

Learning Apache Spark with Python, by Dr. Wenqiang Feng
=> pyspark code index
=> purpose -> browse code to increase awareness of Apache Spark’s available classes, methods, and functions

1.1 About

1.1.1 About this note

This is a shared repository for Learning Apache Spark Notes. The PDF version can be downloaded from HERE. The first version was posted on Github in ChenFeng ([Feng2017]). This shared repository mainly contains the self-learning and self-teaching notes from Wenqiang during his IMA Data Science Fellowship. The reader is referred to the repository https://github.com/runawayhorse001/LearningApacheSpark for more details about the dataset and the .ipynb files.

In this repository, I try to use the detailed demo code and examples to show how to use each main functions. If you find your work wasn’t cited in this note, please feel free to let me know.

Although I am by no means an data mining programming and Big Data expert, I decided that it would be useful for me to share what I learned about PySpark programming in the form of easy tutorials with detailed example. I hope those tutorials will be a valuable tool for your studies.

The tutorials assume that the reader has a preliminary knowledge of programming and Linux. And this document is generated automatically by using sphinx.

1.1.2 About the author

- Wenqiang Feng
- Director of Data Science and PhD in Mathematics
- University of Tennessee at Knoxville
- Email: von198@gmail.com
- Biography
Wenqiang Feng is the Director of Data Science at American Express (AMEX). Prior to his time at AMEX, Dr. Feng was a Sr. Data Scientist in Machine Learning Lab, H&R Block. Before joining Block, Dr. Feng was a Data Scientist at Applied Analytics Group, DST (now SS&C). Dr. Feng’s responsibilities include providing clients with access to cutting-edge skills and technologies, including Big Data analytic solutions, advanced analytic and data enhancement techniques and modeling.
Dr. Feng has deep analytic expertise in data mining, analytic systems, machine learning algorithms, business intelligence, and applying Big Data tools to strategically solve industry problems in a crossfunctional business. Before joining DST, Dr. Feng was an IMA Data Science Fellow at The Institute for Mathematics and its Applications (IMA) at the University of Minnesota. While there, he helped startup companies make marketing decisions based on deep predictive analytics.
Dr. Feng graduated from University of Tennessee, Knoxville, with Ph.D. in Computational Mathematics and Master’s degree in Statistics. He also holds Master’s degree in Computational Mathematics from Missouri University of Science and Technology (MST) and Master’s degree in Applied Mathematics from the University of Science and Technology of China (USTC).
- Declaration
The work of Wenqiang Feng was supported by the IMA, while working at IMA. However, any opinion, finding, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the IMA, UTK, DST, HR & Block and AMEX.

1.2 Motivation for this tutorial

- I was motivated by the IMA Data Science Fellowship project to learn PySpark. After that I was impressed and attracted by the PySpark. And I foud that: 1. It is no exaggeration to say that Spark is the most powerful Bigdata tool.
2. However, I still found that learning Spark was a difficult process. I have to Google it and identify which one is true. And it was hard to find detailed examples which I can easily learned the full process in one file.
3. Good sources are expensive for a graduate student.

1.3 Copyright notice and license info

This Learning Apache Spark with Python PDF file is supposed to be a free and living document, which is why its source is available online at https://runawayhorse001.github.io/LearningApacheSpark/pyspark. pdf. But this document is licensed according to both MIT License and Creative Commons AttributionNonCommercial 2.0 Generic (CC BY-NC 2.0) License. When you plan to use, copy, modify, merge, publish, distribute or sublicense, Please see the terms of those licenses for more details and give the corresponding credits to the author.

=====
Note=> the following numbers correspond to the Dr. Feng’s Learning apache spark pdf book.
=====

A

accuracy() (pyspark.ml.classification.LogisticRegressionSummary property), 441
addGrid() (pyspark.ml.tuning.ParamGridBuilder method), 475
AFTSurvivalRegression (class in pyspark.ml.regression), 411
AFTSurvivalRegressionModel (class in pyspark.ml.regression), 412
aic() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420
ALS (class in pyspark.ml.recommendation), 465
ALSModel (class in pyspark.ml.recommendation), 469
areaUnderROC() (pyspark.ml.classification.BinaryLogisticRegressionSummary property), 430
assignClusters() (pyspark.ml.clustering.PowerIterationClustering method), 463
avgMetrics (pyspark.ml.tuning.CrossValidatorModel attribute), 474

B

baseOn() (pyspark.ml.tuning.ParamGridBuilder method), 475
bestModel (pyspark.ml.tuning.CrossValidatorModel attribute), 474
bestModel (pyspark.ml.tuning.TrainValidationSplitModel attribute), 477
BinaryClassificationEvaluator (class in pyspark.ml.evaluation), 477
BinaryLogisticRegressionSummary (class in pyspark.ml.classification), 430
BinaryLogisticRegressionTrainingSummary (class in pyspark.ml.classification), 431
BisectingKMeans (class in pyspark.ml.clustering), 450
BisectingKMeansModel (class in pyspark.ml.clustering), 451
BisectingKMeansSummary (class in pyspark.ml.clustering), 452
boundaries() (pyspark.ml.regression.IsotonicRegressionModel property), 423
build() (pyspark.ml.tuning.ParamGridBuilder method), 475

C

ChiSquareTest (class in pyspark.ml.stat), 405
clusterCenters() (pyspark.ml.clustering.BisectingKMeansModel method), 451
clusterCenters() (pyspark.ml.clustering.KMeansModel method), 457
ClusteringEvaluator (class in pyspark.ml.evaluation), 478
coefficientMatrix() (pyspark.ml.classification.LogisticRegressionModel property), 440
coefficients() (pyspark.ml.classification.LinearSVCModel property), 437
coefficients() (pyspark.ml.classification.LogisticRegressionModel property), 440
coefficients() (pyspark.ml.regression.AFTSurvivalRegressionModel 413
coefficients() (pyspark.ml.regression.GeneralizedLinearRegressionModel property), 419
coefficients() (pyspark.ml.regression.LinearRegressionModel property), 424
coefficientStandardErrors() (pyspark.ml.regression.GeneralizedLinearRegressionTrainingSummary property), 421
coefficientStandardErrors() (pyspark.ml.regression.LinearRegressionSummary property), 425
computeCost() (pyspark.ml.clustering.BisectingKMeansModel method), 452
computeCost() (pyspark.ml.clustering.KMeansModel method), 457
Configure Spark on Mac and Ubuntu, 17
copy() (pyspark.ml.classification.OneVsRest method), 447
copy() (pyspark.ml.classification.OneVsRestModel method), 448
copy() (pyspark.ml.pipeline.Pipeline method), 470
copy() (pyspark.ml.pipeline.PipelineModel method), 471
copy() (pyspark.ml.tuning.CrossValidator method), 473
copy() (pyspark.ml.tuning.CrossValidatorModel method), 474
copy() (pyspark.ml.tuning.TrainValidationSplit method), 476
copy() (pyspark.ml.tuning.TrainValidationSplitModel method), 477
corr() (pyspark.ml.stat.Correlation static method), 406
Correlation (class in pyspark.ml.stat), 406
count() (pyspark.ml.stat.Summarizer static method), 409
CrossValidator (class in pyspark.ml.tuning), 472
CrossValidatorModel (class in pyspark.ml.tuning), 473

D

DecisionTreeClassificationModel (class in pyspark.ml.classification), 431  
DecisionTreeClassifier (class in pyspark.ml.classification), 432  
DecisionTreeRegressionModel (class in pyspark.ml.regression), 413  
DecisionTreeRegressor (class in pyspark.ml.regression), 413  
degreesOfFreedom() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
degreesOfFreedom() (pyspark.ml.regression.LinearRegressionSummary property), 425  
describeTopics() (pyspark.ml.clustering.LDAModel method), 462  
deviance() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
devianceResiduals() (pyspark.ml.regression.LinearRegressionSummary property), 425  
dispersion() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
DistributedLDAModel (class in pyspark.ml.clustering), 452

E

estimatedDocConcentration() (pyspark.ml.clustering.LDAModel method), 462  
evaluate() (pyspark.ml.classification.LogisticRegressionModel method), 440  
evaluate() (pyspark.ml.evaluation.Evaluator method), 479  
evaluate() (pyspark.ml.regression.GeneralizedLinearRegressionModel method), 419  
evaluate() (pyspark.ml.regression.LinearRegressionModel method), 424  
evaluateEachIteration() (pyspark.ml.classification.GBTClassificationModel method), 433  
evaluateEachIteration() (pyspark.ml.regression.GBTRegressionModel method), 414  
Evaluator (class in pyspark.ml.evaluation), 479  
explainedVariance() (pyspark.ml.regression.LinearRegressionSummary property), 425

F

falsePositiveRateByLabel() (pyspark.ml.classification.LogisticRegressionSummary property), 441  
featureImportances() (pyspark.ml.classification.DecisionTreeClassificationModel property), 432  
featureImportances() (pyspark.ml.classification.GBTClassificationModel property), 433  
featureImportances() (pyspark.ml.classification.RandomForestClassificationModel property), 448  
featureImportances() (pyspark.ml.regression.DecisionTreeRegressionModel property), 413  
featureImportances() (pyspark.ml.regression.GBTRegressionModel property), 415  
featureImportances() (pyspark.ml.regression.RandomForestRegressionModel property), 428  
featuresCol() (pyspark.ml.classification.LogisticRegressionSummary property), 441  
featuresCol() (pyspark.ml.regression.LinearRegressionSummary property), 426  
fMeasureByLabel() (pyspark.ml.classification.LogisticRegressionSummary method), 441  
fMeasureByThreshold() (pyspark.ml.classification.BinaryLogisticRegressionSummary property), 430

G

G GaussianMixture (class in pyspark.ml.clustering), 453  
GaussianMixtureModel (class in pyspark.ml.clustering), 455  
GaussianMixtureSummary (class in pyspark.ml.clustering), 455  
gaussiansDF() (pyspark.ml.clustering.GaussianMixtureModel property), 455  
GBTClassificationModel (class in pyspark.ml.classification), 433  
GBTClassifier (class in pyspark.ml.classification), 434  
GBTRegressionModel (class in pyspark.ml.regression), 414  
GBTRegressor (class in pyspark.ml.regression), 415  
GeneralizedLinearRegression (class in pyspark.ml.regression), 416  
GeneralizedLinearRegressionModel (class in pyspark.ml.regression), 419  
GeneralizedLinearRegressionSummary (class in pyspark.ml.regression), 419  
GeneralizedLinearRegressionTrainingSummary (class in pyspark.ml.regression), 421  
getAlpha() (pyspark.ml.recommendation.ALS method), 466  
getBlockSize() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444  
getCensorCol() (pyspark.ml.regression.AFTSurvivalRegression method), 412  
getCheckpointFiles() (pyspark.ml.clustering.DistributedLDAModel method), 452  
getColdStartStrategy() (pyspark.ml.recommendation.ALS method), 466  
getDistanceMeasure() (pyspark.ml.clustering.BisectingKMeans method), 451  
getDistanceMeasure() (pyspark.ml.clustering.KMeans method), 456  
getDistanceMeasure() (pyspark.ml.evaluation.ClusteringEvaluator method), 479  
getDocConcentration() (pyspark.ml.clustering.LDA method), 459  
getDstCol() (pyspark.ml.clustering.PowerIterationClustering method), 464  
getEpsilon() (pyspark.ml.regression.LinearRegression method), 424  
getFamily() (pyspark.ml.classification.LogisticRegression method), 438  
getFamily() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
getFeatureIndex() (pyspark.ml.regression.IsotonicRegression method), 422  
getFinalStorageLevel() (pyspark.ml.recommendation.ALS method), 466  
getImplicitPrefs() (pyspark.ml.recommendation.ALS method), 467  
getInitialWeights() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444  
getInitMode() (pyspark.ml.clustering.KMeans method), 456  
getInitMode() (pyspark.ml.clustering.PowerIterationClustering method), 464  
getInitSteps() (pyspark.ml.clustering.KMeans method), 457  
getIntermediateStorageLevel() (pyspark.ml.recommendation.ALS method), 467  
getIsotonic() (pyspark.ml.regression.IsotonicRegression method), 422  
getItemCol() (pyspark.ml.recommendation.ALS method), 467  
getK() (pyspark.ml.clustering.BisectingKMeans method), 451  
getK() (pyspark.ml.clustering.GaussianMixture method), 454  
getK() (pyspark.ml.clustering.KMeans method), 457  
getK() (pyspark.ml.clustering.LDA method), 459  
getK() (pyspark.ml.clustering.PowerIterationClustering method), 464  
getKeepLastCheckpoint() (pyspark.ml.clustering.LDA method), 459  
getLayers() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444  
getLearningDecay() (pyspark.ml.clustering.LDA method), 459  
getLearningOffset() (pyspark.ml.clustering.LDA method), 459  
getLink() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
getLinkPower() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
getLinkPredictionCol() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
getLossType() (pyspark.ml.classification.GBTClassifier method), 435  
getLossType() (pyspark.ml.regression.GBTRegressor method), 416  
getLowerBoundsOnCoefficients() (pyspark.ml.classification.LogisticRegression method), 438  
getLowerBoundsOnIntercepts() (pyspark.ml.classification.LogisticRegression method), 438  
getMetricName() (pyspark.ml.evaluation.BinaryClassificationEvaluator method), 478  
getMetricName() (pyspark.ml.evaluation.ClusteringEvaluator method), 479  
getMetricName() (pyspark.ml.evaluation.MulticlassClassificationEvaluator method), 480  
getMetricName() (pyspark.ml.evaluation.RegressionEvaluator method), 481  
getMinDivisibleClusterSize() (pyspark.ml.clustering.BisectingKMeans method), 451  
getModelType() (pyspark.ml.classification.NaiveBayes method), 446  
getNonnegative() (pyspark.ml.recommendation.ALS method), 467  
getNumFolds() (pyspark.ml.tuning.CrossValidator method), 473  
getNumItemBlocks() (pyspark.ml.recommendation.ALS method), 467  
getNumUserBlocks() (pyspark.ml.recommendation.ALS method), 467  
getOffsetCol() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
getOptimizeDocConcentration() (pyspark.ml.clustering.LDA method), 459  
getOptimizer() (pyspark.ml.clustering.LDA method), 459  
getQuantileProbabilities() (pyspark.ml.regression.AFTSurvivalRegression method), 412  
getQuantilesCol() (pyspark.ml.regression.AFTSurvivalRegression method), 412  
getRank() (pyspark.ml.recommendation.ALS method), 467  
getRatingCol() (pyspark.ml.recommendation.ALS method), 467  
getSmoothing() (pyspark.ml.classification.NaiveBayes method), 446  
getSrcCol() (pyspark.ml.clustering.PowerIterationClustering method), 464  
getStagePath() (pyspark.ml.pipeline.PipelineSharedReadWrite static method), 472



G

<continue>  
getStages() (pyspark.ml.pipeline.Pipeline method), 470  
getStepSize() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444  
getSubsamplingRate() (pyspark.ml.clustering.LDA method), 459  
getThreshold() (pyspark.ml.classification.LogisticRegression method), 438  
getThresholds() (pyspark.ml.classification.LogisticRegression method), 439  
getTopicConcentration() (pyspark.ml.clustering.LDA method), 459  
getTopicDistributionCol() (pyspark.ml.clustering.LDA method), 460  
getTrainRatio() (pyspark.ml.tuning.TrainValidationSplit method), 476  
getUpperBoundsOnCoefficients() (pyspark.ml.classification.LogisticRegression method), 439  
getUpperBoundsOnIntercepts() (pyspark.ml.classification.LogisticRegression method), 439  
getUserCol() (pyspark.ml.recommendation.ALS method), 467  
getVariancePower() (pyspark.ml.regression.GeneralizedLinearRegression method), 418

H

hasSummary() (pyspark.ml.classification.LogisticRegressionModel property), 440  
hasSummary() (pyspark.ml.clustering.BisectingKMeansModel property), 452  
hasSummary() (pyspark.ml.clustering.GaussianMixtureModel property), 455  
hasSummary() (pyspark.ml.clustering.KMeansModel property), 457  
hasSummary() (pyspark.ml.regression.GeneralizedLinearRegressionModel property), 419  
hasSummary() (pyspark.ml.regression.LinearRegressionModel property), 425

I

intercept() (pyspark.ml.classification.LinearSVCModel property), 437  
Index 493  
intercept() (pyspark.ml.classification.LogisticRegressionModel property), 440  
intercept() (pyspark.ml.regression.AFTSurvivalRegressionModel property), 413  
intercept() (pyspark.ml.regression.GeneralizedLinearRegressionModel property), 419  
intercept() (pyspark.ml.regression.LinearRegressionModel property), 425  
interceptVector() (pyspark.ml.classification.LogisticRegressionModel property), 440  
isDistributed() (pyspark.ml.clustering.LDAModel method), 462  
isLargerBetter() (pyspark.ml.evaluation.Evaluator method), 479  
IsotonicRegression (class in pyspark.ml.regression), 422  
IsotonicRegressionModel (class in pyspark.ml.regression), 422  
itemFactors() (pyspark.ml.recommendation.ALSModel property), 469

K

KMeans (class in pyspark.ml.clustering), 455  
KMeansModel (class in pyspark.ml.clustering), 457  
KolmogorovSmirnovTest (class in pyspark.ml.stat), 407

L

labelCol() (pyspark.ml.classification.LogisticRegressionSummary property), 441  
labelCol() (pyspark.ml.regression.LinearRegressionSummary property), 426  
labels() (pyspark.ml.classification.LogisticRegressionSummary property), 441  
layers() (pyspark.ml.classification.MultilayerPerceptronClassificationModel property), 443  
LDA (class in pyspark.ml.clustering), 458  
LDAModel (class in pyspark.ml.clustering), 461  
LinearRegression (class in pyspark.ml.regression), 423  
LinearRegressionModel (class in pyspark.ml.regression), 424  
LinearRegressionSummary (class in pyspark.ml.regression), 425  
LinearRegressionTrainingSummary (class in pyspark.ml.regression), 428  
LinearSVC (class in pyspark.ml.classification), 435  
LinearSVCModel (class in pyspark.ml.classification), 437  
load() (pyspark.ml.pipeline.PipelineModelReader method), 471  
load() (pyspark.ml.pipeline.PipelineReader method), 471  
load() (pyspark.ml.pipeline.PipelineSharedReadWrite static method), 472  
LocalLDAModel (class in pyspark.ml.clustering), 462  
LogisticRegression (class in pyspark.ml.classification), 437  
LogisticRegressionModel (class in pyspark.ml.classification), 440  
LogisticRegressionSummary (class in pyspark.ml.classification), 441  
LogisticRegressionTrainingSummary (class in pyspark.ml.classification), 442  
logLikelihood() (pyspark.ml.clustering.GaussianMixtureSummary property), 455  
logLikelihood() (pyspark.ml.clustering.LDAModel method), 462  
logPerplexity() (pyspark.ml.clustering.LDAModel method), 462  
logPrior() (pyspark.ml.clustering.DistributedLDAModel method), 452

M

max() (pyspark.ml.stat.Summarizer static method), 409  
mean() (pyspark.ml.stat.Summarizer static method), 409  
meanAbsoluteError() (pyspark.ml.regression.LinearRegressionSummary property), 426  
meanSquaredError() (pyspark.ml.regression.LinearRegressionSummary property), 426  
metrics() (pyspark.ml.stat.Summarizer static method), 409  
min() (pyspark.ml.stat.Summarizer static method), 410  
module pyspark.ml.classification, 430  
pyspark.ml.clustering, 450  
pyspark.ml.evaluation, 477  
pyspark.ml.pipeline, 470  
pyspark.ml.recommendation, 465  
pyspark.ml.regression, 411  
pyspark.ml.stat, 405  
pyspark.ml.tuning, 472  
MulticlassClassificationEvaluator (class in pyspark.ml.evaluation), 480  
MultilayerPerceptronClassificationModel (class in pyspark.ml.classification), 443  
MultilayerPerceptronClassifier (class in pyspark.ml.classification), 443

N

NaiveBayes (class in pyspark.ml.classification), 445  
NaiveBayesModel (class in pyspark.ml.classification), 446  
normL1() (pyspark.ml.stat.Summarizer static method), 410  
normL2() (pyspark.ml.stat.Summarizer static method), 410  
nullDeviance() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
numInstances() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
numInstances() (pyspark.ml.regression.LinearRegressionSummary property), 426  
numIterations() (pyspark.ml.regression.GeneralizedLinearRegressionTrainingSummary property), 421  
numNonZeros() (pyspark.ml.stat.Summarizer static method), 410

O

objectiveHistory() (pyspark.ml.classification.LogisticRegressionTrainingSummary property), 442  
objectiveHistory() (pyspark.ml.regression.LinearRegressionTrainingSummary property), 428  
OneVsRest (class in pyspark.ml.classification), 446  
OneVsRestModel (class in pyspark.ml.classification), 448

P

ParamGridBuilder (class in pyspark.ml.tuning), 474  
pi() (pyspark.ml.classification.NaiveBayesModel property), 446  
Pipeline (class in pyspark.ml.pipeline), 470  
PipelineModel (class in pyspark.ml.pipeline), 471  
PipelineModelReader (class in pyspark.ml.pipeline), 471

P

<continue>

PipelineModelWriter (class in pyspark.ml.pipeline), 471  
PipelineReader (class in pyspark.ml.pipeline), 471  
PipelineSharedReadWrite (class in pyspark.ml.pipeline), 472  
PipelineWriter (class in pyspark.ml.pipeline), 472  
PowerIterationClustering (class in pyspark.ml.clustering), 463  
pr() (pyspark.ml.classification.BinaryLogisticRegressionSummary property), 430  
precisionByLabel() (pyspark.ml.classification.LogisticRegressionSummary property), 441  
precisionByThreshold() (pyspark.ml.classification.BinaryLogisticRegressionSummary property), 431  
predict() (pyspark.ml.regression.AFTSurvivalRegressionModel method), 413  
predictionCol() (pyspark.ml.classification.LogisticRegressionSummary property),441  
predictionCol() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
predictionCol() (pyspark.ml.regression.LinearRegressionSummary property), 427  
predictions() (pyspark.ml.classification.LogisticRegressionSummary property), 441  
predictions() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
predictions() (pyspark.ml.regression.IsotonicRegressionModel property), 423  
predictions() (pyspark.ml.regression.LinearRegressionSummary property), 427  
predictQuantiles() (pyspark.ml.regression.AFTSurvivalRegressionModel method), 413  
probability() (pyspark.ml.clustering.GaussianMixtureSummary property), 455  
probabilityCol() (pyspark.ml.classification.LogisticRegressionSummary property), 442  
probabilityCol() (pyspark.ml.clustering.GaussianMixtureSummary property), 455  
pValues() (pyspark.ml.regression.GeneralizedLinearRegressionTrainingSummary property), 421  
pValues() (pyspark.ml.regression.LinearRegressionSummary property), 426  
pyspark.ml.classification module, 430  
pyspark.ml.clustering module, 450  
pyspark.ml.evaluation module, 477  
yspark.ml.pipeline module, 470  
pyspark.ml.recommendation module, 465  
pyspark.ml.regression module, 411  
pyspark.ml.stat module, 405  
pyspark.ml.tuning module, 472

R

r2() (pyspark.ml.regression.LinearRegressionSummary property), 427  
r2adj() (pyspark.ml.regression.LinearRegressionSummary property), 427  
RandomForestClassificationModel (class in pyspark.ml.classification), 448  
RandomForestClassifier (class in pyspark.ml.classification), 448  
RandomForestRegressionModel (class in pyspark.ml.regression), 428  
RandomForestRegressor (class in pyspark.ml.regression), 429  
rank() (pyspark.ml.recommendation.ALSModel property), 469  
rank() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420  
read() (pyspark.ml.pipeline.Pipeline class method), 470  
read() (pyspark.ml.pipeline.PipelineModel class method), 471 read() (pyspark.ml.tuning.CrossValidator class method), 473 read() (pyspark.ml.tuning.CrossValidatorModel class method), 474 read() (pyspark.ml.tuning.TrainValidationSplit class method), 476 read() (pyspark.ml.tuning.TrainValidationSplitModel class method), 477 recallByLabel() (pyspark.ml.classification.LogisticRegressionSummary property), 442 recallByThreshold() (pyspark.ml.classification.BinaryLogisticRegressionSummary property), 431 recommendForAllItems() (pyspark.ml.recommendation.ALSModel method), 469 recommendForAllUsers() (pyspark.ml.recommendation.ALSModel method), 469 recommendForItemSubset() (pyspark.ml.recommendation.ALSModel method), 469 recommendForUserSubset() (pyspark.ml.recommendation.ALSModel method), 469 RegressionEvaluator (class in pyspark.ml.evaluation), 480 residualDegreeOfFreedom() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 420 residualDegreeOfFreedomNull() (pyspark.ml.regression.GeneralizedLinearRegressionSummary property), 421 residuals() (pyspark.ml.regression.GeneralizedLinearRegressionSummary method), 421 residuals() (pyspark.ml.regression.LinearRegressionSummary property), 427 roc() (pyspark.ml.classification.BinaryLogisticRegressionSummary property), 431 rootMeanSquaredError() (pyspark.ml.regression.LinearRegressionSummary property), 427 Run on Databricks Community Cloud, 11

S

saveImpl() (pyspark.ml.pipeline.PipelineModelWriter method), 471  
saveImpl() (pyspark.ml.pipeline.PipelineSharedReadWrite static method), 472  
saveImpl() (pyspark.ml.pipeline.PipelineWriter method), 472  
scale() (pyspark.ml.regression.AFTSurvivalRegressionModel property), 413  
scale() (pyspark.ml.regression.LinearRegressionModel property), 425  
Set up Spark on Cloud, 29 setAlpha() (pyspark.ml.recommendation.ALS method), 467  
setBlockSize() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444  
setCensorCol() (pyspark.ml.regression.AFTSurvivalRegression method), 412  
setColdStartStrategy() (pyspark.ml.recommendation.ALS method), 467  
setDistanceMeasure() (pyspark.ml.clustering.BisectingKMeans method), 451  
setDistanceMeasure() (pyspark.ml.clustering.KMeans method), 457  
setDistanceMeasure() (pyspark.ml.evaluation.ClusteringEvaluator method), 479  
setDocConcentration() (pyspark.ml.clustering.LDA method), 460  
setDstCol() (pyspark.ml.clustering.PowerIterationClustering method), 464  
setEpsilon() (pyspark.ml.regression.LinearRegression method), 424  
setFamily() (pyspark.ml.classification.LogisticRegression method), 439  
setFamily() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
setFeatureIndex() (pyspark.ml.regression.IsotonicRegression method), 422  
setFeatureSubsetStrategy() (pyspark.ml.classification.GBTClassifier method), 435  
setFeatureSubsetStrategy() (pyspark.ml.classification.RandomForestClassifier method), 449  
setFeatureSubsetStrategy() (pyspark.ml.regression.GBTRegressor method), 416  
setFeatureSubsetStrategy() (pyspark.ml.regression.RandomForestRegressor method), 430  
setFinalStorageLevel() (pyspark.ml.recommendation.ALS method), 467  
setImplicitPrefs() (pyspark.ml.recommendation.ALS method), 468  
setInitialWeights() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444  
setInitMode() (pyspark.ml.clustering.KMeans method), 457  
setInitMode() (pyspark.ml.clustering.PowerIterationClustering method), 464  
setInitSteps() (pyspark.ml.clustering.KMeans method), 457  
setIntermediateStorageLevel() (pyspark.ml.recommendation.ALS method), 468  
setIsotonic() (pyspark.ml.regression.IsotonicRegression method), 422  
setItemCol() (pyspark.ml.recommendation.ALS method), 468  
setK() (pyspark.ml.clustering.BisectingKMeans method), 451  
setK() (pyspark.ml.clustering.GaussianMixture method), 454  
setK() (pyspark.ml.clustering.KMeans method), 457  
setK() (pyspark.ml.clustering.LDA method), 460  
setK() (pyspark.ml.clustering.PowerIterationClustering method), 464  
setKeepLastCheckpoint() (pyspark.ml.clustering.LDA method), 460  
setLayers() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444  
setLearningDecay() (pyspark.ml.clustering.LDA method), 460  
setLearningOffset() (pyspark.ml.clustering.LDA method), 460  
setLink() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
setLinkPower() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
setLinkPredictionCol() (pyspark.ml.regression.GeneralizedLinearRegression method), 418  
setLossType() (pyspark.ml.classification.GBTClassifier method), 435  
setLossType() (pyspark.ml.regression.GBTRegressor method), 416  
setLowerBoundsOnCoefficients() (pyspark.ml.classification.LogisticRegression method), 439  
setLowerBoundsOnIntercepts() (pyspark.ml.classification.LogisticRegression method), 439  
setMetricName() (pyspark.ml.evaluation.BinaryClassificationEvaluator method), 478  
setMetricName() (pyspark.ml.evaluation.ClusteringEvaluator method), 479



S

<continue>

setMetricName() (pyspark.ml.evaluation.MulticlassClassificationEvaluator method), 480

setMetricName() (pyspark.ml.evaluation.RegressionEvaluator method), 481

setMinDivisibleClusterSize() (pyspark.ml.clustering.BisectingKMeans method), 451

setModelType() (pyspark.ml.classification.NaiveBayes method), 446

setNonnegative() (pyspark.ml.recommendation.ALS method), 468

setNumBlocks() (pyspark.ml.recommendation.ALS method), 468

setNumFolds() (pyspark.ml.tuning.CrossValidator method), 473

setNumItemBlocks() (pyspark.ml.recommendation.ALS method), 468

setNumUserBlocks() (pyspark.ml.recommendation.ALS method), 468

setOffsetCol() (pyspark.ml.regression.GeneralizedLinearRegression method),418

setOptimizeDocConcentration() (pyspark.ml.clustering.LDA method), 460

setOptimizer() (pyspark.ml.clustering.LDA method), 461

setParams() (pyspark.ml.classification.DecisionTreeClassifier method), 433

setParams() (pyspark.ml.classification.GBTClassifier method), 435

setParams() (pyspark.ml.classification.LinearSVC method), 436

setParams() (pyspark.ml.classification.LogisticRegression method), 439

setParams() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 444

setParams() (pyspark.ml.classification.NaiveBayes method), 446

setParams() (pyspark.ml.classification.OneVsRest method), 448

setParams() (pyspark.ml.classification.RandomForestClassifier method), 449

setParams() (pyspark.ml.clustering.BisectingKMeans method), 451

setParams() (pyspark.ml.clustering.GaussianMixture method), 454

setParams() (pyspark.ml.clustering.KMeans method), 457

setParams() (pyspark.ml.clustering.LDA method), 461

setParams() (pyspark.ml.clustering.PowerIterationClustering method), 464

setParams() (pyspark.ml.evaluation.BinaryClassificationEvaluator method), 478

setParams() (pyspark.ml.evaluation.ClusteringEvaluator method), 479

setParams() (pyspark.ml.evaluation.MulticlassClassificationEvaluator method), 480

setParams() (pyspark.ml.evaluation.RegressionEvaluator method), 481

setParams() (pyspark.ml.pipeline.Pipeline method), 470

setParams() (pyspark.ml.recommendation.ALS method), 468

setParams() (pyspark.ml.regression.AFTSurvivalRegression method), 412

setParams() (pyspark.ml.regression.DecisionTreeRegressor method), 414

setParams() (pyspark.ml.regression.GBTRegressor method), 416

setParams() (pyspark.ml.regression.GeneralizedLinearRegression method), 419

setParams() (pyspark.ml.regression.IsotonicRegression method), 422

setParams() (pyspark.ml.regression.LinearRegression method), 424

setParams() (pyspark.ml.regression.RandomForestRegressor method), 430

setParams() (pyspark.ml.tuning.CrossValidator method), 473

setParams() (pyspark.ml.tuning.TrainValidationSplit method), 476

setQuantileProbabilities() (pyspark.ml.regression.AFTSurvivalRegression method), 412

setQuantilesCol() (pyspark.ml.regression.AFTSurvivalRegression method), 412

setRank() (pyspark.ml.recommendation.ALS method), 468

setRatingCol() (pyspark.ml.recommendation.ALS method), 468

setSmoothing() (pyspark.ml.classification.NaiveBayes method), 446

setSrcCol() (pyspark.ml.clustering.PowerIterationClustering method), 464

setStages() (pyspark.ml.pipeline.Pipeline method), 471

setStepSize() (pyspark.ml.classification.MultilayerPerceptronClassifier method), 445

setSubsamplingRate() (pyspark.ml.clustering.LDA method), 461

setThreshold() (pyspark.ml.classification.LogisticRegression method), 439

setThresholds() (pyspark.ml.classification.LogisticRegression method), 439

setTopicConcentration() (pyspark.ml.clustering.LDA method), 461

setTopicDistributionCol() (pyspark.ml.clustering.LDA method), 461

setTrainRatio() (pyspark.ml.tuning.TrainValidationSplit method), 476

setUpperBoundsOnCoefficients() (pyspark.ml.classification.LogisticRegression method), 440

setUpperBoundsOnIntercepts() (pyspark.ml.classification.LogisticRegression method), 440

setUserCol() (pyspark.ml.recommendation.ALS method), 468 setVariancePower()  
(pyspark.ml.regression.GeneralizedLinearRegression method), 419

solver() (pyspark.ml.regression.GeneralizedLinearRegressionTrainingSummary property), 421

subModels (pyspark.ml.tuning.CrossValidatorModel attribute), 474

subModels (pyspark.ml.tuning.TrainValidationSplitModel attribute), 477

Summarizer (class in pyspark.ml.stat), 408

summary() (pyspark.ml.classification.LogisticRegressionModel property), 440

summary() (pyspark.ml.clustering.BisectingKMeansModel property), 452

summary() (pyspark.ml.clustering.GaussianMixtureModel property), 455

summary() (pyspark.ml.clustering.KMeansModel property), 458

summary() (pyspark.ml.regression.GeneralizedLinearRegressionModel property), 419

summary() (pyspark.ml.regression.LinearRegressionModel property), 425

summary() (pyspark.ml.stat.SummaryBuilder method), 410

SummaryBuilder (class in pyspark.ml.stat), 410

T

test() (pyspark.ml.stat.ChiSquareTest static method), 405

test() (pyspark.ml.stat.KolmogorovSmirnovTest static method), 407

theta() (pyspark.ml.classification.NaiveBayesModel property), 446

toLocal() (pyspark.ml.clustering.DistributedLDAModel method), 453

topicsMatrix() (pyspark.ml.clustering.LDAModel method), 462

totalIterations() (pyspark.ml.classification.LogisticRegressionTrainingSummary property), 443

totalIterations() (pyspark.ml.regression.LinearRegressionTrainingSummary property), 428

trainingLogLikelihood() (pyspark.ml.clustering.DistributedLDAModel method), 453

TrainValidationSplit (class in pyspark.ml.tuning), 475

TrainValidationSplitModel (class in pyspark.ml.tuning), 476

trees() (pyspark.ml.classification.GBTClassificationModel property), 434

trees() (pyspark.ml.classification.RandomForestClassificationModel property), 448

trees() (pyspark.ml.regression.GBTRegressionModel property), 415

trees() (pyspark.ml.regression.RandomForestRegressionModel property), 429

truePositiveRateByLabel() (pyspark.ml.classification.LogisticRegressionSummary property), 442

tValues() (pyspark.ml.regression.GeneralizedLinearRegressionTrainingSummary property), 421

tValues() (pyspark.ml.regression.LinearRegressionSummary property), 428

U

userFactors() (pyspark.ml.recommendation.ALSModel property), 470

V

validateStages() (pyspark.ml.pipeline.PipelineSharedReadWrite static method), 472

validationMetrics (pyspark.ml.tuning.TrainValidationSplitModel attribute), 477

variance() (pyspark.ml.stat.Summarizer static method), 410

vocabSize() (pyspark.ml.clustering.LDAModel method), 462

W

weightedFalsePositiveRate() (pyspark.ml.classification.LogisticRegressionSummary property), 442

weightedFMeasure() (pyspark.ml.classification.LogisticRegressionSummary method), 442

weightedPrecision() (pyspark.ml.classification.LogisticRegressionSummary property), 442

weightedRecall() (pyspark.ml.classification.LogisticRegressionSummary property), 442

weightedTruePositiveRate() (pyspark.ml.classification.LogisticRegressionSummary property), 442

weights() (pyspark.ml.classification.MultilayerPerceptronClassificationModel property), 443

weights() (pyspark.ml.clustering.GaussianMixtureModel property), 455 write()  
(pyspark.ml.pipeline.Pipeline method), 471

write() (pyspark.ml.pipeline.PipelineModel method), 471

write() (pyspark.ml.tuning.CrossValidator method), 473

write() (pyspark.ml.tuning.CrossValidatorModel method), 474

write() (pyspark.ml.tuning.TrainValidationSplit method), 476

write() (pyspark.ml.tuning.TrainValidationSplitModel method), 477