HW #2

תאריך הגשה 19.1.18 יום שישי בשעה 16:00

הגשה באיחור – כל יום איחור יוריד 5 נקודות מציון התרגיל – אנא הגישו בזמן

ניתן להגיש לבד או בזוגות. אם עבדת בזוג – לציין בקובץ וורד עם מי עבדתם.

- יש להגיש קובץ וורד המתאר בקצרה את התהליך בקוד ואת התוצאות.
- יש להגיש את קובץ הנתונים איתו עבדתם, אפשר להשתמש בקובץ שאעלה לתיקיה.
- יש להגיש את קובצ/י מטלב.

1. Read Sections I and II of the paper **An ensemble approach for feature selection of Cyber Attack Dataset** to understand the gathered data. The paper is under the HW folder.

2. Prepare a data set:

   - Data – go to **KDD Cup 1999 Data Data Set** in UCI and download kddcup.data_10_percent.gz from the data folder.

   - This file is too long.. randomly erase rows of data. Keep about 10,000 rows.

     Check that the last column contains some **normal** rows and some **attack** rows.

   - For this HW, in the last column: replace **normal** with **0,** replace all other values with **1.**

   - Erase columns 2 and 3 and 4 that contain categorical values for the `protocol_type` (tcp, udp..), service (http, smtp…) and flags (SF, S0, ..).

   - You should end up with a data set that has 10000 rows , 37 feature columns

     and 1 binary label column. Instead, you can use the file I uploaded.

3. Separate the data into train and test by using a 5-fold cross validation procedure

4. Use PCA to project the train data:
    - Normalize the train data and save the means and variance.
    - Apply matlab's ***pca*** function.
    - Plot the eigenvalues and decide how many coordinates to keep.
    - Generate 2 or 3 plots using ***figure; scatter3***(….) . color the dots by the value of their label. (plot for example [Z(:,1), Z(:,2), Z(:,3)], [Z(:,2), Z(:,3), Z(:,4)], [Z(:,5), Z(:,6), Z(:,7)]).

5. Project the test data onto the first PC's:
    - Normalize the test data.
    - Use the matrix W (output of ***princomp***) to project the test data.

6. Use Logistic regression to classify the projected data into **normal vs. attack.**

7. Use SVM to classify the projected data into **normal vs. attack.**
    - Use ***svmtrain*** on the projected train data.
    - Use ***svmclassify*** to classify the projected test data.

    Optional: Repeat with Kernel SVM (choose one kernel)

8. Compute the correct classification (confusion matrix) for the Logistic regression and the svm classifiers.

Good Luck!