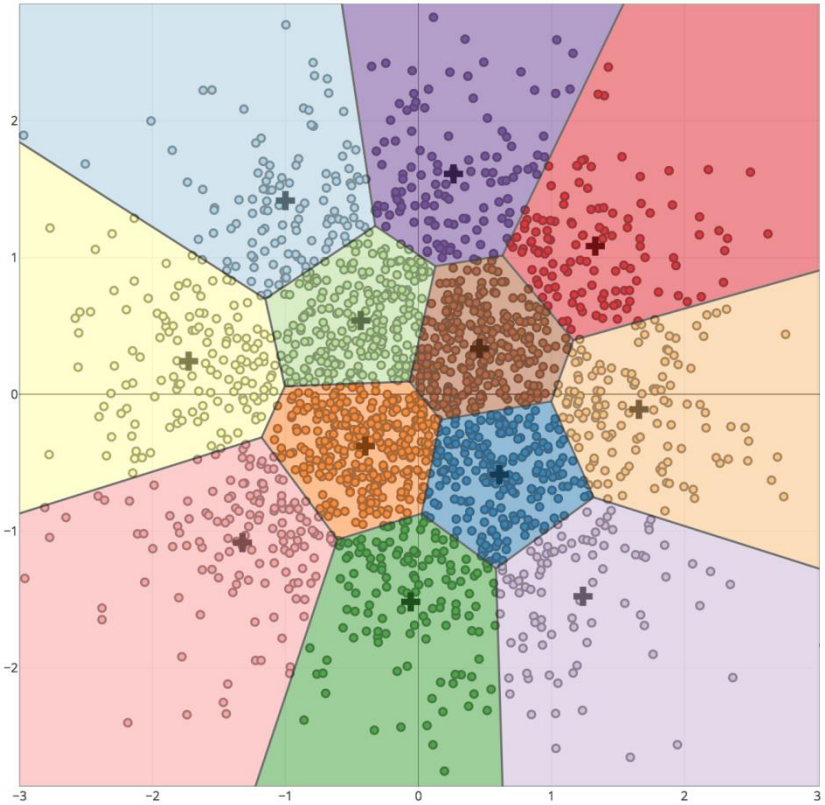
Машинне навчання

Кластерний аналіз

Карта світу машинного навчання

- **Класичне навчання**
  (з вчителем, без вчителя)
- Навчання з підкріпленням
- Ансамблеві методи
- Нейронні мережі і глибоке навчання

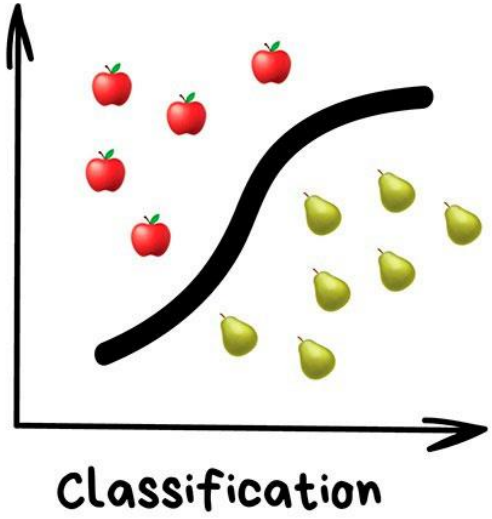
Кластеризація



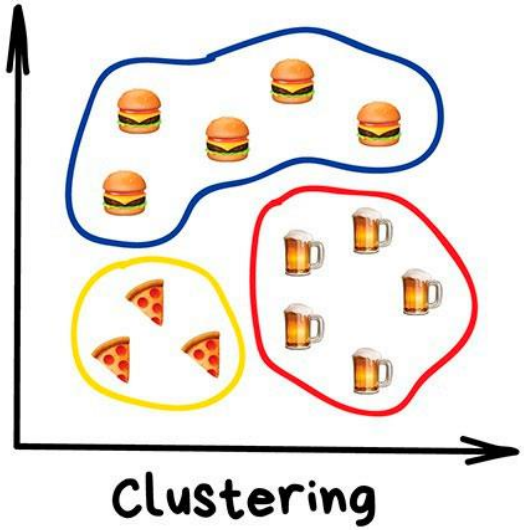
Кластеризація (кластерний аналіз) – це техніка машинного навчання, яка полягає у групуванні точок даних у множини, які називаються **кластерами**

Точки даних, які входять до однієї групи, повинні мати подібні властивості та/або особливості, тоді як точки даних у різних групах повинні мати дуже різні властивості та/або особливості

Кластеризація vs. класифікація



З вчителем
Кількість кластерів відома
Ознаки відомі

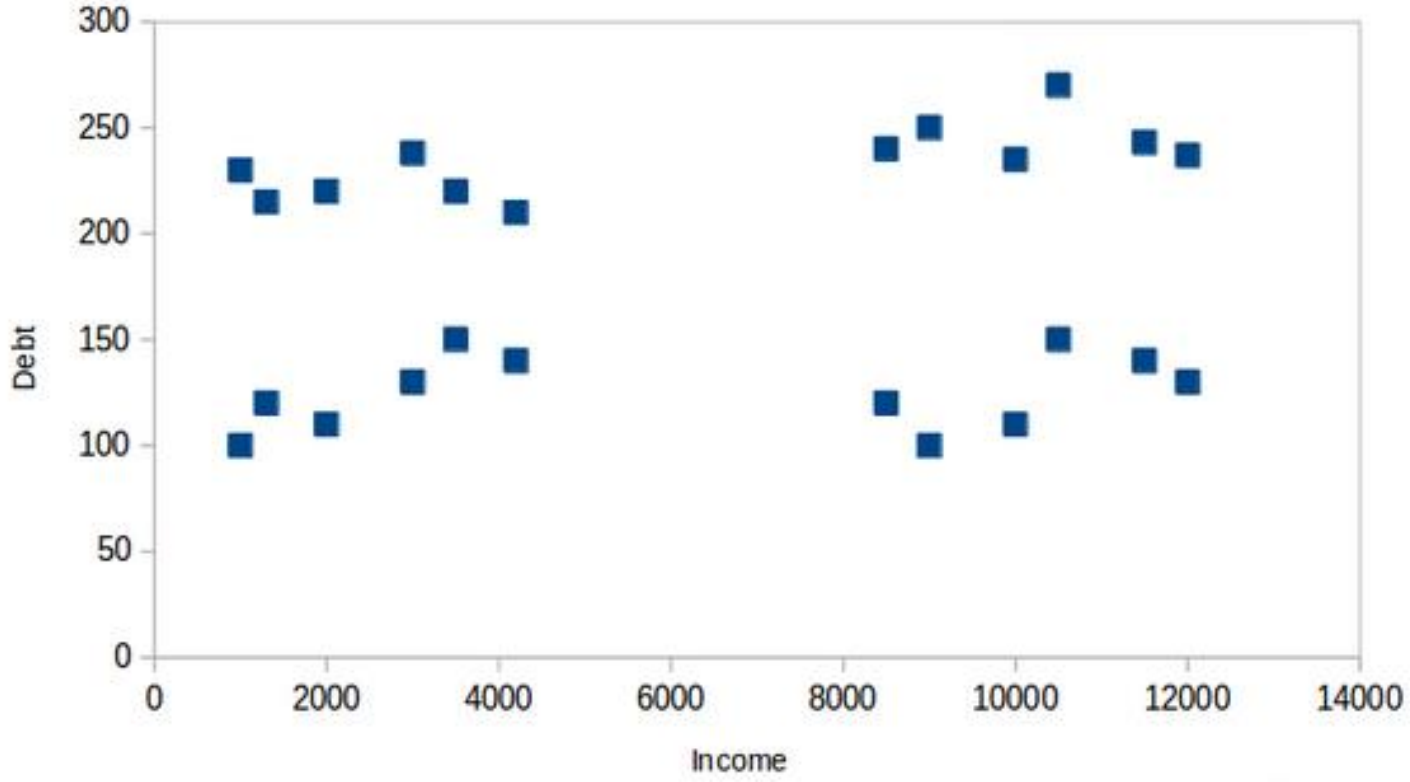


Без вчителя
Кількість кластерів невідома
Ознаки невідомі

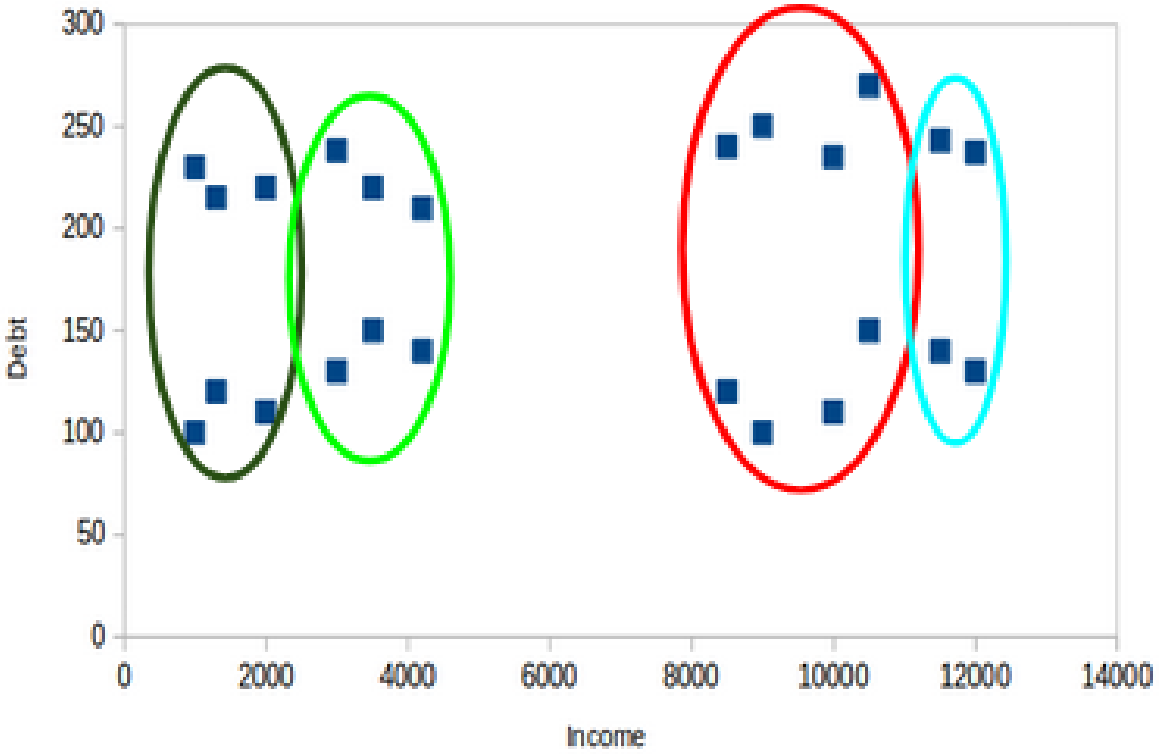
Використання кластеризації

- Сегментація клієнтів
- Сортування документів
- Сегментація зображень
- Рекомендаційні двигуни
- Аналіз ДНК

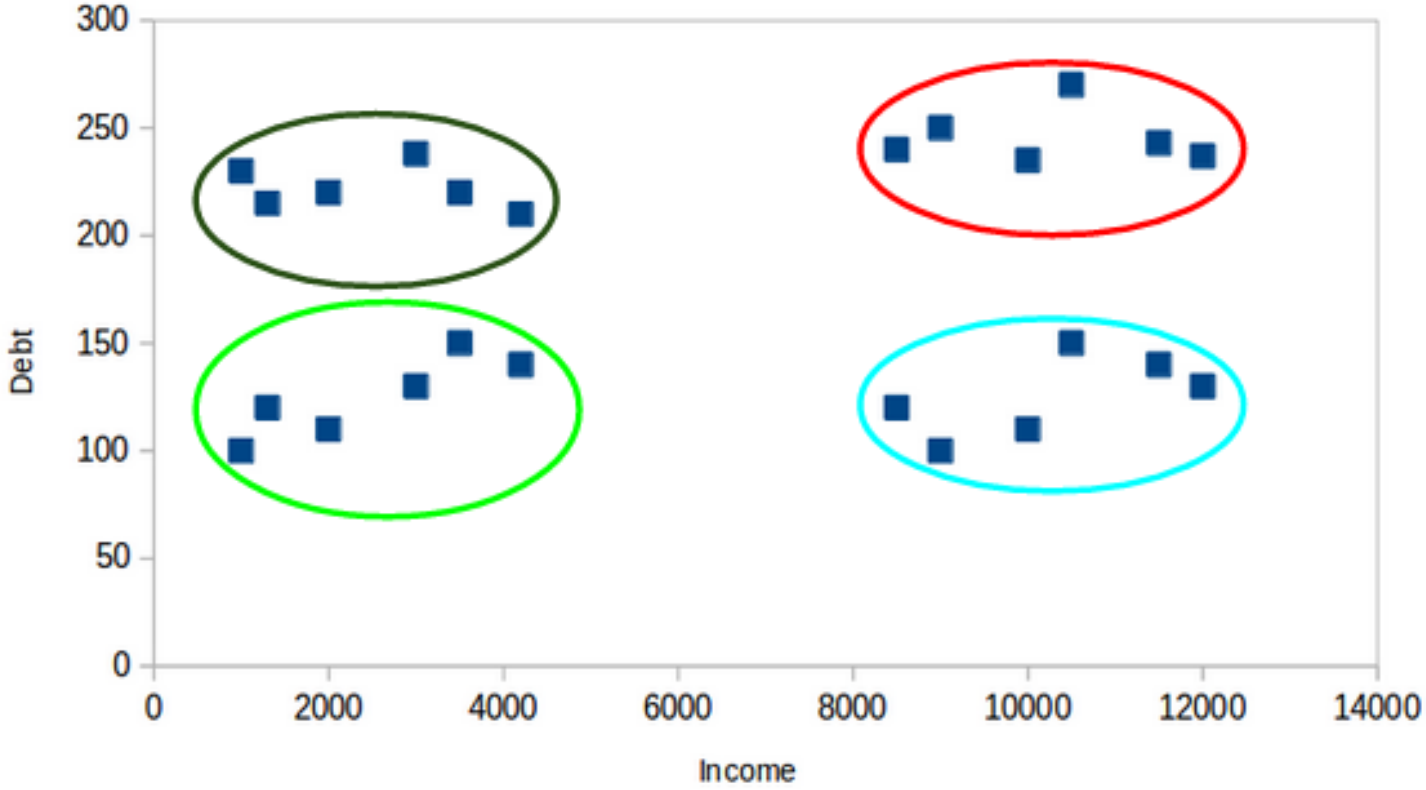
Первинна вибірка даних



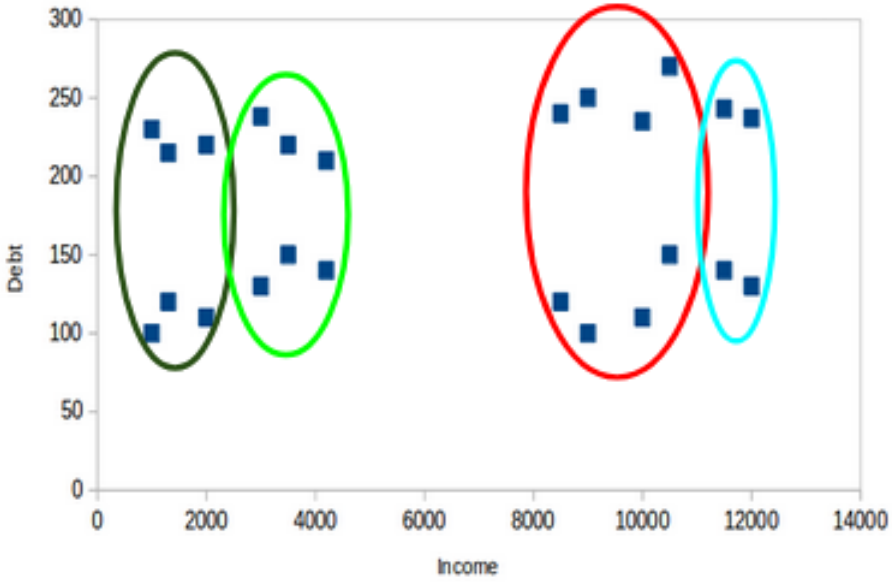
Варіант розподілу клієнтів



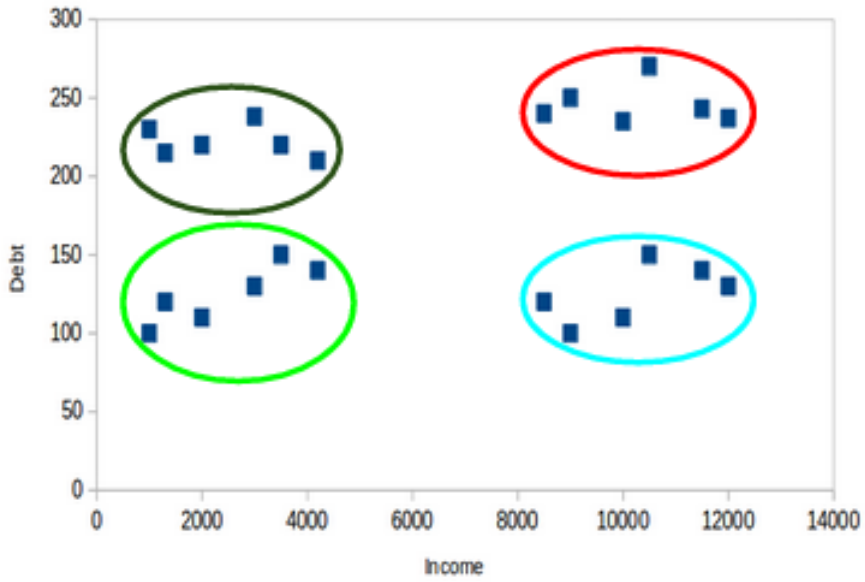
Варіант розподілу клієнтів



Який варіант є найкращим?



Case - I



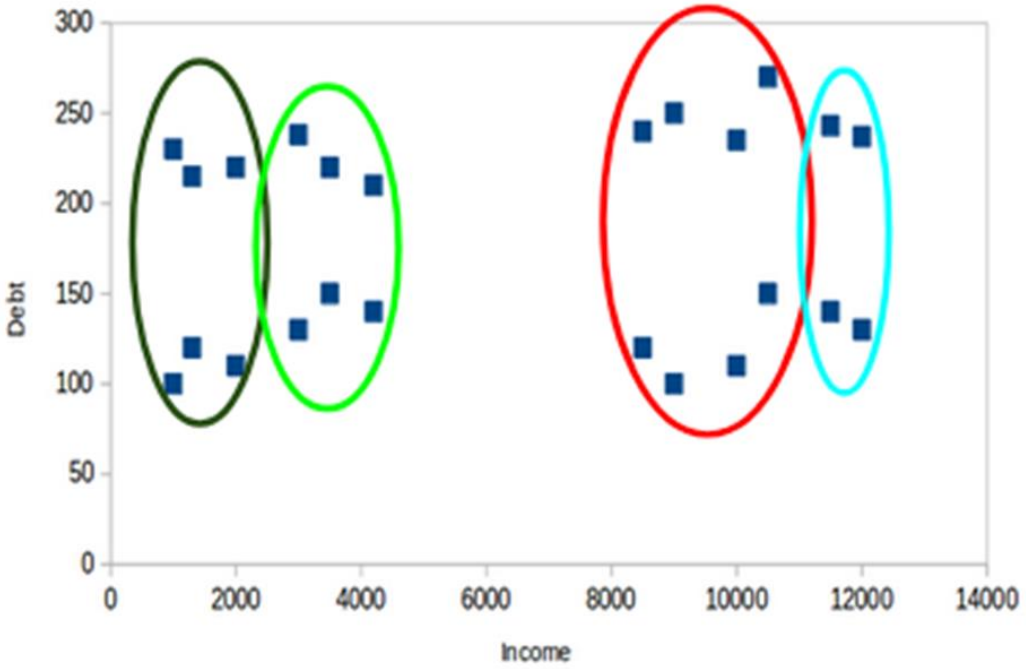
Case - II

Головні властивості кластерів

- Усі точки даних у кластері повинні бути подібними

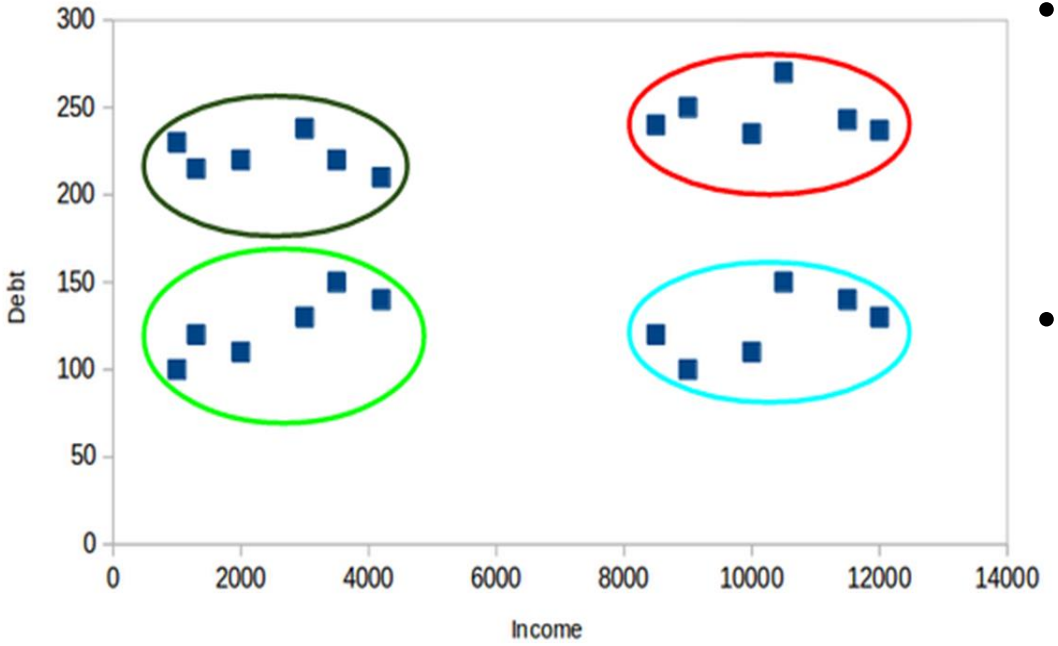
- Точки даних з різних кластерів повинні бути максимально різними

Варіант розподілу клієнтів



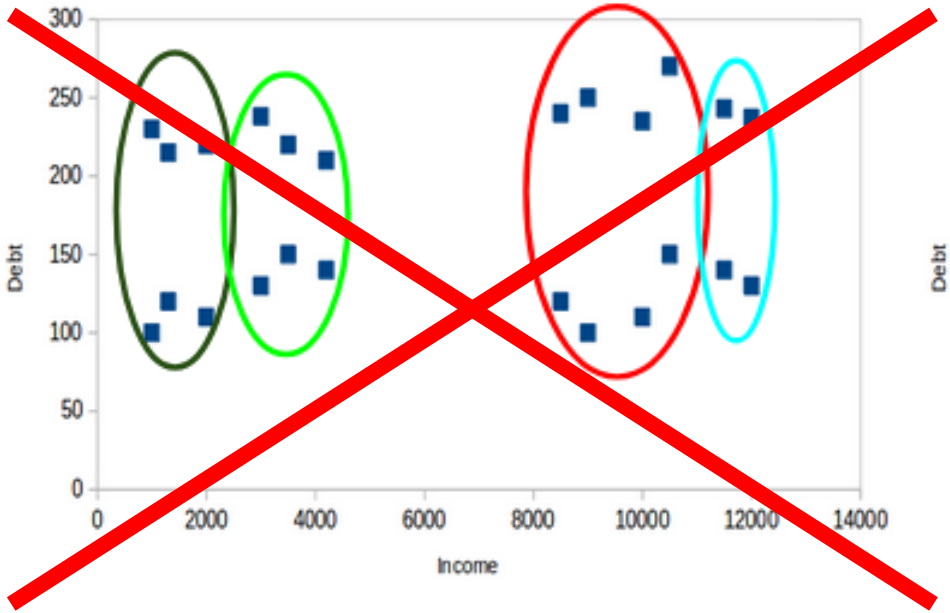
- Клієнти у червоному кластері не схожі між собою (одні мають високій дохід, але великий борг, а інші – високій дохід із малим боргом)
- Клієнти у червоному та синьому кластерах схожі між собою (деякі мають і високий дохід, і великий борг)

Варіант розподілу клієнтів

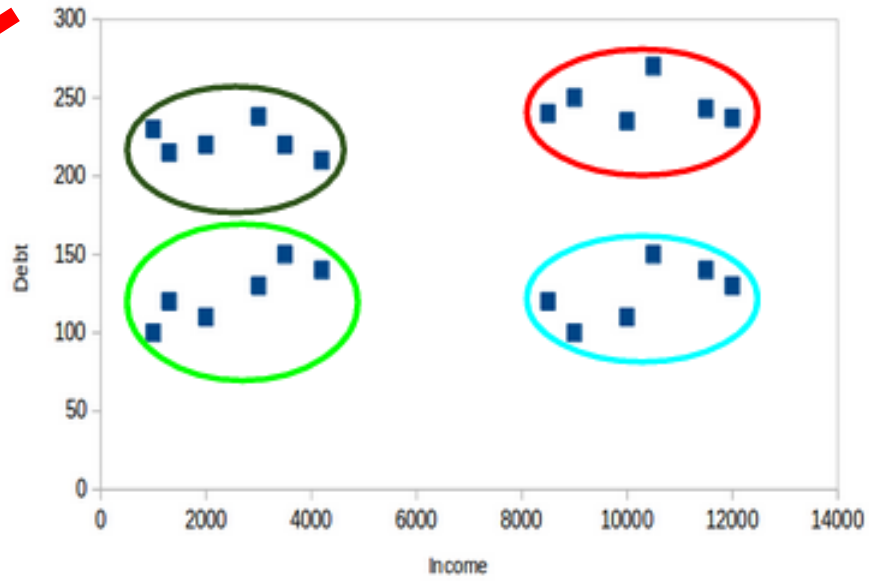


- Усі клієнти у червоному кластері схожі між собою (усі мають високій дохід і великий борг)
- Клієнти у червоному та синьому кластерах не схожі між собою (усі мають високий дохід, але різний рівень боргу)

Який варіант є найкращим?



Case - I

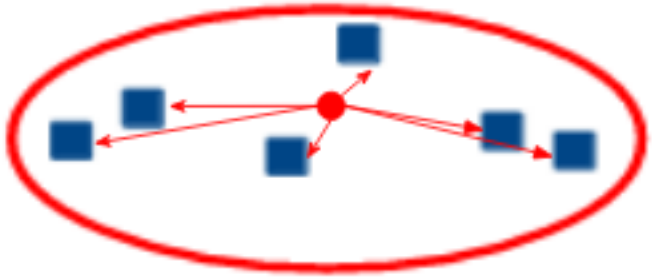


Case - II

Але як виміряти???

Метрики оцінки кластеризації

Інерція

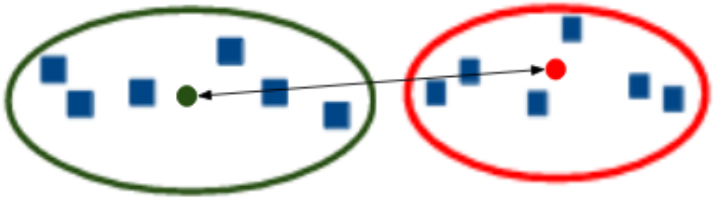


Intra cluster distance

- **Центроїд кластера** – точка із середніми значеннями ознак
- **Внутрішньокластерна відстань** – відстань між елементом та центроїдом
- **Інерція** – сумма внутрішньокластерних відстаней

Чим менше значення інерції, тим кращі кластери

Індекс Данна



Inter cluster distance

$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

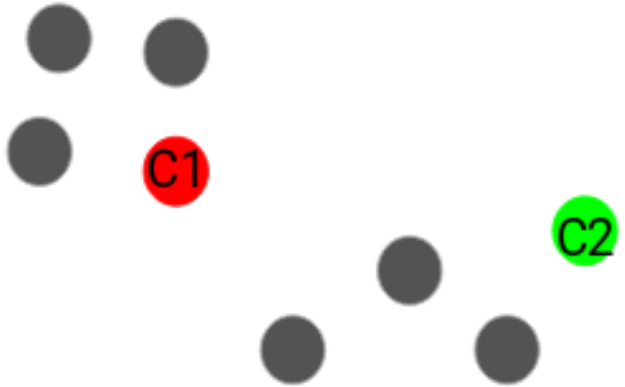
- **Міжкластерна відстань** – відстань між центроїдами різних кластерів
- **Індекс Дана** – відношення мінімальної (серед усіх кластерів) міжкластерної відстані до максимальної (серед усіх кластерів) інерції

Чим більше значення індексу Данна, тим кращими будуть кластери

Алгоритм k-середніх (k-means)

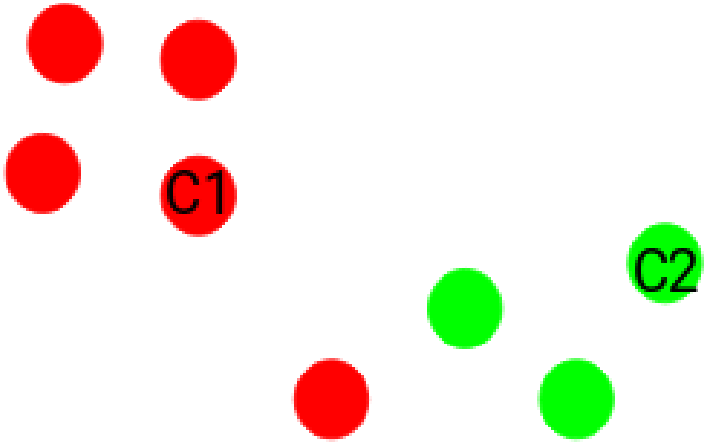
Алгоритм k-середніх

Крок 1. Обираємо кількість кластерів

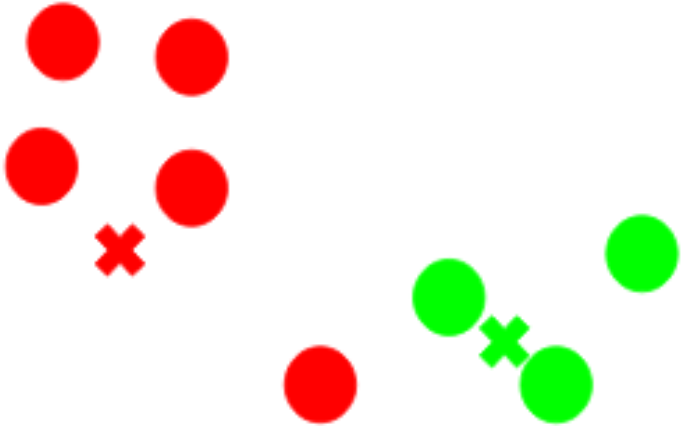


Крок 2. Випадковим чином назначаємо центроіди кластерів

Алгоритм k-середніх



Крок 3. Призначення точок найближчому центроїду кластера



Крок 4. Обчислення центроїдів новоутворених кластерів

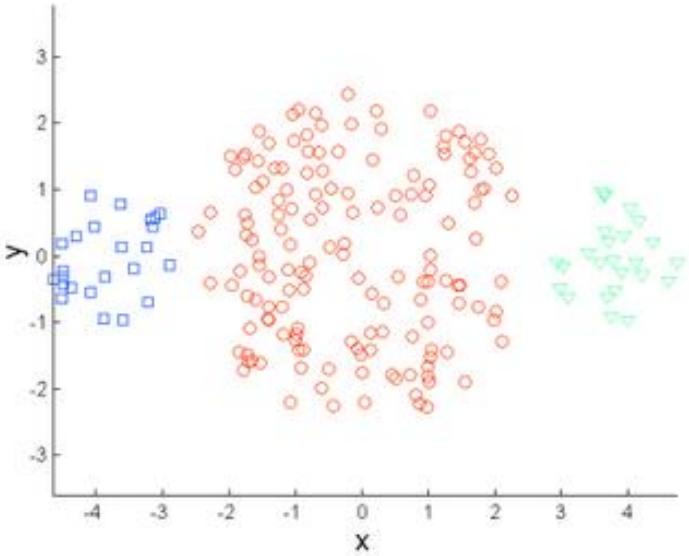
Алгоритм k-середніх

Крок 5. Повторення кроків 3 і 4

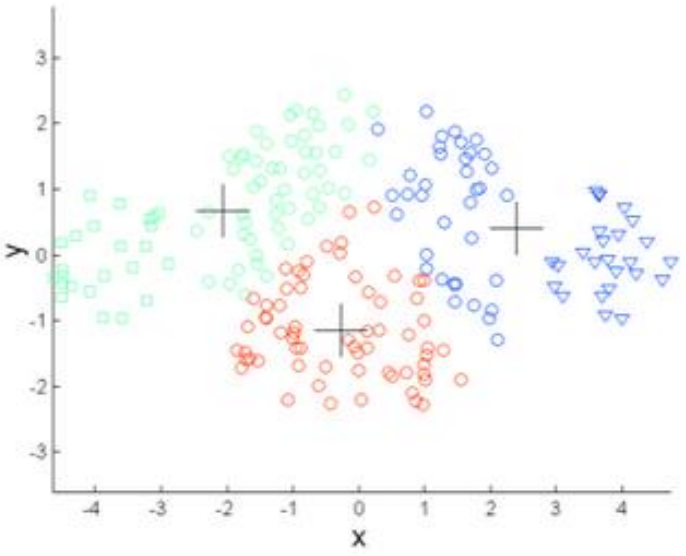
Коли зупинитись?

- Центроїди новоутворених кластерів не змінюються
- Точки залишаються у тому ж скупченні
- Досягнуто максимальної кількості ітерацій

Проблеми з алгоритмом кластеризації k-means



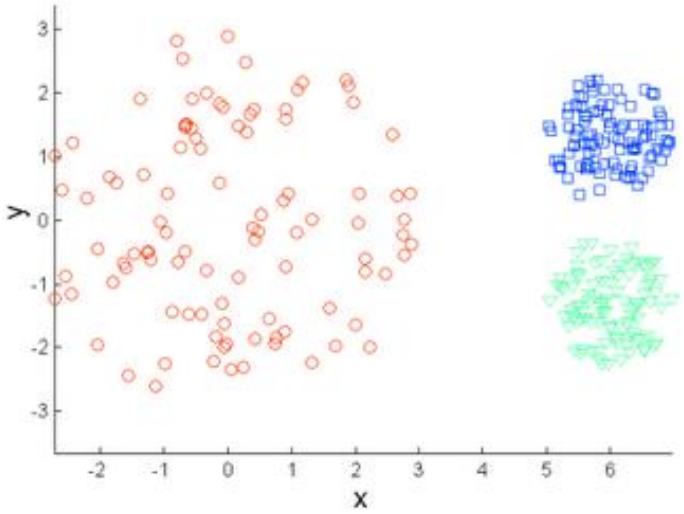
Original Points



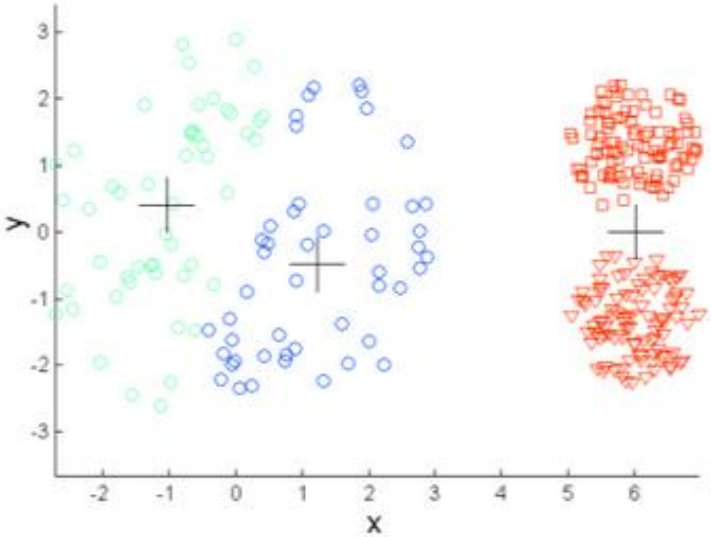
K-means (k = 3)

Різний розмір кластерів

Проблеми з алгоритмом кластерізації k-means



Original Points



K-means (k = 3)

Різна щільність розташування точок

Варіанти вирішення проблем

- Збільшення кількості кластерів
- Відмова від випадкового визначення центроїдів (кожного разу отримуємо різні кластери)
- Варіантом вирішення проблеми є використання алгоритму k-means++

Алгоритм k-середніх++ (k-means++)

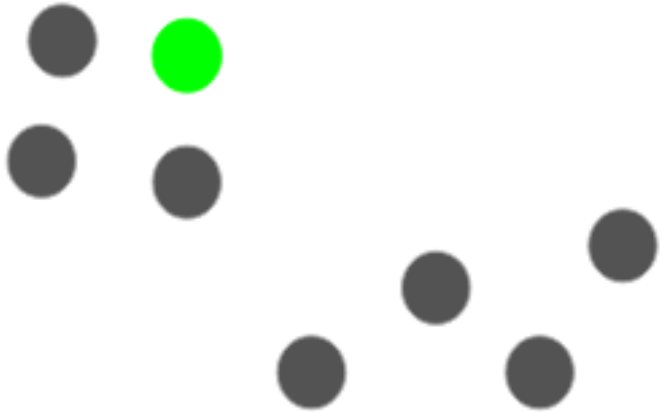
Алгоритм k-середніх++

- Використовуючи алгоритм *k*-means++, ми оптимізуємо крок, коли ми випадковим чином вибираємо центроїд кластера

- Ми, швидше за все, знайдемо рішення, яке буде конкурентоспроможним для оптимального рішення *k*-means, використовуючи ініціалізацію *k*-means++

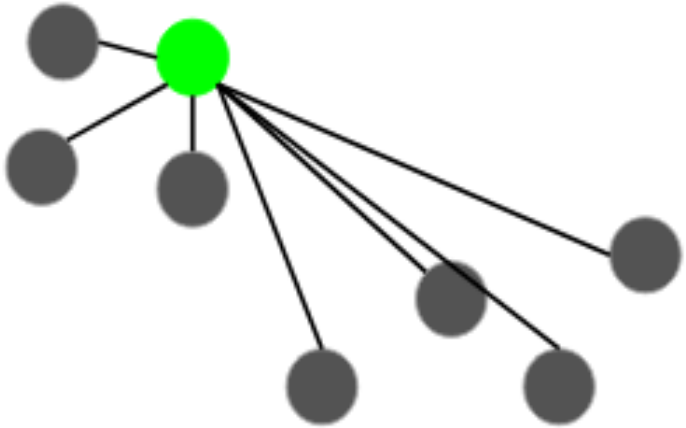
Алгоритм k-середніх++

Крок 1. Обираємо кількість кластерів (k = 3)



Крок 2. Випадково обираємо точку як центроїд кластера

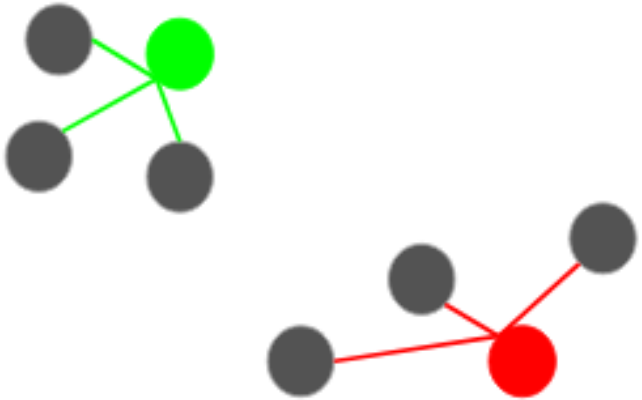
Алгоритм k-середніх++



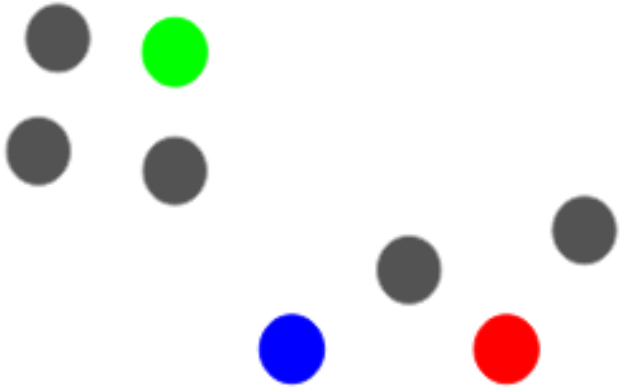
Крок 3. Обчислюємо відстань D(x) до кожної точки даних

Крок 4. Наступним центроїдом буде той, у якого квадрат відстані $D^2(x)$ є найвіддаленішим від поточного центроїда

Алгоритм k-середніх++



Крок 5. Обчислюємо відстань *D(x)* кожної точки даних до найближчого центроїда



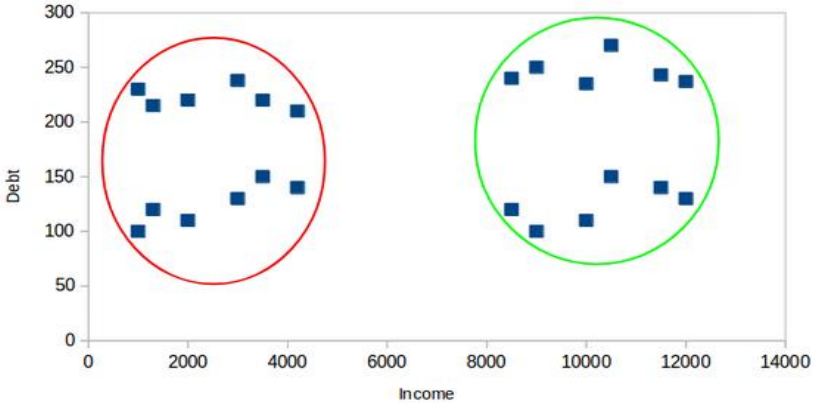
Крок 6. Наступним центроїдом буде той, у якого квадрат відстані $D^2(x)$ є найвіддаленішим від найближчого центроїда

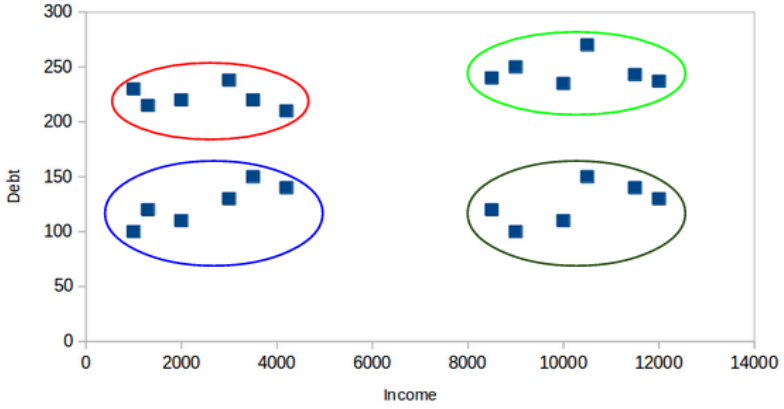
Алгоритм k-середніх++

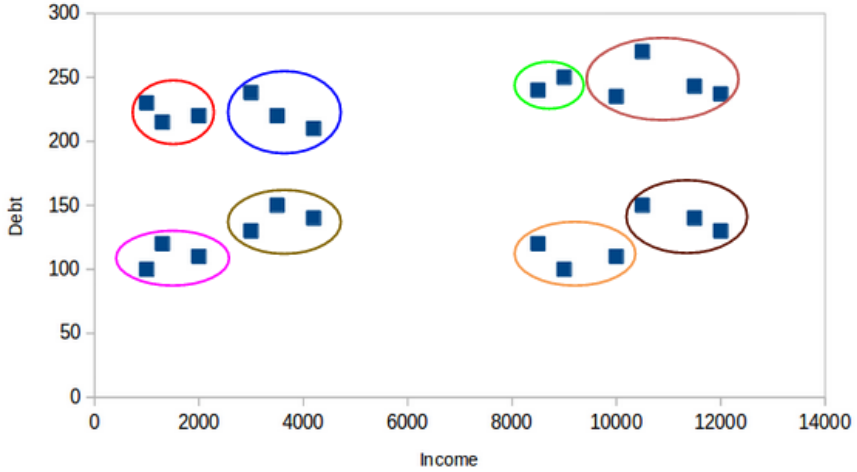
- Крок 7. Застосування алгоритму k-середніх для вже визначених центроідів

А скільки кластерів взагалі потрібно???

Розбиття виборки на різну кількість кластерів



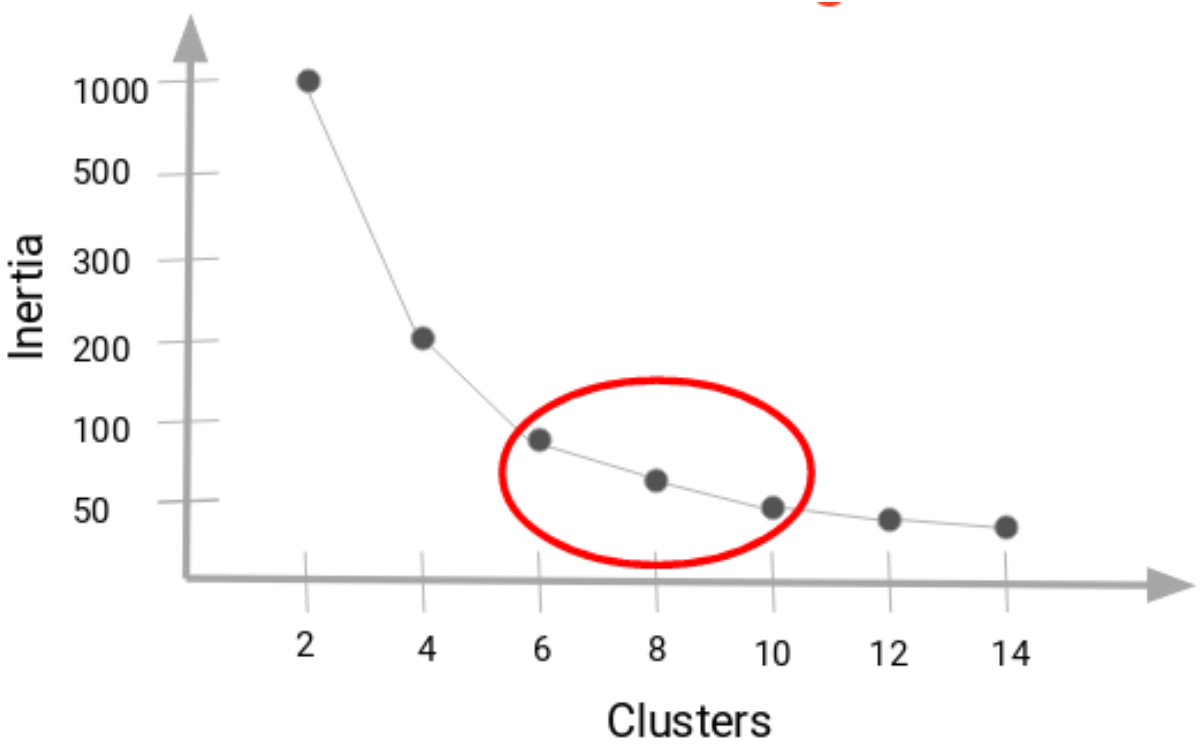




Максимальна кількість кластерів дорівнює кількості точок

А навіщо тоді вона потрібна?

Крива ліктя



Значення кластера, де зменшення значення інерції (або індексу Данна) стає постійним, потрібно вибрати як оптимальне значення кластера для даного набора даних

Висновки

- На відміну від задачі класифікації в задачі кластеризації є невідомим ні кількість категорій, ні ознаки за якими потрібно групувати об’єкти

- Проблема кластеризації не є тривіальною задачею, особливо у випадку даних високої розмірності, з яким необхідно працювати у в більшість реальних застосувань

- Найбільш критичне питання, що виникає під час кластеризації, полягає в тому, щоб зрозуміти, що означає «подібне»

Дякую за увагу!

О.П.Русу

2023 р.