

Day 23

特徵工程

類別型特徵 - 均值編碼



出題教練

陳明佑



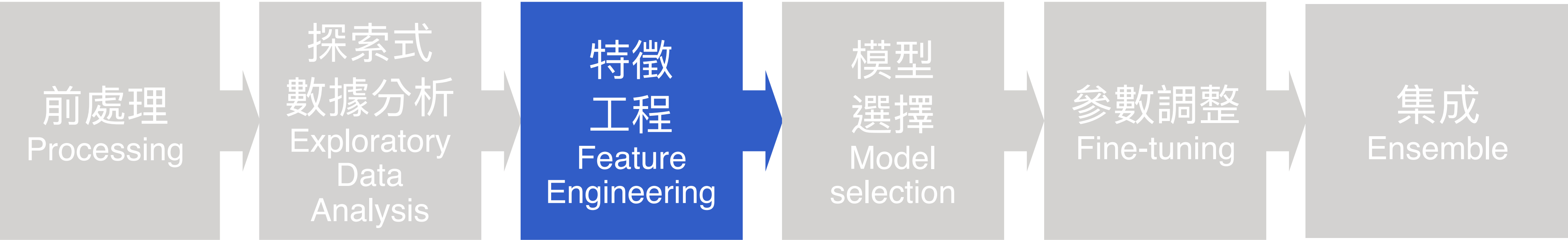
本日知識點目標

- 知道當類別特徵與目標明顯相關時，該用什麼編碼方式
- 知道均值編碼可能有什麼問題
- 知道應該使用何種方式修正均值編碼的問題

知識地圖 特徵工程 類別型特徵 - 均值編碼

機器學習概論 Introduction of Machine Learning

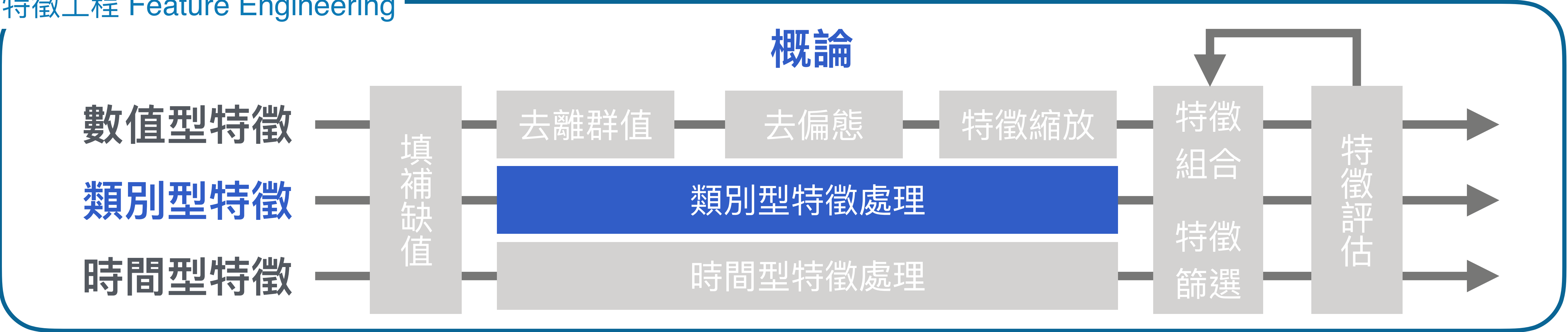
監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



特徵工程 Feature Engineering



均值編碼 (1 / 2)

額外線索：如果類別特徵看起來與目標值有顯著相關，應該如何編碼？

行政區

大安區

南港區

大安區

大安區

南港區

文山區

房產價位

4500萬

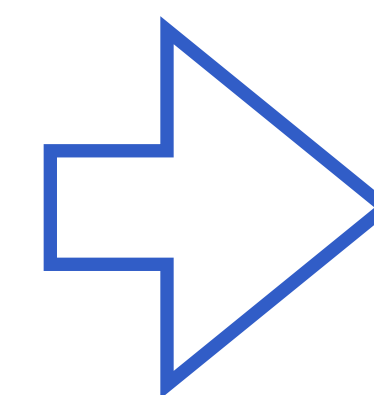
1500萬

3500萬

2500萬

1800萬

2000萬

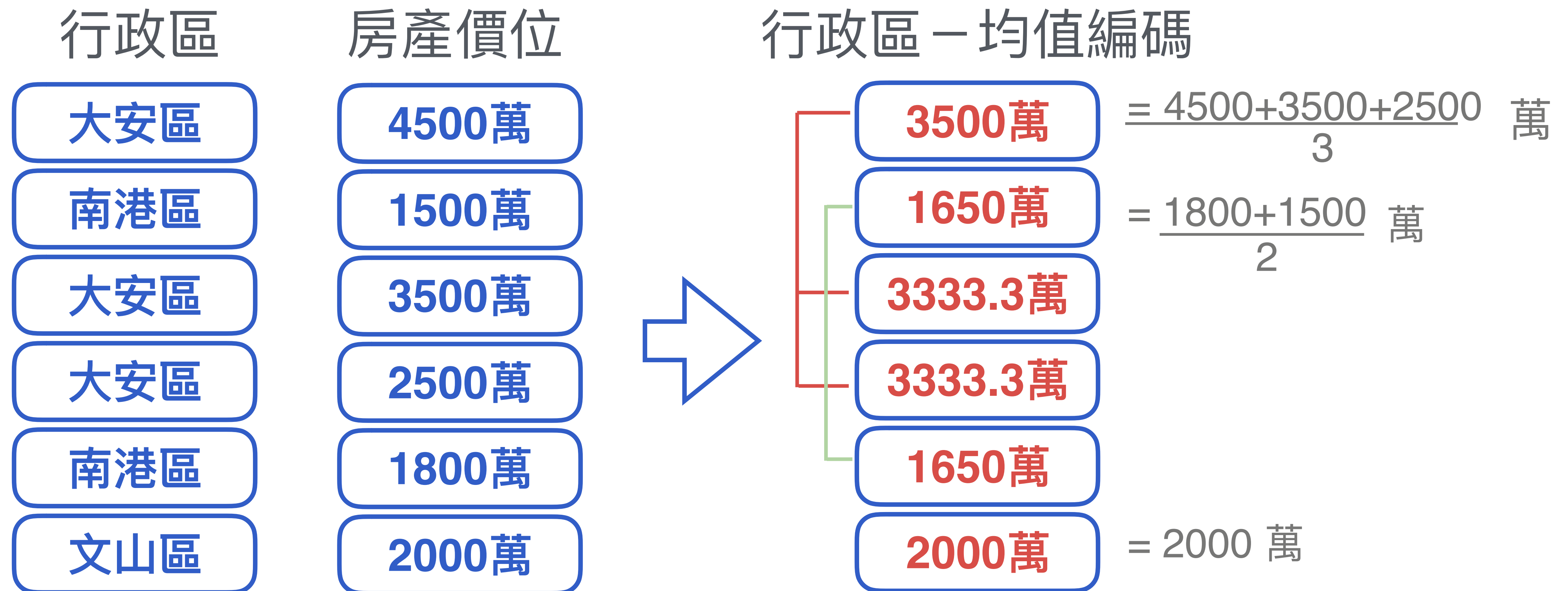


?

均值編碼 (2 / 2)

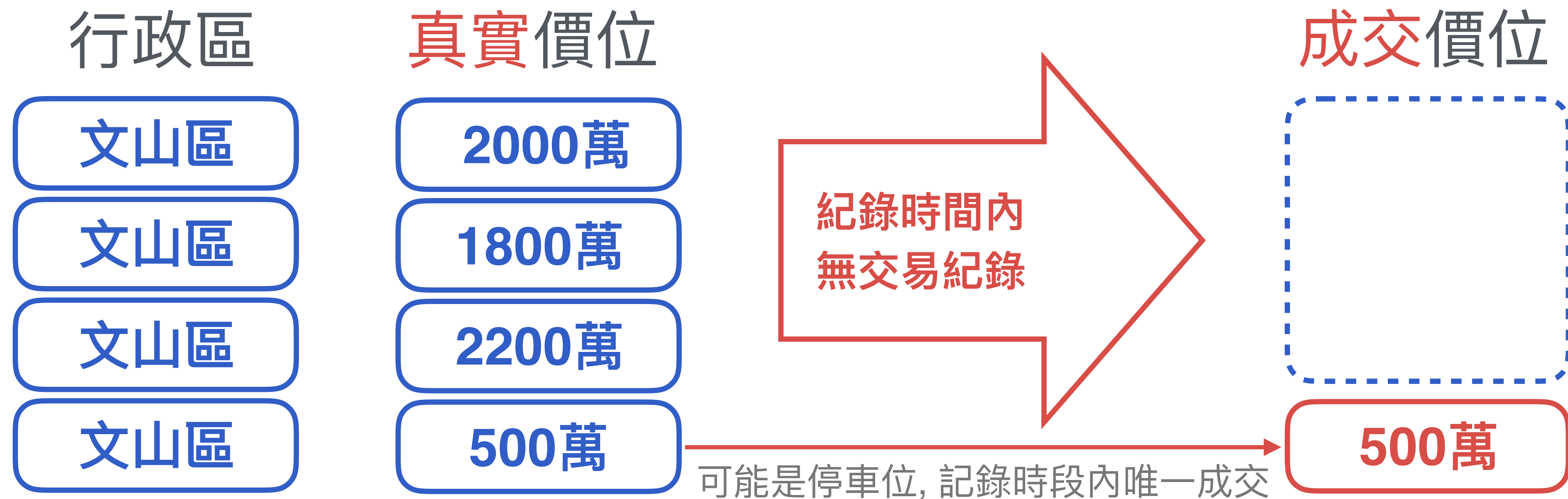
均值編碼 (Mean Encoding)：使用目標值的平均值，取代原本的類別型特徵

*在部分模型中，使用均值編碼作為類別型特徵預設編碼方式



平滑化 (Smoothing) (1 / 2)

如果交易樣本非常少, 且剛好抽到極端值, 平均結果可能會有誤差很大



想想看：這個問題如何解決？

平滑化 (Smoothing) (2 / 2)

因此, 均值編碼還需要考慮紀錄筆數, 當作可靠度的參考

行政區	價位均值	記錄筆數	可靠度
文山區	500萬	X1	低
大安區	3000萬	X100	高
南港區	1700萬	X10	中

- 當平均值的可靠度低時, 我們會傾向相信全部的總平均
- 當平均值的可靠度高時, 我們會傾向相信類別的平均
- 依照紀錄筆數, 在這兩者間取折衷

平滑化公式與小提醒

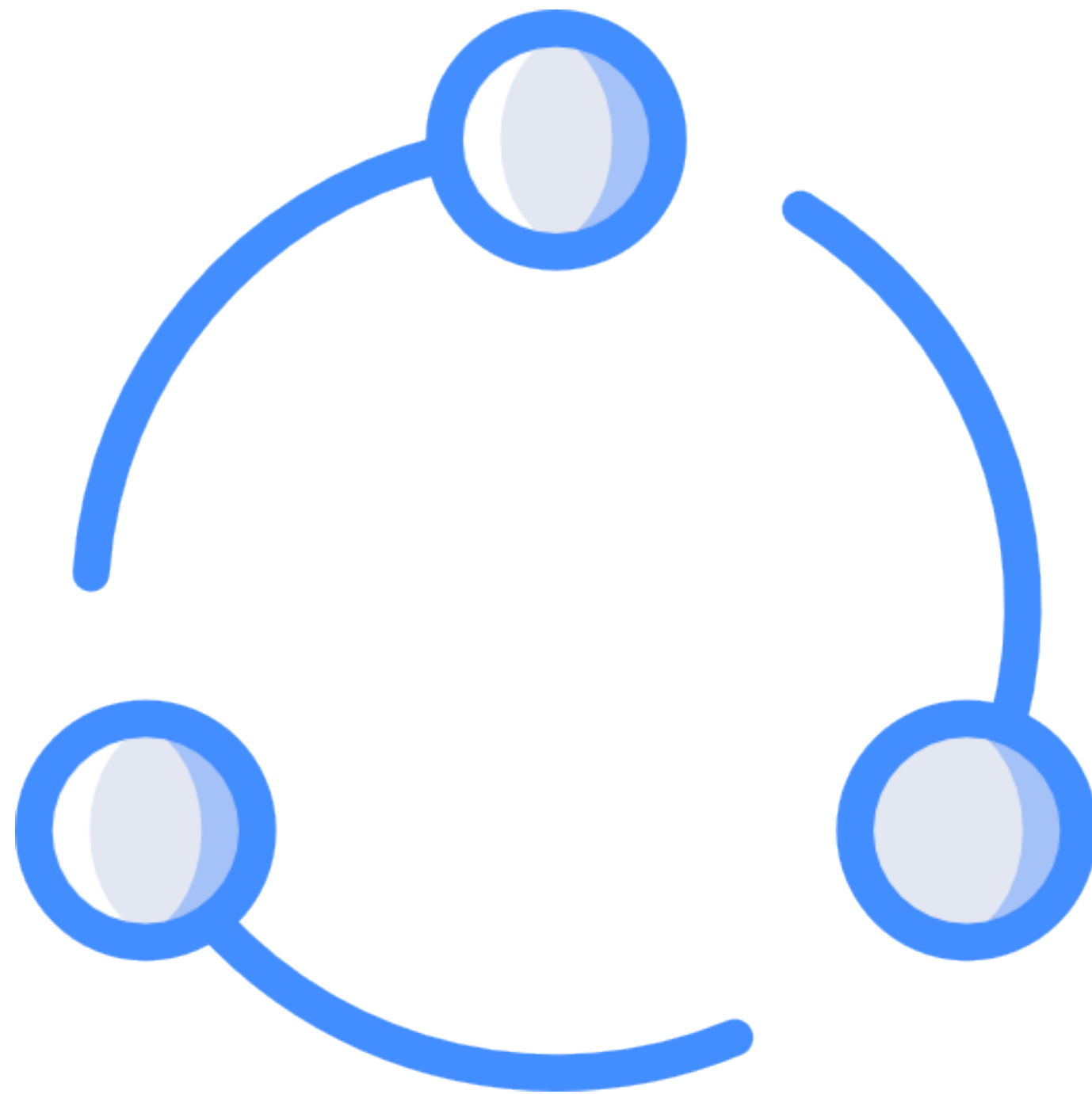
均值編碼平滑化

$$\text{新類別均值} = \frac{\text{原類別平均} * \text{類別樣本數} + \text{全部的總平均} * \text{調整因子}}{\text{類別樣本數} + \text{調整因子}}$$

*調整因子用來調整平滑化的程度，依總樣本數調整

小提醒：均值編碼容易 overfitting

雖然均值編碼符合直覺，並且也是強大的編碼方式
但實際上使用時很容易 overfitting (即使使用了平滑化)
所以需確認是否適合再使用 (用 cross validation 確認使用前後分數)



- 當類別特徵與目標明顯相關時，該考慮採用**均值編碼**
- 知道均值編碼最大的問題，在於**相當容易 Overfitting**
- **平滑化**的方式能修正均值編碼容易 Overfitting 的問題，但效果有限，因此仍須**經過檢驗**後再決定是否該使用均值編碼

解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

