KNN, K-means

Корлякова М.О.

2021

Метрические алгоритмы

Метрические алгоритмы - алгоритмы, основанные на вычислении оценок сходства между объектами.

.

Расстояния между объектами

- Метрики : Минковский
- Меры: Хемминг
- И МНОГО ДРУГИХ МЕТОДОВ!!!!

Метрики

- Функции
- Минковского
- Косинусная

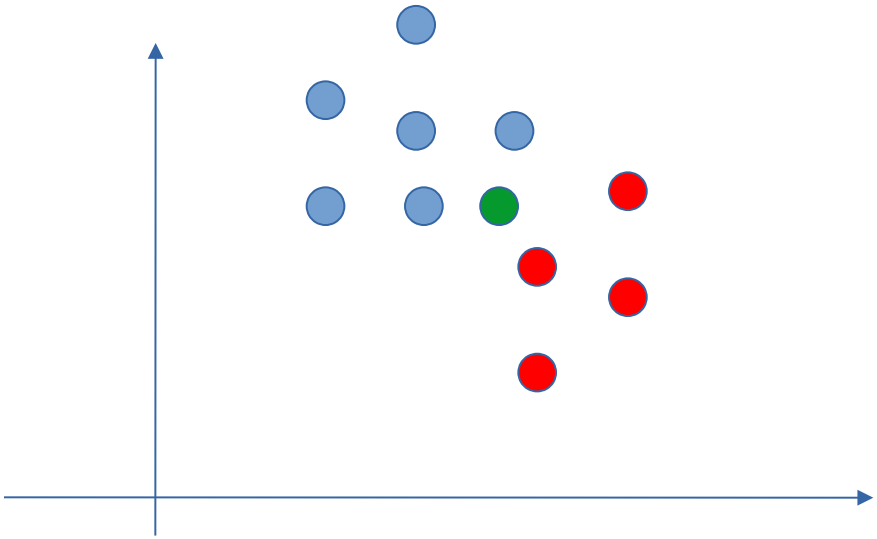
- Цель: Измерить расстояние

Классификация

- T={(Xi,yi)}

Цель: Найти для любого нового X класс y

KNN:



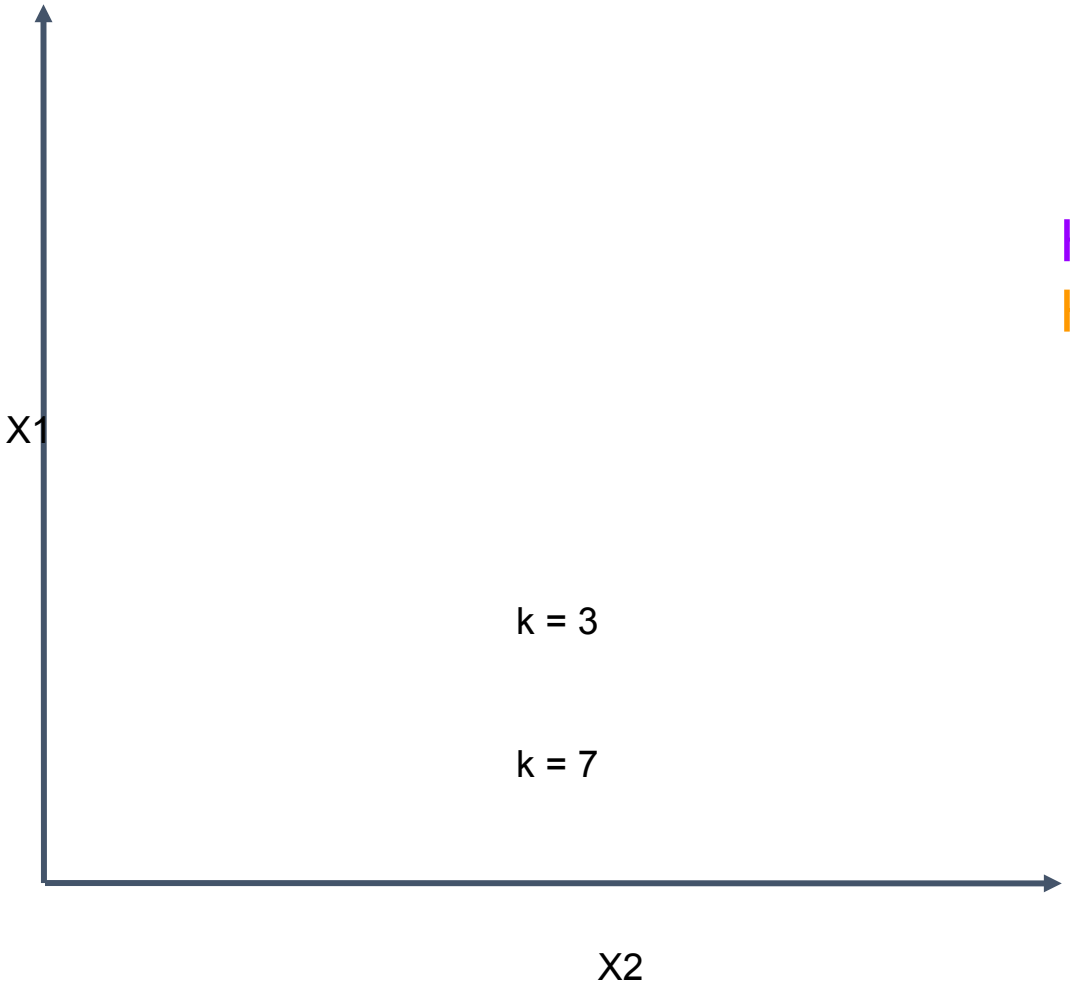
Метрические алгоритмы классификации

Метод k-ближайших соседей

Метрические алгоритмы - алгоритмы, основанные на вычислении оценок сходства между объектами.

Метод k-ближайших соседей - объекту присваивается тот класс, который наиболее распространен среди его k соседей.

Метод k-ближайших соседей



Класс A
Класс B

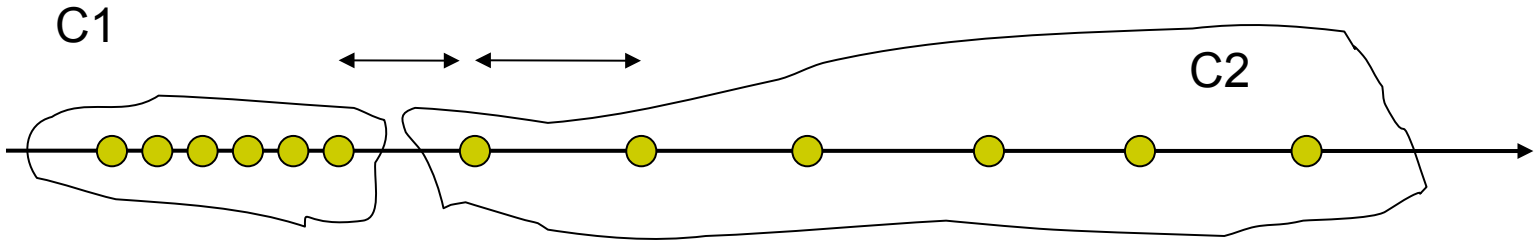
Кластеризация

- T={(Xi)}

- Цель: Найти классы

Гипотеза компактности

- Гипотеза λ-компактности
- Расстояние мало, но есть неоднородность.
- G – полный граф для T={Xi}
- A(a,b) – расстояние от точки a к b – длина ребра
- D=max(A(a,b))



Алгоритм к-средних

- Фиксирует число классов
  1. Номер итерации s=0
  2. Связать с каждым кластером Kj объект Xi из обучающей выборки (случайно).
  3. если число кластеров меньше N, то перейти к процедуре формирования кластеров (п.4.).
  4. Вычислить расстояния от всех объектов до всех центров кластеров.
  5. присоединить объект Xi к кластеру Ck, если $Ck = \min_{j=1..M} d(Xi, Cj)$
  6. повторить для всех объектов выборки.
  7. вычислить новое положение центров кластеров $\mathbf{centr}_b = \frac{1}{N_b} \sum_{i=1}^{N_b} \mathbf{x}_i$
  1. где Nb – число примеров множества Cb.
  2. Повторять от п.4. пока кластер смещается более чем на ε (задано пользователем), иначе остановить процесс.

Достоинства алгоритма k-средних:

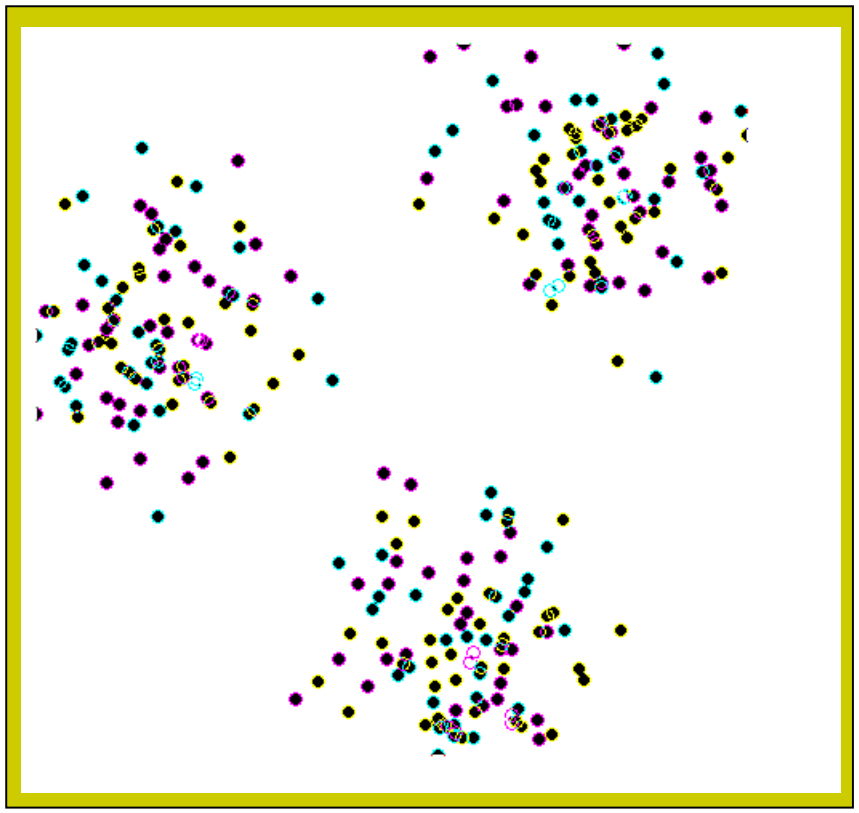
- простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

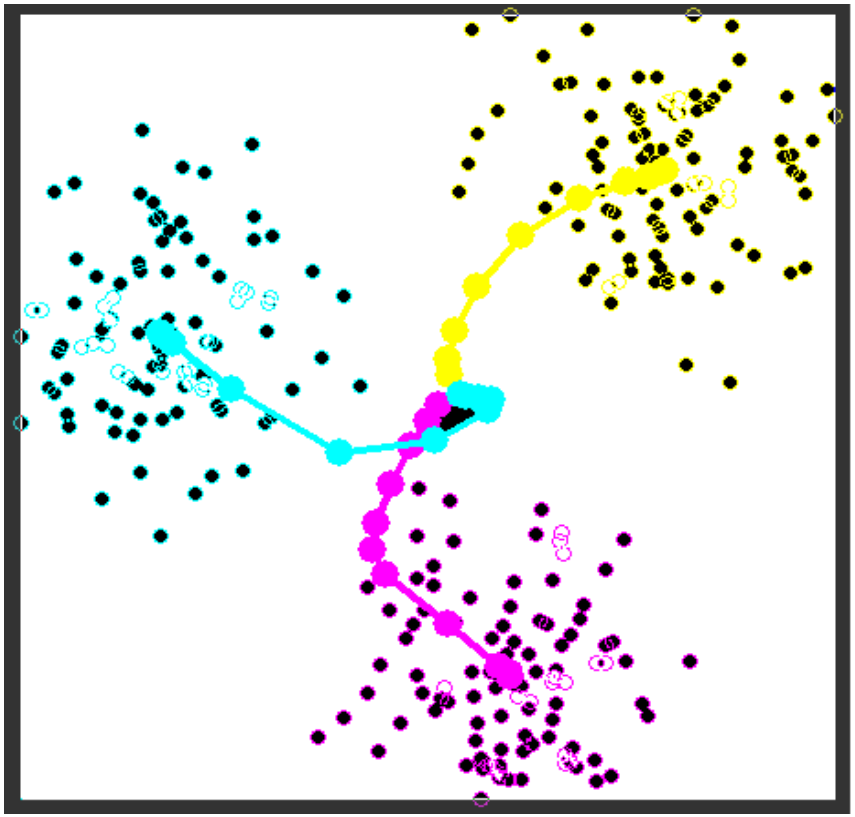
Недостатки алгоритма k-средних:

- Чувствителен к выбросам, которые могут искажать среднее;
- Может медленно работать на больших базах данных.

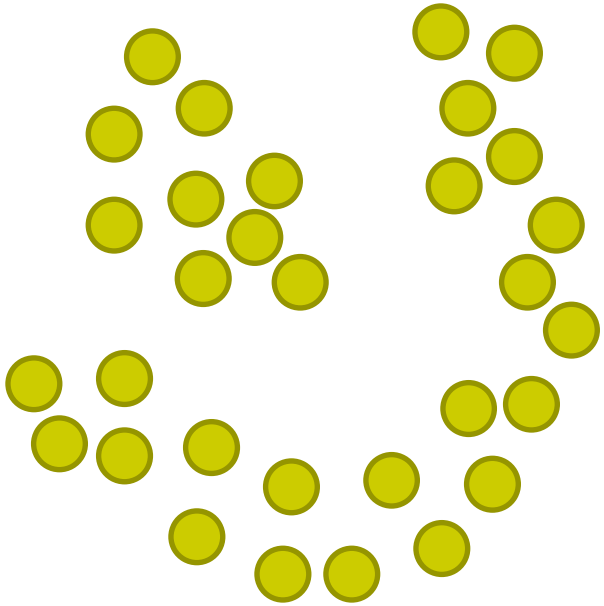
Пример

Исходное множество Результат

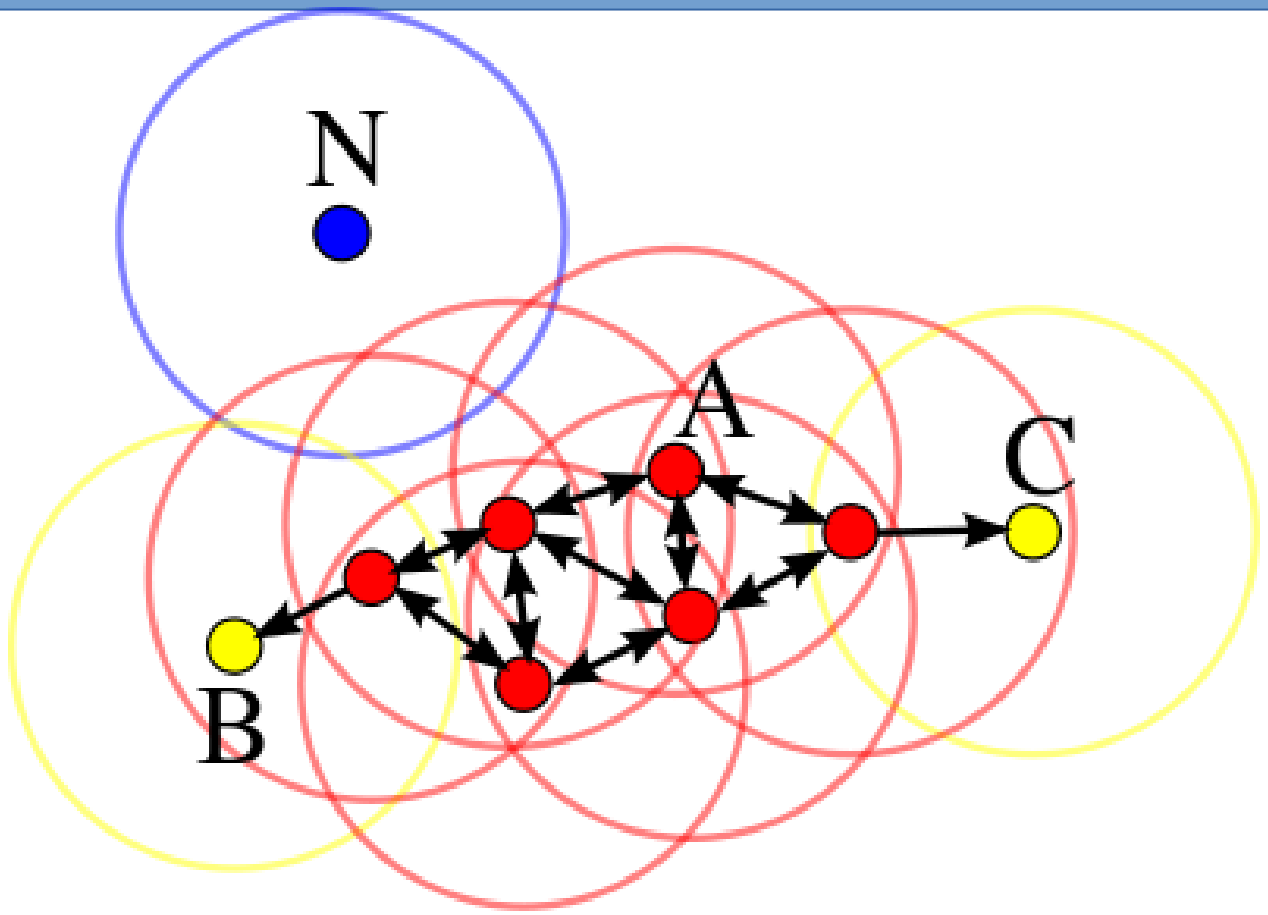




DBSCAN



DBSCAN



Смысловые цели кластеризации

- Минимизировать изменчивость внутри кластеров,
- Максимизировать изменчивость между кластерами.

Расстояние между множествами

- Ближний сосед
- Средний
- Дальний сосед

- Метрика Хаусдорфа

- И МНОГО ДРУГИХ МЕТОДОВ!!!!!

Анализ результатов кластеризации.

- не является ли полученное разбиение на кластеры случайным;
- является ли разбиение надежным и стабильным на подвыборках данных;
- существует ли взаимосвязь между результатами кластеризации и переменными, которые не участвовали в процессе кластеризации;
- можно ли интерпретировать полученные результаты кластеризации.

Процедуры проверки качества кластеризации:

- анализ результатов кластеризации, полученных на определенных выборках;
- кросс-проверка;
- проведение кластеризации при изменении порядка наблюдений в наборе данных;
- проведение кластеризации при удалении некоторых наблюдений;
- проведение кластеризации на небольших выборках.

Использование нескольких методов

- Отсутствие подобия не будет означать некорректность результатов,
- Присутствие похожих групп считается признаком качественной кластеризации.

Как сделать кластер анализ быстрее

- Провести предобработку данных
  - Правильный выбор координат (оценка информативности)
  - Удаление выбросов (статистика и нормализация модели)
  - Редукция размерности
    - Факторный анализ (МЕТОД ГЛАВНЫХ КОМПОНЕНТ)
    - Многомерное шкалирование