01MLIntroduction

November 10, 2018

1 Roadmap

1.1 DataSource

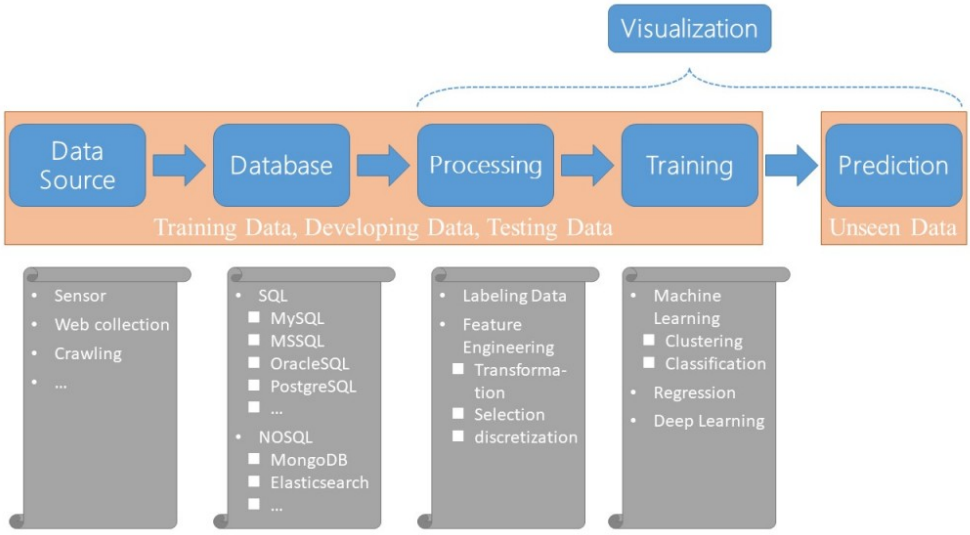
1. 物聯網感測器: 時間序列、空間資料
2. 網頁使用者資料: 很多元
3. 網頁爬蟲: 圖片、文字

1.2 Database

item	SQL(關聯式資料庫)	NoSQL(非關聯式資料庫)
特性	易於紀錄物件關聯性	讀寫效能加，可 scale up
適用情境	物件關聯複雜	物件關聯簡單 (單一 table 可以儲存)
舉例	公司內部員工管理系統、訂單管理系統	爬蟲、IOT 資料

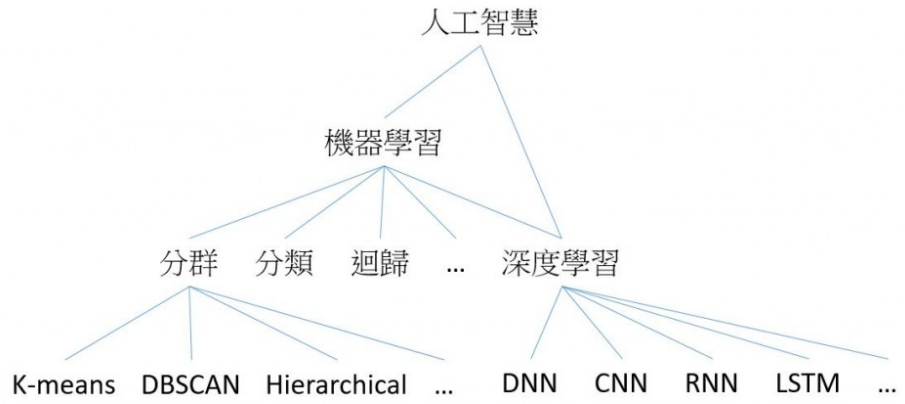
1.3 Processing

1. Data Cleaning(資料清理)
    - Fill in missing values(填入遺漏值)
    - Identify or remove outliers(辨識並移除離群值)
    - Resolve inconsistencies(資料不一致)
2. Integration(資料整合)
    - Join tables
    - Different data sources: csv, json, xml, html, api, datebases
3. Transformation(資料轉換)
    - Normalization(正規化)
    - Aggregation(加總)
    - Feature Engineering(特徵值篩選)
4. Data discretization(資料切片)
    - data reduction(降維)



Roadmap

1.4 Training



1. 分群 (Clustering) - K-means - DBSCAN - Hierarchy
2. 分類 (Classfication)
    - KNN
    - Decision Tree
    - Logistic Regression
    - Bayesian
    - SVM
    - Random Forest
    - XGBoost
    - lightGBM
3. 檢索與推薦 (Information Retrieval)

- Vector Space Model
- BM25
- Bayesian Model

4. 推薦 (Recommendation)

- Collaborative Filtering
- Content-based Rencommendation

1.5 Production

1. Deployment
2. Hardware
3. System

1.6 議題

1. 資料取得有哪幾種手段? 資料各有甚麼特性?
2. 資料庫有分哪兩種類型? 請分別列舉兩個資料庫品牌? 兩種類型的資料庫分別有什麼好處?
3. 甚麼樣的資料需要標記?
4. 應採取什麼樣的標記手段?
5. 標記是否為正確答案?