Apache Sparkによる機械学習入門

株式会社MNU 小峰 央志

自己紹介

小峰 央志 (KOMINE Hisashi)

株式会社MNU 取締役

Facebook: Hisashi KOMINE
Twitter: @hssh

サーバサイドアプリケーション開発、

Webフロントエンド開発、DevOps、スクラムなど

Python、JavaScript、PHP、Rubyなど

機械学習とは

- 与えられたデータからパターンやルールを見つけ出し、未知のデータに対して知見を得たり、分類パターンを見つけたりする
- 大きく分けて以下のような手法がある
  - 教師あり機械学習: 答えの与えられる学習
  - 教師なし機械学習: 答えの与えられない学習
  - 強化学習: 「行動の選択肢」と「報酬」による学習

機械学習アルゴリズムの例

- 回帰分析
  - 教師データからその背景にある関数をモデル化し、既存の教師データに含まれていない未知の入力に対して予期される結果を推定する
- 分類
  - 複数のグループに分類された教師データから母集団におけるグループの分布をモデル化し、未知の入力がどのグループに属するのかを推定
- クラスタリング
  - 教師なしのデータに対して、類似性によりグループ分けを行う

機械学習とビッグデータ

- SNSやIoTなどの普及によって取り扱う必要のあるデータは日々増加してきている
- これらのデータを効率的・効果的に処理するためには、大量のデータを分散処理したり、それらを機械学習によって自動処理するような仕組みが必要
- データの形式も従来のRDBに保存されていたような構造化データ意外にも非構造化データを処理する必要がでてきている

Apache Spark

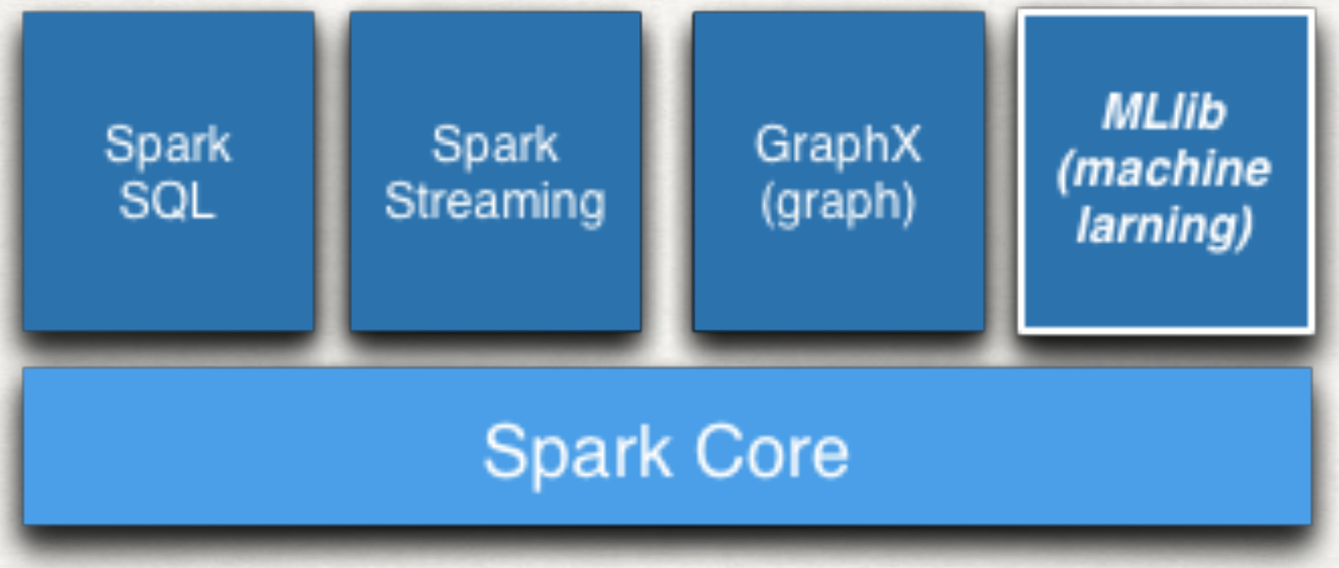
- 大量のデータをオンメモリ高速に処理するためのシステム
- MapReduceのパフォーマンス面の問題などを解決
- 2009年UCバークレーのRAD Labで誕生
- 2010年3月にオープンソース化
- 2013年の6月にはApache Software Foundationへ移行
- 現時点の最新バージョンは 2.0.1

Apache Sparkの特徴

- Speed：Hadoop MapReduceと比較してon-memoryで100倍、on-diskで10倍の速度
- Ease of Use：豊富なAPIによる各種プログラミング言語（Java、Scala、Python、R）での開発
- Generality：SQLや機械学習ライブラリ、グラフ処理などの多様な処理の組み合わせ
- Runs Everywhere：Hadoop、Mesos、スタンドアロンもしくはクラウドでの実行

Apache Spark の汎用性

- Spark SQL: SQL処理用ライブラリ
- Spark Streaming: リアルタイム処理用ライブラリ
- GraphX: グラフ計算、可視化ライブラリ
- MLlib: 機械学習用ライブラリ



IBM Bluemix

- IBM Bluemix
  - https://www.ibm.com/cloud-computing/jp/ja/bluemix/
  - IBMのSaaSサービス (先日PaasのSoftLayerとブランド統合)
  - 初回登録後クレジットカードの登録なしで30日間無料利用可能

Apache Spark on Bluemix

- Apache Spark on Bluemix
  - トライアル期間であれば無料で利用可能
  - Jupyter Notebookによるインタラクティブなデータ分析
- 主なソフトウェア
  - Apache Spark 1.6.0
  - Python 2.7.11
  - Jupyter notebook 4.0.6
  - iPython 4.0.1


IBM Bluemix Ready? Try the new Bluemix | New! Try OpenWhisk
ダッシュボード
ソリューション
カタログ
料金
資料
コミュニティー
1
カタログを選択
組織: komine@usa-mi...
spark
2
sparkを検索
スターター
ボイラープレート
計算
ランタイム
コンテナー
サービス
Watson
モバイル
DevOps
Web とアプリケーション
ネットワーク
統合
データおよび分析
セキュリティー
ストレージ
ビジネス・アナリティクス
サービス // 卓越したすべてのアプリのビルディング・ブロック
データおよび分析
無限の可能性を備えた重要なデータ・サービス
選択のヘルプ
Apache Spark
IBM
3
「Apache Spark」をクリック
さらに詳しい情報
Bluemix Labs Catalog を確認して、試験ランタイムと試験サービスを試してみましょう。
Bluemix Labs Catalog

Apache Spark

IBM

公開日
2016/07/05

作成者
IBM

タイプ
サービス

資料の表示

Apache Spark is an open source cluster computing framework optimized for extremely fast and large scale data processing, which you can access via the newly integrated notebook interface IBM Analytics for Apache Spark. You can connect to your existing data sources or take advantage of the on-demand big data optimization of Object Storage. Spark plans are based on the maximum number of executors available to process your analytic jobs. Executors exist only as long as they're needed for processing, so you're charged only for processing done.

- **Incredibly Fast**

  Apache Spark delivers 100x the performance of Apache Hadoop for certain workloads because of its advanced in-memory computing engine.

- **Easy to Use and Powerful**

  Apache Spark's Streaming and SQL programming models backed by MLlib and GraphX make it incredibly easy for developers and data scientists to build apps that exploit machine learning and graph analytics. Because the service is 100% compatible with Apache Spark, developers can build their apps and run them against the IBM managed service to benefit from operational, maintenance, and hardware excellence.

- **Convenient Data Storage**

  Object Storage enables a convenient way to upload your data from a file for immediate use by your Spark instance. You can set up Object Storage directly from the Spark service interface.

プランの選択　　表示している月々の価格の対象国または地域: 日本

プラン	フィーチャー	
✓ Personal	2 Spark Executors	¥74.00 JPY/Instance-Hour
	An entry level plan to run programs using up to 2 Spark executors	
Reserved Enterprise	30 Spark Executors	-

ご利用条件

サービスの追加

スペース:
dev

アプリ:
アンバインドのまま

サービス名:
Apache Spark-0p

資格情報名:
1 dentials-1
プランはPersonal

選択済みプラン:
Personal

作成
2 「作成」をクリック

IBM Bluemix Ready? Try the new Bluemix | New! Try OpenWhisk

ダッシュボード ソリューション カタログ 料金 資料 コミュニティー

'dev' のダッシュボ...

Apache Spark-mu

資料

Work with Notebooks and Spark

Create and run analytic code in interactive notebooks and share these notebooks with others.

NOTEBOOKS

1 NOTEBOOKS をクリック

Monitor Spark Usage

Get details about your Spark instance, such as the history of jobs and memory usage in notebooks and applications.

Job History

IBM Bluemix

DASHBOARD SOLUTIONS CATALOG PRICING DOCS COMMUNITY

Services

Data

Apache Spark-mu

My Notebooks Object Storage

NEW NOTEBOOK

2 NEW NOTEBOOK をクリック

You have no notebooks. To get started, create a new notebook. You can create a new notebook from scratch, by uploading an existing notebook or by using one of the samples.

IBM Bluemix

DASHBOARD SOLUTIONS CATALOG PRICING DOCS COMMUNITY

Services

Data

Analytics

Exchange

Create Notebook

Blank From File From URL Samples

Name*

1 ノートブック名を記入

Sandbox

43 Characters Remaining

Description

Language*

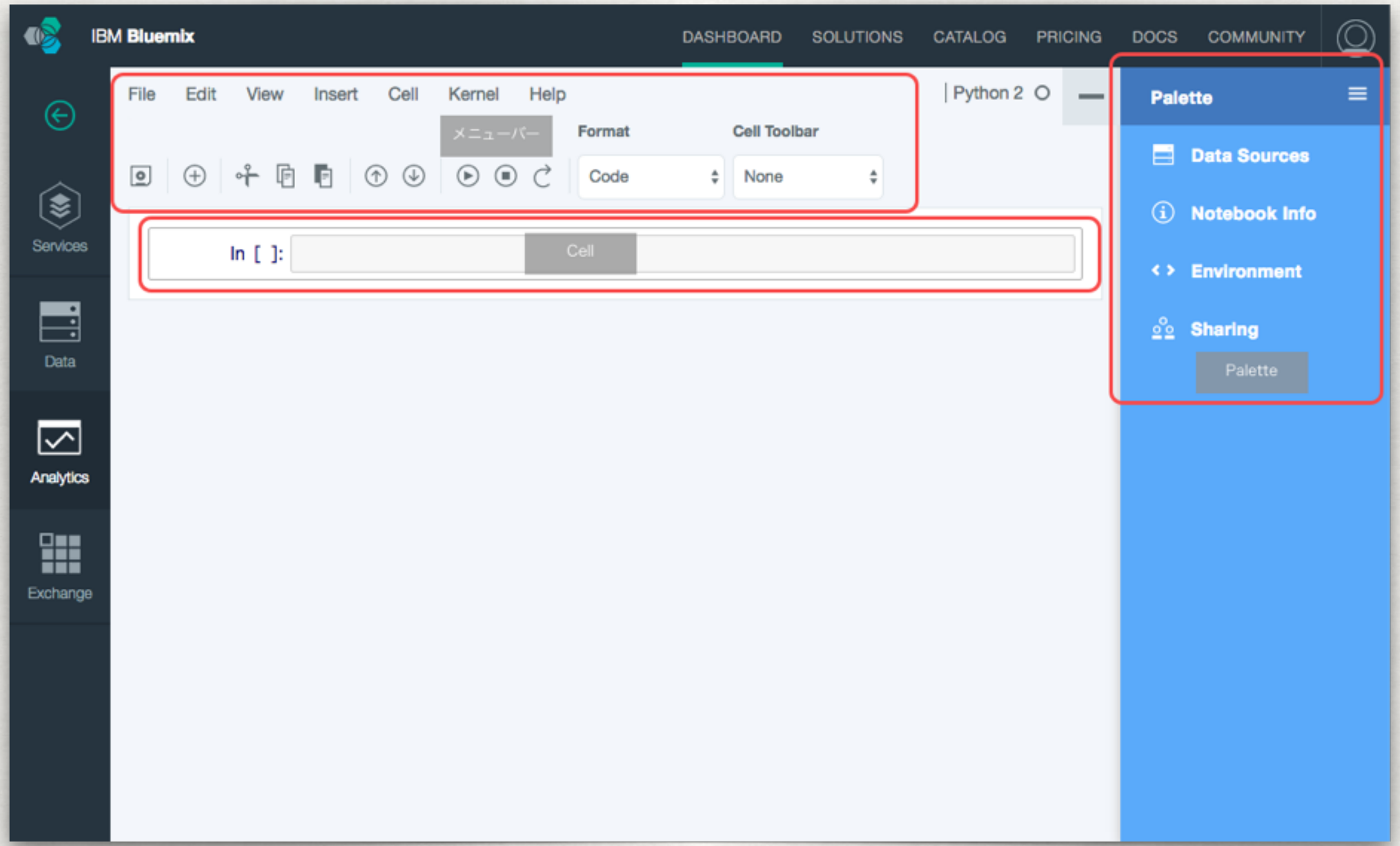
Python Scala R (Technical preview)

2 言語はPythonを選択

3 ノートブックを作成

Cancel

CREATE NOTEBOOK


IBM Bluemix
DASHBOARD
SOLUTIONS
CATALOG
PRICING
DOCS
COMMUNITY
File
Edit
View
Insert
Cell
Kernel
Help
Python 2
メニューバー
Format
Cell Toolbar
Code
None
In [ ]:
Cell
Palette
Data Sources
Notebook Info
Environment
Sharing
Palette
Services
Data
Analytics
Exchange

SparkContext

- SparkContext
  - Sparkの演算クラスタとのやりとりを管理するオブジェクト
  - Notebookでは「sc」という変数名であらかじめ初期化されている
  - RDDを生成するために利用する
- RDD
  - Resilient Distributed Datasets、耐障害性分散データセット
  - Sparkの演算クラスタ上に分散されたデータをコレクションとして表現
  - Sparkでの操作はRDDの生成、変換およびmap()やreduce()などのメソッド呼び出しとして表現される

SparkContextとRDD

Sparkのバージョン確認

```
type(sc)
sc.version
```

RDD生成、処理

```
rdd = sc.parallelize(range(1, 11))
print(rdd.reduce(lambda x, y: x + y))
```

MLlib

- MLlib
  - MLlib Main API
    - DataFrameに対応した機械学習用API群
  - MLlib RDD-Based API
    - RDDに対応した機械学習用API群
    - Spark 2.0.1ではメンテナンスステータス
- DataFrameはRDDよりも機能が充実したデータフォーマットなので
  最近は直接RDDをさわる事はあまりない模様

Case1: 小説書きだし部分による著者推定

- トレーニングデータとして2人の著者(太宰治、宮澤賢治)の小説書き出し部分を利用する
- トレーニングデータから索引語頻度 TF (Term Frequency)を特徴量として抽出
- 抽出した特徴量をラベリングし、ロジスティック回帰により学習を行う

- 索引語頻度 TF (Term Frequency)
  - ある文書の中での各単語の出現頻度

Case1: ロジスティック回帰による著者推定

- dazai.txt

  メロス は 激怒 した 必ず かの 邪智 暴虐 の 王 を 除か なければ ならぬ と 決意 した メロス には 政治 が わからぬ
  昔 の 話 で ある 須々木 乙彦 は 古着屋 へ はいって 君 の ところ に 黒 の 無地 の 羽織 は ないか と 言った
  これ は れい の 飲食店 閉鎖 の 命令 が 未だ 発せ られない 前 の お話 で ある 新宿 辺 も こんど の 戦火 で
  朝 食堂 で スウプ を 一さじ すっ と 吸って お母さま が あ と 幽 な 叫び 声 を お挙げ に なった
  或る とし の 春 私 は 生れて はじめて 本州 北端 津軽 半島 を 凡そ 三週間 ほど かかつて 一周 した ので あるが

- miyazawa.txt

  オツベル と きたら 大した もんだ 稲扱 器械 の 六台 も 据え つけて のんのんのんのんのんのん と 大そろ しない 音を たてて やっている
  をかし な はがき が ある 土曜日 の 夕がた 一郎 の うち に きました
  では みなさん は そういう ふうに 川 だと 言われ たり 乳 の 流れた あと だと 言われ たり していた この ぼんやり と 白い もの が ほんとう は 何 か ご承知 ですか
  グスコーブドリ は イーハトーヴ の 大きな 森 の なか に 生まれ ました おとうさん は グスコーナドリ と いう 名高い 木こり で どんな 大きな 木 でも まるで 赤ん坊 を 寝かし つける ように わけ なく 切って しまう 人 でした
  二人 の 若い 紳士 が すっかり イギリス の 兵隊 の かたち を して ぴかぴか する 鉄砲 を かついで 白熊 の ような 犬 を 二疋 つれて だいぶ 山奥 の 木の葉 の かさかさ した とこ を こんな こと を 云い ながら あるいて おり ました

Case2: イチローの打率推定

- BASEBALL-REFERENCE.comから取得した2001年から2016年までの打率を学習
  - http://www.baseball-reference.com/
- 線形回帰分析を用いて今後の打率データの推定を行う

Case2: イチローの打率推定

- 線形回帰分析
  - 与えられたデータ点に
    フィットする線形モデルを
    推定する