

Анализ данных на практике

Занятие 1. Введение

Кантор Виктор
Осенний семестр 2015 года

Кто ведёт лекции

Работа:

Yandex Data Factory (Data Scientitst), в прошлом: ABBYY (Head of learning group, R&D), сооснователь 2Long2Read и SmartTagger

Преподавание:

- МФТИ, спецкурс по анализу данных: с 02.2012
- МФТИ, Машинное обучение: с 02.2013
- Яндекс, Автоматическая обработка текстов, с 09.2014
- МФТИ, Анализ изображений: с 09.2014
- МФТИ, Автоматическая обработка текстов: с 02.2015
- СберТех, Машинное обучение: с 09.2015
- СберТех, Python и библиотеки для анализа данных: с 02.2015

Область интересов:

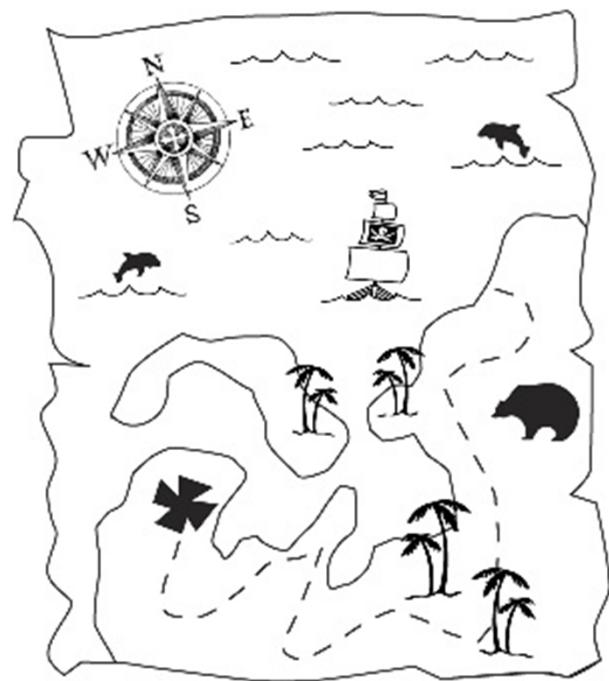
Анализ текстов, некоторые темы анализа изображений,
рекомендательные системы, приложения машинного обучения в бизнесе

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, в качестве IDE рекомендую PyCharm



План



1. Пример задачи машинного обучения
2. Стандартные задачи
3. Работа с признаками
4. Переобучение
5. Задачи на сообразительность
6. Философские вопросы
7. Что будет дальше

1. Пример

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	1	0	0	1	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	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	0	0	1	0	0	1	0	0	1	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	0	1	1
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	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	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	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	60	1	0	0	0	0	05	0	1	1	1	1	0	0	1	0	1	0	0	0	0	1	
4	12	2	30	Attribute 1: Status of existing checking account																					
2	12	2	12	1	... < 0 DM																				
1	48	1	24	2	0 <= ... < 200 DM																				
1	24	1	15	1	... >= 200 DM /																				
1	24	4	24	1	salary assignments for at least 1 year																				
1	30	2	24	1	4																				
4	24	4	9	2	no checking account																				
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	1	0	0	1	1	1	
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	0	1	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	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)



		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	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	0	1	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0	0	0	1
4	12	2	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	1	0	0	1
2	30	4	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	1	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	1	0	0	1	0	1	0	0	0	1
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	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	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	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	0	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	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	0	1
4	10	4	21	5	3	4	1	3	26	3	2	1	1	2	0	0	1	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	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	1	0	0	1	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	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	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	0	1	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	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	1	1
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	1	1	1	0	0	1	0	0	1	2
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	1	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

Answer: 1 – Good, 2 - Bad

Выдача кредита

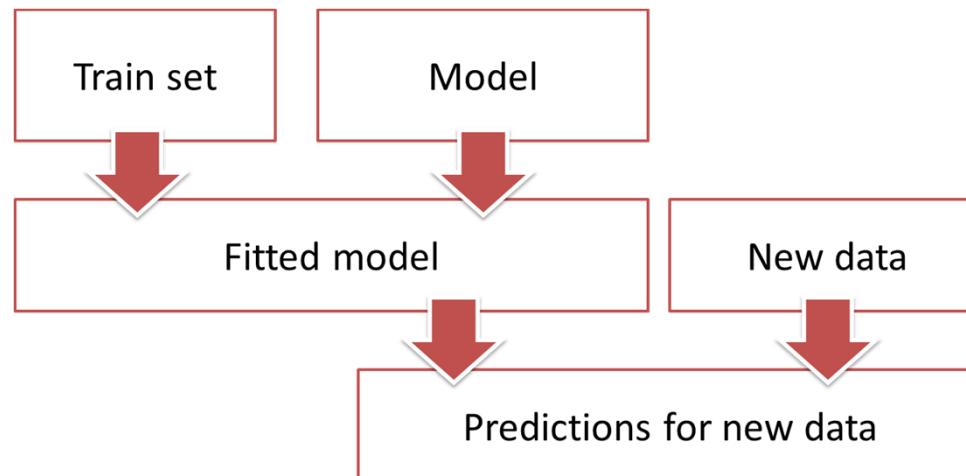
Задача (supervised classification): предсказать класс (1 или 2)

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	0	1	0	0	1	?
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



2. Стандартные задачи

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)

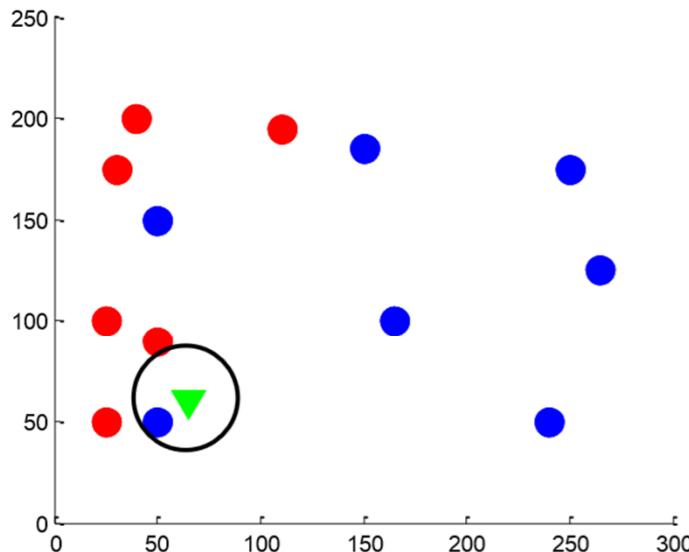
Классификация: обучающая выборка

Fisher's Iris Data

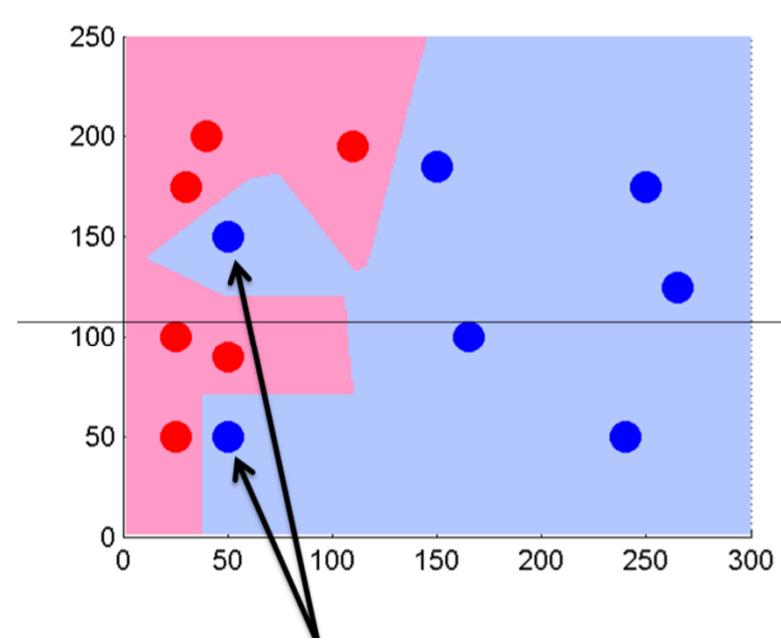
Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	<i>I. versicolor</i>
6.0	2.2	5.0	1.5	<i>I. virginica</i>
6.2	2.2	4.5	1.5	<i>I. versicolor</i>
6.0	2.2	4.0	1.0	<i>I. versicolor</i>
6.3	2.3	4.4	1.3	<i>I. versicolor</i>
5.5	2.3	4.0	1.3	<i>I. versicolor</i>
5.0	2.3	3.3	1.0	<i>I.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



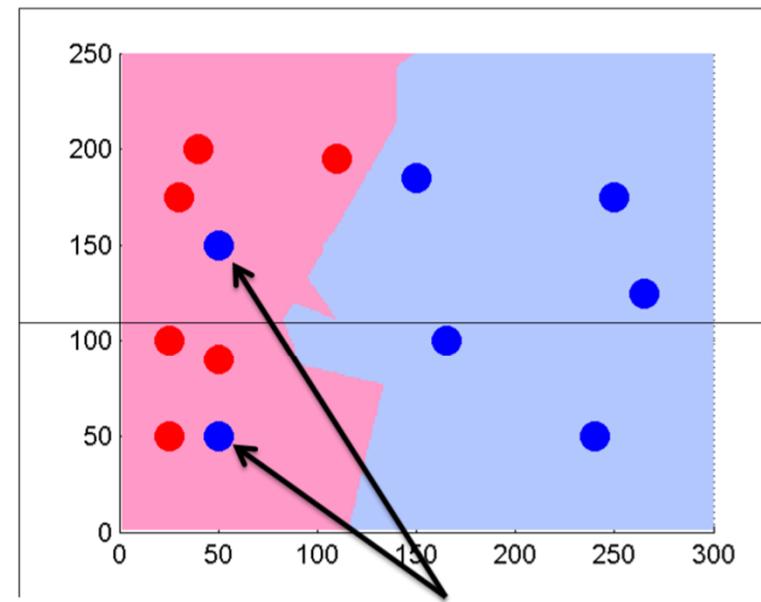
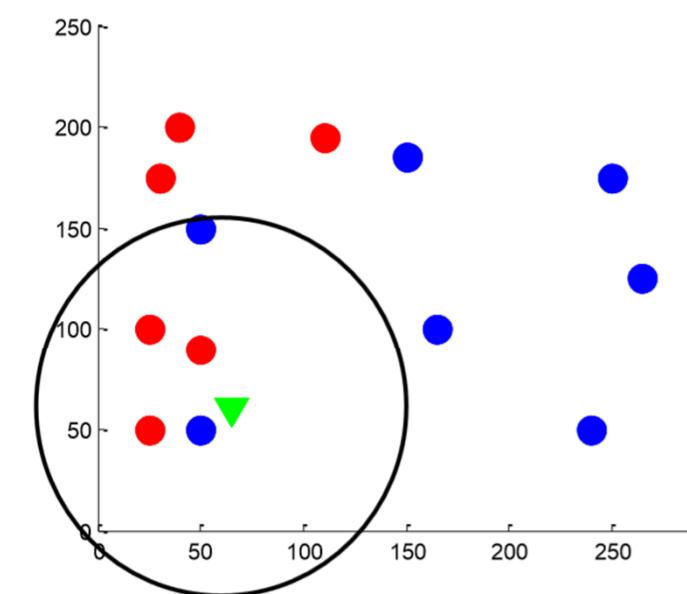
$k = 1$



Шумы? (outliers)

Простой классификатор: kNN

k nearest neighbours

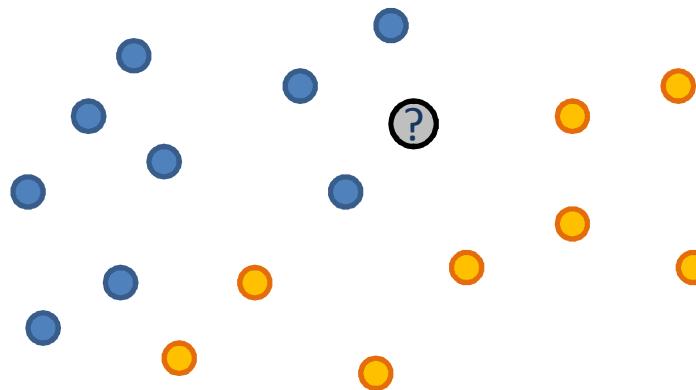


Ошибки?

$$k = 5$$

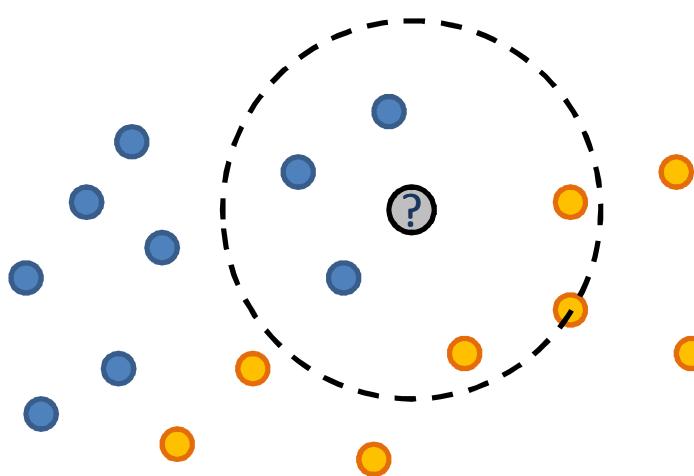
Взвешенный kNN

Пример классификации ($k = 6$):



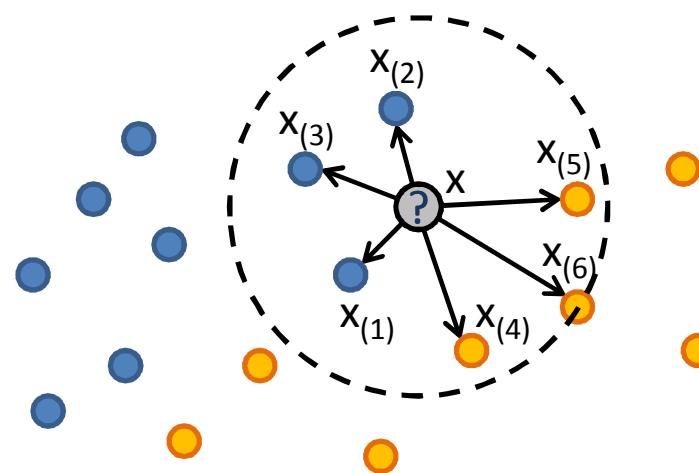
Взвешенный kNN

Пример классификации ($k = 6$):



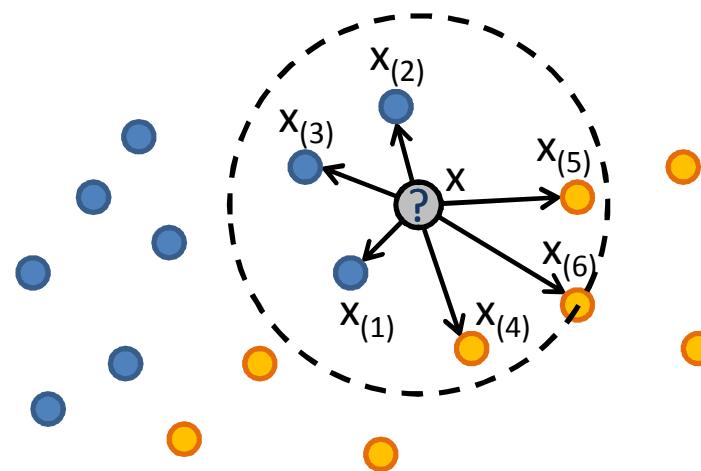
Взвешенный kNN

Пример классификации ($k = 6$):



Взвешенный kNN

Пример классификации ($k = 6$):

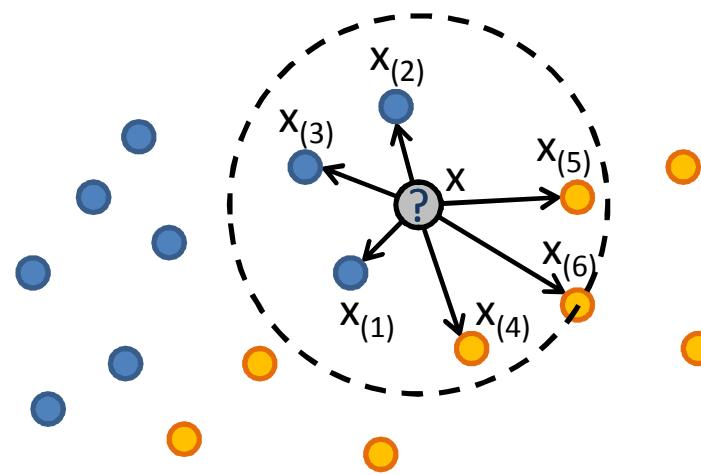


Веса можно определить как функцию от соседа или его номера:

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

Взвешенный kNN

Пример классификации ($k = 6$):



Веса можно определить как функцию от соседа или его номера:

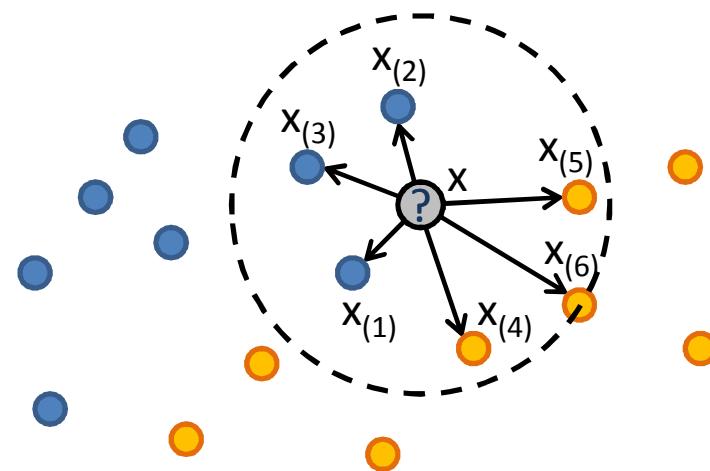
$$w(x_{(i)}) = w(i)$$

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:

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

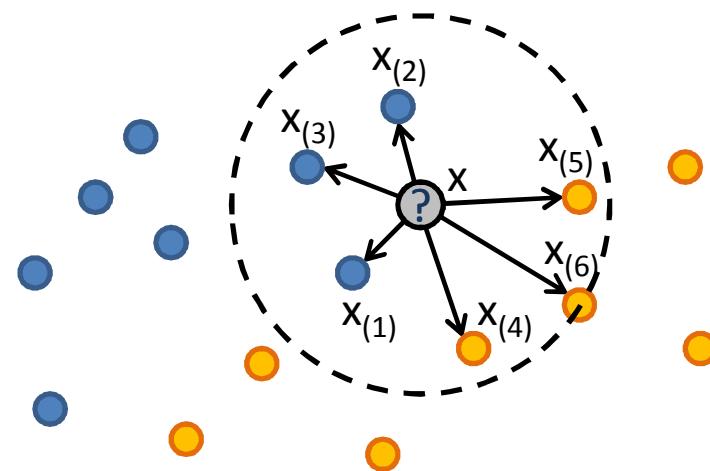
или как функцию расстояния:

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

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

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:

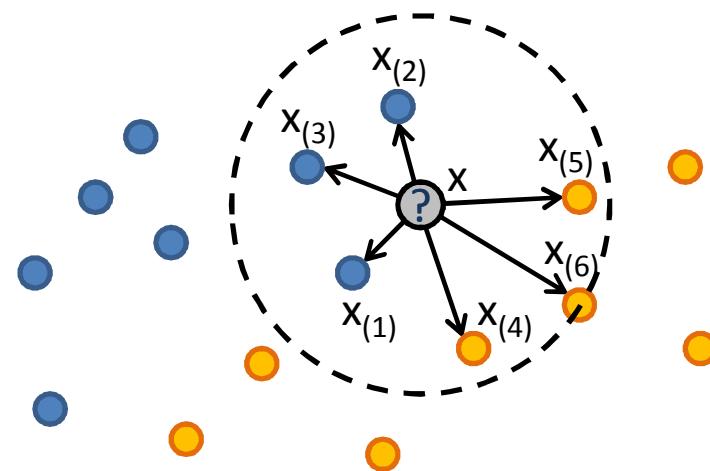
$$w(x(i)) = w(d(x, x_{(i)}))$$

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

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

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:

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

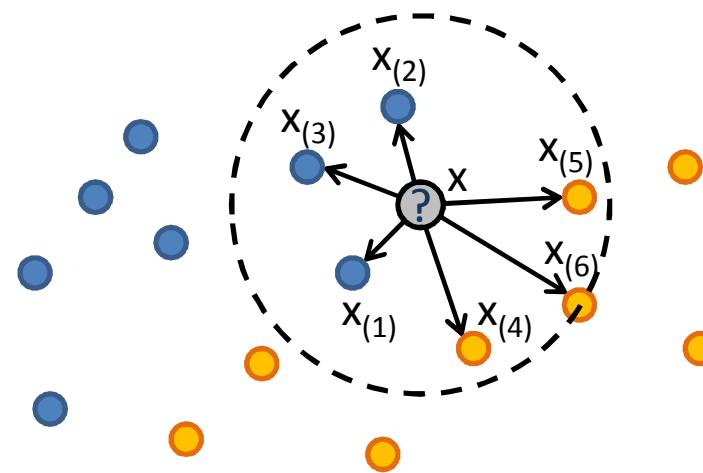
$$Z_{\text{blue}} = \frac{\text{blue shaded area}}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

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

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

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:

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

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

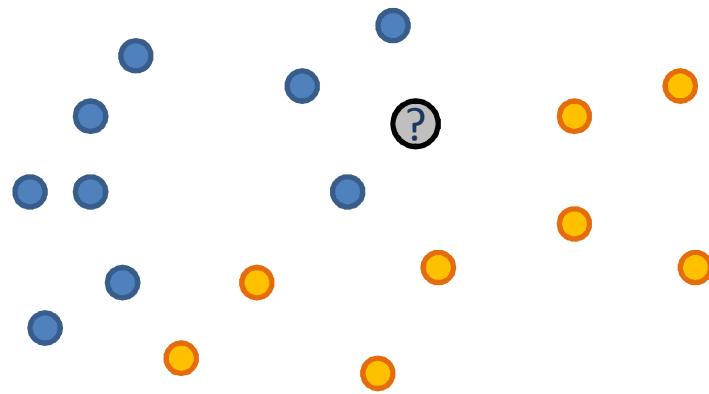
$$Z_{\text{orange}} = \frac{\text{orange shaded area}}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

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

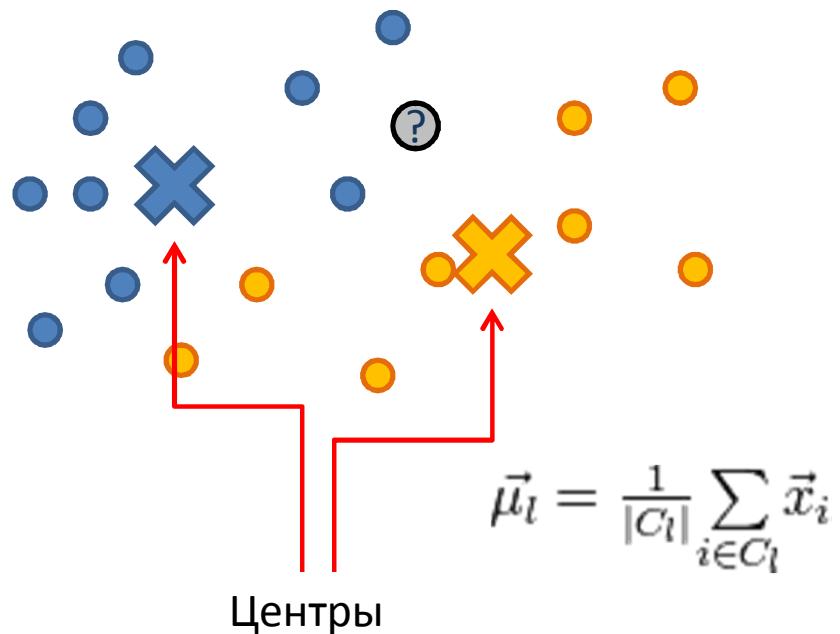
$$\text{if } Z_{\text{orange}} > Z_{\text{blue}} : \quad \text{?} = \text{orange}$$

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

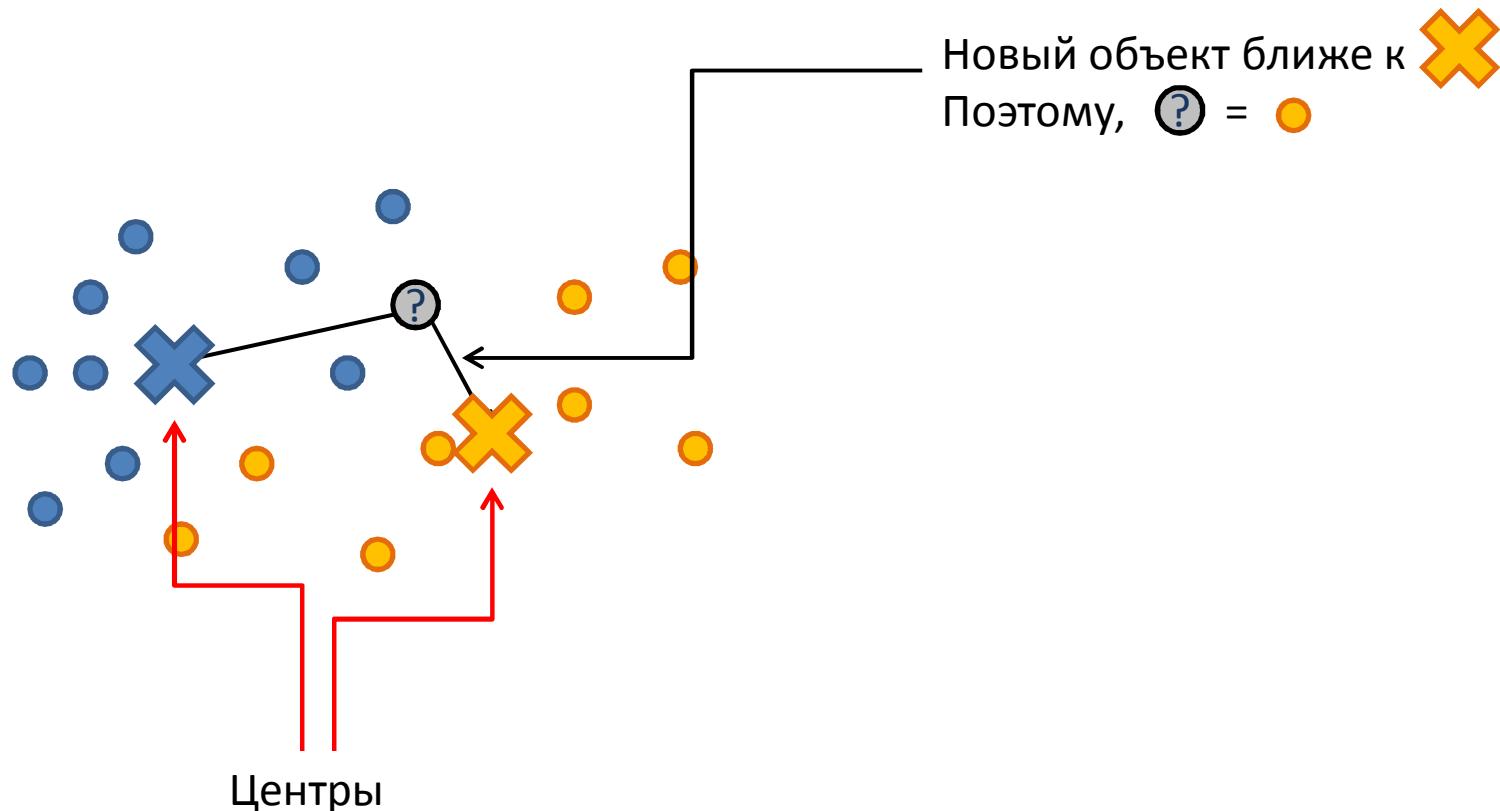
Центроидный классификатор



Центроидный классификатор



Центроидный классификатор



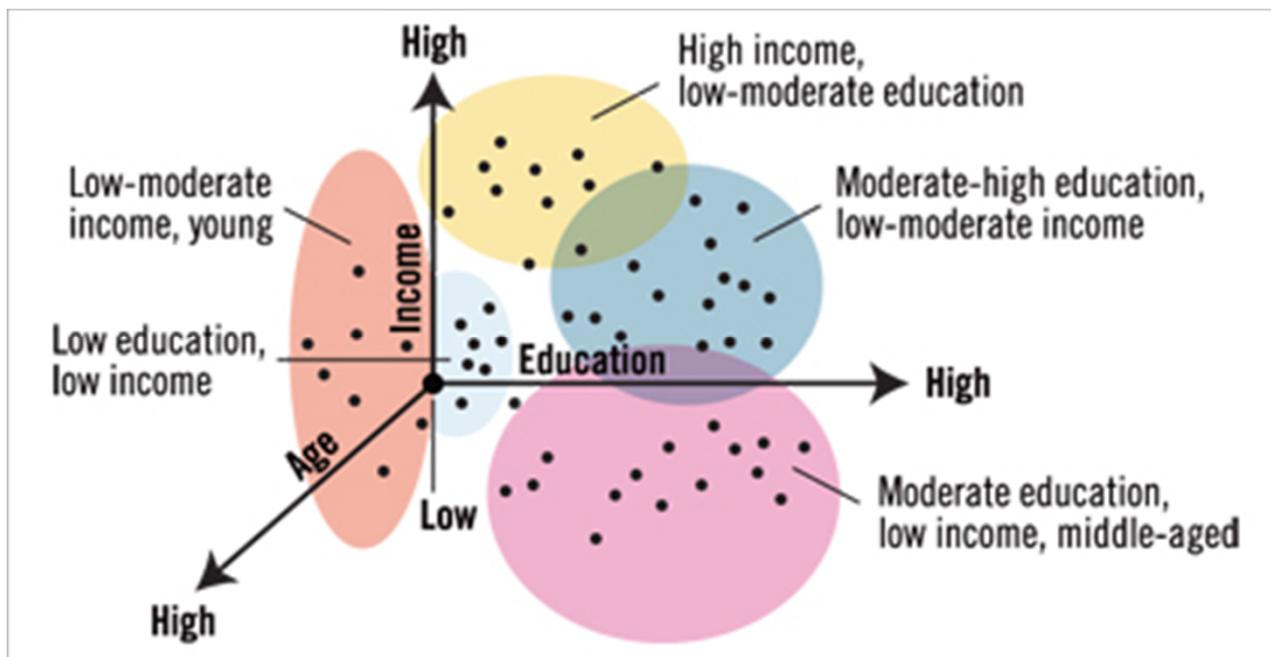
Кластеризация

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

Скриншот с сайта clusters.uk.com:

96% accurate segmentation is how the right marketing reaches the right people

Figure 1. Clusters

96% the right marketing to the right people

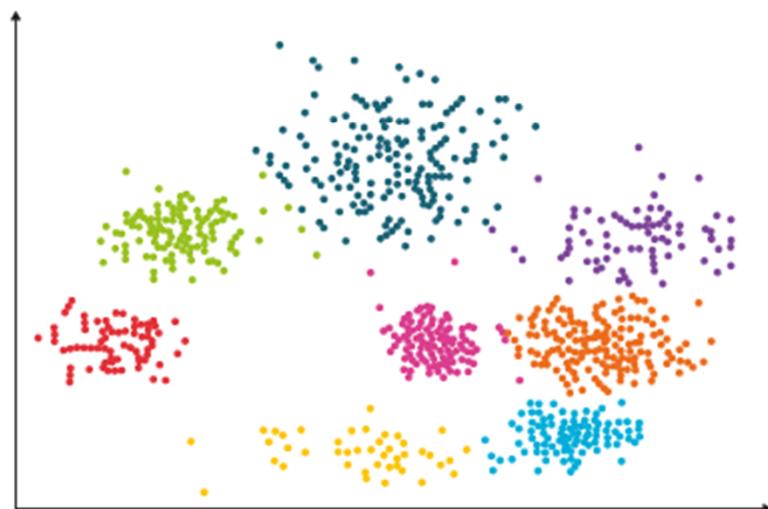


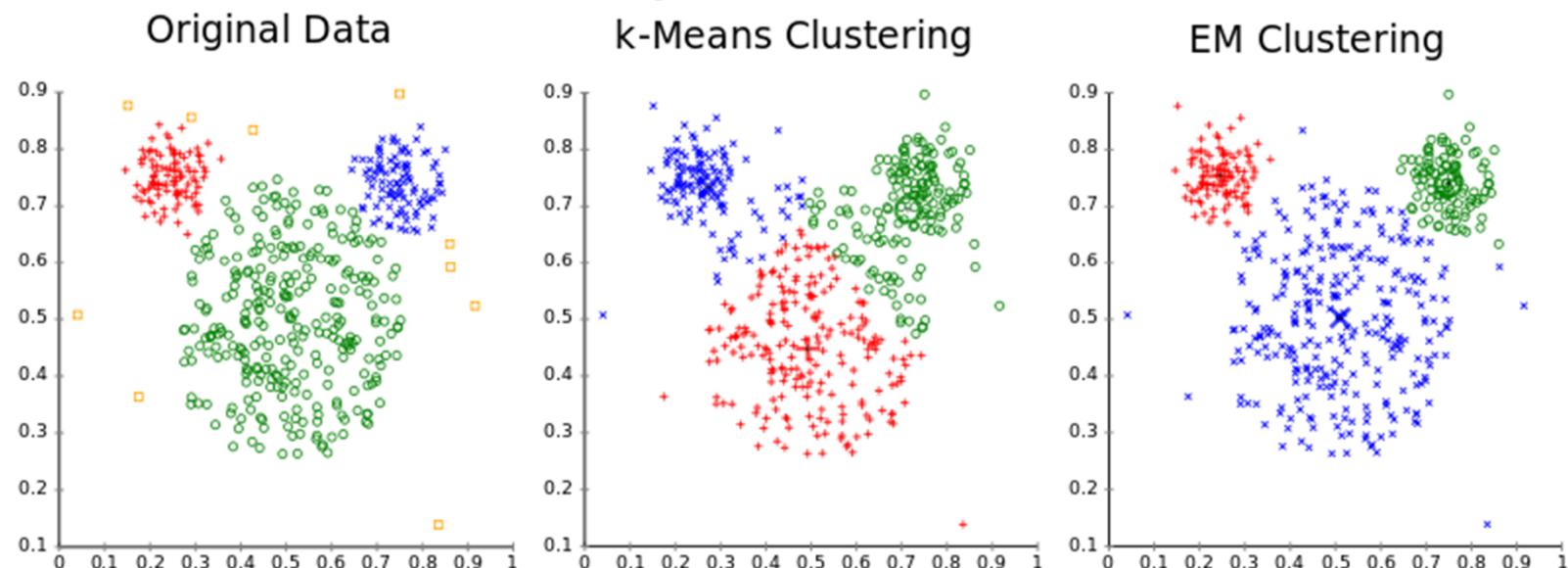
Figure 2. K-Means

62% the right marketing to the right people

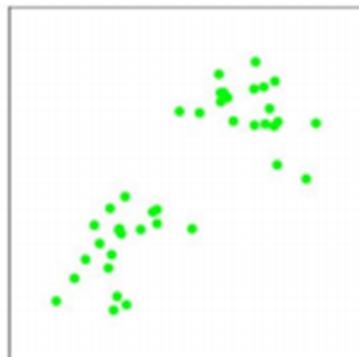


Кластеризация

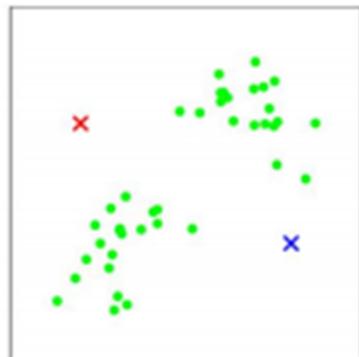
Different cluster analysis results on "mouse" data set:



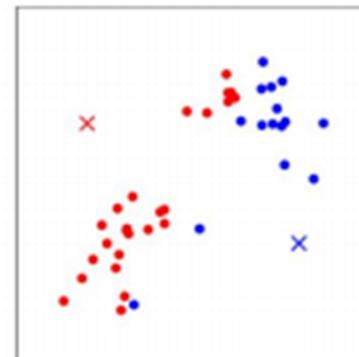
Простой алгоритм кластеризации: kMeans



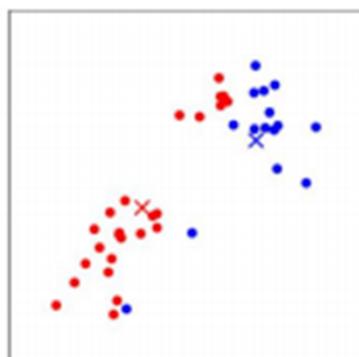
(a)



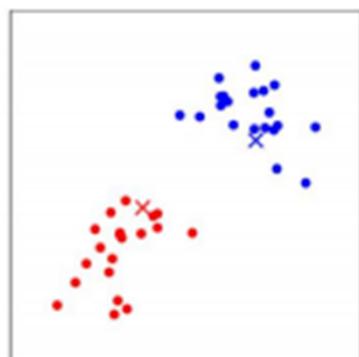
(b)



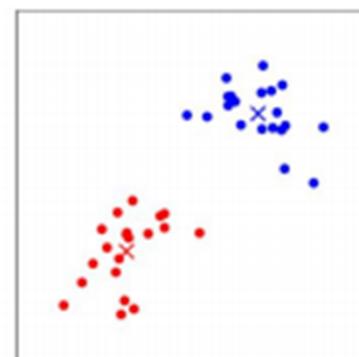
(c)



(d)



(e)



(f)

Регрессия

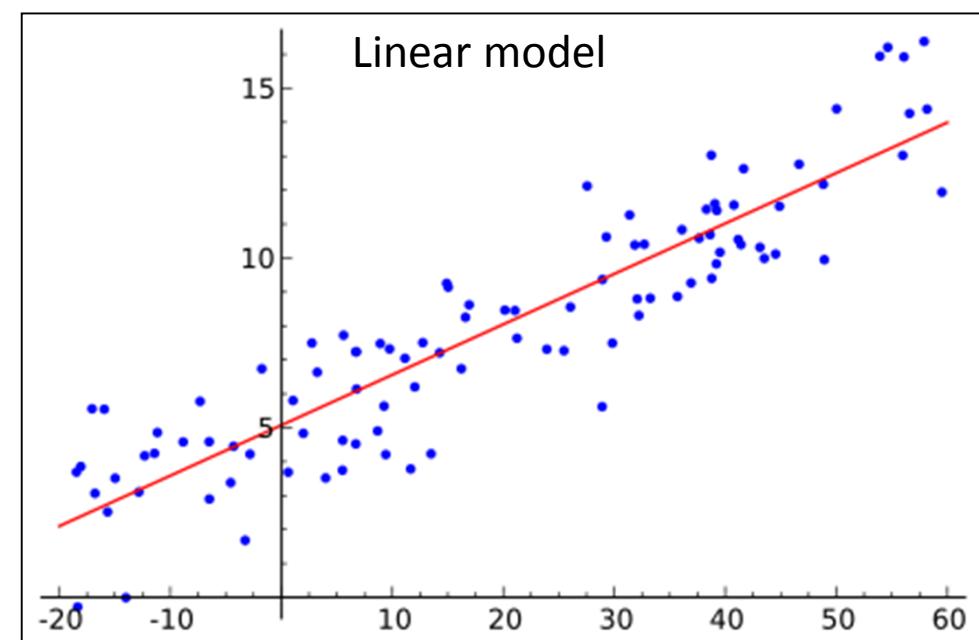
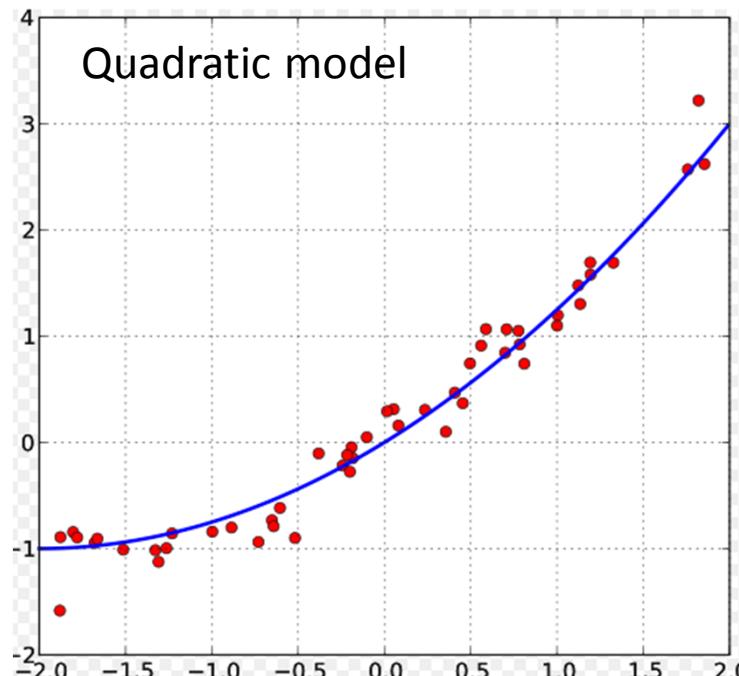
Вход (обучающая выборка):

Признаки N объектов с известными значениями прогнозируемого вещественного параметра объекта

Выход:

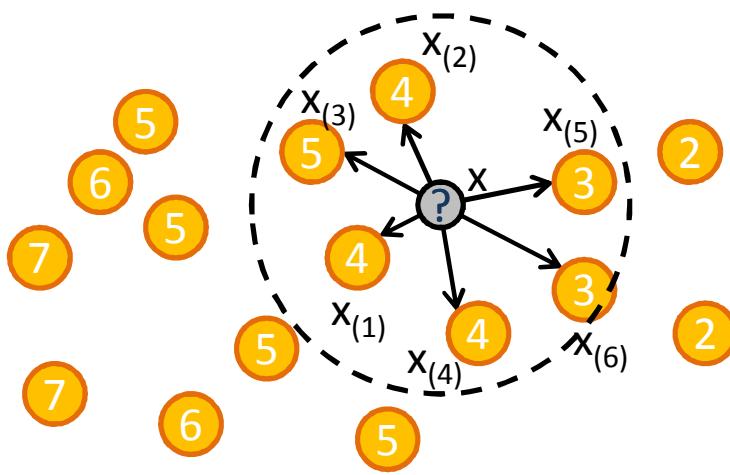
Алгоритм, прогнозирующий значение вещественной величины по признакам объекта

Изображения для случая одного признака: x – признак, y – величина



Взвешенный kNN для регрессии

Пример ($k = 6$):



Веса можно определить как функцию от соседа или его номера:

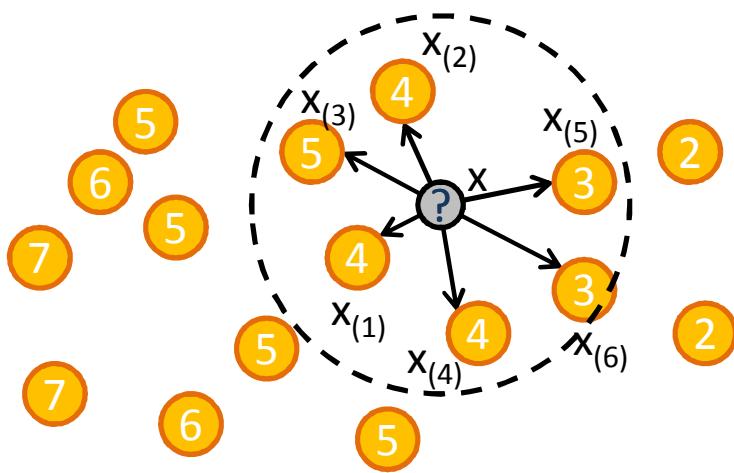
$$w(x_{(i)}) = w(i)$$

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:

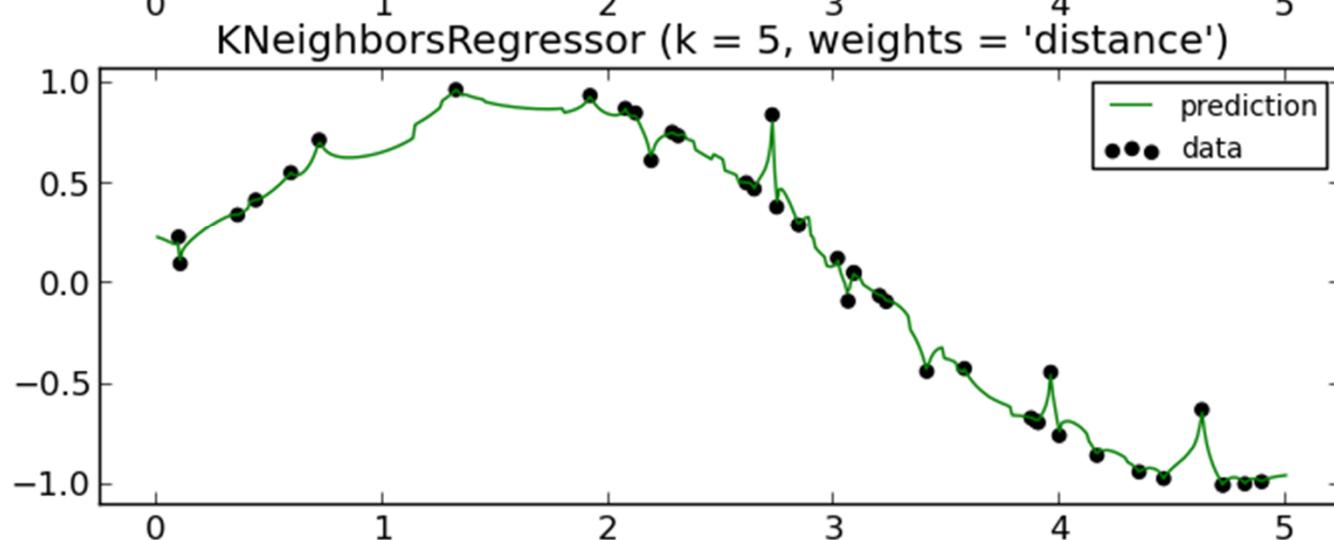
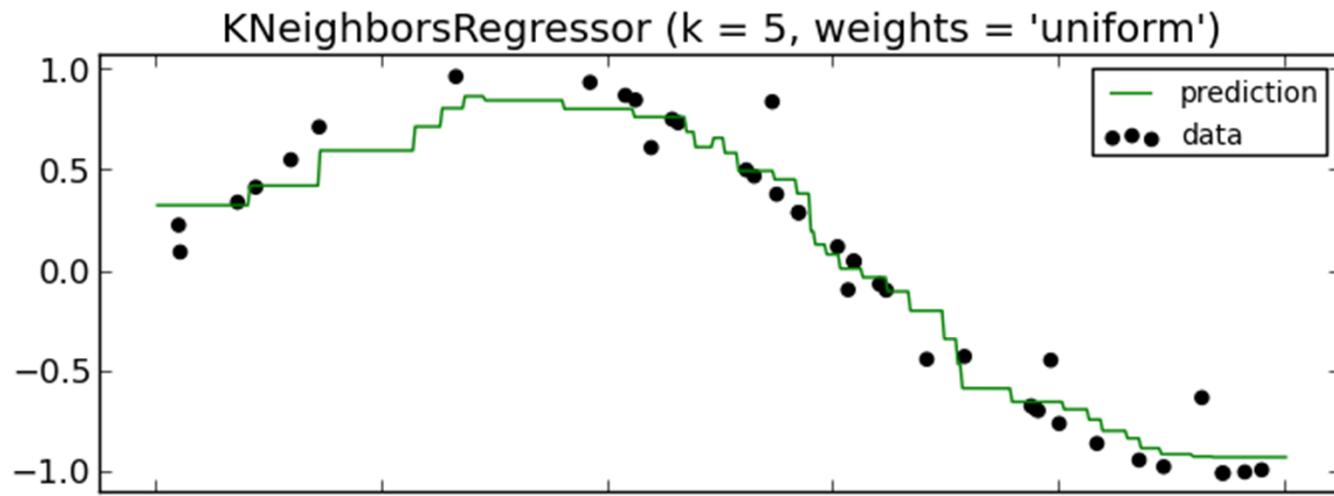
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

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

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

Приимер работы kNN для регрессии



3. Работа с признаками

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: 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).

A couple of things though:

1) In looking around I have yet to see anyone driving this car. I see lots of Honda's and Toyota's.

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

INCLUDES INFORMATION ON DELEGATE DEALS!
(Back by Popular Demand!)

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.

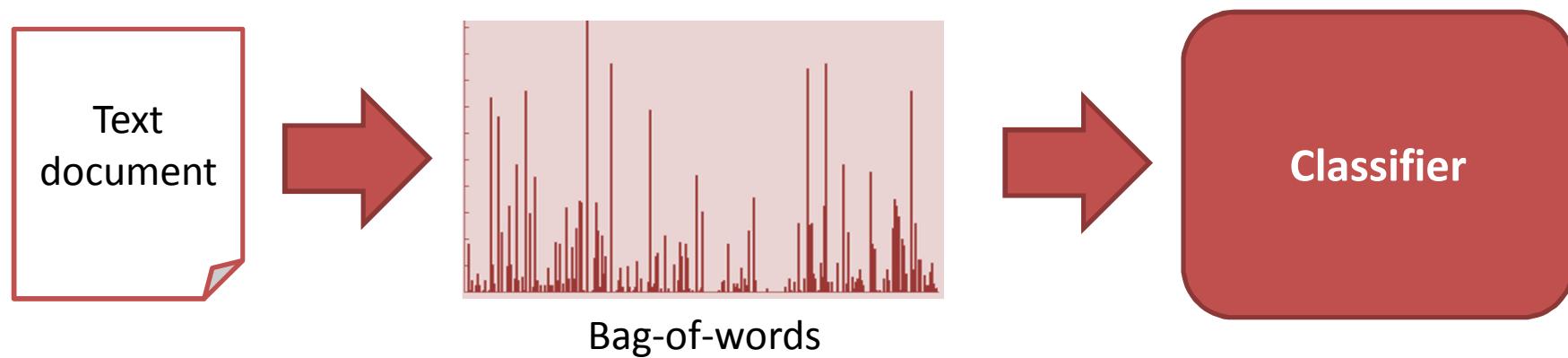
At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

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.

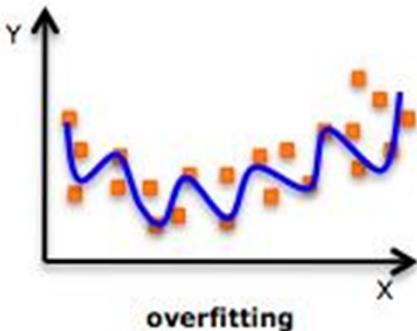
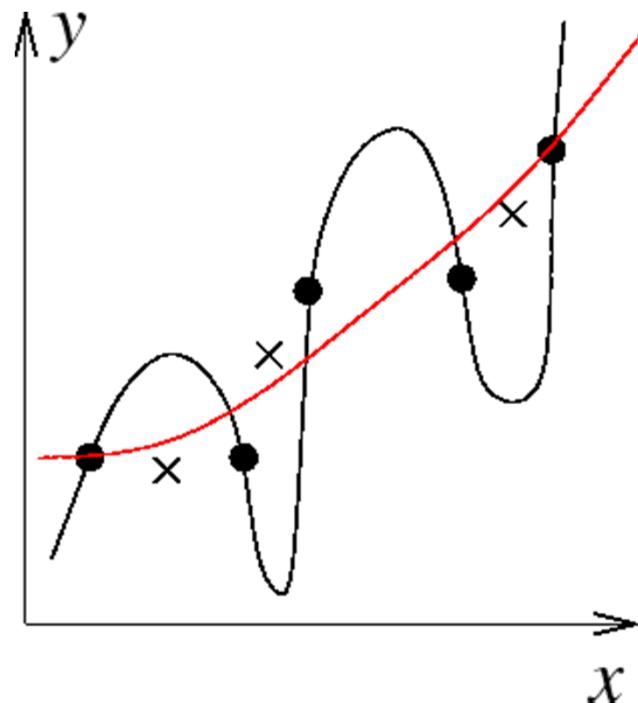
aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

Самый простой классификатор текстов

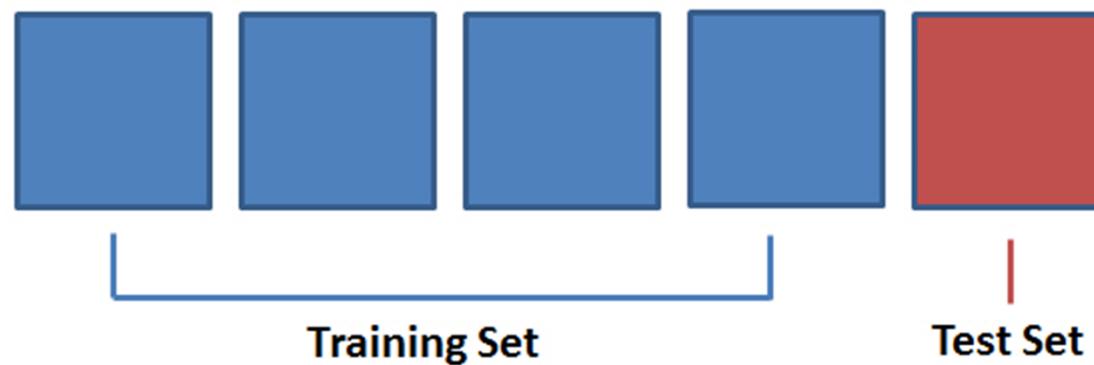


4. Переобучение

Пример: аппроксимация функции одной переменной

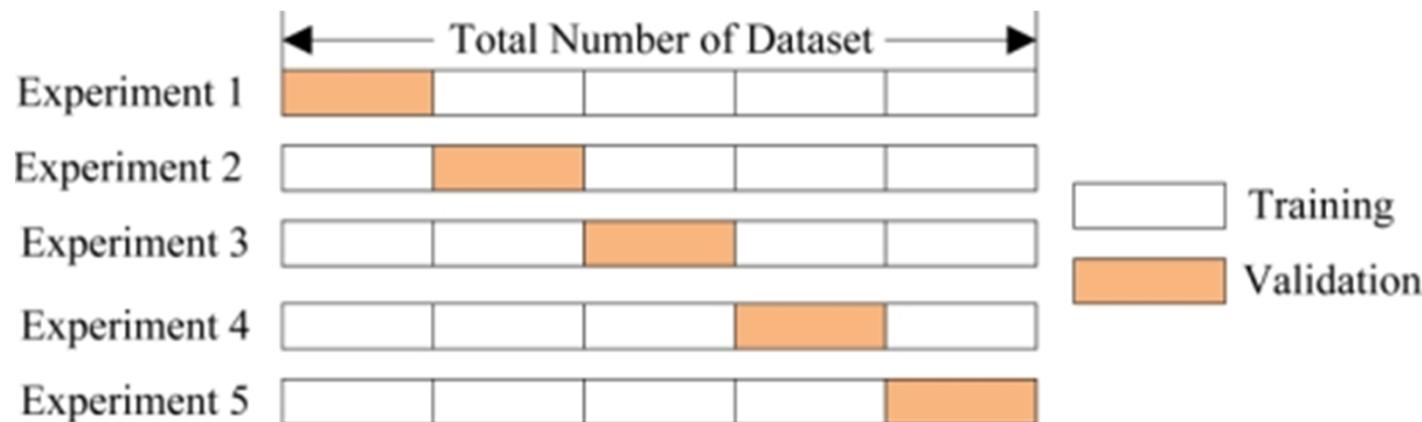


Оценка качества



Кросс-валидация

K-Fold cross validation:



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

5. Задачи на сообразительность

Задача 1

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?

Задача 2

К исследователю обращается сеть магазинов с формулировкой: «Нам кажется, у нас проблемы. И нам кажется, что это очереди». Придумайте, что может сделать исследователь: какие задачи можно решить, какие признаки при этом стоит извлечь.

5. Философские вопросы

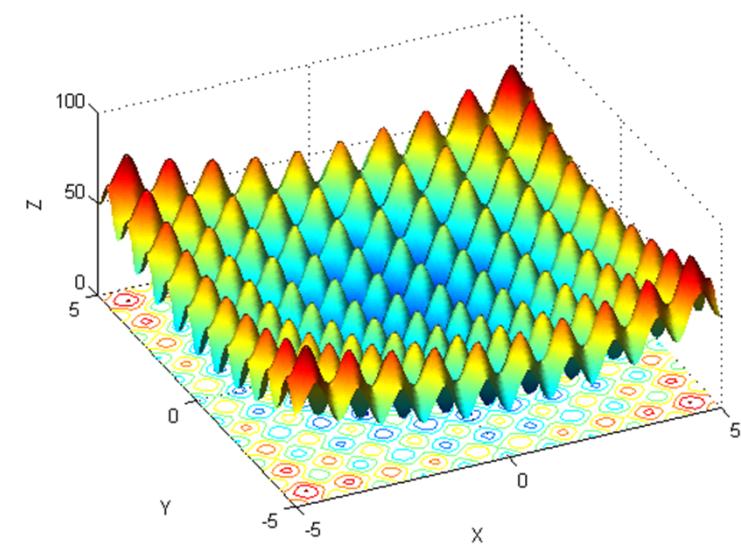
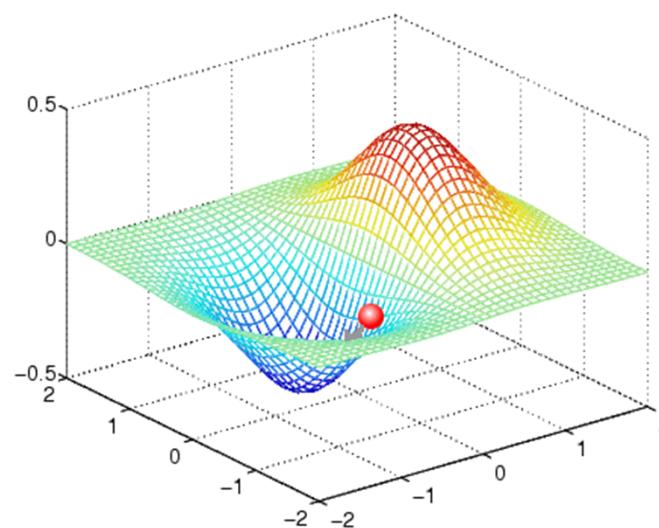
Вклад в качество итогового решения



6. Что будет дальше

Теория: анонс следующего занятия

- Допустим, мы придумали признаки, придумали как строить классификатор, осталось подобрать его параметры
- Логично настраивать параметры, минимизируя ошибку алгоритма
- О том, как же это можно делать, мы и поговорим в следующий раз



Соревнования по анализу данных

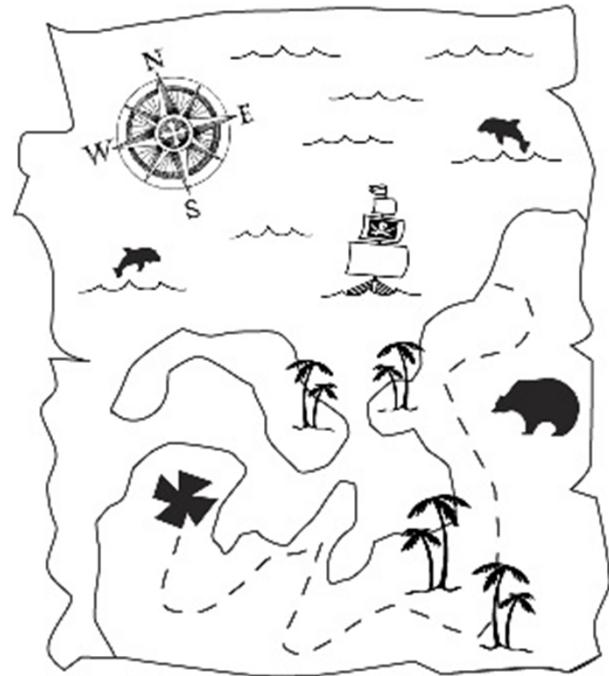
- Kaggle.com
- crowdanalytix.com
- drivendata.org

Active Competitions			
		Helping Santa's Helpers Jingle bells, Santa tells ...	42 days 50 teams \$20,000
		Click-Through Rate Prediction Predict whether a mobile ad will be clicked	2 months 404 teams \$15,000
		BCI Challenge @ NER 2015 A spell on you if you cannot detect errors!	3 months 66 teams \$1,000

Чем можно заняться уже сейчас

- Зарегистрироваться на kaggle.com и разобрать tutorial к контесту про Титаник
- Сделать на kaggle пробные посылки результатов в любой учебных контест (например, по классификации цифр)
- Ответить на вопросы:
 - Будет ли всегда заканчивать работу k-Means?
 - Бывает ли переобучение в задаче обучения без учителя?
 - Есть разные фрагменты текста, извлеченные с визитки, по каким признакам можно понять, что это фрагмент имени? А фрагмент названия компании? Придумайте как можно больше признаков и попробуйте привести примеры, когда на поставленный вопрос невозможно дать корректный ответ.
- Разобрать "[A Crash Course in Python for Scientists](#)"

План



1. Пример задачи машинного обучения
2. Стандартные задачи
3. Работа с признаками
4. Переобучение
5. Задачи на сообразительность
6. Философские вопросы
7. Что будет дальше

Спасибо за внимание!