Машинное обучение

Лекция 12

Кластеризация

Ковалев Евгений

НИУ ВШЭ, 2020

На прошлых лекциях

- Дано: матрица «объекты-признаки» X и, возможно, ответы y
- Найти: подмножество признаков или новые признаки

На прошлых лекциях

- Методы обучения с учителем: линейные модели, решающие деревья, случайные леса, …
- Дано: матрица «объекты-признаки» X и ответы y
- Найти: модель $a(x)$

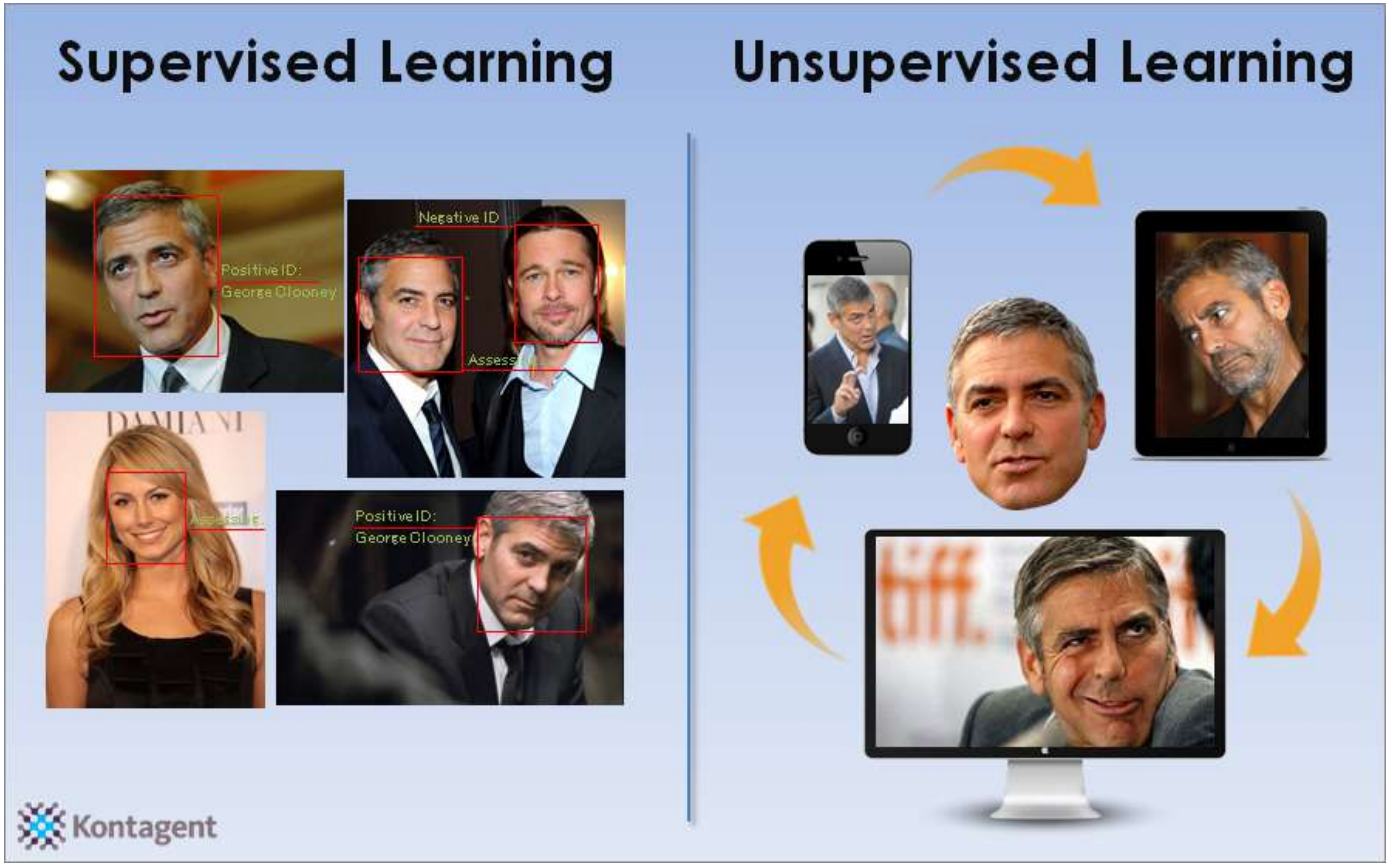
Обучение с учителем (supervised learning)

- Для каждого объекта известен ответ (класс или число)
- Даны примеры объектов с ответами
- Нужно построить модель, которая будет предсказывать ответы для новых объектов

Обучение без учителя (unsupervised learning)

- Даны объекты
- Нужно найти в них внутреннюю структуру
- Примеры:
  - Кластеризация
  - Обнаружение аномалий
  - Тематическое моделирование
  - Визуализация
  - Предсказание следующего кадра видео
  - …
- Ближе к обучению в реальной жизни

Обучение с учителем и без учителя



Обучение без учителя: предсказание кадра

Обучение без учителя: кластеризация

Case 2. Оптимизация воронки продаж

СБЕРБАНК

ШАГ I
Анализ данных, в т.ч. транзакционных
Way4, ЦОД, кред. фабрика

ШАГ II
Выявление паттернов и сегментация клиентов по характеристикам

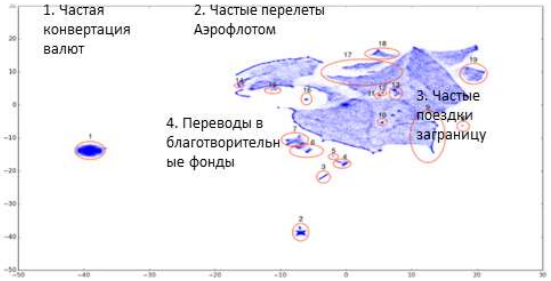
ШАГ III
Формирование продуктовых предложений на базе характеристик клиента

МЕТОДЫ алгоритмы кластеризации, визуализация данных большой размерности с использованием LargeVis

ЭКОНОМИЧЕСКИЙ ЭФФЕКТ

- ✓ Рост эффективности воронки продаж
- ✓ Рост лояльности клиентов

КЛАСТЕРИЗАЦИЯ КЛИЕНТОВ ПО ХАРАКТЕРУ ТРАНЗАКЦИЙ



В ЗАВИСИМОСТИ ОТ КЛАСТЕРА КЛИЕНТА ПРЕДЛОЖИТЬ РЕЛЕВАНТНЫЙ ПРОДУКТ

Паттерн	Продукт
1. Частая конвертация валют	Мультивалютный счет
2. Частые перелеты Аэрофлотом	Карта «Аэрофлот Бонус»
3. Частые поездки заграницу	Страховка для выезжающих за рубеж
4. Переводы в благотворительные фонды	Карта «Подари жизнь»

https://habrahabr.ru/article/318152/

Кластеризация

- Дано: матрица «объекты-признаки» X
- Найти:
  1. Множество кластеров Y
  2. Алгоритм кластеризации $a(x)$, который приписывает каждый объект к одному из кластеров

- Каждый кластер состоит из похожих объектов
- Объекты из разных кластеров существенно отличаются

Отличия

Обучение с учителем

- Цель: минимизация функционала ошибки
- Множество ответов известно заранее
- Конкретные способы измерения качества

Кластеризация

- Нет строгой постановки
- Множество кластеров неизвестно
- Правильные ответы отсутствуют (в большинстве случаев) — нельзя измерить качество

Зачем кластеризовать?

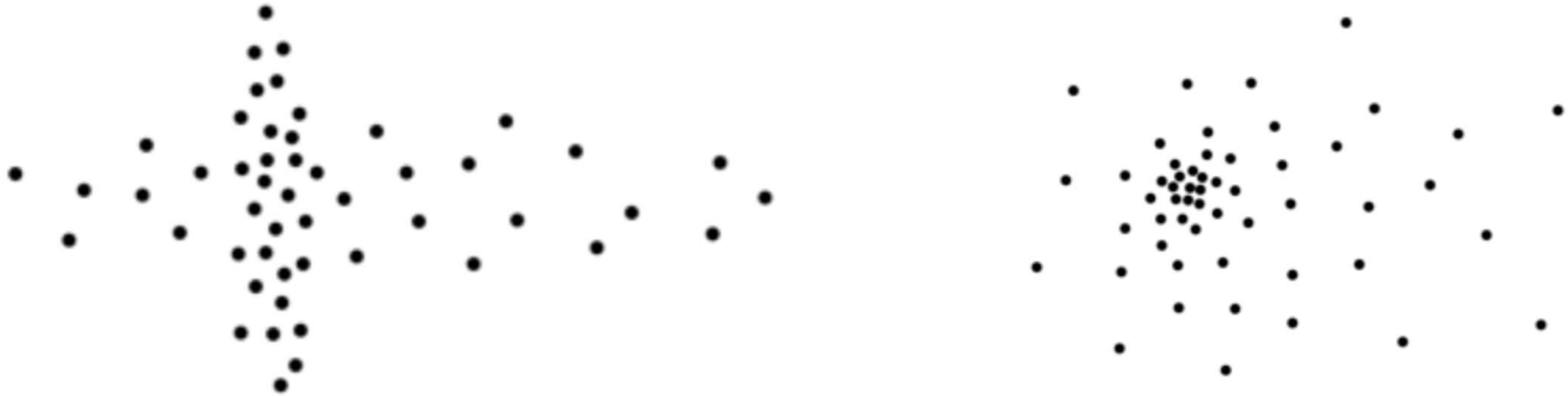
- Маркетинг: искать похожих клиентов
- Модерация: проверять только одно сообщение из кластера
- Соц. опросы: выделять группы схожих анкет
- Соц. сети: искать сообщества

- Выявлять типы людей и формировать поведенческие паттерны для каждого типа

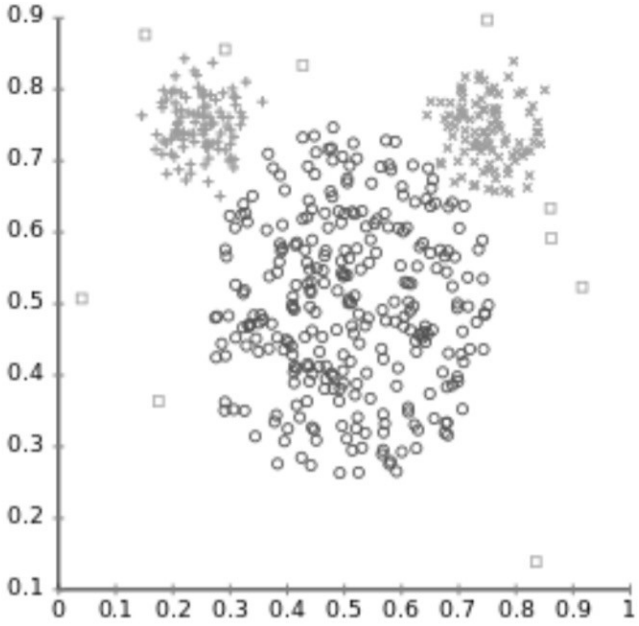
Виды кластеризации

Форма кластеров

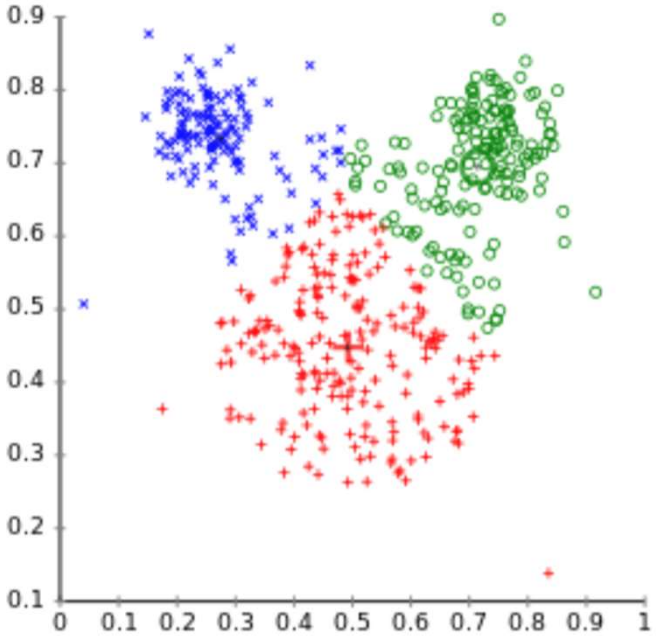
Форма кластеров



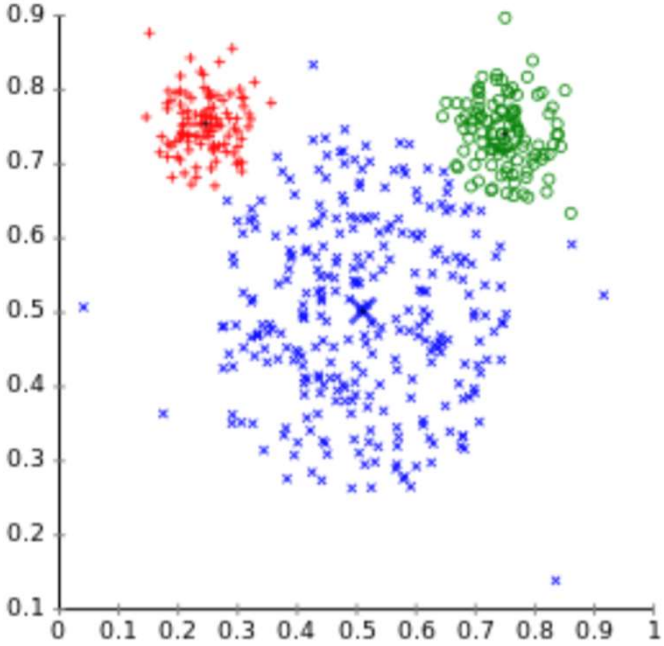
Различия в результатах работы



Исходная выборка
("Mouse" dataset)

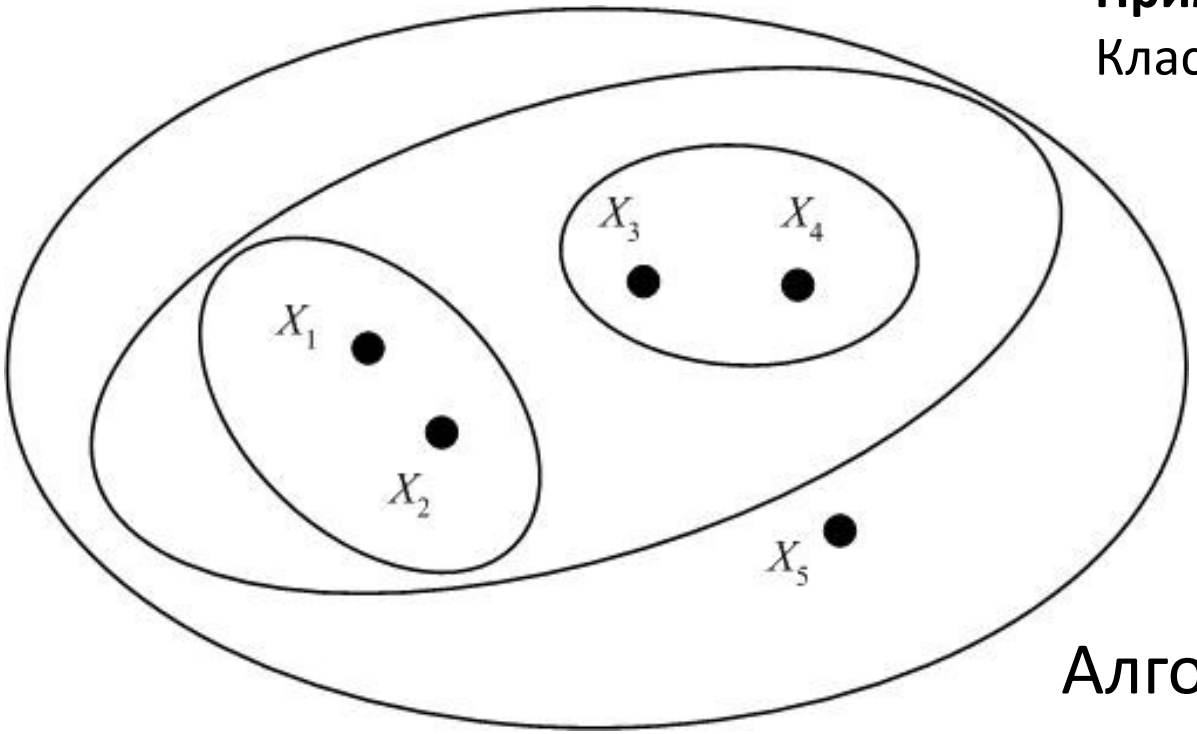


Метод 1

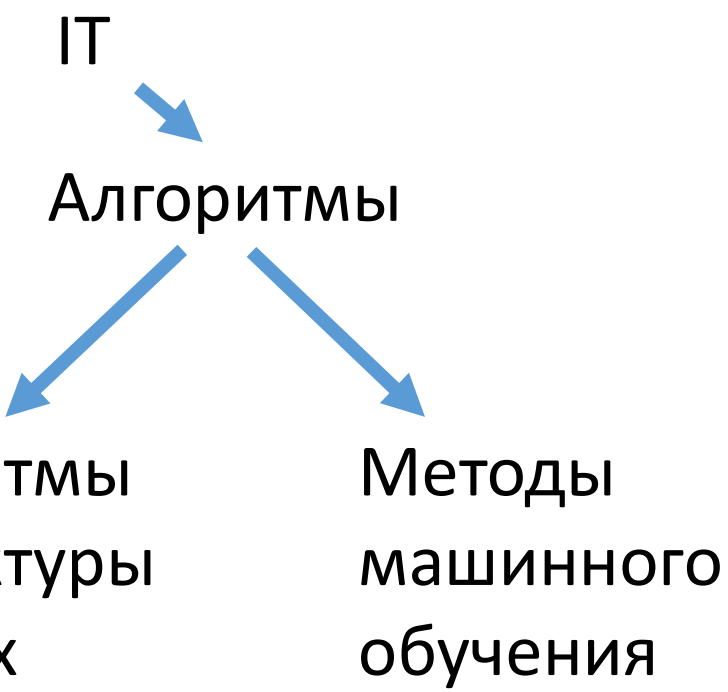


Метод 2

Иерархическая кластеризация



Пример:
Кластеризация статей на Хабре



Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 1: в один кластер должны попадать новости на одну тему

Батыршин сыграет вместо Хабарова у «Магнитки» в матче с «Салаватом»

Место в третьей паре защиты «Магнитки» на третью встречу плей-офф Кубка Гагарина с «Салаватом Юлаевым» занял защитник Рафаэль Батыршин, сообщает из Уфы корреспондент «Чемпионата» Павел Панышев. Травмированный Ярослав Хабаров выбыл на неопределённый срок. Для форварда Оскара Осалы сезон закончен.

Футболисты ЦСКА проиграли «Долгопрудному» в товарищеском матче

Футболисты московского ЦСКА со счетом 2:3 проиграли клубу второго дивизиона "Долгопрудный" в товарищеском матче, который состоялся в Москве на стадионе "Октябрь". У армейцев забитыми мячами отличились Александр Цауня (15-я минута) и Сергей Ткачев (54).

Скриншот с сайта Яндекс.Новости (news.yandex.ru)

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 2: в один кластер должны попадать новости об одном «большом» событии

Скриншот с сайта РИА Новости (ria.ru)

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 3: в один кластер должны попадать тексты об одной и той же новости

11:41, 08 ФЕВРАЛЯ 2014

Открытие Олимпиады в Сочи посмотрели несколько миллиардов человек

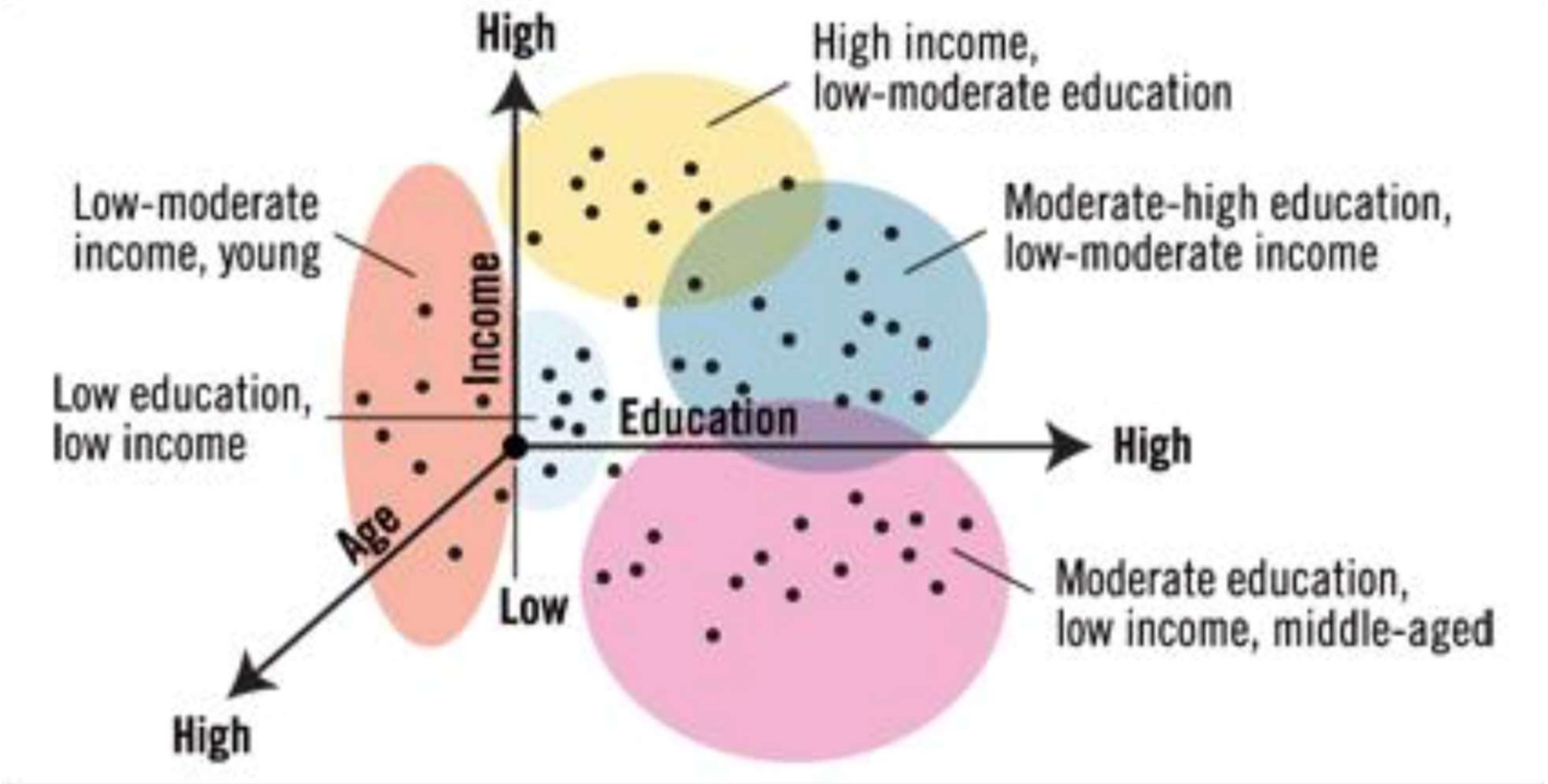
Олимпиада в Сочи открыта

Церемония открытия Олимпиады в Сочи. Онлайн-репортаж

Требования к кластерам

- Чтобы проверить, выполняются ли требования, нужно делать разметку данных
- Для новостей: показывать асессору пары документов и спрашивать, относятся ли они к одному кластеру

Кластеризация как основная задача

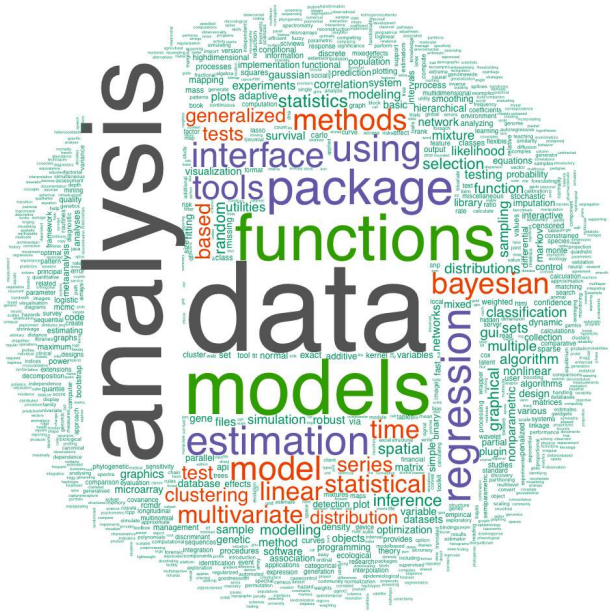


Кластеризация как вспомогательная задача

Цель: улучшение распознавания

«Жесткая» и «мягкая» кластеризации

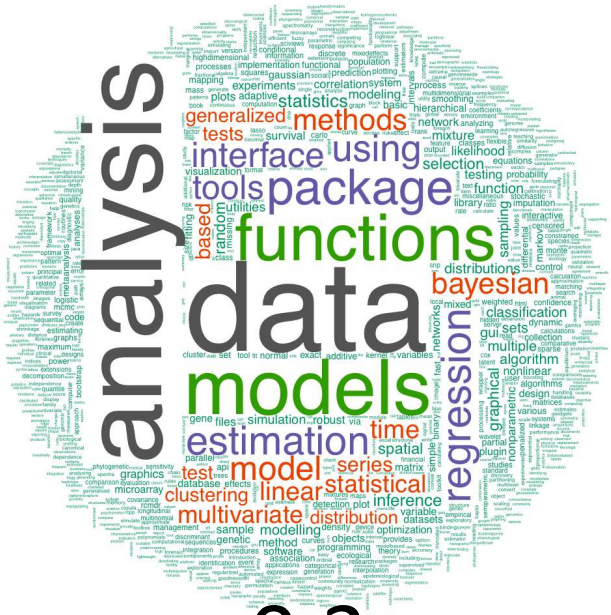
Кластеризация для выделения «тем»



«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

0.2



0.3

0.5

Типы задач кластеризации

- Форма кластеров, которые нужно выделять
- Плоская или древовидная структура
- Размер кластеров
- Конечная задача или вспомогательная
- Жесткая или мягкая кластеризация

K-Means

K-Means

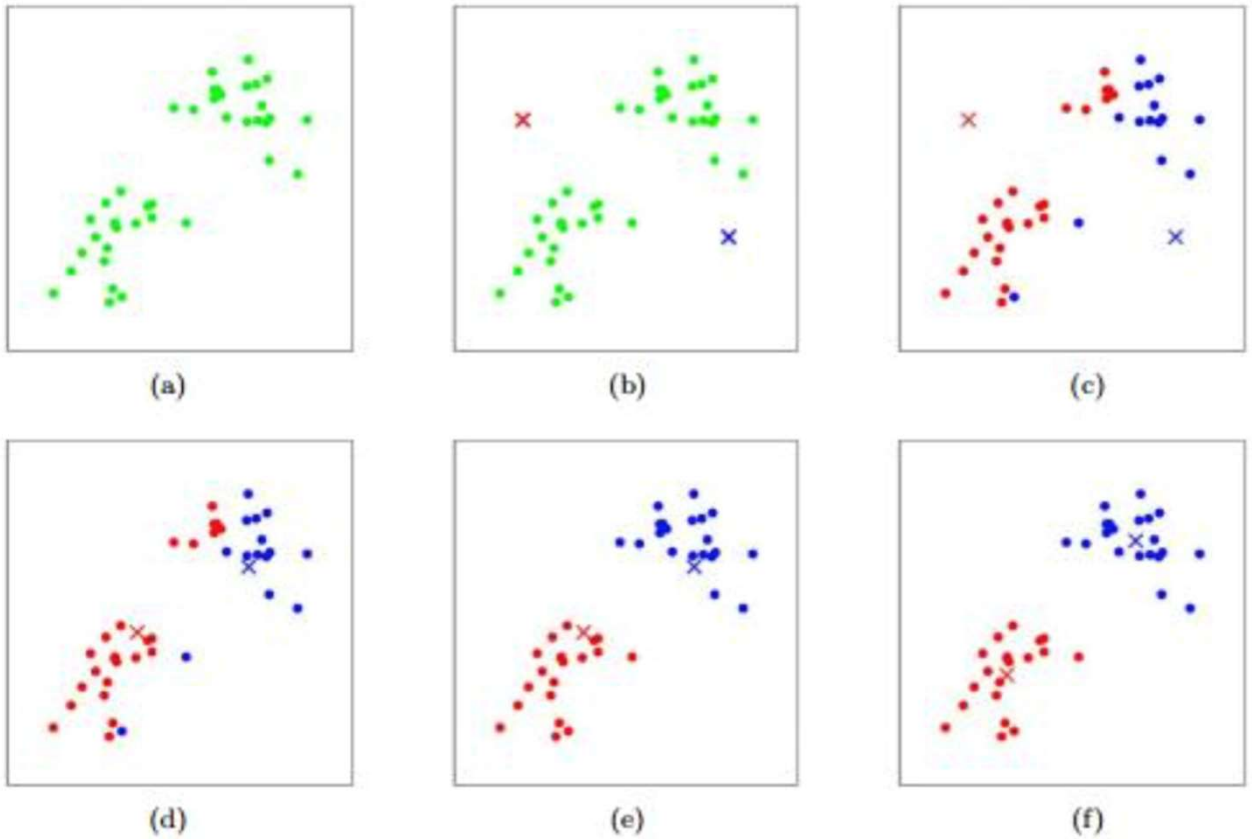
- Дано: выборка $x_1, \dots, x_\ell$
- Параметр: число кластеров K
- Начало: случайно выбрать K центров кластеров $c_1, \dots, c_K$
- Повторять по очереди до сходимости:
  - Шаг А: отнести каждый объект к ближайшему центру

$$y_i = \arg\min_{j=1,\dots,K} \rho(x_i, c_j)$$

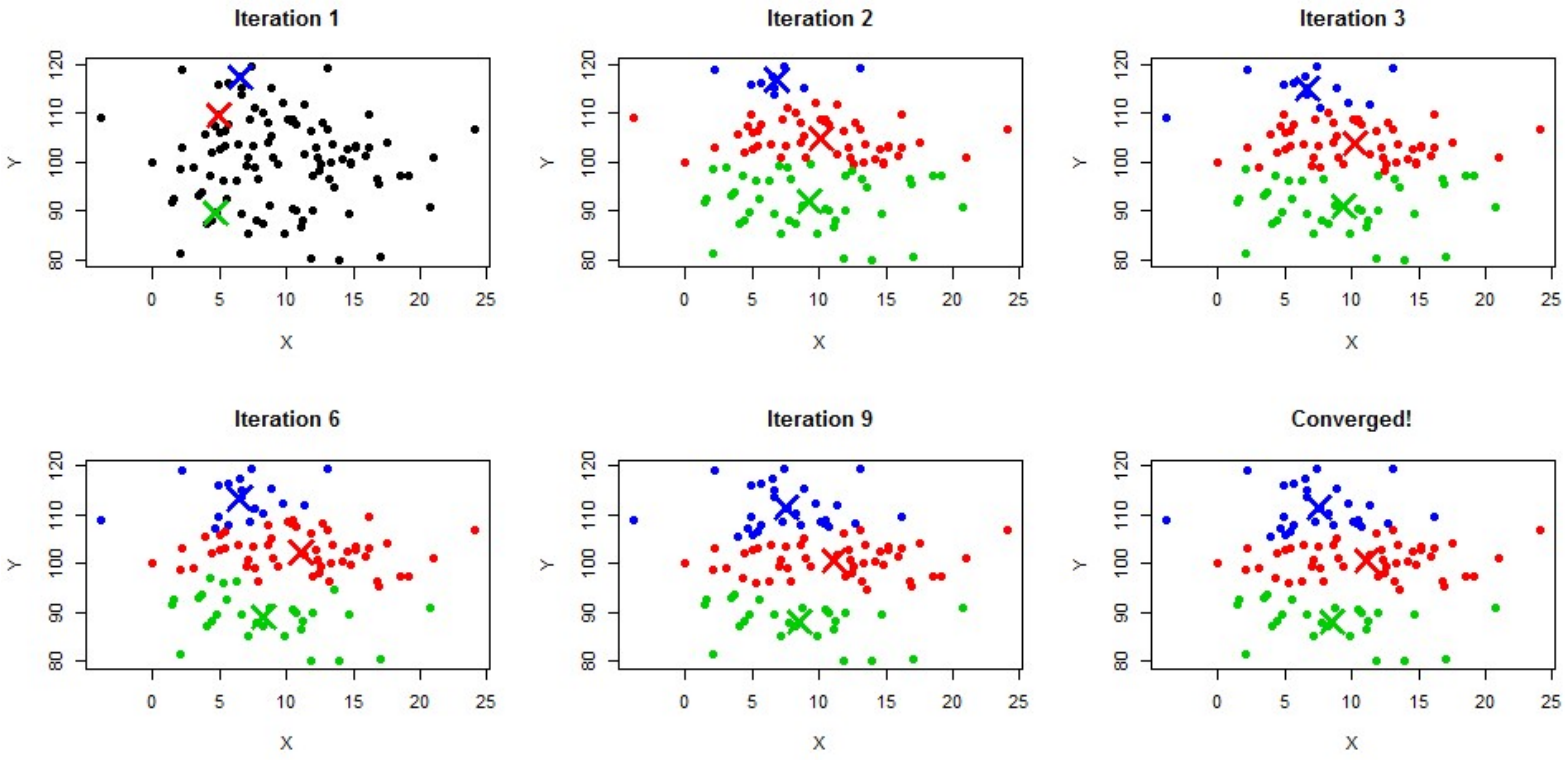
  - Шаг Б: переместить центр каждого кластера в центр тяжести

$$c_j = \frac{\sum_{i=1}^{\ell} x_i [y_i = j]}{\sum_{i=1}^{\ell} [y_i = j]}$$

K-Means



K-Means


Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
X
Y
0
5
10
15
20
25
80
90
100
110
120

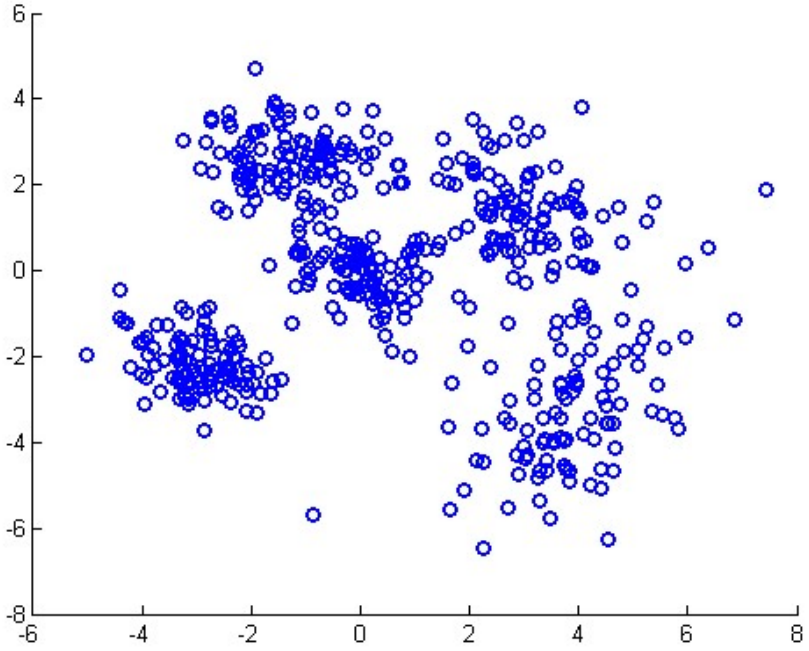
Выбор числа кластеров

- Качество кластеризации: внутрикластерное расстояние

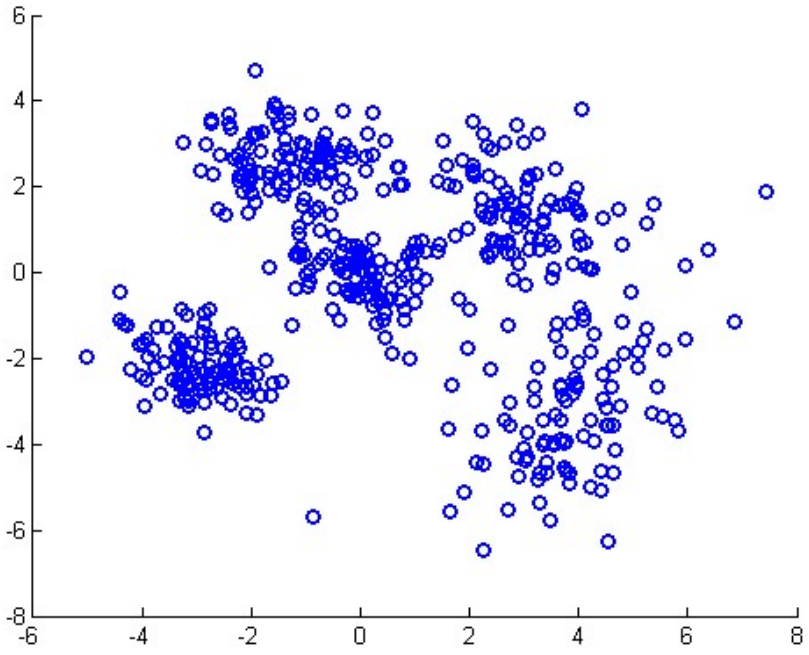
$$J(C) = \sum_{i=1}^{\ell} \rho\left(x_i, c_{y_i}\right)$$

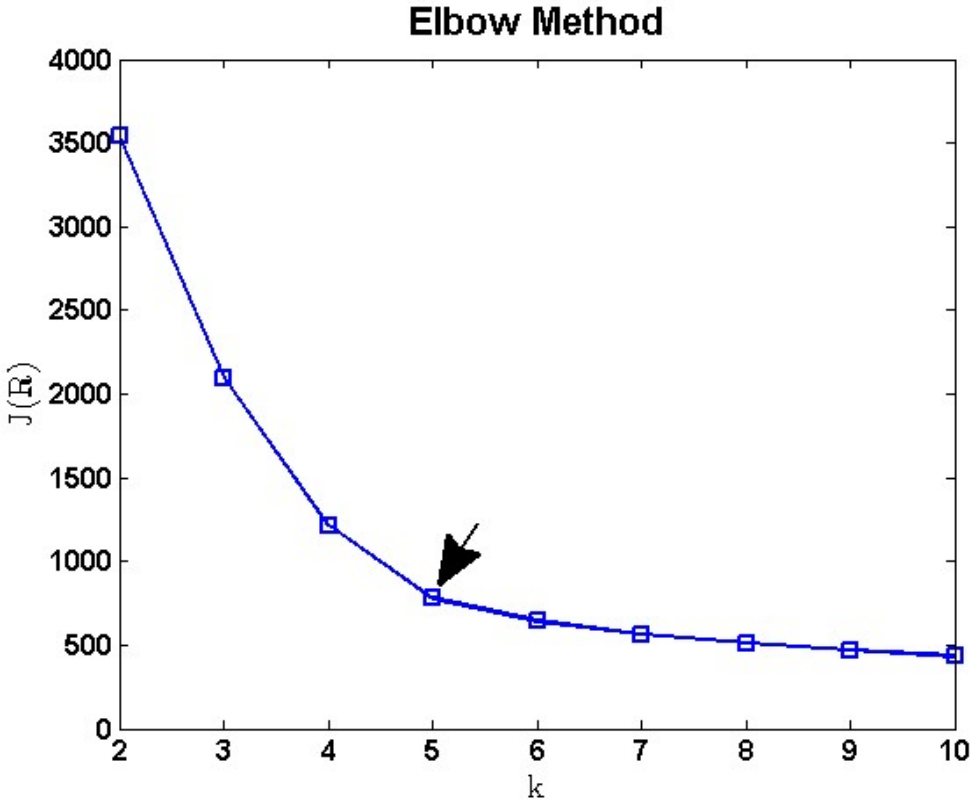
- Зависит от K
- Нужно подобрать такое K, после которого качество меняется не слишком сильно

Выбор числа кластеров



Выбор числа кластеров



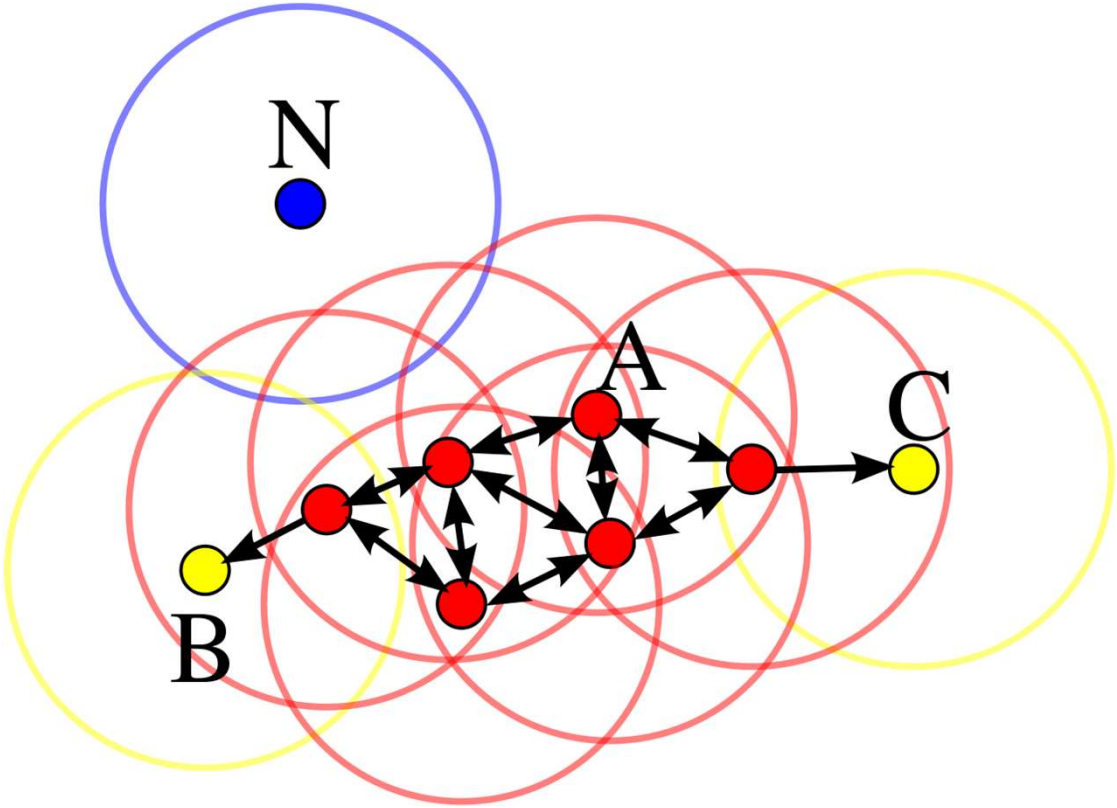


Особенности K-Means

- Может работать с большими объёмами данных
- Подходит для кластеров с простой геометрией
- Требует выбора числа кластеров

Density-based clustering

Основные, граничные и шумовые точки

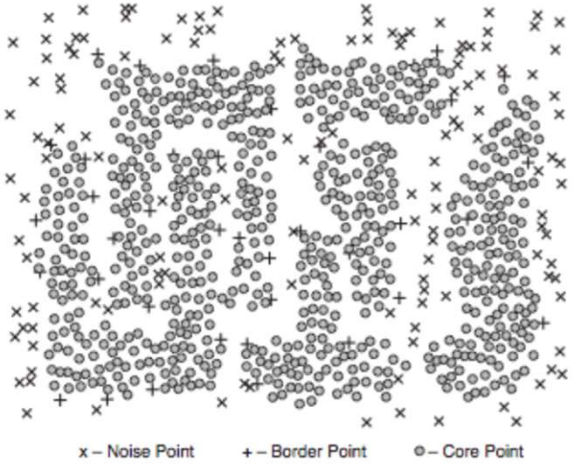


Параметры DBSCAN

- Размер окрестности (eps)
- Минимальное число объектов в окрестности — для определения основных точек

DBSCAN

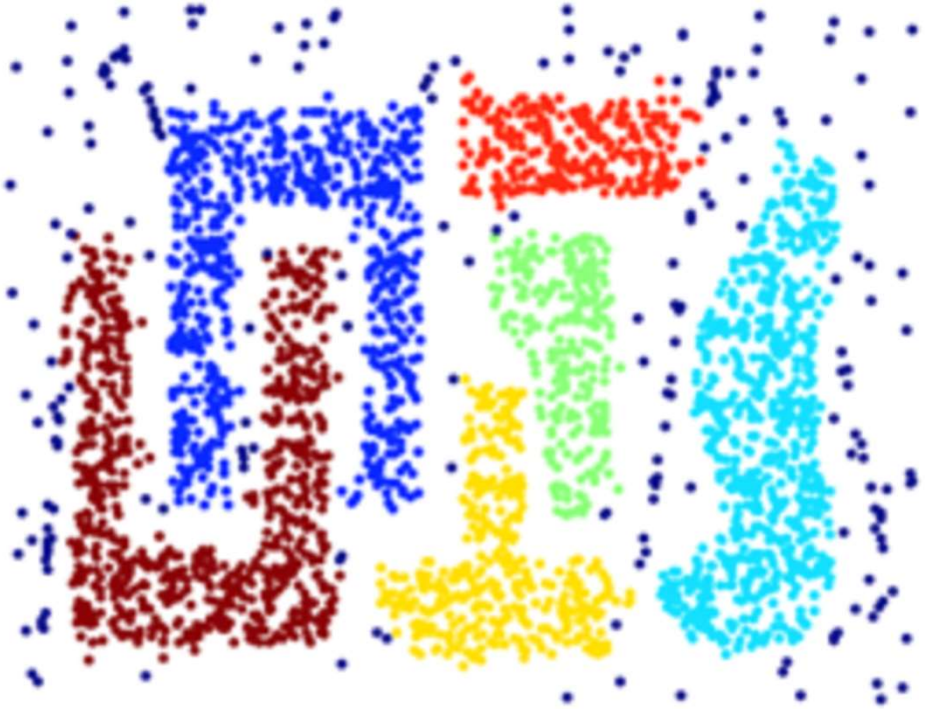
(a) Clusters found by DBSCAN.



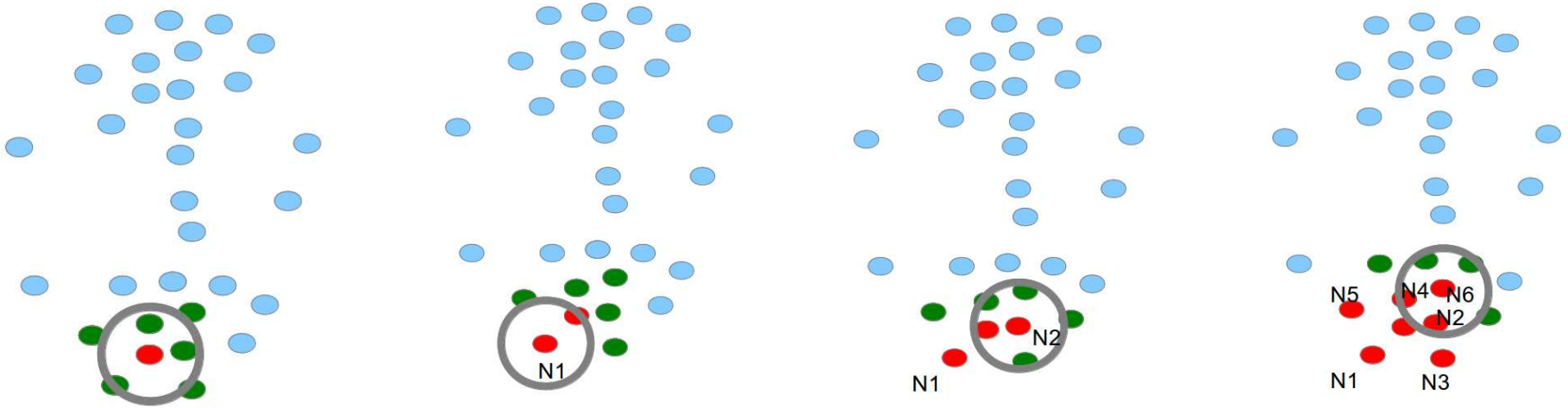
(b) Core, border, and noise points.

1. Выбрать точку без метки
2. Если в окрестности меньше N точек, то пометить как шумовую
3. Создать новый кластер, поместить в него текущую точку
4. Для всех точек из окрестности S: (а) если точка шумовая, то отнести к данному кластеру, но не использовать для расширения; (б) если точка основная, то отнести к данному кластеру, а её окрестность добавить к S
5. Перейти к шагу 1

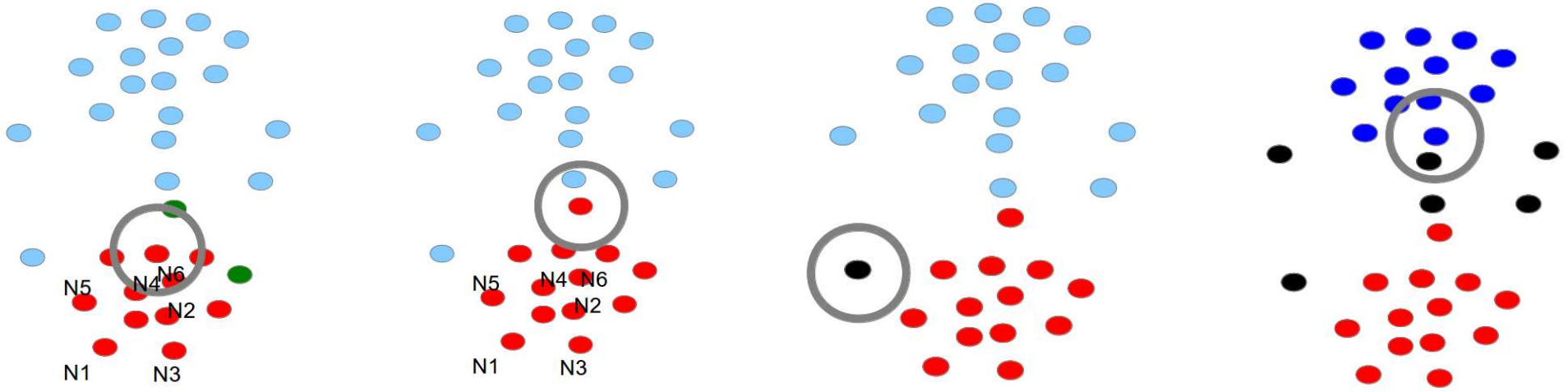
DBSCAN: результаты работы



Пример



Пример



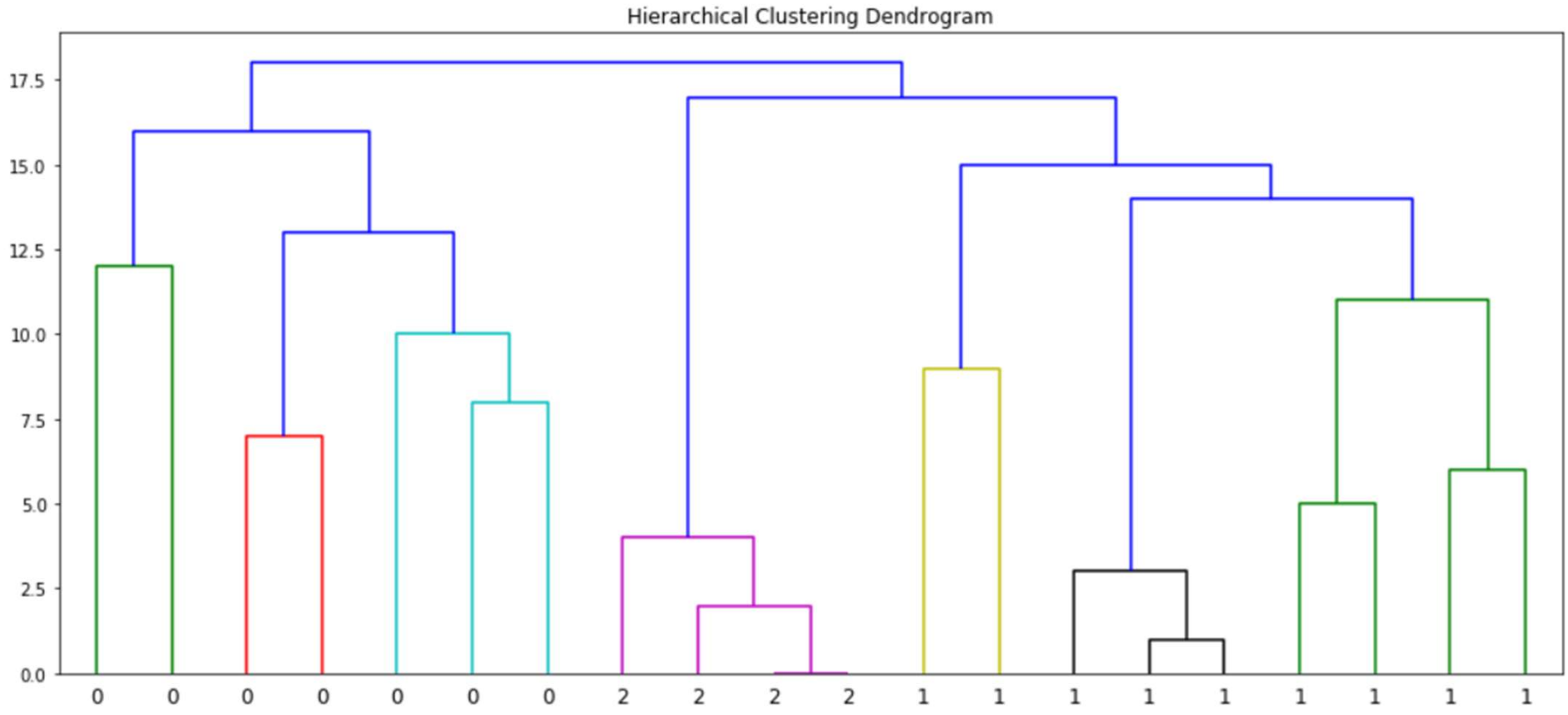
Особенности DBSCAN

- Находит кластеры произвольной формы
- Может работать с большими объёмами данных
- Нужно подбирать размер окрестности (eps) и минимальное число объектов в окрестности

Иерархическая кластеризация

Виды иерархической кластеризации

- Аггломеративная – на каждой итерации объединяем два меньших кластера в один побольше
- Дивизивная – на каждой итерации делим один большой кластер на два поменьше


Hierarchical Clustering Dendrogram
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0 0 0 0 0 0 0 2 2 2 2 1 1 1 1 1 1 1 1 1

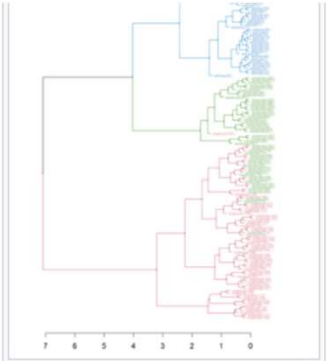
Аггломеративная кластеризация

1. Инициализация – каждая точка = кластер
2. Самые близкие (относительно какой-то метрики) кластеры объединяются
3. Повторяем до того момента, когда все точки будут в одном кластере
4. Останавливаемся, когда достигаем фиксированного числа кластеров, либо когда расстояние между кластерами больше заданного порога

Метрики расстояния

Для построения матрицы сходства (различия) необходимо задать меру расстояния между двумя кластерами. Наиболее часто используются следующие методы определения расстояния (англ. *sorting strategies*)[2]:

1. **Метод одиночной связи** (англ. *single linkage*), также известен, как «метод ближайшего соседа». Расстояние между двумя кластерами полагается равным минимальному расстоянию между двумя элементами из разных кластеров: $\min\{\, d(a,b) : a \in A,\ b \in B \,\}$, где $d(a,b)$— расстояние между элементами a и b, принадлежащими кластерам A и B
2. **Метод полной связи** (англ. *complete linkage*), также известен, как «метод дальнего соседа». Расстояние между двумя кластерами полагается равным максимальному расстоянию между двумя элементами из разных кластеров: $\max\{\, d(a,b) : a \in A,\ b \in B \,\}$;
3. **Метод средней связи** (англ. *pair-group method using arithmetic mean*):
   - Невзвешенный (англ. *UPGMA*). Расстояние между двумя кластерами полагается равным среднему расстоянию между элементами этих кластеров: $\frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$, где $d(a,b)$— расстояние между элементами a и b, принадлежащими кластерам A и B, а $|A|$ и $|B|$— мощности кластеров.
   - Взвешенный (англ. *WPGMA*).
4. **Центроидный метод** (англ. *pair-group method using the centroid average*):
   - Невзвешенный (англ. *UPGMC*). Расстояние между кластерами полагается равным расстоянию между их центроидами (центрами массы)[3]: $\|c_A - c_B\|$, где c_A и c_B— центройды A и B.
   - Взвешенный (англ. *WPGMC*).
5. **Метод Уорда** (англ. *Ward's method*). В отличие от других методов кластерного анализа, для оценки расстояний между кластерами здесь используются методы дисперсионного анализа. В качестве расстояния между кластерами берётся прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения[4]: $\Delta = \sum_i (x_i - \bar{x})^2 - \sum_{x_i \in A} (x_i - \bar{a})^2 - \sum_{x_i \in B} (x_i - \bar{b})^2$. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению дисперсии. Этот метод применяется для задач с близко расположенными кластерами.



Дендрограмма кластеризации ирисов Фишера

Для первых трёх методов существует общая формула, предложенная А. Н. Колмогоровым для мер сходства[5]:

$$K_\eta([i,j],k) = \left[\frac{(n_i K(i,k)^\eta + (n_j K(j,k)^\eta)}{n_i + n_j}\right]^{\frac{1}{\eta}}, -1 \leqslant \eta \leqslant +1$$

где $[i,j]$ — группа из двух объектов (кластеров) i и j; k — объект (кластер), с которым ищется сходство указанной группы; n_i — число элементов в кластере i; n_j — число элементов в кластере j. Для расстояний имеется аналогичная формула Ланса — Вильямса[6].

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

Обучение без учителя и текстовые данные

Похожие слова

- «Идти» и «шагать» — синонимы
- Для компьютера это разные строки
- Как понять, что они похожи?

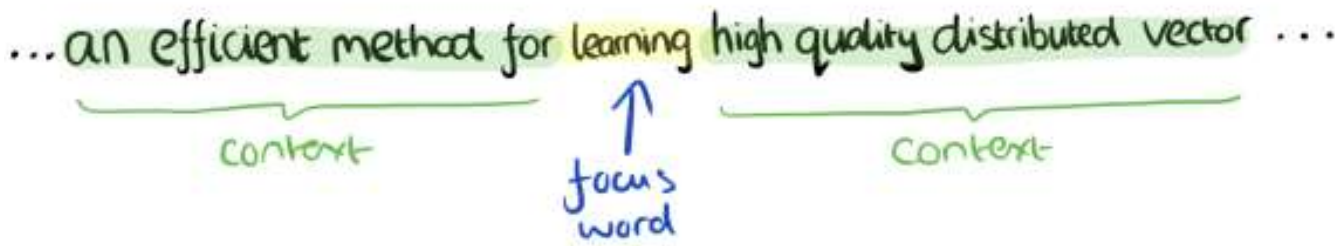
Похожие слова

- «Идти» и «шагать» — синонимы
- Для компьютера это разные строки
- Как понять, что они похожи?

- На основе данных!
- Слова со схожим смыслом часто идут в паре с одними и теми же словами
- У них похожие контексты

Дистрибутивная семантика

- У похожих по смыслу слов похожие *контексты*
- Контекст — окрестность слова



Векторные представления слов

Хотим представить каждое слово в виде вещественного вектора:

$$w \to \vec{w} \in \mathbb{R}^d$$

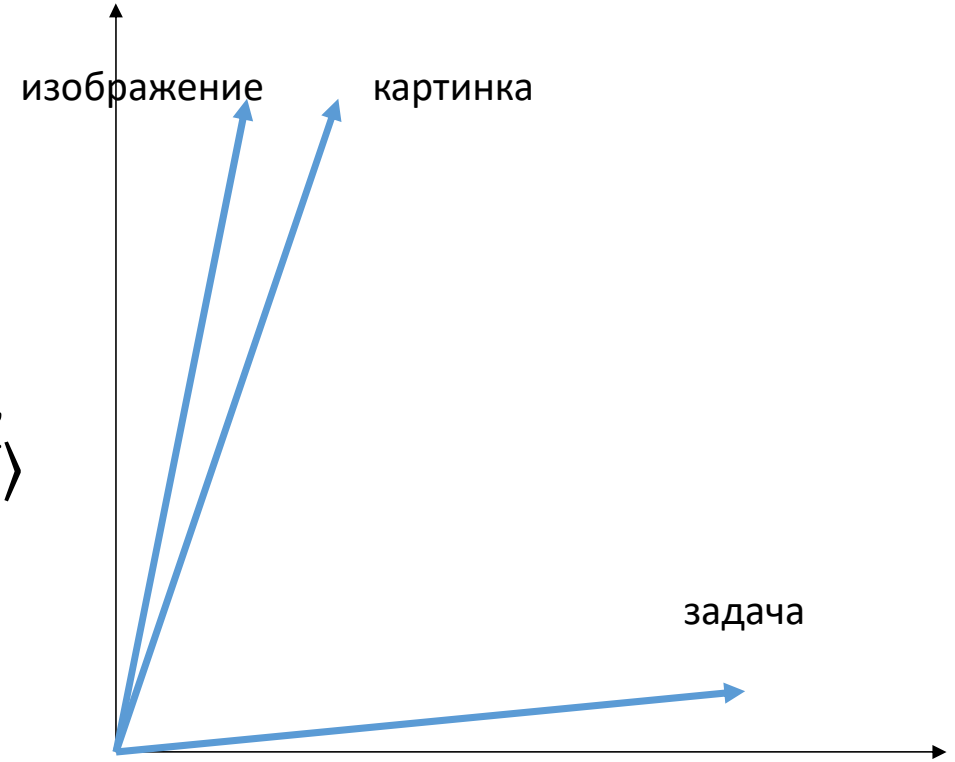
Требования к представлениям (embeddings):

- Размерность d должна быть не очень большой
- Похожие слова должны иметь близкие векторы
- Арифметические операции над векторами должны иметь смысл

word2vec

Задача:

- Для каждого слова w построить вектор $\vec{w}$
- Если два слова w_1 и w_2 идут рядом, то скалярное произведение $\langle \overrightarrow{w_1}, \overrightarrow{w_2} \rangle$ должно быть большим



word2vec

Если два слова w_1 и w_2 идут рядом, то скалярное произведение $\langle \overrightarrow{w_1}, \overrightarrow{w_2} \rangle$ должно быть большим:

$$p(w_i \mid w_j) = \frac{\exp(\langle \vec{w}_i, \vec{w}_j \rangle)}{\sum_{w \in W} \exp(\langle \vec{w}, \vec{w}_j \rangle)}$$

$$\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \sum_{\substack{k=-K \\ k \neq 0}}^{K} \log p(\vec{w}_{j+k} \mid \vec{w}_j) \to \max_{\{\vec{w}\}_{w \in W}}$$

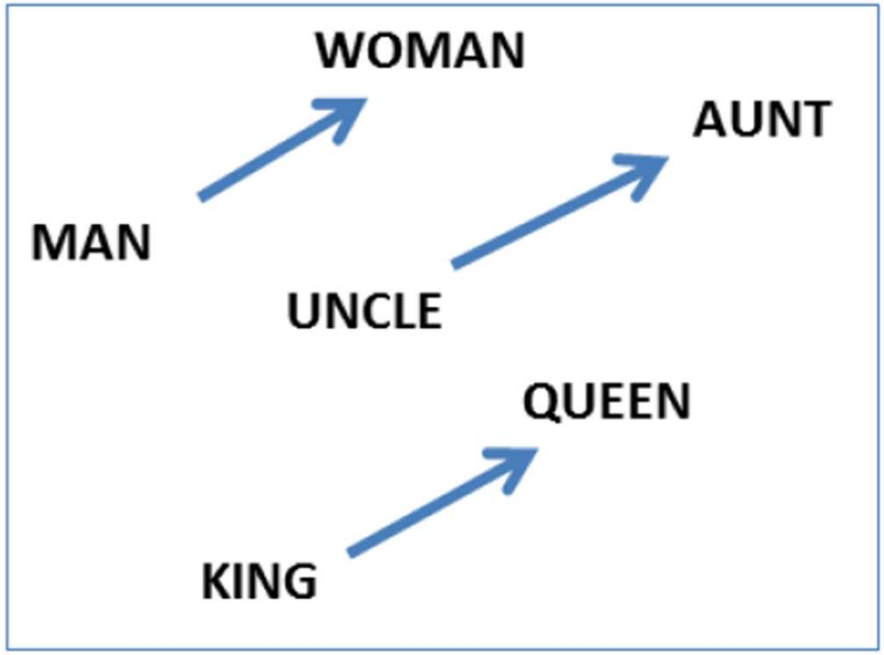
word2vec

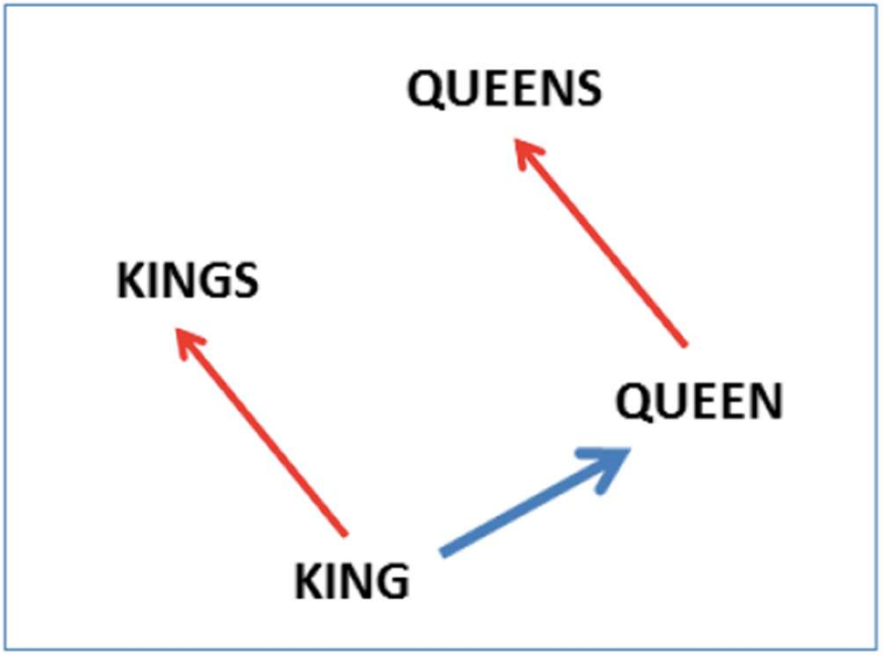
Векторы можно прибавлять и вычитать:

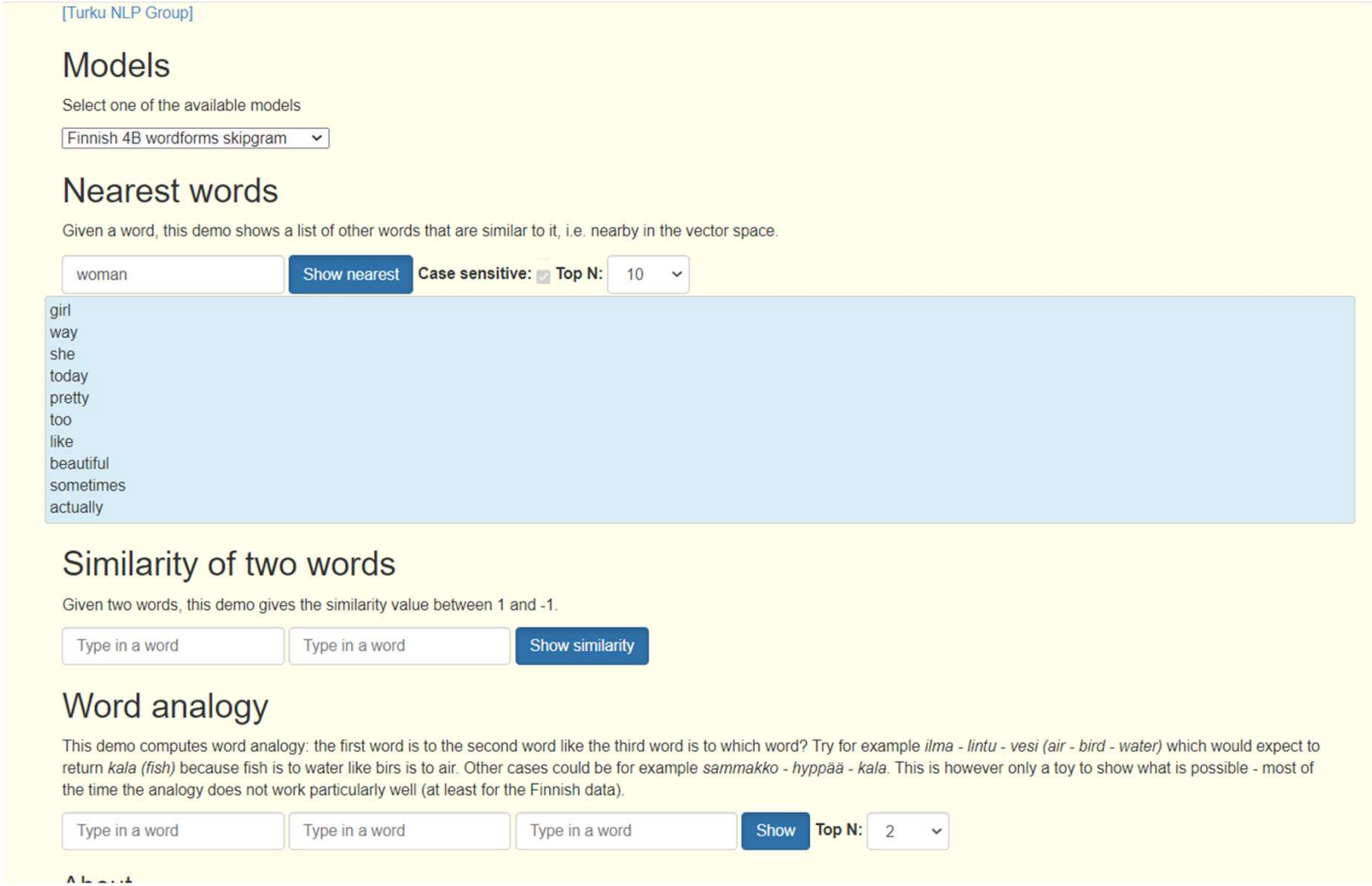
- $\overrightarrow{\text{король}} - \overrightarrow{\text{мужчина}} + \overrightarrow{\text{женщина}} \approx \overrightarrow{\text{королева}}$
- $\overrightarrow{\text{медведь}} - \overrightarrow{\text{Россия}} + \overrightarrow{\text{Австралия}} \approx \overrightarrow{\text{кенгуру}}$

Можно переводить слова:

- $\overrightarrow{\text{математика}} + \left(\overrightarrow{\text{word}} - \overrightarrow{\text{слово}}\right) \approx \overrightarrow{\text{math}}$
- $\overrightarrow{\text{король}} + \left(\overrightarrow{\text{word}} - \overrightarrow{\text{слово}}\right) \approx \overrightarrow{\text{king}}$
- $\overrightarrow{\text{корова}} + \left(\overrightarrow{\text{word}} - \overrightarrow{\text{слово}}\right) \approx \overrightarrow{\text{cow}}$


WOMAN
AUNT
MAN
UNCLE
QUEEN
KING


QUEENS
KINGS
QUEEN
KING

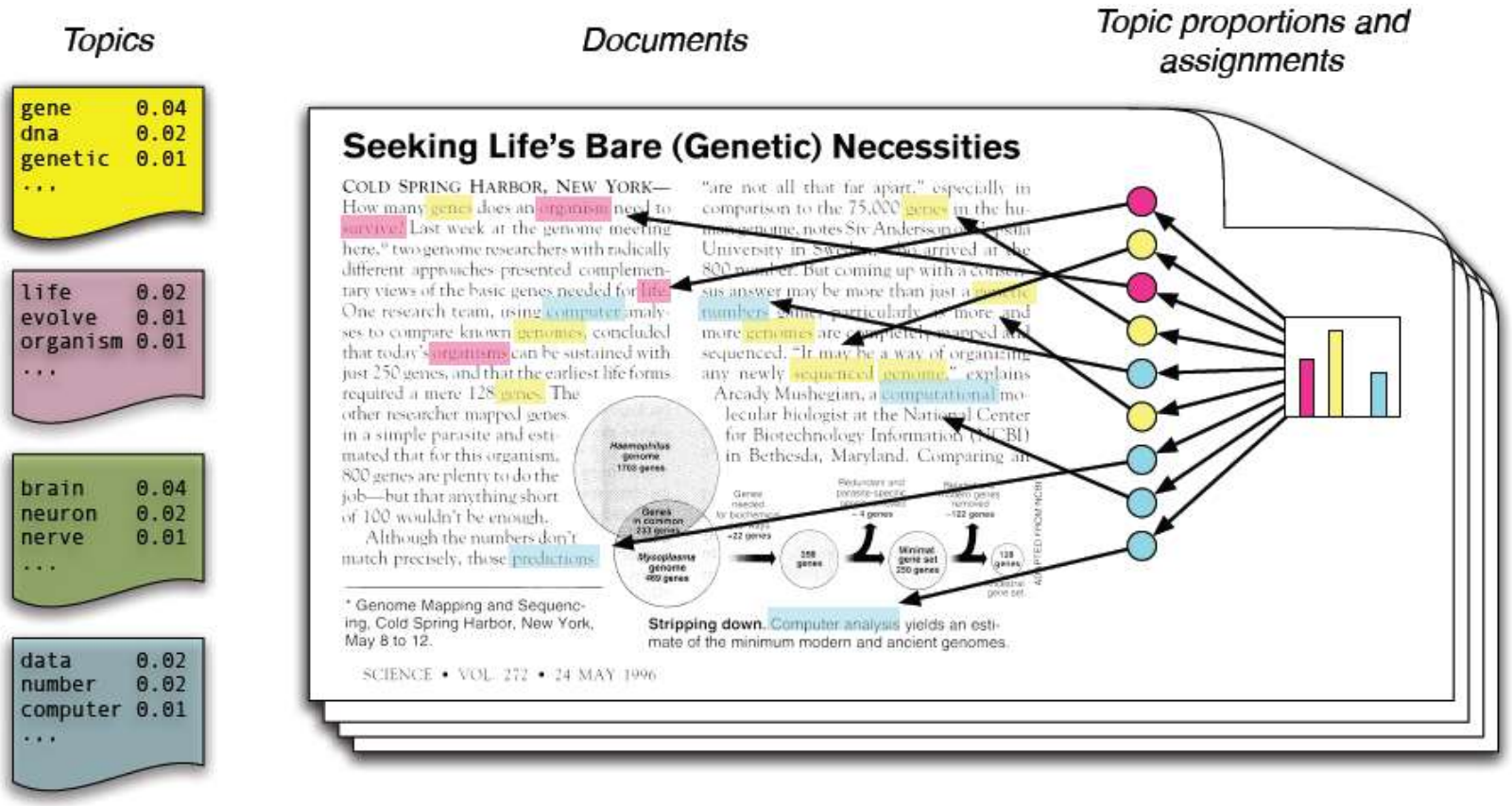


http://bionlp-www.utu.fi/wv_demo/

Тематическое моделирование

- Рассматриваем каждый документ как мешок слов
- Всего K тем
- Тема — распределение на словах
- Документ — распределение на темах

Тематическое моделирование



Модель PLSA

- Probabilistic Latent Semantic Analysis

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}$$

- T — множество тем
- $p(w|t) = \varphi_{wt}$ — распределение слов в теме t
- $p(t|d) = \theta_{td}$ — распределение тем в документе d

Модель PLSA

- Probabilistic Latent Semantic Analysis

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d) \to \max_{\varphi_{wt}, \theta_{td}}$$

Ограничения: $\varphi_{wt} \geq 0, \theta_{td} \geq 0, \sum_{w \in W} \varphi_{wt} = 1, \sum_{t \in T} \theta_{td} = 1$

- D — множество документов
- W — множество слов

Пример

- Данные: новостные заголовки

Topic: 0 Word: 0.008*"octob" + 0.006*"search" + 0.006*"miss" + 0.006*"inquest" + 0.005*"stori" + 0.005*"jam" + 0.004*"john" + 0.004*"harvest" + 0.004*"australia" + 0.004*"world"
Topic: 1 Word: 0.006*"action" + 0.006*"violenc" + 0.006*"thursday" + 0.005*"domest" + 0.005*"cancer" + 0.005*"legal" + 0.005*"union" + 0.005*"breakfast" + 0.005*"school" + 0.004*"student"
Topic: 2 Word: 0.023*"rural" + 0.018*"govern" + 0.013*"news" + 0.012*"podcast" + 0.008*"grandstand" + 0.008*"health" + 0.007*"budget" + 0.007*"busi" + 0.007*"nation" + 0.007*"fund"
Topic: 3 Word: 0.030*"countri" + 0.028*"hour" + 0.009*"sport" + 0.008*"septemb" + 0.008*"wednesday" + 0.007*"commiss" + 0.006*"royal" + 0.006*"updat" + 0.006*"station" + 0.005*"bendigo"
Topic: 4 Word: 0.014*"south" + 0.009*"weather" + 0.009*"north" + 0.008*"west" + 0.008*"coast" + 0.008*"australia" + 0.006*"east" + 0.006*"queensland" + 0.006*"storm" + 0.005*"season"
Topic: 5 Word: 0.008*"monday" + 0.008*"august" + 0.006*"babi" + 0.005*"shorten" + 0.005*"hobart" + 0.004*"victorian" + 0.004*"donald" + 0.004*"safe" + 0.004*"scott" + 0.004*"donat"
Topic: 6 Word: 0.022*"interview" + 0.013*"market" + 0.009*"share" + 0.008*"cattl" + 0.008*"trump" + 0.008*"turnbul" + 0.007*"novemb" + 0.007*"michael" + 0.006*"australian" + 0.006*"export"
Topic: 7 Word: 0.019*"crash" + 0.014*"kill" + 0.009*"fatal" + 0.009*"dead" + 0.007*"die" + 0.007*"truck" + 0.007*"polic" + 0.006*"attack" + 0.006*"injur" + 0.006*"bomb"
Topic: 8 Word: 0.008*"drum" + 0.007*"abbott" + 0.007*"farm" + 0.006*"dairi" + 0.006*"asylum" + 0.006*"tuesday" + 0.006*"water" + 0.006*"labor" + 0.006*"say" + 0.005*"plan"
Topic: 9 Word: 0.017*"charg" + 0.014*"murder" + 0.011*"court" + 0.011*"polic" + 0.009*"woman" + 0.008*"assault" + 0.008*"jail" + 0.008*"alleg" + 0.007*"accus" + 0.007*"guilti"

Резюме

- Кластеризация — задача без строгой постановки и без строгих критериев качества
- Много разновидностей в подходах
- Методы: K-Means, DBSCAN, иерархическая кластеризация и т.д.
- Обучение без учителя — гораздо более широкая область