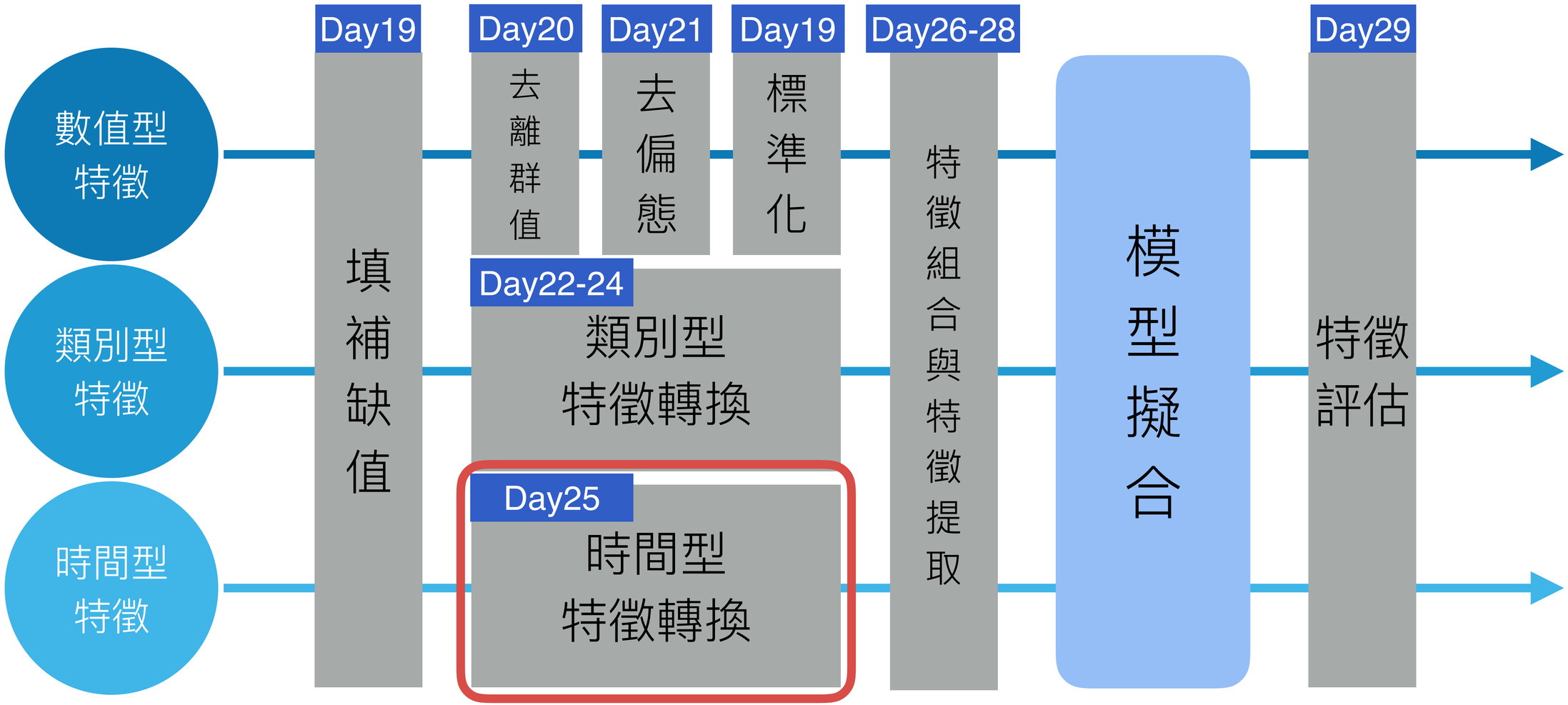
Day 25 特徵工程

時間型特徵

特徵工程 - 學習地圖

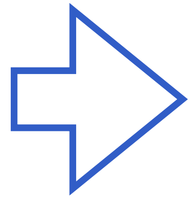

數值型特徵
類別型特徵
時間型特徵
Day19
填補缺值
Day20
去離群值
Day21
去偏態
Day19
標準化
Day22-24
類別型特徵轉換
Day25
時間型特徵轉換
Day26-28
特徵組合與特徵提取
模型擬合
Day29
特徵評估

時間特徵分解 (1 / 2)

最常見的特殊欄位是時間欄位，想想看應該怎樣編碼?

時間戳記

- 2014-06-12 03:25:56
- 2011-07-16 01:19:59
- 2011-10-21 23:54:10
- 2015-02-03 10:42:03
- 2009-06-13 16:10:54
- 2010-08-09 14:46:03



?

時間特徵分解 (2 / 2)

最直覺的方式，就是依照原意義分欄處理，或加上第幾周或星期幾
但某些欄(例：分、秒)與目標值的關係很低，有沒有更**有意義**的特徵呢?

時間戳記	年	月	日	時	分	秒
2014-06-12 03:25:56	2014	6	12	3	25	56
2011-07-16 01:19:59	2011	7	16	1	19	59
2011-10-21 23:54:10	2011	10	21	23	54	10
2015-02-03 10:42:03	2015	2	3	10	42	3
2009-06-13 16:10:54	2009	6	13	16	10	54
2010-08-09 14:46:03	2010	8	9	14	46	3

週期循環特徵 (1 / 2)

時間也有週期的概念, 可以用週期合成一些重要的特徵
聯想看看：有哪幾種時間週期, 可串聯到一些可做特徵的性質?

- 年週期 與春夏秋冬季節溫度相關
- 月週期 與薪水、繳費相關
- 周週期 與周休、消費習慣相關
- 日週期 與生理時鐘相關

週期循環特徵 (2 / 2)

前述的週期所需數值都可由時間欄位組成, 但還**首尾相接**

因此週期特徵還需以**正弦函數(sin)**或**餘弦函數(cos)**加以組合

例如：**年週期 (正：冷 / 負：熱)**

cos((月/6 + 日/180)π)

周週期 (正：精神飽滿 / 負：疲倦)

sin((星期幾/3.5 + 小時/84)π)

日週期 (正：精神飽滿 / 負：疲倦)

sin((小時/12 + 分/720 + 秒/43200)π)

*註：此處小時是24小時制

時段特徵

短暫時段內的事件計數，也可能影響事件發生的機率
如：網站銷售預測，點擊網站前 10分鐘 / 1小時 / 1天 的累計點擊量

以一筆 17:05 發生的網站瀏覽事件為例
同樣是1小時的統計，基礎分解會統計當日 17 時整個小時的點擊量
時段特徵則是會統計 16:05-17:04 的點擊量
兩者相比，後者較前者更為合理

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題