Macau Talent Program 2018

Introduction to Computational Drug Discovery Techniques

Use R to develop Machine Learning Models

【前言】

本次試驗涉及 R 語言的使用，fasta 文件的讀取和 Random Forest 與 SVM 預測模型的建立。

R 是用於統計計算和圖形的免費軟件環境。 它編譯並運行在各種 UNIX 平台，Windows 和 MacOS 上。

FASTA 格式中的一條完整序列，包含開頭的單行描述行和多行序列數據。描述行行首前置半角大於號（ “>” ）以和數據行區分。 “>”後緊接的內容為該序列的標識符，該行剩餘部分則為序列的描述（標識符與描述均非必須）。 “>”和標識符之間不應有空格，且建議將單行內容限制在80字符以內。序列的結束以下一條序列的“>”出現為標識。如下為 FASTA 格式一條序列的示例：

>gi|31563518|ref|NP_852610.1| microtubule-associated proteins
MKMRFFSSPCGKAAVDPADRCKEVQQIRDQHPSKIPVIIERYKGEKQLPVLDKTKFLVPDHVNMSELVKI
IRRRLQLNPTQAFFLLVNQHSMVSVSTPIADIYEQEKDEDGFLYMVYASQETFGFIRENE

【步驟】

1. 在桌面空白處單擊鼠標右鑑，點擊 Open Terminal 打開終端。
2. 觀察 fasta 文件
   - 輸入指令： vi ~/Desktop/lab4/LEE-positive.fasta 觀察 fasta 格式的文件
   - 使用上下箭頭移動光標
   - 輸入指令：:q!（冒號 q 感歎號）退出文件
3. 從 fasta 文件中讀取氨基酸序列
   - 輸入指令 R，進入 R 命令環境
   - 輸入指令 require(seqinr)， 引用 seqinr 程序包處理 fasta 文件。
   - 輸入 protdata = read.fasta('~/Desktop/lab4/LEE-positive.fasta',seqtype="AA",as.string=TRUE)，讀取 fsata 文件。
   - 輸入 seqs=unlist(unlist(getSequence(protdata,as.string=T),recursive=F))，得到所有的氨基酸序列。
   - 輸入 head(seqs,3)觀察上一步得到的序列
     - 結果應如下：

       [1] "ACDCRGDCFCGGGGIVRRADRAAVP" "AFGMALKLLKKVL"
       [3] "AIGKFLHSAKKFGKAFVGEIMNS"

- 結果中的三個字符串分別表示三個不同的氨基酸序列（抗菌肽肽鏈）。氨基酸共有20種，分別用20不同的大寫字母表示，每個大寫字母代表一種氨基酸，多個氨基酸組合在一起形成肽鏈，多個肽鏈結合形成蛋白質。

4. 由氨基酸序列生成特徵值
   - 輸入 source('~/Desktop/lab4/lab4.R')，讀入 lab4.R 文件中的所有函數。
   - 輸入 pn=read_pn_seq()， 同時讀取所有的 AMP 序列（LEE-positive.fasta）和所有的 non-AMP 序列（LEE-negative.fasta）。
   - 輸入 data_aac=gene_ftdata(pn,'AAC')，生成特徵值。
   - 輸入 head(data_aac,3)觀察上一步得到的特徵值。
     - 結果應如下：

```
        A  R         N  D  C         E Q         G         H
1 0.1600000 0.16 0.00000000 0.12 0.16 0.00000000 0 0.20000000 0.00000000
2 0.1538462 0.00 0.00000000 0.00 0.00 0.00000000 0 0.07692308 0.00000000
3 0.1304348 0.00 0.04347826 0.00 0.00 0.04347826 0 0.13043478 0.04347826
         I         L         K         M         F  P         S T W Y
1 0.04000000 0.00000000 0.0000000 0.00000000 0.04000000 0.04 0.00000000 0 0 0
2 0.00000000 0.30769231 0.2307692 0.07692308 0.07692308 0.00 0.00000000 0 0 0
3 0.08695652 0.04347826 0.1739130 0.04347826 0.13043478 0.00 0.08695652 0 0 0
          V class
1 0.08000000     1
2 0.07692308     1
3 0.04347826     1
```

     - 第一列的1，2，3代表行名稱，每行對應一個肽鏈，第一行是對應3中 head(seqs,3) 輸出的第一個肽鏈（"ACDCRGDCFCGGGGIVRRADRAAVP"）的20個特徵值和相應的類別（class）；
     - A，R，N，D … … W，Y，V 代表特徵名稱，本例中他们代表二十个氨基酸分别在当前肽链的比例。
       - 例如：肽链"ACDCRGDCFCGGGGIVRRADRAAVP"（3中讀入的第一個氨基酸序列）共含有25個氨基酸(可輸入 nchar(seqs[1])) 得到此序列的長度，也可手動數出序列長度)，其中 A（丙氨酸）共4個，A 所佔比例：4/25 = 0.16。
     - 最後一列 class 代表此行的類別，本例中 class 值只包括0和1，1代表此行對應的肽鍊是抗菌肽，0代表不是抗菌肽。

5. 建立預測模型
   - 建立 random Forest 的預測模型：
     - 輸入 data_aac$class=factor(data_aac$class), 將 data 的 class 列由整數型轉變成因子型便於接下來的分類。
     - 輸入 flds <- createFolds(data_aac$class, k = 2, list = TRUE, returnTrain = FALSE)，將 data 分成兩份。

- 輸入 **train=data_aac[flds[[1]],]**，選第一份做訓練集
- 輸入 **test=data_aac[flds[[2]],]**，選第二份做測試集
- 輸入 rf.mdl <- randomForest(class ~., train,proximity=TRUE) , 得到訓練模型
- 輸入 rf.mdl 观察訓練模型，結果如下：
  - Call:
  - randomForest(formula = class ~ ., data = train, proximity = TRUE)
  - Type of random forest: classification
  - Number of trees: 500
  - No. of variables tried at each split: 4
  - OOB estimate of error rate: 12.56%
  - Confusion matrix:
  - 0 1 class.error
  - 0 197 14 0.06635071
  - 1 39 172 0.18483412
- 輸入 **rf.pre=predict(rf.mdl,test[,-21]),** 得到 random forest 的預測結果。
- confusion matrix

	Actual Positive **實際為陽性**	Actual Negative **實際為陰性**
Predict Positive **預測為陽性**	True Positive (TP) **正確的預測為陽性**	False **Positive**(FP) **錯誤的預測為陽性**
Predict Negative **預測為陰性**	False Negative(FN) **錯誤的預測為陰性**	True Negative(TN) **正確的預測為陰性**

- ACC 正確率(accuracy)公式： ACC = (TP+TN)/(TP+TN+FP+FN)
- Sn 敏感性(sensitivity) 公式：Sn= TP/P = TP/(TP + FN) (P：實際為陽性的總個數)
- Sp 特異性(specificity) 公式：Sp= TN/N = TN/(FP+TN) (N：實際為陰性的總個數)
- 輸入 **original=test$class,** 得到實際種類
- 輸入 prediction=rf.pre, 得到預測種類
- 假設0是陰性 negative，假設1是陽性 positive
- 輸入 TP=sum(original[which(prediction==1)]==1)，得到 TP。
  - prediction==1，判斷預測類別是否為陽性
  - which(prediction==1)，找到預測類別為陽性所在的位置
  - original[which(prediction==1)]，尋找預測類別為陽性的相同的位置的實際類別

- original[which(prediction==1)]==1，判斷預測類別為陽性的相同的位置的實際類別是否為陽性
  - sum(original[which(prediction==1)]==1)，統計預測類別預測結果為陽性的相同的位置的實際類別是否為陽性的總個數，既 TP **正確的預測為陽性**
  - 輸入 **TN=sum(original[which(prediction==0)]==0)**，得到 TN。
  - 請自行計算 FP，FN，答案在最後
  - 利用公式計算 ACC 和 Sn, Sp
  - 輸入 **rf.mat=list(ACC=ACC,Sn=Sn,Sp=Sp,TP=TP,TN=TN,FP=FP,FN=FN)** 將7個指標放入名為 rf.mat 的表（list）中
  - 輸入 **rf.mat**，观察结果
  - 輸入 **rf.mat$ACC** 调取 ACC 值。
- 建立 svm 的預測模型：
  - 輸入 **svm.mdl <- svm(class~., data=train, kernel = "radial", type="nu-classification")**, 得到訓練模型
  - 輸入 **svm.mdl** 观察訓練模型，結果如下：
    - Call:
    - svm(formula = class ~ ., data = train, kernel = "radial", type = "nu-classification")
    -
    -
    - Parameters:
    - SVM-Type: nu-classification
    - SVM-Kernel: radial
    - gamma: 0.05
    - nu: 0.5
    -
    - Number of Support Vectors: 240
  - 輸入 **svm.pre=predict(svm.mdl,test[,-21]),** 得到 svm 的預測結果
  - 輸入 **original=test$class,** 得到實際種類
  - 輸入 **prediction=svm.pre,** 得到預測種類
  - 請自行計算 SVM 的 TP，TN，FP，FN，答案在最後
  - 利用公式計算 ACC 和 Sn，Sp
  - 輸入 **svm.mat=list(ACC=ACC,Sn=Sn,Sp=Sp,TP=TP,TN=TN,FP=FP,FN=FN)** 將7個指標放入名為 svm.mat 的表（list）中
  - 輸入 **svm.mat**，观察结果, 輸入 **svm.mat$Sn** 调取 Sn 值。

6. 比較 random forest 和 svm 兩個模型的優劣

7. 答案
    - 輸入 TP=sum(original[which(prediction==1)]==1)，得到 TP。
    - 輸入 TN=sum(original[which(prediction==0)]==0)，得到 TN。
    - 輸入 FP=sum(original[which(prediction==1)]==0)，得到 FP。
    - 輸入 FN=sum(original[which(prediction==0)]==1)，得到 FN。
    - 輸入 ACC=(TP+TN)/(TP+TN+FP+FN)，得到 ACC。
    - 輸入 Sn = TP/(TP + FN)，得到 Sn。
    - 輸入：Sp = TN/(FP+TN)，得到 Sp。
    - 我的 rf.mat 结果
        - $ACC
        - [1] 0.8270142
        - $Sn
        - [1] 0.7962085
        - $Sp
        - [1] 0.8578199
        - $TP
        - [1] 168
        - $TN
        - [1] 181
        - $FP
        - [1] 30
        - $FN
        - [1] 43
    - 我的 svm.mat 结果
        - $ACC
        - [1] 0.8507109
        - $Sn
        - [1] 0.7630332
        - $Sp
        - [1] 0.9383886
        - $TP

- [1] 161
-
- $TN
- [1] 198
-
- $FP
- [1] 13
-
- $FN
- [1] 50

課後練習

請 email 你的答案給老師：shirleysiu@umac.mo

1. 請找出現時抗菌肽在醫學或藥物發展上的用途的例子。
2. 機器學習以至深度學習真的能應機器解決所有問題嗎？你能找到一兩個機器學習的局限性嗎？