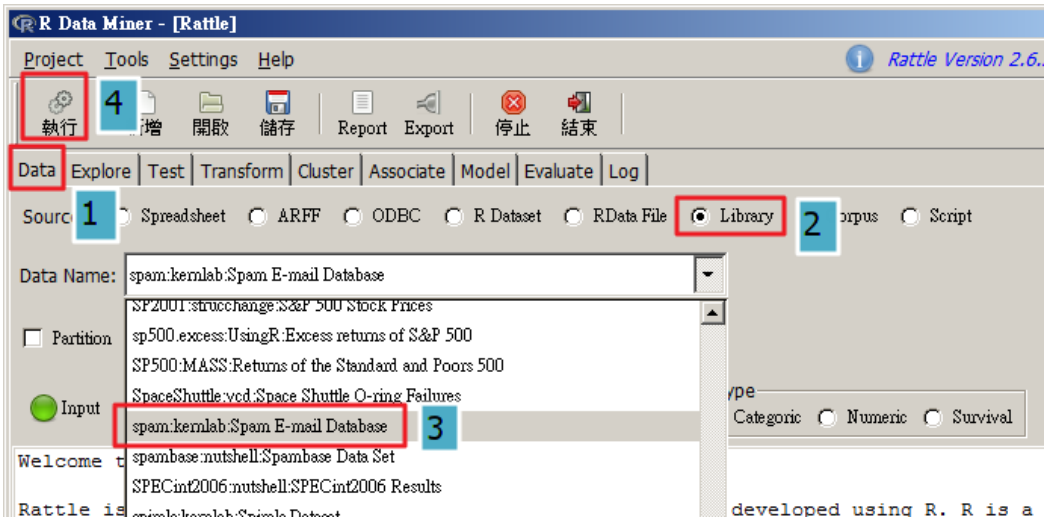
title: Data mining with Rattle – ROC curve
date: 2013.8.10
author: Ming-Chang Lee

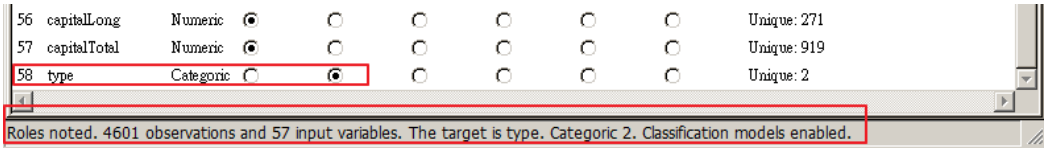
本範例說明採用 R 圖形化使用者介面 rattle 套件 執行 ROC curve 比較
步驟 1
安裝並執行 rattle 套件－資料探勘使用者介面

```
install.packages("rattle")
library(rattle)
rattle()
```

步驟 2
匯入 kernlab 套件的資料集 spam, 一般採用 spam {kernlab} 表示, 步驟如下:
Data \ Source: 選取 Library \ Data Name: 選取 「spam:kernlab:Spam E-mail Database」 \ 按 執行.



匯入 spam 資料結果, 此為垃圾郵件資料集, 其中第 58 個變數是目標變數且為類別型資料, 在 R 中屬於因子(factor)資料物件, 在最下列訊息區顯示全部有 4601 筆觀測值, 58 個輸入變數, 可使用分類模型, 詳如下圖所示:



查詢 spam 資料集指令如下:

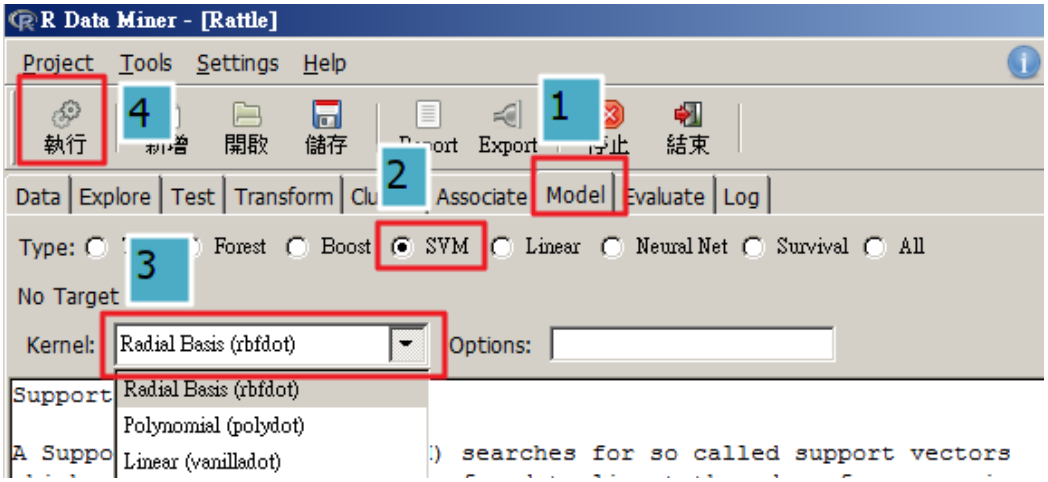
```
library(kernlab)  # 如果之前已輸入過一次, 則此行可省略.
?spam
```

步驟 3

Rattle 的 Support Vector Machines 方法，採用 kernlab 套件的 ksvm 方法，在 R console 視窗中輸入以下指令可查詢其使用說明.

```
library(kernlab)
?ksvm
```

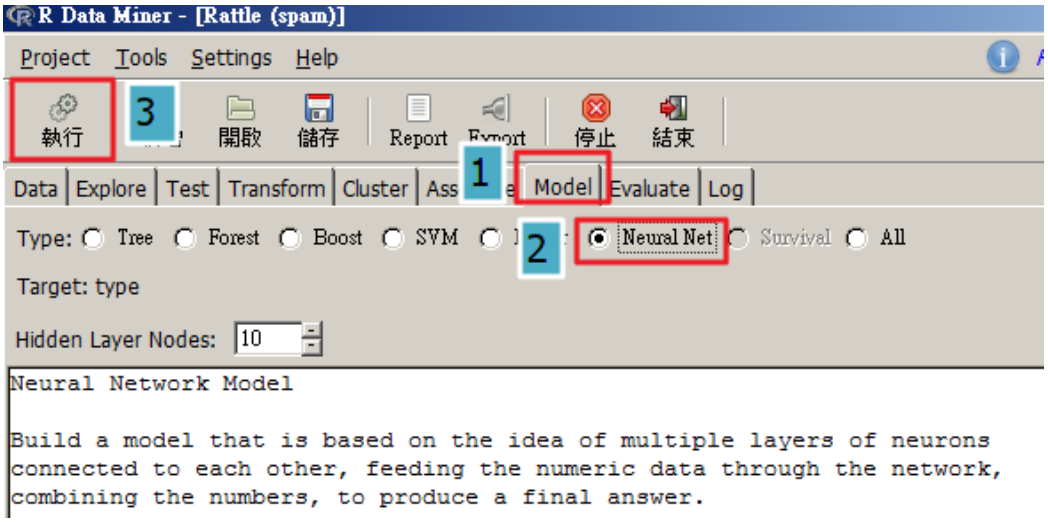
Model \ Type 選取 SVM \ Kernel 採用預設值 Radial Basis (rbfdot) \ 按 執行.



步驟 4

使用 Neural Network 方法，Rattle 採用 nnet 套件的 nnet 方法，該方法採用 Feed-forward neural networks 且為一個隱藏層 (hidden layer)，預設節點數是十個.

Model \ Type 選取 Neural Net \ 按 執行.



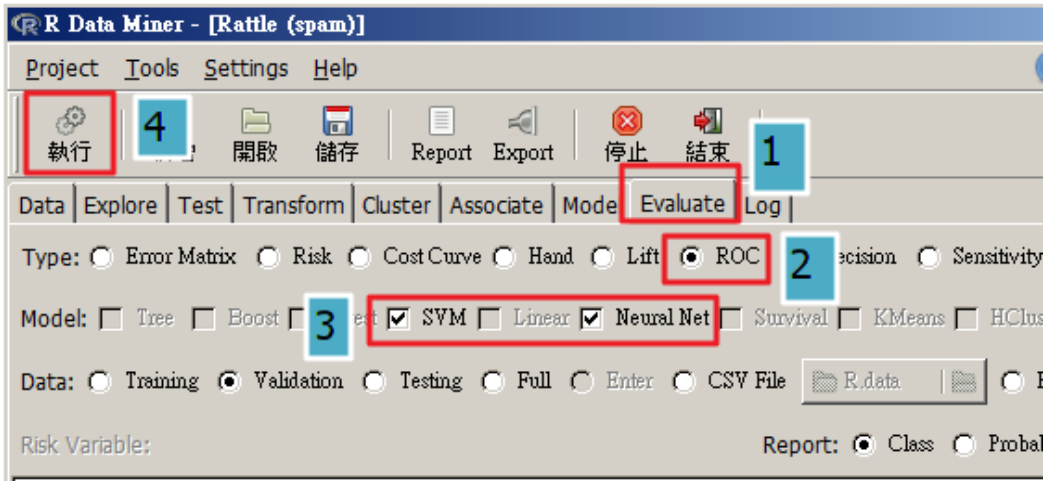
步驟 5

模式比較主要採用 ROC curve 方法.

Link: http://web.ydu.edu.tw/~alan9956/docu/refer/roc_introduction.pdf

Rattle 會針對之前已完成的模型進行效益評估.

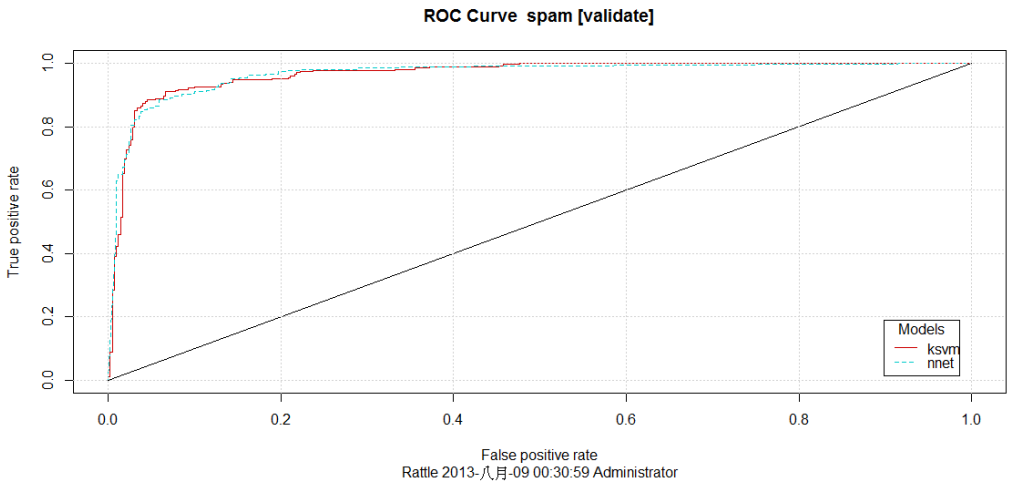
Evaluate \ Type: 選取 ROC \ Model: SVM , Neural Net 打勾 \ 按 執行.



同理上述 Type 有多種選項可供選取, 例如 : Error Matrix 等.

步驟 6

考慮實際結果有二種情形(Yes, No), 在 ROC curve 視窗中 x 軸表示 False Positive (FP) rate (實際為 N, 但預模型分類為 Y, 此時 FP 愈小愈好), y 軸表示 TP rate (實際為 Y, 且預模型分類為 Y, 此時 FP 愈大愈好), 因此, ROC 曲線愈偏向左上角愈好, 即曲線以下面積較大者較佳. 由圖形可知 ksvm 較偏向於左上角位置, 因此採用 svm 較佳。



參考輸出結果可知 svm 面積較大.

```
Area under the ROC curve for the ksvm model on spam [validate] is 0.9664

Rattle timestamp: 2013-08-09 00:30:59 Administrator
============================================================================
Area under the ROC curve for the nnet model on spam [validate] is 0.9663

Rattle timestamp: 2013-08-09 00:30:59 Administrator
============================================================================
```

end