SIT384 Cyber security analytics

High Distinction Task 6.3HD: Comparing Classification Models

Task description:

You are asked to apply different classification techniques on a given Spambase dataset, to potentially enhance the accuracy of the learnt models via selecting better parameters, and/or pre-processing etc., to compare the results, and to summarize your findings in a report.

The Spambase Data Set can be retrieved from https://archive.ics.uci.edu/ml/datasets/Spambase or be directly downloaded from task resources.

The classification algorithms to apply are:

1. K-Nearest Neighbours
2. Logistic Regression
3. Random Forest
4. Support Vector Machines

You must compare and interpret the results of using different approaches on the dataset. Other requirements are:

- In your report, briefly discuss how you have selected the parameters for each classification algorithm and how these parameters affect the accuracy of the algorithm. It is ideal to use experimental results to support your discussion.
- In your report after comparing the experimental results of the given classification algorithms, write a paragraph or two trying to explain/speculate why, in your opinion one classification algorithm outperformed the others.
- Finally, at the end of your report provide a 1-2 paragraphs, summarizing the most important findings of this task.
- Try your best to improve the accuracy of each algorithm.

The github website of the prescribed textbook has quite some useful supplemental material (**code examples**, IPython notebooks, etc.), available at https://github.com/amueller/introduction_to_ml_with_python, especially chapter 2.

Submission:

Submit the following files to OnTrack:

1. Your program source code (e.g. task6_3.py)
2. A screen shot of your program running
3. Your comparison and discussion report

Check the following things before submitting:

1. Add proper comments to your code