記事
Minoru Horita · 2020年11月30日 6m read

ML モデルを InterSystems IRIS に読み込む

みなさん、こんにちは。 今回は ML モデルを IRIS Manager にアップロードしてテストしようと思います。

注意: Ubuntu 18.04、Apache Zeppelin 0.8.0、Python 3.6.5 で以下を実行しました。

はじめに

最近では実にさまざまなデータマイニングツールを使用して予測モデルを開発し、これまでにないほど簡単にデータを分析できるようになっています。 InterSystems IRIS Data Platform はビッグデータおよび高速データアプリケーション向けに安定した基盤を提供し、最新のデータマイニングツールとの相互運用性を実現します。

この連載記事では、InterSystems IRIS
で利用できるデータマイニング機能について説明
します。最初の記事ではインフラストラクチャを構成し、作業を開始する準備をしました。2
番目の記事では、Apache Spark と Apache Zeppelin
を使用して花の種を予測する最初の予測モデルを構築しました。 この記事では KMeans PMML
モデルを構築し、InterSystems IRIS でテストします。

Intersystems IRIS は PMML の実行機能を提供しています。
そのため、モデルをアップロードし、SQLクエリを使用して任意のデータに対してそのモデルをテストできます。
正解率、適合率、F スコアなどが表示されます。

要件の確認

まず、jpmml をダウンロードし（表を確認して適切なバージョンを選択してください）、それを任意のディレクトリに移動します。 Scala を使用しているのであれば、それで十分でしょう。

> PySpark2PMML must be paired with JPMML-SparkML based on the following compatibility matrix:

Apache Spark version	JPMML-SparkML development branch	JPMML-SparkML uber-JAR file
2.0.X	`1.1.X`	1.1.20
2.1.X	`1.2.X`	1.2.12
2.2.X	`1.3.X`	1.3.8
2.3.X	`master`	1.4.5

Python を使用している場合は、ターミナルで次のコマンドを実行してください。

```
pip3 install --user --upgrade git+https://github.com/jpmml/pyspark2pmml.git
```

正常にインストールされたことを確認したら **Spark Dependencies** に異動し、ダウンロードした jpmml
に次のように依存関係を追加してください。

KMeans モデルの作成

PMML
ビルダーはパイプラインを使用しますので、ここでは以前の記事で書いたコードに若干の変更を加えました。
次のコードを Zeppelin で実行します。

```
%pyspark
  from pyspark.ml.linalg import Vectors
  from pyspark.ml.feature import VectorAssembler
  from pyspark.ml.clustering import KMeans
  from pyspark.ml import Pipeline
  from pyspark.ml.feature import RFormula
  from pyspark2pmml import PMMLBuilder

dataFrame=spark.read.format("com.intersystems.spark"). \
  option("url", "IRIS://localhost:51773/NEWSAMPLE").option("user", "dev"). \
  option("password", "123"). \
  option("dbtable", "DataMining.IrisDataset").load() # iris データセットをロード

(trainingData, testData) = dataFrame.randomSplit([0.7, 0.3]) # データを 2 セットに分割
  assembler = VectorAssembler(inputCols = ["PetalLength", "PetalWidth", "SepalLength", "SepalWidth"],
outputCol="features") # features を含む新しいカラムを追加

kmeans = KMeans().setK(3).setSeed(2000) # 使用するクラスタリングアルゴリズム

pipeline = Pipeline(stages=[assembler, kmeans]) # 渡されたデータはまず assembler
に対して実行され、その後に kmeans に対して実行されます。
  modelKMeans = pipeline.fit(trainingData) # トレーニングデータを渡す

pmmlBuilder = PMMLBuilder(sc, dataFrame, modelKMeans)
  pmmlBuilder.buildFile("KMeans.pmml") # pmml モデルを作成
```

上記により、PetalLength / PetalWidth / SepalLength / SepalWidth を特徴として使用して Species を予測するモデルが作成されます。 このモデルは PMML フォーマットを使用します。

PMML は XML ベースの予測モデル交換フォーマットであり、分析アプリケーションがデータマイニングおよび機械学習アルゴリズムによって生成された予測モデルを記述し、交換する方法を提供します。
これにより、モデルの構築とモデルの実行を切り離すことができます。

出力には PMML モデルへのパスが表示されます。

PMML モデルのアップロードとテスト

IRIS Manager から [**Menu**] -> [**Manage Web Applications**] を開き、目的の**ネームスペース**をクリックしてから [**Analytics**] を有効にしてから [**Save**] をクリックします。

次に、[**Analytics**] -> [**Tools**] -> [**PMML Model Tester**] に移動します。

次の画像のように表示されるはずです。

[**New**] をクリックしてクラス名を書き、PMML
ファイル（パスは出力に表示されていました）をアップロードし、[**Import**] をクリックします。その後、次の SQL クエリを [**Custom data source**] に貼り付けます。

```
SELECT PetalLength, PetalWidth, SepalLength, SepalWidth, Species,
  CASE Species
  WHEN 'Iris-setosa' THEN 0
  WHEN 'Iris-versicolor' THEN 2
  ELSE 1
  END
  As prediction
  FROM DataMining.IrisDataset
```

KMeans クラスタリングではクラスタが数値（0、1、2）として返されるため、ここでは CASE を使用しています。また、種を数値に置換しなかった場合は誤ってカウントされてしまいます。クラスタの番号を種の名前に置換する方法をご存じの方はコメントをお願いします。

結果は以下のとおりです。

結果には詳細な分析データが表示されています。

真陽性や**偽陰性**などの知識を深めたい方は、「適合率と再現率」を参照してください。

まとめ

PMML Model Tester がデータに対してモデルをテストできる非常に便利なツールであることが分かりました。
このツールは詳細な分析データ、グラフ、SQL 実行機能を提供しますので、
別途ツールを用意しなくてもモデルをテストすることができます。

リンク

前の記事

PySpark2PMML

JPMML

機械学習パイプライン

Apache Spark のドキュメント

#Artificial Intelligence (AI) #API #Python #ビッグデータ #分析 #初心者 #Machine Learning (ML) #InterSystems IRIS

ソースURL:
https://jp.community.intersystems.com/post/ml-%E3%83%A2%E3%83%87%E3%83%AB%E3%82%92-intersystems-iris-%E3%81%AB%E8%AA%AD%E3%81%BF%E8%BE%BC%E3%82%80