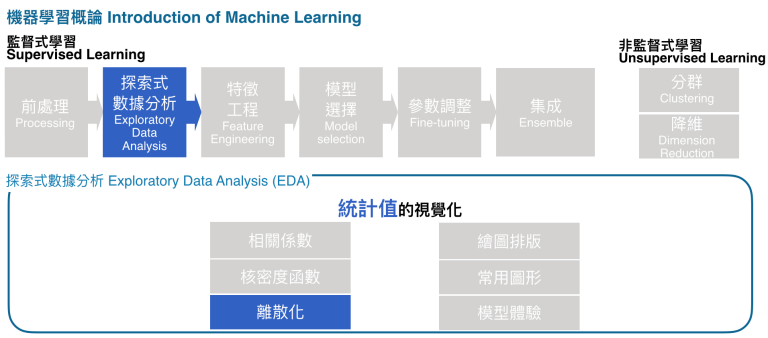
D17：EDA: 把連續型變數離散化

簡報閱讀 範例與作業 問題討論

把連續型變數離散化

知識地圖



本日知識點目標

- 了解離散化連續數值的意義以及方法

連續型變數離散化

Goal

1. 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)
2. 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

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

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

除了以上的主要方法，也會因需求而需要自己定義離散化的方式，如何離散化是一門學問！

- 離散化的目的是讓事情變簡單、減少 outlier 對分析以及訓練模型的影響
- 主要的方法是等寬劃分 (對應 pandas 中的 cut) 以及等頻劃分 (對應 pandas 中的 qcut)
- 可以依實際需求來自己定義離散化的方式

推薦延伸閱讀

連續特徵的離散化：在什麼情況下可以獲得更好的效果(知乎)

> 連續特徵的離散化：在什麼
> 連續特徵的離散化：在什麼情況下將連續的特徵離散化之後可以獲得更好
> www.zhihu.com

這個網頁是個討論串，經由幾個網友的討論與補充，很好地說明了離散化的理由：儲存空間小，計算快，降低異常干擾與過擬合(ovefitting)的風險，主要想請同學參考第1位的回答，至於其他的討論則請同學參考即可。

解題時間

Sample Code &作業
開始解題

下一步：閱讀範例與完成作業