Day 27 特徵工程

特徵組合 - 類別與數值組合

知識地圖 特徵工程 特徵組合 - 類別與數值組合

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

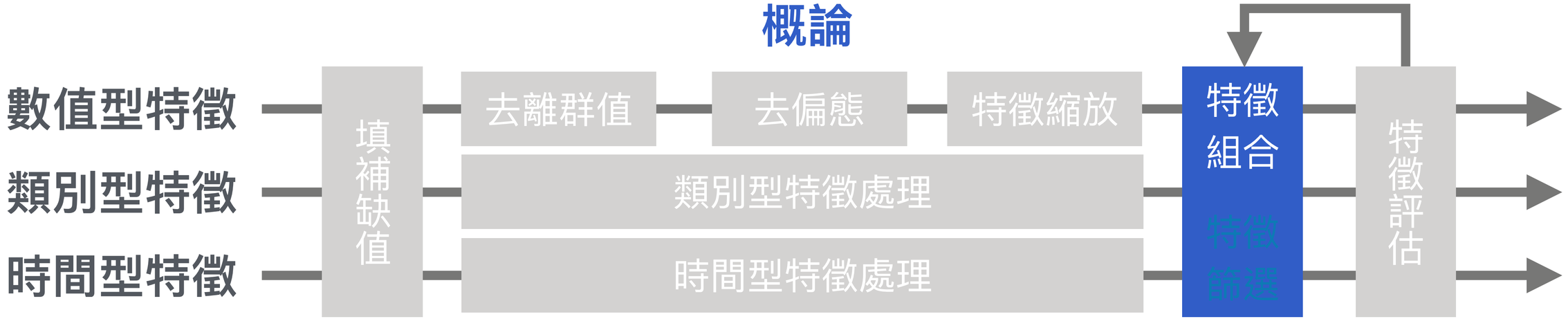
前處理 Processing → 探索式數據分析 Exploratory Data Analysis → **特徵工程 Feature Engineering** → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

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

- 類別型特徵也能和數值特徵組成新特徵嗎?
- 群聚編碼有哪些操作運算可以使用?
- 群聚編碼與之前的均值編碼最主要有什麼不同?

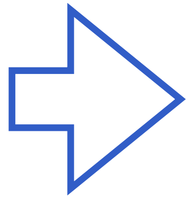
群聚編碼 (1 / 2)

既然數值型特徵之間能合成新特徵，那麼**類別與數值型之間**也能嗎?

例 : 保費續約預測中，已知險種 (類別型) 與前一年度保費 (數值型) 都是重要特徵

是否可以組合兩者，形成新特徵?

險種	前一年保費
第三責任險	2600
竊盜險	1600
第三責任險	2500
第三責任險	2300
竊盜險	1400



?

群聚編碼 (2 / 2)

類似均值編碼的概念，可以取類別平均值 (Mean) 取代險種作為編碼
但因為比較像性質描寫，因此還可以取其他統計值，如中位數 (Median)，眾數 (Mode)，最大值(Max)，最小值(Min)，次數(Count)...等

險種	前一年保費	平均值	最大值	次數
第三責任險	2600	2466.7	2600	3
竊盜險	1600	1500	1600	2
第三責任險	2500	2466.7	2600	3
第三責任險	2300	2466.7	2600	3
竊盜險	1400	1500	1600	2

群聚編碼 (Group by Encoding)

- 數值型特徵對文字型特徵最重要的特徵組合方式
- 常見的有 mean, median, mode, max, min, count 等
- 與均值編碼 (Mean Encoding) 的比較

名稱	均值編碼 Mean Encoding	群聚編碼 Group by Encoding
平均對象	目標值	其他數值型特徵
過擬合 (Overfitting)	容易	不容易
對均值平滑化 (Smoothing)	需要	不需要

群聚編碼的常見疑問

Q1 : 什麼時候需要群聚編碼?

Ans : 與數值特徵組合相同,

先以 **領域知識** 或 **特徵重要性** 挑選強力特徵後, 再將特徵組成更強的特徵

兩個特徵都是數值就用特徵組合, 其中之一是類別型就用聚類編碼

*特徵重要性會於 Day 29 再與各位詳述

Q2 : 聚類編碼時, 該如何挑選平均 / 最大值 / 次數 ... 等統計值?

Ans : 依照 **領域知識** 挑選,或亂槍打鳥後再以 **特徵重要性** 挑選

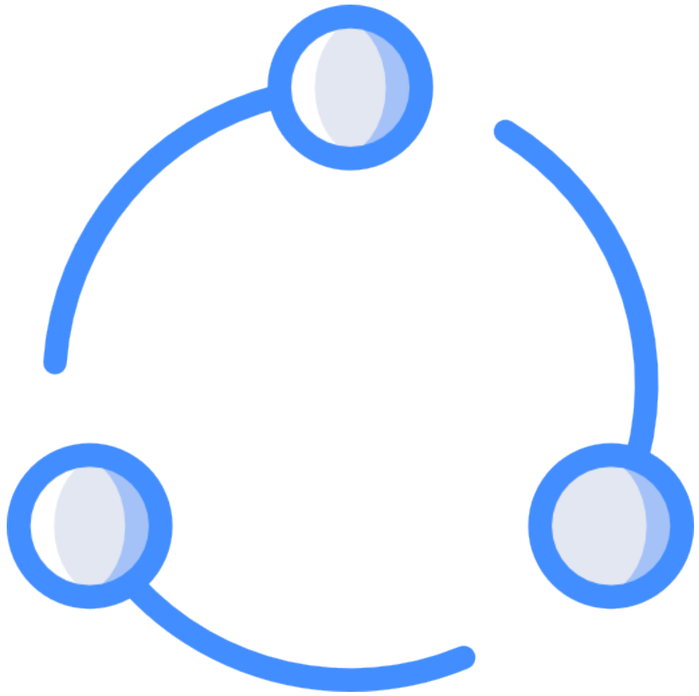
Q3 : 亂槍打鳥? 不會造出無用的特徵嗎?

Ans : 機器學習的特徵是 **寧濫勿缺** 的,

因為以前以非樹狀模型為主, 為了避免共線性, 會很注意類似的特徵不要增加太多

但現在強力的模型都是樹狀模型, 所以只要有可能就通通做特徵囉

重要知識點複習



- 類別特徵與數值特徵，可以使用**群聚編碼**組合出新的特徵
- 群聚編碼最常使用的運算是 mean, 除此之外還有 median、mode、max、min、count等統計量可以使用
- 群聚編碼與之前的均值編碼最主要的差異，一個是特徵彼此之間與特徵目標值之間的差異，另一個最大的差異是：群聚編碼因為與目標值無關，因此**不容易 Overfitting**，也因此比均值編碼使用頻率高得多

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題