Day 26 特徵工程

特徵組合 - 數值與數值組合

陳明佑
出題教練

知識地圖 特徵工程 特徵組合 - 數值與數值組合

CUPOY

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

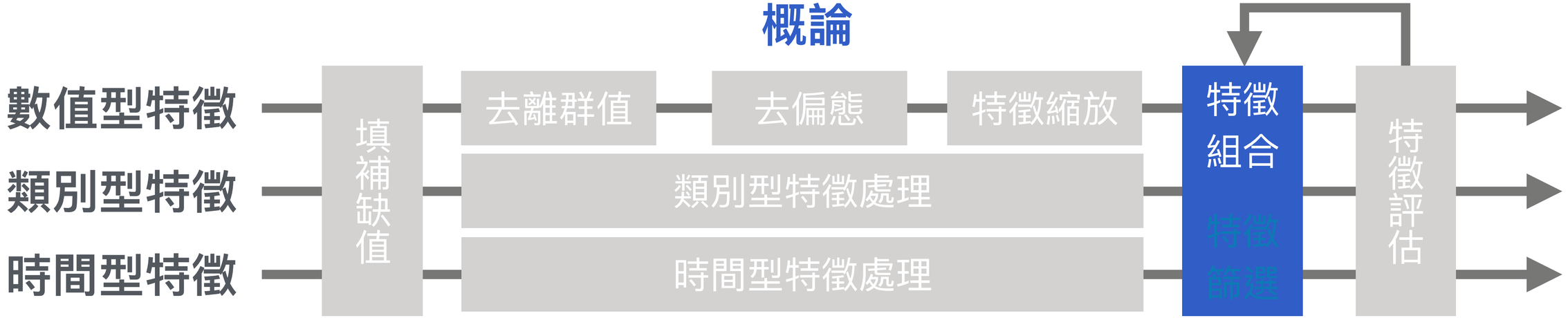
前處理 Processing → 探索式數據分析 Exploratory Data Analysis → **特徵工程 Feature Engineering** → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

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

- 數值與數值的特徵組合，除了基礎的加減乘除等四則運算，最關鍵的部分是什麼?
- 機器學習的關鍵又是什麼?

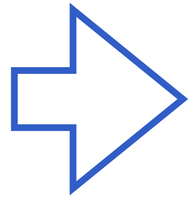
特徵組合 (1 / 3)

CUPOY

在計程車費預估中，有四個欄位分別表示起終點的經緯度

想想看，是否可以用這些組合出與車費更有相關的特徵?

起點經度	起點緯度	終點經度	終點緯度
-73.99058	40.76107	-73.98112	40.75863
-73.98840	40.72343	-73.98964	40.74169
-74.01578	40.71511	-74.01202	40.70788
-73.97732	40.78727	-73.95803	40.77883

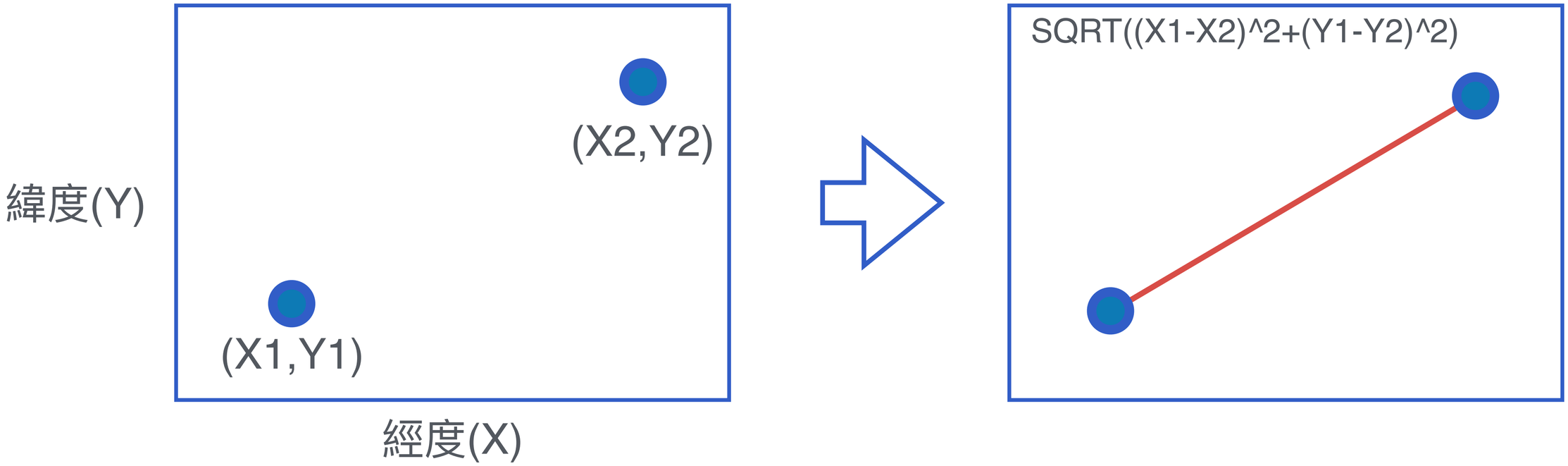


?

特徵組合 (2 / 3)

合理的想法是：將這四個特徵看成座標

因此用平面座標距離組合出來的特徵，更有預測力也非常合理

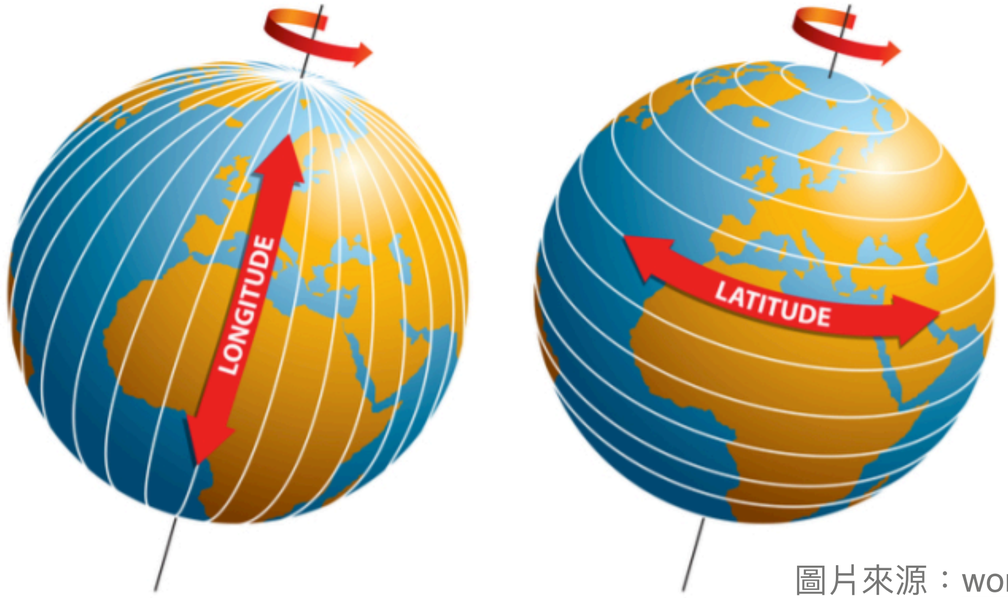


想一想：還有沒有可能合成更強力的特徵呢?

特徵組合 (3 / 3)

CUPOY

事實：經緯度每一度並不一樣長



圖片來源：worldatlas

觀察資料緯度集中在 40.75 度附近

可以算得經度與緯度代表的長度比為 cos(40.75度) : 1 = 0.75756 : 1

由此校正後的兩地距離，預測正確度更高

特徵工程的核心概念：領域知識

CUPOY

- 機器學習的關鍵在特徵工程
- 特徵工程的關鍵在領域知識

回想一下：

只是知道有四個數值欄位，預測力有限

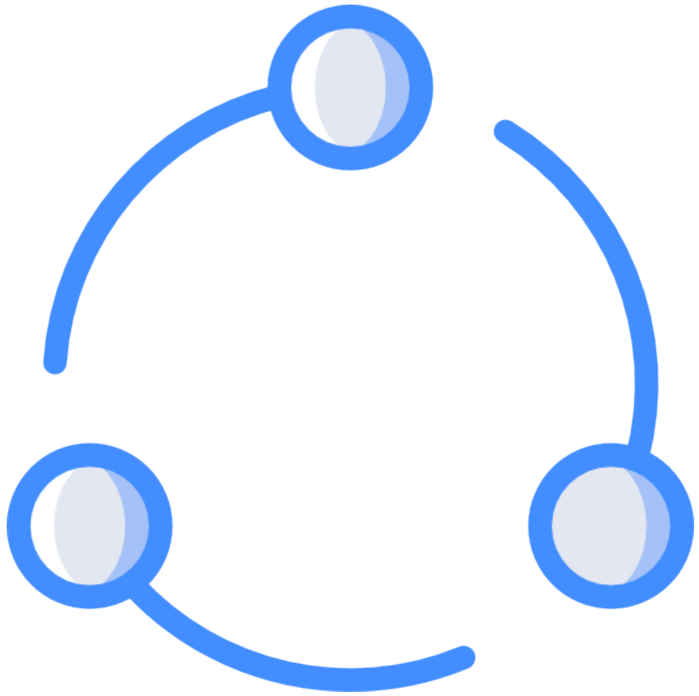
加上知道這四個數值是座標相關，就可以使用高斯距離合成特徵

再加上知道這是經緯度，就可以得到更精確的結果

因此，對問題領域知識的了解，才是特徵工程最重要的環節

*Day26所說時間的幾種週期，也可視為我們對「時間」的知識

重要知識點複習



- 數值與數值的特徵組合，最關鍵的部分是領域知識
- 機器學習的關鍵是特徵工程，當然其餘部分仍然很重要，但是各部分都熟悉之後，最有效提升模型預測力的部分就是特徵工程

**註：好的資料能夠更有效提升預測力，特徵工程最有效的前提是資料集固定時(例如競賽)

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題