

Day 19

特徵工程

數值型特徵 補缺失值與標準化



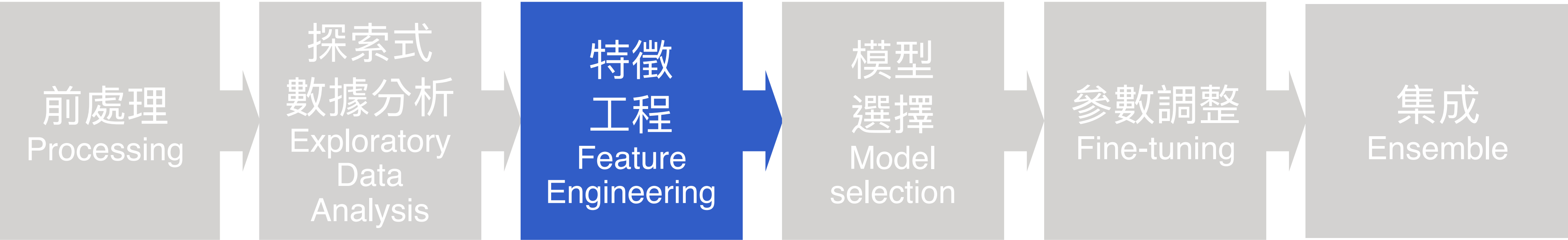
出題教練

陳明佑

知識地圖 特徵工程 數值型特徵 - 補缺失值與標準化

機器學習概論 Introduction of Machine Learning

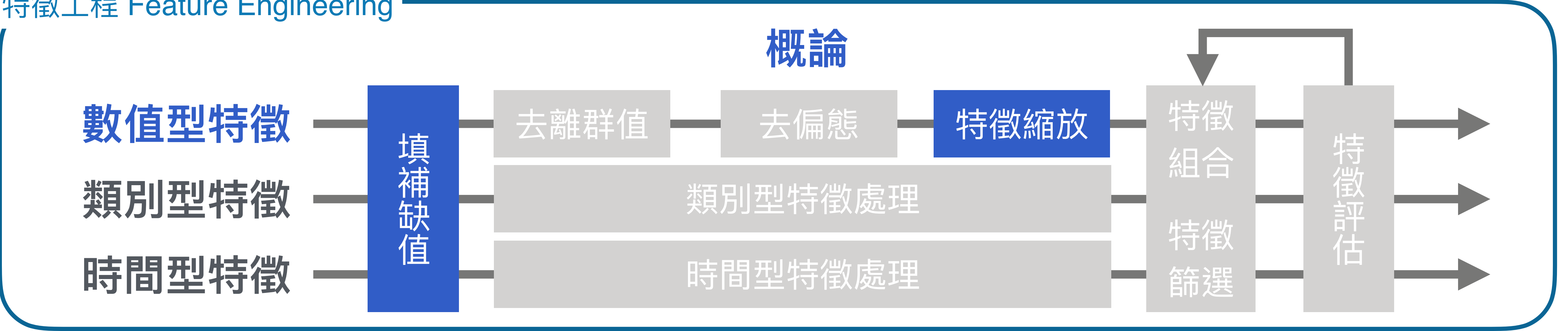
監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



特徵工程 Feature Engineering



本日知識點目標

1

資料當中，缺失值應該怎麼補？

2

補缺失值時該注意什麼？

3

將資料標準化的意義在哪裡？

4

什麼時候該用標準化？什麼時候又該用最大最小化呢？

填補缺值 (1 / 2)

看答案前先想一想：下列幾種缺失值怎麼補最好？(問號表示缺失值)

停車位	房間數	屋齡	行政區
True	1	11	?
?	?	20	南港區
?	2	32	大安區
True	1	?	南港區
True	?	25	?
?	1	?	文山區

填補缺值 (2 / 2)

停車位	房間數	屋齡	行政區
True	1	11	南港區(眾數)
False	0	20	南港區
False	2	32	大安區
True	1	屋齡總平均	南港區
True	0	25	南港區(眾數)
False	1	屋齡總平均	文山區

沒有False但應該有, 推測False表示成為空白

沒有0但應該有, 推測0表示成為空白

屋齡不可能為空值, 推測應該是資料遺失或者漏填, 故可取總平均或中位數

行政區不可能為空值, 推測應該是漏填, 故可取行政區眾數或另創一值

填補缺值
最重要的是欄位的領域知識與欄位中的非缺數值

複習：填補缺值的方式

- 填補統計值
 - 填補平均值(Mean)：數值型欄位，偏態不明顯
 - 填補中位數(Median)：數值型欄位，偏態很明顯
 - 填補眾數(Mode)：類別型欄位
- 填補指定值 - 需對欄位領域知識已有了解
 - 補 0：空缺原本就有 0 的含意，如前頁的房間數
 - 補不可能出現的數值：類別型欄位，但不適合用眾數時
- 填補預測值 - 速度較慢但精確，從其他資料欄位學得填補知識
 - 若填補範圍廣，且是重要特徵欄位時可用本方式
 - 本方式須提防overfitting：可能退化成為其他特徵的組合

為何要標準化 (1 / 2)

想一想：競賽中的給分，如果發生下列情形... 要如何修正呢？

評審1	評審2	評審3	評審4
0	6	10	8.1
10	7	7	8.5
10	5	8	7.6
0	8	9	8.3
0	6	9	8.1
10	7	8	8.6

為何要標準化 (2 / 2)

評審1

原始值	修正後
MIN 0	0
MAX 10	1
10	1
0	0
0	0
10	1

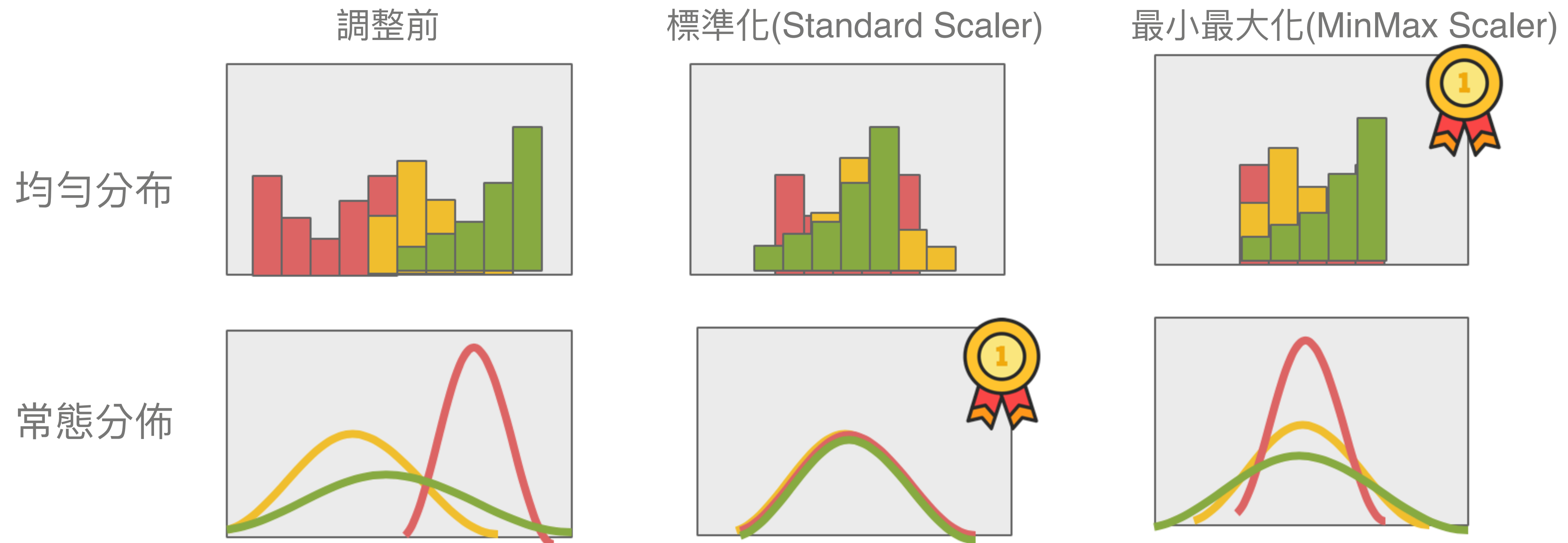
評審4

原始值	修正後
8.1	0.5
8.5	0.9
MIN 7.6	0
8.3	0.7
8.1	0.5
MAX 8.6	1

標準化：以合理的方式，平衡特徵間的影響力
此處範例為「最大最小化」，一般常用的方法還有標準化

複習：標準化 / 最小最大化

- 標準化 (Standard Scaler)：假定數值為常態分佈，適合本方式平衡特徵
- 最小最大化 (MinMax Scaler)：假定數值為均勻分佈，適合本方式平衡特徵



如果離群值有處理 (處理方式將於 Day20 課程介紹)，兩者差異不太大

標準化 / 最小最大化適用場合

- 樹狀模型或非樹狀模型(參考今日練習題)
 - 非樹狀模型：如線性迴歸, 羅吉斯迴歸, 類神經...等，標準化 / 最小最大化後對預測會有影響
 - 樹狀模型：如決策樹, 隨機森林, 梯度提升樹...等，標準化 / 最小最大化後對預測不會有影響
- 標準化 / 最小最大化 使用上的差異
 - 標準化：轉換不易受到極端值影響
 - 最小最大化：轉換容易受到極端值影響

註：因此，去過離群值的特徵，比較適用最大最小化

- 補缺失值的方法因**特徵類型**與**缺的意義**不同，會有許多不同補法，需要因資料調整，無法一概而論
- 除了上面兩點，補缺失值還要注意盡量**不要破壞資料分布**
- 標準化的意義：**平衡**數值特徵間的影響力
- 因為**最大最小化**對**極端數值**較敏感，所以如果資料不會有極端值，或已經去極端值，就適合用最大最小化，否則請用標準化

解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

