Інтелектуальний аналіз даних. Задачі та методи

Machine Learning,
Data Mining,
Knowledge Discovery

Інформаційний вибух

- Постійне збільшення швидкості та обсягів публікацій (обсягу інформації)
- Генеруються величезні обсяги даних:
  - Банківські, телекомунікаційні та інші види транзакцій.
  - Наукові дані: астрономія, біологія, фізика тощо.
  - Веб, електронна комерція, текстова інформація
- Щоб використовувати накопичені дані, необхідні методи виявлення знань та знаходження прихованих структур у даних, які перетворюють дані в інформацію (Knowledge Discovery).

Сфери застосування IAД

- Наука
  - астрономія, біоінформатика, розробка лікарських засобів, медична діагностика.
- Бізнес та виробництво
  - реклама, управління взаємовідносинами з клієнтами (CRM), інвестиції, виробництво, спорт/індустрія розваг, телекомунікації, електронна комерція, цільовий маркетинг, охорона здоров'я, аналіз кредитних ризиків, залучення та утримання клієнтів, аналіз відмов, прогнозування пікових навантажень, детекція підлог коштів, контроль якості …
- Веб
  - пошукові системи, bots, електронна комерція (рекомендації)…
- Уряд
  - Правоохорона діяльність

Визначення ІАД

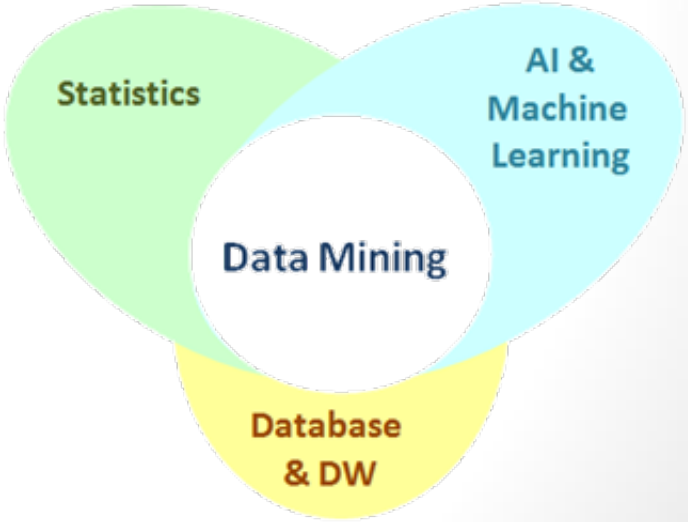
- Інтелектуальний аналіз даних (Data Mining) – це процес знаходження у необробленних даних
  - нових
  - нетривіальних
  - значимих
  - практично корисних
  - доступних інтерпретації
- шаблонів (patterns) чи знань, необхідних для прийняття рішень у різних сферах людської діяльності.

- Термін запроваджено Григорієм П'ятецьким-Шапіро 1989 році.

Споріднені області

- Data Mining - мультидисциплінарна область, що виникла і розвивається на базі таких наук як статистика, розпізнавання образів, штучний інтелект, теорія баз даних та ін.

- **Машинне навчання** - це наука, яка вивчає комп'ютерні алгоритми, що автоматично покращуються під час роботи
- **Штучний інтелект** - науковий напрям, у якого ставляться і вирішуються завдання апаратного чи програмного моделювання видів людської діяльності, традиційно вважаються інтелектуальними



Основні етапи ІАД

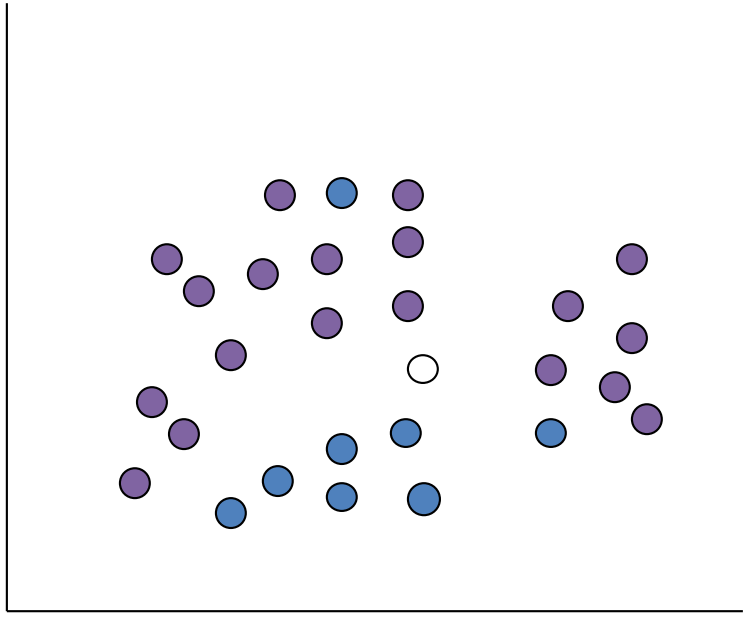
- Розуміння та формулювання завдання аналізу
- Збір даних
- Первинне дослідження даних (розвідувальний аналіз) та їх підготовка до інтелектуального аналізу
- Побудова моделі
- Оцінка моделі
- Інтерпретація моделі людиною
- Аналіз аномалій - виявлення та пояснення аномалій, знайдених у закономірностях

Основні завдання ІАД

- **Класифікація:** передбачення класу об'єкта
- **Регресія:** передбачення числового атрибута об'єкта
- **Кластеризація:** знаходження груп у даних
- **Пошук асоціативних правил:** аналіз споживчих кошиків; асоціації між атрибутами об'єктів
- **Візуалізація:** графічний образ даних
- **Підбиття підсумків** (summarization): опис групи
- **Аналіз відхилень** (винятків, викидів): знаходження відхилень від «норми»
- **Передбачення числових рядів:** передбачення числового значення

Задача класифікації (Classification)

Навчити метод передбачати клас нового об'єкта на основі інформації про раніше класифіковані об'єкти



Передбачувана модель із категоріальним виходом.

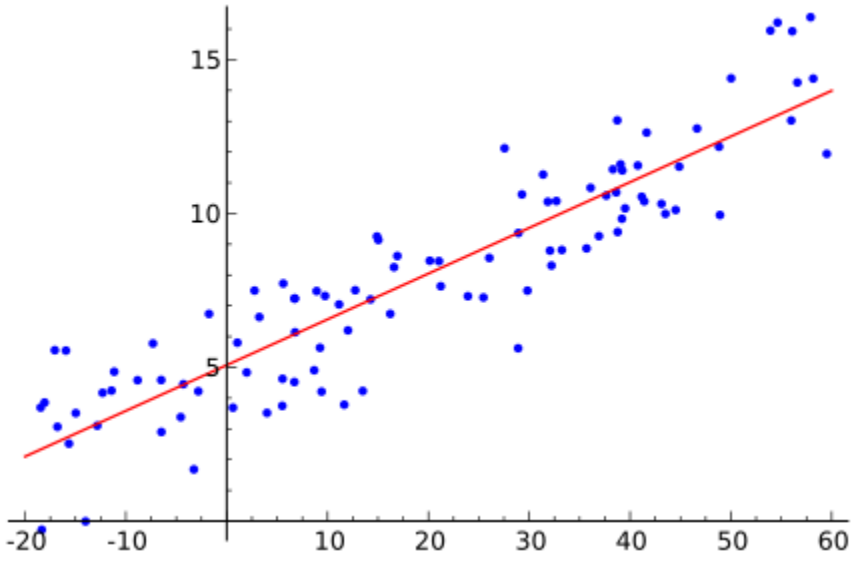
Класи заздалегідь визначені

Маємо набір точок двох класів

Який клас нової точки?

Задача регресії (Regression)

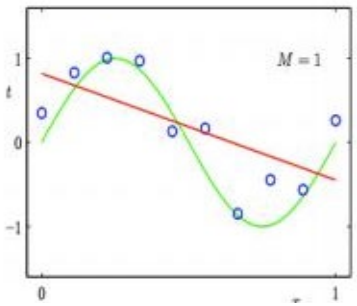
Передбачити значення числового атрибута об'єкта на основі інформації про об'єкти, для яких значення цього атрибуту відоме

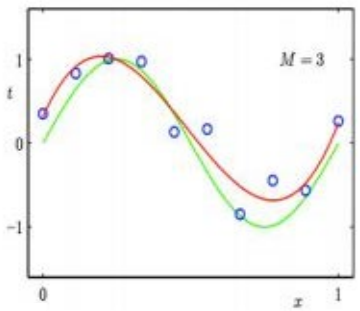


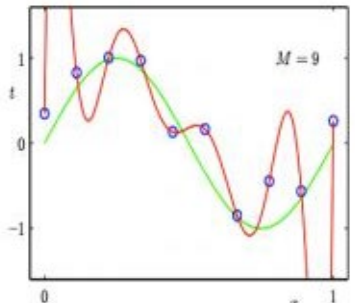
Передбачувана модель з числовим виходом

Перенавчання

Регресія



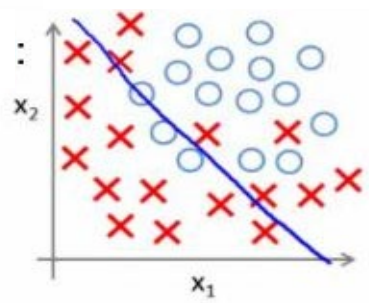


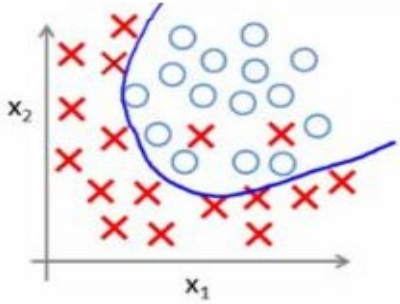


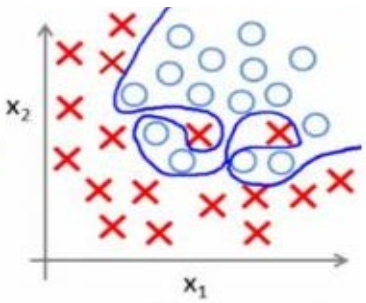
Модель не може «знайти» шаблон

Модель відображає шум у даних

Класифікація :







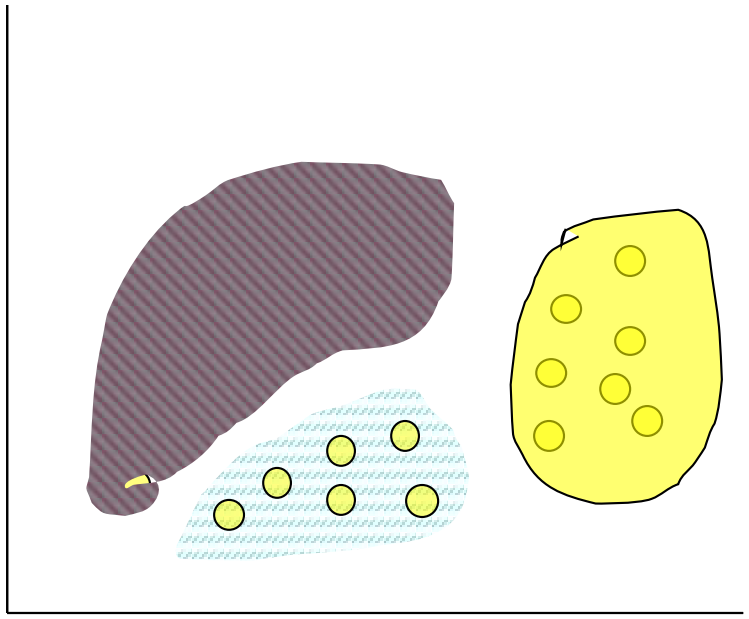
Занадто проста модель; не може пояснити дані

Занадто складна модель

Бритва Оккама: з двох моделей, які дають однакову точність, вибираємо простішу

Задача кластеризації (Clustering)

Розбити об'єкти на групи так, щоб у групі об'єкти були схожі, а об'єкти з різних груп не схожі



Описова модель. Угруповання об'єктів, пошук найбільш схожих однорідних груп

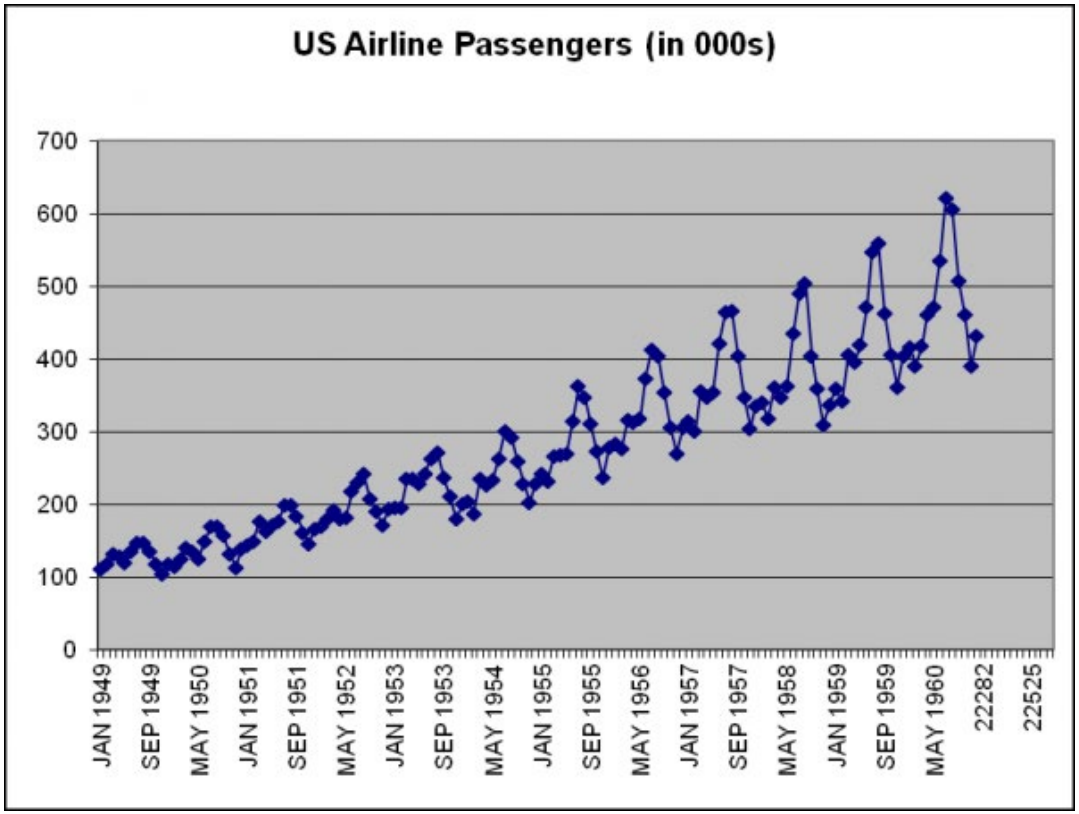
Пошук асоціативних правил (Assosiations)

Пошук цікавих асоціацій серед спостережень (які спостереження слідують разом), асоціації між атрибутами

Сиквенційний аналіз (важлива також послідовність спостережень/подій)

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Передбачення числових рядів (Time-series forecasting)



Передбачення майбутніх значень числового ряду