Homework-5: Spark Machine Learning

Deadline: October 24th, 11:59 PM ET

In this homework, you are expected to practice spark machine learning library.

Your homework submission should be on GitHub. Use the following GitHub classroom to access the assignment and create your assignment repository: https://classroom.github.com/a/xfON7v-1

You should submit the URL for your GitHub repository on Canvas. Grading penalty will be applied if otherwise.

Cite any external sources you use. External sources shouldn't exceed more than 30% of the final solution.

Submit your answers in one (or more) Jupyter notebook(s) **but clearly identify in your ReadMe file the location of the answers for every question**. If you need to include output or screenshots, you may include them in your notebook or attach them as separate image files using the naming convention qx.png (where x is the question number). If you have more than one image for a given question, name the files as qx_1.png, qx_2.png, etc.

Q1. (20%) SparkML builds on the concept of Transformers, Estimators, and Pipelines. In Lecture 8, we created a pipeline for data engineering and preprocessing, but the pipeline does not include machine learning models.

In this question, please build a **single Pipeline** that conducts not just data engineering and preprocessing as in the lecture, but also machine learning (including hyper-parameter tuning) using the logistic regression model. You may use the same data engineering stages as in the lecture, and you may choose your own hyper-parameter grid.

Fit your pipeline to the raw training dataframe (loaded from file) and then use the fitted pipeline to transform the raw test dataframe. Print the schema and the first few rows of the transformed test dataframe.

Note that **only one single pipeline** should be created. The input to the fitted pipeline should be the raw dataframe loaded from file, and the output should contain predictions made by the tuned model. In your submission, include a screenshot of your code that shows all the stages in the Pipeline you created.

Q2. (80%) Multi-Class Classification. In this question, you will use sparkML to predict not just whether there is an attack, but also the type of the attack. In other words, this is a multi-class classification problem with 5 possible categories (normal, DOS, R2L, U2R, probing)

- Q2-1 (15%) Create a preprocess pipeline that conducts the usual data engineering steps (as in the lecture), and also create an outcome column that indicates which of the 5 categories the record belongs to.
- Q2-2 (20%) Select 2 machine learning models. For each of the machine learning models, train it on the training dataset. Calculate the train and test accuracy, and plot the confusion matrix (for the predictions on the test dataset).
- Q2-3 (20%) For each of the 2 machine learning models, identify at least one hyper parameter, build a parameter grid and conduct hyper-parameter tuning using cross-validation, with accuracy as the metric. Calculate the test accuracy after tuning. (**Hint**: for evaluating accuracy, the you may need to find the appropriate Evaluator with the appropriate metric name).
- Q2-4 (5%) In your own words, explain why you chose the two machine learning models, and for each of the two models, why you chose the hyper-parameter(s) to tune, and how you designed the parameter grid. Also, please include a discussion on the comparison of the two models.
- Q2-5(10%) Pick one of the two ML models and run the training process (don't need to do hyper-parameter tuning) on the google cloud DataProc cluster in ***cluster*** mode. Show a screenshot of the spark history server, including the Jobs page and the Executor page.
- Q2-6 (10%) Continuing from Q2-5, on the DataProc cluster, pick a dataframe and show how many partitions the underlying RDD has.