K-means и K-means++

Введение в кластеризацию

Кластеризация — это метод обучения без учителя для группировки данных по схожести.

Позволяет находить скрытые структуры в данных, объединяя объекты в кластеры (группы).

Выделяет группы с похожими свойствами без заранее указанных свойств.

K-means

K-means разбивает данные на K кластеров, минимизируя внутрикластерную дисперсию (насколько точки внутри одной группы разбросаны друг от друга).

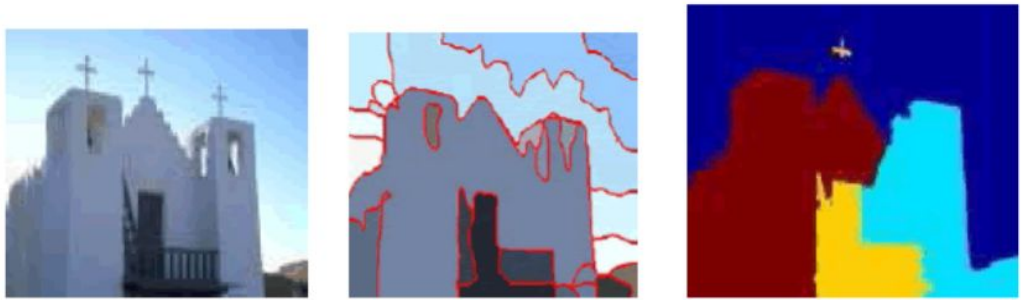


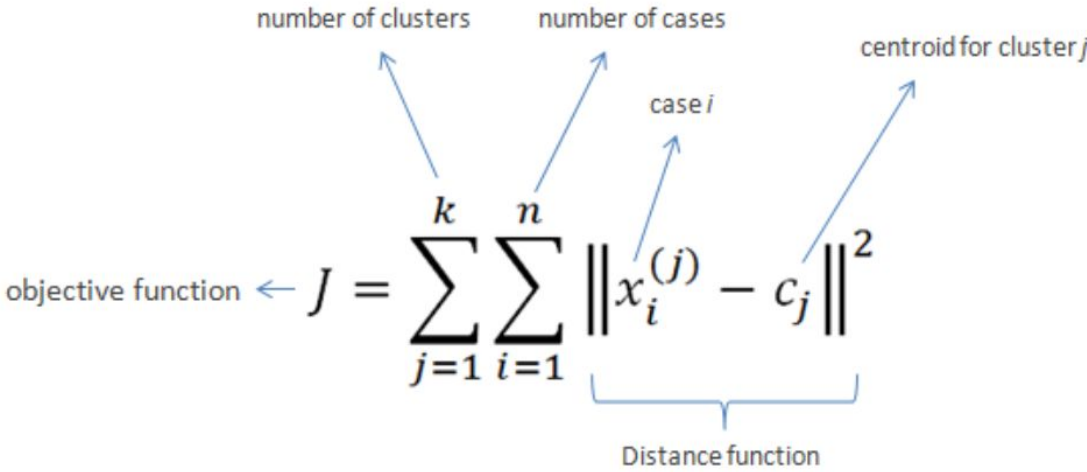
Figure 1: (a) is the original image; (b) and (c) are the segmentation results.

Алгоритм популярен из-за своей вычислительной эффективности (линейная сложность O(n * *K* * I), где n — число точек, I — итерации), но требует от пользователя заранее определить K, что не всегда тривиально.

Функция потерь

Цель K-means: минимизировать расстояние от точек до центроидов

1. Инициализировать K центроидов.
2. Присвоить точки ближайшим центроидам.
3. Пересчитать центроиды.
4. Повторять до сходимости.

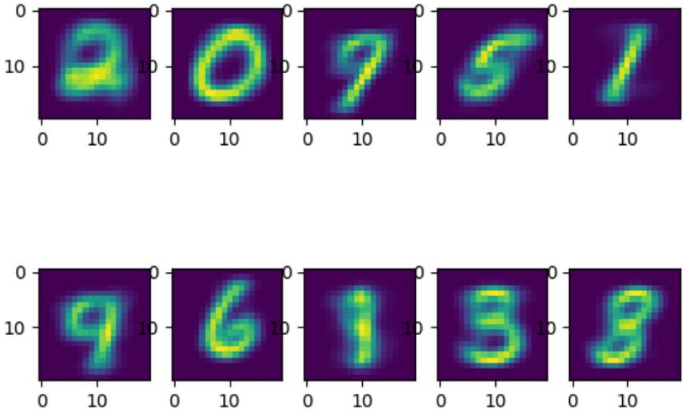


Преимущества и недостатки

- **Плюсы**: Просто, быстро, работает с большими данными.

- **Минусы**: Надо заранее знать k, зависит от стартовых точек, только для чисел.

Случайная инициализация может привести к плохим локальным минимума



пример центроидов найденных K-means алгоритмом

K-means++: улучшенная инициализация

K-means++ выбирает центроиды не случайно, а с вероятностью, пропорциональной расстоянию. Стартовые точки выбираются далеко друг от друга

- результат лучше
- алгоритм быстрее выполняется

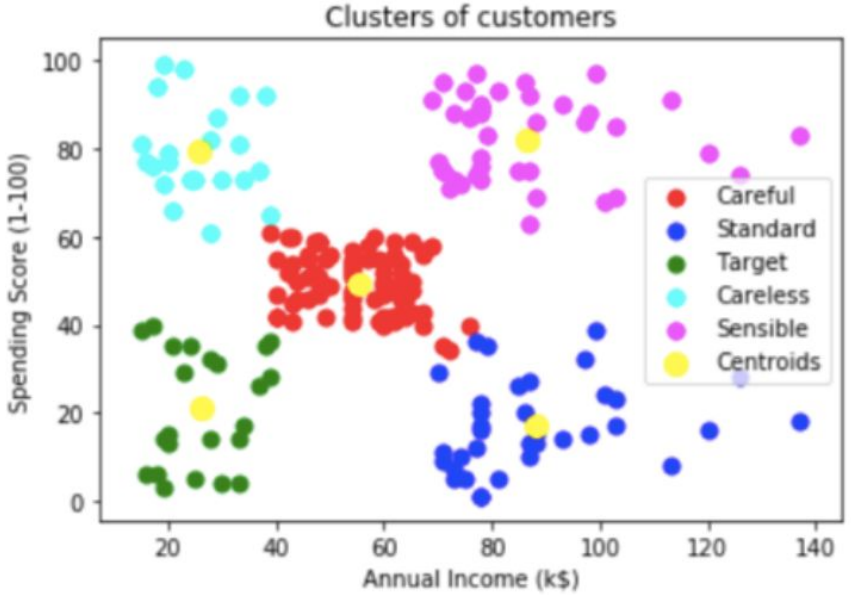
Алгоритм K-means++ в деталях

1. Выбрать первый центр случайно.

2. Расчитать расстояние от каждой точки до ближайшего центра.

3. Выбрать следующий центр там, где точки дальше всего.

4. Повторить, пока не наберем k центров.



K-Means clustering example

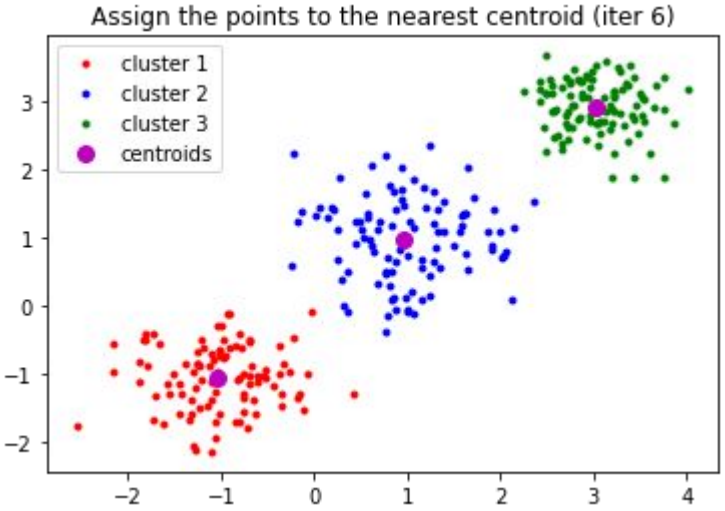
Сравнение производительности

K-means++ быстрее сходится и даёт меньшую дисперсию J.

Ограничения K-means

Чувствителен к выбросам, предполагает сферические кластеры.

K-means и k-means++ — полезные методы, но важно знать их слабые места.



Практическое применение

Сегментация клиентов по предпочтениям, сжатие изображений, анализ генов.

Улучшенные версии K-means

K-medoids - медианы вместо средних

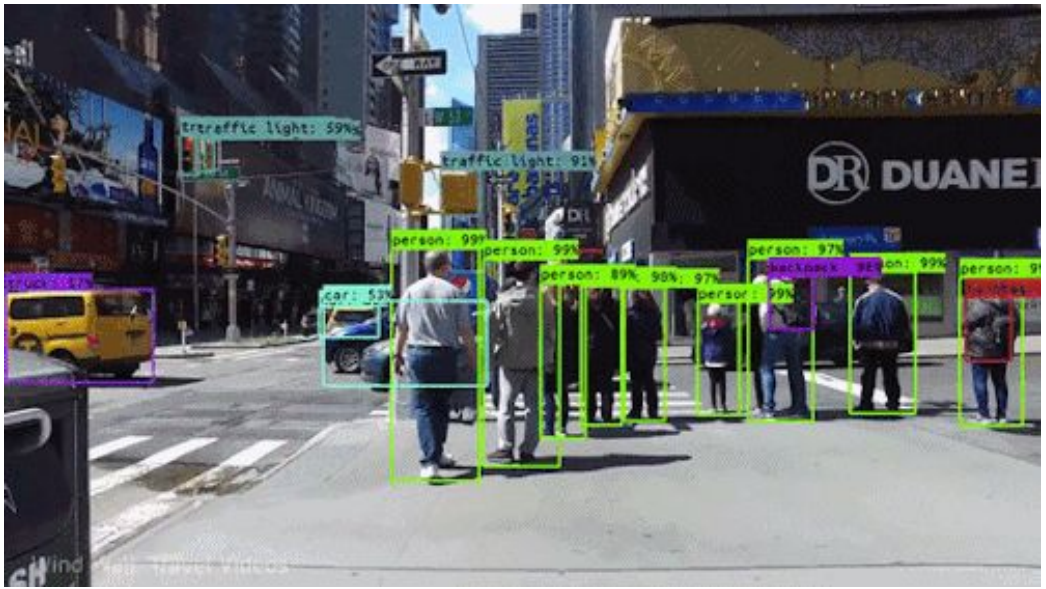
Bisecting K-means - делит данные иерархически

Kernel K-means - проецирует данные в пространство высокой размерности через ядровую функцию

Альтернативы K-means (спектральная кластеризация)

DBSCAN - группирует точки по плотности, не требуя K, и устойчив к шуму

GMM - моделируют кластеры как многомерные нормальные распределения



Выводы и перспективы

K-means и K-means++ — основа кластеризации, но их развитие продолжается.

Оба метода ограничены предположениями о данных