***Кластеризация – разбиение множества объектов на группы (кластеры) так, чтобы объекты одной группы были более похожи друг на друга чем объекты из разных групп.***

Необходимо договориться, что означает более похожи.

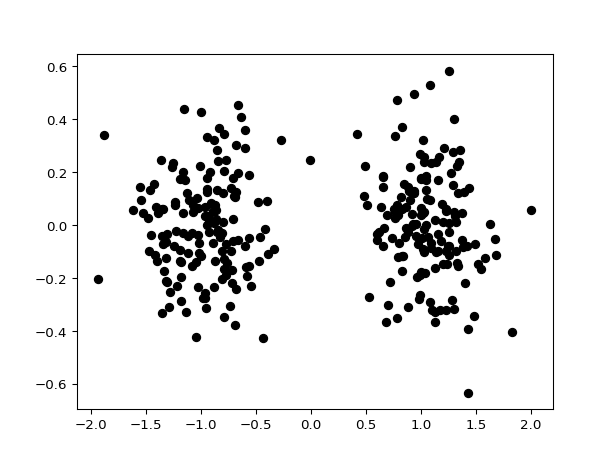
***Формализация задачи, построение математической модели:***

* Реальные объекты -> набор характеристик объекта -> множество элементов как совокупности характеристик.
* Пространство характеристик, вводим метрику в пространстве характеристик. Схожесть объектов ⬄ расстояние между точками в пространстве характеристик.

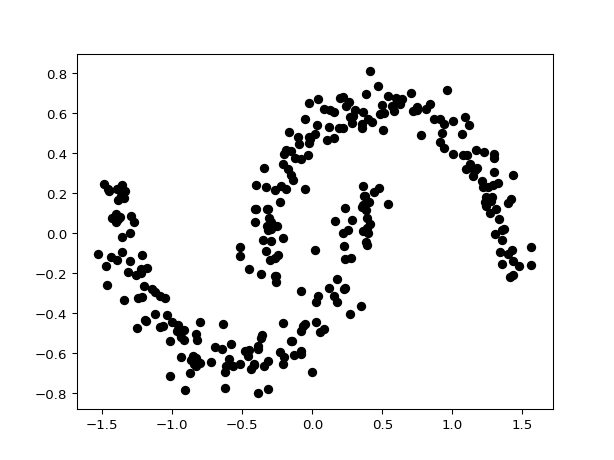
Можно ввести формальное определение кластеризации данных:

здесь – набор кластеров, – расстояние между парой объектов в соответствии с выбранной метрикой.

При таком определении определении требуется задать число кластеров. Такое определение кластеризации даёт хорошие результаты для кластеров вида



В некоторых случаях может приводить к неудовлетворительным результатам:



В подобных ситуациях используются другие определения кластеризации. Например, метод ближайших соседей или кластеризация, основанная на плотности.

**Метод** **К-средних (K-means)**

В методе К-средних каждый кластер представлен координатами его центра (и радиуса, если необходимо) в пространстве характеристик. Решение задачи кластеризации сводится к минимизации квадратичных отклонений точек от центра кластера:

где – координаты центра кластера .

При этом, центр кластера определяется как среднее арифметическое координат входящих в него точек:

Радиус кластера может быть определён через разброс точек от центра кластера:

Алгоритм кластеризации можно представить в виде следующей последовательности шагов:

1. Задаём число кластеров и положение центров кластеров
2. Каждую точку входного множества включаем в тот кластер, центр которого оказался ближайшим
3. Вычисляем новые значения центров кластеров, в соответствии с распределением точек, полученным на предыдущем шаге
4. Повторяем шаги 2 и 3 до тех пор, пока распределение точек по кластерам не стабилизируется

Алгоритм завершается за конечное число шагов так как количество разбиений конечного множества на заданное число кластеров конечно и на каждой итерации суммарное квадратичное отклонение уменьшается.

Недостатки алгоритма:

* Применим не для всех видов кластеров (см. картинки выше).
* Требуется задавать число кластеров.
* Не гарантируется достижение глобального минимума суммарного квадратичного отклонения.
* Результат зависит от выбора исходных центров кластеров, оптимальный выбор неизвестен.

Нечёткая кластеризация (Fuzzy Classifier Means - FCM)

В основе алгоритма FCM лежит использование матрицы , описывающей вероятность того, что i-й элемент принадлежит кластеру с номером k:

Параметр q задаёт "разброс" относительно центра кластера: чем больше значение q, тем больше вероятность включить в кластер далёкие точки.

Сам алгоритм аналогичен алгоритму К-средних, описанному ранее, с очевидной заменой критерия принадлежности объекта кластеру на 2-м шаге алгоритма.