一、预测任务与目标:

在给定时间内，使用给予的训练集中的因子和 label，构造预测模型（树，神经网络，LR 等任意模型）。预测效果的评判标准为在样本外测试集（测试集不给予笔试者）中，**预测值与短周期标签的相关系数**，相关系数越大，预测效果越好。

相关系数的公式为

$$r(pred, label) = \frac{\mathrm{Cov}(pred, label)}{\sqrt{\mathrm{Var}(pred)\mathrm{Var}(label)}}$$

二、数据集

1. 训练数据集为 data.npy 文件，包含所有样本信息，形状为(3355055, 204)。数据集第一维度为样本，第二维度为因子与标签，因此意味着该数据集包含 3355055 条样本，列数中第一列为脱敏后 的时间，按顺序从 0-24 代表时间先后顺序；第 2-201 列为提供的因子数据；第 202 列为短周期 标签，第 203 列为中周期标签，最后一列为长周期标签。

2. 测试数据集为 test.npy 文件,形状为(424606,201),没有最后三列标签信息。提交需要给出，可选择三种的任一种给出。 模型最后的评判标准是对短周期标签的预测力。

三、提交 笔试需要应试者至少提供两类文件：

1. 文字文件：简要说明自己模型构造的思路，并展示最终模型预测效果的文件，字数无要求，讲清 楚即可，包括 CV 的划分，特征的处理分析，一些其他发现等。

2. 模型文件+代码：模型文件推荐提供 pth 或 pkl 格式的文件，其他文件也可，只要能跑通。代码 文件为示例使用文件，即如何使用你的模型对数据进行预测的示例代码，方便我们评估更多样本外 数据的表现，请确保该文件能跑通。

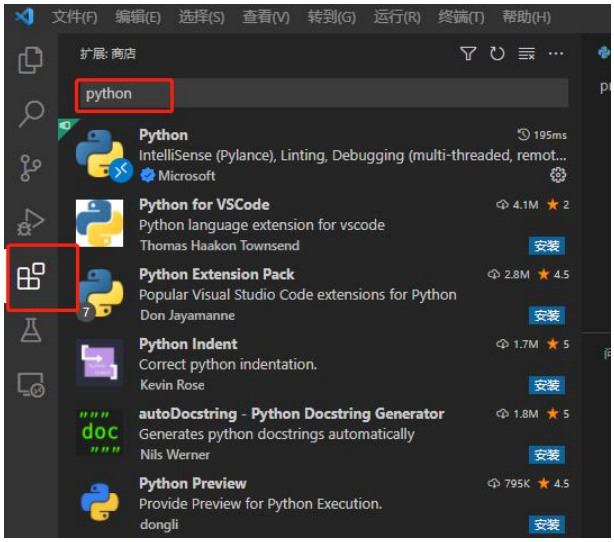
3. 提交 prediction.npy，行数为 test 的行数，列数可以 1~3 个周期的标签。

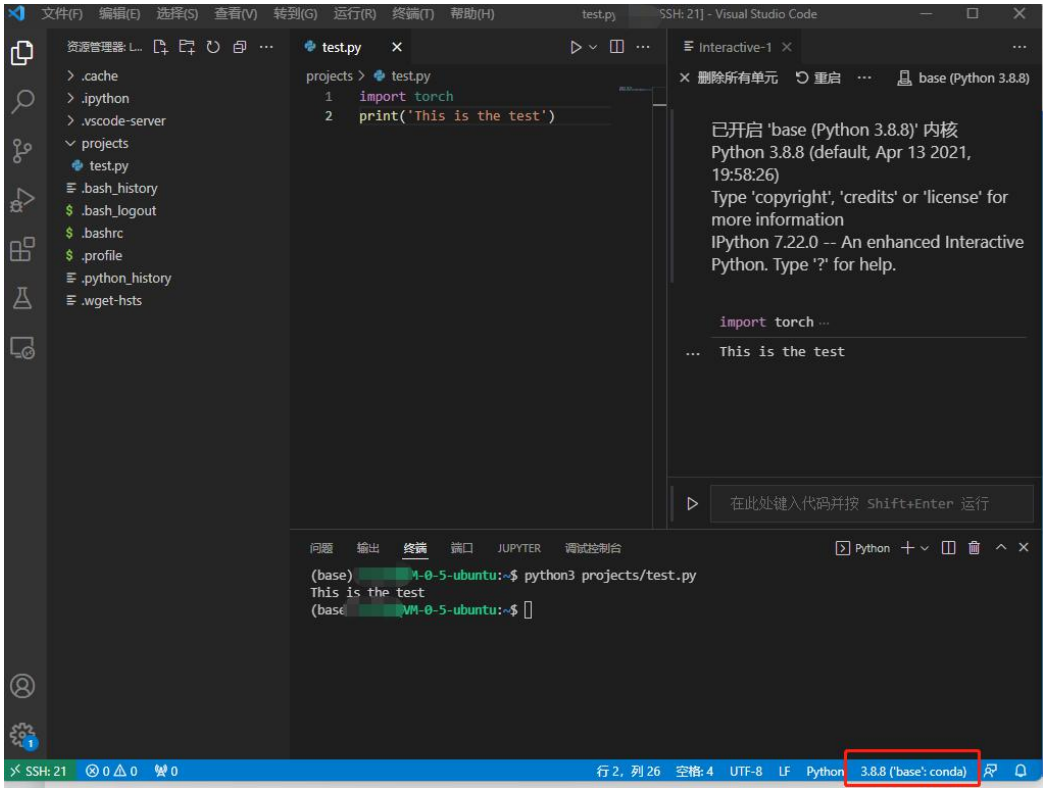
四 . 服务器使用

1. 提供训练云算力，请使用给定的主机 IP，账号，密码登录，勿私自修改密码。可以使用 xshell, pycharm,vscode 等,

2. 如果不了解如何使用服务器，一点使用 vscode 的小建议:

- 可以参考 https://blog.csdn.net/zhaxun/article/details/120568402 来配置远程连接服务器,注意配置 VScode 时有个选择 remote host platform 时用第一个，Linux。其他的基本也是选择第一个选项向后走。
- 新建自己的项目，python 文件等。
- 安装 Python 扩展或者还有其他的扩展，见下图:





- 之后就可以正常使用 shell 或者 jupyter 来调试服务器上的代码，别忘了及时保存。解释器可以像上图红色圈出部分，使用 /usr/local/anaconda3/bin/python

3. 性能：CPU 40 core,GPU Tesla T4*2,160G 内存，100G 磁盘。注意资源会被共享，可根据 GPU 使用状态选定 device 0 或 1. 如算力不够，请及时告知调整。

4. 如果想要安装不同版本的包，运行下面代码，请在报告中明确：

source /usr/share/virtualenvwrapper/virtualenvwrapper.sh

mkvirtualenv *yourProjectName*

workon

这样就是相应的虚拟环境

5. 已经预装 torch 等常用包，其他需要包的可以参照以下安装：

pip install torch==1.7.1 -i https://pypi.tuna.tsinghua.edu.cn/simple/

6. 所用到的数据存储位置，只读:

- /wydata/data.npy

-/wydata/test.npy

读取时 np.load('/wydata/data.npy')即可。

7. 其他个人产生的所有文件和数据应该在当前目录下（pwd 查看位置）

五、特别提醒

1，最后仅仅考察预测值与短周期标签（即标签的第一列）的相关系数，另外两个标签笔试者可以选择不用，也可以选择辅助训练。

2，笔试者最后的得分是基于样本外的数据集，请笔试者小心使用数据，避免过拟合。

3，该试题严禁泄露给第三方，如有发现，将追究法律责任。

4，如果放弃本次笔试，请及时联系 HR 告知原因，我们将回收资源以防浪费。