第一章 概论

任浩

材料物理系

renh@upc.edu.cn

课程背景

- 现有少量相关课程开设参照
  - Utah（https://github.com/sp8rks/MaterialsInformatics）
  - Cambridge (no online material)
  - Georgia Tech (https://www.coursera.org/learn/material-informatic)
  - 东京工大（T0300A404, no online material）
  - NCSU、MIT……
- 无教材
- 无明确课程内容
  - 机器学习、数据库（搭建/使用）
  - 工程材料（工程+商业）

先修基础（你应该会啥）

课程所需基础	培养方案对应	是不是真的会?
单变量微积分	高等数学（必，2）	？？？
多变量微积分	高等数学（必，2）	？？？
线性代数	线性代数（必，3）	？？？
概率与统计	概率论与数理统计（选，3）	？？？
计算机基础	大学计算机（必，2）	~0
程序设计	程序设计（必，1）	~0
物理、化学基础	大学物理、物理化学、材料基	？？？

课程思路

课程定位：一门工具性课程，强调数据获取、分析和挖掘的具体思想、技能和实践操作

- 计算机工具：python基础（随学）、机器学习框架（scikit-learn、TensorFlow、PyTorch……）
- 常用机器学习算法
  - 背景及特性（分类、优缺点、适用范围）
  - 动手操作（通用数据集讲解，材料数据集实操，现有材料数据库及工具包）

参考资料：机器学习&统计学习

- James et al. “An Introduction to Statistical Learning”, 2nd ed, Aug. 2021 (2nd web edition, https://www.statlearning.com/resources-second-edition)
- McKinney (Pandas作者), “Python for Data Analysis”, 2nd ed, Aug. 2022 (3rd web edition, https://wesmckinney.com/book/)
- Géon, “Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow”, O’Reilly, 2nd ed, 2019
- Murphy, “Probabilistic Machine Learning: An Introduction”, MIT Press, 2022. (https://probml.github.io/pml-book/book1.html)

参考资料：数学相关

线性代数：

- D. C. Lay, “Linear Algebra and Its Applications”, （《线性代数及其应用》，机械工业出版社，刘深泉等译，初级）
- P. D. Lax, “Linear Algebra and Its Applications”, （《线性代数及其应用》，人民邮电出版社，傅莺莺等译，进阶）

概率、统计：

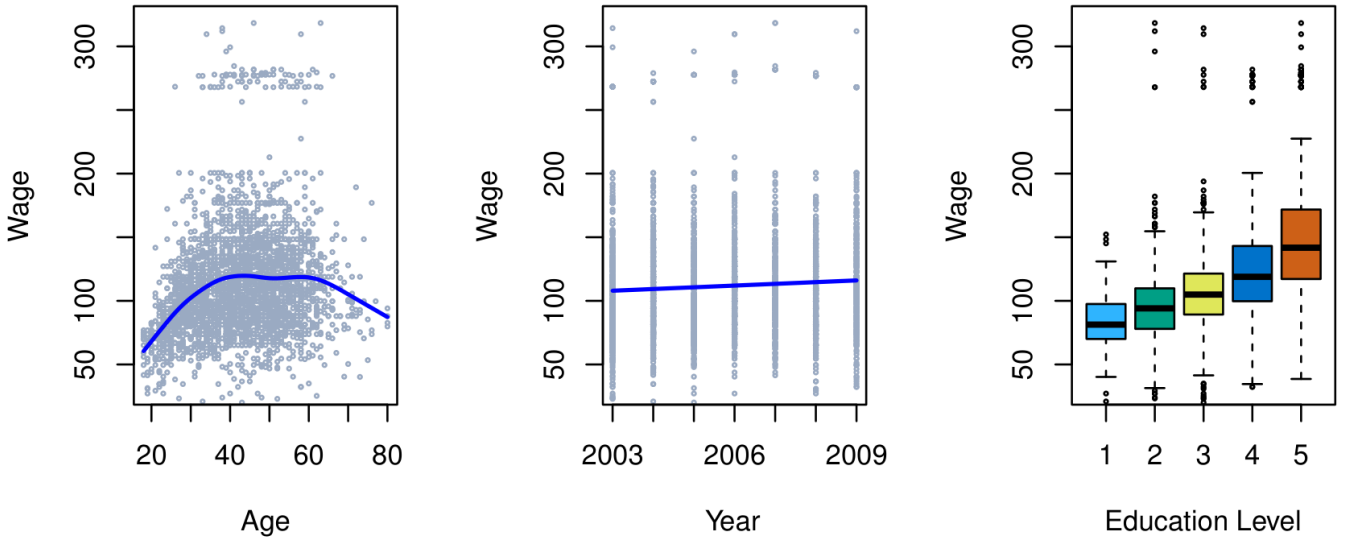
- D. P. Bertsekas, “Introduction to Probability”, （《概率导论》，人民邮电出版社，郑忠国等译，初级）
- S. M. Ross, “A First Course in Probability”, （《概率论基础教程》，机械工业出版社，童行伟等译，初级）

中译未必靠谱，购买前请试读！！

用于理解数据的一套工具集

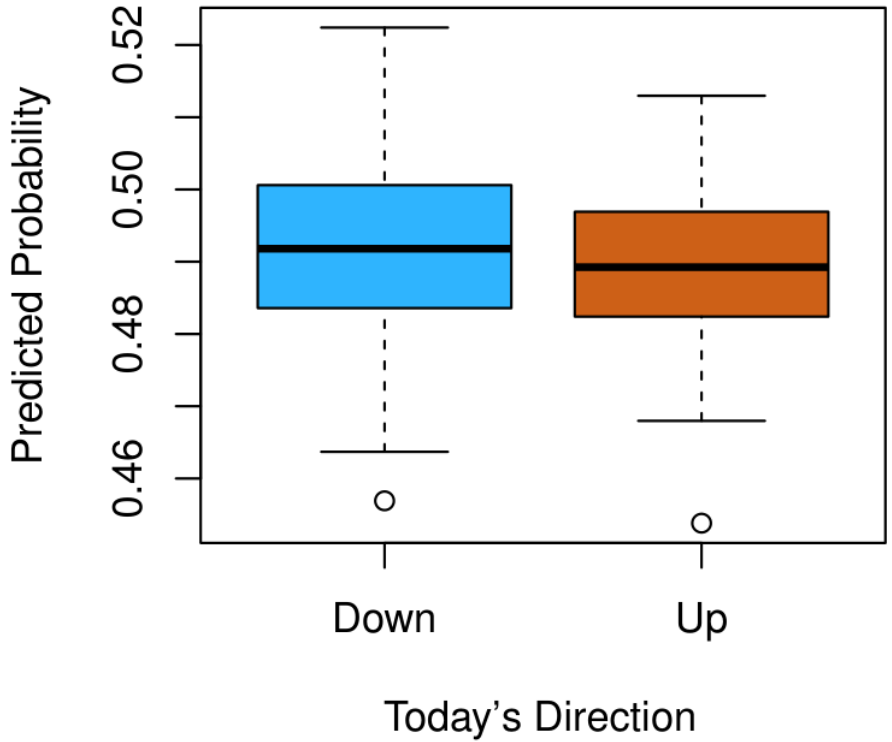
- 监督学习（supervised learning）
  - 构建一个统计模型，该模型基于一个或多个输入进行预测或估计，得到输出
  - 回归（regression）和分类（classification）
- 非监督学习（unsupervised learning）
  - 有输入而无输出
  - 从现存数据中寻找数据内部的关联或结构
  - 聚类（clustering）、关联规则（association rules）、降维（dimension reduction）

实例数据集：Wage



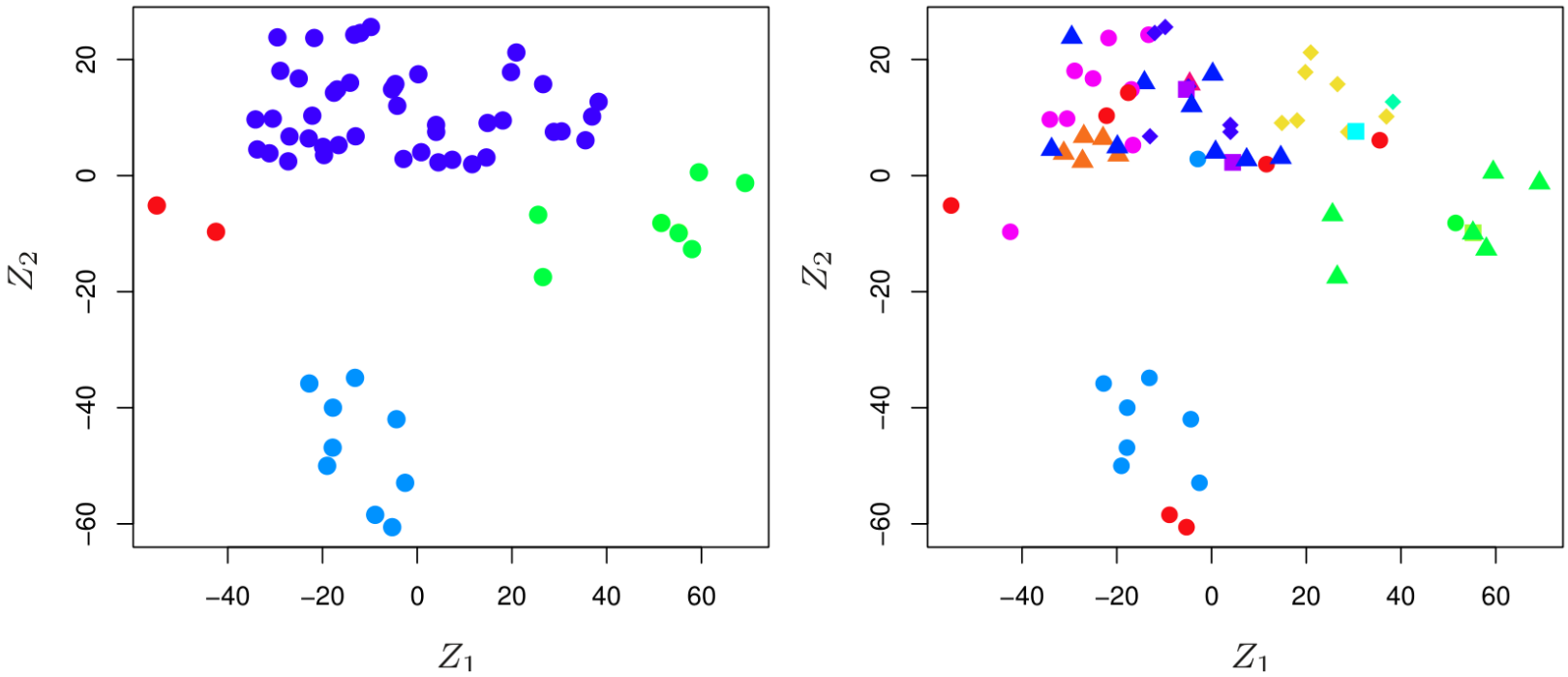
某地区在2003-2009年间，3000个样本（自然人）的工资数据。每个样本包含年份、年龄、教育程度、种族等16个特征。用于本课程中的回归任务。

实例数据集：Smarket



- 2001~2005年标准普尔（Standard & Poor's, S&P）500指数。
- 在本课程中用于构建预测指数上涨/下跌的统计学习模型（分类）。

实例数据集：NCI60



- 64个癌细胞系的6830组基因表达数据
- 用于课程中的无监督学习，上图为降维结果

统计学习历史的简单回顾

“统计学习”（statistical learning）这一名字本身相对新，但涉及到的概念具有较久远的历史

- 19世纪初，最小二乘法➔ 线性回归
- 1936，线性决策分析（linear discriminant analysis）
- 1940s，逻辑回归（logistic regression）
- 1970s，广义线性模型（generalized linear model）
- 1980s，非线性方法，分类和回归树，广义可加模型，神经网络
- 1990s，支持向量机（support vector machines）

本课程的教学策略

- 多数模型广泛用于多个领域（包括材料科学），从最简单、最常用的模型讲起，打好基础
- 学习时，不应将各类模型视作“黑箱”。了解黑箱的内部机制，要对模型提出的背景、思路、所用假设、优缺点有了解
- （暂时）不必掌握各个模型的具体实现细节，需要会用
- 将模型应用于实例（生活、学习、项目等）

本课程使用的符号及惯例

- 通常用n表示数据点（样本）个数
- p表示用于预测的特征个数
- 如Wage数据集中，$n = 3000, p = 11$
- p可能很大，如
  - 基于晶体组成和结构预测性质
  - 基于基因数据预测生物性状
  - 基于用户数据进行广告定点投放
- x_{ij}表示第i个样本的第j个特征。

$$\mathbf{X} = \overbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}}^{p\text{个特征}} \Bigg\} n\text{个样本}$$

本课程使用的符号及惯例

- 若只关注**X**的行，即某个样本的所有特征，可写作：$x_1, x_2, \cdots, x_n$（小写，斜体），此时为一长度为p的矢量，也可写作

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

- 通常用列来表示矢量（列矢量）

本课程使用的符号及惯例

- 若只关注$\mathbf{X}$的列，即所有样本的某个相同特征，可写作：$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$（小写、正体、加粗），此时为一长度为$n$的矢量，也可写作

$$\mathbf{x}_i = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

- 如在Wage中，$\mathbf{x}_1$长度为3000，包含3000个样本（人）的第一个特征
- 矩阵$\mathbf{X}$可以写作 $\mathbf{X} = (\mathbf{x}_1 \ \ \mathbf{x}_2 \ \ \text{或} \mathbf{x}_p)$ $\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$

本课程使用的符号及惯例

- 通常用y_i表示第i个样本的观测结果，也称为target或label，如Wage数据集中的工资（wage）栏

$$\mathbf{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- 则整个数据集可表示为 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

本课程使用的符号及惯例

小写粗体，如 $\mathbf{a}$，表示矢量

大写粗体，如 $\mathbf{A}$，表示矩阵

实数域上的标量： $a \in \mathbb{R}$

实数域上的长度为k的矢量： $\mathbf{a} \in \mathbb{R}^k$

实数域上的维度为$r \times s$的矩阵： $\mathbf{A} \in \mathbb{R}^{r \times s}$

Wage数据集

文本：

```
"year","age","maritl","race","education","region","jobclass","health","health_ins","logwage","wage"
2006,18,"1. Never Married","1. White","1. < HS Grad","2. Middle Atlantic","1. Industrial","1. <=Good","2. No",4.31806333496276,75.0431540173515
2004,24,"1. Never Married","1. White","4. College Grad","2. Middle Atlantic","2. Information","2. >=Very Good","2. No",4.25527250510331,70.4760196469445
2003,45,"2. Married","1. White","3. Some College","2. Middle Atlantic","1. Industrial","1. <=Good","1. Yes",4.8750612633917,130.982177377461
2003,43,"2. Married","3. Asian","4. College Grad","2. Middle Atlantic","2. Information","2. >=Very Good","1. Yes",5.04139268515823,154.68529299563
2005,50,"4. Divorced","1. White","2. HS Grad","2. Middle Atlantic","2. Information","1. <=Good","1. Yes",4.31806333496276,75.0431540173515
2008,54,"2. Married","1. White","4. College Grad","2. Middle Atlantic","2. Information","2. >=Very Good","1. Yes",4.84509804001426,127.115743812184
2009,44,"2. Married","4. Other","3. Some College","2. Middle Atlantic","1. Industrial","2. >=Very Good","1. Yes",5.13302127864665,169.528538036679
2008,30,"1. Never Married","3. Asian","3. Some College","2. Middle Atlantic","2. Information","1. <=Good","1. Yes",4.7160033436348,111.720849360989
2006,41,"1. Never Married","2. Black","3. Some College","2. Middle Atlantic","2. Information","2. >=Very Good","1. Yes",4.77815125038364,118.884359339886
2004,52,"2. Married","1. White","2. HS Grad","2. Middle Atlantic","2. Information","2. >=Very Good","1. Yes",4.85733249643127,128.680488220624
2007,45,"4. Divorced","1. White","3. Some College","2. Middle Atlantic","2. Information","1. <=Good","1. Yes",4.76342799356294,117.146816914805
2007,34,"2. Married","1. White","2. HS Grad","2. Middle Atlantic","1. Industrial","2. >=Very Good","2. No",4.39794000867204,81.2832532842527
2005,35,"1. Never Married","1. White","2. HS Grad","2. Middle Atlantic","2. Information","2. >=Very Good","1. Yes",4.49415459401844,89.4924795180001
```

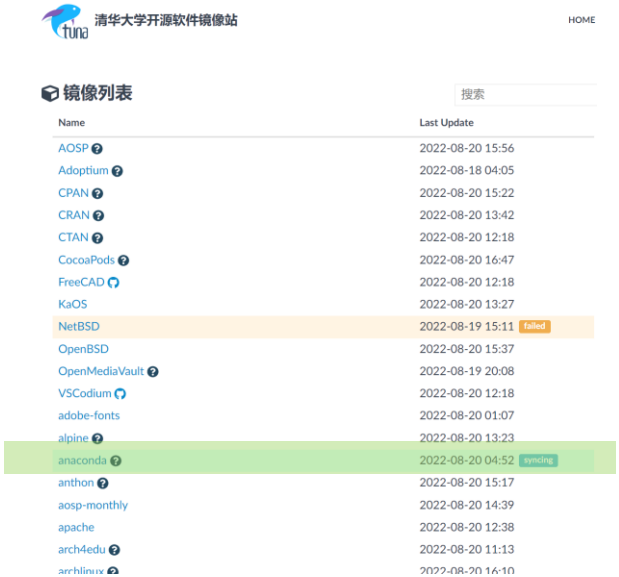
excel：

	A	B	C	D	E	F	G	H	I	J	K
1	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
2	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
3	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Go	2. No	4.255273	70.47602
4	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.9822
5	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Go	1. Yes	5.041393	154.6853
6	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063	75.04315
7	2008	54	2. Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Go	1. Yes	4.845098	127.1157
8	2009	44	2. Married	4. Other	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Go	1. Yes	5.133021	169.5285
9	2008	30	1. Never Married	3. Asian	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.716003	111.7208
10	2006	41	1. Never Married	2. Black	3. Some College	2. Middle Atlantic	2. Information	2. >=Very Go	1. Yes	4.778151	118.8844
11	2004	52	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Go	1. Yes	4.857332	128.6805
12	2007	45	4. Divorced	1. White	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.763428	117.1468
13	2007	34	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Go	2. No	4.39794	81.28325
14	2005	35	1. Never Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Go	1. Yes	4.494155	89.49248

文件格式：csv, comma separated values

学习环境和工具

- python，numpy，pandas，matplotlib，jupyter，……
- 建议使用anaconda或miniconda环境
- 可用教育网镜像（清华、科大、上交、北外……）
- 如：https://mirrors.tuna.tsinghua.edu.cn/

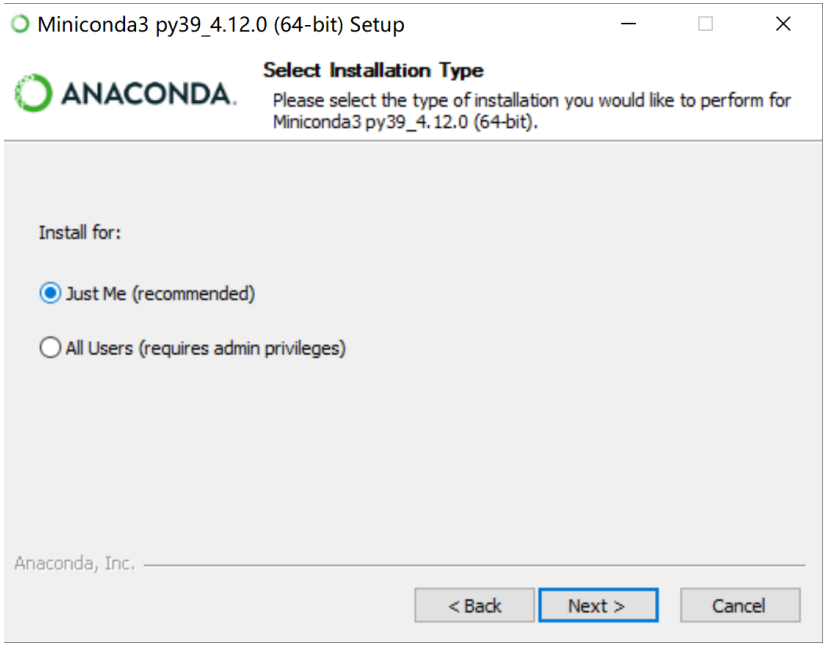


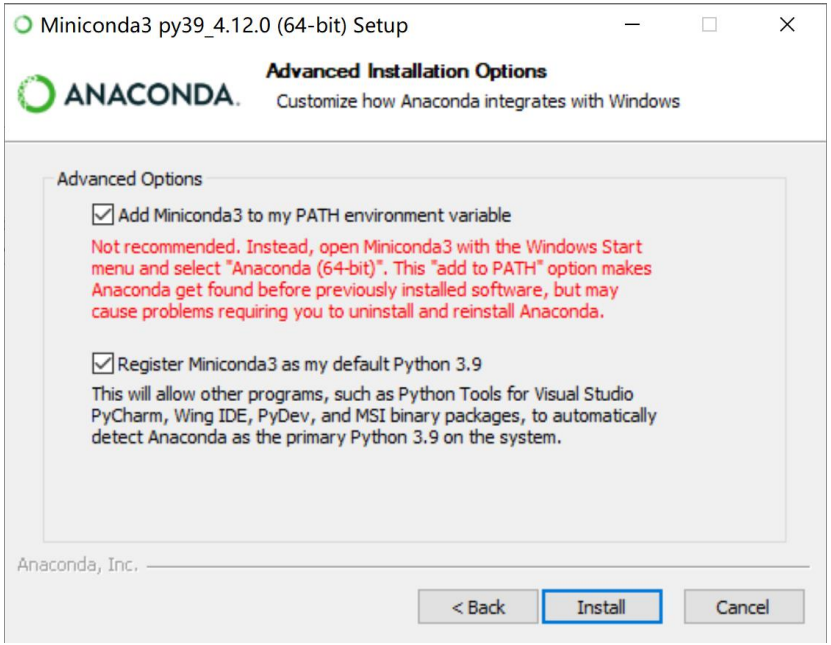
学习环境和工具

Miniconda3-latest-Linux-aarch64.sh	75.3 MiB	2022-05-17 04:01
Miniconda3-latest-Linux-armv7l.sh	29.9 MiB	2017-01-31 01:54
Miniconda3-latest-Linux-ppc64le.sh	74.3 MiB	2022-05-17 04:01
Miniconda3-latest-Linux-s390x.sh	69.2 MiB	2022-05-17 04:01
Miniconda3-latest-Linux-x86.sh	62.7 MiB	2019-01-03 00:11
Miniconda3-latest-Linux-x86_64.sh	73.1 MiB	2022-05-17 04:01
Miniconda3-latest-MacOSX-arm64.pkg	63.5 MiB	2022-06-02 03:47
Miniconda3-latest-MacOSX-arm64.sh	52.2 MiB	2022-06-02 03:47
Miniconda3-latest-MacOSX-x86.sh	26.0 MiB	2017-01-31 01:55
Miniconda3-latest-MacOSX-x86_64.pkg	62.7 MiB	2022-05-17 04:01
Miniconda3-latest-MacOSX-x86_64.sh	56.0 MiB	2022-05-17 04:01
Miniconda3-latest-Windows-x86.exe	67.8 MiB	2022-05-17 04:01
Miniconda3-latest-Windows-x86_64.exe	71.2 MiB	2022-05-17 04:01
Miniconda3-py37_4.10.1-Linux-aarch64.sh	104.5 MiB	2021-06-02 07:46
Miniconda3-py37_4.10.1-Linux-s390x.sh	84.1 MiB	2021-06-02 07:46

- 点击下载最新版miniconda，注意操作系统版本，通常为 Windows-x86_64
- Mac用户需清楚自己所用硬件版本，最新几年的机器通常为MacOSX-arm64
- Linux用户通常技术水平比较高，不再赘述

学习环境和工具：安装miniconda





学习环境和工具

- 安装完成，打开cmd或powershell

```
PS C:\Users\einsl> conda -V
conda 4.12.0
PS C:\Users\einsl>
```

- 修改默认源

```
PS C:\Users\einsl> conda config --set show_channel_urls yes
PS C:\Users\einsl> dir


    目录: C:\Users\einsl


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----         2022/8/6     16:54                .config
d-r---         2022/7/6     14:27                3D Objects
d-r---         2022/7/6     14:27                Contacts
d-r---        2022/7/24      8:55                Desktop
d-r---        2022/8/20     10:47                Documents
d-r---        2022/8/20     17:07                Downloads
d-r---         2022/7/6     14:27                Favorites
d-----         2022/7/7     22:56                Gaussian
d-r---         2022/7/6     14:27                Links
d-----        2022/8/20     17:24                miniconda3
d-r---         2022/7/6     14:27                Music
dar--l         2022/7/6     15:55                OneDrive
d-r---         2022/7/6     14:28                Pictures
d-r---         2022/7/6     14:27                Saved Games
d-r---         2022/7/6     14:28                Searches
d-----         2022/7/7     10:26                Sync
d-r---        2022/7/22     22:26                Videos
d-----        2022/8/19     23:53                Zotero
-a----        2022/8/20     17:31             25 .condarc
```

Anaconda 镜像使用帮助

Anaconda 是一个用于科学计算的 Python 发行版，支持 Linux, Mac, Windows, 包含了众多流行的科学计算、数据分析的 Python 包。

Anaconda 安装包可以到 https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/ 下载。

TUNA 还提供了 Anaconda 仓库与第三方源（conda-forge、msys2、pytorch等，查看完整列表）的镜像，各系统都可以通过修改用户目录下的 `.condarc` 文件。Windows 用户无法直接创建名为 `.condarc` 的文件，可先执行 `conda config --set show_channel_urls yes` 生成该文件之后再修改。

注：由于更新过快难以同步，我们不同步 `pytorch-nightly`，`pytorch-nightly-cpu`，`ignite-nightly` 这三个包。

```
channels:
  - defaults
show_channel_urls: true
default_channels:
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/r
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/msys2
custom_channels:
  conda-forge: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
  msys2: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
  bioconda: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
  menpo: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
  pytorch: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
  pytorch-lts: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
  simpleitk: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
```

即可添加 Anaconda Python 免费仓库。

运行 `conda clean -i` 清除索引缓存，保证用的是镜像站提供的索引。

运行 `conda create -n myenv numpy` 测试一下吧。

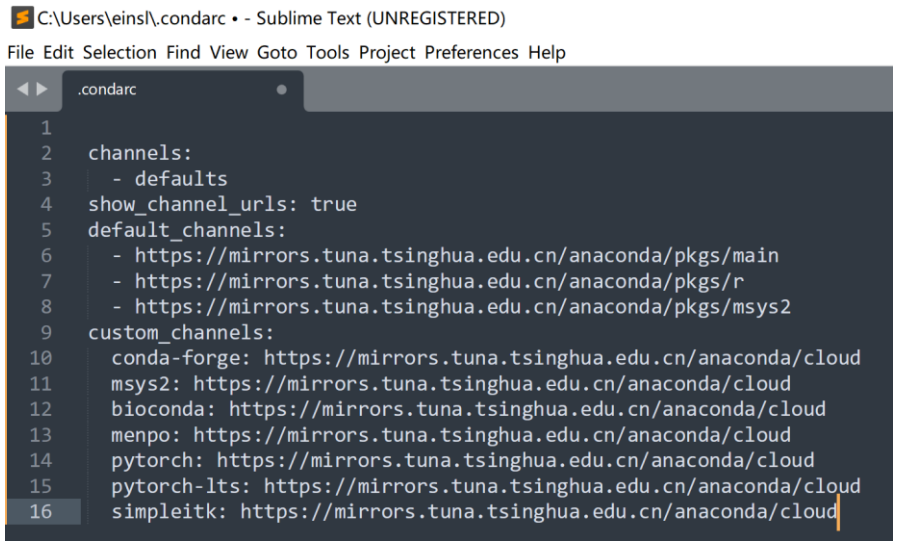
Miniconda 镜像使用帮助

Miniconda 是一个 Anaconda 的轻量级替代，默认只包含了 python 和 conda，但是可以通过 pip 和 conda 来安装所需要的包。

Miniconda 安装包可以到 https://mirrors.tuna.tsinghua.edu.cn/anaconda/miniconda/ 下载。

学习环境和工具


此电脑 › 本地磁盘 (C:) › 用户 › einsl
名称
修改日期
类型
大小
.config 2022/8/6 16:54 文件夹
3D 对象 2022/7/6 14:27 文件夹
Gaussian 2022/7/7 22:56 文件夹
miniconda3 2022/8/20 17:24 文件夹
OneDrive 2022/7/6 15:55 文件夹
Sync 2022/7/7 10:26 文件夹
Zotero 2022/8/19 23:53 文件夹
保存的游戏 2022/7/6 14:27 文件夹
联系人 2022/7/6 14:27 文件夹
链接 2022/7/6 14:27 文件夹
视频 2022/7/22 22:26 文件夹
收藏夹 2022/7/6 14:27 文件夹
搜索 2022/7/6 14:28 文件夹
图片 2022/7/6 14:28 文件夹
文档 2022/8/20 10:47 文件夹
下载 2022/8/20 17:07 文件夹
音乐 2022/7/6 14:27 文件夹
桌面 2022/7/24 8:55 文件夹
.condarc 2022/8/20 17:31 CONDARC 文件 1 KB


C:\Users\einsl\.condarc • - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
.condarc
channels:
- defaults
show_channel_urls: true
default_channels:
- https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main
- https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/r
- https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/msys2
custom_channels:
conda-forge: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
msys2: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
bioconda: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
menpo: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
pytorch: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
pytorch-lts: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud
simpleitk: https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud

学习环境和工具

```
PS C:\Users\einsl> conda create -n matgen python=3.10
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: C:\Users\einsl\miniconda3\envs\matgen

  added / updated specs:
    - python=3.10


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    bzip2-1.0.8                |       he774522_0         113 KB  defaults
    ca-certificates-2022.07.19 |       haa95532_0         123 KB  defaults
    certifi-2022.6.15          |  py310haa95532_0         153 KB  defaults
    libffi-3.4.2               |       hd77b12b_4         107 KB  defaults
    openssl-1.1.1q             |       h2bbff1b_0         4.8 MB  defaults
    pip-22.1.2                 |  py310haa95532_0         2.5 MB  defaults
    python-3.10.4              |       hbb2ffb3_0        15.9 MB  defaults
    setuptools-61.2.0          |  py310haa95532_0         1.0 MB  defaults
    sqlite-3.39.2              |       h2bbff1b_0         805 KB  defaults
    tk-8.6.12                  |       h2bbff1b_0         3.1 MB  defaults
    tzdata-2022a               |       hda174b7_0         109 KB  defaults
    vc-14.2                    |       h21ff451_1           8 KB  defaults
    vs2015_runtime-14.27.29016 |       h5e58377_2        1007 KB  defaults
    wheel-0.37.1               |     pyhd3eb1b0_0          33 KB  defaults
    wincertstore-0.2           |  py310haa95532_2          15 KB  defaults
    xz-5.2.5                   |       h8cc25b3_1         246 KB  defaults
    zlib-1.2.12                |       h8cc25b3_2         116 KB  defaults
    ------------------------------------------------------------
                                           Total:        30.2 MB

The following NEW packages will be INSTALLED:

  bzip2              anaconda/pkgs/main/win-64::bzip2-1.0.8-he774522_0
  ca-certificates    anaconda/pkgs/main/win-64::ca-certificates-2022.07.19-haa95532_0
  certifi            anaconda/pkgs/main/win-64::certifi-2022.6.15-py310haa95532_0
  libffi             anaconda/pkgs/main/win-64::libffi-3.4.2-hd77b12b_4
  openssl            anaconda/pkgs/main/win-64::openssl-1.1.1q-h2bbff1b_0
  pip                anaconda/pkgs/main/win-64::pip-22.1.2-py310haa95532_0
  python             anaconda/pkgs/main/win-64::python-3.10.4-hbb2ffb3_0
  setuptools         anaconda/pkgs/main/win-64::setuptools-61.2.0-py310haa95532_0
  sqlite             anaconda/pkgs/main/win-64::sqlite-3.39.2-h2bbff1b_0
  tk                 anaconda/pkgs/main/win-64::tk-8.6.12-h2bbff1b_0
  tzdata             anaconda/pkgs/main/noarch::tzdata-2022a-hda174b7_0
  vc                 anaconda/pkgs/main/win-64::vc-14.2-h21ff451_1
  vs2015_runtime     anaconda/pkgs/main/win-64::vs2015_runtime-14.27.29016-h5e58377_2
  wheel              anaconda/pkgs/main/noarch::wheel-0.37.1-pyhd3eb1b0_0
  wincertstore       anaconda/pkgs/main/win-64::wincertstore-0.2-py310haa95532_2
  xz                 anaconda/pkgs/main/win-64::xz-5.2.5-h8cc25b3_1
  zlib               anaconda/pkgs/main/win-64::zlib-1.2.12-h8cc25b3_2


Proceed ([y]/n)?
```

```
    wincertstore-0.2           | py310haa95532_2          15 KB  defaults
done
#
# To activate this environment, use
#
#     $ conda activate matgen
#
# To deactivate an active environment, use
#
#     $ conda deactivate

PS C:\Users\einsl> conda activate matgen
PS C:\Users\einsl>
```

安装 mamba 加速安装...

```
conda install mamba -n matgen -c conda-forge
```

学习环境和工具

```
mamba install ipython numpy matplotlib scipy jupyter pandas
```

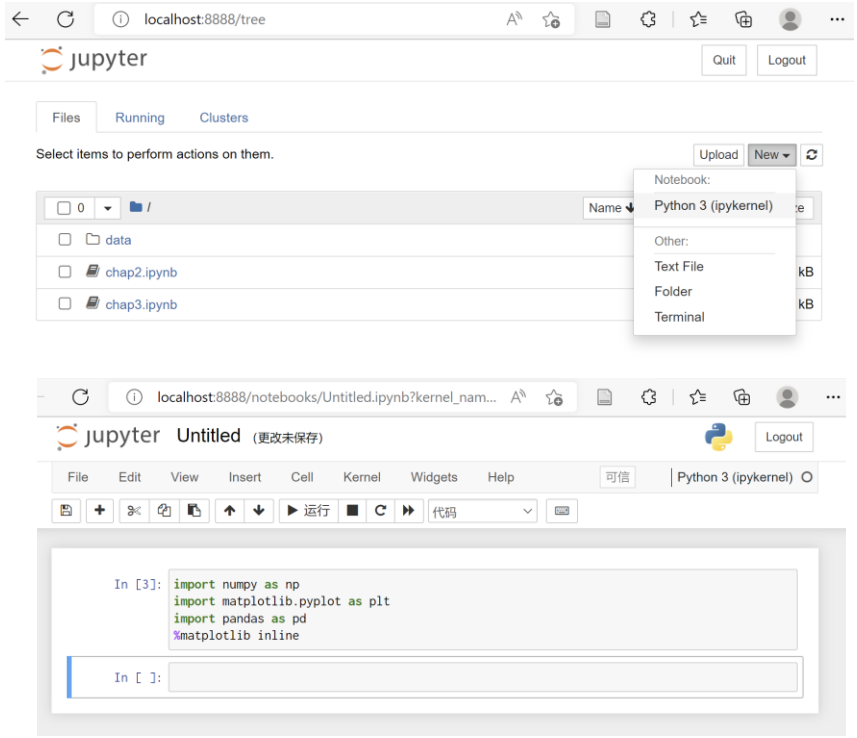
学习环境和工具

ipython：一个更好用的 python shell

```
(matgen) C:\Users\einsl>where ipython
C:\Users\einsl\miniconda3\envs\matgen\Scripts\ipython.exe

(matgen) C:\Users\einsl>ipython
Python 3.10.4 | packaged by conda-forge | (main, Mar 30 2022, 08:38:02) [MSC v.1916 64
bit (AMD64)]
Type 'copyright', 'credits' or 'license' for more information
IPython 8.4.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: import numpy as np

In [2]: import pandas as pd

In [3]: np.__version__
Out[3]: '1.23.1'

In [4]: pd.__version__
Out[4]: '1.4.3'

In [5]:
```

学习环境和工具

```
(matgen) C:\Users\einsl>cd C:\current\ISLRv2

(matgen) C:\current\ISLRv2>dir
 驱动器 C 中的卷没有标签。
 卷的序列号是 10EA-3026

 C:\current\ISLRv2 的目录

2022/08/14  17:05    <DIR>          .
2022/08/14  17:05    <DIR>          ..
2022/07/06  20:09    <DIR>          .ipynb_checkpoints
2022/08/14  17:05           399,369 chap2.ipynb
2021/11/03  19:34           178,867 chap3.ipynb
2022/07/06  20:09    <DIR>          data
               2 个文件        578,236 字节
               4 个目录 707,671,687,168 可用字节

(matgen) C:\current\ISLRv2>jupyter notebook --no-browser
[I 2022-08-20 18:02:06.271 LabApp] JupyterLab extension loaded from C:\Users\einsl\mini
conda3\envs\matgen\lib\site-packages\jupyterlab
[I 2022-08-20 18:02:06.271 LabApp] JupyterLab application directory is C:\Users\einsl\m
iniconda3\envs\matgen\share\jupyter\lab
[I 18:02:06.277 NotebookApp] Serving notebooks from local directory: C:\current\ISLRv2
[I 18:02:06.277 NotebookApp] Jupyter Notebook 6.4.12 is running at:
[I 18:02:06.278 NotebookApp] http://localhost:8888/?token=459482c314334f6979392c0693368
ae0aefbfc16734127f6
[I 18:02:06.278 NotebookApp]  or http://127.0.0.1:8888/?token=459482c314334f6979392c069
3368ae0aefbfc16734127f6
[I 18:02:06.278 NotebookApp] Use Control-C to stop this server and shut down all kernel
s (twice to skip confirmation).
[C 18:02:06.282 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///C:/Users/einsl/AppData/Roaming/jupyter/runtime/nbserver-11292-open.html
    Or copy and paste one of these URLs:
        http://localhost:8888/?token=459482c314334f6979392c0693368ae0aefbfc16734127f6
     or http://127.0.0.1:8888/?token=459482c314334f6979392c0693368ae0aefbfc16734127f6
```



建议设置浏览器，代码使用等宽字体显示

学习环境和工具

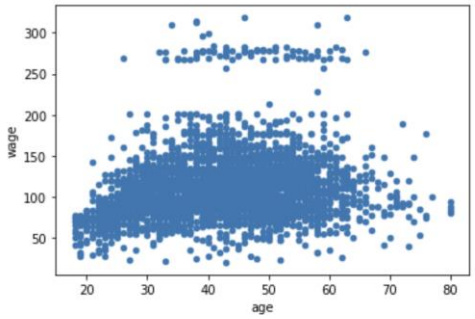
```
In [4]: wage = pd.read_csv('data/all_data/Wage.csv')

In [5]: wage
```

Out[5]:

	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	
0	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	7
1	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	7
2	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	13
3	2003	43	2. Married	3.	4. College	2. Middle	2.	2. >=Very	1. Yes	5.041393	15

```
In [8]: wage.plot.scatter('age','wage')
```

Out[8]: <AxesSubplot:xlabel='age', ylabel='wage'>



作业：

设置好python环境，并使用pandas重现第8页Wage数据图。

提交方式：

jupter notebook或pdf，包含代码及运行结果。

坚果云收作业（群内发链接）。

请于11月13日24点前提交，过期、抄袭无效。