Day 24 特徵工程

類別型特徵 - 其他進階處理

計數編碼（Counting）

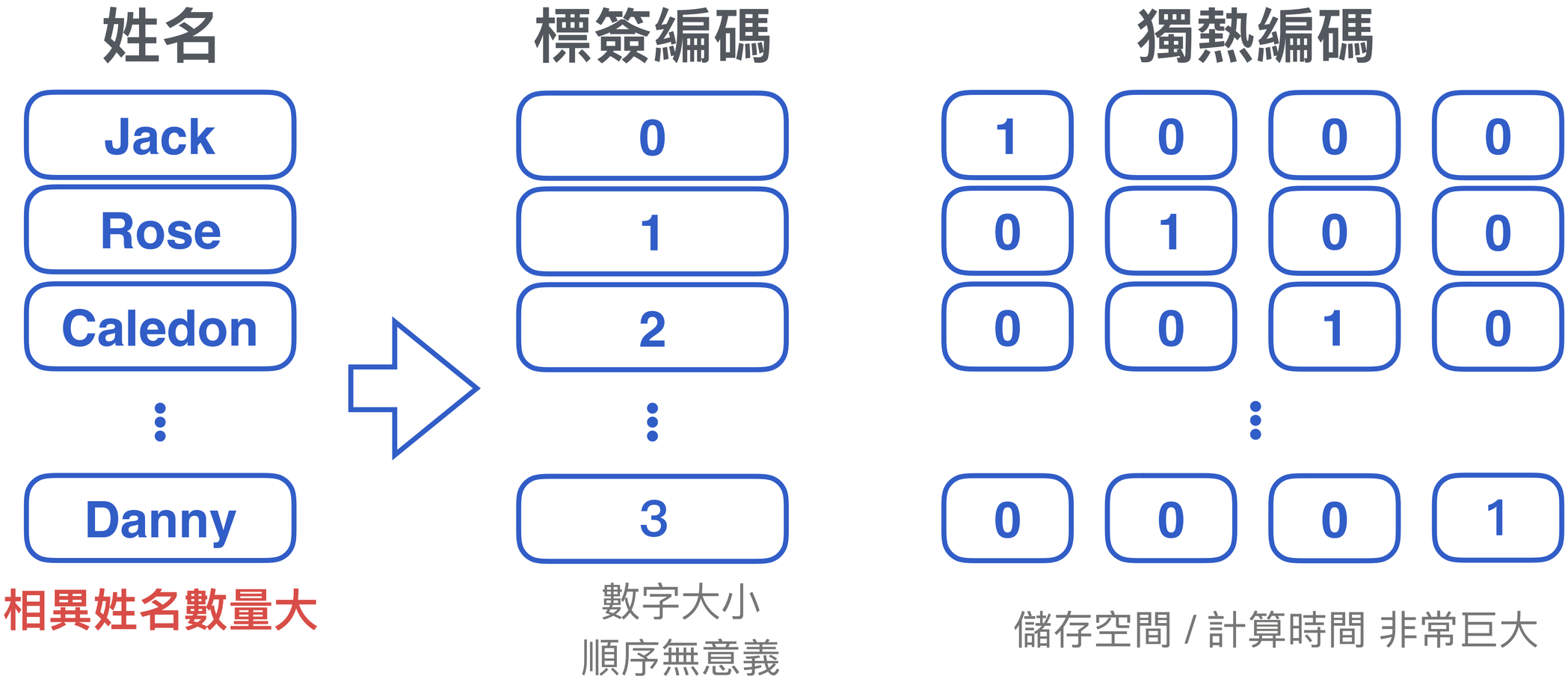
如果類別的目標均價**與類別筆數呈正相關**（或負相關），也可以將**筆數本身**當成特徵
例如：購物網站的消費金額預測

*自然語言處理時，字詞的**計數編碼**又稱**詞頻**，本身就是一個很重要的特徵

特徵雜湊 (Feature Hash) (1 / 2)

類別型特徵最麻煩的問題：**相異**類別的**數量非常龐大**, 該如何編碼?

*舉例：鐵達尼生存預測的旅客姓名



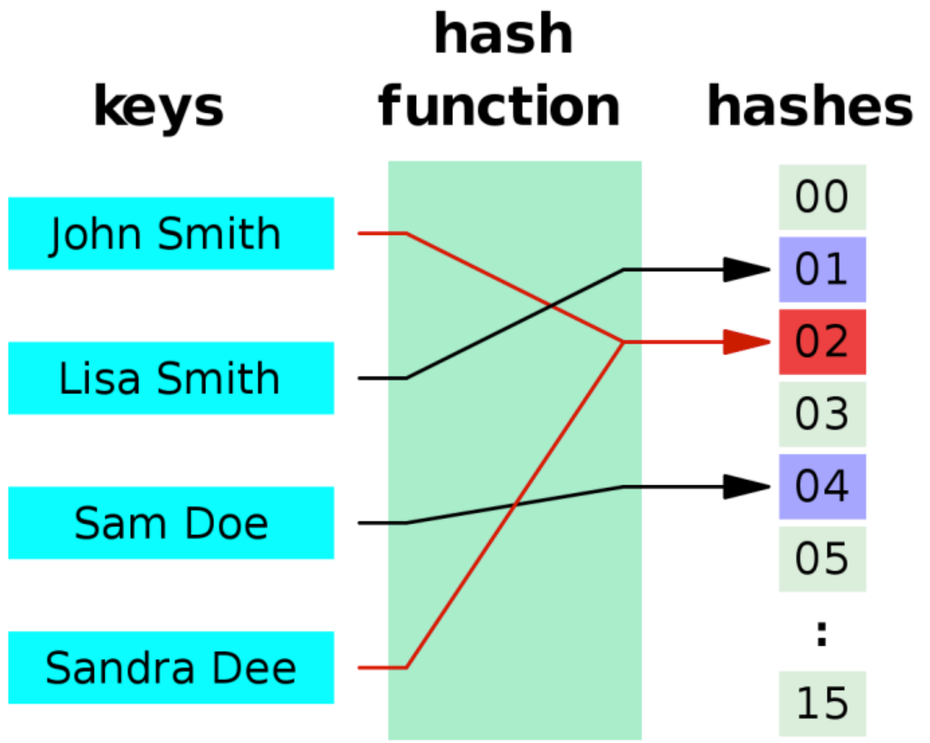
特徵雜湊 (Feature Hash) (2 / 2)

這個問題沒有很好的通用解法...只能採折衷方案或個別情況解決

特徵雜湊

- 特徵雜湊是一種折衷方案
- 將類別由雜湊函數定應到一組數字

- 調整雜湊函數**對應值的數量**
- 在**計算空間/時間**與**鑑別度**間取折衷
- 也提高了訊息密度, 減少無用的標籤



圖片來源：維基百科 https://en.wikipedia.org/wiki/Hash_function

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題