Day 1

機器學習概論

資料介紹與評估指標

出題教練

游為翔 / 杜靖愷

知識地圖 機器學習概論 資料介紹與評估資料

CUPOY

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

前處理 Processing → 探索式數據分析 Exploratory Data Analysis → 特徵工程 Feature Engineering → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

- 分群 Clustering
- 降維 Dimension Reduction

機器學習概論 Introduction of Machine learning

- **機器學習的限制**
- 機器學習可分析的幾類問題
- 機器學習流程
- 數據分析流程

本日知識點目標

準備進入資料科學領域的概念與流程與關鍵

學習路徑

CUPOY

挑一個有趣的問題, 並從解決一個簡單的問題開始

試圖改進你的原始解決方案並從中學習 (如代碼優化、速度優化、演算法優化)

不斷在一系列不同的問題上反覆練習

在這個題目上做一個原型解決方案 (prototype solution)

紀錄是一個好習慣，試著紀錄並分享你的解決方案歷程

認真地參與一場比賽

首次面對資料，我們應該思考哪些問題？

CUPOY

Questions	Explanation	Examples
為什麼這個問題重要？ (Why it is important)	A. 好玩 B. 企業的核心問題 C. 公眾利益 / 影響政策方向 D. 對世界很有貢獻	A. 預測生存 (吃雞) 遊戲誰可以活得久, PUBG B. 用戶廣告投放, ADPC C. 停車方針, 計程車載客優化 D. 肺炎偵測
資料從何而來？ (Where do data come from)	• 來源與品質息息相關 • 根據不同資料源，我們可以合理的推測/懷疑異常資料異常的理由與頻率	資料來源如： 網站流量、購物車紀錄、網路爬蟲、格式化表單、Crowdsourcing、紙本轉電子檔
資料的型態是什麼？ (What are they)	A. 結構化資料需要檢視欄位意義以及名稱 B. 非結構化資料需要思考資料轉換與標準化方式	A. 結構化：數值, 表格, ...etc B. 非結構化：圖像、影片、文字、音訊, … etc
我們可以回答什麼問題 ？ 問題：指標 (What is our goal)	每個問題都應該要可以被驗證 → 有一個可供衡量的數學評估指標 (Evaluation Metrics)	常見的衡量指標如： 分類問題：正確率, AUC, MAP, ...etc 迴歸問題：MAE, RMSE, ...etc 補充資料：衡量指標

範例一：我們應該要 / 可以回答什麼問題？

CUPOY

生存 (吃雞) 遊戲

- 玩家排名：平均絕對誤差 (Mean Absolute Error, MAE)
- 怎麼樣的人通常活得久/不久 (如加入遊戲的時間、開始地點、單位時間內取得的資源量, ...) → 玩家在一場遊戲中的存活時間：迴歸 (Mean Squared Error, MSE)

範例二：我們應該要 / 可以回答什麼問題？

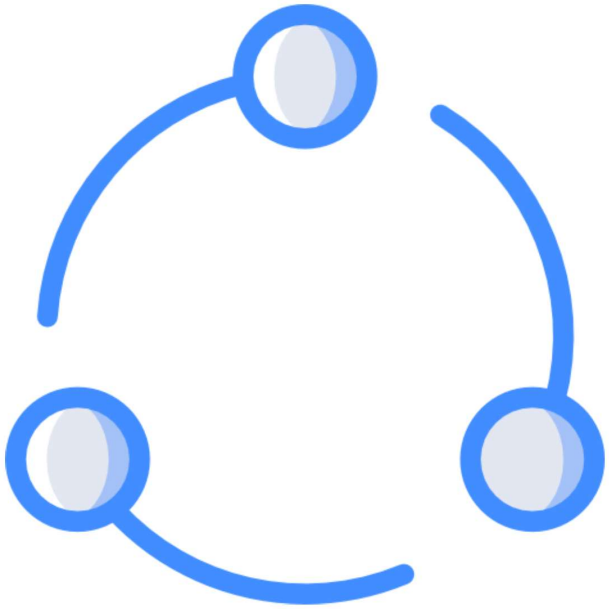
CUPOY

廣告投放

- 不同時間點的客群樣貌如何 → 廣告點擊預測 → 預測哪些受眾會點擊或行動：Accuracy / Receiver Operating Curve, ROC
- 哪些素材很好/不好 → 廣告點擊預測 → 預測在版面上的哪個廣告會被點擊：ROC / MAP@N (eg. MAP@5, MAP@12)

重要知識點複習

CUPOY



- 初入資料科學的探索流程
- 找到問題 → 初探 → 改進 → 分享 → 練習 → 實戰
- 面對問題需要思考的關鍵點
  - 為什麼這個問題重要
    - 資料從何而來
    - 資料的型態是什麼
    - 回答問題的關鍵指標是什麼

解題時間
It's Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題