Лабораторная работа № 3.

Тема: Кластеризация методом k-средних.

Выполнил: студент группы ИУ7-71

Кочуркин И.А.

Теория

**k-means** (иногда называемый **k-средних**) - наиболее популярный метод [кластеризации](http://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F). Алгоритм представляет собой модификацию [EM-алгоритма](http://ru.wikipedia.org/wiki/EM-%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC) для разделения смеси [гауссиан](http://ru.wikipedia.org/wiki/%D0%93%D0%B0%D1%83%D1%81%D1%81%D0%B8%D0%B0%D0%BD%D0%B0). Он разбивает [множество](http://ru.wikipedia.org/wiki/%D0%9C%D0%BD%D0%BE%D0%B6%D0%B5%D1%81%D1%82%D0%B2%D0%BE) элементов [векторного пространства](http://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D1%81%D1%82%D0%B2%D0%BE) на заранее известное число кластеров *k*. Действие алгоритма таково, что он стремится минимизировать [дисперсию](http://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%B8%D1%8F) на точках каждого кластера:

V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 

где *k* - число кластеров, *Si* - полученные кластеры, i = 1, 2, \dots, kи μ*i* - центры масс векторов x_j \in S_i.

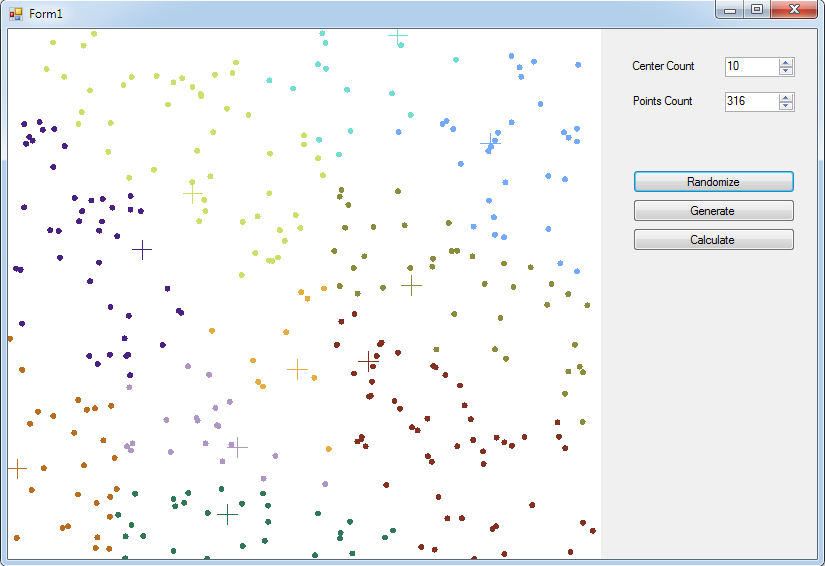
Основная идея заключается в том, что на каждой [итерации](http://ru.wikipedia.org/wiki/%D0%98%D1%82%D0%B5%D1%80%D0%B0%D1%86%D0%B8%D1%8F) перевычисляется [центр масс](http://ru.wikipedia.org/wiki/%D0%A6%D0%B5%D0%BD%D1%82%D1%80_%D0%BC%D0%B0%D1%81%D1%81) для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбраной [метрике](http://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0). Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

Как показали Д.Артур и С.Вассилвицкий, на некоторых классах множеств [сложность алгоритма](http://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D0%BE%D0%B6%D0%BD%D0%BE%D1%81%D1%82%D1%8C_%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC%D0%B0) по времени, нужному для сходимости, равна 2^{\Omega(\sqrt{n})}.[[1]](http://ru.wikipedia.org/wiki/K-means#cite_note-0)

Постановка задачи

Написать программу, реализующую кластеризацию точек по координатам методом k-средних.

Реализация



Программа написана на языке C# в среде Microsoft Visual Studio с использованием библиотек платформы .NET Framework 4.0.

Входными данными являются заданные пользователям точки и центроиды кластеров.

Выходными данными являются кластеры.

Руководство пользователя

Для генерации случайного количества кластеров и точек, нажмите кнопку Randomize

Для генерации кластеров и точек в количествах, указанных в полях, нажмите кнопку Generate