"""

Random Forest Classification Example.

"""

from \_\_future\_\_ import print\_function

import sys

from pyspark import SparkContext

from pyspark.mllib.tree import RandomForest, RandomForestModel

from pyspark.mllib.util import MLUtils

if \_\_name\_\_ == "\_\_main\_\_":

sc = SparkContext(appName="PythonRandomForestClassificationExample")

# Load and parse the data file into an RDD of LabeledPoint.

trainingData = MLUtils.loadLibSVMFile(sc, 'data/mllib/train\_RF\_reg\_100\_changed.txt')

testData = MLUtils.loadLibSVMFile(sc, 'data/mllib/test\_gre\_only.txt')

model = RandomForest.trainClassifier(trainingData, numClasses=10, categoricalFeaturesInfo={},

numTrees=4, featureSubsetStrategy="auto",

impurity='entropy', maxDepth=5, maxBins=32)

# Evaluate model on test instances and compute test error

predictions = model.predict(testData.map(lambda x: x.features))

pred = predictions.collect()

print('Prediction: '+ str(pred))

labelsAndPredictions = testData.map(lambda lp: lp.label).zip(predictions)

testErr = labelsAndPredictions.filter(lambda (v, p): v != p).count() / float(testData.count())

print('Test Error = ' + str(testErr))