【2018 機器學習 Homework 3】

- **注意事項**

1. 請使用 python 語言，配合anaconda python內含的開發軟體實作本次作業並進行測試，並安裝、使用scikit-learn函式庫。
2. 請依據作業規定設定專案名稱，若未依照規定，將根據狀況扣分。
3. 嚴禁抄襲其他同學作業，參與者(抄襲與被抄襲)皆以零分計算。
4. 請對你的程式碼有深入瞭解，demo 時助教會問。
5. 逾期以零分計算，不接受補交，有任何因素導致無法如期繳交，請事先告知。
6. Demo 時間會另外通知。

- **作業規定與上傳**

1. 專案名稱：Student ID_HW　　EX:M053040086_HW3

2. 作業請繳交專案之 tar 或 zip archive 並上傳至網路大學。

   請於 2018 年 4 月 17 日(週二) 23:59 前
   上傳完畢，逾期以零分計算，不接受補交，有任何因素導致無法如期繳交，請事先告知。

data知識:

每筆資料為一個網絡連接定義為在某個時間內從開始到結束的 TCP 數據包序列，並且在這段時間內，數據在預定義的協議下（如 TCP、UDP）從源 IP 地址到目的 IP 地址的傳遞。每個網絡連接被標記為正常（normal）或異常（attack），異常類型被細分為 4 大類共 39 種攻擊類型，其中 22 種攻擊類型出現在訓練集中，另有 17 種未知攻擊類型出現在測試集中。

4 種異常類型分別是：

1. DOS(denial-of-service)拒絕服務攻擊，例如 ping-of-death, syn flood, smurf 等。
2. R2L(unauthorized access from a remote machine to a local machine) 來自遠程主機的未授權訪問，例如 guessing password。
3. U2R(unauthorized access to local superuser privileges by a local unpivileged use)未授權的本地超級用戶特權訪問，例如 buffer overflow attacks。
4. PROBING, surveillance and probing, 端口監視或掃描，例如 port-scan, ping-sweep 等。

作業目的:請分出每筆連線是一般連線 還是 攻擊連線中的哪一類型(DOS、R2L、U2R、PROBING)，沒限定分類器、前處理步驟，自由發揮，準錯度越高作業分數越高。

參考提示:

1.在 NSL_KDD 裡面中有三種大小的 traindata，你可以試試看不同種大小，出來的效能如何，三種大小的資料名稱: KDDTrain+、20 Percent Training Set、Small Training Set。

2.如果你有網路攻擊知識相關的概念，你可以先試著找找看有沒有多餘的特徵，先進行手動刪除，再交由特徵選擇，選出重要的特徵。

3.你可以嘗試各種不同的分類器、各種不同的特徵選擇、各種不同的降維方法，找出你最佳的組合。

4.HW3 作業的前處理相當重要，請觀察資料看是否需要標準化、正規化……等，請觀察特徵是連續還是離散，請思考如何對連續資料、離散資料進行處理，例如:有些特徵需要進行編碼再丟入分類器，請試試看如何編碼，有編碼效果是否會比較好?

5. 41 種特徵解說:https://blog.csdn.net/com_stu_zhang/article/details/6987632