SVM, RBF SVM, Gaussian Process, Decision Tree, Random Forest, and Neural Net. Each model plot includes a numerical accuracy score (e.g., 97, 98, 95, 96, 78, 82, 90). Below the grid is a horizontal navigation bar with a left arrow, a series of blue dots, and a right arrow.

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

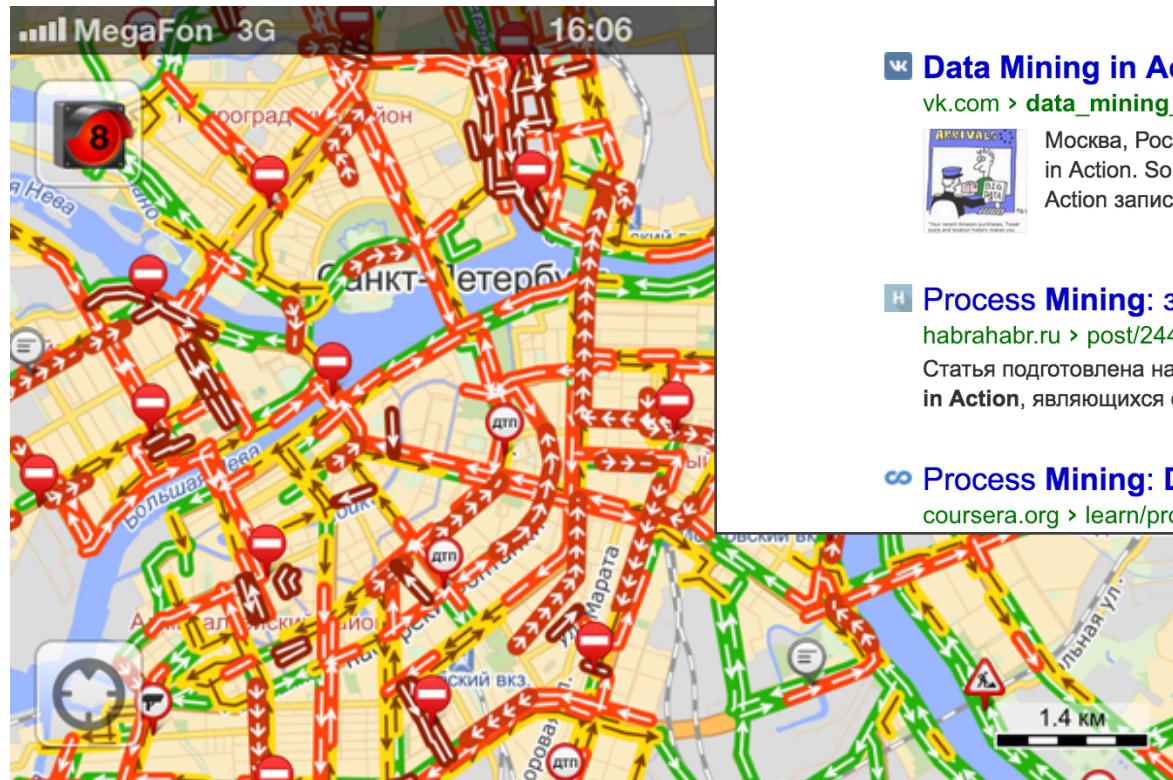
Машинное обучение в несколько строк

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

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

Миссия курса

Предпосылки



Яндекс Data Mining in Action Найти Войти

ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ ДИСК МУЗЫКА ЕЩЁ

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

habrahabr.ru [Process Mining: знакомство](#)
habrahabr.ru > post/244879/ ▾
Статья подготовлена на основе материалов книги Data Mining in Action, являющихся собственностью издательства

coursera.org [Process Mining: Data science for process mining](#)
coursera.org > learn/process-mining

Нашлось 8 млн результатов
Дать объявление Показать все

Задача: нести культуру в массы



Спасибо за внимание!