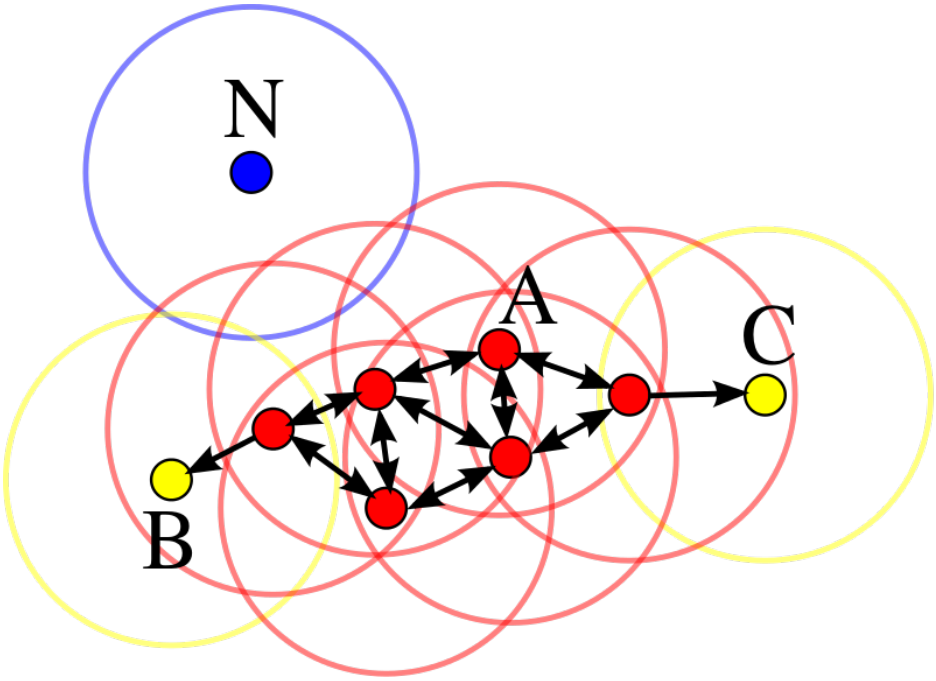
Applied Mutivariate Analysis
Final Report

DBSCAN



109024501 林建佑
109024505 王軒宇
109024510 黃仕傑
109024517 高梓傑

June 16 2021

目錄 Contents

1 簡介 Introduction **1**

2 方法 Methodology **2**
2.1 Definition 2
2.2 Alogorithm 3

3 模擬 Simulation **4**
3.1 Compare parameters 4
3.2 Compare Methods 6

4 資料分析 Real Data Analysis **7**
4.1 個案背景 7
4.2 Goal 7
4.3 Data Source 8
4.4 Exploratory Data Analysis 8
4.5 Data Preprocess 8
4.6 DBSCAN 9
4.7 Image Segmentation 10
4.8 SVM 11

5 結論 Conclusion **12**

6 Reference & Dataset **13**

1 簡介 Introduction

Unsupervised Learning 在統計上是一個在沒有 label 的資料，找出該資料的未知模式所採行的統計方法，像是 Dimension Reduction, Clustering 等皆是 Unsupervised Learning 的一種。而此次主題「DBSCAN」正是一種 Unsupervised Learning 的方式，並且更詳細的說，他是一個 「Clustering」的方式。Clustering 所要解決的問題是想要將相近的樣本分成群，以幫助我們解決決策問題，例如在商業分析當中，客群往往會是各家企業會注重的，但往往客群在討論上會流於抽象概念，爲了具體化這樣的概念，就可以利用「Clustering」具體化問題；此外，在衛星影像分析中，我們有可能會遇到的問題是將影像中臨近居住地分成一個一個聚落以進行政府政策施行。

倘若我們要將分群應用在與空間有關的資料，如以下種類的資料和影像等等：

- Spatial data
- Satelite
- X-ray

在此種資料的分析中，目標經常會是需要分辨出任意形狀的特徵，如衛星影像裡的河流、照片裡的車子等等，群聚演算法在此種任務中扮演重要的角色。

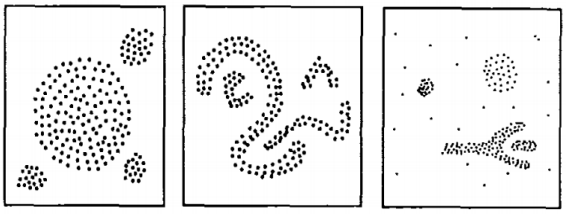


figure 1: Sample databases

在分析此種空間型資料型態時需要以下要求:

- 處理大型空間資料時，對於domain knowledge的要求較小
- 可以分辨出任意形狀的資料型態
- 在分析大型資料時仍然有好的計算效率

傳統的分群分法如:K-means、hierarchical clustering等等較難以分辨出任意型態的空間資料，又或是需要較多的domain knowledge來事先決定參數，無法在上述的要求下提供好的解決方案，於是作者提出了DBSCAN的方法以解決以上問題。

DBSCAN 在解決 Domain knowledge 影響的 input parameters 的問題上，只要定義臨近區域大致會有多近的距離即可進行分群。比起 K-means 得先定義分成幾群才可進行分群，DBSCAN 可以不用太多的 Domain knowledge，甚至可以直接藉由觀看影像決定。 DBSCAN 在解決找到任意形狀的分群上與 K-means 等其餘的分群方式不同，這個會在後續的章節介紹。 DBSCAN 在解決計算效率上也有好的演算法支持。 以上三點是 DBSCAN 比起傳統分群方式在 Large Spatial Databases 上優勢的原因。

2 方法 Methodology

2.1 Definition

DBSCAN是一種基於密度的分群演算法，這類密度分群演算法一般假定群可以通過樣本分布的緊密程度決定。相似度高的樣本，他們之間是緊密相連的，也就是說，在該群任意樣本周圍不遠處一定有相似度高的樣本存在。通過將緊密相連的樣本劃爲一類，這樣就得到了一個分群。通過將所有各組緊密相連的樣本劃爲各個不同的群，則我們就得到了最終的所有分群結果。

ϵ：鄰域的半徑

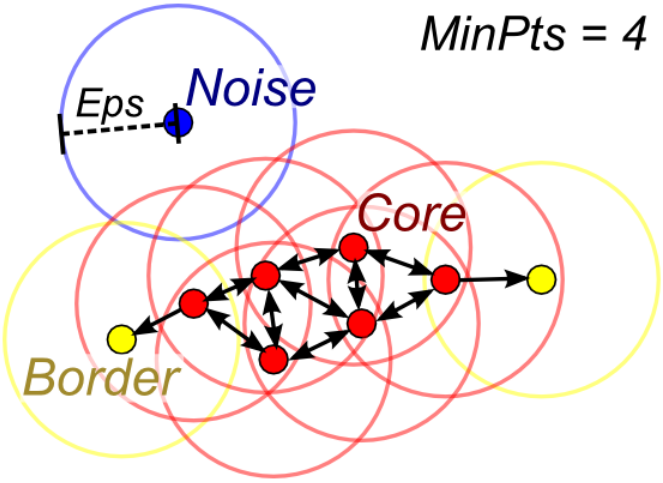
MinPts：形成高密度區域所需要的最少點數

核心點 (core point)：這些點在基於密度的cluster內部。點的鄰域由距離函數和用户指定的距離參數 eps 決定。核心點的定義是，如果該點的給定鄰域內的點的個數超過給定的閾值MinPts，其中MinPts也是一個用户指定的參數。

邊界點 (border point)：邊界點不是核心點，但它落在某個核心點的鄰域內。

雜訊點 (noise point)：雜訊點是即非核心點也是非邊界點的任何點。

如圖所示：



DBSCAN 密度定義：

DBSCAN是基於一組鄰域來描述樣本集的緊密程度的，參數（ϵ，MinPts）用來描述鄰域的樣本分布緊密程度。其中 ε 描述了某一樣本的鄰域距離閾值，MinPts 描述了某一樣本的距離爲 ε 的鄰域中樣本個數的閾值。

假設樣本集是$D = (x_1, x_2, \ldots, x_m)$ 則DBSCAN 具體的密度描述定義如下：

1. ϵ **- 鄰域**：對於 $x_j \in D$，其中 ϵ - 鄰域包含樣本及D中與 x_j 的距離不大於 ϵ 的子樣本集，即$N_\epsilon(x_j) = \{x_i \in D | d(x_i, x_j) \leq \epsilon\}$，這個子樣本集的個數記爲$|N_\epsilon(x_j)|$。

2. **核心對象**：對於任一樣本 $x_j \in D$ ，如果其 ϵ - 鄰域對應的 $N_\epsilon(x_j)$ 至少包含MinPts個樣本，即如果$|N_\epsilon(x_j)| \geq$ MinPts，則 x_j 是核心對象。簡單來說就是若某個點的密度達到演算法設定的閾值則其爲核心點。

3. **密度直達**：如果 x_i 位於 x_j 的 ϵ - 鄰域中，且 x_j 是核心對象，則稱 x_i 由 x_j 密度直達。反之不一定成立，即不能說 x_j 由 x_i 密度直達，除非 x_i 也是核心對象。

4. **密度可達**：對於 x_i 和 x_j ，如果存在樣本序列 $p_1, p_2, \ldots, p_T$，滿足$p_1 = x_i$, $p_T = x_j$，且p_{t+1}由 p_t 密度直達，則稱x_j 由 x_i 密度可達。也就是說，密度可達滿足傳遞性。此時序列中的傳遞樣本 $p_1, p_2, \ldots, p_{T-1}$均爲核心對象，因爲只有核心對象才能使其他樣本密度直達。注意密度可達也不滿足對稱性，這個可以由密度直達的不對稱性得出。

5. **密度相連**：對於 x_i 和 x_j ，如果存在核心對象樣本 x_k，使 x_i 和 x_j 均由 x_k 密度可達，則稱 x_i 和 x_j 密度相連。注意密度相連關係是滿足對稱性的。

2.2 Alogorithm

1. 參數設定：決定距離(半徑) ϵ 與最少點MinPts(門檻值)。

2. 任意選取一個樣本當作中心點，以步驟1設定好半徑畫圓。
   若圓內樣本數大於等於門檻值，則此樣本爲核心點，標記可達到圓內任一點。
   若圓內樣本數小於門檻值，則此樣本爲非核心點，不可達到任何點。

3. 對每一個樣本重複步驟2的動作，直至所有樣本都當過中心點爲止。

4. 分群：將有連結性(雙向可達)的樣本點劃分爲一群，其他局外點可檢視是否單向可達，劃分爲不同群體。

3 模擬 Simulation

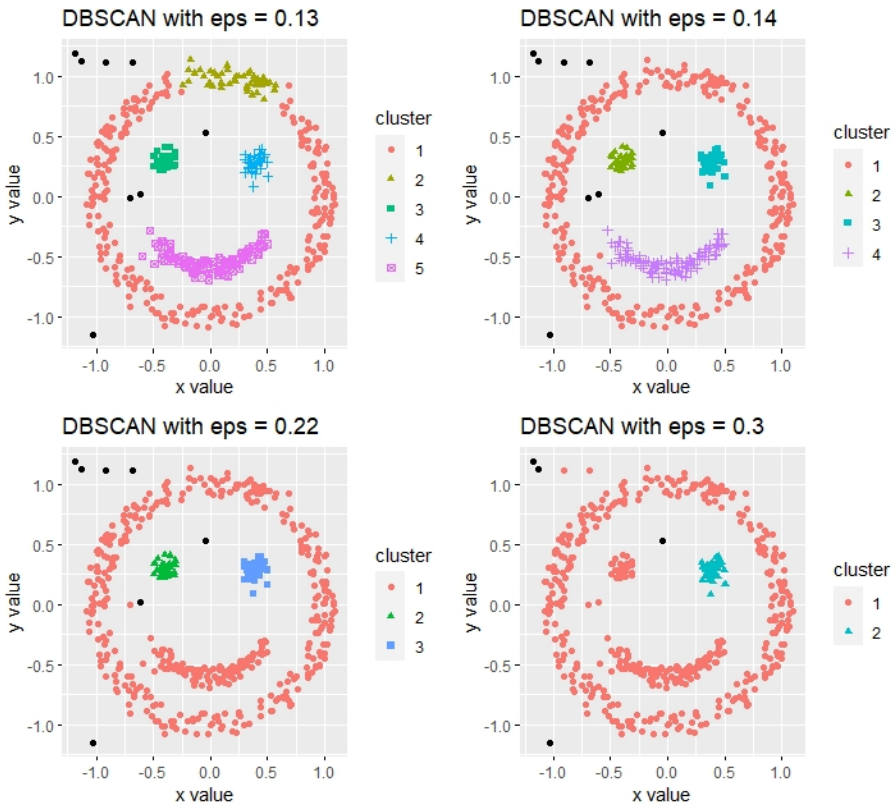
在使用 DBSCAN 時，可以調整 2 個參數 ϵ 和 MinPts ，本章節會介紹在不同參數設定下，以及不同方法之間的 Clustering 效果。

3.1 Compare parameters

3.1.1 ϵ

改變 ϵ 設定:

分別設定 $\epsilon = 0.13,\ 0.14,\ 0.22,\ 0.3$

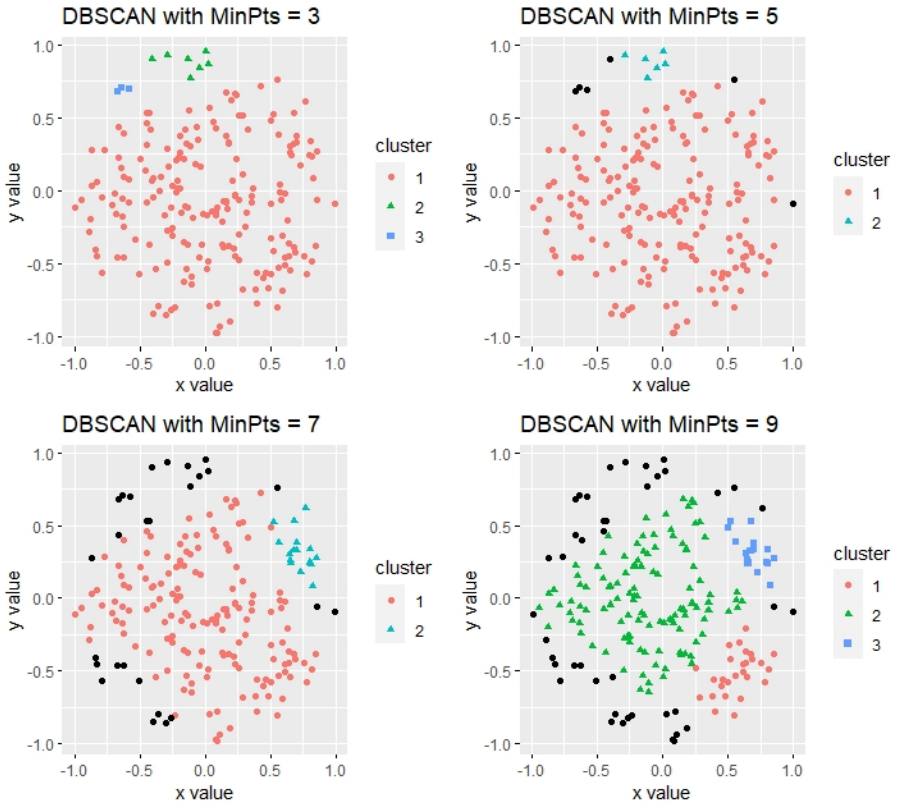


若 ϵ 設定過小，可能導致大部分資料不能聚類；相反地，若 ϵ 設定過大，大部分資料會被歸到同一群中。

3.1.2 MinPts

改變 MinPts 設定:

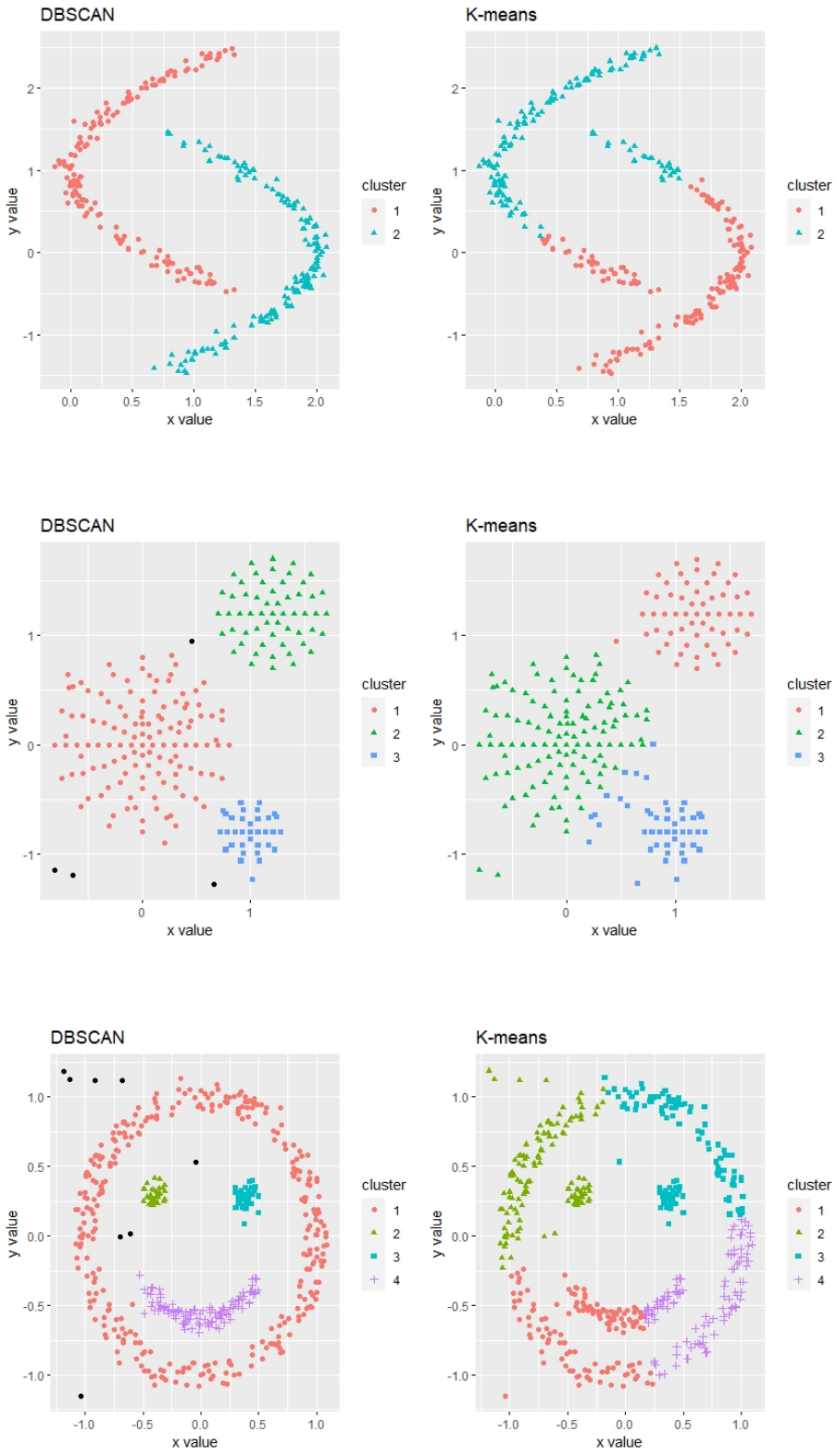
分別設定 MinPts = 3, 5, 7, 9



可以從上圖觀察到，隨著 MinPts 變大，core point 的鄰域密度要達到 MinPts，所以對 noise point 會更敏感。

3.2 Compare Methods

DBSCAN 和 K-means 兩種方法 Clustering 效果比較



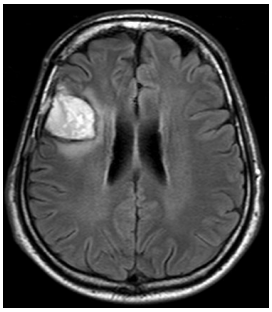
從上方三張圖形比較，可以觀察到 DBSCAN 的 Clustering 效果較優。

4 資料分析 Real Data Analysis

4.1 個案背景

腦部屬於中樞神經系統，組成的細胞種類多樣，是人最重要的器官之一。另外，醫學上所稱「腫瘤」，是不正常的細胞增生，侵犯了周圍或遠處的細胞組織，影響正常生理功能。
腦腫亦有良性與惡性之分，各包括多種細胞類型。惡性的腫瘤即所謂的癌症，癌細胞生長快速，並侵犯周圍組織。 良性腦瘤雖然不會侵犯到鄰近組織，但也可能壓迫腦組織的敏感區域並造成症狀，當一個良性腫瘤位於腦部的重要功能區域，並影響到神經與其他生理功能時，即使它不含癌細胞，臨床上仍應被視爲惡性的，故腦部腫瘤無論其本質爲良性與惡性，均可能造成神經功能的傷害，因此一般在臨床上均統稱爲「腦瘤」，從以上我們可以得知腦瘤對於人的影響巨大，故在醫學上，我們要盡可能即早發現這些腦瘤，並對症下藥。

根據台灣癌症防治網的介紹，目前對於腦瘤之診斷幾乎完全仰賴影像檢查。藉由不同之影像檢查我們可以知道患者是否有腦瘤、腦瘤之位置及大小，其與週邊腦血管神經之關係，及腦瘤之血管性等，進而預做判斷此腦瘤爲良性或是惡性，其中腦部電腦斷層檢查（Brain CT）以及磁振造影檢查（MRI）是幾種主要影像檢查方式。MRI 對於診斷腦疾病之優點，除了不具放射性外，對於不同組織之鑑別程度，亦較CT高甚多。也因此在醫療機構中常常以 MRI 影像 來診斷病患是否患有腦瘤，以下爲 MRI 影像：



4.2 Goal

由於目前 MRI 影像還是透過人的肉眼判讀，倘若 MRI 影像數量一多，就會導致醫護人員的工作量劇增，在疲勞下可能會導致誤判。故本次個案目標有二：

- 以分群方式標註 MRI 影像中腫瘤位置
- 根據這些標註，將之分類成病患與否

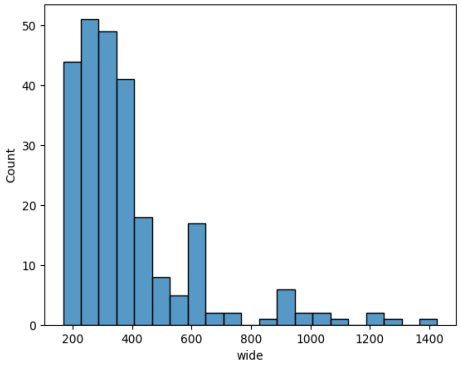
4.3 Data Source

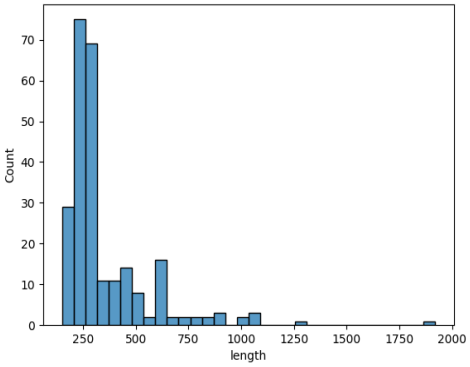
本次資料來源由 Kaggle 上的 Brain MRI Images for Brain Tumor Detection 資料集，總共有 253 張腦部 MRI 影像，其中 155 張有腫瘤的 MRI 影像，98 張沒有腫瘤的 MRI 影像。

4.4 Exploratory Data Analysis

在影像資料中，每張照片會有寬：m 個像素 (Pixel)、長：n 個像素。而在這些 $m \times n$ 個像素中，都有三個數值 - R,G,B， 分別代表著紅色、綠色、藍色的數值多寡，這些數值從 $0 \sim 255$ 的整數。所以在影像資料分析當中，我們會使用每個像素中的 R,G,B 數值作統計分析，也因此一張照片共會有 $m \times n \times 3$ 個變數。

從以下直方圖可以看到每張照片的大小不一。



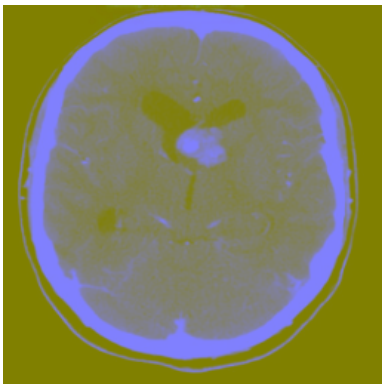


4.5 Data Preprocess

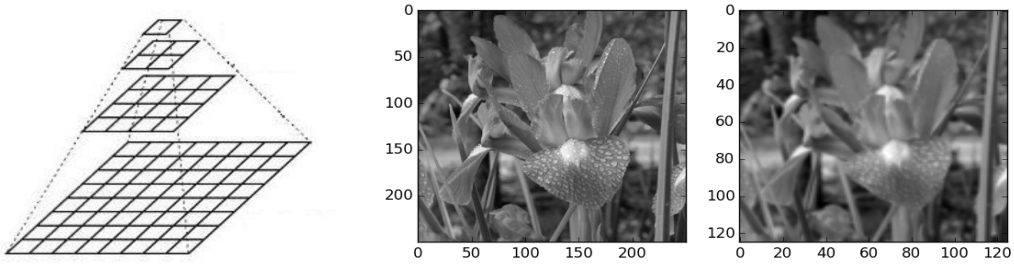
我們的 Data Preprocess 有三個步驟：

1. Image Resize： 由於每張照片的大小不一，但爲了後續分類時每張照片的 Covariate 個數相同，我們使用差值法將資料轉換成同樣大小：256 Pixel × 256 Pixel 。
2. 色彩空間轉換 : Hunter (1948) 創造了 L, a, b 色彩空間。其中 L 代表顏色的深淺，a 代表顏色的紅綠方向，b代表顏色的黃藍方向，用L a b就可以表示任何實物樣品的反射色或者透射色。 其好處是可以表示的顏色範圍更大且更貼近人眼所觀察到的顏色，多用於高階調色與設計軟體 (Adobe Photoshop)。
    - L : 以 $0 \sim 100$ 決定明亮度 數值由小到大，由黑到白
    - A : 以 $-128 \sim 127$ 代表顏色對立的維度 數值由小到大，由綠到紅
    - B : 以 $-128 \sim 127$ 代表顏色對立的維度 數值由小到大，由藍到黃

    故我們將原本的 RGB 色彩空間轉至 Lab 色彩空間，希望透過貼近人眼所觀察到的顏色幫助在分群以及分類上有所幫助。



3. 圖像金字塔：透過附近像素的加權平均計算出附近的特徵，也因此經過圖像金字塔過後的圖片會較模糊，影像大小也會變小，如附圖所示。



4.6 DBSCAN

完成Data precessing後，在此階段會使用分群演算法來達成影像分割(Image segmentation)，將圖像中我們感興趣的部分(腫瘤)突顯出來，並減少圖片中的雜訊，以利進行後續的分類。

由於常見的分群演算法如:K-Means、Hierarchical clustering等等會將所有樣本都分進某個群，無法做到偵測雜訊，故在此使用DBSCAN演算法不僅可做到影像分割，且有辦法去除影像中的雜訊。

DBSCAN遭遇的問題

- DBSCAN在分群和過濾雜訊上有令人驚豔的效果，但是模型的性能很大程度上取決於參數的選擇，若是在每張影像分割都要手動調整參數則會需要大量人力。

解決辦法

- 爲了解決上述問題，我們使用了一種動態調整參數的方法，此方法的主要想法是認爲Minpts和Eps參數存在一種函數型關係，更仔細來說的話可根據Minpts和k nearest neighbor來找出，以此來達成對Eps的自動調整。

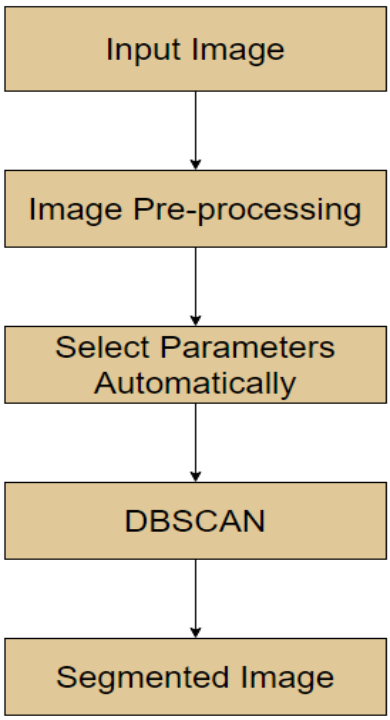
演算法

具體自動調整參數的做法如下:

1. 事先決定參數Minpts的值，通常與影像的size有關，在此我們選擇爲3。
2. 設定k=3，則對每張影像，計算出pixel間的距離，並取出各個pixel間的k-dist，其中k-dist表示距離第k近的點的距離。
3. 選擇所有pixels的k-dist中最大的值作爲Eps，且設定Eps爲 [1,3] 間的正整數，若是超出此區間則取邊界值。

4.7 Image Segmentation

解決了參數設定的問題後，則可在每次輸入新的影像時自動調整DBSCAN所需的參數，自動調整參數也讓每張影像的分割效果變得更好且穩定，需注意的是在此我們是使用L,a,b色彩空間作爲covarites對pixels進行分群，並無考慮pixel與pixel間在圖案上的空間訊息，對於每張圖片，則可使用以下做法來執行影像分割:



影像分割結果

以下爲輸入的影像和對影像做完影像分割的結果，可以看出DBSCAN演算法有確實將腦部和腫瘤的部分作出切割，完成影像的分割後，則會對每個pixels都標記上分群產生的labels，後續可根據這些labels來對圖片建立分類模型。

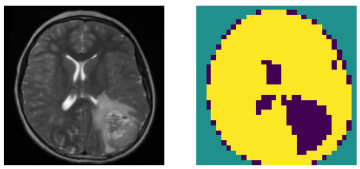


Figure 1: MRI with Brain Tumor

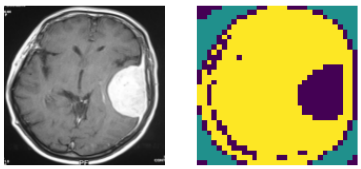


Figure 2: MRI with Brain Tumor

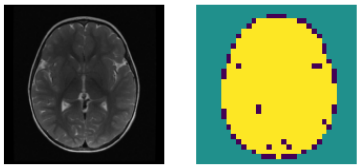


Figure 3: MRI without Brain Tumor

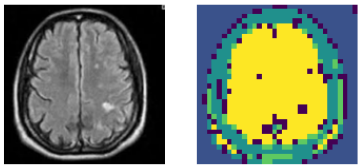


Figure 4: MRI without Brain Tumor

4.8 SVM

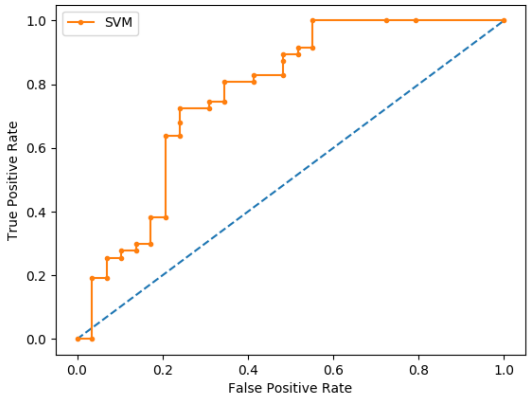
我們第二個目標是根據這些標註，將每個病患的 MRI 影像分類成淺在病患與否。我們的具體作法如下：

- 資料集：每個影像中的 label 做爲我們每個影像的 Covariate，再將這個病患是否罹患腦癌作爲 Response。
- 分類模型：SVM
- 模型訓練步驟：
  - **Training and Testing Data Spliting**
    將資料按 7 : 3 分成 Training and Testing Data
  - **Cross Validation**
    依照 K-fold (K = 5) 選擇 SVM 的 Soft Margin Penalty 參數（C），最後我們選擇的 C = 1.5。
  - **Fitting SVM model**
    根據 Cross Validation 的結果作爲最終 SVM 模型，透過全部的訓練資料配適 SVM 模型。

- Result :
  - Accuracy : 75 %
  - Confusion Matrix

Confusion Matrix		
眞實 \ 預測	沒有罹患	罹患
沒有罹患	14	15
罹患	4	43

  - AUC and ROC curve
    AUC : 77 %



可以觀察到最終模型對於眞實狀況是病患的預測準確率相較於眞實狀況並非是病患的預測準確率高上許多。我們認爲這樣的狀況是好的，由於腦瘤屬於嚴重的疾病，寧可將沒有罹患的病患先預測爲罹患再進行下一步的人工確認。 另外，如果想在提升模型的準確率，我們可以多考慮病患的病情衡量的指標，如:過去病史、家人病史等等，增加這些變數可以增加準確率。

5 結論 Conclusion

DBSCAN 優點

1. 相比 K-平均算法，DBSCAN 不需要預先聲明聚類數量。
2. DBSCAN 可以找出任何形狀的聚類，甚至能找出一個聚類，它包圍但不連接另一個聚類，另外，由於 MinPts 參數，single-link effect （不同聚類以一點或極幼的線相連而被當成一個聚類）能有效地被避免。
3. DBSCAN 能分辨噪音（局外點）。
4. DBSCAN 只需兩個參數，且對資料庫內的點的次序幾乎不敏感（兩個聚類之間邊緣的點有機會受次序的影響被分到不同的聚類，另外聚類的次序會受點的次序的影響）。

DBSCAN 缺點

1. 如果樣本集的密度不均勻，cluster間距離相差很大時，cluster品質較差，這時用DBSCAN一般不適合。
2. 如果資料庫裡的點有不同的密度，而該差異很大，DBSCAN 將不能提供一個好的聚類結果，因爲不能選擇一個適用於所有聚類的 MinPts- ε 參數組合。
3. 如果沒有對資料和比例的足夠理解，將很難選擇適合的 ϵ 參數。

資料分析

透過 DBSCAN 可以很好地將 MRI 影像中的腫瘤位置標註出來，並且在我們分類結果上有約 75% 的準確率以及77% 的 AUC。由於我們的分類屬於較保守的情況，所以我們認爲這樣的分類結果是可以接受的。

6 Reference & Dataset

A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (1996),Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu

Segmentation of Brain Tumour from MRI image – Analysis of K-means and DBSCAN Clustering (2013), Samir Kumar Bandyopadhyay, Tuhin Utsab Paul

Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, Jörg Sander, Martin Ester, Hans-Peter Kriegel, Xiaowei Xu

Mining Biomedical Images with Density-based Clustering, M. Emre Celebi and Y. Alp Aslandogan, Paul R. Bergstresser

Segmentation of Images using Density-Based Algorithms (2015), Atrayee Dhua, Debjani Nath Sarma, Sneha Singh, Bijoyeta Roy

A dynamic Method for Discovering Density Varied Clusters (2013), Mohammed Elbatta, Wesam Ashour

Dataset
Brain MRI Images for Brain Tumor Detection - Kaggle