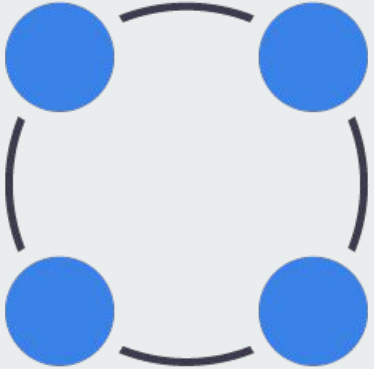
Кластеризация

Лекция 3

(19.02.2022)



Примерный план лекции

- Многомерные данные
- Общие понятия о кластеризации
- Кластеризация:
  - K-means
  - Affinity Propagation
  - Иерархическая кластеризация (агломеративная и дивизивная)
- Метрики качества кластеризации
- tSNE

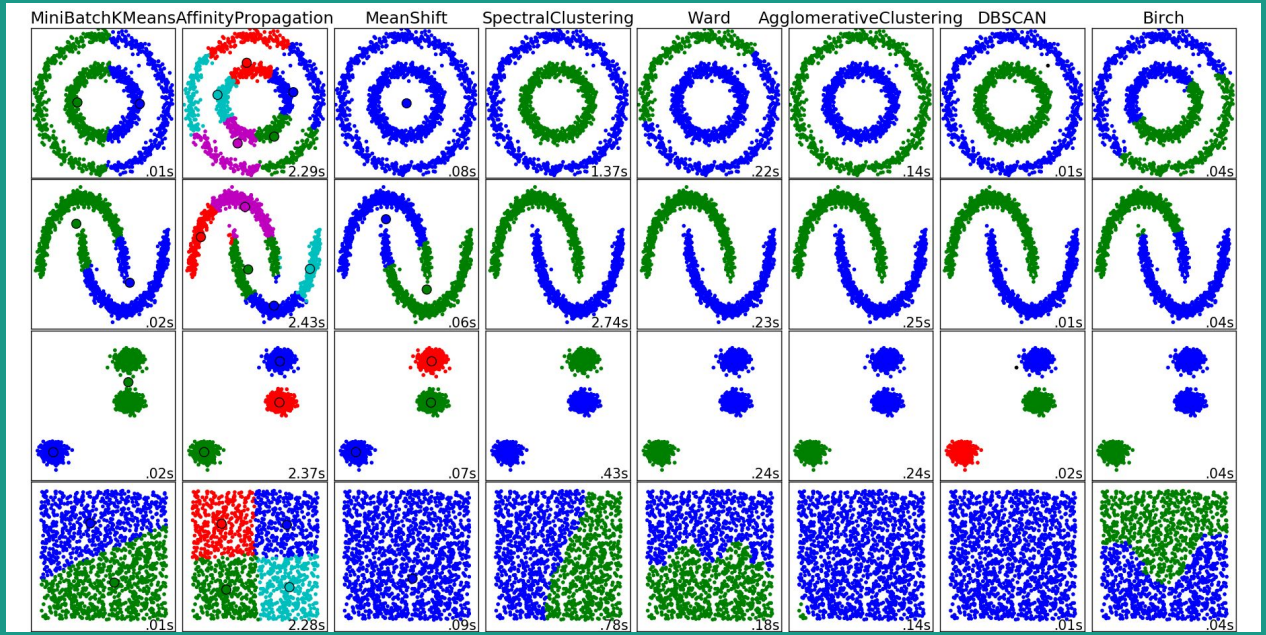
Объекты могут описываться множеством признаков

- Сообщества (признаки - виды)
- Форма тела (признаки - размеры тех или иных частей)
- Социальная активность животного (признаки - проявление того или иного паттерна)
- Транскриптом или протеом (признаки - транскрипты или белки)

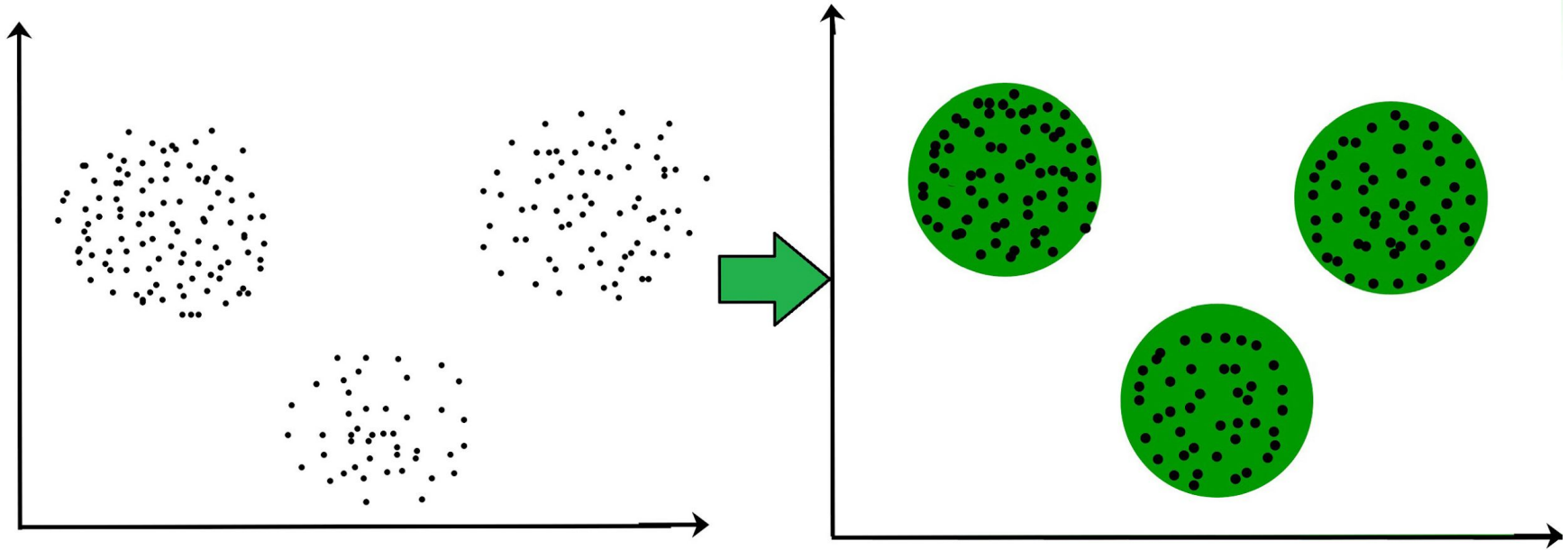
Объекты могут описываться множеством признаков

Элементы анализа	(син)экология	морфология	молекулярная биология
Объекты	площадки, пробы	особи, клетки и т. п.	особи, клетки и т. п.
Собственные свойства объектов	численность/биомасса особей разных видов	измерения, свойства	интенсивность экспрессии гена/пептида
Внешние факторы	свойства среды	свойства среды, особей или клеток	свойства среды, особей или клеток

Кластеризация



Основная идея кластеризации



Основная идея кластеризации

Если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

Гипотеза компактности

Исходя из этого -- близким объектам соответствуют близкие ответы.

Гипотеза непрерывности

Основная идея кластеризации

Если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

Гипотеза компактности

Исходя из этого -- близким объектам соответствуют близкие ответы.

Гипотеза непрерывности

Близкие точки похожи!

Меры сходства и различия

- Сходство (S) достигает максимума, когда объекты обладают идентичными признаками, различия (D), наоборот - достигает минимума.
- Обычно (но не всегда) коэффициенты сходства распределены от 0 до 1.
- Для всех метрик будут верны следующие свойства:
  a. Если $a=b$, то $D(a,b)=0$
  b. Симметричность $D(a,b)=D(b,a)$
  c. Справедливо неравенство треугольника $D(a,b)+D(b,c) \geq D(a,c)$

Значимы ли нули

Меры различия *не учитывающие* двойные нули. Эти коэффициенты не изменяются если в данные будут добавлены двойные нули (например, при увеличении количества описанных объектов).

Примеры:

- Евклидово расстояние
- расстояние по Манхеттену

Меры различия *учитывающие* двойные нули. Эти коэффициенты изменяются при появлении двойных нулей. Сходство возрастает за счет того, что отсутствие признака считается тоже сходством.

Пример:

- Коэффициенты корреляции (непараметрические)

Меры расстояния

Нестандартизованные

Евклидово расстояние

$$D = \sqrt{\sum (x_{i,j} - x_{i,k})^2}$$

Расстояние по манхеттену

$$D = \sum |x_{i,j} - x_{i,k}|$$

Стандартизованные

Расстояние по Канберре

$$D = \frac{1}{p} \sum \frac{|x_{i,j} - x_{i,k}|}{x_{i,j} + x_{i,k}}$$

Расстояние Хи-квадрат

$$\chi^2 = \sqrt{\sum \frac{1}{c_i} (x_{i,j} - x_{i,k})^2}$$

Евклидово расстояние, вычисленное по относительным величинам.

Методы кластеризации

1. Методы, основанные на плотности (*DBSCAN (Density-Based Spatial Clustering of Applications with Noise) , OPTICS (Ordering Points to Identify Clustering Structure)*)
2. Иерархические методы:
   a. Агломеративные (снизу вверх)
   b. Дивизные (сверху вниз)
3. Методы разбиения (k-means)
4. Нейросети

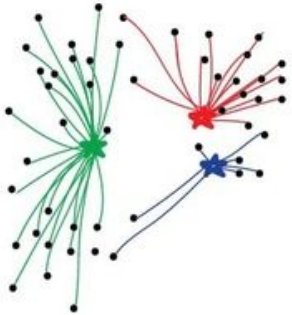
K-means

1. Выбрать количество кластеров ***k***, которое нам кажется оптимальным для наших данных.
2. Высыпать случайным образом в пространство наших данных ***k*** точек (центроидов).
3. Для каждой точки нашего набора данных посчитать, к какому центроиду она ближе.
4. Переместить каждый центроид в центр выборки, которую мы отнесли к этому центроиду.
5. Повторять последние два шага фиксированное число раз, либо до тех пор пока центроиды не "сойдутся" (обычно это значит, что их смещение относительно предыдущего положения не превышает какого-то заранее заданного небольшого значения).

Ставим три ларька с шаурмой оптимальным образом

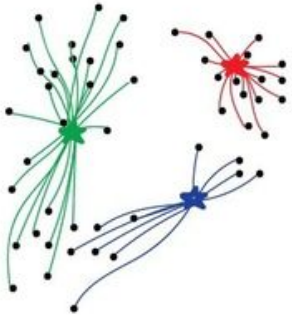
(иллюстрируя метод K-средних)

1. Ставим ларьки с шаурмой в случайных местах

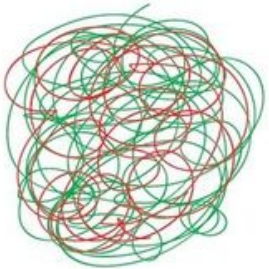


2. Смотрим в какой кому ближе идти

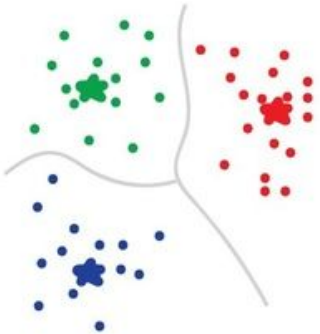
3. Двигаем ларьки ближе к центрам их популярности



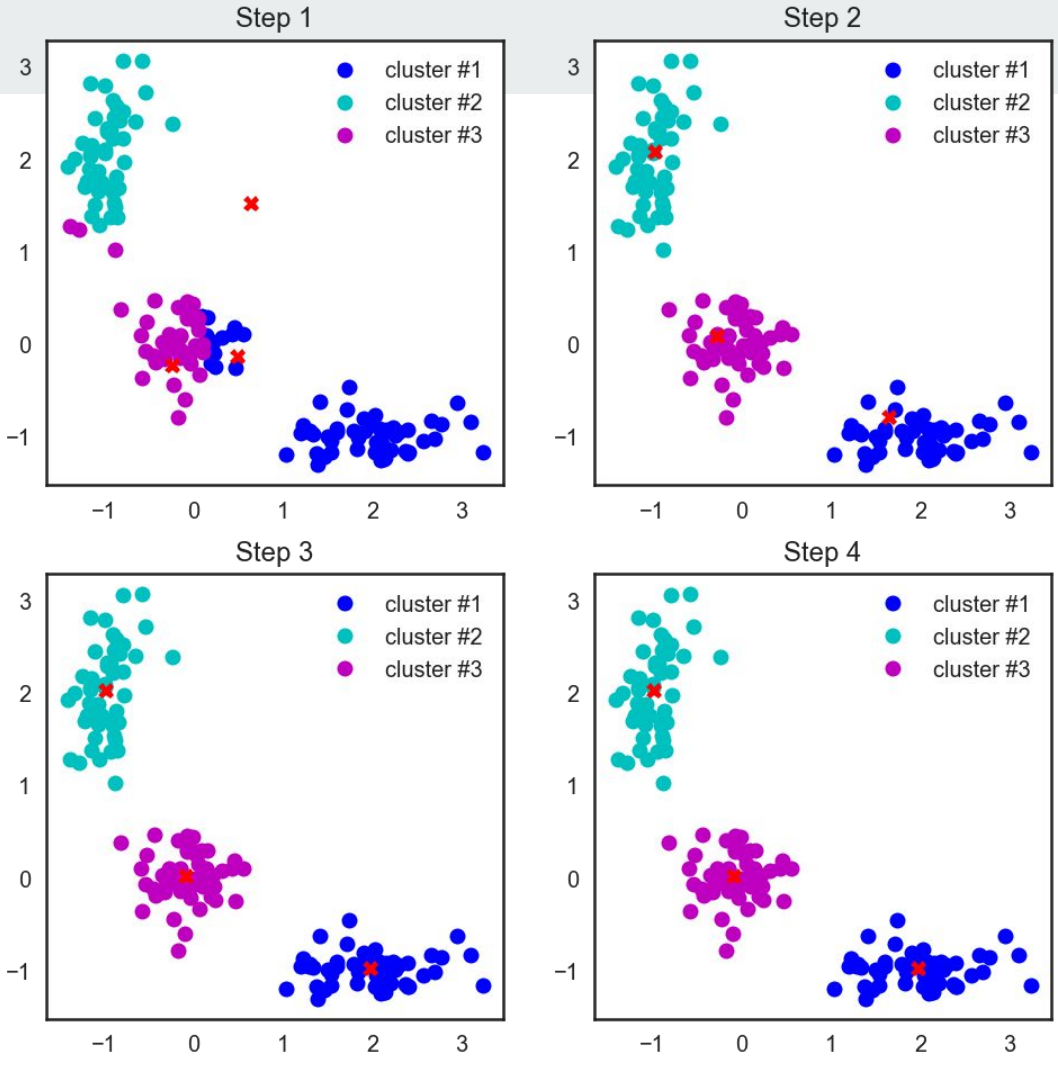
4. Снова смотрим и двигаем



5. Повторяем много раз



6. Готово, вы великолепны!


Step 1
Step 2
Step 3
Step 4
cluster #1
cluster #2
cluster #3
3
2
1
0
−1
−1
0
1
2
3

Выбор числа кластеров для k-means

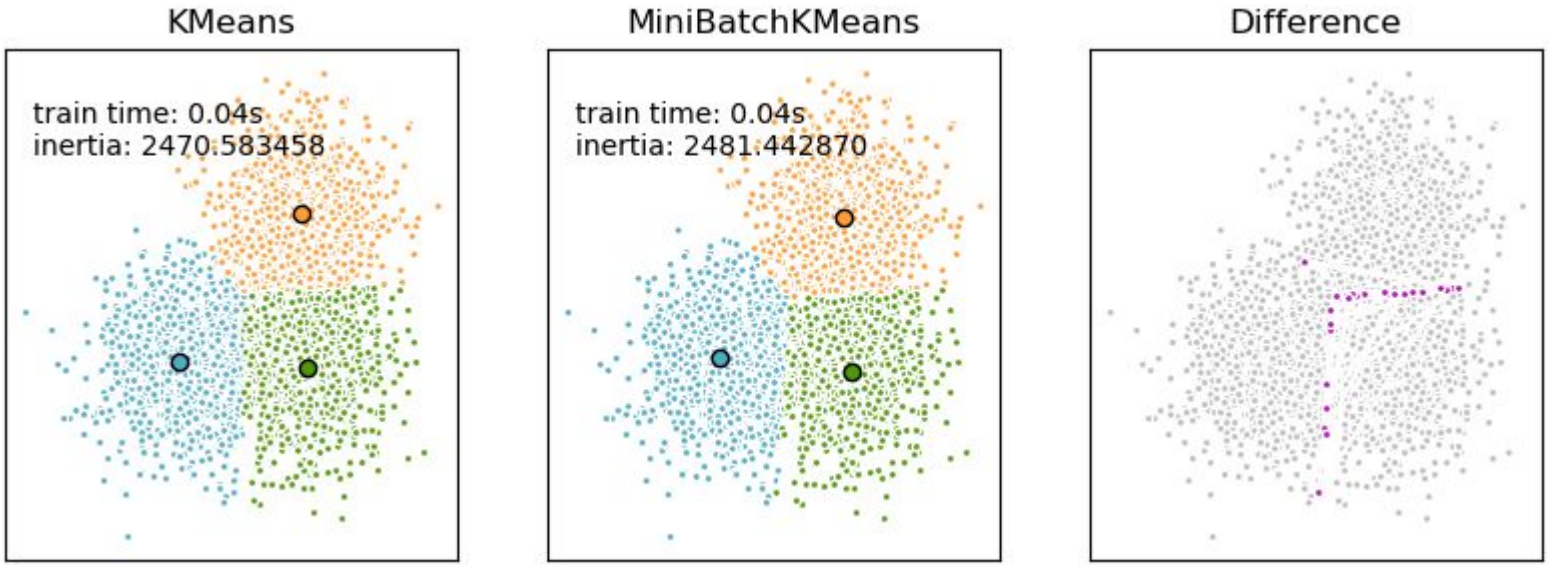
$$J(C) = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2 \to \min_C,$$

- сумма квадратов расстояний от точек до центроидов кластеров, к которым они относятся.

$$D(k) = \frac{|J(C_k) - J(C_{k+1})|}{|J(C_{k-1}) - J(C_k)|} \to \min_k$$

В идеале -- каждая точка это кластер, поэтому оцениваем, когда сумма квадратов расстояний начинает уменьшаться минимально

Mini Batch k-means



для обучения используется не весь датасет целиком, а лишь маленькие его порции (batch) и центроиды обновляются используя среднее за всю историю обновлений центроида от всех относящихся к нему точек.

Affinity propagation

Оценка того, насколько сами наблюдении похожи друг на друга - необходима метрика:

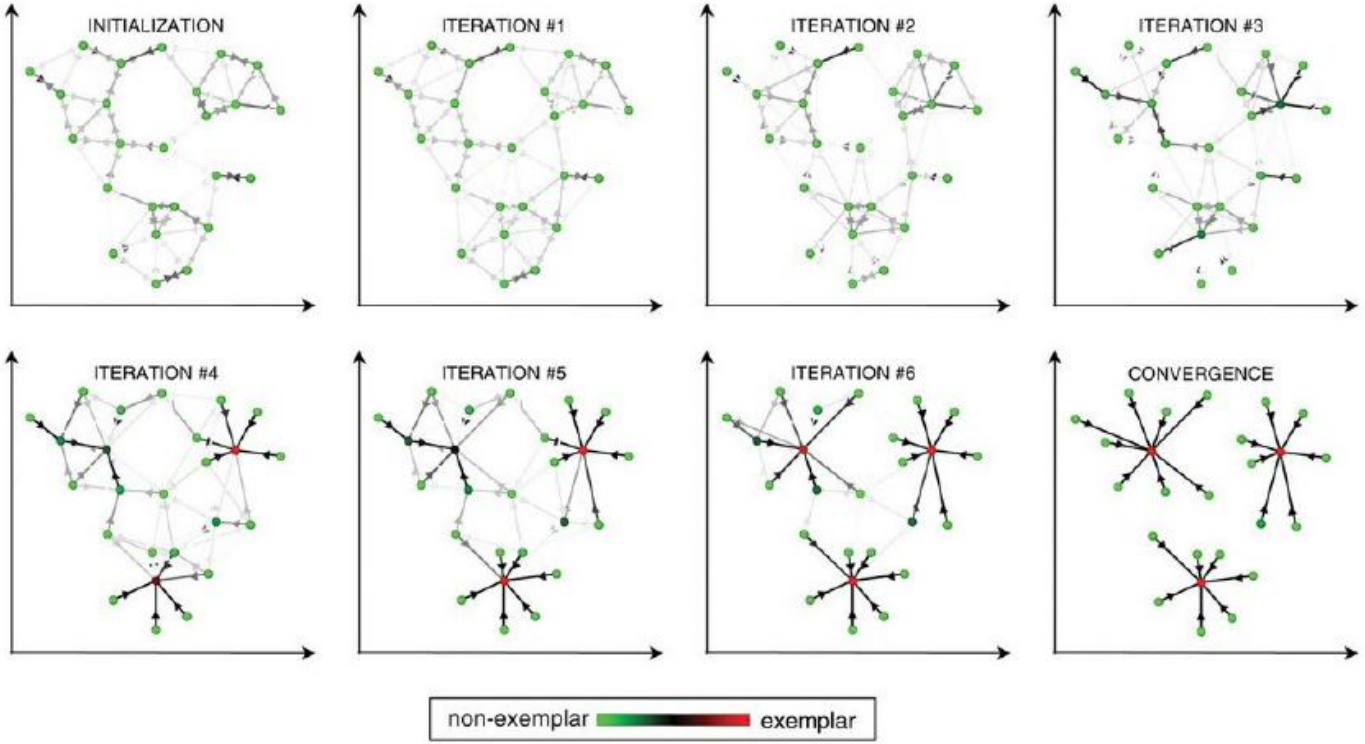
$$s(x_i, x_j) = -||x_i - x_j||^2$$

$$r_{i,k} \leftarrow s_(x_i, x_k) - \max_{k' \neq k} \left\{ a_{i,k'} + s(x_i, x'_k) \right\}$$

$$a_{i,k} \leftarrow \min \left(0, r_{k,k} + \sum_{i' \notin \{i,k\}} \max(0, r_{i',k}) \right), \quad i \neq k$$

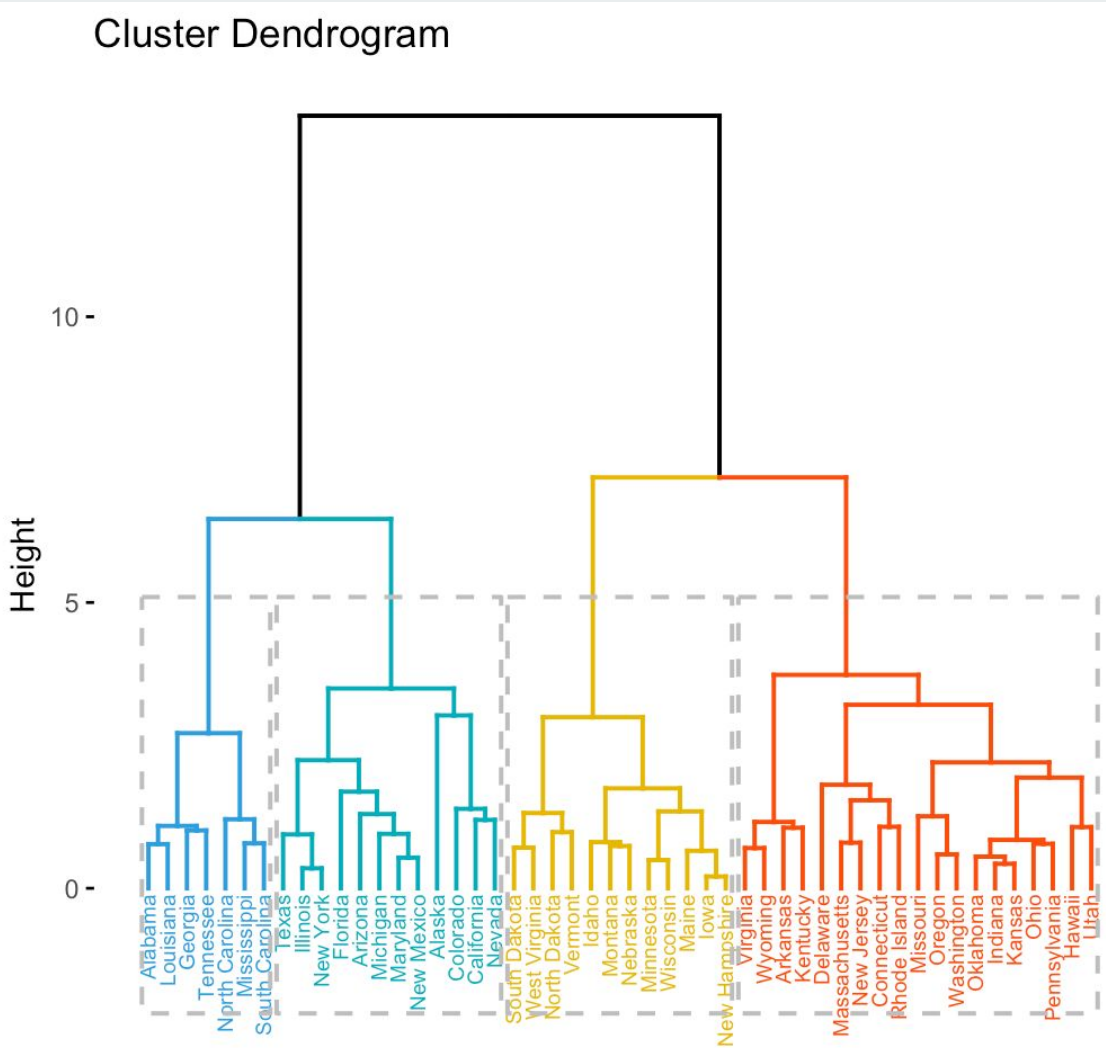
$$a_{k,k} \leftarrow \sum_{i' \neq k} \max(0, r_{i',k})$$

Affinity propagation



Агломеративная кластеризация

1. Начинаем с того, что высыпаем на каждую точку свой кластер
2. Сортируем попарные расстояния между центрами кластеров по возрастанию
3. Берём пару ближайших кластеров, склеиваем их в один и пересчитываем центр кластера
4. Повторяем п. 2 и 3 до тех пор, пока все данные не склеятся в один кластер


Cluster Dendrogram
Height
10
5
0
Alabama
Louisiana
Georgia
Tennessee
North Carolina
Mississippi
South Carolina
Texas
Illinois
New York
Florida
Arizona
Michigan
Maryland
New Mexico
Alaska
Colorado
California
Nevada
South Dakota
West Virginia
North Dakota
Vermont
Idaho
Montana
Nebraska
Minnesota
Wisconsin
Maine
Iowa
New Hampshire
Virginia
Wyoming
Arkansas
Kentucky
Delaware
Massachusetts
New Jersey
Connecticut
Rhode Island
Missouri
Oregon
Washington
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Hawaii
Utah

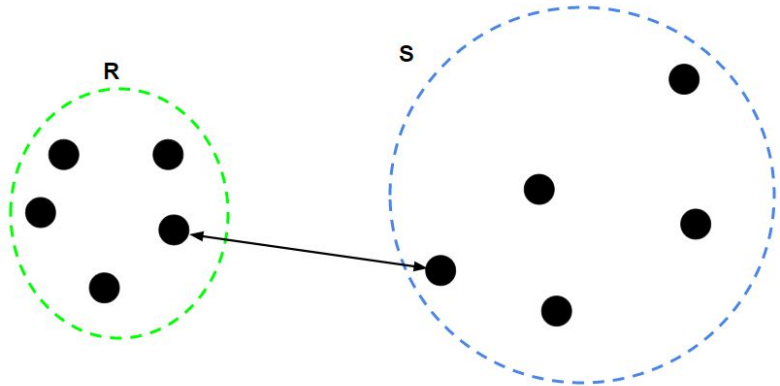
Возможные варианты реализации

1. Single linkage — минимум попарных расстояний между точками из двух кластеров
2. Complete linkage — максимум попарных расстояний между точками из двух кластеров
3. Average linkage — среднее попарных расстояний между точками из двух кластеров (UPGMA)
4. Centroid linkage — расстояние между центроидами двух кластеров
5. Ward distance — на основании расстояния Варда

Метод ближайшего соседа (single linkage)

Как работает

- к кластеру присоединяется ближайший к нему кластер/объект
- кластеры объединяются в один на расстоянии, которое равно расстоянию между ближайшими объектами этих кластеров



Особенности

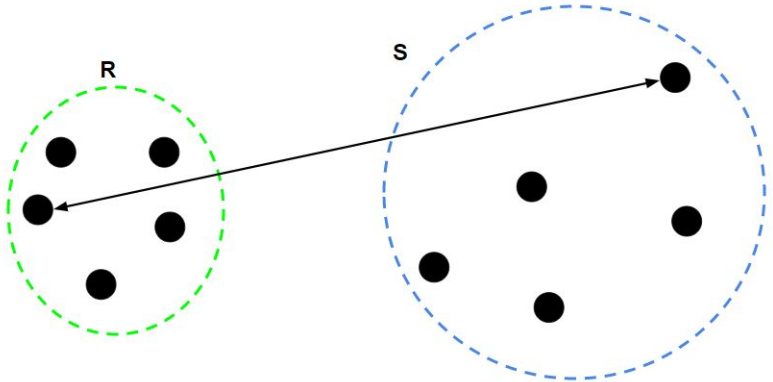
- может быть сложно интерпретировать, если нужны группы
  a. объекты на дендрограмме часто не образуют четко разделенных групп
  b. часто получаются цепочки кластеров (объекты присоединяются как бы по-одному)
- хорош для выявления градиентов

$$L(R,S) = min(D(i,j)), i \epsilon R, j \epsilon S$$

Метод отдаленного соседа (Complete linkage)

Как работает

- к кластеру присоединяется отдаленный кластер/объект
- кластеры объединяются в один на расстоянии, которое равно расстоянию между самыми отдаленными объектами этих кластеров (следстви чем более крупная группа, тем сложнее к ней присоединиться)



$$L(R,S) = max(D(i,j)), i \epsilon R, j \epsilon S$$

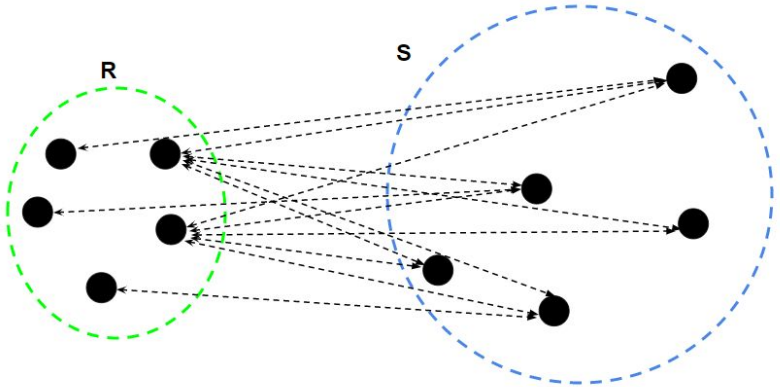
Особенности

- на дендрограмме образуется много отдельных некрупных групп
- хорош для поиска дискретных групп в данных

Average linked

Как работает

- кластеры объединяются в один на расстоянии, которое равно среднему значению всех возможных расстояний между объектами из разных кластеров.



Особенности

- может приводит к инверсиям на дендрограммах

$$L(R,S) = \frac{1}{n_R+n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i,j), i \epsilon R, j \epsilon S$$

Метрики качества кластеризации

- Adjusted Rand Index (ARI)
- Adjusted Mutual Information (AMI)
- Гомогенность, полнота, V-мера
- Силуэт

Adjusted Rand Index (ARI)

- Нам известны метрики объектов
- RI - доля объектов, для которых эти разбиения (исходное и полученное в результате кластеризации) "согласованы"
- Не зависит от значений и перестановок меток
- Интерпретация
  - -1 - независимое разбиение на кластеры
  - 0 - случайные разбиения
  - 1 - два разбиения схожи

$$\mathrm{RI} = \frac{2(a+b)}{n(n-1)}.$$

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}.$$

Adjusted Mutual Information (AMI)

- Не зависит от значений и перестановок меток
- Взаимная информация измеряет долю информации, общей для обоих разбиений: насколько информация об одном из них уменьшает неопределенность относительно другого.
- Интерпретация
  - -1 - независимое разбиение на кластеры
  - 0 - случайные разбиения
  - 1 - два разбиения схожи

Гомогенность, полнота, V-мера

- Определяются с использованием функций энтропии и условной энтропии, рассматривая разбиения выборки как дискретные распределения
- K - результат кластеризации, C - истинное разбиение выборки на классы
- V-мера - среднее гармоническое -- показывает сходство кластеров

$$h = 1 - \frac{H(C \mid K)}{H(C)}, c = 1 - \frac{H(K \mid C)}{H(K)},$$

Силуэт

- не предполагает знания истинных меток объектов,
- a - среднее расстояние от данного объекта до объектов из того же кластера
- b - среднее расстояние от данного объекта до объектов из ближайшего кластера (отличного от того, в котором лежит сам объект)
- Интерпретация
  - -1 - кластеры перекрываются
  - 1 - хорошо, четко выделяемые кластеры

$$s = \frac{b - a}{\max(a, b)}.$$

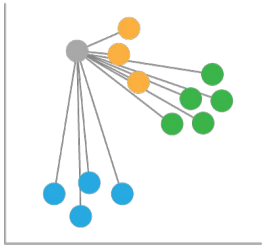
t-SNE

t-SNE (t-distributed stochastic neighbor embedding)

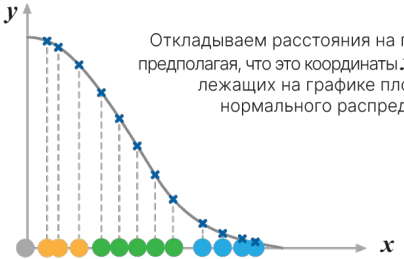
Идея состоит не в том, чтобы напрямую максимизировать дисперсию, а найти такое пространство в котором расстояние между объектами будет сохраняться или по крайне мере не сильно меняться. При этом будем больше беспокоиться о расстоянии между близкими объектами, нежели о расстоянии между далекими

Описываем расстояния в исходном пространстве

1.



Считаем все расстояния от заданной точки до остальных



Откладываем расстояния на прямой, предполагая, что это координаты x точек, лежащих на графике плотности нормального распределения

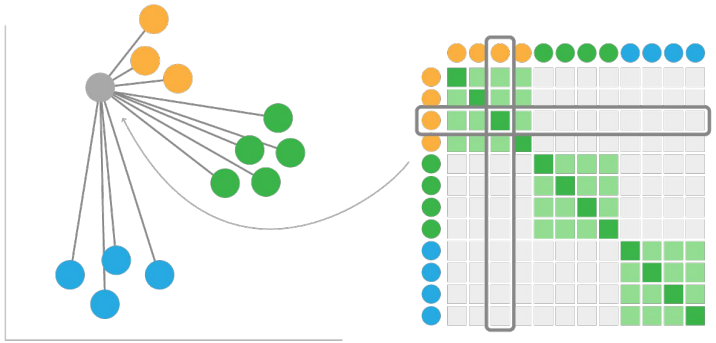
И в качестве ненормированных расстояний используем координаты y этих точек

2.

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||_2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||\mathbf{x}_i - \mathbf{x}_k||_2 / 2\sigma_i^2)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

3.

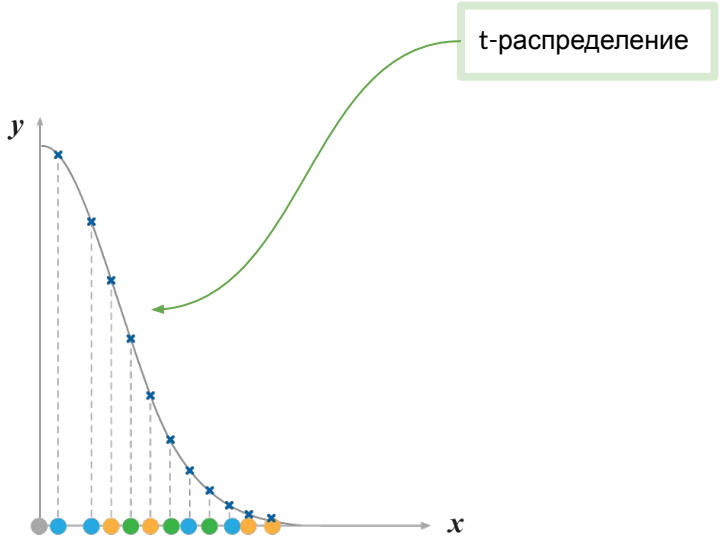


Высокая similarity

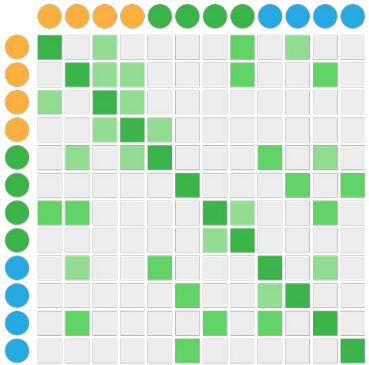
Низкая similarity

Описываем расстояния в пространстве низкой размерности

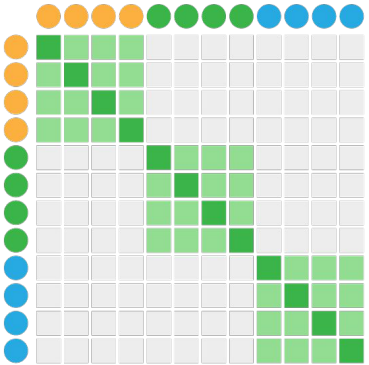
Снова считаем “похожести”...
на этот раз назовем их не p, а q



Но они же не похожи...



Матрица расстояний
в пространстве низкой размерности



Матрица расстояний
в пространстве высокой размерности

$$Loss = KL(P||Q) = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}}$$

Важные параметры tSNE

perplexity
Определяет то, как подбирается стандартное отклонение для распределения расстояний для каждой точки. Чем больше perplexity - тем более на глобальную структуру мы смотрим

metric
Как считаются расстояния между точками - metric. По-умолчанию используется евклидово расстояние, но часто помогают и другие (например, косинусное)

learning_rate
Шаг градиентного спуска, тоже влияет на полученное представление

Минусы tSNE

1. Стохастичность
2. Добавление новых точек
3. Расстояния между кластерами точек могут ничего не значить (плохо сохраняются далекие расстояния)
4. Размеры кластеров ничего не значат
5. Можно увидеть артефактные кластеры
6. Можно увидеть не ту структуру, которая по идее должна быть

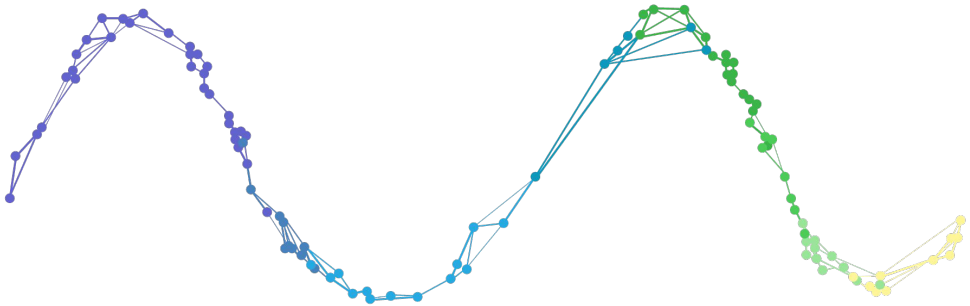
Tricks

1. Инициализация при помощи PCA
2. Kernel PCA

UMAP

UMAP (uniform manifold approximation and projection)

Внутри себя метод строит граф, в котором ребрами соединены между собой k ближайших соседей. При этом эти ребра неравноправны - если для данной пары точек расстояние между ними сильно больше, чем расстояния между ними и другими точками - то и ребро будет иметь маленький вес.



Далее задача состоит в том, чтобы в пространстве более низкой размерность получился граф похожий на тот, который был в высокой размерностью. Для этого опять же, оптимизируем низкоразмерное представление градиентным спуском

Плюсы

1. Быстрее чем tSNE
2. Можно добавлять новые данные