MACHINE LEARNING AND AI

ЛЕКЦИЯ 8

РАЗНОВИДНОСТИ ML

Обучение с учителем

- Регрессия
- Классификация
- Рекомендательные системы

Обучение без учителя

- Кластеризация
- Задачи сокращения размерности
- Обнаружение аномалий

КЛАСТЕРИЗАЦИЯ

Кластерный анализ выполняет следующие основные задачи:

Разработка типологии или классификации.

Исследование полезных концептуальных схем группирования объектов.

Порождение гипотез на основе исследования данных.

Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

КЛАСТЕРИЗАЦИЯ

Кластерный анализ предполагает следующие этапы:

Отбор выборки для кластеризации. Подразумевается, что имеет смысл кластеризовать только количественные данные.

Определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства.

Вычисление значений той или иной меры сходства (или различия) между объектами.

Применение метода кластерного анализа для создания групп сходных объектов.

Проверка достоверности результатов кластерного решения.

ЦЕЛИ КЛАСТЕРИЗАЦИИ

- Понимание данных путём выявления кластерной структуры
- Сжатие данных
- Обнаружение новизны (англ. novelty detection)

МЕТОДЫ КЛАСТЕРИЗАЦИИ

Иерархическая кластеризация

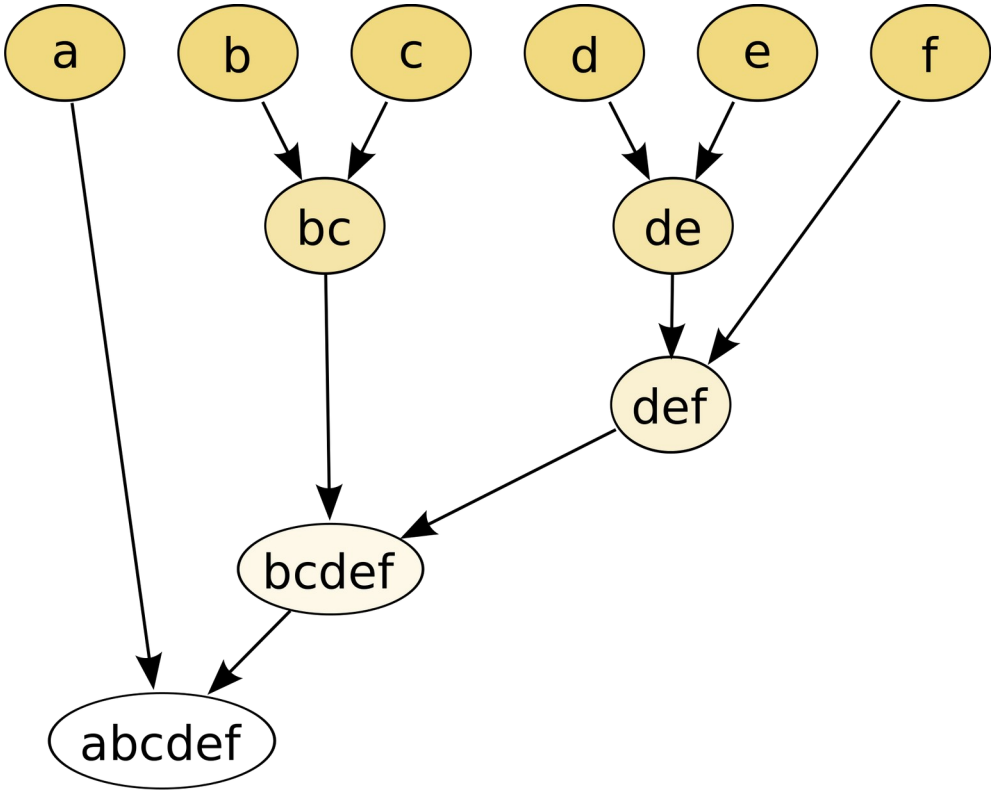
Графовые алгоритмы

Сети Кохонена

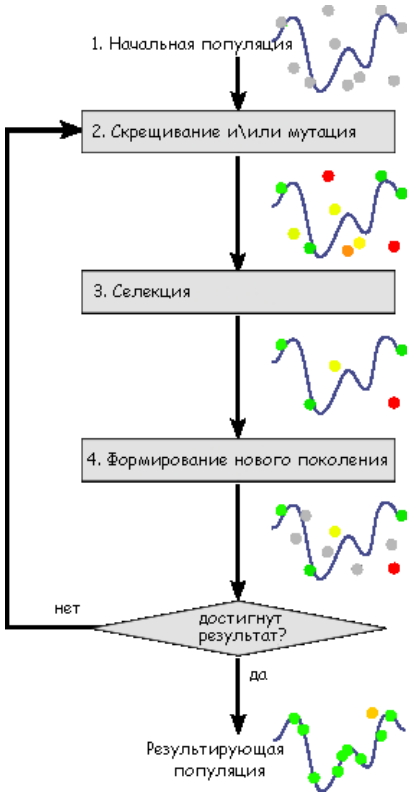
Генетические алгоритмы

Вероятностный подход

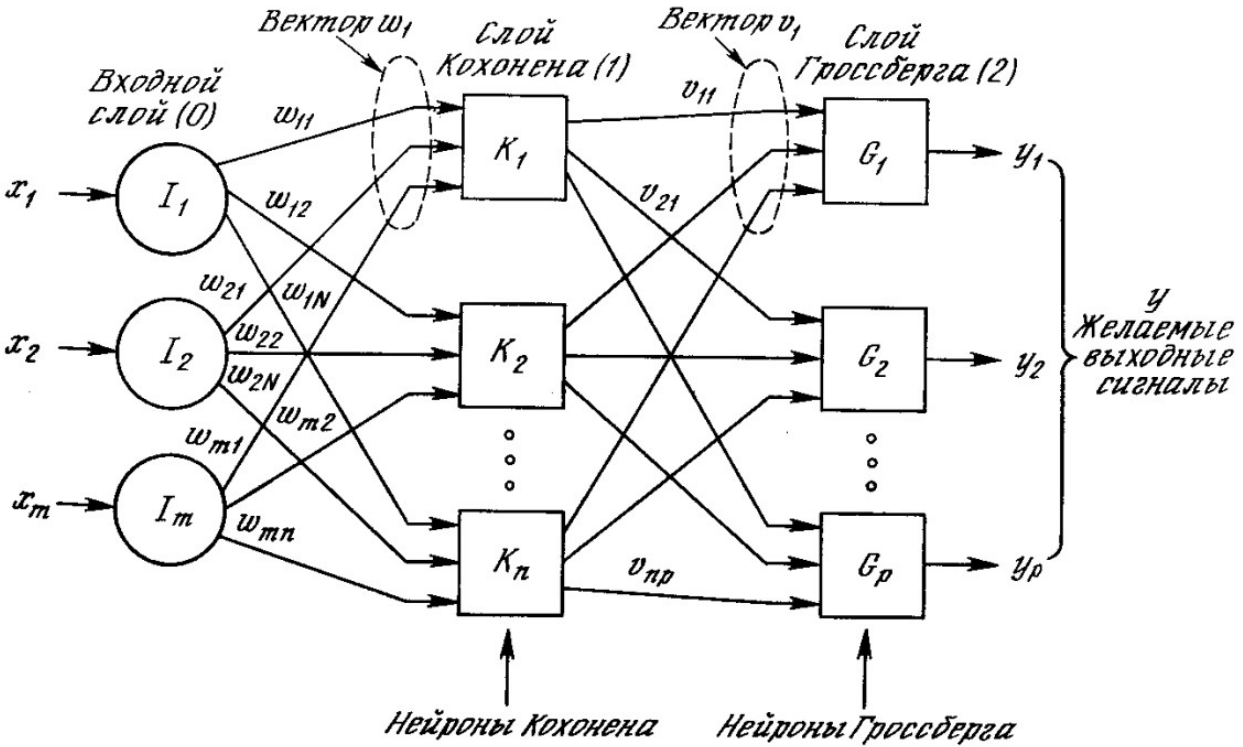
ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ



ГЕНЕТИЧЕСКИЙ АЛГОРИТМ

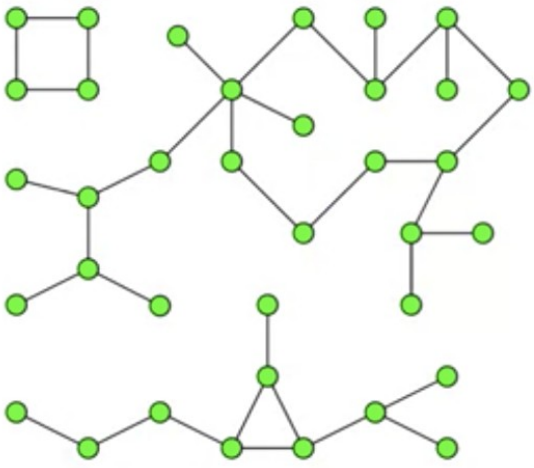


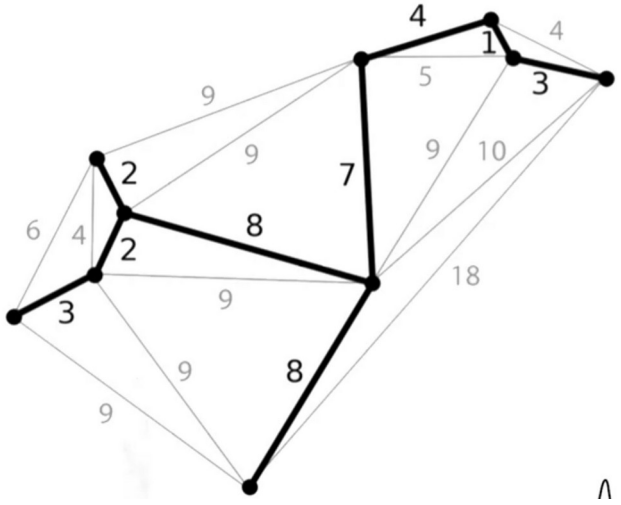
СЕТЬ КОХОНЕНА



ГРАФОВЫЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ

Суть таких алгоритмов заключается в том, что выборка объектов представляется в виде графа G=(V, E), вершинам которого соответствуют объекты, а ребра имеют вес, равный «расстоянию» между объектами



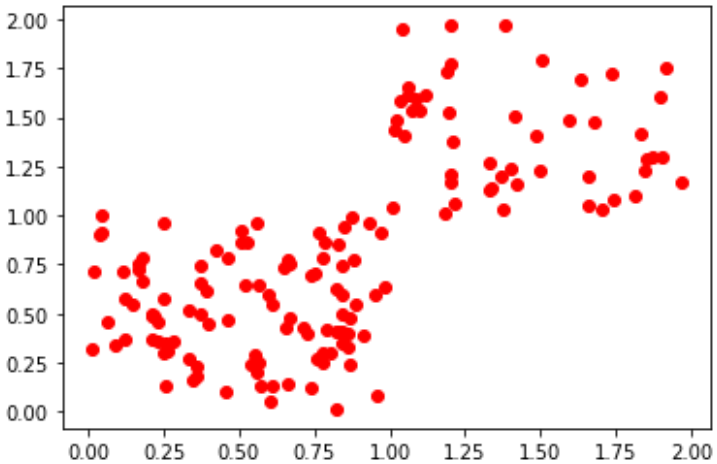


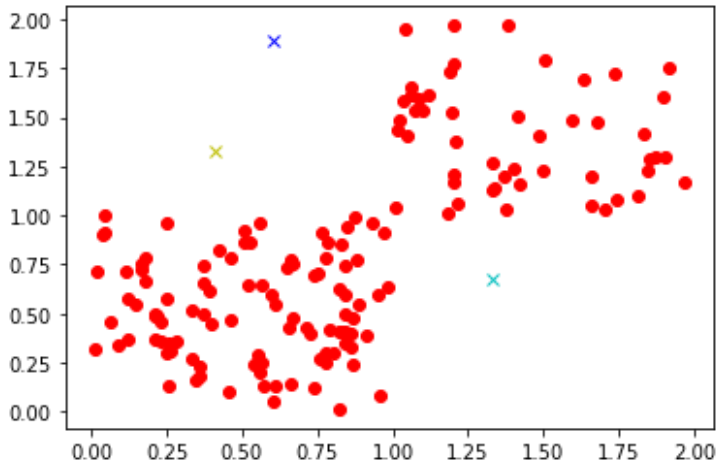
АЛГОРИТМ K-means (K-средних)

1. Генерируем случайным образом центры кластеров
2. Для каждого примера вычисляем ближайший центроид
3. Перемещаем центроиды вглубь точек, принадлежащих этому кластеру
4. Повторяем п. 2-3 до полной сходимости
5. При необходимости посторяем п. 1-4 несколько раз

АЛГОРИТМ K-means (K-средних)

Генерируем центроиды





АЛГОРИТМ K-means (K-средних)

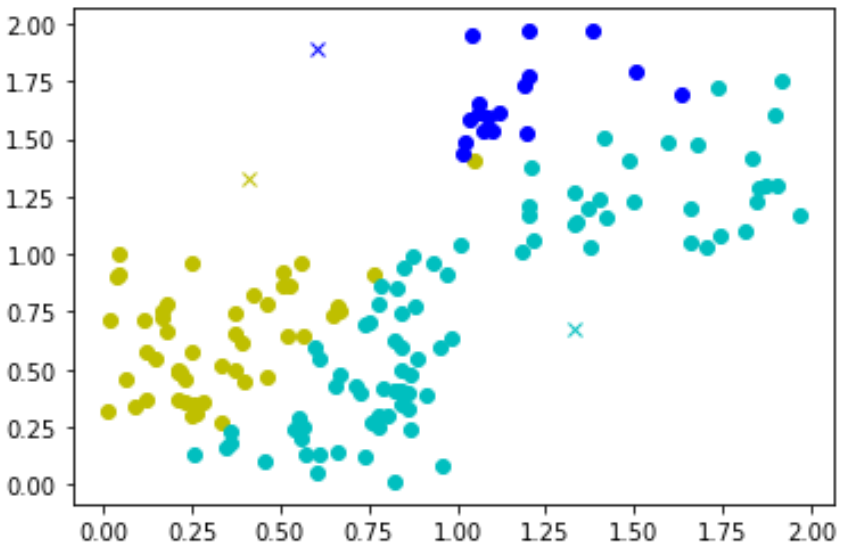
Для каждой точки считаем расстояние до каждого центроида

$$r_{i,j} = \sqrt{\left(x_i - c_{j,0}\right)^2 + \left(y_i - c_{j,1}\right)^2}$$

Номером кластера будет номер с наименьшим расстоянием

$$cluster_i = argmin\left(r_{i,0}, r_{i,1}, r_{i,2}\right)$$

АЛГОРИТМ K-means (K-средних)

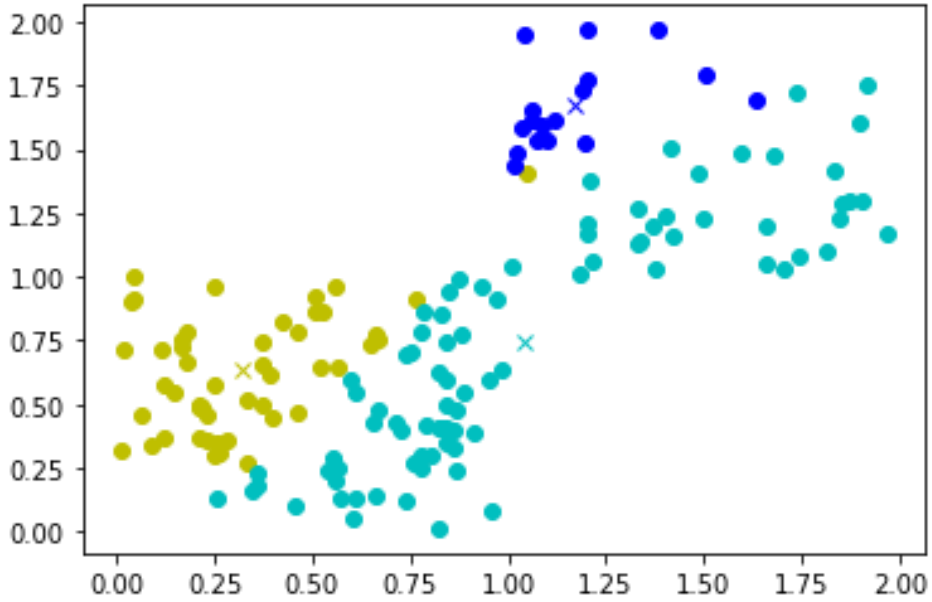


АЛГОРИТМ K-means (K-средних)

Обновляем центры кластеров

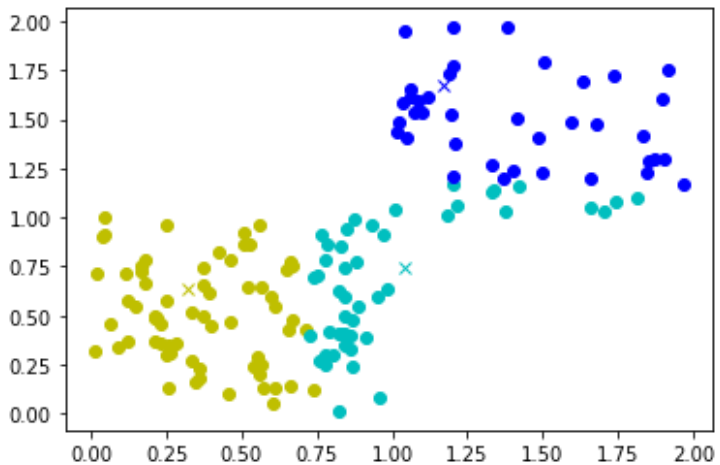
$$\forall (x, y) \in cluster_j \rightarrow c_{j,0} = \frac{\sum_{i=1}^{n_j} x_i}{n_j}$$

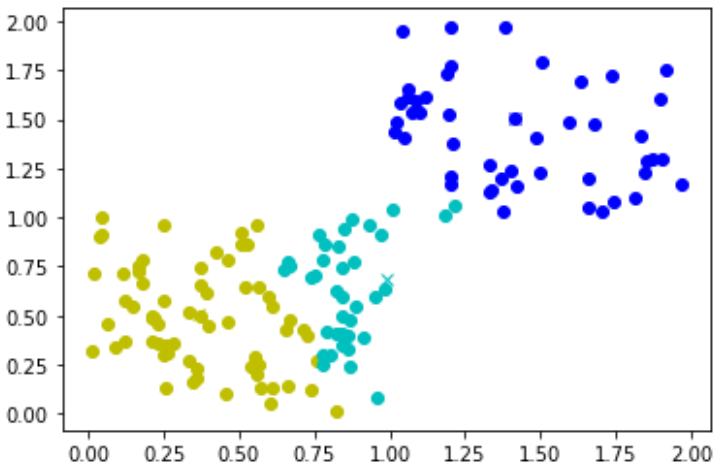
$$\forall (x, y) \in cluster_j \rightarrow c_{j,1} = \frac{\sum_{i=1}^{n_j} y_i}{n_j}$$



АЛГОРИТМ K-means (K-средних)

Следующая итерация





АЛГОРИТМ K-means (K-средних)

Конечный результат