极客大学

极客大学机器学习训练营
Python 和 R 基础

王然

众微科技 AI Lab 负责人

二〇二一年四月十八日

大纲

极客大学

1. 环境搭建
2. 开发工具
3. Python 和 R 基础
4. 预习内容
5. 参考文献和阅读材料

大纲

极客大学

两种选择

极客大学

- 采用CoLab
  - 不需要花费大量时间搭建环境
  - 免费（低费用）使用 GPU/TPU 和多核
  - 网速快
  - 环境标准；助教容易帮忙复现问题
  - 缺点：需要科学上网
- 采用自己搭建的环境
  - 环境搭建十分困难
  - API 经常会变
  - 有利于本地实验
  - 生产环境中必须学会搭建环境

本课程建议采用 CoLab

一般搭建环境步骤

极客大学

- 基本 GCC 开发环境
- Anaconda 或 Miniconda 安装
- Conda 环境下 R 的安装
- Docker 的安装
- CUDA 安装

基本 GCC 开发环境

极客大学

- 根据 Linux 系统版本不一样，采用命令不一样。
- 以 Ubuntu 为例，命令为 sudo apt-get install build-essential。
- 注意采用至少 gcc-4.8 以上版本。

插曲：如何从源代码安装

极客大学

- 将源代码下载
- 运行.configure 或.bootstrap 命令
- 运行 make -j4 命令。其中-j4 表示采用四个线程进行编译。
- 运行 sudo make install 命令。

Conda 安装

极客大学

- 可以选择 Anaconda 或 Miniconda。
- 下载文件后直接运行即可 (.sh)。
- 安装完之后运行 source ~/.bashrc 命令修改环境变量。
- 建议安装 miniconda。后续采用 pip 安装其他软件包。
- 本课程采用 python 3.7 版本。需要修改版本请用 conda install python=3.7。

安装 R

极客大学

- 注意不要直接用 package manager 安装。
- 采用 conda install r-essentials r-base 命令。
- 好处是可以在 jupyter notebook 当中使用 R 和 Python。

Docker 安装

极客大学

- Docker 是类似于一种虚拟机的环境隔离方式。Docker 比虚拟机更轻，所以是微服务的核心组件之一。
- 安装 Docker 请参照官方文档。注意执行 post installation step。
- 需要使用 docker 镜像时，请使用 docker pull image:tag 命令。
- 如果需要在 docker 中运行 GPU，则需要安装 NVidia Docker。请根据官网命令进行安装。

Docker 使用方法

极客大学

- 基本命令 docker run -it –rm image:tag。
- -it 表示采用交替式的运行。
- –rm 表示运行完后删除 container。节省硬盘资源。
- 其他命令
  - -v 将本地文件映射到 docker 文件中。
  - -p 将端口进行映射。使用 Jupyter Notebook 时候有用。

CUDA 安装

极客大学

- 版本选择：最新版或适合你的深度学习框架的版本。
- 注意：当 host 系统上的 CUDA 版本高于 docker 内的时候，可以进行运行；反过来不行。
- 安装步骤：
  - disable nouveau。不同系统不一样，需要查看不同命令。
  - 重启系统并进入到 physical shell。ubuntu 的快捷键（可能是）CTRL+ALT+F2。
  - 关闭 xserver。命令可以有很多，例如 sudo init 3。
  - 官网下载 runfile 文件！
  - 运行时候结尾加上–no-opengl-libs 命令。否则会造成登陆循环。
- 注意：CUDA 安装十分危险。安装失败后补救措施一般只能采用重装系统的方式。

大纲

开发工具

极客大学

本课程我们主要使用 PyCharm 和 Jupyter Notebook 进行开发。

- PyCharm 优点：代码补全，重命名（SHIFT+F6），提供 debugger，可以提取函数（refactor→ extract_methods），可以自动实现 PEP8。建议搭配 VIM 使用。在学习源代码时候可以进行跳进（很容易找到源代码所在地点）。
- Jupyter 优点：交互式运行，避免重复读取大量数据。注意：CoLab 就是一个很类似于 Jupyter 的程序。

建议开发方法：将代码在 IDE 进行修改，然后粘贴进行测试。注意，安装一个新的 python 包之后，Jupyter 服务需要重启才能读进去。

Jupyter 魔法函数

极客大学

Jupyter 当中有很多 (?) 很方便的魔法函数。具体请见 jupyter notebook 文件。

大纲

极客大学

1. 环境搭建
2. 开发工具
3. Python 和 R 基础
4. 预习内容
5. 参考文献和阅读材料

学习新语言的步骤

极客大学

- 变量和赋值。
- 控制循环。
- 函数定义。
- 类的定义。
- 常见函数、数据结构。

我们不会对此进行完整的复习，只会讲解重点

为什么要学习 R

极客大学

- 操作方便：例如??dplyr 包当中可以实现迅速的数据探索性分析；
- 很多统计性常用的包：例如使用多重填充后用，GLM+ 样条 +L1 损失构造变量，这种特征构造对于和线性模型有关的（例如深度学习）有很大关系。注意 R 当中不需要考虑模型不收敛问题，解决这个问题用 C（python 速度过慢）大概需要 1000 行；

Python 语法重点

此部分请结合 *Jupyter Notebook* 进行学习。

- 函数定义（type hint，args 和 kwargs）；
- python 中异常处理的方式（exception 和 monad）；
- python 中常见的数据结构；注意 set 和 dict 为无序的；
- python 风格的类，核心为内置函数（built-in function）；
- dataclass 和多参数传递的方式；文档的撰写；
- 装饰器。

大纲

极客大学

1. 环境搭建
2. 开发工具
3. Python 和 R 基础
4. 预习内容
5. 参考文献和阅读材料

- 第二周：对于有 C 基础的同学，请简单复习一下 C 的内容和语法。没有基础不需要复习。可以参考C++ 教程
- 第四周：请阅读 All of Statistics(Wasserman 2013) 的 1，2，3，6，9，11 章。十分重要，看不懂地方请一定注明。

大纲

极客大学

1. 环境搭建
2. 开发工具
3. Python 和 R 基础
4. 预习内容
5. 参考文献和阅读材料


Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Thanks!