Machine Learning in Databricks with Spark ML

Brandon Ritchie, Rachelle Ceron, Spencer Carillo, Gavin South, Trevor Neri

Why Spark machine learning?

- Scale
  - Process more data than can fit in any one machine
  - More data == performant models
- Works with pre-existing pipelines and tools
  - Spark (streaming, ETL, ad hoc analysis, reporting)
  - Frameworks (sklearn, Tensorflow and Horovod, R)
  - Languages (Python, R, Scala, SQL, Java)
- Model training and production model serving

Spark MLLib (Dataframe Based) vs Spark MLLib (RDD Based)

MLLib (RDD)

- Deprecated spark machine learning library
- Built on RDD's (https://spark.apache.org/docs/latest/ml-guide.html)

MLLib (DataFrame)

- Built on dataframes that translate to the RDD structure
- More user friendly API.

Important Note

Both methods are actually inefficient with smaller datasets. If you are working with smaller data, sci-kit learn is the way to go with RAM stored data.

What is an RDD?

Resilient Distributed Dataset - A collection of elements of data, partitioned across nodes that allows processes to be run in parallel.

The dataframe structure in Spark ML allows for seamless translation to RDD's.

(https://databricks.com/glossary/what-is-rdd)

5 Minute Machine Learning Review Dump

Preprocess Data - Manipulating data to enhance performance of model

This link will be your best friend!!! (https://spark.apache.org/docs/1.4.1/ml-features.html)

- Feature engineering (done)
- Quantify categorical data
  - One-hot encoding
    - https://stackoverflow.com/questions/42295001/how-to-interpret-results-of-spark-onehotencoder
  - Label encoding
  - Clustering
- Interpolation to fill NA's
  - https://www.analyticsvidhya.com/blog/2021/06/power-of-interpolation-in-python-to-fill-missing-values/
- Split data into training and testing
  - https://spark.apache.org/docs/2.1.0/ml-tuning.html#train-validation-split
- Feature scaling (https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf)
  - Prevents large features from dominating the model
  - Standardization - Centers data with mean 0 and a standard deviation of 1
  - Normalization - Rescales values from 0-1 -Spark ML StandardScalar()

5 Minute Machine Learning Review Dump

Fit Model to Data

Refer to documentation for Spark ML models under MLlib (DataFrame-based) tab

https://spark.apache.org/docs/latest/api/python/reference/pyspark.ml.html

5 Minute Machine Learning Review Dump

Evaluate Model

- Accuracy
  - Am I predicting more correctly than incorrectly
  - (TP + TN) / (TP + TN + FP + FN)
- Precision
  - How well can you pick 5 bad apples out of 1,000,000 total apples
  - TP / (TP + FP)
- Recall
  - How many bad apples are predicted out of all of the bad apples
  - TP / FN
- F-Score (No strong goal and want to do well with both precision and recall)
  - Harmonic mean of precision and recall
  - Less interpretable, but big number good small number bad
- Validation Curves
  - Plot error by training iteration for training and validation data

Coding Walkthrough

Follow the link below and make sure to clone it to your repository!

[Put link to walkthrough]