Day 8 資料清理數據前處理

EDA之資料分布

出題教練 游為翔 / 杜靖愷

知識地圖 機器學習前處理 EDA之資料分布

CUPOY

機器學習前處理

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

前處理 Processing

資料讀取 → 格式調整 → 填補缺值 → 去離群值 → 特徵縮放

本日知識點目標

了解如何通過基本的統計數值以及畫圖來了解資料

EDA - 統計量化的方式？

CUPOY

以單變量分析來說，量化的分析方式可包含

- **計算集中趨勢**
  - 平均值 Mean
  - 中位數 Median
  - 眾數 Mode
- **計算資料分散程度**
  - 最小值 Min
  - 最大值 Max
  - 範圍 Range
  - 四分位差 Quartiles
  - 變異數 Variance
  - 標準差 Standard deviation

基本上使用上述統計特徵就可以讓我們初步了解資料的樣子，並且觀察是否有異樣

EDA視覺化的方式？

有句話「一畫勝千言」，除了數字，視覺化的方式也是一種很好觀察資料分佈的方式，可參考 python 中常用的視覺化套件

畫圖沒靈感的時候可以到這兩個套件的範例網頁逛逛！

matplotlib

seaborn

重要知識點複習

- 資料大部分時候都是非常多的，我們沒辦法用眼睛一筆一筆都看完，平均值、標準差、最大最小值等統計數值能幫助我們迅速對資料有初步的了解。
- 了解統計數值後，把資料的圖畫出來除了能夠更全面地了解資料，也能幫我們快速觀察到異常的地方
- pandas 有許多已經寫好用來做以上這些觀察的函數，熟悉這些函數的使用能加速觀察資料的過程

延伸閱讀

除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有多餘時間，可再補充延伸閱讀文章內容。

推薦延伸閱讀

敘述統計與機率分布

吳漢銘老師

網頁連結

要做出足夠深入的 EDA，對於統計的理解是必須的，這份教材可以提供同學了解統計觀念的機會，由於這份教材的範圍太廣，牽涉到太多預備知識，並不適合同學完整閱讀，只建議在不熟悉名詞時，回頭當作工具書參考即可。

Source: By Calvin.Andrus (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

推薦延伸閱讀

常見的統計分佈 (英文)

healthknowledge.com

網頁連結

這個網頁描述了幾個常見的分布：常態分布 / 二項式分布 / 卜瓦松分布，其中常態分布是我們最常使用到的，這個網頁建議同學大致上知道常態分布的形狀 (右圖) 即可，至於機率密度函數等其他相關知識，可以等到有需要時再查詢。

解題時間

It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題