

環境構築

プログラミング言語を使用する上での環境構築とは、機械語への変換に必要なツールを自分のローカル環境(PC)上に落としてきて、使えるようにする作業を言います。

R と Python のコードが使えるように環境構築の方法を説明しています。

R の環境設定方法

R と RStudio のインストール


R はフリーの、オープンソースのプログラミング言語で、様々な環境上で動きます。(Linux 系、Windows、MacOS など OS を問いません)。また、統計解析・作図にも強い点が特徴です。R のデフォルトの機能だけでも便利に使うことができますが、R はパッケージ(package)を活用することで、その機能を拡張することができます。

RStudio は、R をより使いやすくするための統合開発環境(IDE)です。

R は直接操作するのではなく、RStudio 経由で R を操作するほうが利便性が良いため RStudio のインストールも必要です。

R のインストール

・統計数理研究所の CRAN ミラーサイト <https://cran.ism.ac.jp/>



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)
[About R](#)

The Comprehensive R Archive Network

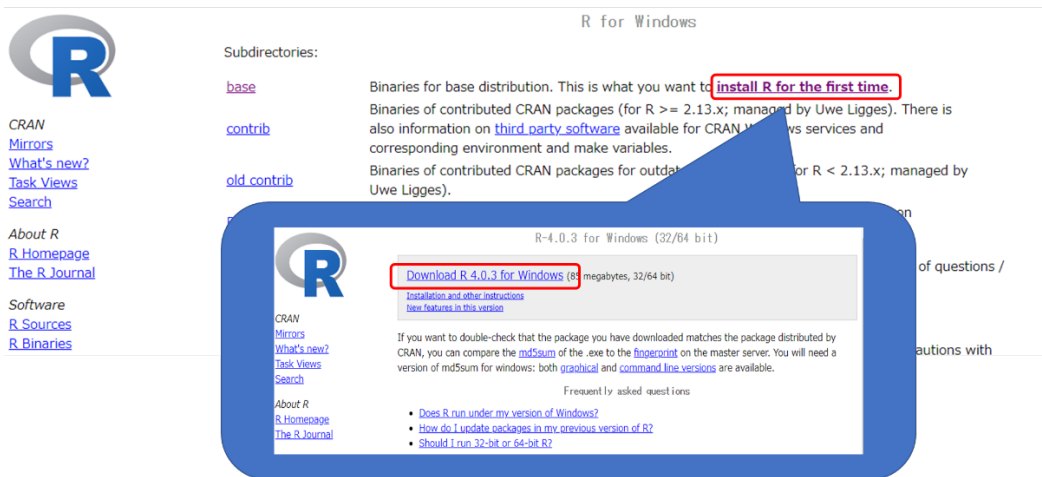
Download and Install R

Precompiled binary distributions of the base system and contributed packages,
Windows and Mac users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

自分の環境(Linux、MaxOS、Windows)に適したリンクを選択します。



「install R for the first time」をクリックします。(上記反転している部分)。最新バージョンの R をダウンロードするための画面に移動します。「Download～」をクリックするとインストーラーがダウンロードされます。

ダウンロードの完了後、インストーラーを起動したら、「OK」と「次へ」ボタンだけを押してインストールを進めてください。

※設定の変更は不要です。もし変更する場合は、インストール場所にはご注意ください。「パスに日本語が含まれる場所」や「同期しているクラウドストレージ上」にインストールすると正常に動作しないことがあります。

Rstudio のインストール

RStudio の公式サイトから、最新版の RStudio インストーラーをダウンロードしてください。Installers for Supported Platforms の中から、各自の環境に合ったインストーラーを選んでください。

<https://rstudio.com/products/rstudio/download/#download>

RStudioデスクトップ1.4.1103 .リリースノート

1.1。 Rをインストールします。 RStudioにはR3.0.1+が必要です。

2.2。 RStudioデスクトップをダウンロードします。 お使い

のシステムに推奨
[RSTUDIO FOR WINDOWSをダウンロードする](#)
 1.4.1103 | 156.96MB

Windows 10/8/7 (64ビット) が必要



すべてのインストーラー

Linuxユーザーは、オペレーティングシステムのセキュリティポリシーによっては、インストール前にRStudioの公開コード署名キーをインポートする必要がある場合があります。

RStudioには64ビットのオペレーティングシステムが必要です。32ビットシステムを使用している場合は、古いバージョンのRStudioを使用できます。

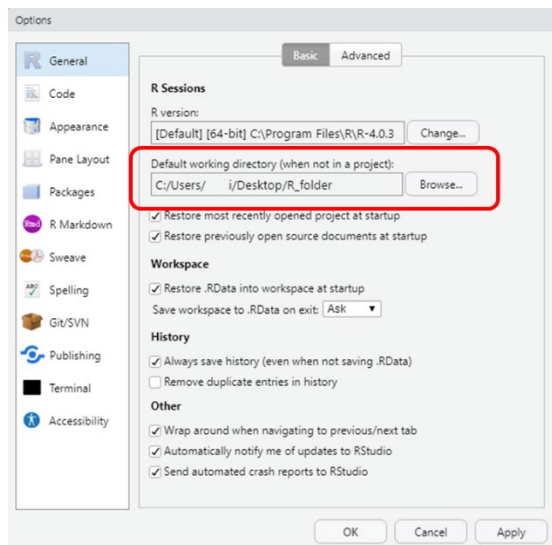
インストーラーを起動したら、R と同様に他の設定は一切変更せず、「次へ」と「インストール」のボタンだけを押してインストールを進めてください。

インストールが完了したら、次に作業ディレクトリの設定と文字コードの設定を行います。

作業ディレクトリの設定

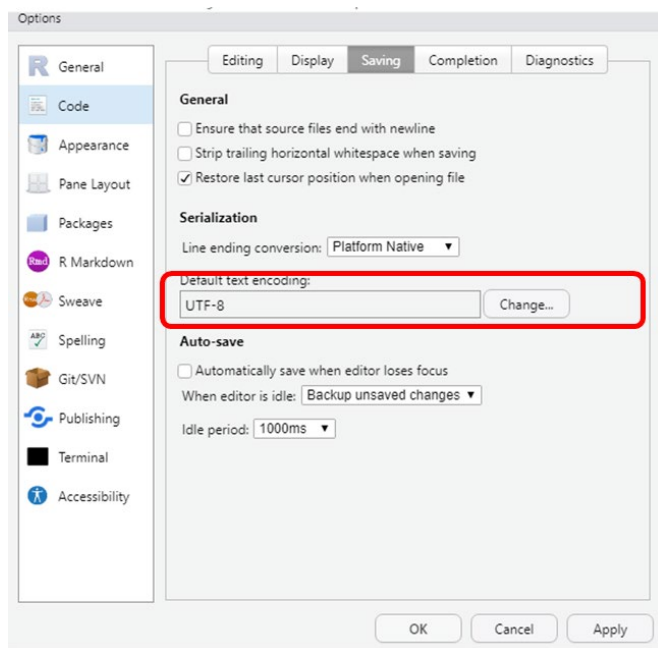
ファイルを参照する場所を指定します。

「Browse...」ボタンを押して、任意のフォルダを作業ディレクトリに設定しましょう。ここではデスクトップ上に「R_folder」というフォルダを新しく作り、そこをデフォルトの作業ディレクトリに指定しました。



文字コードの設定

プログラムで、日本語のようにマルチバイト文字を使用する場合、文字コードによっては文字化けが生じることがあります。そのようなことが無いようにまず文字コードを設定いたします。左のカラム上から2つ目の「Code」、上部のタブ左から3つ目の「Saving」を選び、以下の画面を開きます。「Change...」ボタンを押すことで、任意の文字コードを指定できます。ここでは UTF-8 を選んでいます。



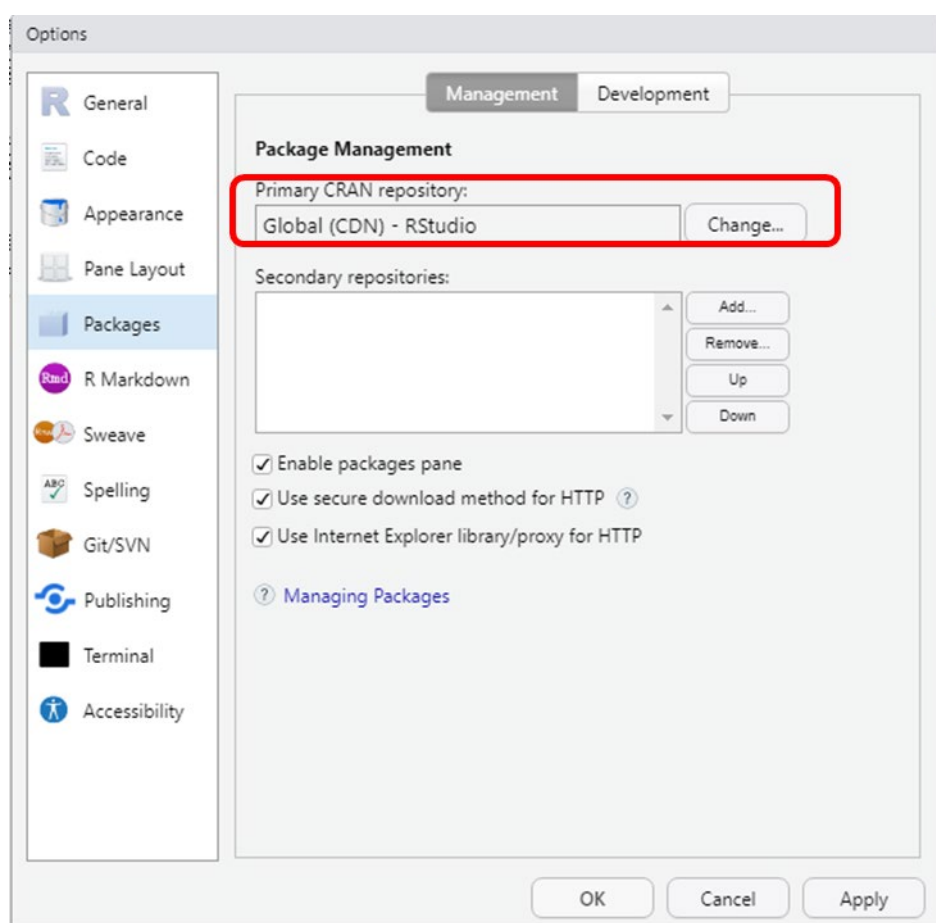
パッケージをダウンロードする CRAN リポジトリの設定

R はパッケージを次々にインストールしていくことによって、その機能を拡張することが出来ます。パッケージは、CRAN リポジトリからインターネット経由でダウンロード&インストールすることが出来ます。その都度リポジトリを選ぶことも出来ますが、ここで設定しておいた方が後々手間を省けます。

「Change...」ボタンを押すとリポジトリを変更可能です。日本国内では、以下の機関が CRAN リポジトリを管理しているので、いずれかを選ぶと良いでしょう。

統計数理研究所(Japan (Tokyo) [[https](https://www.imstat.org/)] - The Institute of Statistical Mathematics, Tokyo)

山形大学(Japan (Yonezawa) [[https](https://www.yamagata-u.ac.jp/)] - Yamagata University)



ここまで出来れば設定完了です。

PYTHON の環境設定方法

Google Colaboratory(略称: Colab)は、ジュピター・ノートブックをクラウド上で動かしますが、Python や Numpy など、機械学習に必要なほぼ全ての環境がすでに構築されています。必要なのはブラウザのみで、すぐに実行できるサービスです。

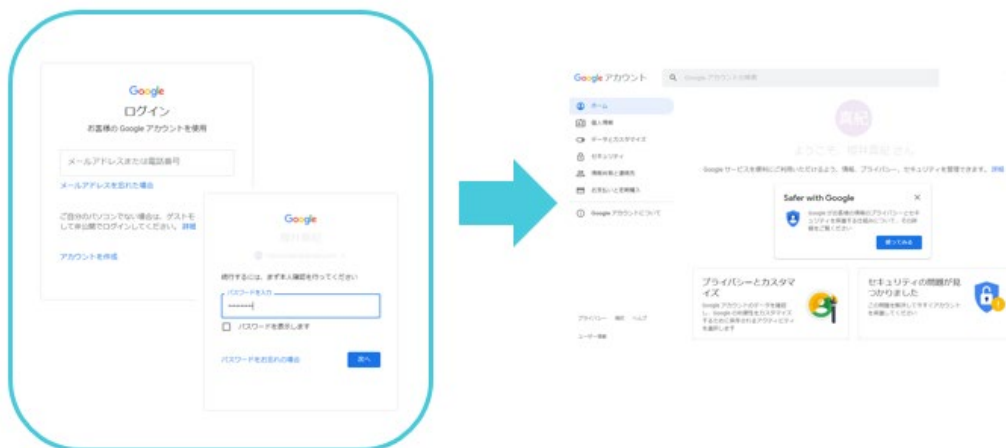
Google Colaboratory のメリットは、先ほど説明した通り、Google アカウントがあれば、環境構築が不要ですすぐに開始できます。また Google Colab で書いたコード(ノートブック)は、グーグルドライブで保存されま

す。ですので、チーム内でノートブックの共有などが非常に簡単で、かつ権限管理など Google Drive 上で行えるので安心でもあります。

最後に、GPU(Tesla K80 GPU)を無料で使用できます。機械学習では大規模なデータを、高負荷がかかる計算をすることが多くあります。自身のパソコンで処理を行う場合、訓練に 12 時間かかることも多々ありますが、Google Colab の GPU 環境を使うことで、時間短縮が可能です。

以下、Colaboratory(略称: Colab)の基本設定になります。

1. Google アカウントにログイン



2. Google Drive で Google Colab 用の新規フォルダを作成



3. 作成した Google Drive 上のフォルダと Google Colab の連携

フォルダ名のドロップダウンをクリックして「アプリで開く」>「アプリを追加」をクリック

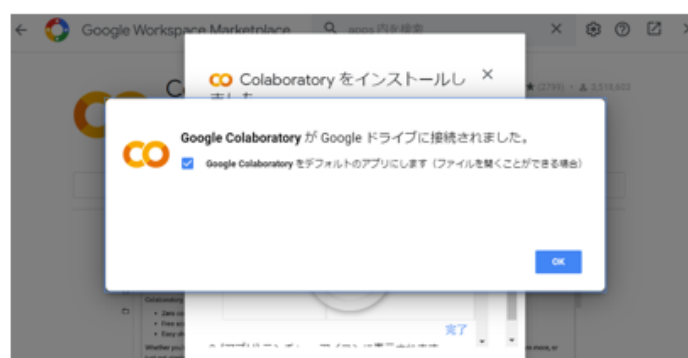


ポップアップの検索窓に「colaboratory」と入力、Google Colab のアプリが絞り込まれます。こちらの「インストール」をクリック



インストールを進めると以下のポップアップが出ますので、こちらの「OK」をクリックで完了です。

インストールを進めると以下のポップアップが出ますので、こちらの「OK」をクリックで完了です。



4. Colaboratory で新規ファイルを作成
「右クリック」>「その他」>「Colaboratory」で新規ファイルを作成



作成されたファイル



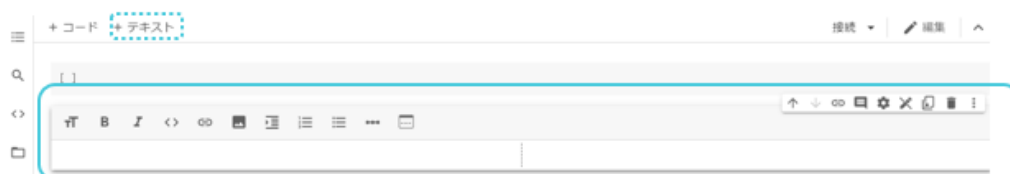
セルには主に「+コード」と「+テキスト」の二種類があります。

デフォルトでは、上記のとおり、コードセルが表示されており、コードを入力できる状態になっています。

- ・「+コード」⇒ 新しいコードセルをノートブックに追加
- ・「+テキスト」⇒ 新しいテキストセルをノートブックに追加

Colaboratory

「+テキスト」ボタンをクリックするとテキストセルが追加されます。



GPU の設定

- デフォルトは CPU となっています。
- ノートブックのメニューバーから「編集」>「ノートブックの設定」をクリックしましょう。こちらの画面で GPU の選択が可能ですので、GPU を選択しましょう。



GPU の設定(確認)

- 以下の通りコードを入力して実行すると確認することができます。

```
# TensorFlow経由でデバイス設定の確認が可能です
from tensorflow.python.client import device_lib
device_lib.list_local_devices()

[name: "/device:CPU:0"
 device_type: "CPU"
 memory_limit: 268435456
 locality {
 }
 incarnation: 1246946331566648751, name: "/device:GPU:0"
 device_type: "GPU"
 memory_limit: 14674281152
 locality {
   bus_id: 1
   links {
   }
 }
 incarnation: 13577881043160061023
 physical_device_desc: "device: 0, name: Tesla T4, pci bus id: 0000:00:04.0, compute capability: 7.5"]
```


ライブラリの紹介

ライブラリとは、いくつかのパッケージをまとめてインストールできるようにしたものです。（関数やクラス、モジュール、パッケージなどを総称してライブラリと呼ぶこともあります。）ここでは、R と Python の代表的なライブラリを紹介します。

R の代表的なライブラリ

R 言語では、通常はライブラリと呼ばれるファイルのことをパッケージといいます。

R 言語にも様々なパッケージが準備されています。機械学習に便利な代表的パッケージをいくつか紹介します。

【R 言語パッケージ】

dplyr

dplyr(ディプライヤー)はデータフレーム操作のパッケージです。データの絞り込み、追加や並べ替え、グルーピングなどを関数との組み合わせでビックデータを扱うときにも効率的に行うことができます。

stringr

stringr(ストリンガー)は文字列操作のパッケージです。文字列の置換や正規表現による検索など、テキストを扱う際によく利用する機能が収められています。

ggplot2

ggplot2(ジージープロットツー)はグラフ描画のパッケージです。先に紹介したように R 言語には plot コマンドがありますが、それよりも綺麗で複雑な描画が行えるものです。

Caret (Classification And REgression Training)

caret(キャレット)は機械学習のタスクを効率化するパッケージです。機械学習のアルゴリズムが組み込まれたものでもあります。caret は上の 3 つのように RStudio のサイトからはダウンロードできません。CRAN のパッケージ一覧から探してダウンロードして下さい。



PYTHON の代表的なライブラリ

Python ライブラリには、Python をインストールした時点で一緒にインストールされる「標準ライブラリ」と、ダウンロードなど追加インストールが必要な「外部ライブラリ」が存在します。

事前に定義せずに利用できる関数は標準ライブラリの中で定義されています。

また、Python は外部ライブラリが豊富に存在していることから、AI と親和性が高いとも言われています。

【代表的なライブラリ】

Pandas: SQL や R 言語のようなデータフレームの加工が可能

NumPy:行列演算用で C/C++ と Fortran で実行されるため非常に高速

Matplotlib, seaborn:グラフ描画

scikit-learn:機械学習のアルゴリズム

Chainer:深層学習のアルゴリズム

OpenCV:画像処理のアルゴリズム

※ Python で使うことができるパッケージ(ライブラリ)は、サイトにまとめられています。
(<https://pypi.python.org/pypi>)

参考書籍

機械学習、データサイエンスを入門的に学ぶ上で、おすすめの書籍を記載いたします。

著者:工藤卓哉

これからデータ分析を始めたい人のための本

発行元:PHP 研究所

著者:椎名 洋 (著), 姫野 哲人 (著), 保科 架風 (著), 清水 昌平 (編集)

データサイエンスのための数学 (データサイエンス入門シリーズ) 単行本

発行元:講談社

著者:金明哲

R によるデータサイエンス データ解析の基礎から最新手法まで

発行元:森北出版

著者:高橋威知郎、安宅和人、河本薫、吉田隆光、北川拓也

トップデータサイエンティストが教える データ活用実践教室

発行元:日経 BP

著者:谷尻かおり

文系プログラマーのための Python で学び直す高校数学

発行元:日経 BP

著者:下山輝昌、三木孝行、伊藤淳二

Python 実践機械学習システム 100 本ノック

発行元:秀和システム

著者:下山輝昌、三木孝行、松田雄馬

Python 実践データ分析 100 本ノック

発行元:秀和システム

引用・参考文献

[書籍]

監修:加藤公一 秋庭伸也・杉山阿聖・寺田学[共著]
見て試してわかる 機械学習アルゴリズムの仕組み 機械学習図鑑
翔泳社

機械学習研究会著・株式会社 ALBERT データ分析部 安達章浩・青木健児監修
60分でわかる! 機械学習&ディープラーニング超入門
技術評論社

大城 信晃(監修・著者)
AI・データ分析プロジェクトのすべて[ビジネス力×技術力=価値創出]
技術評論社

西川 仁(著), 佐藤 智和(著), 市川 治(著), 清水 昌平(編集)
テキスト・画像・音声データ分析(データサイエンス入門シリーズ)
講談社

椎名 洋(著), 姫野 哲人(著), 保科 架風(著), 清水 昌平(編集)
データサイエンスのための数学
講談社

江崎貴裕(著)
分析者のためのデータ解釈学入門 データの本質をとらえる技術
ソシム

江崎 貴裕(著)
データ分析のための数理モデル入門 本質をとらえた分析のために
ソシム

合同会社 アイキューバータ共同代表 下山 輝昌(著), 松田 雄馬(著), 三木 孝行(著)
Python 実践データ分析 100本ノック
秀和システム

[WEB ページ]

総務省「人工知能(AI)の現状と未来 第1部特集 IoT・ビッグデータ・AI～ネットワークとデータが創造する新たな価値～」
<https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/html/nc142000.html>

データサイエンティスト協会 データサイエンス 100本ノック(構造化データ加工編)
<https://github.com/The-Japan-DataScientist-Society/100knocks-preprocess>