K-means (Евдокимов)

Описание программы

Задача программы состоит в кластеризации точек на плоскости с помощью алгоритма k-means в Python.
Для начала по заданным данным строится модель k-means с помощью библиотеки sklearn. Затем для определения эффективности модели или корректности вводимых данных, проверям получившиеся кластеры на наличие точек пересечения. Затем программа сверяет изначально заданные кластеры с полученными при помощи k-means, и демонстрирует совпадения. Если мы имеем 100% совпадений, то наш k-means отлично сработала для этих данных. После этого выводятся кластеры, полученные метод Уорда (иерархическая кластеризация), и тоже сравниваются с изначальными.

Тесты и эксперементы

(Можно самостоятельно запустить тесты в файле test.py) Проверим программу на разных данных:

Тест на случайных точках

Генирируются 2 случайных кластера по 300 точек. С большой вероятностью никогда не получится так, что модель когда-нибудь со 100% точностью сопоставит кластеры. Но чисто теоритически, это может когда-то произойти.

Тестируем на простых квадратах

Оба вида кластеризации хорошо справляются с этой задачей. Явные 3 квадрата были разделены на 3 кластера.
Результат:
Запускаю k-means
Одинаковые кластеры
[[2, 2], [2, 3], [2, 4], [3, 2], [3, 3], [3, 4], [4, 2], [4, 3], [4, 4]]
[[2, 2], [2, 3], [2, 4], [3, 2], [3, 3], [3, 4], [4, 2], [4, 3], [4, 4]]
|-|
Одинаковые кластеры
[[6, 6], [6, 7], [6, 8], [7, 6], [7, 7], [7, 8], [8, 6], [8, 7], [8, 8]]
[[6, 6], [6, 7], [6, 8], [7, 6], [7, 7], [7, 8], [8, 6], [8, 7], [8, 8]]
|-|
Одинаковые кластеры
[[8, 2], [8, 3], [8, 4], [9, 2], [9, 3], [9, 4], [10, 2], [10, 3], [10, 4]]
[[8, 2], [8, 3], [8, 4], [9, 2], [9, 3], [9, 4], [10, 2], [10, 3], [10, 4]]
|-|
Данный вид кластеризации работает отлично
Запускаю иерархическую кластеризацию по методу Уорда
Одинаковые кластеры
[[2, 2], [2, 3], [2, 4], [3, 2], [3, 3], [3, 4], [4, 2], [4, 3], [4, 4]]
[[2, 2], [2, 3], [2, 4], [3, 2], [3, 3], [3, 4], [4, 2], [4, 3], [4, 4]]
|-|
Одинаковые кластеры
[[6, 6], [6, 7], [6, 8], [7, 6], [7, 7], [7, 8], [8, 6], [8, 7], [8, 8]]
[[6, 6], [6, 7], [6, 8], [7, 6], [7, 7], [7, 8], [8, 6], [8, 7], [8, 8]]
|-|

Одинаковые кластеры
[[8, 2], [8, 3], [8, 4], [9, 2], [9, 3], [9, 4], [10, 2], [10, 3], [10, 4]]
[[8, 2], [8, 3], [8, 4], [9, 2], [9, 3], [9, 4], [10, 2], [10, 3], [10, 4]]
|-|
Данный вид кластеризации работает отлично

Ромб, треугольник, квадрат

Три впритык расположенные фигуры. Верно определил k-means только квадрт, остальные точки фигур перемешались в кластерах. Уорд не справился вовсе.
Результат:
Запускаю k-means
Одинаковые кластеры
[[3, 2], [4, 2], [5, 2], [3, 3], [4, 3], [5, 3], [3, 4], [4, 4], [5, 4]]
[[3, 2], [4, 2], [5, 2], [3, 3], [4, 3], [5, 3], [3, 4], [4, 4], [5, 4]]
|-|
Полученные кластеры
[[1, 1], [3, 1], [0, 0], [1, 0], [2, 0], [3, 0], [4, 0]]
[[0, 3], [1, 2], [2, 3], [1, 4], [2, 2]]
[[3, 2], [4, 2], [5, 2], [3, 3], [4, 3], [5, 3], [3, 4], [4, 4], [5, 4]]
Запускаю иерархическую кластеризацию по методу Уорда
Полученные кластеры
[[1, 2], [1, 1], [3, 1], [0, 0], [1, 0], [2, 0], [3, 0], [4, 0]]
[[4, 2], [5, 2], [4, 3], [5, 3], [3, 4], [4, 4], [5, 4]]
[[0, 3], [2, 3], [1, 4], [3, 2], [3, 3], [2, 2]]

Окружность с точкой внутри

Никто не справился. Было необходимо поместить точку в центре окружности в один кластер, а саму окружность в другой.
Результат:
Запускаю k-means
Полученные кластеры
[[-5, 0], [0, -5], [-3, -4], [-4, -3]]
[[5, 0], [0, 5], [3, 4], [4, 3], [0, 0]]
Запускаю иерархическую кластеризацию по методу Уорда
Полученные кластеры
[[5, 0], [0, 5], [3, 4], [4, 3], [0, 0]]
[[-5, 0], [0, -5], [-3, -4], [-4, -3]]

“Маленькие” данные

-1<=x<=1 и -1<=y<=1
Строим два треугольника в первой и третей координатной четвертях и прямую, их раздяляющую.
Никто не справился с задачей.
Результат:
Запускаю k-means
Полученные кластеры
[[1, 1], [1, 0], [0, 1], [0, 0]]
[[-1, -1], [-1, 0], [-1, 1]]
[[0, -1], [1, -1]]
Запускаю иерархическую кластеризацию по методу Уорда
Полученные кластеры
[[0, 1], [-1, -1], [-1, 0], [0, 0], [-1, 1]]
[[0, -1], [1, -1]]
[[1, 1], [1, 0]]

Выводы

Смотря на результаты тестов, можно утверждать, что для разных данных требуются тщательный анализ и разные подходы к кластеризации, так как разные типы имеют разную эффективность для определенного сорта задач.