Spark for Machine Learning & AI

with Dan Sullivan

Lynda.com FROM LINKEDIN

DataFrame Commands	Example	Description
spark.read.csv	spark.read.csv("/Users/mydir/exercises/employee.txt",header=True,inferSchema=True)	Loads text file data into a DataFrame
.printSchema()	df.printSchema	Prints a DataFrame schema
.column	df.column	Returns a list of DataFrame column names
.take(n)	df.take(10)	Returns the first n rows of a DataFrame
.count()	df.count	Returns number of rows in a DataFrame
.filter()	emp_df.filter("salary >= 100000")	Returns rows of DataFrames that pass the Boolean filter
.show()	emp_mgrs_df.select("salary").show()	Prints rows of a DataFrame

Commonly Used Modules	Description
pyspark.ml.classification	Contains classification algorithms
pyspark.ml.clustering	Contains clustering algorithms
pyspark.ml.evaluation	Contains evaluation algorithms
pyspark.ml.feature	Contains preprocessing functionality
pyspark.ml.linalg	Contains Vectors module
pyspark.ml.regression	Contains regression algorithms
pyspark.sql.function	Contains SQL commands for manipulating DataFrames

Preprocessing Functions	Description
MinMaxScaler	Normalizes data to the 0 to 1 range
StandardScaler	Standardizes data to the –1 to 1 range with a mean of 0
Bucketizer	Groups continuous data into partitions, such as a histogram
Tokenizer	Splits strings into words
HashingTF	Computes term frequencies
IDF	Computes inverse document frequency
VectorAssembler	Used to create a single feature vector from multiple input columns
StringIndexer	Used to map a categorical variable to a numeric index

ML Algorithms	Description
KMeans	A basic clustering algorithm; a good choice for exploring data sets
Bisecting KMeans	Hiearchical KMeans that is faster than KMeans with large data sets
NaiveBay	Fast classifier that can work well if attributes are independent
MultilayerPerceptronClassifier	Classifier to use if data cannot be separated into categories using a linear description
DecisionTreeClassifier	Good, general purpose classifier
LinearRegression	Basic linear regression algorithm
DecisionTreeRegressor	Another fast regression algorithm that may work well in cases not suited for LinearRegression
GBTRegressor	A gradient-boosted tree regressor can give good results but may take longer to build a model than other regression algorithms

Spark Command Lines	Description
pyspark	Python interface to Spark
spark-shell	Scala interface to Spark
sparkR	R interface to Spark (assumes R is installed)