MACHINE LEARNING USING SPARK AT SHARCNET

Jose Nandez, PhD
jnandez@sharcnet.ca
Big Data Specialist
SHARCNET/Western University
March 2017
 WHAT IS APACHE SPARK?

- Apache Spark or Spark is a fast and general engine for processing large-scale datasets
- Spark extends the MapReduce model, supporting interactive queries and stream processing
- Spark has the ability to run computations in memory or disk (MapReduce) depending on the complexity of the problem
- Spark is designed to work on batch applications, iterative algorithms, interactive queries, and streaming.
- It has API for Python, Scala, Java, R, and SQL
• Spark SQL lets you query structured data
• Spark Streaming lets you ingest live data streams (such as Twitter data)
• MLlib is a scalable machine learning library (this will check today)
• GraphX is for graphs and graph-parallel computation for graph analysis (such as Facebook)
MLlib is Spark’s machine learning (ML) library. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as:

- **ML Algorithms**: common learning algorithms such as classification, regression, clustering, and collaborative filtering
- **Featurization**: feature extraction, transformation, dimensionality reduction, and selection
- **Pipelines**: tools for constructing, evaluating, and tuning ML Pipelines
- **Persistence**: saving and load algorithms, models, and Pipelines
- **Utilities**: linear algebra, statistics, data handling, etc.
- **Dataframes**: The Spark 2+ API uses DataFrame from Spark SQL as an ML dataset.
ML ALGORITHMS

Classification:
- Logistic regression
- Decision tree classifier
- Random forest classifier
- Gradient-boosted tree classifier
- Multilayer perception classifier
- One-vs-Rest classifier
- Naïve Bayes

Regression:
- Linear regression
- Generalized linear regression
- Decision Tree regression
- Random forest regression
- Gradient-boosted tree regression
- Survival regression
- Isotonic regression

Clustering:
- K-means
- Latent Dirichlet allocation
- Bisection k-means
- Gaussian Mixture Model

Collaborative Filtering:
- Alternating Least Squares (ALS)
They include feature transformers:

- This could take a Dataframe, read certain columns and map it into a new one
- The output can the feature vectors, or a column for further transformation

Transformer also include learning models:

- A learning model could take a Dataframe and predict a the label (this is a transformation)

A transformer implements the transform() function

It converts a Dataframe into a new Dataframe

There are some Feature Transformers, Feature Extractors, Feature Selectors which are part of the so-called “Featurization”. These are functions meant to transform your data for optimal use of the Spark ML.
ESTIMATORS

- This is used for learning algorithms or any algorithm that fits or trains on data
- It used the fit() function
- Accepts a Dataframe and produces a Model => Transformer
It is used for running a sequence of algorithms to process or learn the data.

The workflow is represented by Pipeline.

The sequence is given by PipelineStages, sequence of Transformers and Estimators.

A pipeline is an estimator, then uses fit() function. This will get a Transformer.

Pipelines are a concept from sklearn from Python. There is also an R pipeline model, but it is not well tested as the sklearn.

Note that the Spark Pipeline model was inspired by sklearn.
EXAMPLE OF PIPELINE

- Problem: Classify whether the income of an Adult is “<=50” (0) or “>50” (1).
- Data: age, education, marital status, year of education, nationality, race. Label = {“<=50”, “>50”}.
- Solution: Transform the label column into an index (0 or 1). Create a Feature Col (this a single vector column). Then apply Logistic Regression Model. Optional: change the predicted column into the original label.
PipelineModel (Transformer)

PipelineModel.transform()

Raw Cols → Indexed Label → Single Vector Col

StringIndexer → VectorAssembler → Logistic Regression Model

Predictions
WHY SPARK ML?

- Spark uses fault tolerant data structure
- Spark ML is a distributed ML library. This means that the ML algorithm can in multiple nodes, making the training and prediction method faster for really large data sets (PB of data).
- Spark can read CSV, JSON, Parquets, text files, JDBC, and then apply ML algorithms.
WHERE TO FIND HELP IN SHARCNET?

- https://www.sharcnet.ca/help/index.php/Apache_Spark
- help@sharcnet.ca
- Or email me (jnandez@sharcnet.ca) if you want to know more about Spark
- https://www.sharcnet.ca/help/index.php/JUPYTER (this links will help you set up a notebook on vdi-fedora23)
REFERENCES

- Learning Spark: Lightning-Fast Big Data Analysis By Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia
- Advanced Analytics with Spark Patterns for Learning from Data at Scala By Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills
- http://spark.apache.org/