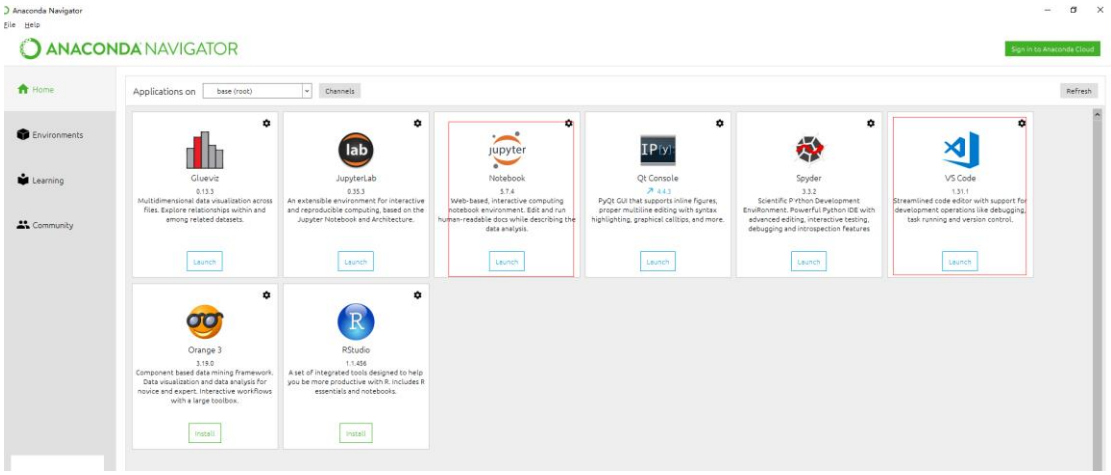
一、实验环境

1、软件环境：Anaconda3（内置 Python3.7 ）

建议使用该软件中的 vscode 工具或者 Jupyter Notebook

安装后的环境如下：



建议使用图中红框内的软件。

说明：**vscode** 工具具有调试功能和代码提示

Jupyter Notebook 是一款网页版的开发工具，占用内存少，无调试功能和代码提示

2、所需工具包

sklearn 机器学习工具包，**Anaconda3** 已经集成。若需下载以及学习 **sklearn** 中 **API** 相关的内容，网址为：https://scikit-learn.org/

3、所用数据集

数据集在 **sklearn.datasets** 库中，本实验所用的 fetch_california_housing 和 fetch_20newsgroups_vectorized 数据集，在导入后程序会自动下载到计算机中。

使用方法：

```
from sklearn.datasets import fetch_california_housing
```

4、评价指标

评价指标在 **sklearn.metrics** 中，本实验所用的评价指标如表 **1** 所示。

表 **1** 与评级指标相关的函数

函数名	说明
mean_squared_error,	MSE（Mean Squared Error）均方误差，对应 **ppt** 第 **13** 页
mean_absolute_error	MAE（Mean Absolute Error）平均绝对误差
r2_score	R^2 分数

公式参考：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$\mathbf{MAE} = \frac{1}{\mathbf{m}}\sum_{\mathbf{i=1}}^{\mathbf{m}}(\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i})$$

其中，m 是样本数，y_i 是真实值，$\hat{y}_i$是预测值。

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y_i})^2}$$

其中，m 是样本数，yi 是真实值，$\hat{y}_i$是预测值，$\overline{y_i}$是真实值的平均值。

二、实验中用到的函数

本实验所用的函数均来自线性模型，在 sklearn.linear_model 库中。

表 2 实验中用到的函数

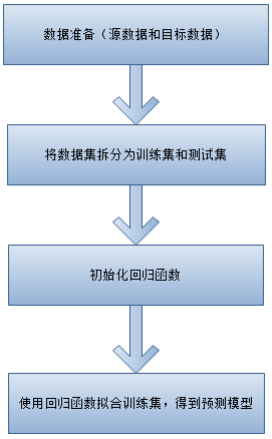
函数名	说明
LinearRegression	线性回归
Ridge	岭回归
Lasso	套索回归
LogisticRegression	逻辑回归
ElasticNet	弹性网回归

所有函数的说明都可以参考 https://scikit-learn.org/中所提供的 API

三、应用举例

1、建立预测模型

模型建立步骤分为数据准备、拆分数据集、初始化回归函数、使用回归函数拟合训练集，最终得到预测模型。



得到预测模型后，通过该模型能得到线性模型中的直线系数以及截距，从而实现线性模型的方程。

2、评测预测模型

使用 sklearn.metrics 中的评价指标评测模型。

3、实例

```
#导入线性回归模型
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.datasets import make_regression
from sklearn.metrics import mean_squared_error
#1.数据准备，获取源数据、目标数据
#生成有 1 个特征和 50 个样本的数据
data,target= make_regression(n_features=1,n_samples=50)
#2、将数据集拆分为训练集和测试集，测试集的比例为 20%
x_train,x_test,y_train,y_test=train_test_split(data,target,test_size=0.2,rand
```

```
om_state=1)
#3、初始化 LinearRegression 函数
lr=LinearRegression()
#4、使用 lr 拟合训练集，得到预测模型
predict_model=lr.fit(x_train,y_train)

#从预测模型中获取直线系数和截距
coef=predict_model.coef_                    #直线系数
intercept=predict_model.intercept_          #截距
print(coef,intercept)
print("coef={},intercept={}".format(coef,intercept))
print("y={}x+{}".format(coef,intercept))     #通过模型得到的直线方程
#使用 MSE 评测模型
#获取预测值
y_predict=predict_model.predict(x_test)
mse=mean_squared_error(y_test,y_predict)
print("mse={}".format(mse))
```

四、实验内容：

1、线性回归

所用数据集：fetch_california_housing

实验内容：分别使用线性回归、Ridge 回归、LASSO 回归、ElasticNet 回归预测房价

实验步骤：

（1）使用 fetch_california_housing 的全部特征来预测房价，分别使用 R^2、MSE（Mean Squared Error）、MAE(Mean Absolute Error)来评估实验结果，并写出预测模型

（2）对 LASSO 回归的参数调优（使用 GridSearchCV）

2、Logisitic 回归

所用数据集：fetch_20newsgroups_vectorized

实验内容：使用 Logistic 回归（LogisticRegression）对新闻分类