Day 6

資料清理數據前處理

EDA與Outlier檢查

出題教練

游為翔 / 杜靖愷

知識地圖 機器學習前處理 Outlier 及處理

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

前處理 Processing

資料讀取 → 格式調整 → 填補缺值 → **去離群值** → 特徵縮放

本日知識點目標

CUPOY

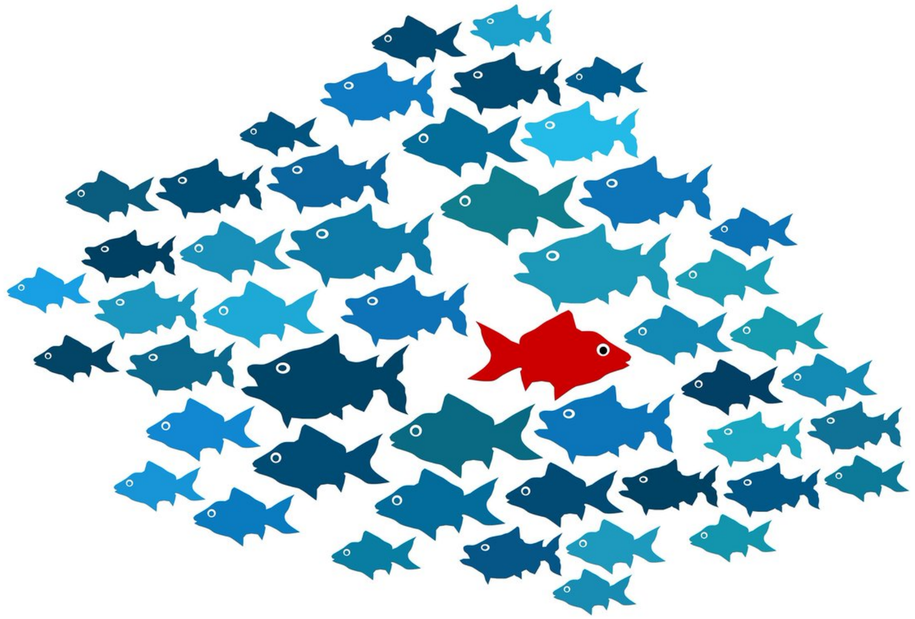
目標
知識點

了解什麼是例外值 (outlier)

獲得
知識點

完成今日課程後你應該可以了解

- 學會如何透過資料探勘方法找到例外值



圖片來源: Sergio Santoyo

Dell電腦標價錯誤

1

異常值 (Outliers) 出現的可能原因

1. 所以未知值，隨意填補 (約定俗成的代入)
   如年齡 = -1 或 999, 電話是 0900-123-456

2. 可能的錯誤紀錄/手誤/系統性錯誤
   如某本書在某筆訂單的銷售量 = 1000 本

2

檢查 Outliers 的流程與方法

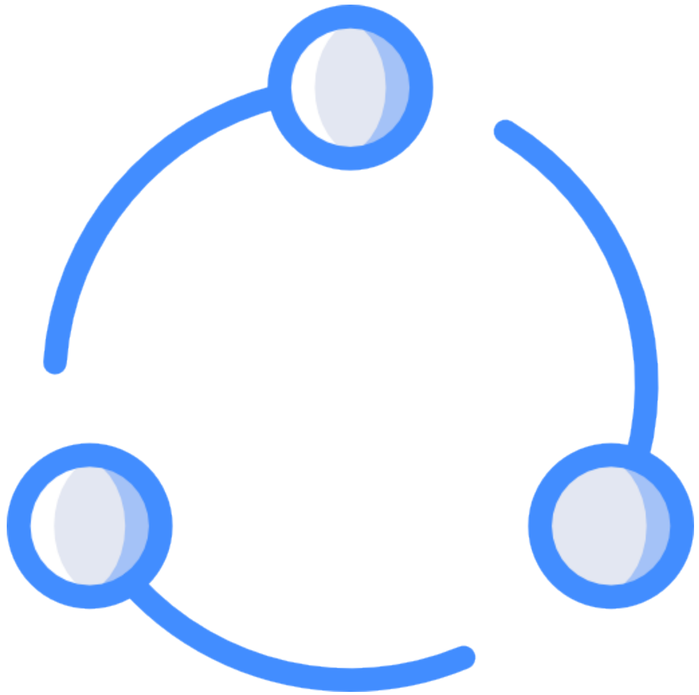
- 盡可能確認每一個欄位的意義 (但有些競賽資料不會提供欄位意義)
- 透過檢查數值範圍 (五值、平均數及標準差) 或繪製散點圖 (scatter)、分布圖 (histogram) 或其他圖檢查是否有異常。

3

對 Outliers 的處理方法

- 新增欄位用以紀錄異常與否
- 填補 (取代)
- 視情況以中位數, Min, Max 或平均數填補(有時會用 NA)

重要知識點複習



- 檢查異常值的方法
  - 統計值：如平均數、標準差、中位數、分位數
  - 畫圖：如直方圖、盒圖、次數累積分布等
- 處理異常值
  - 取代補值：中位數、平均數等
  - 另建欄位
  - 整欄不用

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題