Project

Notes 1: Consider it as fun part of your learning and don't take it as a burden or assignment with a forced deadline. Do it by yourself.

Notes 2: There are is no solution to this project. I will provide material in the form of scripts that you will give you a quick start.

Note 3: If you feel difficulties in understanding some task of the project, post your question in the Q/A section of the course. Do not forget to mention the task number you are querying about.

Note 4: If you feel you have some exciting tasks, please inbox to me and I will add to the task list there after review. This will help you fellows to have more practice and fun.

Note 5: **Have Fun**

Project: Malware Dataset Analysis

In this project we are going to analyze the malware data. Follow the video tutorials and complete the following tasks.

Task 1: Description

Objective: The objective in this task to to learn how to compare different classifiers and how to show the results in the form of a bar graph. This task has two parts. Complete these tasks based on the Holdout method with 80 -20 split.

Task 1a: Show the bar graph of accuracies of all the classifiers with default options. Share your results with me.

Task 1b: Show the bar graphs of precision and recall of all the classifiers with default options for the malicious category, i.e., category 0. Two graphs are required one for precision and another one for the recall.

Hint: use the script Classification_template.m which is in the folder

\Machine Learning for Data Science using MATLAB\Classification\K-Nearest Neighbor

And comment out the visualization part since the data contains more than two variable.

Task 2: Description

Objective: The objective in this task to learn how to customize the classifiers based on your dataset.

Task 2a: Customize each of the classifiers covered in the course based on its respective customization options and report the best accuracy results for each classifier in the form of a bar graph. (Note: use the customization script for the classifiers)

Task 2b: Based on the customization done in Task 2a, (the customization option which leads to best accuracy) report the precision and recall values for each of the customized classifiers for the malicious category in the form of the bar graph.

Note: Again assume the validation method of Holdout with 80 -20 split.

Task 3: Description

Objective: The objective in this task is to learn how to apply different validation method and what is its impact on the customization options.

Task 3a: Now change the validation method to kfold with k = 5 and compute the overall accuracies (not individual fold accuracy) of all the classifiers covered in the course in the form of a bar graph.

Task 3b: Customize each classifier based on the kfold with k = 5. The customization is based on best accuracy. Based on the selected customization,

A. Report whether the customization for each classifier is the same as the customization achieved with holdout,
B. Report a bar graphs of accuracies for customized classifiers.
C. Report a bar graph of the precision and recall for the malicious category.

Note: To complete this task use the template Validating_Classifier_Performance.m which is in folder

\Machine Learning for Data Science using MATLAB\Classification\Evaluation

Task 4: Description

Objective: The objective in this task to visualize the data. However, since the data contains more than two features, therefore we are unable to visualize the data directly. So we therefore see how to apply PCA and what are its results in terms of classification accuracy.

Task 4a: Apply the PCA algorithm on the data and show the visualization of the decision boundary for all the classifiers. Show the bar graph of accuracies of all the clasasifiers with 80 - 20 split and default options. Use the script PCA_algorithm.m

Task 4b: Now with changed validation method, i.e., kfold with k = 5 and customized options selected for kfold in Task 3b apply the PCA algorithm and compute the overall accuracy of each classifier. How does these accuracies compare with those found without PCA in task 3b.

Task 5: Description

Objective: The objective in this task is to look at the different window sizes and its impact on accuracy.

Task 5a: Grab the datasets corresponding to different window sizes. Use the setting of kfold with k = 5 and best customization options for each classifier selected in Task 3. Show the accuracies of different classifiers in the form of a bar graph based on different settings of the window sizes (one graph for each window size having all the classifers). You may also want to apply the PCA. The choice of whether or not to use PCA may be justified based on your analysis in task 4.