IT-Academy

Кластеризация

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

Вопросы занятия

IT-Academy

1. Задача кластеризации: постановка и примеры;

2. Основные алгоритмы;

3. Метрики качества кластеризации.

В конце занятия научимся:

- производить кластеризацию данных;

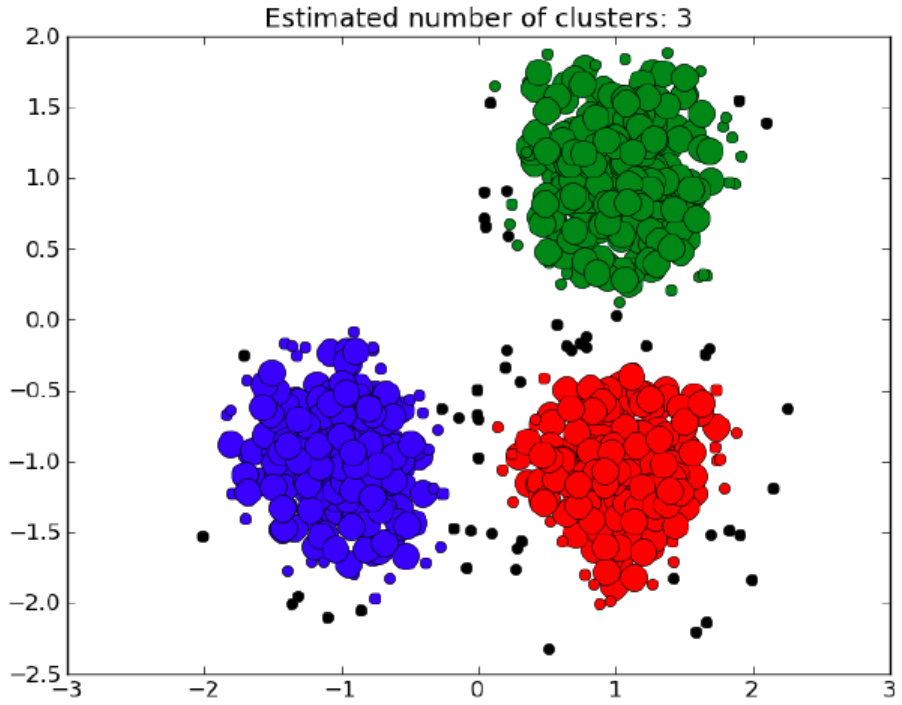
- выбирать наиболее подходящий алгоритм для задачи.

ЗАДАЧА КЛАСТЕРИЗАЦИИ

IT-Academy

ТИПЫ ЗАДАЧ

- классификация
- ранжирование
- регрессия
- **кластеризация**



IT-Academy

ПРИМЕРЫ ЗАДАЧ КЛАСТЕРИЗАЦИИ

Пользовательская сегментация. Как выглядят типичные пользователи? (находим сектора, работаем с ними отдельно)

Логистика. Где расположить магазины, чтобы охватить большее кол-во покупателей?

Новости. О чём сейчас пишут СМИ? (Новостной портал кластеризует новости и выдает их отдельными темами)

EDA. Есть 100млн обращений пользователей. О чём они пишут?

IT-Academy

ДОПОЛНИТЕЛЬНЫЕ ПРИЛОЖЕНИЯ

Создание дополнительных фич. Можно дополнить имеющиеся данные метками принадлежности к одному из классов;

Разметка данных. Если нет проставленных классов, но нужно сделать классификатор, то в создании разметки для обучающей выборки сильно поможет кластеризация;

Поиск структуры данных как часть эксплоративного анализа.

ЗАДАЧА КЛАСТЕРИЗАЦИИ

IT-Academy

Постановка задачи кластеризации

Пусть X — множество объектов, Y — множество идентификаторов (меток) кластеров. На множестве X задана функция расстояния между объектами $\rho(x,x')$.

Дана конечная обучающая выборка объектов $X^{**}m=\{x1,\ldots,xm\}\subset X$. Необходимо разбить выборку на подмножества (кластеры), то есть каждому объекту $x_i \in X^{**}m$ сопоставить метку $y_i \in Y$, таким образом чтобы объекты внутри каждого кластера были близки относительно метрики ρ, а объекты из разных кластеров значительно различались.

Определение

Алгоритм кластеризации — функция $a:X\to Y$, которая любому объекту $x \in X$ ставит в соответствие идентификатор кластера $y \in Y$.

ЗАДАЧА КЛАСТЕРИЗАЦИИ

IT-Academy

Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации.
Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки объектов из обучающей выборки y_i изначально не заданы, и даже может быть неизвестно само множество Y.

Решение задачи кластеризации объективно неоднозначно по ряду причин:

- Не существует однозначного критерия качества кластеризации. Известен ряд алгоритмов, осуществляющих разумную кластеризацию "по построению", однако все они могут давать разные результаты. Следовательно, для определения качества кластеризации и оценки выделенных кластеров необходим эксперт предметной области;
- Число кластеров, как правило, заранее не известно и выбирается по субъективным критериям. Даже если алгоритм не требует изначального знания о числе классов, конкретные реализации зачастую требуют указать этот параметр;
- Результат кластеризации существенно зависит от метрики. Однако существует ряд рекомендаций по выбору метрик для определенных классов задач.

Число кластеров фактически является гиперпараметром для алгоритмов кластеризации.

Теорема невозможности Клейнберга

IT-Academy

Клейнберг постулировал три простых свойства в качестве аксиом кластеризации и доказал теорему, связывающую эти свойства.

- Алгоритм кластеризации *a* является **масштабно инвариантным** (англ. *scale-invariant*), если для любой функции расстояния ρ и любой константы $\alpha>0$ результаты кластеризации с использованием расстояний ρ и $\alpha \cdot \rho$ совпадают.

- **Полнота** (англ. *Richness*). Множество результатов кластеризации алгоритма aa в зависимости от изменения функции расстояния ρρ должно совпадать со множеством всех возможных разбиений множества объектов *X*.

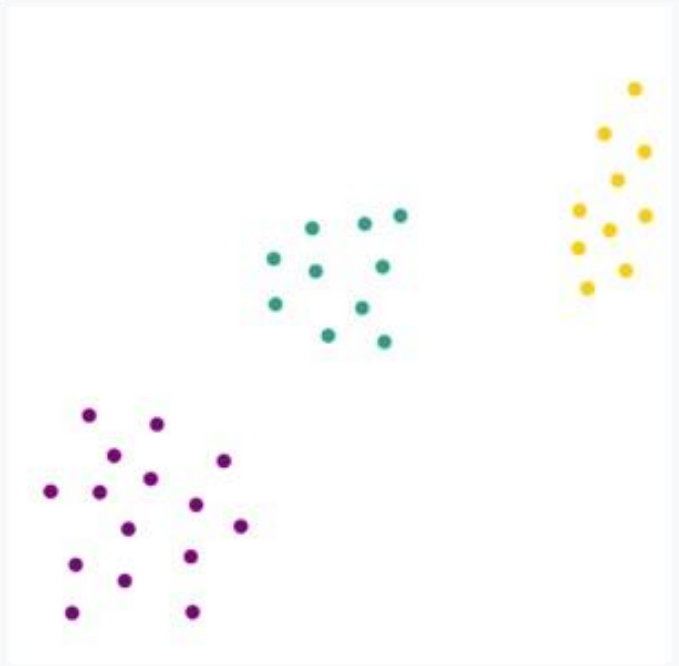
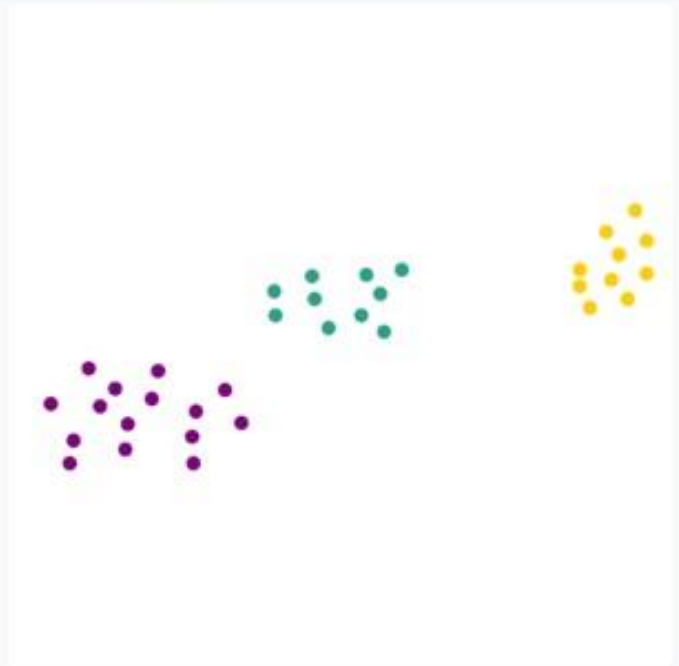
- Алгоритм кластеризации является **согласованным** (англ. *consistent*), если результат кластеризации не изменяется после допустимого преобразования функции расстояния.

Исходя из этих аксиом Клейнберг сформулировал и доказал теорему:

Для множества объектов, состоящего из двух и более элементов, не существует алгоритма кластеризации, который был бы одновременно масштабно-инвариантным, согласованным и полным.

Теорема невозможности Клейнберга

IT-Academy

Примеры преобразований с сохранением кластеров		
		
Исходное расположение объектов и их кластеризация	Пример масштабной инвариантности. Уменьшен масштаб по оси ординат в два раза.	Пример допустимого преобразования. Каждый объект в два раза приближен к центроиду своего класса. Внутриклассовое расстояние уменьшилось, межклассовое увеличилось.

Типология задач кластеризации

IT-Academy

Типы входных данных

- Признаковое описание объектов. Каждый объект описывается набором своих характеристик, называемых признаками (англ. *features*). Признаки могут быть как числовыми, так и категориальными;
- Матрица расстояний между объектами. Каждый объект описывается расстоянием до всех объектов из обучающей выборки.

Цели кластеризации

- Классификация объектов. Попытка понять зависимости между объектами путем выявления их кластерной структуры. Разбиение выборки на группы схожих объектов упрощает дальнейшую обработку данных и принятие решений, позволяет применить к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»). В данном случае стремятся уменьшить число кластеров для выявления наиболее общих закономерностей;
- Сжатие данных. Можно сократить размер исходной выборки, взяв один или несколько наиболее типичных представителей каждого кластера. Здесь важно наиболее точно очертить границы каждого кластера, их количество не является важным критерием;
- Обнаружение новизны (обнаружение шума). Выделение объектов, которые не подходят по критериям ни в один кластер. Обнаруженные объекты в дальнейшем обрабатывают отдельно.

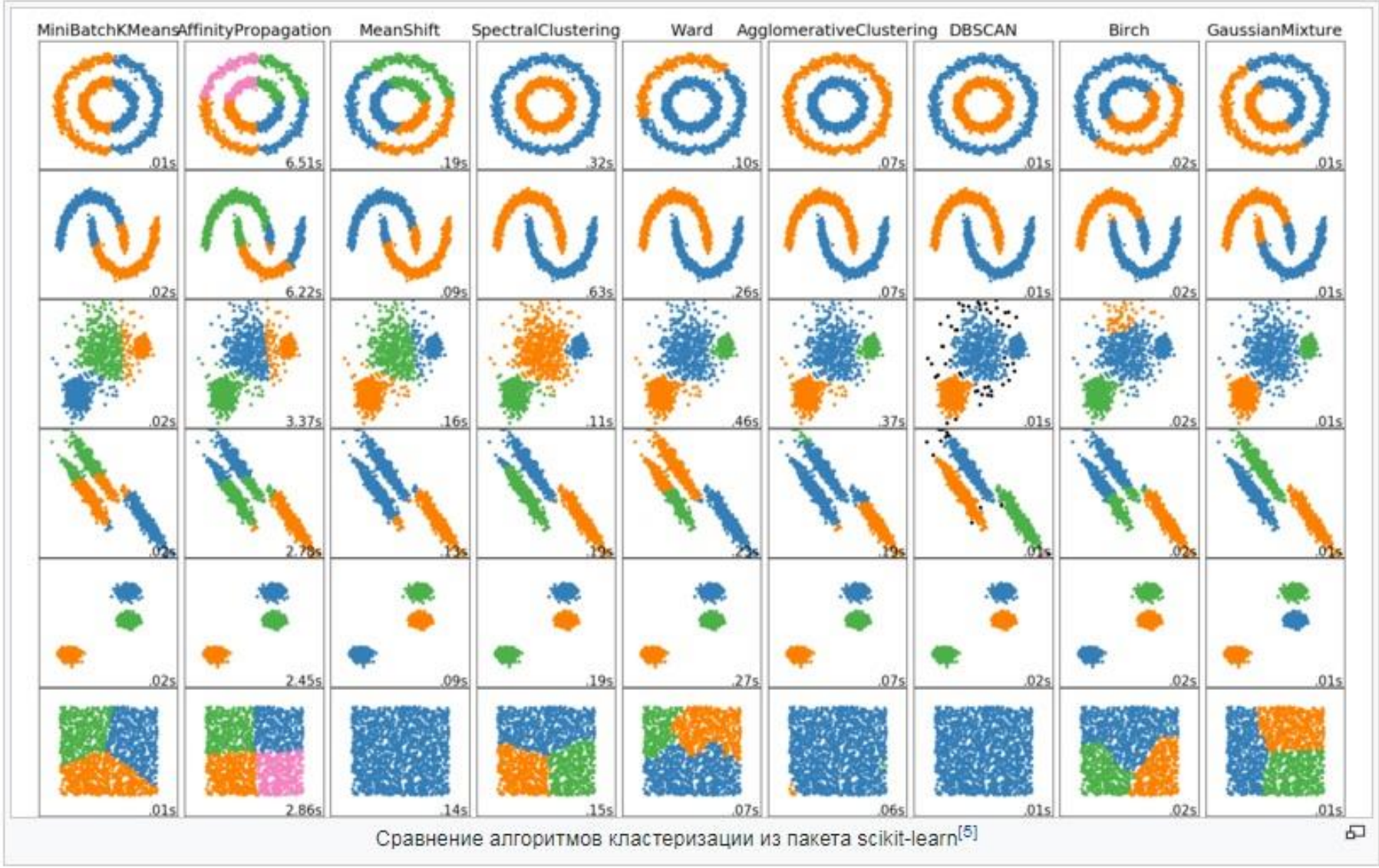
Разнообразие алгоритмов кластеризации

Методы кластеризации

•Графовые алгоритмы кластеризации. Наиболее примитивный класс алгоритмов. В настоящее время практически не применяется на практике;
•Вероятностные алгоритмы кластеризации. Каждый объект из обучающей выборки относится к каждому из кластеров с определенной степенью вероятности;
•Иерархические алгоритмы кластеризации. Упорядочивание данных путем создания иерархии вложенных кластеров;
•Алгоритм *k*-средних (англ. *k-means*). Итеративный алгоритм, основанный на минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров;
•Распространение похожести (англ. *affinity propagation*). Распространяет сообщения о похожести между парами объектов для выбора типичных представителей каждого кластера;
•Сдвиг среднего значения (англ. *mean shift*). Выбирает центроиды кластеров в областях с наибольшей плотностью;
•Спектральная кластеризация (англ. *spectral clustering*). Использует собственные значения матрицы расстояний для понижения размерности перед использованием других методов кластеризации;
•Основанная на плотности пространственная кластеризация для приложений с шумами (англ. *Density-based spatial clustering of applications with noise*, *DBSCAN*). Алгоритм группирует в один кластер точки в области с высокой плотностью. Одиноко расположенные точки помечает как шум.

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

IT-Academy



Сравнение алгоритмов кластеризации из пакета scikit-learn[5]

Как и в других задачах:

разные алгоритмы справляются лучше с разными формами зависимостей

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

Меры качества кластеризации

Для оценки качества кластеризации задачу можно переформулировать в терминах задачи дискретной оптимизации. Необходима так сопоставить объектам из множества X метки кластеров, чтобы значение выбранного функционала качества приняло наилучшее значение.

В качестве примера, стремятся достичь минимума среднего внутрикластерного расстояния или максимума среднего межкластерного расстояния.

IT-Academy

ТИПЫ КЛАСТЕРИЗАЦИИ

Жёсткая кластеризация

(1 объект - 1 класс)

Мягкая (fuzzy) кластеризация

(1 объект - несколько (или 0) классов)

Иерархическая кластеризация

(объект внутри кластера 2.1 -> внутри кластера 2)

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. PREPROCESSING

IT-Academy

PREPROCESSING

Все методы кластеризации основываются на метриках и потому крайне чувствительны к одному масштабу данных, поэтому

StandardScaler - must have

IT-Academy

K-MEANS

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS

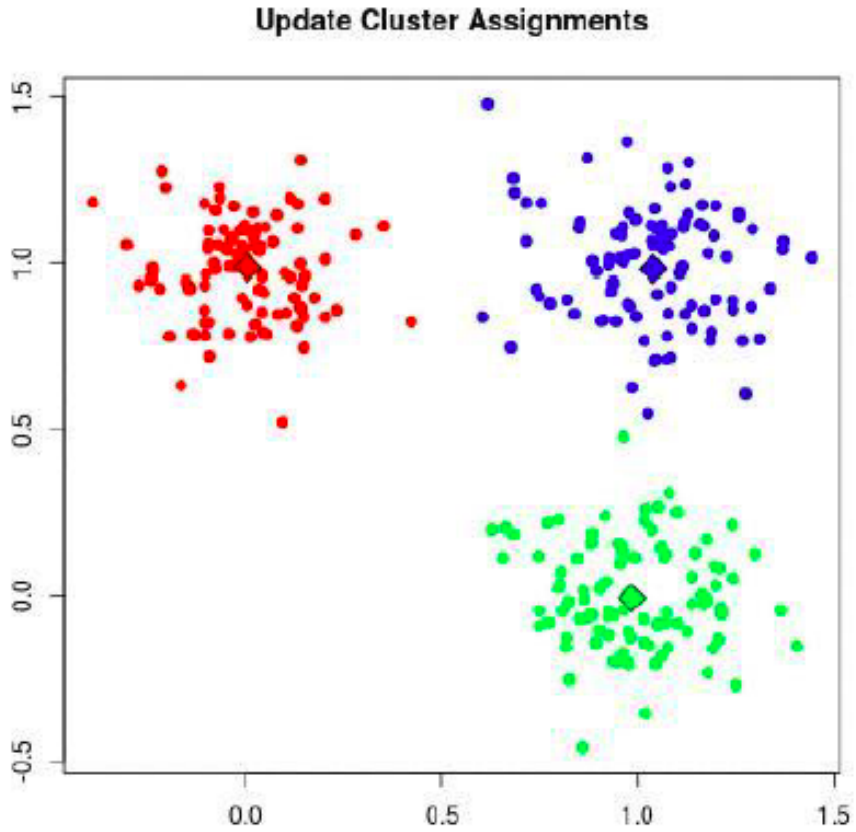
IT-Academy

АЛГОРИТМ

Задать начальные значения центроидов кластеров

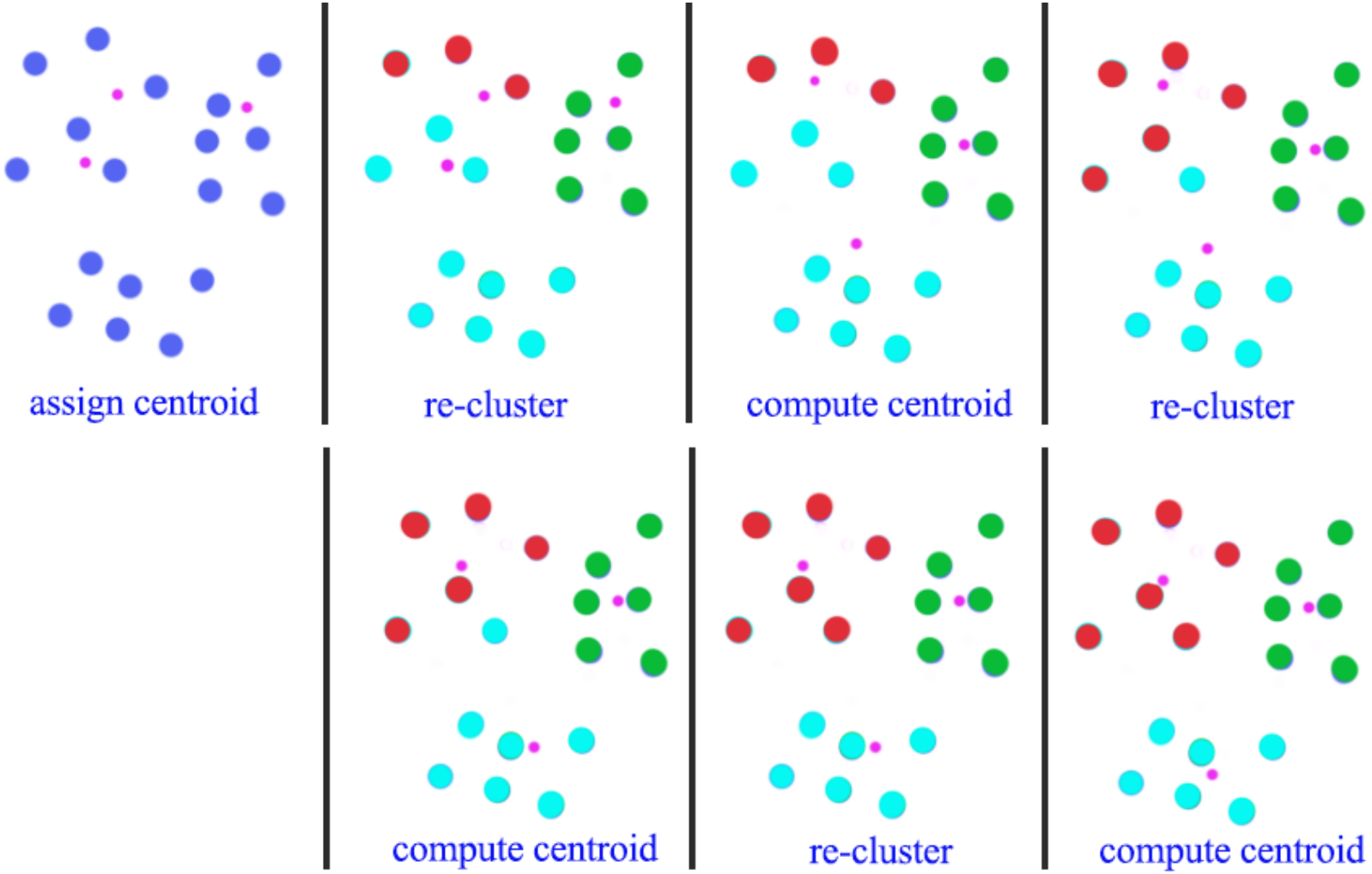
Повторять, пока центроиды смещаются:

* присвоить наблюдениям номер кластера с **ближайшим** к ним центром
* передвинуть центроиды кластеров к среднему значению координат членов кластера



АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS

IT-Academy



АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS

IT-Academy

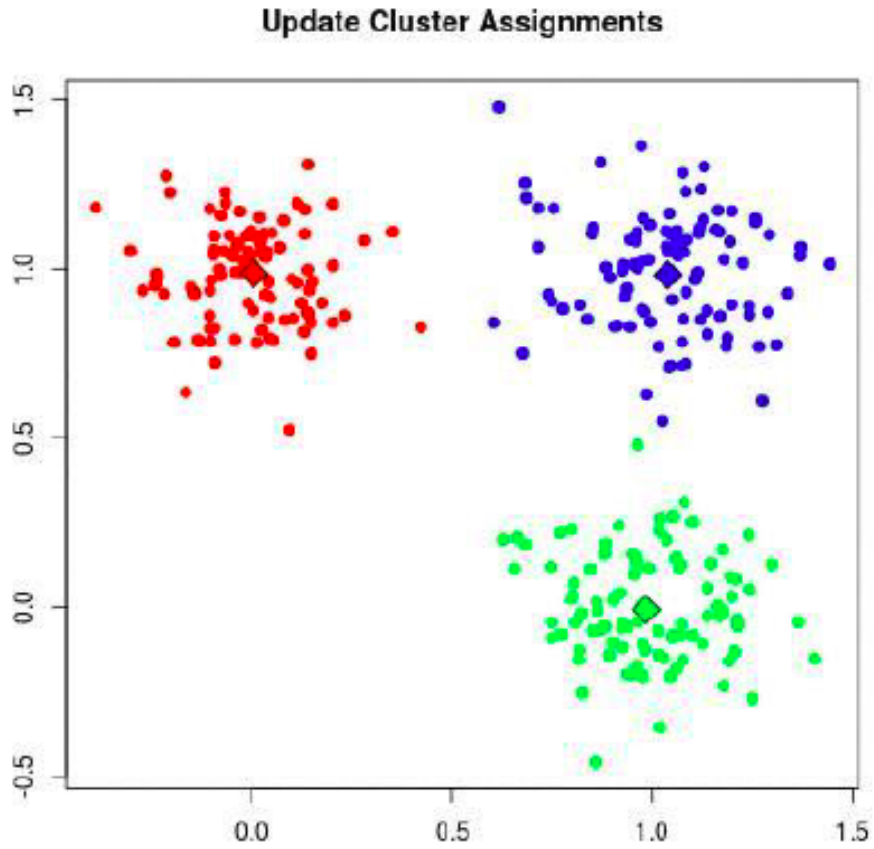
ЦЕЛЬ

Минимизировать внутриклассовые отличия от центроида:

$$\sum_{i=0}^{n} \min_{\mu_j}(\|x_i - \mu_j\|)^2$$

Связанные с этим проблемы:

* предположение о выпуклости и однородности кластеров
* проклятие размерности

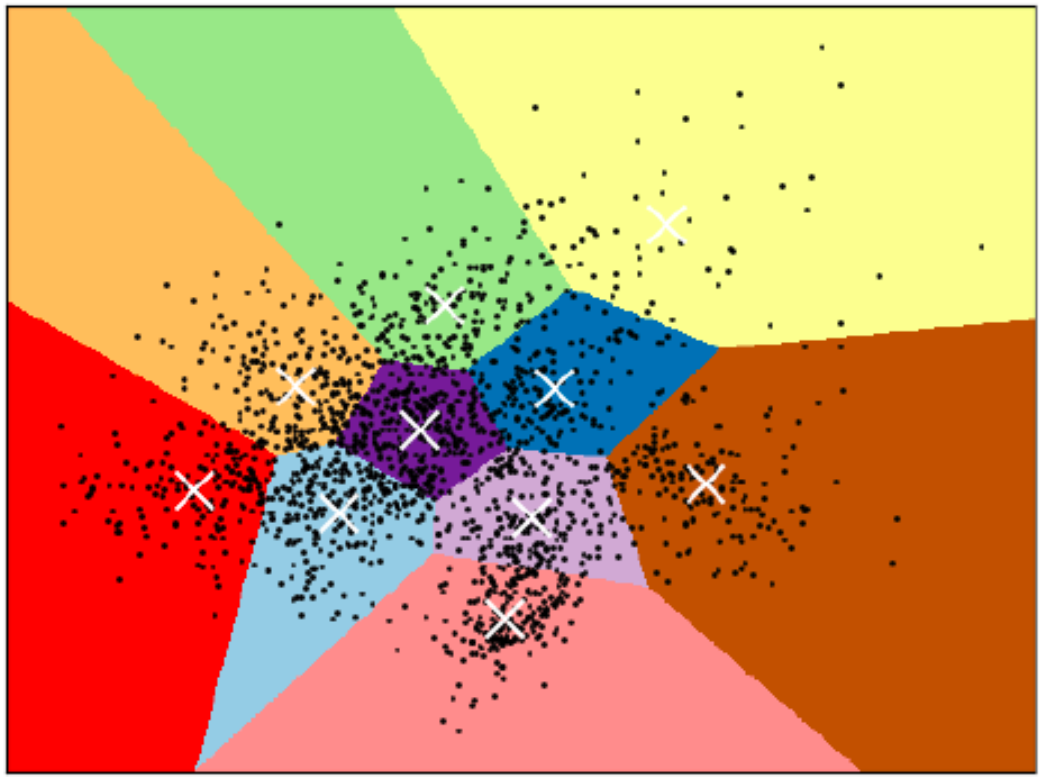


АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS

IT-Academy

ИТОГ

K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



Пространство нарезается на лоскуты из прямых гиперплоскостей

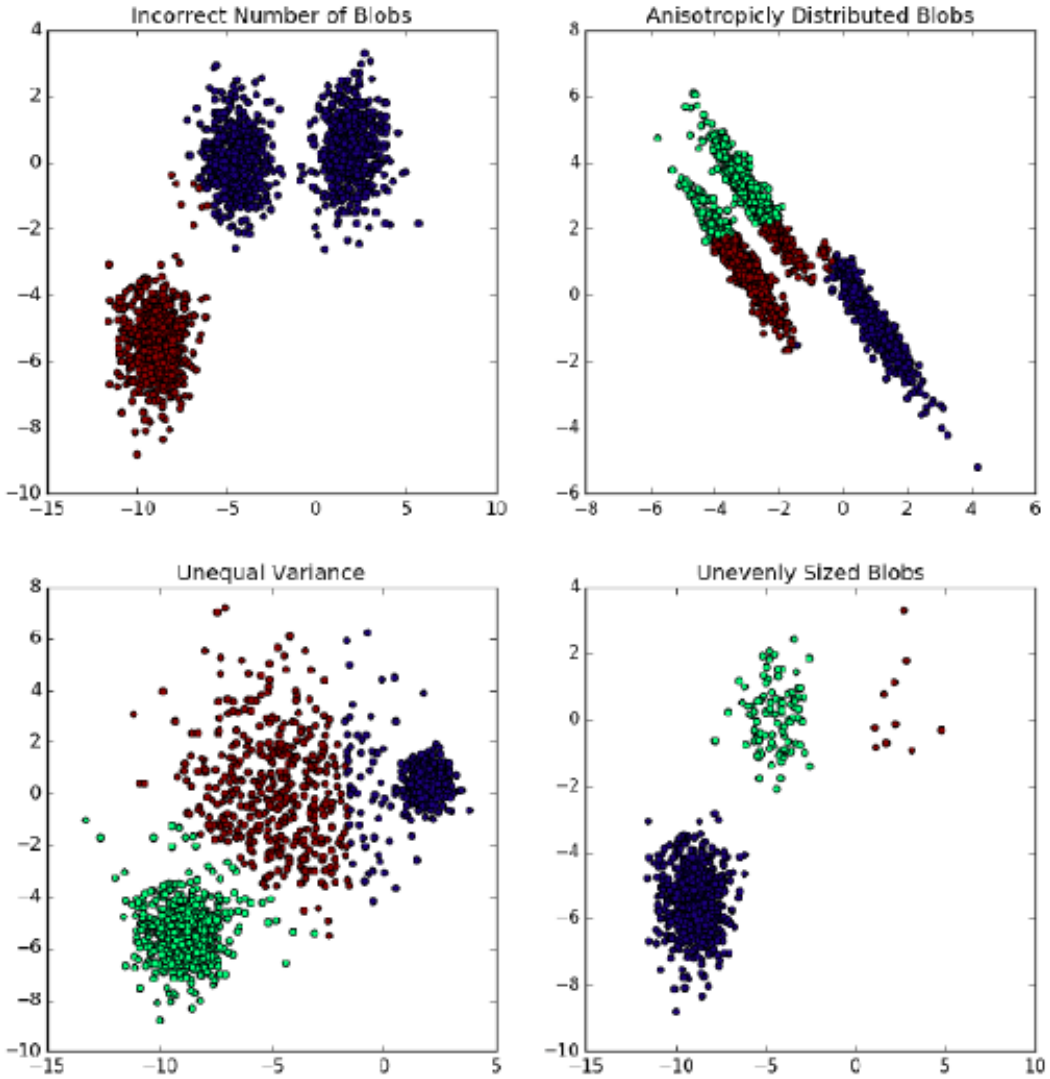
АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS

IT-Academy

ОГРАНИЧЕНИЯ

Алгоритм может выдавать контринтуитивные результаты

1. Если указано не то число кластеров
2. Кластеры - не выпуклые и близко расположены
3. Разная дисперсия близких кластеров



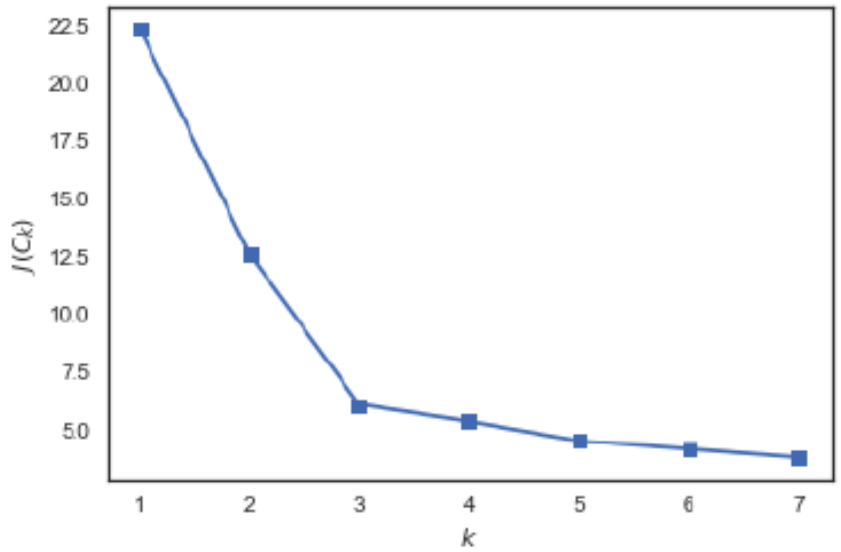
IT-Academy

КОЛИЧЕСТВО КЛАСТЕРОВ

Идея: перебирать от 1 до N кластеров, засечь, с какого момента *качество* перестанет быстро улучшаться

$$(\text{т.е. } k^* = \operatorname{argmin}(\Delta J(C_{k+1}) / \Delta J(C_k)))$$

Качество - сумма квадратов расстояний от точек до центроидов кластеров



IT-Academy

НАЧАЛЬНОЕ ПРИБЛИЖЕНИЕ

Алгоритм очень зависит от начального приближения: метод сойдётся всегда, но к разным локальным минимумам.

Какие точки выбрать?

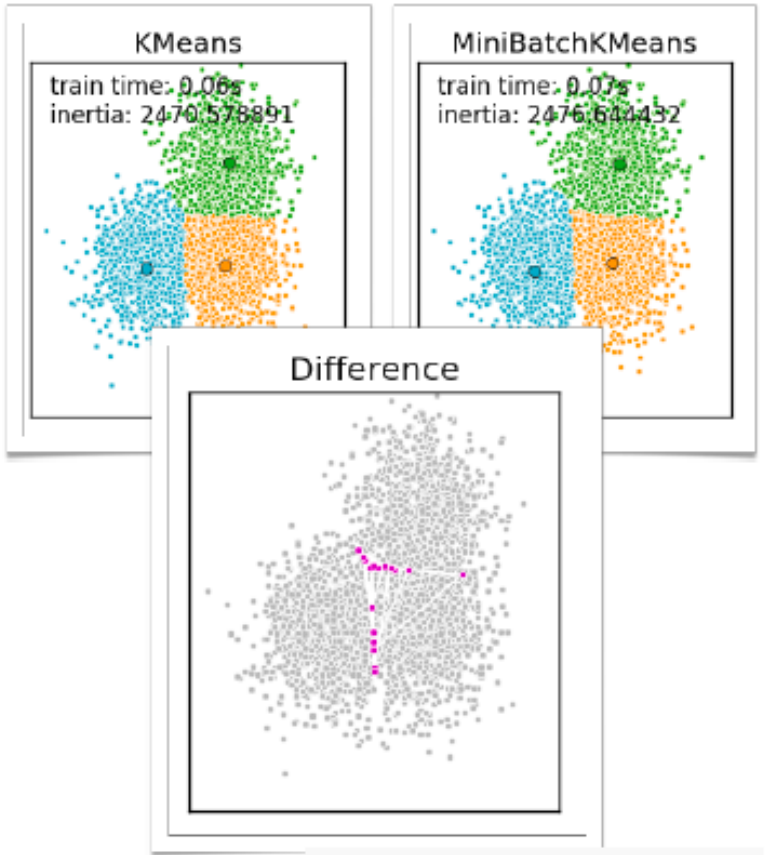
* **Мультистарт**: N наборов начальных приближений, выбор лучшего
* **Наиболее удалённые** друг от друга точки:
  * удалить аномалии (посчитать среднее расстояние до q ближайших соседей, отбросить δ% самых удалённых)
  * взять 2 самые дальние друг от друга точки, они составят множество U
  * k-2 раз добавлять в U по 1 точке, расстояние которой до ближайшей из старых точек U будет максимально большим

IT-Academy

УСКОРЕНИЕ. MINI BATCH KMEANS

Способ: каждый шаг брать не все точки, а лишь подмножества (batch), обновляя центроиды как среднее признаков объектов кластера как текущего, так и всех предыдущих шагов

Результат: рост скорости с мизерным падением качества



IT-Academy

РАЗВИТИЕ. МЯГКИЙ ВАРИАНТ (EM)

Более мягкий вариант KMeans: каждому объекту ставить в соответствие не 1 кластер, а вектор близости к каждому кластеру

Повторять, пока центроиды смещаются:

* оценить близость каждого объекта к каждому центроиду кластеров
* присвоить объектам номер кластера с ближайшим к ним центром
* передвинуть центроиды кластеров к **средневзвешенному** значению координат всех объектов, взвешивая по близости объекта к текущему центроиду кластера

IT-Academy

РЕАЛИЗАЦИЯ В SKLEARN

sklearn.cluster.KMeans

* n_clusters=8
* init='k-means++'
* n_init=10
* max_iter=300
* tol=0.0001
* precompute_distances='auto'
* verbose=0
* random_state=None
* copy_x=True
* n_jobs=1
* algorithm='auto'

Основные параметры

* n_clusters - количество кластеров для разбиения
* init: 'k-means+', 'random', ndarray - начальное приближение
* max_iter - кол-во итераций
* n_jobs - кол-во процессоров (-1 - max)

Основные характеристики

* 11 параметров
* по умолчанию: 10 начальных умных запусков на 1 процессоре, кластеризация на 8 групп

Основные методы

* fit, fit_predict, fit_transform, transform, predict

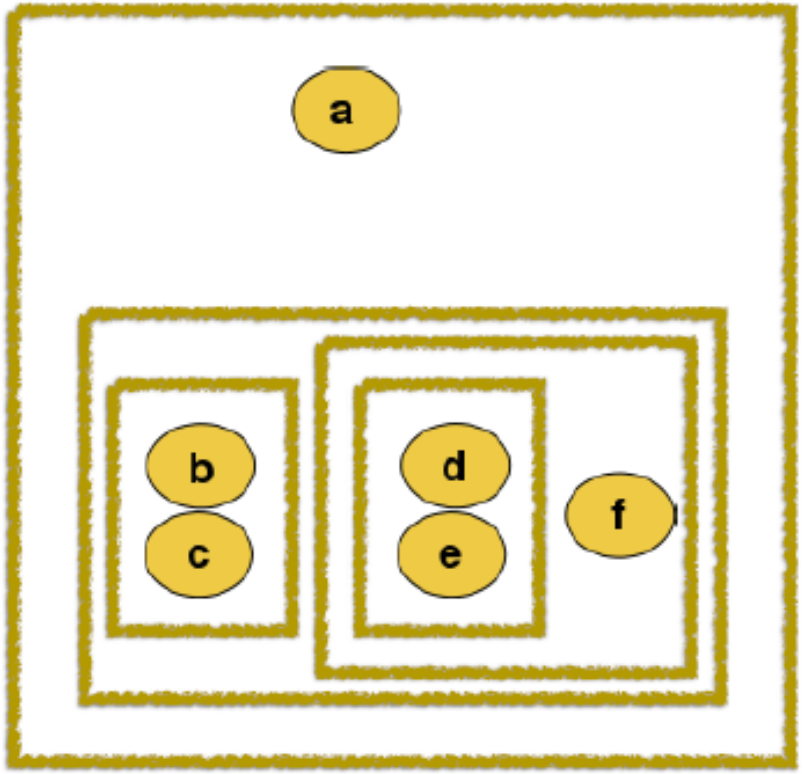
IT-Academy

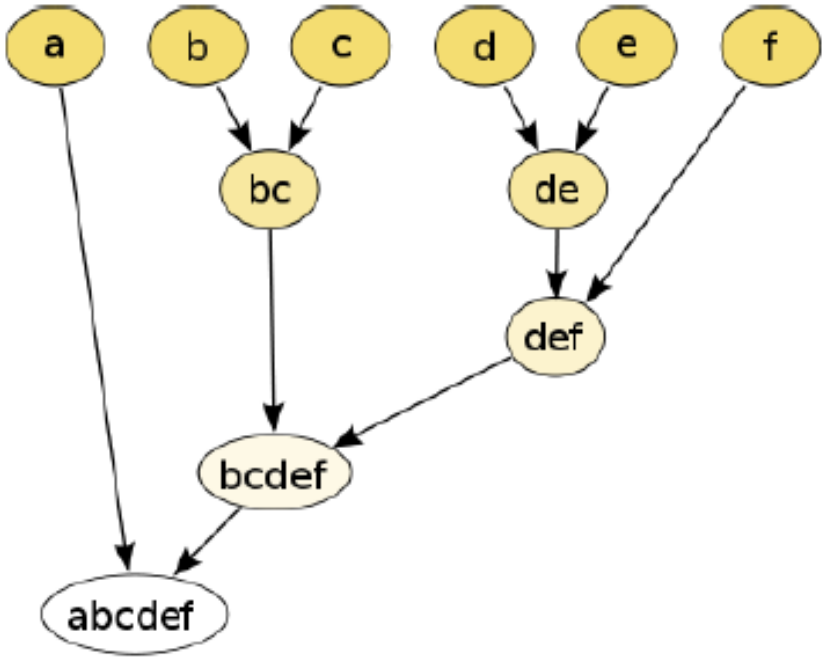
HIERARCHICAL CLUSTERING

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING

IT-Academy

ИДЕЯ





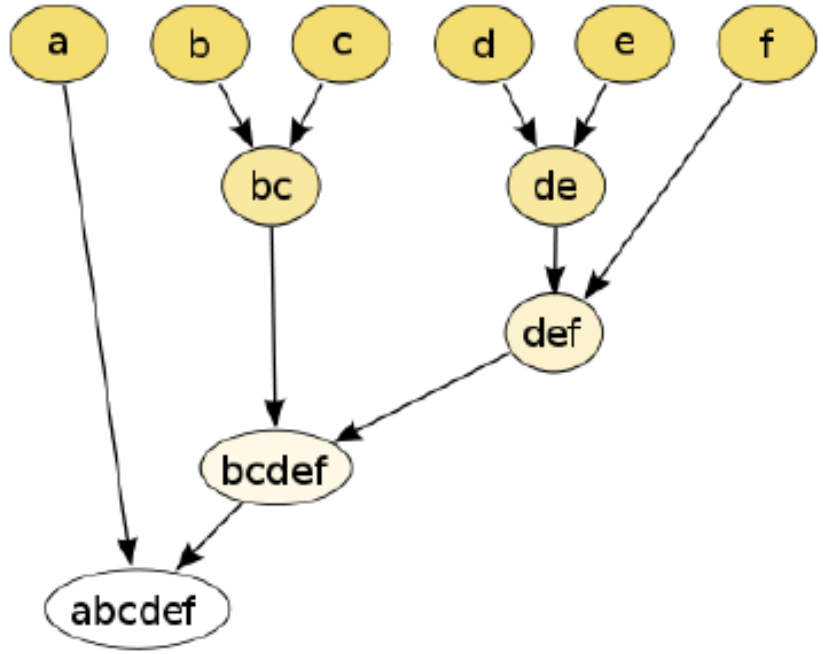
IT-Academy

АЛГОРИТМ

Все объекты - отдельные кластеры

Повторять, пока > 1 кластера:

* соединить 2 **ближайших** кластера



АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING

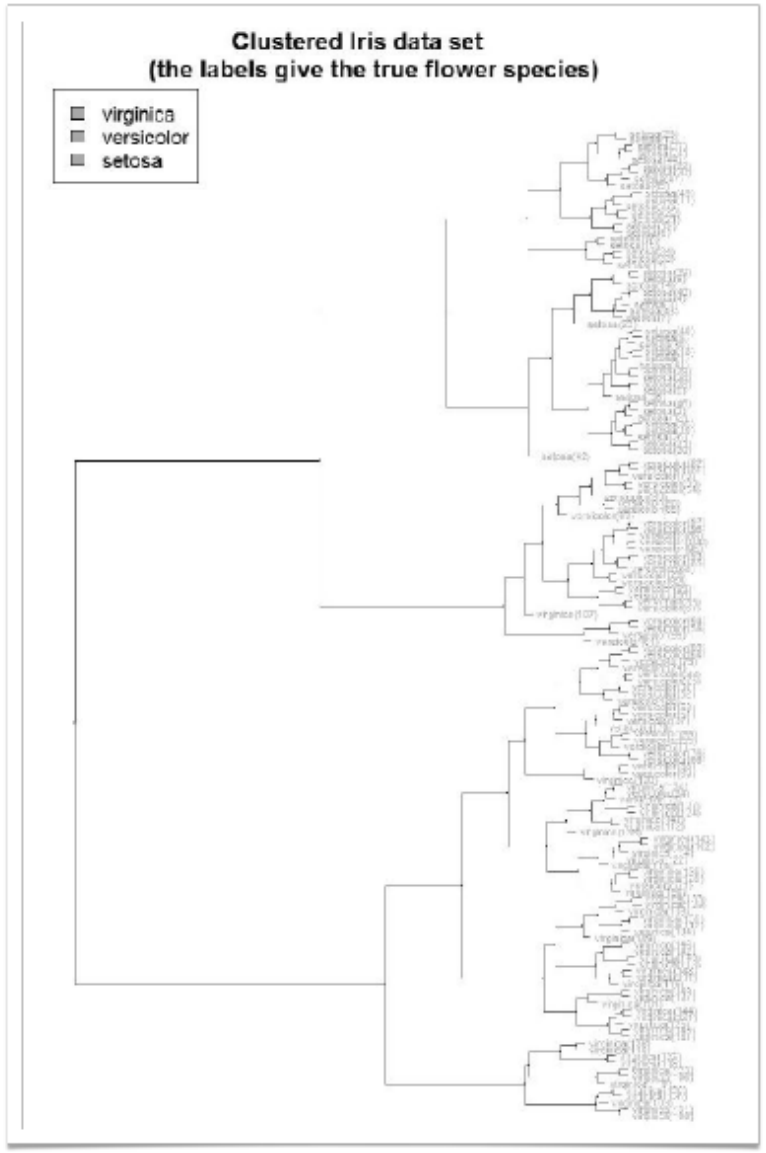
IT-Academy

ПРИМЕР

Дендрограмма кластеризации цветков ириса.

Проведена иерархическая кластеризация, визуально отображаемая в виде дендрограммы.

На картинке цветом линии отмечены 3 кластера, а цветом надписи - настоящий вид цветка



АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING

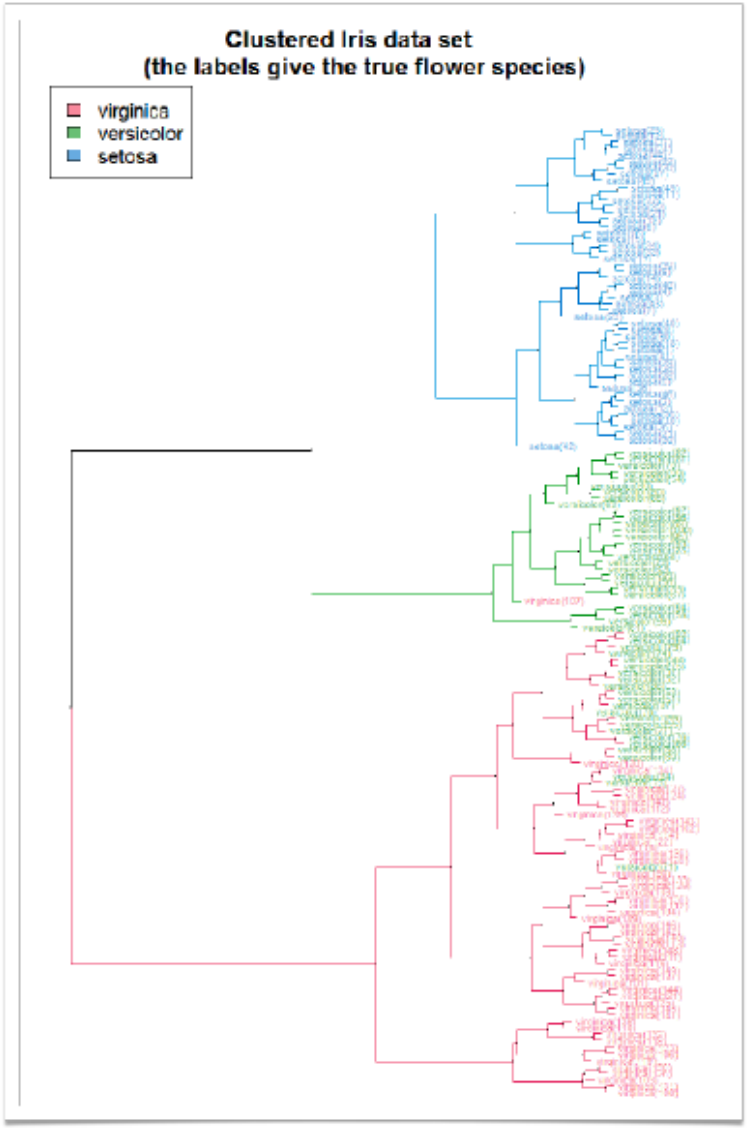
IT-Academy

ПРИМЕР

Дендрограмма кластеризации цветков ириса.

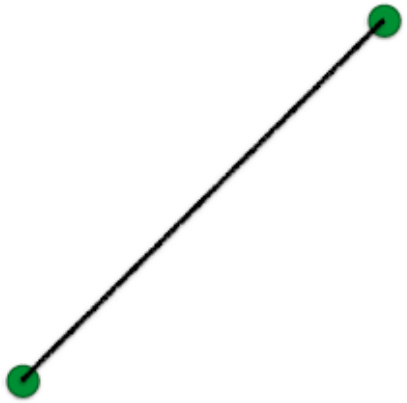
Проведена иерархическая кластеризация, визуально отображаемая в виде дендрограммы.

На картинке цветом линии отмечены 3 кластера, а цветом надписи - настоящий вид цветка

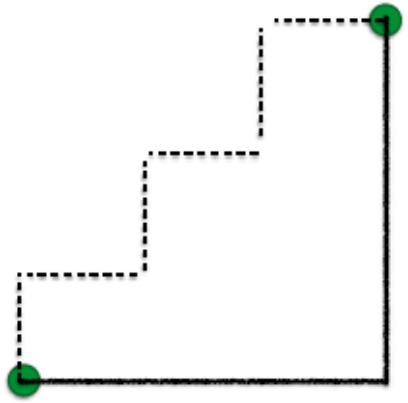


IT-Academy

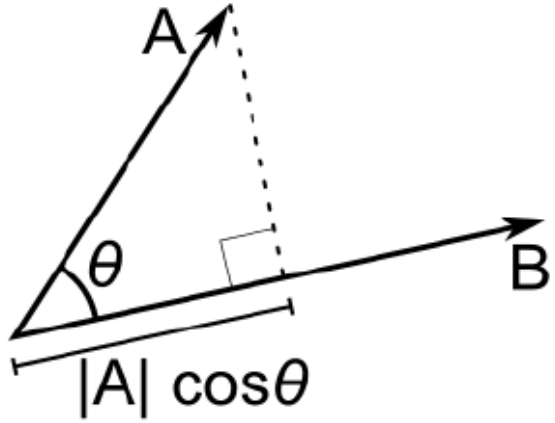
РАССТОЯНИЕ МЕЖДУ ОБЪЕКТАМИ



euclide (l2)



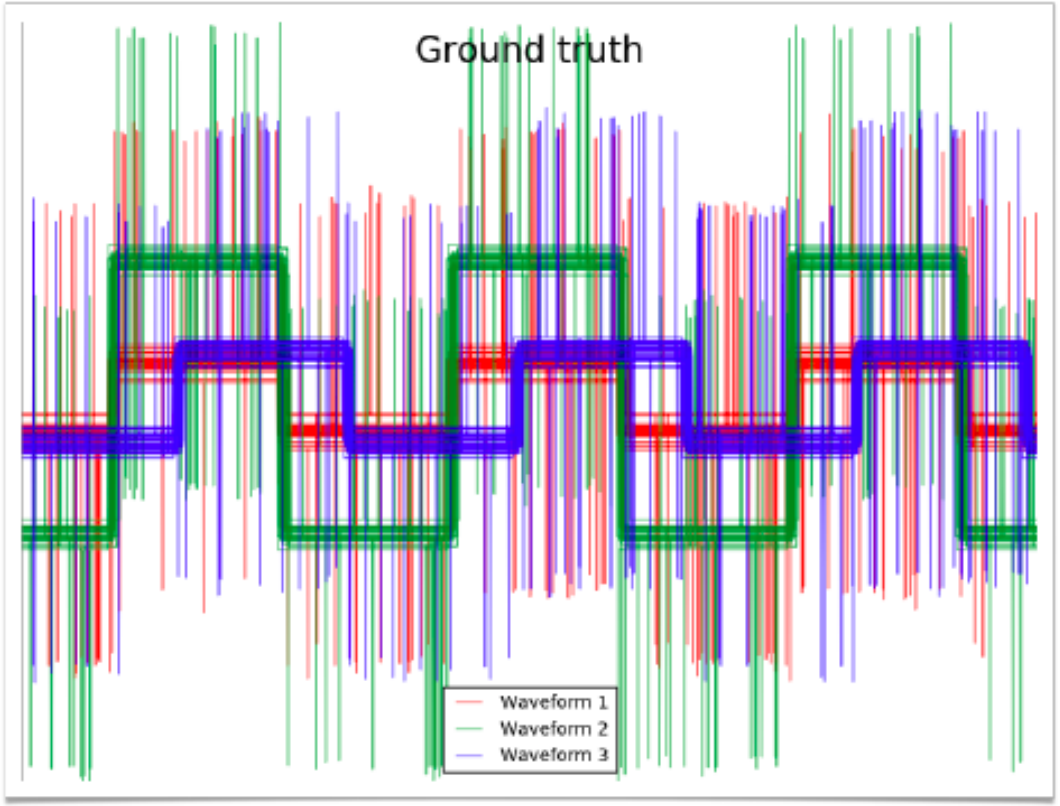
manhattan (l1)



cosine

IT-Academy

РАССТОЯНИЕ МЕЖДУ ОБЪЕКТАМИ



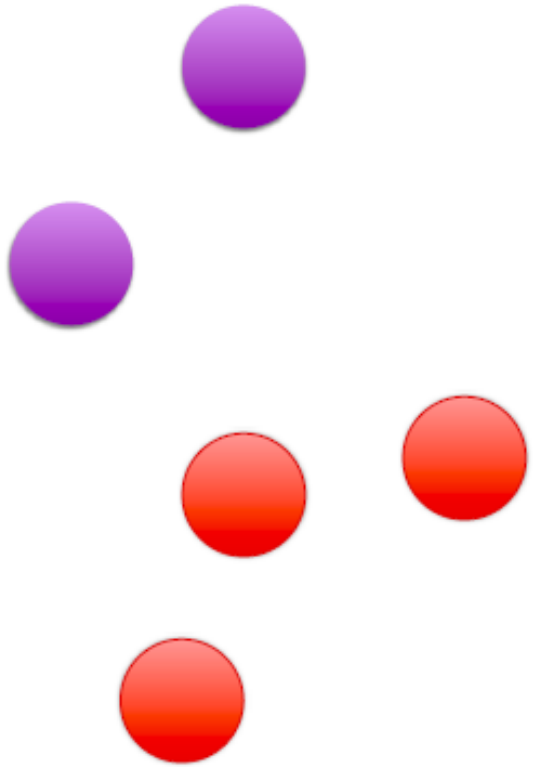
Межкластерное расстояние основывается на расстоянии между объектами. Если с межкластерным расстоянием есть рекомендация брать Уорда, то выбор функции расстояния между объектами более зависит от данных. Слева представлен пример оригинальных данных 3 сигналов, с которыми не справляется косинусное и евклидово расстояние, однако справляется расстояние городских кварталов (l1)

- * sklearn, clustering example

IT-Academy

РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

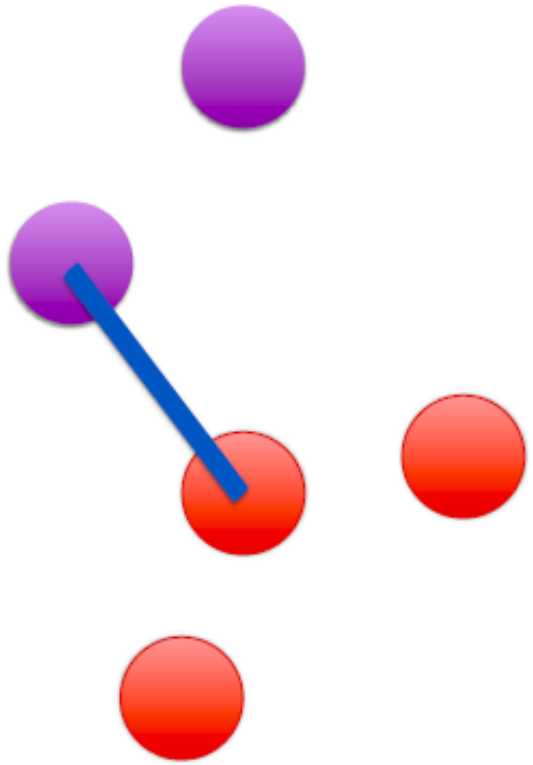
* Ближнего соседа
* Дальнего соседа
* Групповое среднее
* Расстояние между центрами
* Расстояние Уорда



IT-Academy

РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

* **Ближнего соседа**
* Дальнего соседа
* Групповое среднее
* Расстояние между центрами
* Расстояние Уорда



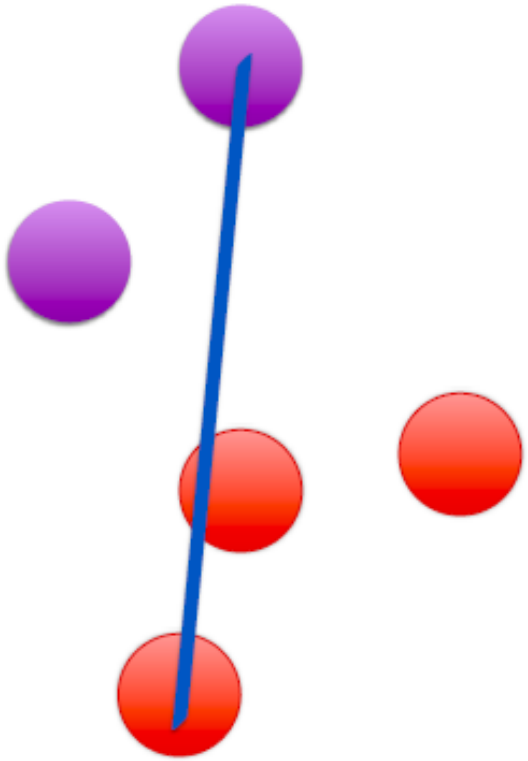
$$R^{б}(W,S) = \min_{w,s} \rho(w,s)$$

IT-Academy

РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

* Ближнего соседа
* **Дальнего соседа**
* Групповое среднее
* Расстояние между центрами
* Расстояние Уорда

$$R^{Д}(W,S) = \max_{w,s} \rho(w,s)$$

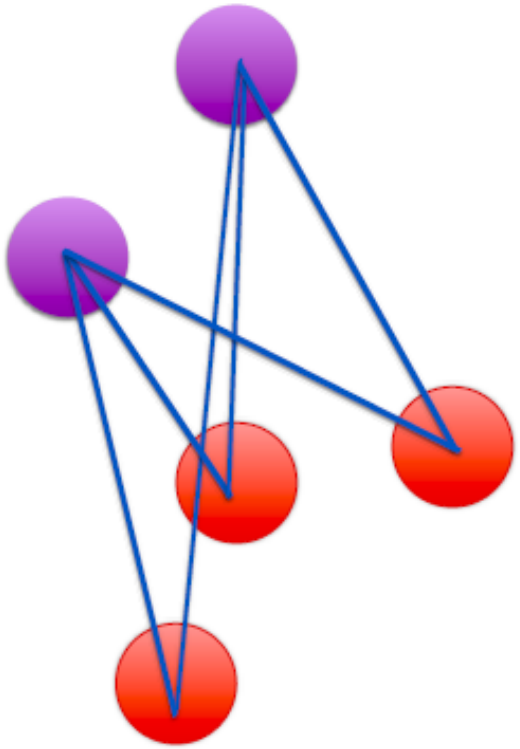


IT-Academy

РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

* Ближнего соседа
* Дальнего соседа
* **Групповое среднее**
* Расстояние между центрами
* Расстояние Уорда

$$R^{\Gamma}(W,S) = \frac{1}{|W| * |S|} \sum_{w} \sum_{s} \rho(w,s)$$

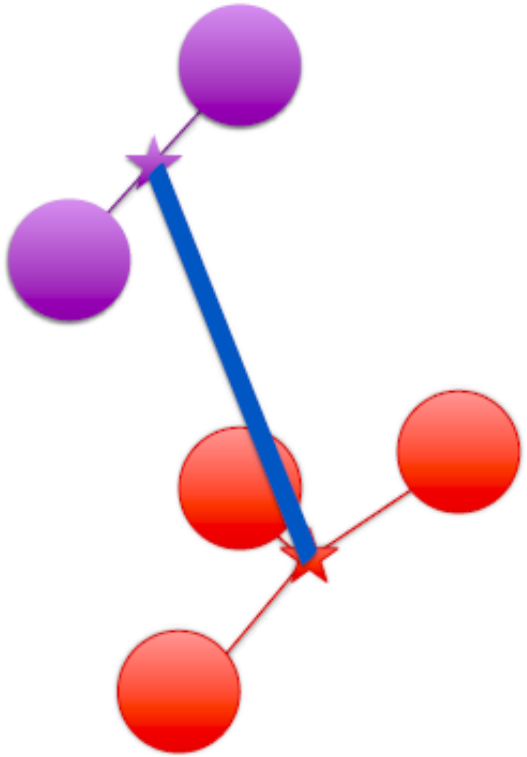


IT-Academy

РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

* Ближнего соседа
* Дальнего соседа
* Групповое среднее
* **Расстояние между центрами**
* Расстояние Уорда

$$R^{\text{ц}}(W,S) = \rho^2\left(\sum_w \frac{w}{|W|}, \sum_s \frac{s}{|S|}\right)$$

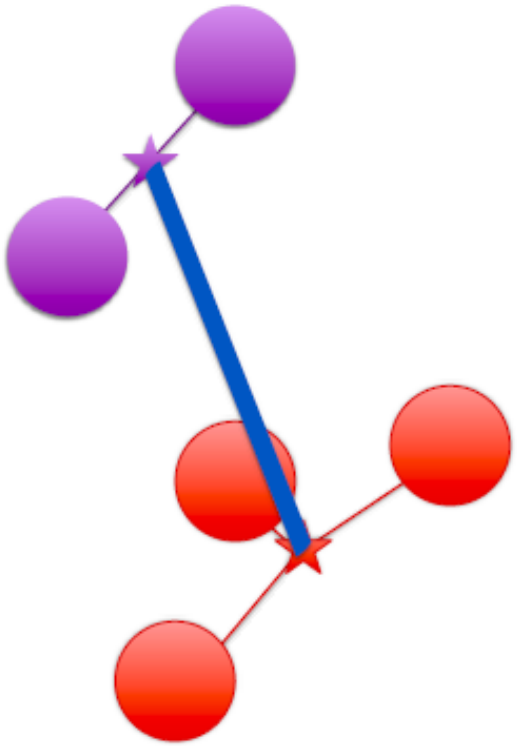


IT-Academy

РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

* Ближнего соседа
* Дальнего соседа
* Групповое среднее
* Расстояние между центрами
* **Расстояние Уорда**

$$R^{y}(W,S) = \frac{|W| * |S|}{|W| + |S|}\rho^{2}\left(\sum_{w}\frac{w}{|W|}, \sum_{s}\frac{s}{|S|}\right)$$



IT-Academy

СВОЙСТА РАССТОЯНИЙ

Расстояние **монотонно, если** при каждом слиянии расстояние между кластерами растёт: $R_2<=R_3<=R_4$..

Расстояние между центрами - не монотонно. Остальные - да

Расстояние **растягивающее,** если при каждом слиянии увеличение расстояний между кластерами растёт:
$R_3-R_2 <= R_4-R_3 <= R_5-R_4$..

Расстояние дальнего соседа и Уорда - растягивающие

IT-Academy

РЕКОМЕНДУЕМОЕ РАССТОЯНИЕ

Расстояние Уорда (Ward)

Оно:

* монотонное
* растягивающее
* работает с центрами кластеров

IT-Academy

РЕАЛИЗАЦИЯ В SKLEARN

AgglomerativeClustering

* n_clusters=2
* affinity='euclidean'
* memory=Memory(cachedir=None)
* connectivity=None
* compute_full_tree='auto'
* linkage='ward'
* pooling_func=<function mean>

Основные параметры

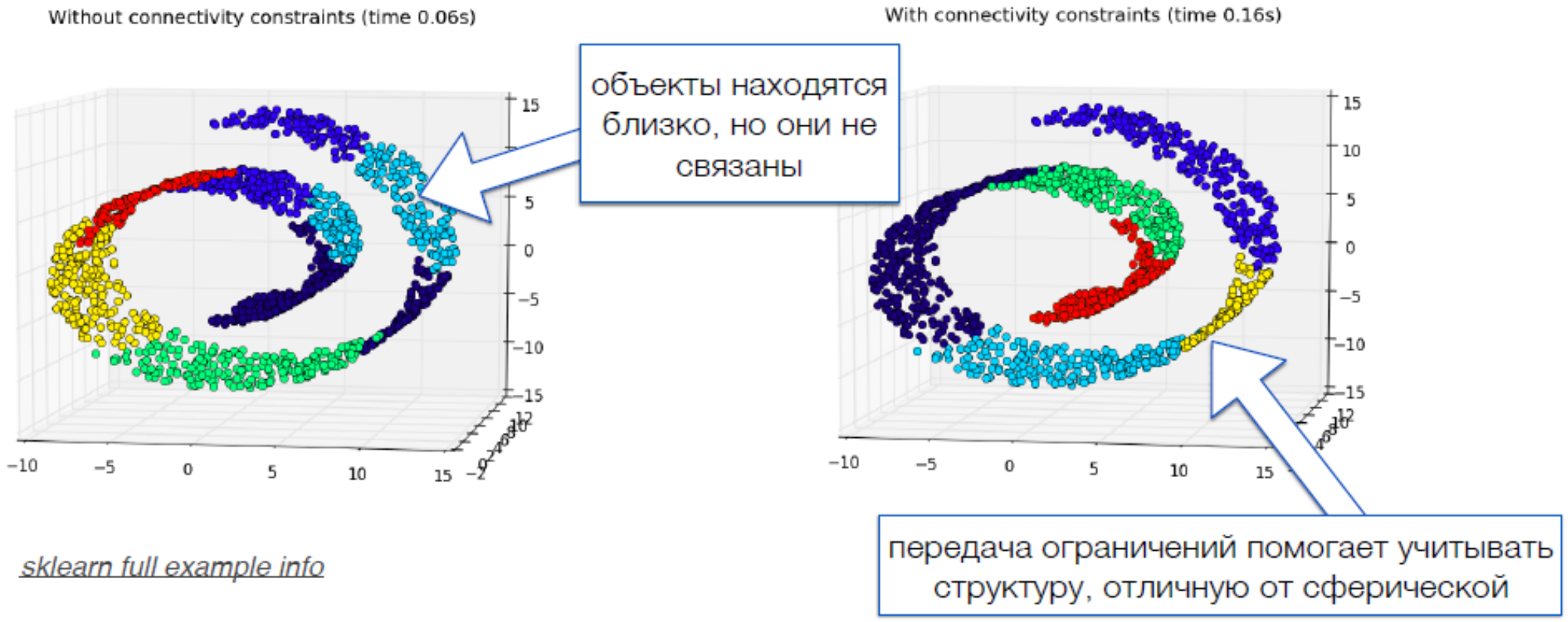
* n_clusters - количество кластеров для разбиения
* linkage: «ward», «complete», «average»
* affinity: "euclidean", "l1", "l2", "manhattan", "cosine", «precomputed» (для linkage = «ward» доступно только евклидово)
* connectivity - априорные знания о структуре данных, подробнее на следующем слайде

Основные методы

* fit, fit_predict

IT-Academy

РЕАЛИЗАЦИЯ В SKLEARN. CONNECTIVITY



sklearn full example info

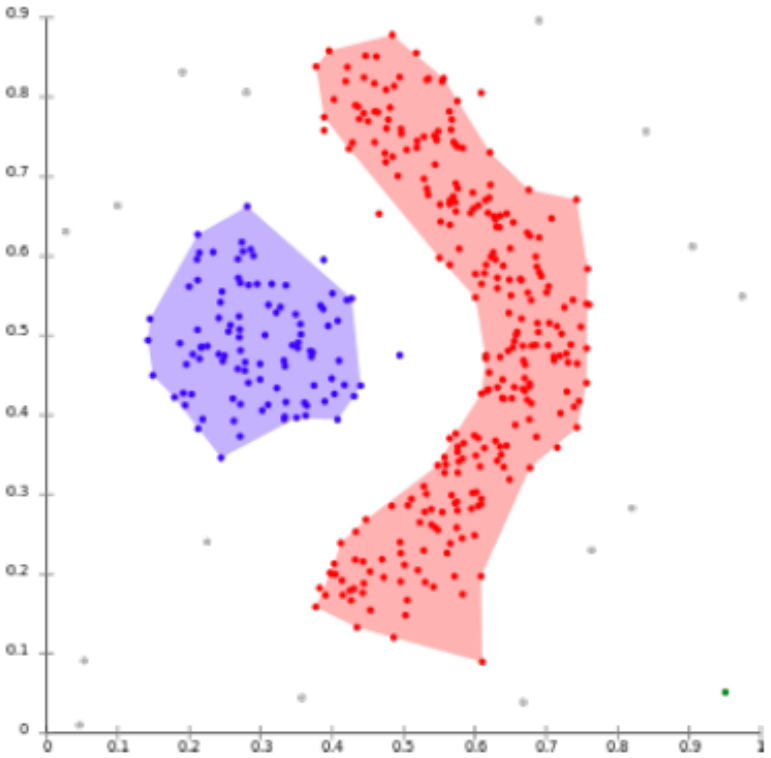
IT-Academy

DBSCAN

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. DBSCAN

IT-Academy

ИДЕЯ *Density-Based Spatial Clustering of Applications with Noise*



Рассматриваем объекты как ядра, вокруг которых собираются другие объекты

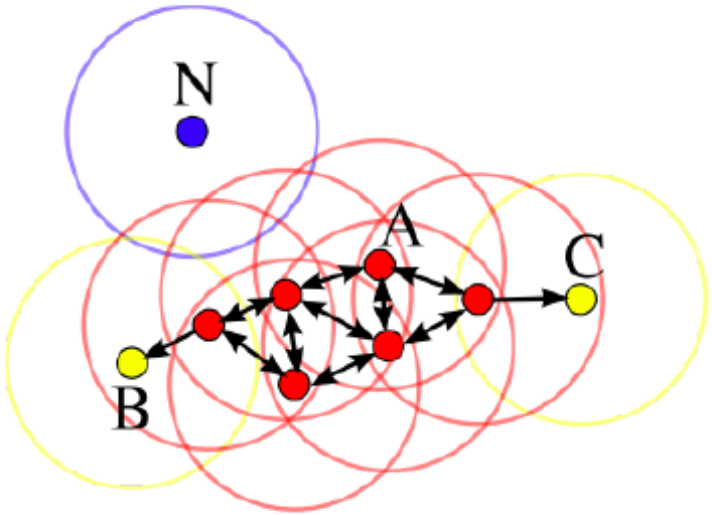
Если не собираются - это выброс

Если ядра связаны - то они и достижимые из них объекты образуют кластер

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. DBSCAN

IT-Academy

ИДЕЯ *Density-Based Spatial Clustering of Applications with Noise*



Все точки делятся на 3 типа:

* ядра
  (в eps-окрестности >= N точек)
* достижимые из ядра
  (в eps-окрестности < N точек, > 0 ядер)
* выбросы
  (остальные)

Ядра и достижимые из них точки образуют кластеры

Выбросы не принадлежат ни одному кластеру

IT-Academy

РЕАЛИЗАЦИЯ В SKLEARN

DBSCAN

* eps=0.5
* min_samples=5
* metric='euclidean'
* algorithm='auto'
* leaf_size=30
* p=None
* n_jobs=1

Основные параметры

* eps - размер окрестности
* min_samples - кол-во точек в окрестности ядра
* n_jobs - кол-во процессоров для расчёта (-1 - max)

Основные методы

* fit, fit_predict

IT-Academy

ДОСТОИНСТВА И НЕДОСТАТКИ

Достоинства:

* не нужно указывать кол-во кластеров
* произвольная форма данных
* обнаруживает выбросы

Недостатки:

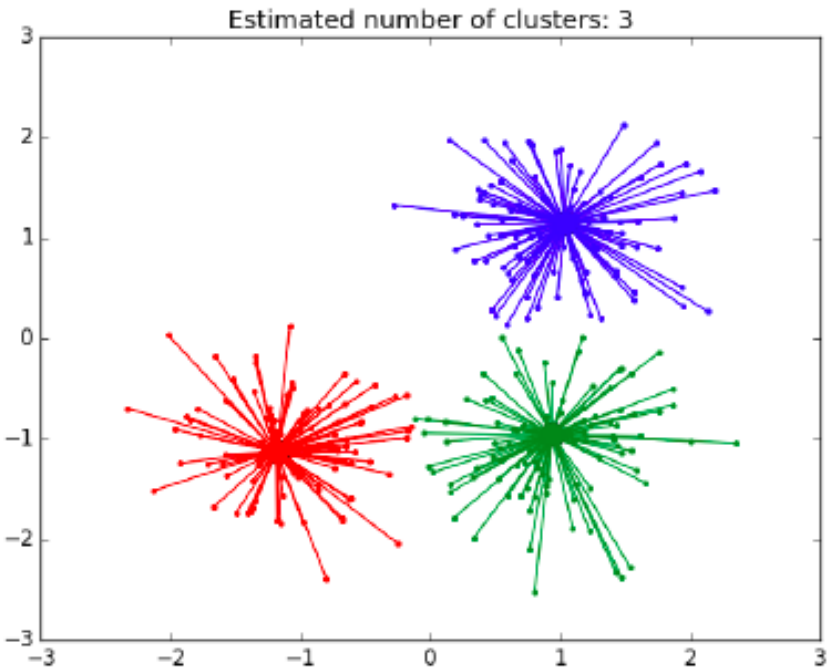
* сложность выбора eps
* плохо работает с кластерами разной плотности

IT-Academy

AFFINITY PROPAGATION

IT-Academy

ИДЕЯ



Объекты обмениваются двумя видами сообщений:

* насколько объект 1 готов быть экземпляром объекта 2
* насколько объект 2 готов предоставить право быть объекту 1 своим экземпляром

Итог:

K объектов - представителей кластеров

IT-Academy

РЕАЛИЗАЦИЯ В SKLEARN

Affinity Propogation

* damping=0.5
* max_iter=200
* convergence_iter=15
* copy=True
* preference=None
* affinity='euclidean'
* verbose=False

Основные параметры

* preference - априорные знания о возможности быть экземпляром
* damping - скорость затухания [0.5-1]
* convergence_iter - условие останова, сколько должно пройти итераций без изменений

Основные методы

* fit, fit_predict

IT-Academy

КАКОЙ АЛГОРИТМ ВЫБРАТЬ?

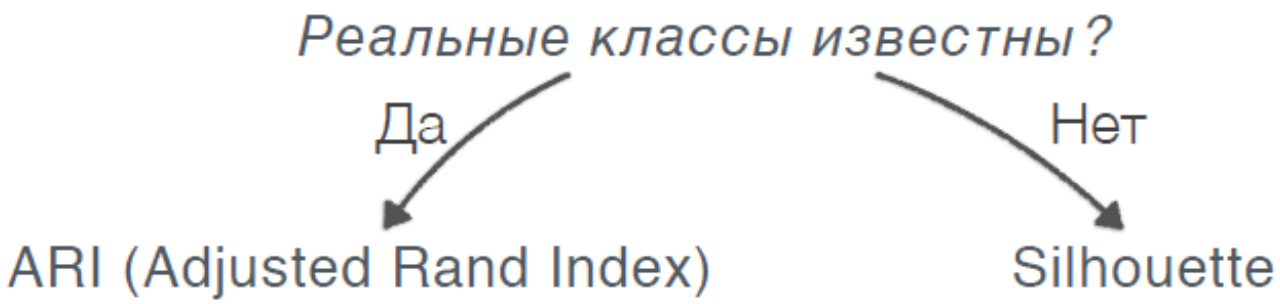
Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points

** sklearn, сравнение кластеризаторов*

МЕТРИКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ

IT-Academy

МЕТРИКИ КАЧЕСТВА



IT-Academy

ARI: ADJUSTED RAND INDEX

Дано:

y_pred - вектор меток кластеризации [0, 0, 0, 1, 1, 1]

y_true - реальные кластеры [2, 2, 2, 7, 7, 7]

ARI $\in$ [-1, 1];

1 - точное соответствие

0 - случайное разбиение кластеров

ARI(y_pred, y_true) = 1

Метрике не важны названия кластеров

IT-Academy

СИЛУЭТ

нет знания правильных кластеров.

Оценим, насколько сильно **один объект** сидит внутри своего кластера и далеко от ближайшего соседнего:

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

a - среднее расстояние до объектов внутри кластера

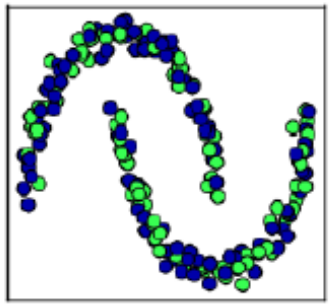
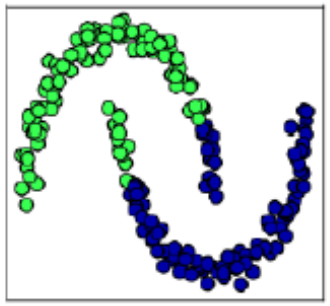
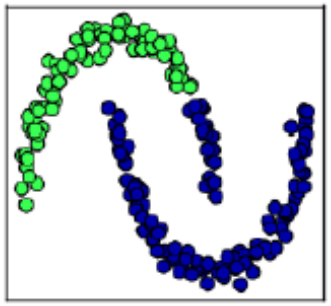
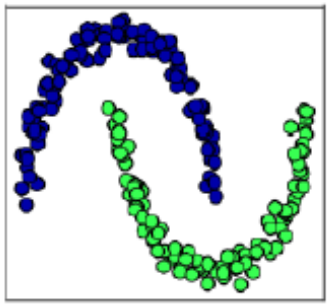
b - среднее расстояние до объектов ближайшего кластера

$$s = mean(s)$$

среднее значение по всем объектам - силуэт

IT-Academy

СРАВНЕНИЕ МЕТРИК

ARI	0.00	0.50	0.61	1.00
Silhouette	0.00	0.49	0.46	0.38

IT-Academy

ПРАКТИКА

clustering.ipynb

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

IT-Academy

1. Кластеризация позволяет **находить структуру** в незамеченных данных, что может послужить **дополнительными признаками** обучения или являться **самодостаточной целью**
2. В задаче кластеризации **нет правильного решения**. Метрики качества служат лишь слабым приближением для создания новых алгоритмов или нахождениям критерия останова
3. Разные алгоритмы кластеризации принципиально **работают по-разному**, для конкретного набора данных необходимо выбирать наиболее подходящий

IT-Academy

Кластеризация

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ