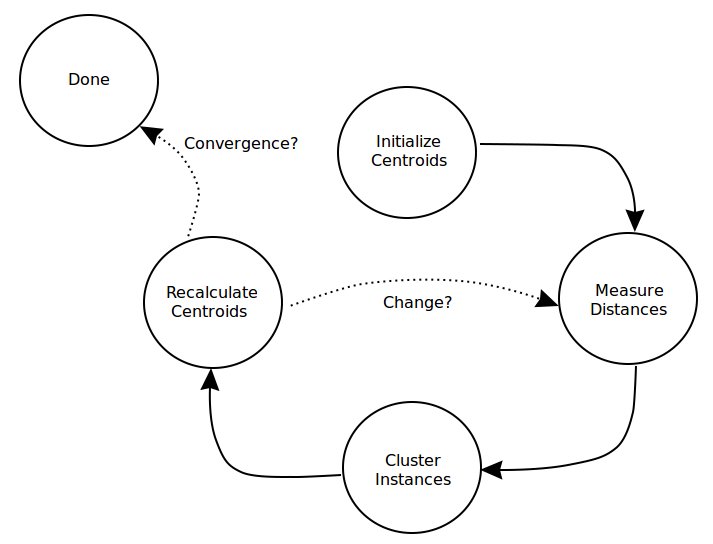
**資料科學作業二心得報告**

A10515001 李大祥

2018-06-06

## **1.演算法流程**

本次作業,我實現的 clustering 演算法是 K-means , 基本流程下面圖所示：



[圖片來源](https://www.kdnuggets.com/2017/06/machine-learning-workflows-python-scratch-part-2.html)

這次作業要求是在 subject.csv 中判斷兩個 pokemon 是否為相同屬性，是為1，不是為0。

所以從 HW2\_pokemon.csv 中讀出資料後，應該要根據預設的 otal,HP,Attack,Defense 等屬性去做分群的動作，也就是將每個觀測分配到聚類中並使得組內平方和（WCSS）達到最小，然後再通過評估距離更新(Update)每個聚類的新的均值點。如此循環往復，直至達成停止條件，例如新的均值點不再變動或者變動範圍極小。

## **2.程式架構**

第1部分先引用 pandas ，用 read\_csv 讀取資料。

第2部分是預處理資料，寫在報告的第三部分。

第3部分是我們使用 sklearn.cluster 的 KMeans() 方法進行分群，通過 inertia\_ 屬性查看每次分群後的效果，inertia\_ 越低表示同一群的點之間距離更近、分的效果越好。

第4部分是訓練 model，並使用訓練好的 model 去預測測試資料。

第5部分就是把預測的結果生成 kaggle 要求的提交格式。

## **3.資料預處理**

資料預處理部分除了利用 pandas 的屬性轉換方法 map 直接轉換 Body\_Style 裡的string 到數值，另外主要重點是使用 sklearn.preprocessing 中的 StandardScaler做處理，也就是減去均值，然後除以標准差，並用 fit\_transform 做一個標準化處理。

## **4.資料分析過程中遇到的困難及嘗試**

遇到的困難主要是不知道需要分成多少群，以及哪些屬性會比較重要，所以需要挨個嘗試，並調整這些屬性的權重。

開始的時候是直接先使用 Total,HP,Attack,Defense,Sp\_Atk,Sp\_Def,Speed 這些只有數字值的屬性，調整群組數到 160 次左右，成績就大概到 0.91 左右，基本再調整分群數目也沒有用了。後來加入了 Height\_m, Weight\_kg 發現成績可以提高到 0.92 多，但是群組數目也提高到 200 左右。最後干脆轉換 string 之後，把 Body\_Style 也加入計算，調整群組數目後，最後成績提高到 0.92741

## **5.程式碼執行說明(包含執行環境等資訊）**

執行環境是 Ubuntu 18.04，Python3 ，需要預裝 Keras 以及 sklearn、numpy、pandas、matplotlib等。

執行就是改一下程式碼檔案名字，然後直接運行 A10515001.py 就好。