CSCE 5300 Introduction to Big data and Data Science

ICE-9

Lesson Title: Machine Learning

Lesson Description: machine learning models (decision tree, random forest, Naive Bayes, SVM)

In Class Exercise:

You can use google colab (https://colab.research.google.com) or Jupiter notebook (run spark on your own laptop) run different on the given dataset and explain the models or algorithms.

Source code given on Canvas.

1. Run decision tree model on the dataset. Change the depth of tree to improve the performance. Finish coding parts of evaluation methods. Explain how model work and reason of improvement of performance.

2. Run random forest model on the dataset. Change the number of trees to improve the performance. Finish coding parts of evaluation methods. Explain how model work and reason of improvement of performance.

3. Run naïve bayes model on the dataset. Change the thresholds to improve the performance. Finish coding parts of evaluation methods. Explain how model work and reason of improvement of performance.

4. Run SVM model on the dataset. Change the max number of iterations to improve the performance. Finish coding parts of evaluation methods. Explain how model work and reason of improvement of performance.

ICE Submission Guidelines

1. ICE Submission is individual.

2. ICE code has to be properly commented.

3. The documentation should include the screenshots of your code/queries and results.

4. Provide the explanation of the exercise for each question as per your understanding.

5. The similarity score for your document should be less than 15%.

6. Submit the source code (if any) properly commented and documentation (.pdf/.doc)

with explanation and screenshot of source code/queries having input logic and output

results.

7. Submission after the deadline is considered as late submission.

References:

Jupiter+spark:

https://www.sicara.ai/blog/2017-05-02-get-started-pyspark-jupyter-notebook-3-minutes