数据挖掘实训新人攻略

实验环境熟悉

Linux/Ubuntu

如果你选择使用Linux/Ubuntu，那么恭喜你不用踩微软的那些环境上的巨坑[微笑]

Ubuntu 中的 python 是自带，所以啊，只用动动命令行就行了。
首先，安装写基本包：

sudo apt-get install build-essential libssl-dev libevent-dev libjpeg-dev libxml2-dev libxslt-dev

然后就可以安装 pip，这个类似 python 库的 apt-get，Mac OS 下的 brew
sudo apt-get install python-pip

如果你想安装一个 python 的库，那么请在 terminal 中键入
sudo pip install xxx
E.g sudo pip install xgboost

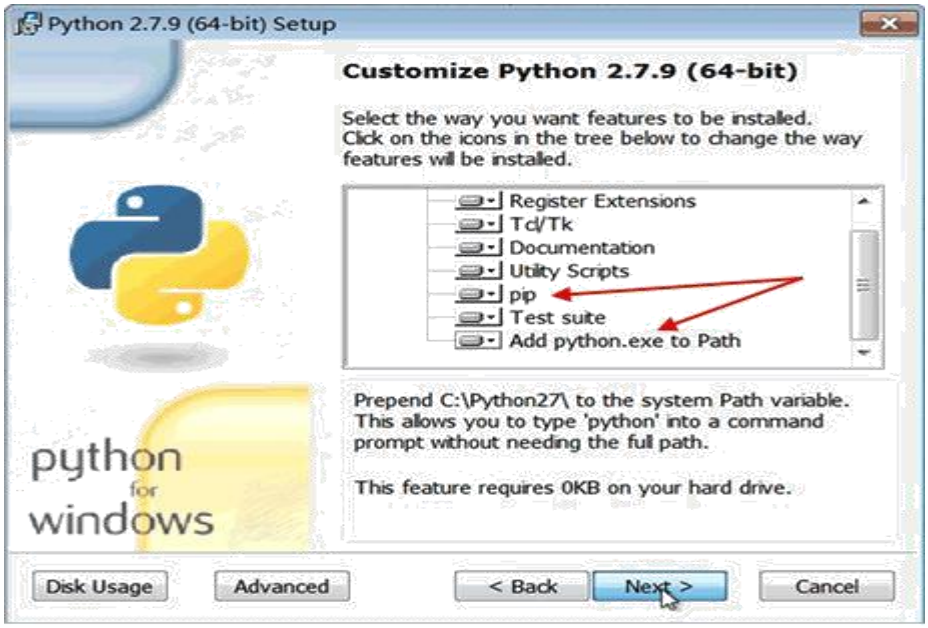
数据挖掘几个基础的包大概是：numpy, scipy, pandas, scikit-learn, statsmodels, matplotlib, xgboost, jupyter。如果大家觉得一个个安装的很麻烦，请使用我写的 requirements.txt，使用方法就是切换到该文件的路径，执行：
pip install -r requirements.txt
包应该就自动安装好了

接下来就是 IDE，我个人推荐 Jupyter，这个是一个网页应用，可以用来写代码。具体的使用见 Two Sigma Connect Example

Windows（不推荐）

a. Python 安装教程

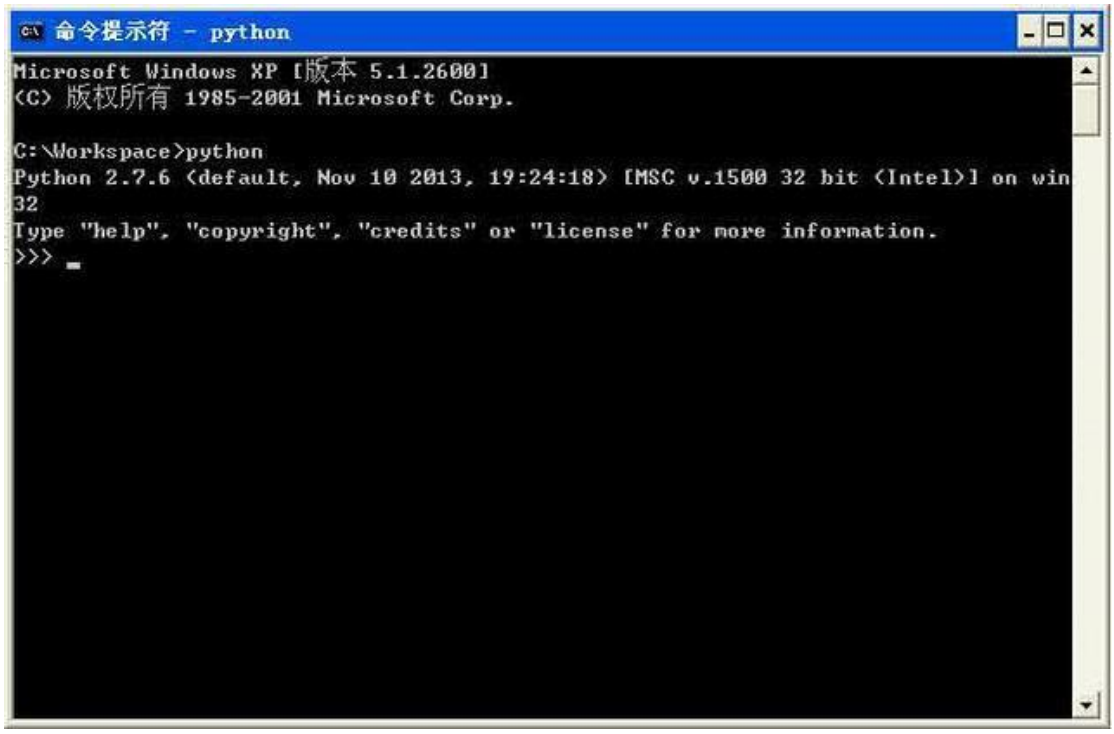
首先，从 Python 的官方网站 python.org 下载最新的 2.7 版本，网速慢的同学请移步国内镜像。然后，运行下载的 MSI 安装包，在选择安装组件的一步时，勾上所有的组件：



特别要注意选上 `pip` 和 `Add python.exe to Path`，然后一路点“Next”即可完成安装。

安装成功后，默认会安装到 C:\Python27 目录下，然后打开命令提示符窗口，敲入 python 后，会出现两种情况：

情况一：



出现上述画面说明安装成功了。

情况二：

'python'不是内部或外部命令，也不是可运行的程序或批处理文件。

出现此种情况后，说明 python 路径没有加入环境变量中。在我的电脑右键属性后出现下述画面：

点击高级系统设置->环境变量：

在系统变量里的 Path 路径里（如果没有则新建 Path 路径）添加 Python 的安装路径。

完成上述操作后再在命令提示符中输入 Python 检查安装是否成功。

b. Pandas 及 numpy 的安装简单使用

windows 安装 pip 后可以直接利用 pip 来安装所需要的 package。如果没有安装 pip，可以按照下面这个链接安装 pip：

c. http://pip-cn.readthedocs.io/en/latest/installing.html

d. pip install --user numpy scipy pandas

e. 键入上述命令后系统会帮忙自动下载所需要的的 packge 和依赖 package，然后还会自动给你安装上去。

f. 如果发现上述下载源的速度比较慢，推荐使用国内的源进行下载：

g. pip install matplotlib -i http://pypi.douban.com/simple --trusted-host pypi.douban.com

h. pip install numpy -i http://pypi.douban.com/simple --trusted-host pypi.douban.com

i. pip install pandas -i http://pypi.douban.com/simple --trusted-host pypi.douban.com

j. pip install seaborn scipy -i http://pypi.douban.com/simple --trusted-host pypi.douban.com

k. 在 pandas 的使用上，推荐一个教程 10 分钟学会 pandas：
http://pandas.pydata.org/pandas-docs/stable/10min.html

l. 在 numpy 的使用上，推荐一个官方的学习文档：

m. https://docs.scipy.org/doc/numpy-dev/user/quickstart.html

n. Pandas 和 numpy 是处理数据的利器，在做好数据挖掘的同时必须要掌握这两个利器。

Xgboost 的安装

以下内容参照：http://stackoverflow.com/questions/33749735/how-to-install-xgboost-package-in-python-windows-platform/41274589#41274589

首先要下载 xgboost 编译过的文件，可以去
https://drive.google.com/file/d/0B6HXvVF1p1HWcTdwZ2dGYkVyZG8/view
也可以去实验室的 BBS:
http://bbs.inpluslab.com/thread/shu-ju-wa-jue-zi-yuan-fen-xiang-339/

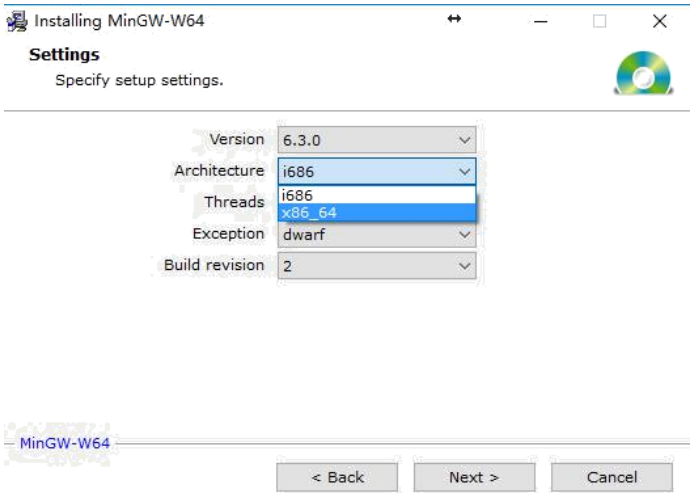
下载之后解压完成之后得到 xgboost/ 文件夹
Xgboost 依赖：NumPy, Scipy 首先确保满足依赖

1. 如果是独立的 python 解释器, 需要将上述得到的 xgboost/ 移动到 site-packages/ 中
2. 如果是 Anaconda，需要将 xgboost/ 移动到 pkgs/ 中

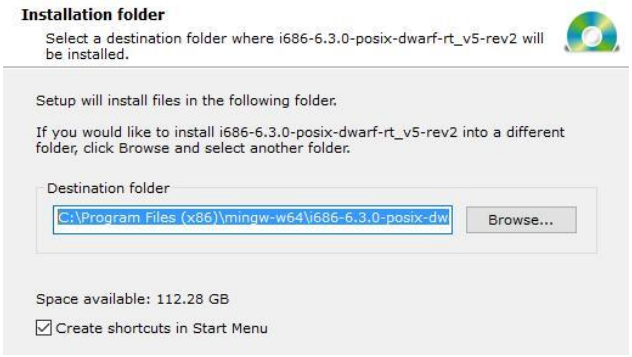
然后命令行进入 xgboost/ 中的 python-packages/ 中，执行如下命令：
python setup.py install
如果发现如下错误：

"WindowsError: [Error 126] "

说明你还没有安装 mingw-64
可以去 http://sourceforge.net/projects/mingw-w64/ 下载安装
安装过程中



这一步选择 x86_64
然后将这里的路径后面加上 \mingw64\bin 添加到系统环境变量中：



然后安装完成后，重新打开一个终端，进入 Python 命令行，输入：
import xgboost
如果没有报错，说明安装成功了

Sklearn 安装

sklearn 是一个开源的 Python 的科学计算库。包含了非常多数据挖掘常用的一些功能模块的函数实现，如 分类，回归，聚类，数据降维，数据预处理，模型选择等，它的官方网站：http://scikit-learn.org/stable/
官网对其介绍如下:

- 是数据挖掘和数据分析中简单有效的工具
- 面向大众，且对不同的代码文件都可复用
- 基于 NumPy, SciPy, 和 matplotlib
- 开源，社区可用

以下内容参照：http://scikit-learn.org/stable/install.html

sklearn 是在 anaconda 中有集成的，所以直接在 Python 终端输入

import sklearn

就可以将这个模块引入。

如果觉得 anaconda 中的 sklearn 版本太低，可以使用如下命令更新：

conda update scikit-learn

入门推荐去看官方的 tutorial:

http://scikitlearn.org/stable/tutorial/basic/tutorial.html

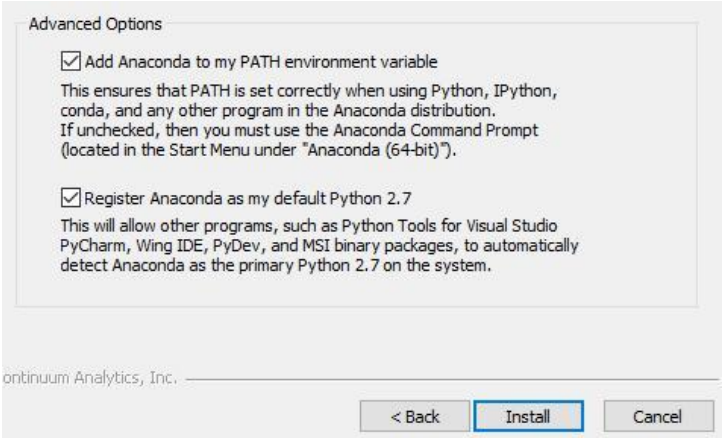
Anaconda 的安装和使用

Anaconda 是一个用于科学计算的 Python 发行版，支持 Linux, Mac, Windows 系统，提供了包管理与环境管理的功能，可以很方便地解决多版本 python 并存、切换以及各种第三方包安装问题。Anaconda 利用工具/命令 conda 来进行 package 和 environment 的管理，并且已经包含了 Python 和相关的配套工具。

以下内容以 Windows 为例：

下载页面 https://www.continuum.io/downloads/

选择系统版本和 Python 版本下载。下载之后安装过程中的这一步：



提示将会把 Anaconda 加入系统的环境变量，同时将 Anaconda 自带的 python 作为系统默认的解释器，这样以后命令行的 python 就是

Anaconda 的自带解释器了。点击 Install,安装完成就可以了。

安装完成后:

进入终端输入 python:

发现已经是 Anaconda 的解释器，说明安装成功。

Anaconda 集成了很多数据挖掘中常用的 Python 库，基本可以满足需求，如果需要额外的依赖，或者删除一些库，可以使用：

conda install xxx

conda uninstall