108 學年度第二學期
資料採礦: 期末考 (TAKE HOME)

日期: 2020/07/25(六), 24:00 前繳交
授課教師：吳漢銘 (臺北大學統計學系副教授)

請仔細閱讀每一個注意事項 (禁止討論)

1. 考試答題要點
   (a) 可參考課本、上課講義 (包含電子檔) 及其它資料。
   (b) 不可與別人 (或同學) 討論，請全部自己做，不可參考同學的答案，不可抄襲。
   (c) 程式設計題，若程式碼直接複製 (或照抄) 講義上的以不給分為原則。
   (d) 請依照「R 程式作業繳交方式」，複製 Console「程式執行及結果」至答案卷。
   (e) 圖形複製，請注意大小，內容數字文字需可辨識。
   (f) 請參照下列文件第 2 ∼ 4 頁寫作規定，不按照規定作答者，會扣分。
   `http://www.hmwu.idv.tw/web/teaching/doc/R-how-homework.pdf`
2. 下載題目卷，上傳答題檔案:
   (a) 於課程網站下載題目卷。
   (b) 上傳答題檔案: 於教師網站首頁登入 [作業考試上傳區]，帳號: `dm108`。密碼: `xxx` (上課教室號碼)。
   (c) 請上傳「學號-姓名-DM-FinalExam.docx」。(改成自己的學號及姓名)(請注意「正確目錄」)
   (d) 若傳錯，請最終要上傳一份正確的的答題檔案。
   (e) 若上傳檔案格式錯誤，內容亂碼，空檔等等問題。請自行負責。
   (f) 若要重覆上傳 (第 2 次以上)，請在檔名最後加「−2」、「−3」，例如: 「學號-姓名-DM-FinalExam-2.docx」、「學號-姓名-DM-FinalExam-3.docx」等等。
   (g) 上傳兩次 (含) 以上、格式不合等等酌量扣分。
   (h) 如果上傳網站出現「You can modify the html file, but please keep the link 'www.wftpserver.com' at least.」，請將滑鼠移至「網址列」後，按「Enter」即可。若再不行，請換 (IE/Edge/Firefox/Chrome)。

我已經仔細閱讀上述各注意事項，若有違背，會自行負責。

R: 資料探勘

1. **維度縮減**

   資料來源 (UCI): `https://archive.ics.uci.edu/ml/datasets/glass+identification`

   Glass Identification Data Set 是玻璃識別資料集，共有 214 個觀察值，具有 9 個變數 (RI, Na, Mg, Al, Si, K, Ca, Ba, Fe) 及一個類別變數 (class，6 個類別)。

   (a) 使用 `PCA {FactoMineR}` 對此資料進行主成份分析。((1) 印出 Eigenvalues/Variances, 畫出 scree plot。(2) 畫出 Circles of Correlation 圖，圖上各點以 square cosine (cos2 values) 之值為顏色標示出來。(3) 畫出 214 觀察值投影於前兩個主成份之散佈圖，並以不同顏色標示 6 個類別。)

   (b) 對此資料進行 MDS 及 ISOMAP 維度縮減方法，並畫出維度縮減後的資料於二維平面的投影散佈圖，同時以不同顏色標示 6 個類別 (請自行選用合適之輸入參數)

2. **群集分析**

   資料來源 (UCI): `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra`

   Breast Cancer Coimbra Data Set, clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls.

   此資料共有 116 人，10 個變數。選取其中 8 個連續變數為解釋變數，記作 $X_{116\times 8}$，(Quantitative Attributes: BMI (kg/m2)，Glucose (mg/dL)，Insulin (µU/mL)，HOMA，Leptin (ng/mL)，Adiponectin (µg/mL)，Resistin (ng/mL)，MCP-1(pg/dL); 剩餘 2 個變數為反應變數，其中 1 個為類別變數 `Labels` (1=Healthy controls，2=Patients)，另 1 個為連續變數 `Age` (years)。

   (a) 以 R 套件 `pheatmap` 畫出此資料 ($X_{116\times 8}$)(列及欄皆未排序) 的熱圖, 並於此熱圖旁加上 `Labels` 及 `Age` 之色條。(請各選合適之色階，需先進行變數標準化，圖上之列及欄位名稱皆需清晰可辨識)

   (b) 同上小題，對此資料 ($X_{116\times 8}$) 進行階層式群集分析 (two-ways hierarchical clustering, complete-linkage)。(請選合適之色階及距離或相關量測指標)

   (c) 同上小題，利用 R 套件 `clValid`，對此資料 (116 人) 進行群集驗証 (cluster validation，internal and stability)，以下列三種分群方法來比較: K-means, PAM, and the hierarchical clustering。

3. **分類法則**

資料集: Microarray gene expression dataset from Khan et al., 2001.

This dataset (subset of 306 genes) can be obtained from R/Bioconductor package "`made4`": `https://www.bioconductor.org/packages/release/bioc/html/made4.html` Khan contains gene expression profiles of four types ("EWS", "BL-NHL", "NB" and "RMS") of small round blue cell tumours of childhood (SRBCT) published by Khan et al. (2001). The `khan$train` data.frame is a subset of Khan that contains 306 genes with 64 arrays.

(a) 以 Fisher criterion (BSS/WSS) 選出前 30 個 BSS/WSS 值最大之 genes。(C03: 分類法則 Classification，第 35/141 頁)。

(b) (承上小題) 利用 R 套件 caret，使用下列分類方法 (`"knn", "rpart"`，`"rf"`，`"adaboost"`，`"svmRadial"`，`"xgbTree"`) 預測此資料 SRBCT 之四個子型，並印出各分類方法之 10-fold CV error rates。(請自行選用合適之輸入參數)

4. **關聯性分析**

資料來源: `swiss datasets` 資料名稱: swiss: Swiss Fertility and Socioeconomic Indicators (1888) Data

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. A data frame with 47 observations on 6 variables, each of which is in percent(百分比), i.e., in [0, 100].

(a) 請依照「C04: 關聯性分析 Association Rules」中「AdultUCI」之範例，將 `swiss` 資料集，轉成"`transactions`" 之類別。(可將各變數離散化，例如取「低、中、高」)

(b) (承上小題) 利用 `apriori {arules}` 找出依 lift 排序之下，前 10 個關聯法則。(請自行選用合適之輸入參數)