Алгоритм кластеризации k-means

Задача кластеризации

Кластеризация – это задача поиска в данных определенных групп – кластеров. Кластеры характеризуются внутренней однородностью и внешней изолированностью. Кластеризация – задача обучения без учителя, то есть метки классов для каждого объекта заранее не определены. Пример задачи кластеризации – выделение категорий клиентов банка. Существуют различные алгоритмы кластеризации, в том числе иерархические и неиерархические. Самым известным неиерархическим алгоритмом является **k-means** (**k-средних**).

Задача на практику

Предположим, требуется сформировать 2 группы студентов (УТ-11 и УТ-12) для обучения на специальности У. Известны оценки абитуриентов за тесты по физике и математике:

№ студента	Физика	Математика
1	4	4
2	3	3
3	5	3
4	2	3
5	5	5
6	3	2
7	2	4
8	4	5
9	5	4
10	2	2

Требуется реализовать алгоритм **k-means**, с помощью которого выделить 2 кластера, описывающих формируемые группы студентов.

Заготовка кода на языке **Python** приведена в файле **k_means.py**. Ниже приведено словесное описание алгоритма **k-means** и необходимые пояснения, которые могут пригодиться в работе.

Рекомендуется активно использовать консоль **Python** для того, чтобы понять, как работает та или иная функция и конструкция языка, а также команду **help** для получения справки по заданным функциям.

Алгоритм k-means

1. Задается количество кластеров **k**, которые требуется обнаружить

2. Центры кластеров изначально инициализируются случайным образом

3. Каждый из объектов приписывается к ближайшему кластеру

4. На основании объектов, вошедших в каждый кластер, центры кластеров пересчитываются

5. Шаги 3 и 4 повторяются до тех пор, пока центры кластеров не стабилизируются, то есть на очередной итерации объекты будут принадлежать тем же кластерам, что и до этого.

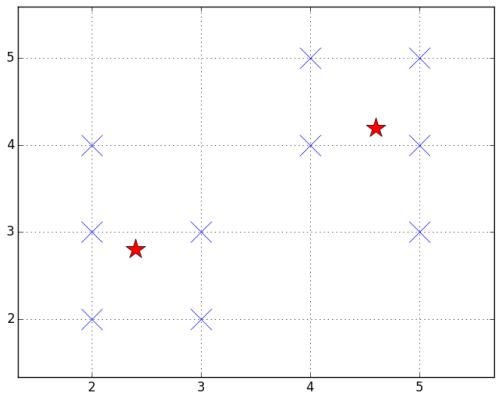
На самом деле, существуют и другие критерии остановки, например, выполнить не более заданного числа итераций или достигнуть приемлемой ошибки. Мы их пока рассматривать не будем.

Для запуска программы необходимо в консоли Windows выполнить следующий код:

```
> python run.py
```

В файле **k_means.py** необходимо найти все комментарии, начинающиеся с **TODO**, и дополнить код так, как описано в комментарии.

Если все фрагменты кода написаны правильно, после запуска файла **run.py** вы должны увидеть следующий график:



Детально разберитесь в коде, приведённом в файлах **run.py** и **k_means.py**, и объясните полученные результаты.