Exercise 3 Report

Akash Shingha Bappy [2307938]

In this exercise, we were introduced to Spark MLlib which is a scalable machine-learning library that offers a wide range of algorithms optimized for distributed computing. It runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud, against diverse data sources. MLlib supports model building, hyperparameter tuning, and evaluation, making it ideal for big data machine learning tasks.

In machine learning, ML Pipelines are like pathways that help us build and train models efficiently. One really good thing about using ML Pipelines is that they help us tune the settings, called hyperparameters, of the models. For example, In our ExampleData dataset, there are lots of different features like date, humidity, wind speed, and others where the task was to predict something, like air temperature, based on these features.

Now, tuning hyperparameters is like finding the perfect settings for the model. It's super important because it can make the model perform much better. For example, if we're predicting air temperature, we want the model to be as accurate as possible so that we can make a better prediction. Hyperparameters control how the model learns from the data, and finding the best ones can make the predictions much more accurate.

In Spark, there are tools called ML Pipelines that help us with this. As the accuracy of a model varies on different data based on the parameters manually tweaking it takes a lot of time and effort. With ML Pipeline, they let us try out different combinations of hyperparameters automatically, saving us a ton of time and effort. So, it makes it easier for us to find the best settings for the models, which leads to more accurate predictions.

In our specific case, where we are trying to predict air temperature from features like date, humidity, wind speed, and msl, using ML Pipelines with models like Gradient-Boosted Trees or Random Forests is really helpful. These models are powerful, but their performance depends a lot on having the right hyperparameters set. With the right environment and data, it can assure good efficiency, consistency, scalability and Reproducibility.

After training the models with different hyperparameters using ML Pipelines, we can evaluate them using metrics like RMSE (Root Mean Squared Error), R-square, and MAE (Mean Absolute Error) to see which one performs the best. Then, we can use libraries like Seaborn and Matplotlib to visualize the results, making it easier to understand how well the models are doing and which one is the best for predicting air temperature.

Therefore, it can be said that hyperparameter optimization is indeed a big benefit of using ML Pipeline as with proper utilization, it can ensure great accuracy in less time and effort.

Reference

1. MLlib | Apache Spark
2. ML Pipelines - Spark 3.5.1 Documentation (apache.org)
3. Classification and regression - Spark 3.5.1 Documentation (apache.org)
4. Machine Learning Pipelines: Benefits, Challenges, Use Cases (plat.ai)