GeekBrains

Кластеризация и решение задачи группировки данных в машинном обучении

Урок 10

На этой лекции вы найдете ответы на такие вопросы как:

-Что такое кластеризация
- Понятие расстояния и меры сходства
- Алгоритмы кластеризации
- Методы оценки качества кластеризации

Булгакова Татьяна

Преподаватель в GeekBrains, Нетология,Skillfactory

С 2010 года занимаюсь DataScience и NN. Фрилансер

- Участвовала в разработке программы по настройке оборудования для исследования пространственного слуха китообразных НИИ ИПЭЭ РАН
- Участвую в разработке рекомендательных систем по настройке нейростимуляторов для медицинских центров
- Работаю над курсом по нейронным сетям

План курса

1. Первичный и визуальный анализ данных
2. Описательные статистики в контексте EDA. Корреляция и корреляционный анализ
3. Регрессия и использование библиотеки Scikit-learn в задачах обучения с учителем
4. Классификация и использование логистической регрессии в задачах классификации
5. Функционалы ошибки и поиск оптимальных параметров в задачах машинного обучения
6. Проблема переобучения и недообучения модели. Кросс-валидация и регуляризация.
7. Ансамблирование и использование деревьев решений в задачах машинного обучения
8. Генерация признаков. Методы отбора признаков. Подбор гиперпараметров.
9. Обучение без учителя. Понижение размерности. Алгоритмы понижения размерности
10. Кластеризация и решение задачи группировки данных в машинном обучении

Что будет на уроке сегодня

? Что такое кластеризация

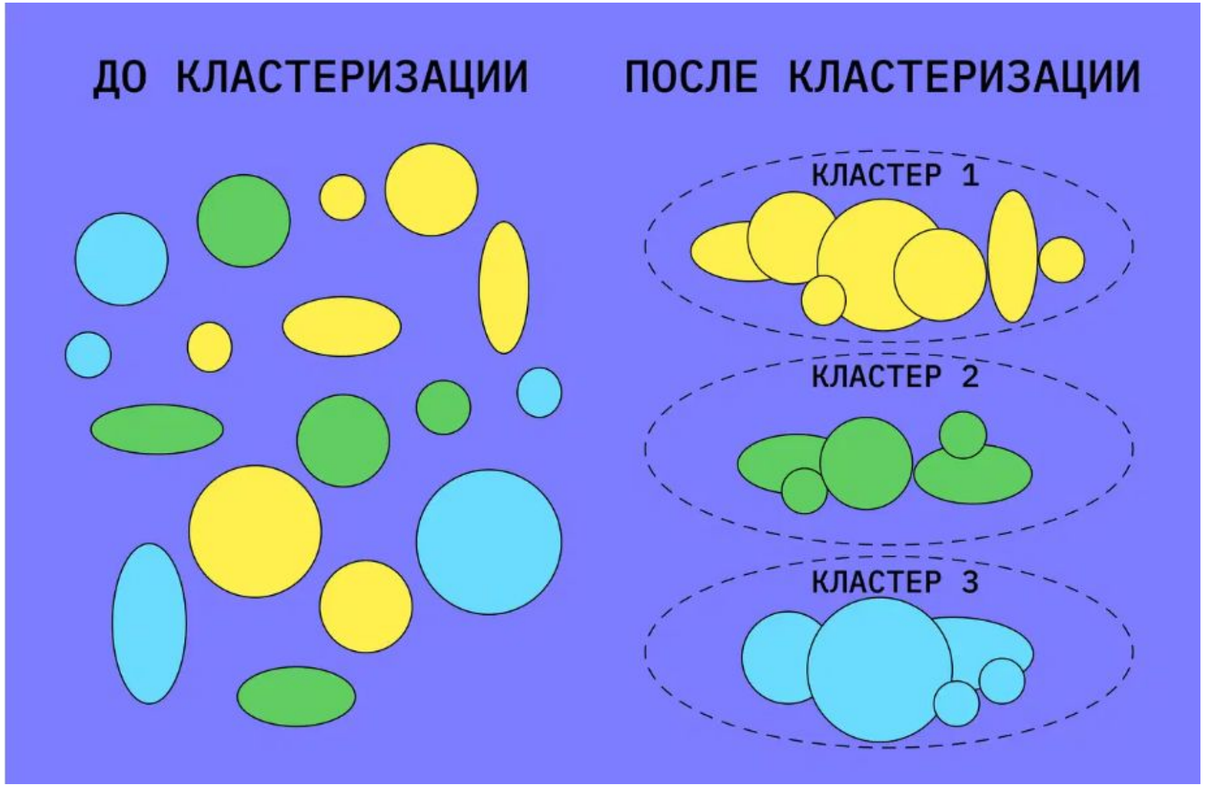
? Понятие расстояния и меры сходства

? Алгоритмы кластеризации

? Методы оценки качества кластеризации

Кластеризация.

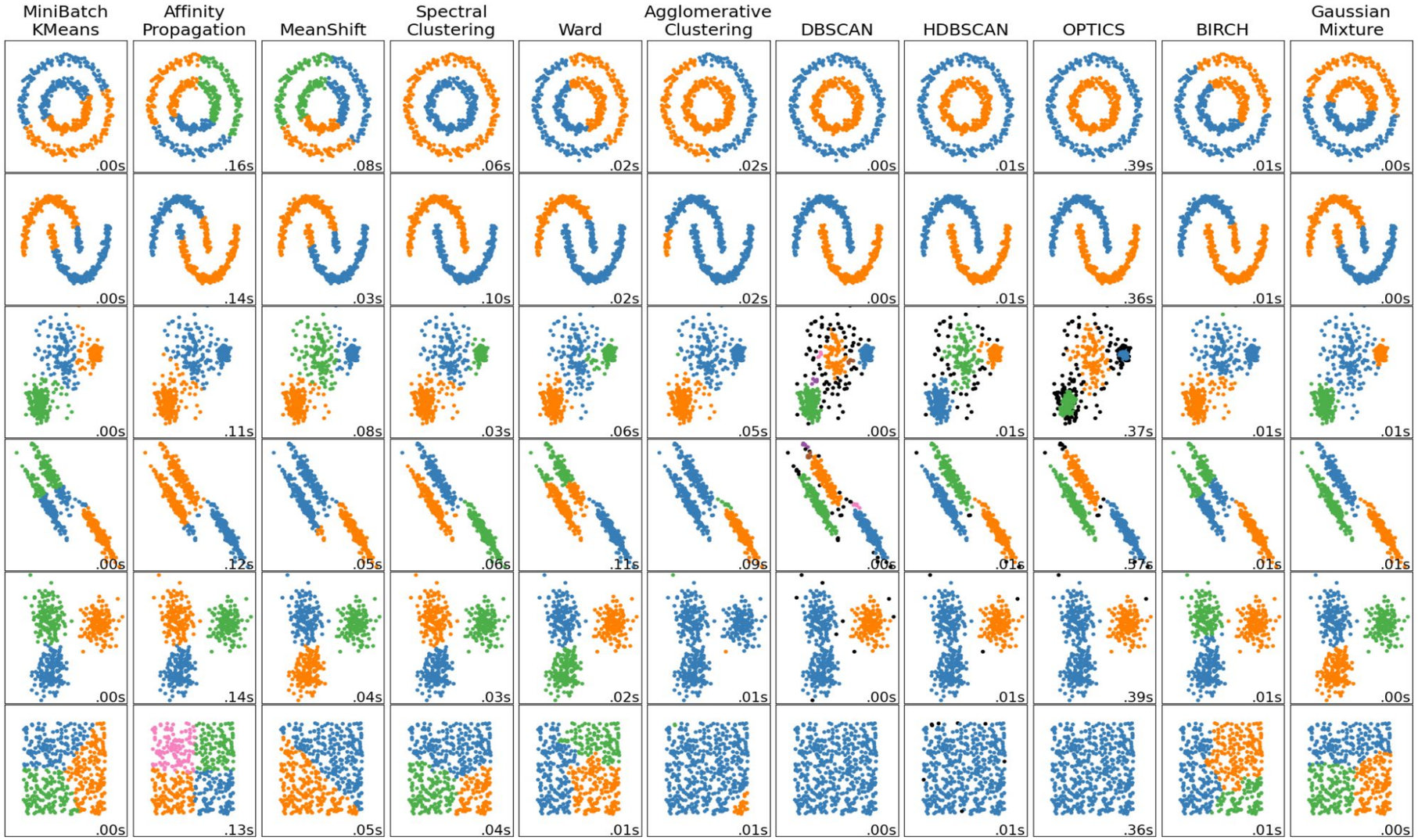
Кластеризация – это процесс разделения набора данных на группы или кластеры, состоящие из схожих объектов.



Кластеризация.

Кластерный анализ данных предоставляет возможность проводить кластеризацию не только один раз, а множество раз.

Кластеризация.



Кластеризация

Сферы применения кластерного подхода включают:

- 👍 Анализ поведения клиентов
- 👍 Исследование рынка
- 👍 Анализ мнений и предпочтений
- 👍 Формирование тематик страниц сайта

Кластеризация

Кластерный подход имеет широкий спектр применения и может быть полезен во многих сферах.

👍 Гораздо больше сфер применения открывается при использовании кластеризации для обработки различных файлов разных форматов.

👍 Кластеризацию можно успешно применять не только к текстовым данным, но и к изображениям, аудиофайлам и видеофайлам.

👍 Удобство обработки собранных файлов разных форматов становится особенно важным при работе с огромными объемами информации.

Понятие расстояния и меры сходства.

Меры расстояния используются для измерения расстояния между двумя объектами или точками данных

Евклидово расстояние:

$$\rho\left(x, x^{'}\right) = \sqrt{\sum_{i}^{n}\left(x_i - x_i^{'}\right)^2}$$

x = (1, 2, 3) y = (4, 6, 5)

$$\rho(x, y) = \sqrt{(1-4)^2 + (2-6)^2 + (3-5)^2} = \sqrt{9 + 16 + 4} = \sqrt{29} = 5.29$$

Понятие расстояния и меры сходства.

Меры расстояния используются для измерения расстояния между двумя объектами или точками данных

Квадрата евклидова расстояния

$$\rho\left(x, x^{'}\right) = \sum_{i}^{n}\left(x_i - x_i^{'}\right)^2$$

В отличие от евклидова расстояния, здесь опускается операция извлечения квадратного корня.

Понятие расстояния и меры сходства.

Меры расстояния используются для измерения расстояния между двумя объектами или точками данных

Манхэттенское расстояние

$$\rho\left(x, x^{'}\right) = \sum_{i}^{n} \left|x_i - x_i^{'}\right|$$

В отличие от евклидова расстояния, здесь не вычисляется квадрат разности координат, а просто суммируются модули разностей по каждой координате.

Понятие расстояния и меры сходства.

Меры расстояния используются для измерения расстояния между двумя объектами или точками данных

Расстояние Чебышева

$$\rho\left(x, x^{'}\right) = max(\left|x_i - x_i^{'}\right|)$$

x = (1, 3, 2) y = (2, 6, 4)

p(x, y) = max(|1 - 2|, |3 - 6|, |2 - 4|) = max(1, 3, 2) = 3

Понятие расстояния и меры сходства.

Меры сходства наоборот измеряют степень схожести или близости между объектами.

Косинусное расстояние

$$\Delta cos(x, y) = 1 - cos(x, y)$$

x = (3, 1, 0, 1, 2) y = (1, 0, 1, 1, 1)

$$cos(x, y) = \frac{(31 + 1 + 1 + 11 + 2 * 1)}{\left(\sqrt{3^2 + 2^1 + 0^2 + 1^2 + 2^2}\right) * \sqrt{(1^2 + 0^2 + 1^2 + 1^2 + 1^2)}} = 0.57$$

dcos(x, y) = 1 - 0.57 = 0.43

Понятие расстояния и меры сходства.

Меры сходства наоборот измеряют степень схожести или близости между объектами.

Расстояние Хэмминга

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

1010001
1000101
d=2

25687914
24657934
d=3

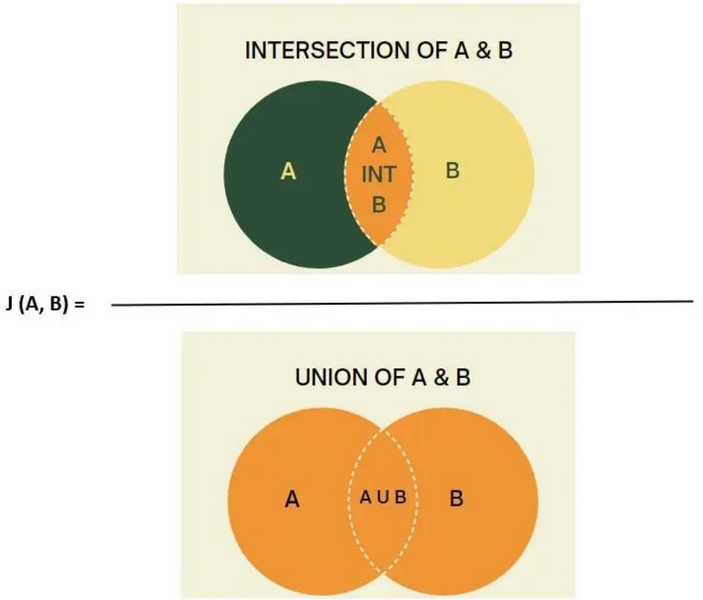
пакет
багет
d=2

Понятие расстояния и меры сходства.

Меры сходства наоборот измеряют степень схожести или близости между объектами.

Метрика Жаккара

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Например, пусть:

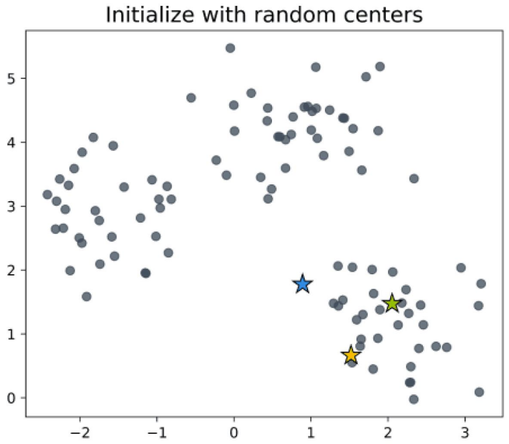
A = {1, 2, 3} B = {2, 3, 4}

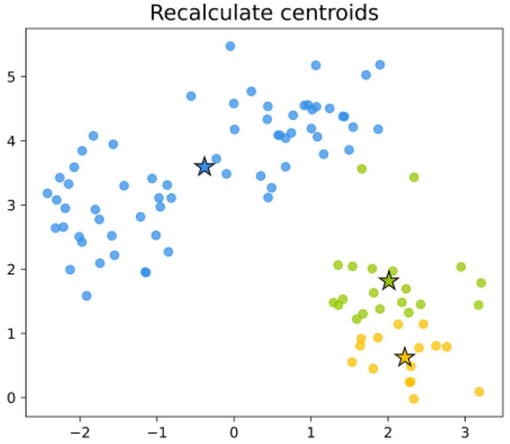
Тогда:

|A ∩ B| = 2 (общие элементы 2 и 3) |A ∪ B| = 4 (все различные элементы 1, 2, 3, 4)

J(A, B) = 2/4 = 0.5

Кластеризация на основе центроидов





Минимизировать внутриклассовые отличия от центроида:

Кол-во кластеров

Кол-во наблюдений

i-ое наблюдение

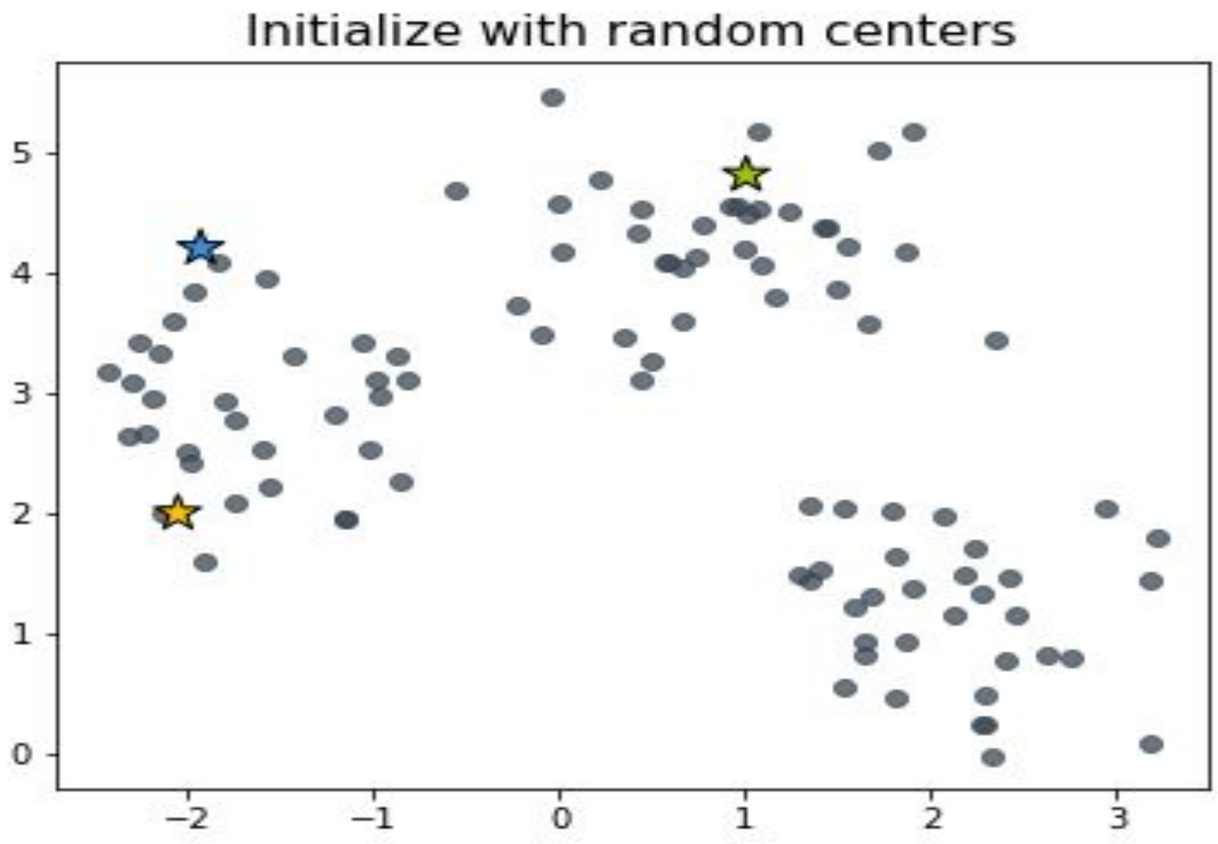
центроид j-ого кластера

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \min(||x_i^{(j)} - c_j||)^2$$

Функция расстояния

Функция потерь
(еще говорят целевая функция, objective function)

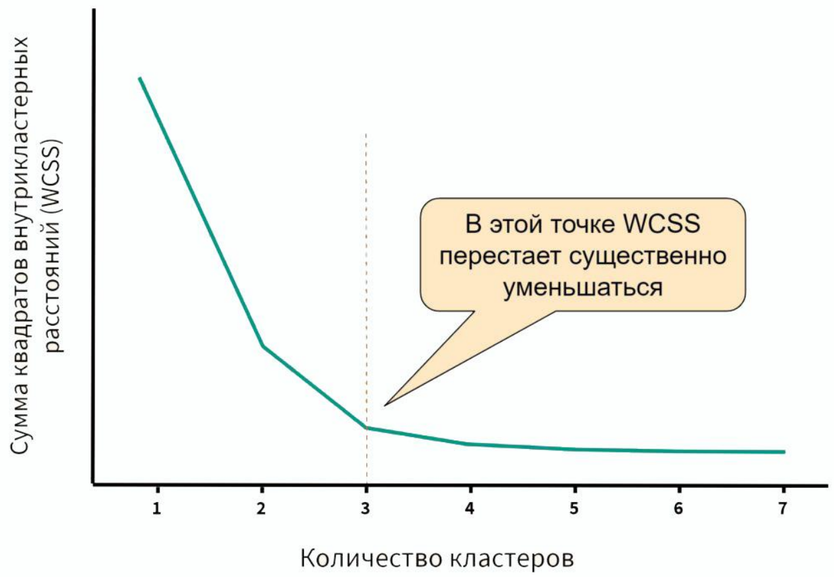
Кластеризация на основе центроидов



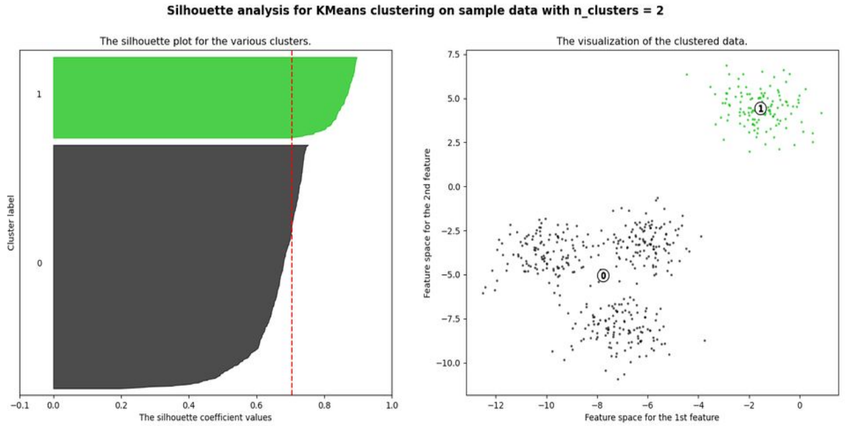
Количество кластеров
Метод локтя

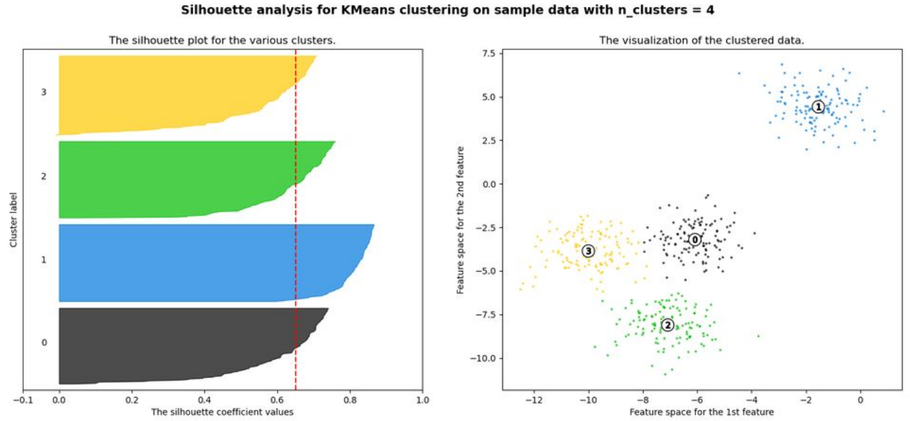
Идея:
перебирать от 1 до N кластеров, засечь, с какого момента качество перестанет быстро улучшаться

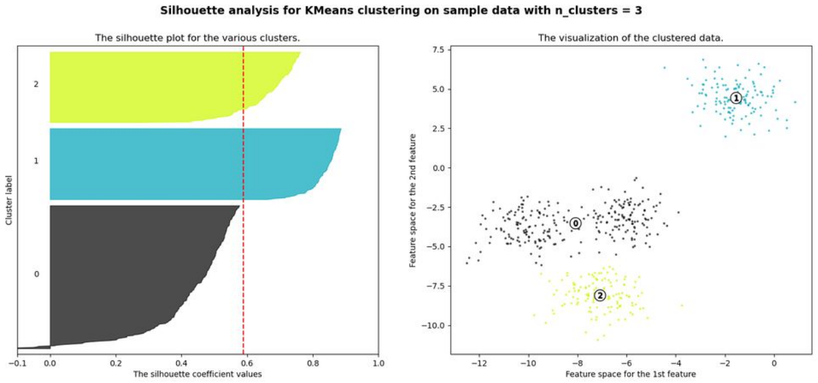
Качество - сумма квадратов расстояний от точек до центроидов кластеров

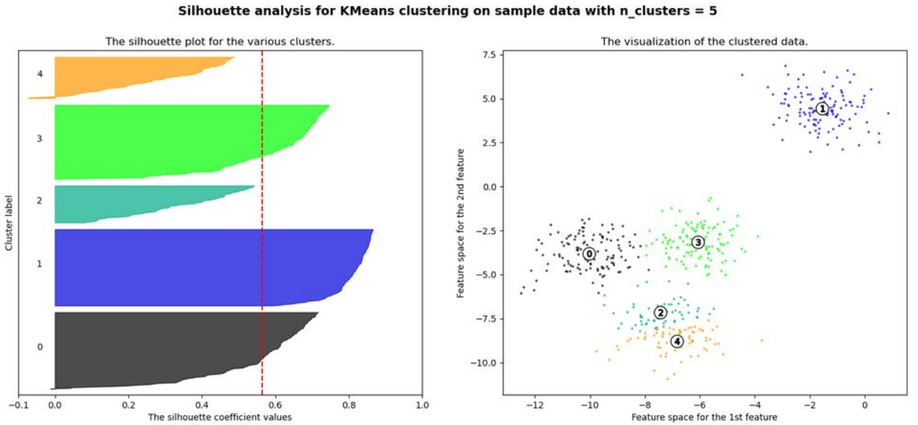


Количество кластеров
Метод силуэта

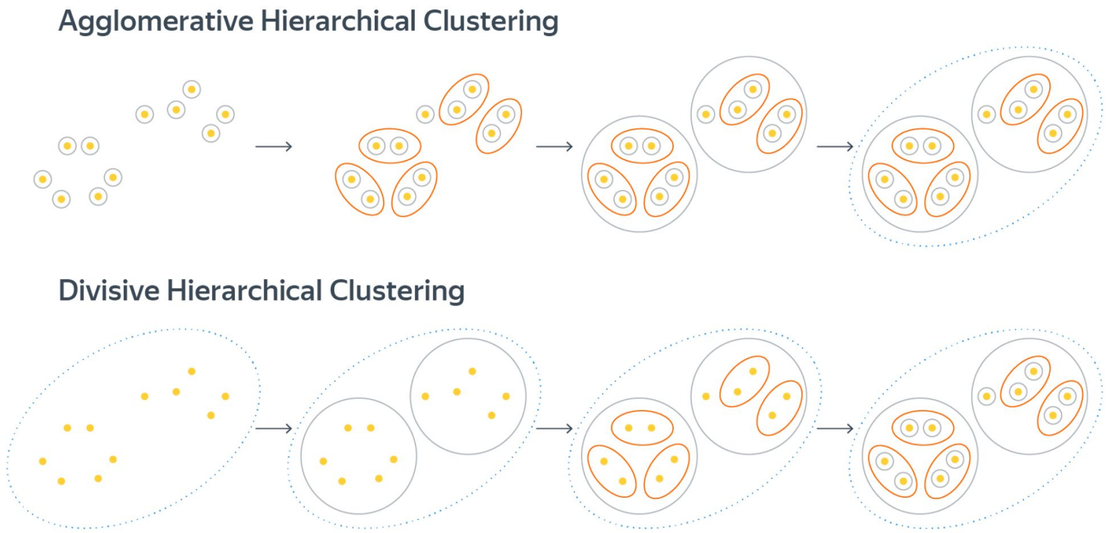








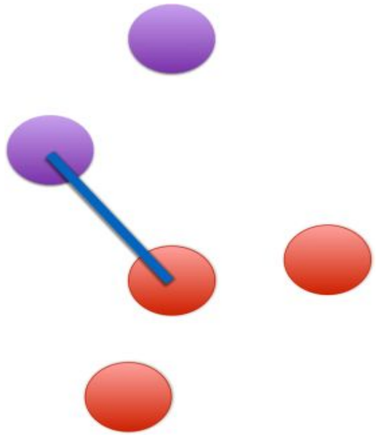
Иерархическая кластеризация



Метрики объединения кластеров

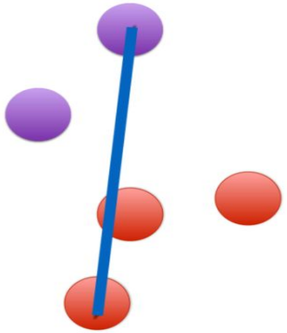
Одиночная связь

$$R^{б}(W,S) = \min_{w,s} \rho(w,s)$$



Полная связь

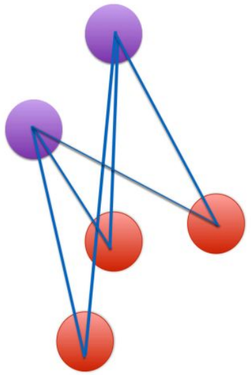
$$R^{д}(W,S) = \max_{w,s} \rho(w,s)$$



Метрики объединения кластеров

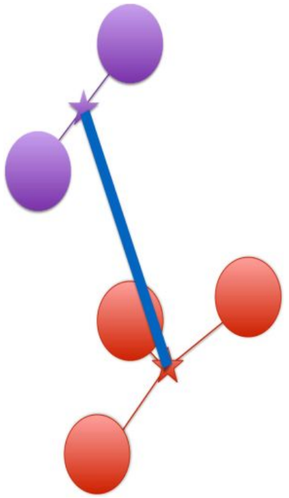
Невзвешенное
попарное среднее

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



Невзвешенный
центроидный метод

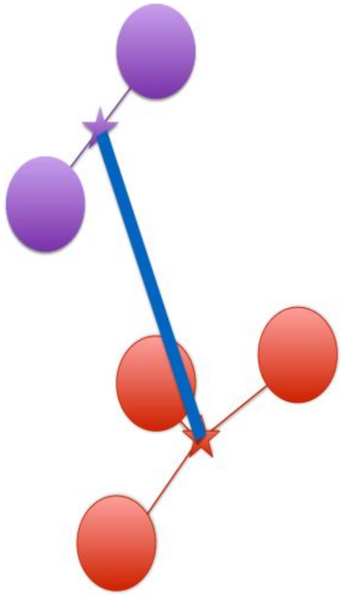
$$R^{\text{ц}}(W,S) = \rho^2\left(\sum_{w} \frac{w}{|W|}, \sum_{s} \frac{s}{|S|}\right)$$



Метрики объединения кластеров

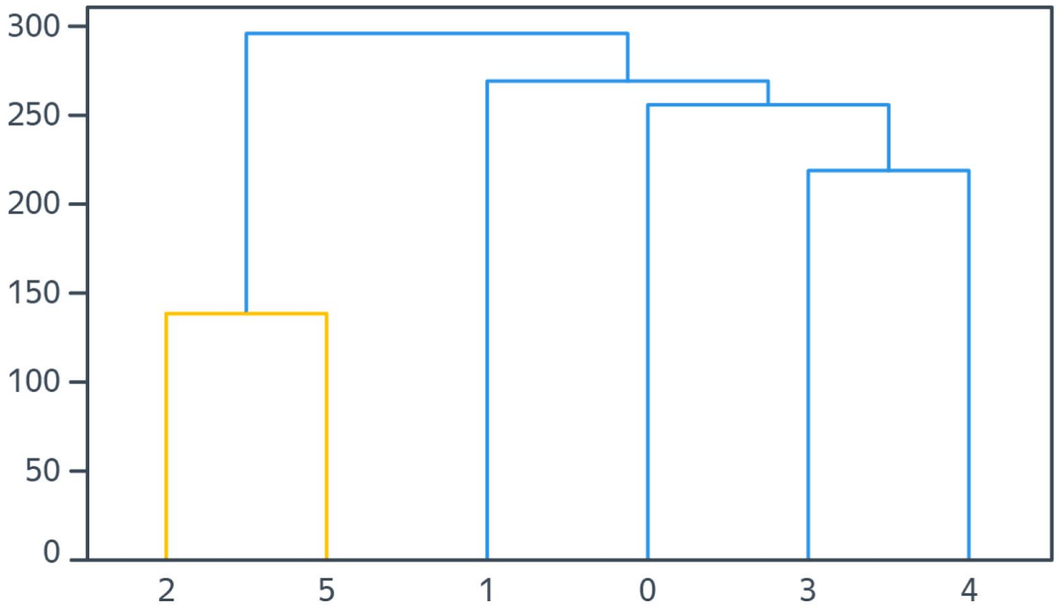
Взвешенный центроидный метод (расстояние Уорда)

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



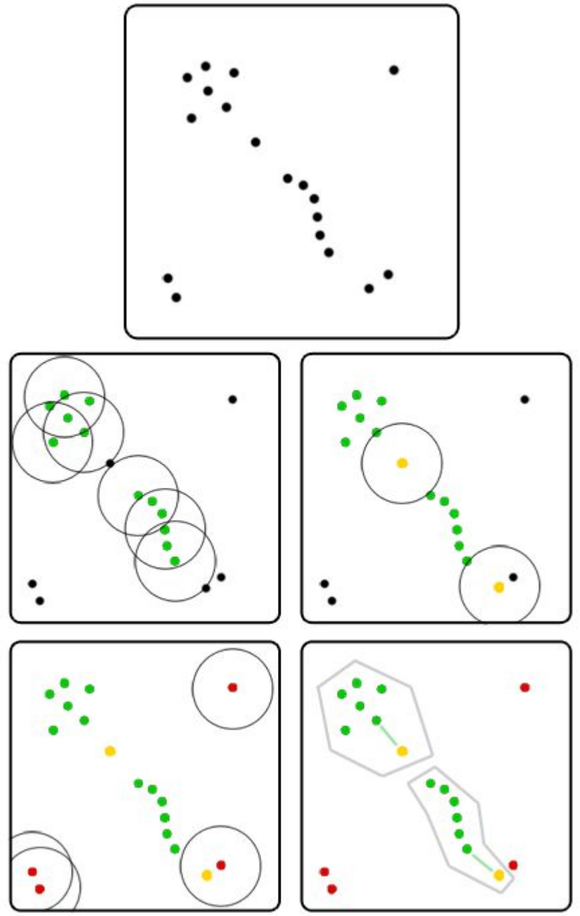
Дендрограмма

Дендрограмма - это схема, которая отображает процесс объединения кластеров в алгоритме кластеризации.

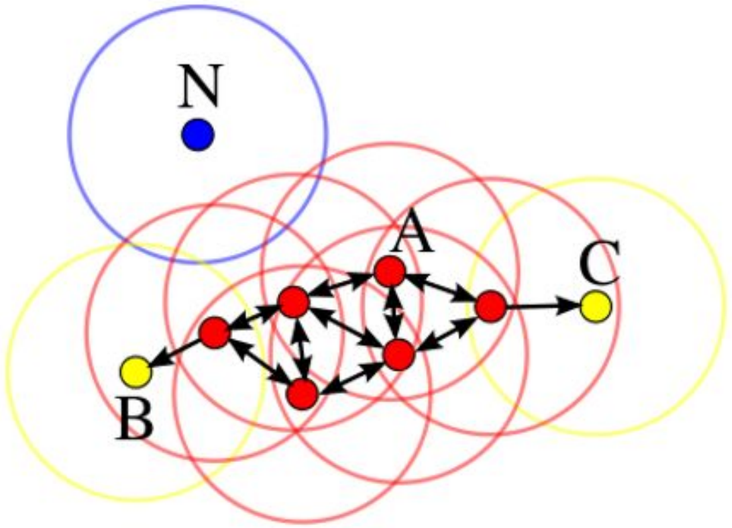


Кластеризация на основе плотности

Основная идея метода заключается в том, чтобы определить области высокой плотности данных и использовать их для определения кластеров



DBSCAN



1. Основные точки A: Идентифицируются по наличию более n объектов в их радиусе.

2. Граничные точки B и C: Хотя рядом есть основные точки, общее число соседей меньше n.

3. Шумовые точки N: Не имеют основных точек поблизости и окружены менее чем n объектами.

Алгоритмы кластеризации на основе плотности

- OPTICS (Ordering Points To Identify the Clustering Structure)

- DENCLUE (DENsity CLUstEring кластеризация на основе плотности)

- HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise пространственная кластеризация приложений с шумом на основе иерархической плотности)

- ST-DBSCAN (Spatial-Temporal Density-Based Clustering Кластеризация приложений на основе пространственно-временной плотности с шумом)

- SUBCLU (Density-Connected Subspace Clustering - кластеризация подпространств)

- VDBSCAN (Varied Density-Based Spatial Clustering of Applications with Noise пространственная кластеризация приложений с шумом на основе вариационной плотности)

Среднее внутрикластерное расстояние (average intra-cluster distance)

Эта метрика позволяет измерить, насколько близко находятся объекты внутри каждого кластера.

$$F_0 = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} \rho(x_i, x_j) \mathbb{I}[a(x_i) = a(x_j)]}{\sum_{i=1}^{n} \sum_{j=i}^{n} \mathbb{I}[a(x_i) = a(x_j)]}$$

Среднее межкластерное расстояние (average inter-cluster distance)

Эта метрика показывает среднее расстояние между центроидами (или средними значениями) каждой пары кластеров.

$$F_1 = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} \rho(x_i, x_j) \mathbb{I}[a(x_i) \neq a(x_j)]}{\sum_{i=1}^{n} \sum_{j=i}^{n} \mathbb{I}[a(x_i) \neq a(x_j)]}$$

Гомогенность (homogeneity)

Она измеряет, насколько хорошо каждый кластер состоит из объектов одного и того же истинного класса.

$$Homogeneity = 1 - \frac{H_{class|clust}}{H_{class}}$$

Полнота

$$Completeness = 1 - \frac{H_{clust|class}}{H_{clust}}$$

V-мера

$$V_\beta = \frac{(1+\beta) \cdot Homogeneity \cdot Completeness}{\beta \cdot Homogeneity + Completeness}$$

Итоги

Тема кластеризации имеет практическую значимость в контексте машинного обучения по нескольким причинам:

- ✅ Позволяет обнаруживать скрытые паттерны и закономерности в данных в обучении без учителя.
- ✅ Используется в предобработке данных для выявления аномалий и шума.
- ✅ Применяется в сжатии данных, поскольку кластеры меньше исходного пространства.
- ✅ Полезна в визуализации и интерпретации больших массивов данных.
- ✅ Помогает определить оптимальное количество классов в задачах классификации.

Спасибо за внимание