111-2 資料探勘期末報告

各位修習資料探勘的同學們好，以下為本學期期末報告相關內容，期末報告與期中報告相同，以組為單位，同學可以採取兩種方式進行本次作業，不得獨自脫組進行報告，報告時間為 6/1（四）。

1. 程式 ：

本次作業有兩個資料集，請同學任選一個作為本次作業的資料集：

第一個為 Arrhythmia Dataset，該資料集中含有 7 種類別(含正常共 8 種類別)，測試集中除了訓練集含有的 8 種類別外，加入了 5 類訓練集中不曾出現過的類別。

第二個為 Gene Expression Cancer RNA-Seq Dataset，該資料集中含有 3 種類別，測試集中除了訓練集含有的 3 種類別外，加入了 2 類訓練集中不曾出現過的類別。

由於測試集中含有訓練集中沒有的類別，請同學進行「先分類後分群」方法，將分類器較為不確定的樣本點分為未知類，再使用分群方法將未知類中不熟悉的樣本點進行分群並加入分類結果中，不限程式語言、不限分類方法與分群方法，目標為使自己的方法能運用訓練集進行調整或訓練，先後使用分類及分群方法，整合出測試集的所有類型之分類結果，並盡量取得高準確率。

資料集連結：

https://drive.google.com/drive/folders/18YTmzdCbupRUgBqsqzauG1NhEtiPTB9E?u sp=sharing

2. 論文閱讀報告:

挑選一篇使用分類及分群或資料探勘相關論文上台報告。決定好要報告的論文後請先寄信給老師告知「組別、論文標題、論文網址」，確認該論文是否合適，並在網路大學公告的網址中留言，論文不可重複，先搶先贏。

◆ 報告資料請於 6/4（日)前上傳至網路大學，需上傳繳交資料為:

1. 書面報告（程式流程圖， 實驗結果， 分析與討論）
2. 報告投影片
3. 程式(論文報告同學免交)
4. 海報(雲端有提供模板參考)

※ 每組只需一個組員上傳至網路大學，不用每個人都上傳。

註 1:程式與論文閱讀均需上台報告，以組別倒序來報告。請於上課前提前來將投影片或程式放入報告的電腦內以節省時間(若持續實施遠距上課則免)。報告時間約 8~10 分鐘(報告 8 分鐘、問答 2 分鐘)。

註 2:選擇程式報告或論文報告均需製作投影片及書面報告，書面報告依照給定格式撰寫(兩欄)，程式報告組別書面至少 800 字以上，論文報告組別書面至少 1500 字以上。書面報告或程式碼經比對系統確認為抄襲，該次作業 0 分。

註 3:書面報告及上台報告分數各佔 50%。

註 4:實現的演算法越多個或閱讀越多篇報告, 成績越佳。

註 5:禁止使用測試集直接對演算法或模型調整及訓練。

註 6:禁止使用網路上的分群程式庫，可以使用「分類」的相關套件。想以深度學習方法完成作業的同學可以使用深度學習框架進行撰寫(tensorflow、pytorch、keras 等)，但嚴禁下載網路(github 等)上任何已經搭建完成且成熟的模型使用。

註 7:未使用分類及分群兩種方法將斟酌扣分。

註 8:請不要直接將黑色為底的程式碼直接截圖放在報告，請調整為適合列印且清楚的格式。

註 9:書面報告請採用黑白及雙面列印，並於上台報告前繳交給老師。海報不需列印。