Spark ML

Taryn Heilman
Kristie Wirth
Adam Richards
March 15, 2018

galvanize

Learning Objectives

galvanize

- Review feature engineering in pyspark using spark SQL
- Review the concept of ml pipelines and introduce the spark implementation
- Describe one difference between Spark & Sklearn.
- Explain the concept of a transformer.
- Explain the concept of an estimator.

Review relevant concepts

galvanize

- What is a pipeline in sci-kit learn? How do they work?
- Compare and contrast Spark and Hadoop MapReduce
- Compare and contrast RDDs and DataFrames in spark
- What does a partition and how do they affect performance?

Machine Learning on Spark

galvanize

- 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.

Timing of Algorithms in Spark

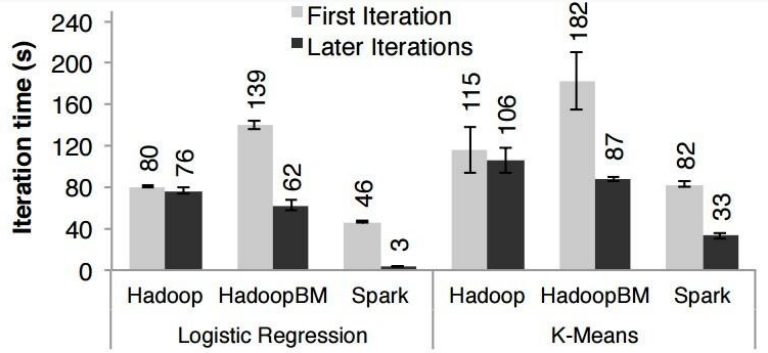
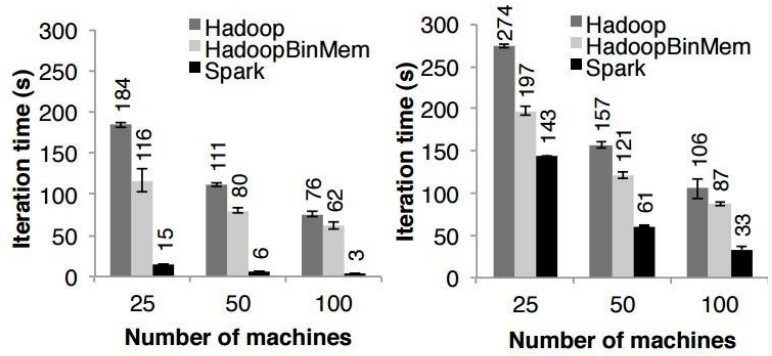


Figure 7: Duration of the first and later iterations in Hadoop, HadoopBinMem and Spark for logistic regression and k-means using 100 GB of data on a 100-node cluster.



(a) Logistic Regression (b) K-Means

Figure 8: Running times for iterations after the first in Hadoop, HadoopBinMem, and Spark. The jobs all processed 100 GB.

Spark Machine Learning Pipeline Terms

galvanize

Pipeline

- Running a sequence of algorithms in a set order to process & learn from data
- Many Data Science workflows can be described as a pipeline, i.e. just a sequential application of various Transforms and Estimators

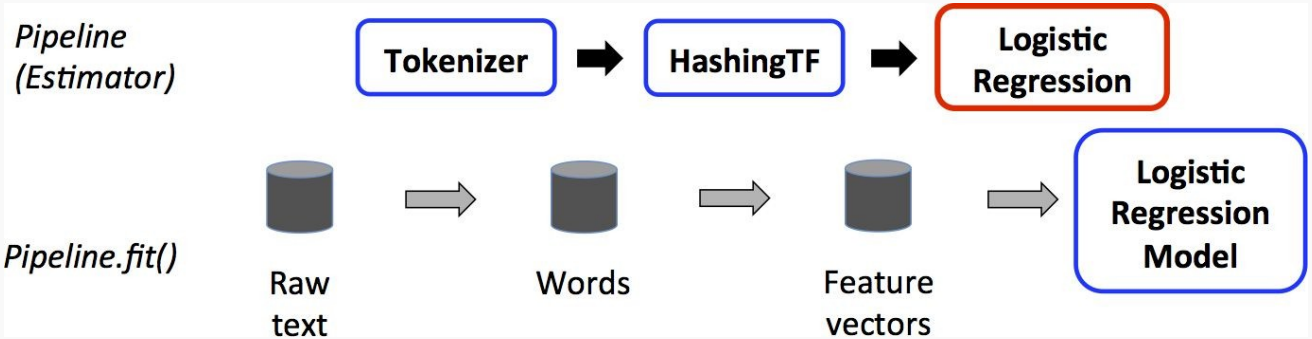
Transformers

- They implement a transform() method
- They convert one DataFrame into another, usually by adding columns
- For example, this is how you get predictions, through using a transform method and adding a column of predictions to your DataFrame
- Examples of transformers: VectorAssembler, Tokenizer, StopWordsRemover, and many more

Estimators

- Any algorithm that fits or trains on data
- They implement a fit() method whose argument is a DataFrame
- The output of fit() is another type called a Model, which is actually a Transformer
- Examples of estimators: LogisticRegression, DecisionTreeRegressor, and many more

Example Pipeline



Blue = Transformers | Red = Estimator | Cylinders = DataFrames

1. Tokenizer.**transform()** - splits the raw text documents into words and **adds a new column** with those words to the DataFrame
2. HashingTF.**transform()** - converts the word column into feature vectors and **adds a new column** with those vectors to the DataFrame
3. LogisticRegression.**fit()** - **trains on the data** and produces a Logistic Regression Model

General rules

galvanize

- Spark ml requires data to be in a very specific format…
- Booleans and ints are not allowed, must be explicitly converted to floats
- Machine learning estimators only take one feature column as input (+ one target column, if you are doing supervised learning). All feature columns must be assembled into a single vector column using the vector assembler. This should be the last step in your pipeline

https://spark.apache.org/docs/latest/api/python/pyspark.ml.html#pyspark.ml.feature.VectorAssembler

Machine Learning Libraries

galvanize

- In the past, there was a trade-off between using the two different machine learning libraries available - Spark MLlib and Spark ML
- In general, spark-ml is newer and is designed to be used with dataframes.
- Mllib is older and designed for use with RDDs. In general you should avoid this, but there are still a few functions that are only available for use in mllib
- The RDD-based API is expected to be removed in Spark 3.0
- You can read more here: https://spark.apache.org/docs/latest/ml-guide.html

Check in Questions

galvanize

- What is a transformer?
- What is an estimator?

Recap: Learning Objectives

galvanize

- Review feature engineering in pyspark using spark SQL
- Review the concept of ml pipelines and introduce the spark implementation
- Describe one difference between Spark & Sklearn.
- Explain the concept of a transformer.
- Explain the concept of an estimator.

Additional Resources

galvanize

- Pyspark machine learning reference: http://spark.apache.org/docs/latest/api/python/pyspark.ml.html
- Machine learning with Spark examples and code reference: https://spark.apache.org/docs/latest/ml-guide.html
- Cross validation and train-test splitting in Spark: https://spark.apache.org/docs/latest/ml-tuning.html