Day 12 資料清理數據前處理

把連續型變數離散化

連續型變數離散化

Goal

- 變得更簡單 (可能性變少了)
  - 假設年齡 0-99 (100 種可能性) >> 每 10 歲一組 (10 種可能性)
- 離散化的變數較穩定，假設年齡 > 30是 1，否則 0。
  如果沒有離散化，outlier 「年齡 300歲」 會給模型帶來很大的干擾。

關鍵點

- 組的數量
  - 一樣以年齡為例子，每 10 歲一組就會有 10 組
- 組的寬度
  - 一組的寬度是 10 歲

連續型變數離散化

主要的方法

- **等寬劃分**：按照相同寬度將資料分成幾等份。缺點是受到異常值的影響比較大。
- **等頻劃分**：將資料分成幾等份，每等份資料裡面的個數是一樣的。
- **聚類劃分**：使用聚類演算法將資料聚成幾類，每一個類為一個劃分。

如何離散化是一門學問！

充電時間
Brain Charge

請跳出PDF至官網Sample Code＆作業
進行今日作業