Day 20 特徵工程

數值型特徵-去除離群值

出題教練 陳明佑

知識地圖 特徵工程 數值型特徵 - 去除離群值

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

特徵工程 Feature Engineering

概論

- **數值型特徵**
- **類別型特徵**
- **時間型特徵**

→ 填補缺值 →
- 去離群值 → 去偏態 → 特徵縮放
- 類別型特徵處理
- 時間型特徵處理

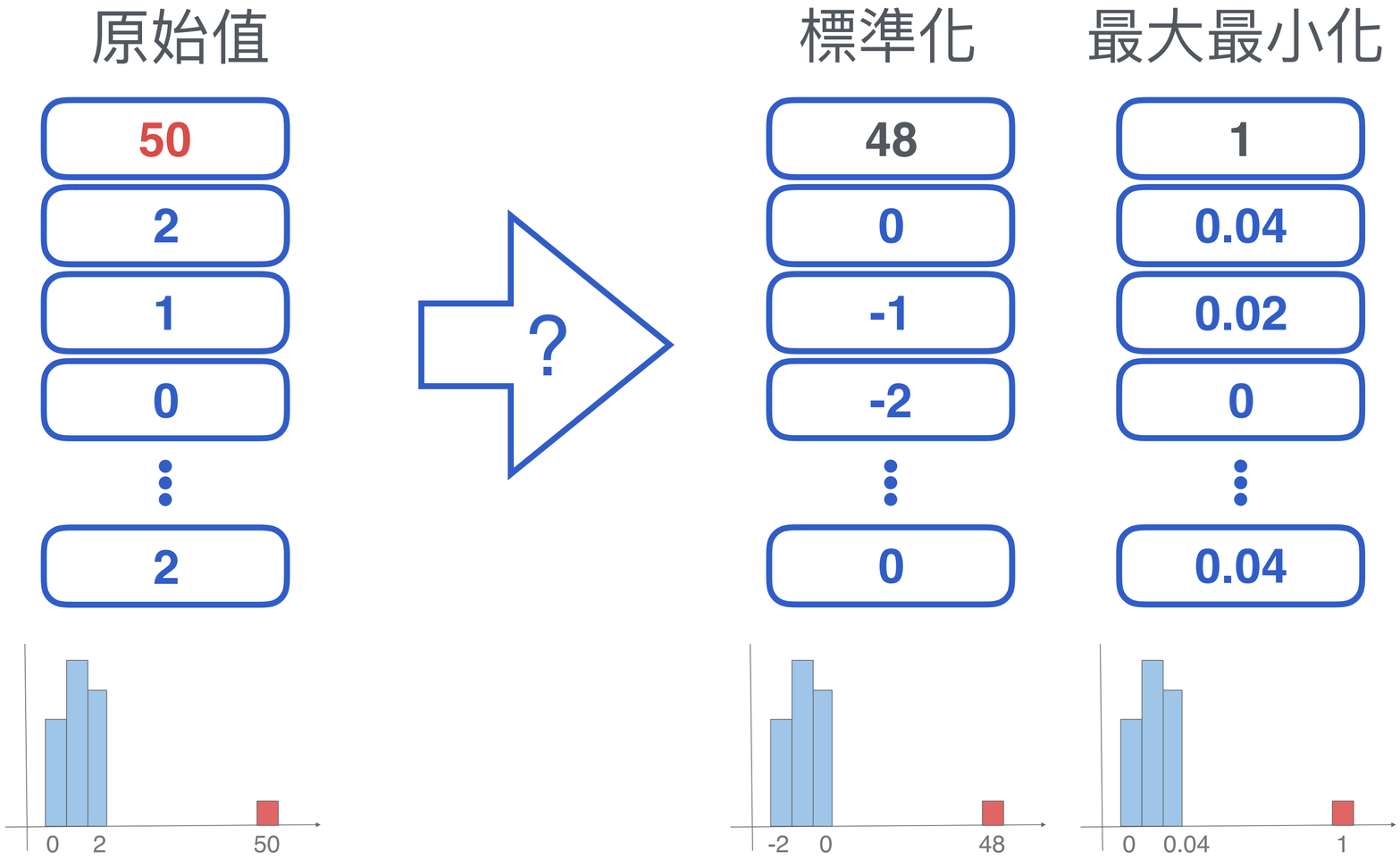
→ 特徵組合 / 特徵篩選 → 特徵評估

本日知識點目標

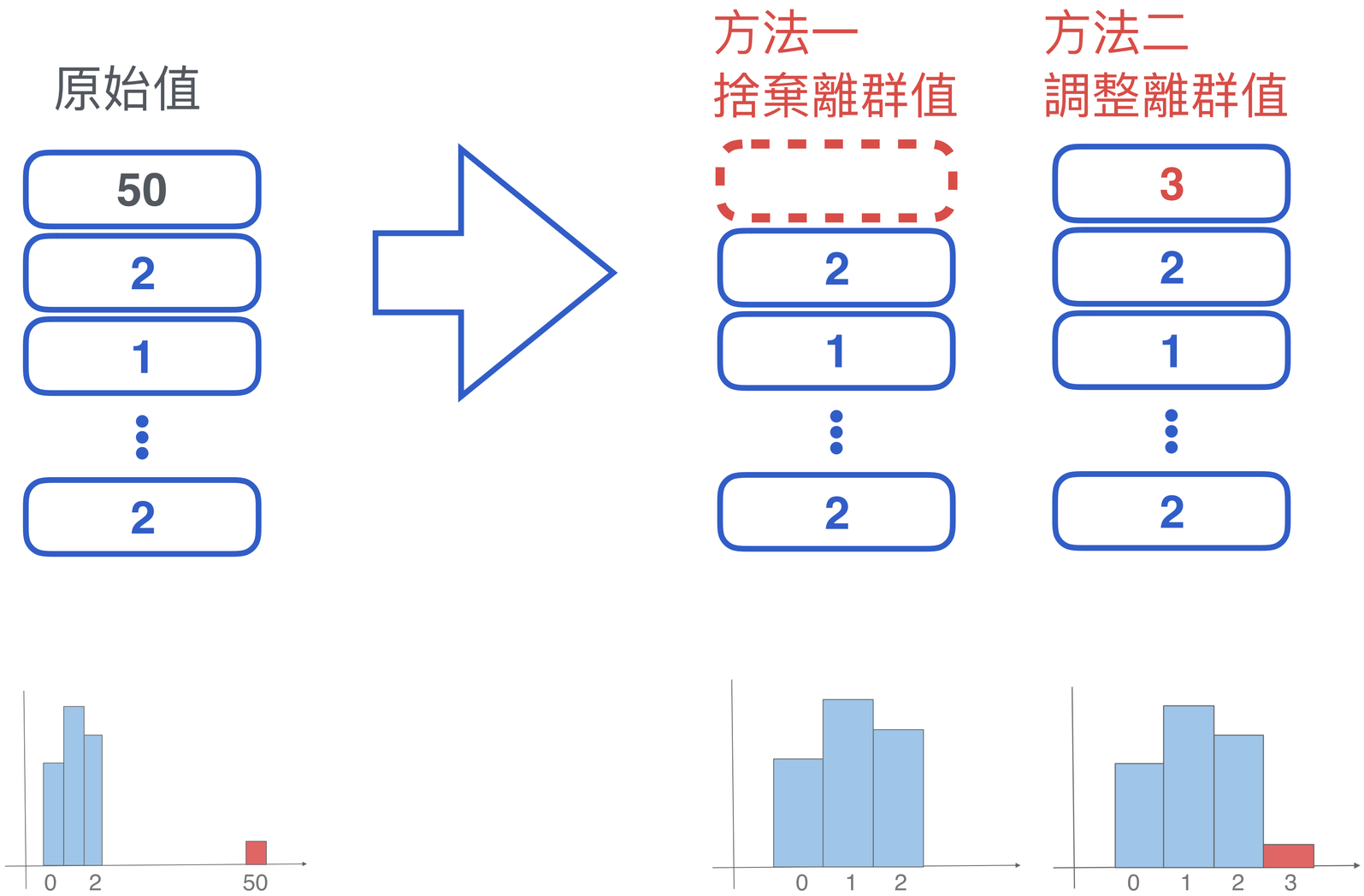
- 初步理解什麼是離群值? 出現時會有什麼問題?
- [複習] 離群值處理，會有哪些優缺點?

去除離群值 (1 / 2)

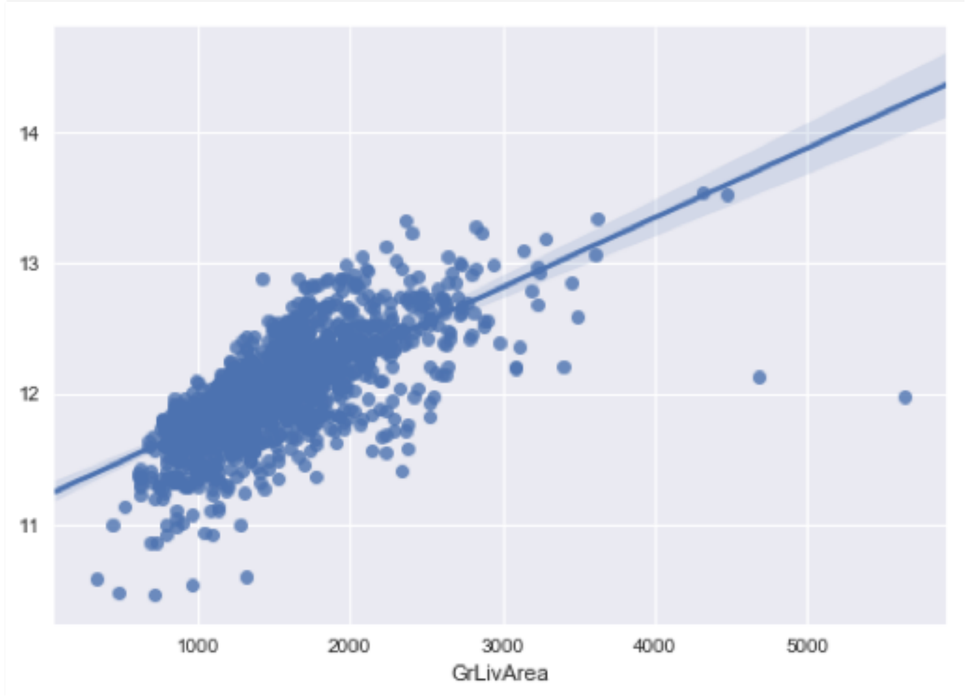
如果只有少數幾筆資料跟其他數值差異很大，標準化無法處理

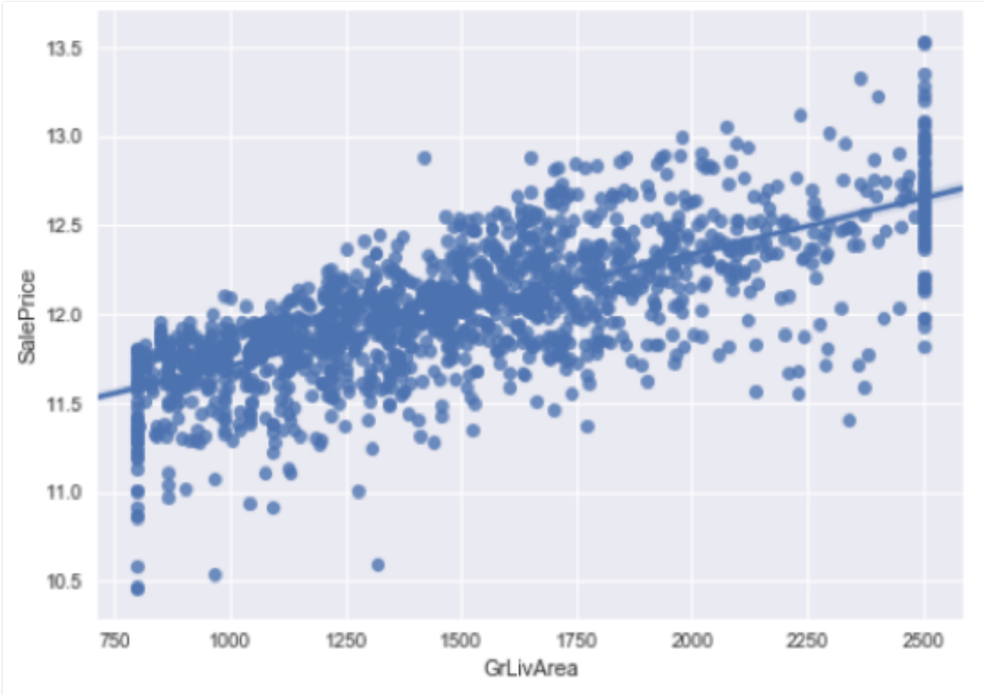


去除離群值 (2 / 2)



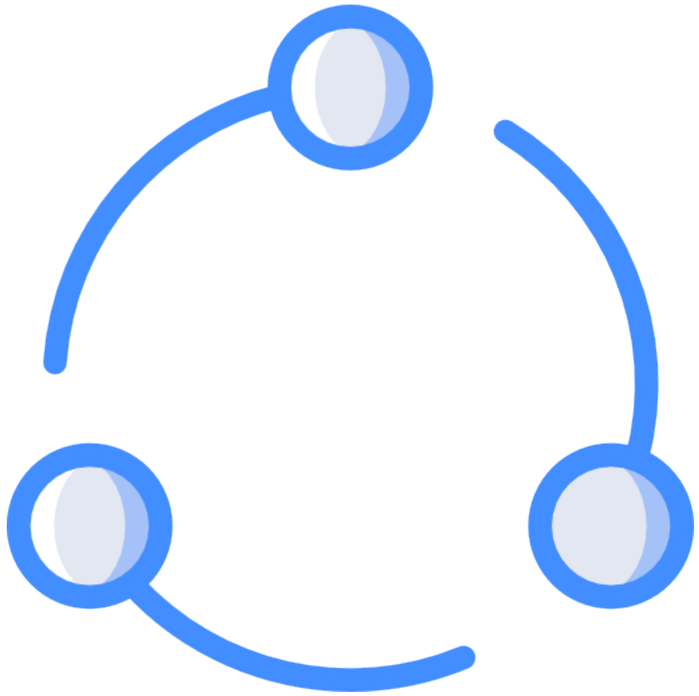
複習：去除離群值





數值型欄位，有時會有與其他數值差距很大的離群值存在
只要離群值數量夠少，除去離群值，將可使得模型預測較為準確
(參考圖中的迴歸直線以及今日範例)

重要知識點複習



- 離群值是與正常數值**偏離較遠**的數值群，如果不處理則**特徵縮放**(標準化 / 最小最大化)就會出現很大的問題
- 處理離群值之後，好處是剩餘資料中模型較為**單純且準確**，壞處是有可能**刪除掉重要資訊**，因此刪除前最好能先了解該數值會離群的可能原因

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題