НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Дисциплина: «Анализ данных»

Домашнее задание на тему:
«Лабораторная работа №13»

Выполнил: Осипов Лев,
студент группы 301ПИ (1).

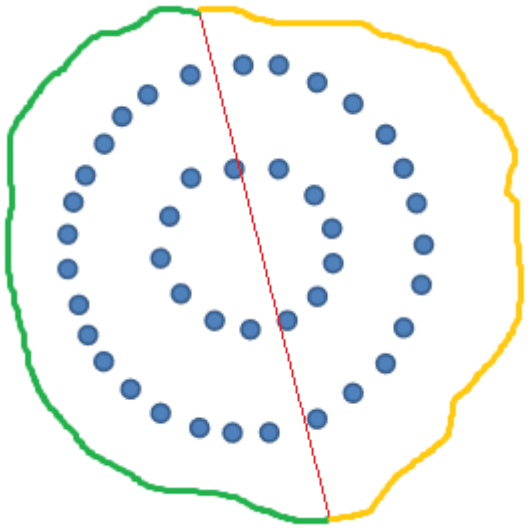
Москва, 2015 г.

СОДЕРЖАНИЕ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

ЗАДАНИЕ 2

Результат кластеризации при k=2:



Для того чтобы разделить объекты во внешнем и внутреннем круге, можно ввести в алгоритм полиномиальное ядро. С его помощью мы сможем спроецировать точки на трехмерное пространство, там уже кластер выделится как фигура, которая при проекции на плоскость даст нам окружность, которая разделит объекты нужным образом.

ЗАДАНИЕ 3

Что касается формальной алгоритмической сложности, в обоих вариантов алгоритма она будет равна O(n*d*k*i), где n – количество объектов, d – количество признаков, k – количество кластеров, i – количество итераций алгоритма. По сути эти алгоритмы отличаются лишь методом подсчета расстояния: расстояние Манхеттена (L1) считается как сумма модулей разности координат, тогда как в Евклидом варианте расстояние считаем евклидовое. На первый взгляд может показаться, что Манхеттен должен работать быстрее, так как арифметические действия в нем менее затратны. Но это вовсе не означает, что при этих двух способах подсчета количество итераций алгоритма будет равное. Поэтому сказать что-то определенное о фактическом времени работы сложно. Разницу стоит мерить эмпирически и находить закономерности.

ЗАДАНИЕ 4

Если мы имеем только расстояния между объектами, можно поступить следующим образом: изначально брать за центры случайные k точек. Затем формировать кластеры, а потом, на основе существующих расстояний, выбирать среди остальных точек геометрическую медиану (точку, у которой сумма расстояний до остальных минимальна). И затем снова формировать кластеры и т.д.

Встает вопрос о сложности нахождения медианы – первый вариант, который приходит в голову (грубая сила) – O(n^2). Я не смог найти алгоритма, который делает это за меньшее время. Может быть, здесь могла бы помочь какая-либо вариация генетического алгоритма.

СПИСОК ЛИТЕРАТУРЫ

1) **Анализ данных (Программная инженерия)** – http://wiki.cs.hse.ru/Анализ_данных_(Программная_инженерия)