Day 22 特徵工程

類別型特徵 - 基礎處理

出題教練 陳明佑

知識地圖 特徵工程 類別型特徵 - 基礎處理

CUPOY

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

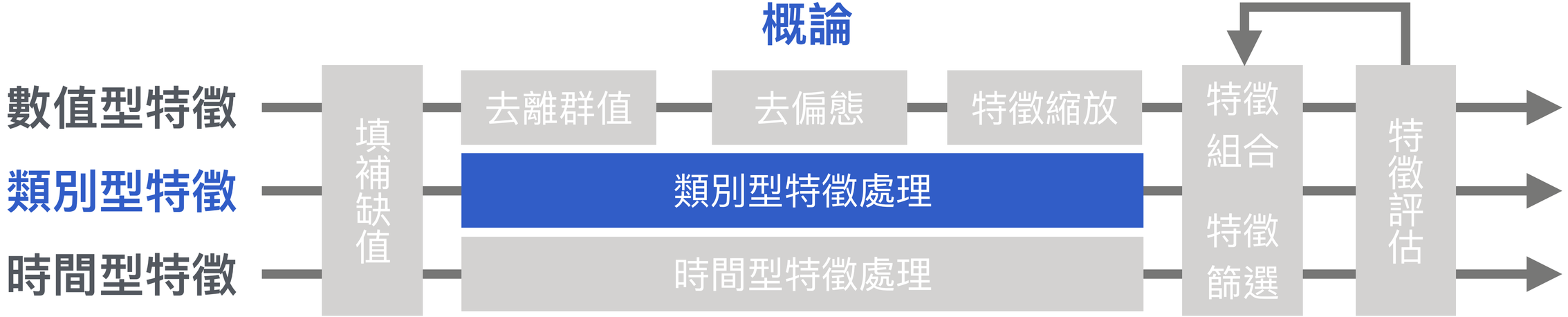
前處理 Processing → 探索式數據分析 Exploratory Data Analysis → **特徵工程 Feature Engineering** → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

- 分群 Clustering
- 降維 Dimension Reduction

特徵工程 Feature Engineering

概論

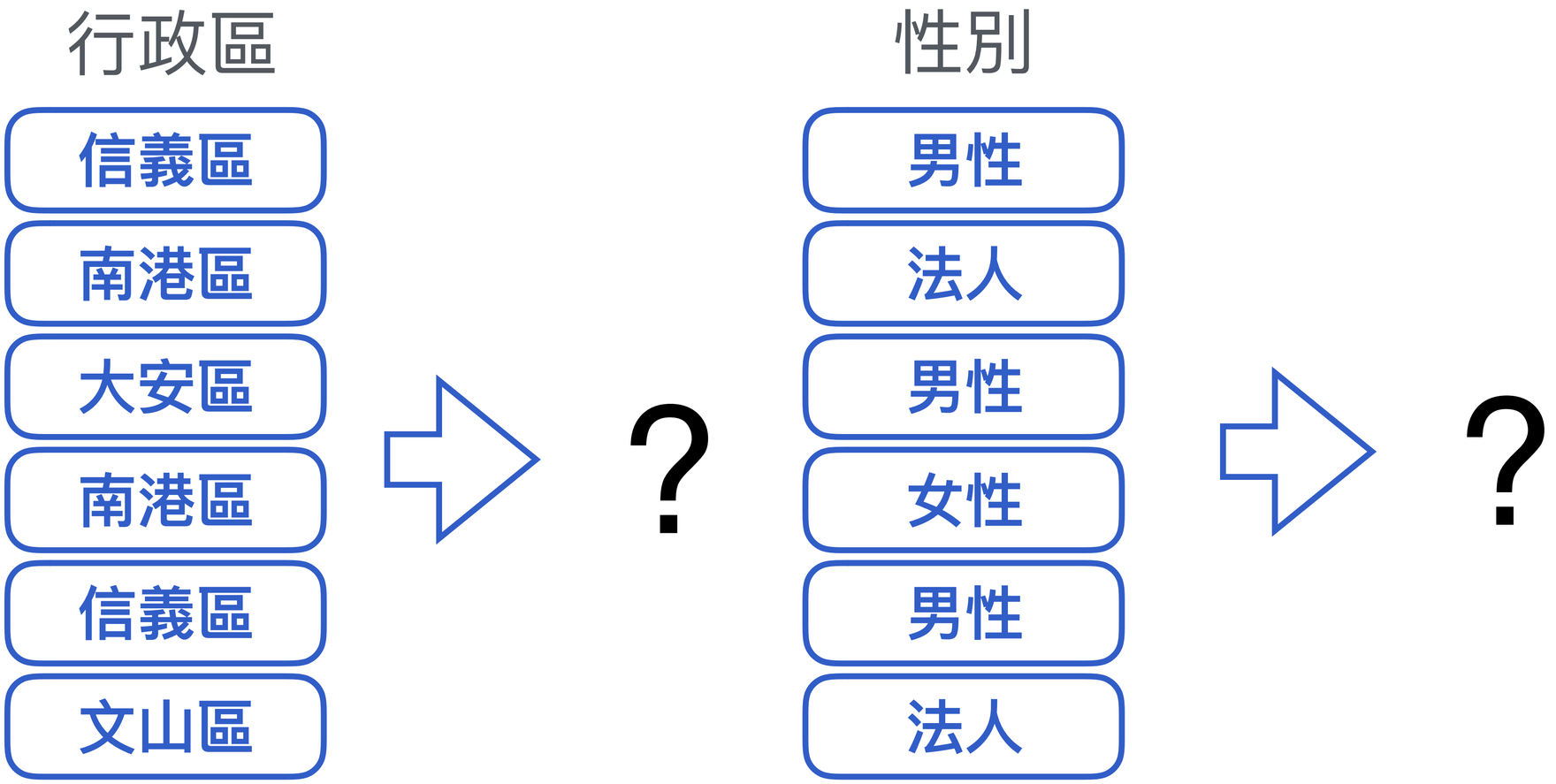


本日知識點目標

- 類別型特徵有哪兩種基礎編碼方式?
- 兩種基礎編碼方式中，哪一種比較常用? 為什麼?
- 在什麼情況下，比較適合獨熱編碼?

類別型特徵的處理

前面提過：特徵工程是事實到對應分數的轉換
請先回憶一下，已學過哪些類別型特徵的轉換方式，您是否可以想到其他的轉換方法?



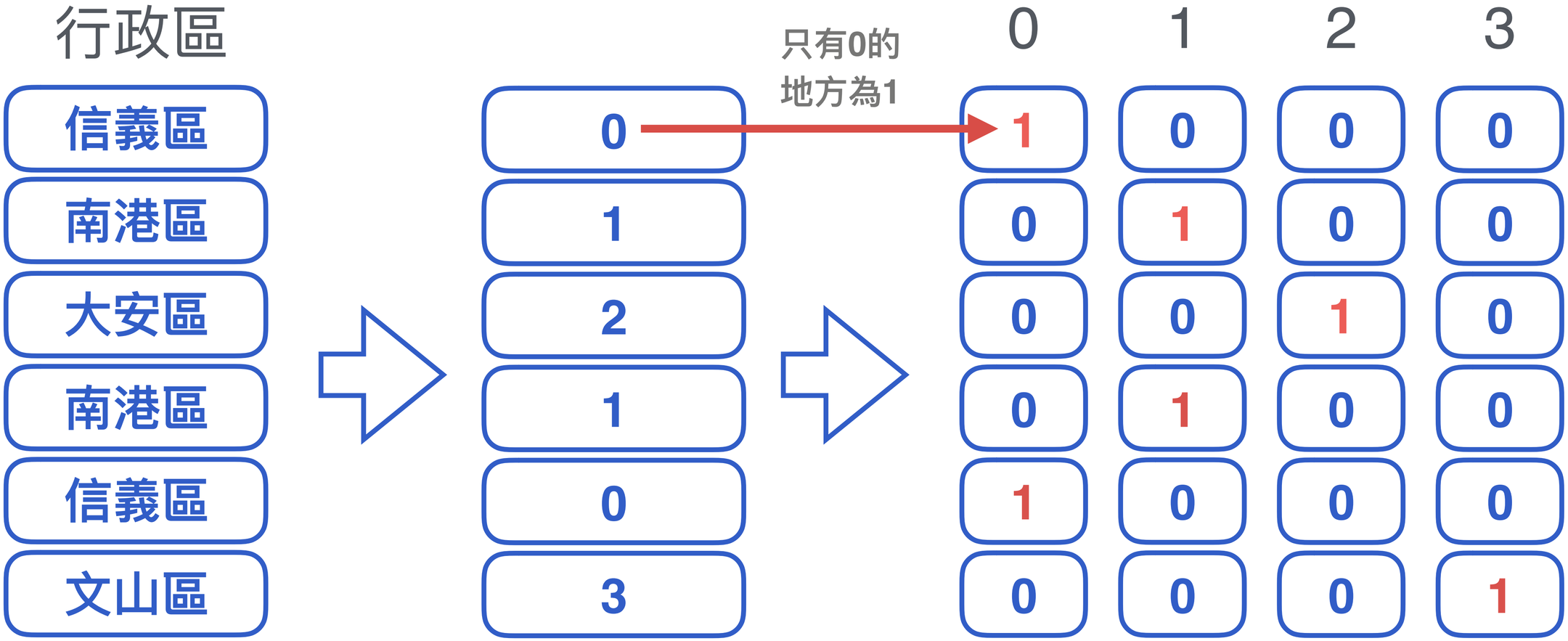
基礎編碼 1：標籤編碼（Label Encoding）

CUPOY

- 類似於流水號，依序將新出現的類別依序編上新代碼，已出現的類別編上已使用的代碼
- 確實能轉成分數，但缺點是分數的大小順序沒有意義

基礎編碼 2：獨熱編碼 (One Hot Encoding)

- 為了改良數字大小沒有意義的問題，將不同的類別分別獨立為一欄
- 缺點是需要較大的記憶空間與計算時間，且類別數量越多時越嚴重



標籤編碼 / 獨熱編碼的比較

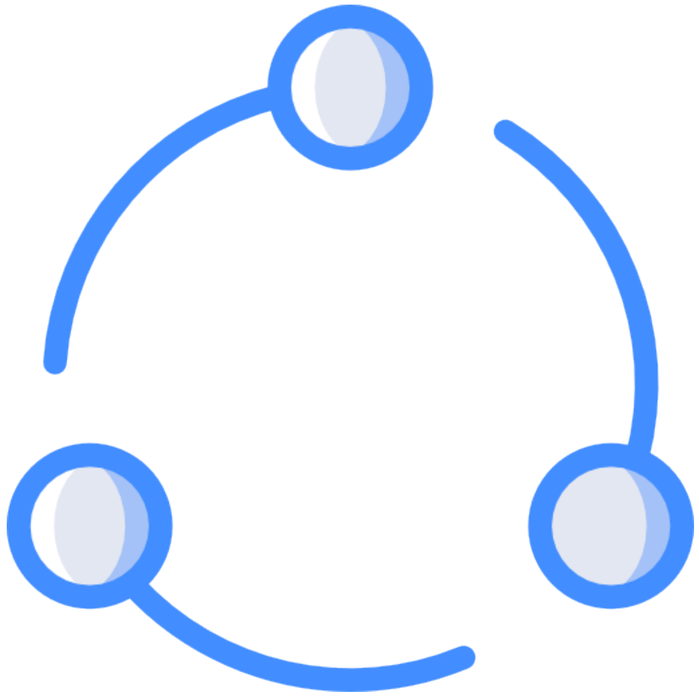
CUPOY

	大小有無意義	儲存空間/計算時間	適用模型
標籤編碼 Label Encoding	無意義	小	樹狀模型
獨熱編碼 One Hot Encoding	有意義	較大	非樹狀模型

綜合建議

- 類別型特徵建議預設採用標籤編碼
- 除非該特徵重要性高，且可能值較少(獨熱編碼時負擔較低) 時，才應考慮使用獨熱編碼

重要知識點複習



- 類別型特徵有**標籤編碼** (Label Encoding) 與**獨熱編碼** (One Hot Encoding) 兩種基礎編碼方式
- 兩種編碼中標籤編碼比較常用
- 當**特徵重要性高**，且**可能值較少**時，才應該考慮獨熱編碼

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題