Proposal

Group3

1 資料概述

本次報告使用的資料來自於 DepMap Consortium (DMC)，包含基因表達量、代謝體濃度、癌症轉移方向，分別有 1405、264、264 筆 (cells) 資料。以 DepMap ID 將資料整合成 264 筆 (cells)，整理後的資料包含 19221 個基因 (表達量)、225 個代謝體 (濃度)、5 種器官 (癌細胞是否轉移) 以及主要癌症，且皆無遺失值。

2 資料前處理

2.1 降維

由於變數個數過多，所以必須對變數做降維，第一個我們嘗試的方法為 PCA，但因為變數個數大於樣本數，PCA 出來的結果並不理想，前幾個主成分並不能解釋大部分原始資料的變異，因此我們使用其他統計方法對資料做降維。

2.1.1 對 gene 做 clustering

這邊我們可以先對 gene 做 clustering，這樣我們可以看到所有 gene 中，哪些 gene 之於所有細胞有相近的性質，有了每個 gene 的 cluster 之後，我們可以用平均或中位數的方式把同一組的 cluster 的 gene 合併成一個新的變數以達到 gene 的降維。

至於如何決定要降到的維度，我們可以使用一些視覺化的工具，如 t-sne 來看看能不能看出要分成幾個 gene 的 cluster。

2.2 變數選擇

2.2.1 Lasso

因為變數個數過多，所以使用 Lasso regression 做變數選擇，Lasso regression 可以將不重要的變數的係數變為 0，達到降維以及解決 overfitting 的問題。

3 模型選擇

3.1 目標

本組主要目標是以基因表達量和代謝體濃度建構一模型，來預測五種器官 (骨、腦、腎、肺、肝) 癌細胞是否轉移，並試圖找出那些代謝體的濃度變化與五種器官癌細胞是否轉移有顯著相關，解釋其可能的生物機制，進而降低癌症轉移風險。

3.2 logistic regression

由於我們的目標是要預測各個癌症是否轉移，在使用 logistic regression 時反應變數放的是 binary 的，所以在這筆資料中，我們可以使用 logistic regression 來對各個癌症是否轉移分別建立模型。

此外，由於我們有五個不同的癌症放在反應變數中，所以我們也可以把五個反應變數綁在一起看，這樣與其我們去建立五個個別的 logistic regression，我們可以在這裡建立出有 32 個反應變數的 multinomial logistic regression，透過這樣的方法建立出來的模型，對於得到每個癌症間的關聯性更加的有比較性。這個方法會有 32 個反應變數，根據之後的分析情況可能會視結果合併某些反應變數。

3.3 Deep learning model

3.3.1 依轉移順序建模

(可搭配 Logistic 或其他機器學習方法) 由於我們需要預測五個目標變數（骨、腦、腎臟、肝臟和肺）的轉移情況，且推測這些變數存在依賴關係，因此決定考量轉移順序建立模型。此方法假設相同疾病有一致的轉移順序，每個目標變數的預測依賴於之前轉移情況的預測。

挑戰: 需找出每個疾病的轉移順序，並確認相同疾病有一致的轉移順序，依順序分組建模。若有太多種轉移順序的組合，會導致模型數量過多及樣本數不足。

3.3.2 多頭神經網絡 (Multi-head Neural Network)

前幾層為共享層，將會從代謝體資料中取出通用的特徵，這些特徵對於所有目標變數的預測都是有用的。在共享層之後，網路會分成五個分支，每個分支會根據對應的目標變數做進一步的特徵處理，而每個分支的末端也會有一個獨立的輸出頭，預測對應的目標變數。

此方法可以從代謝體資料提取通用的特徵，並同時預測骨骼、腦、腎臟、肝臟和肺的轉移情況。