Machine Learning Project (2)

A) d-index and other classification measures (50 points)

- What's the advantage of d-index?
- How can you calculate d-index for multi-class classification?
- Do SVM classification for Iris data with the first 70% training and remaining 30% for test and calculate the following classification measures and explain their meaning. Which ones are more representative, why?
  - d-index
  - f1-micro
  - f1-macro
  - balanced_accuracy
  - roc_auc_score
  - Hint: `https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py`

B) Bagging-SVM (50 points)

Implement a Bagging SVM for three datasets by following the requirements

credit_risk_small_data

credit_data_simulate.

cybersecurity_data

1. Use first 80% data for training and the remaining 20% for test. Do SVM prediction
2. Randomly pick 1/2 data from training data to form three training datasets: training_1, training_2, and training_3
3. run SVM using training_1, training_2, and training_3 to predict the test data: we can them SVM_1, SVM_2, and SVM_3
4. Determinate the final label for each test entry by doing the following voting
   (a) Pick the predicted label with maximum votes, say 1 (svm_1), -1 (svm_2), 1 (svm_3), then final label should be 1
5. Compare the performance of SVM and the Bagging-SVM for the two datasets and draw your conclusion

What should you turn in?

- 1. A folder that contains
  - A ppt to show details of your analytics (at LEAST 40 pages)
  - your data
  - source files
  - corresponding related output.
  - A link to your presentation video
- 2. Send the zipped file (.zip instead of ,rar)