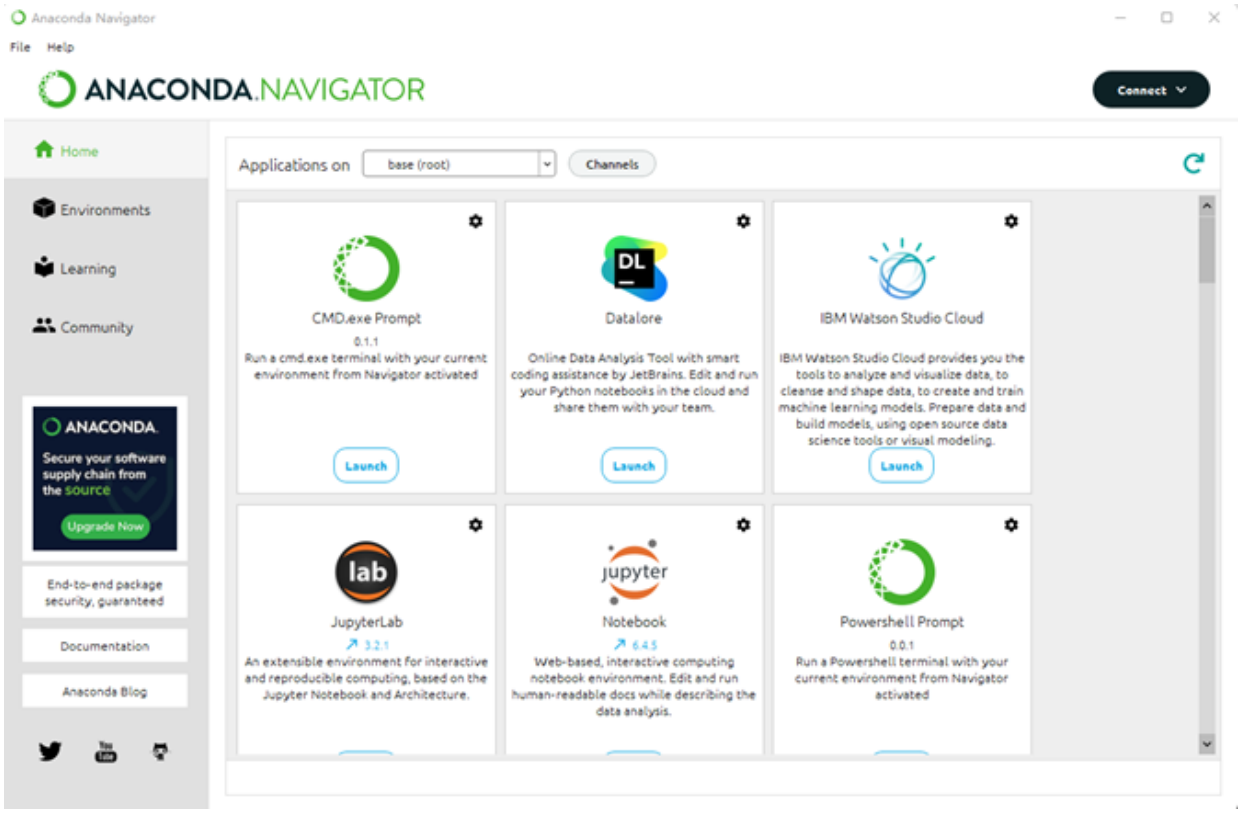
sklearn简介&数据集加载

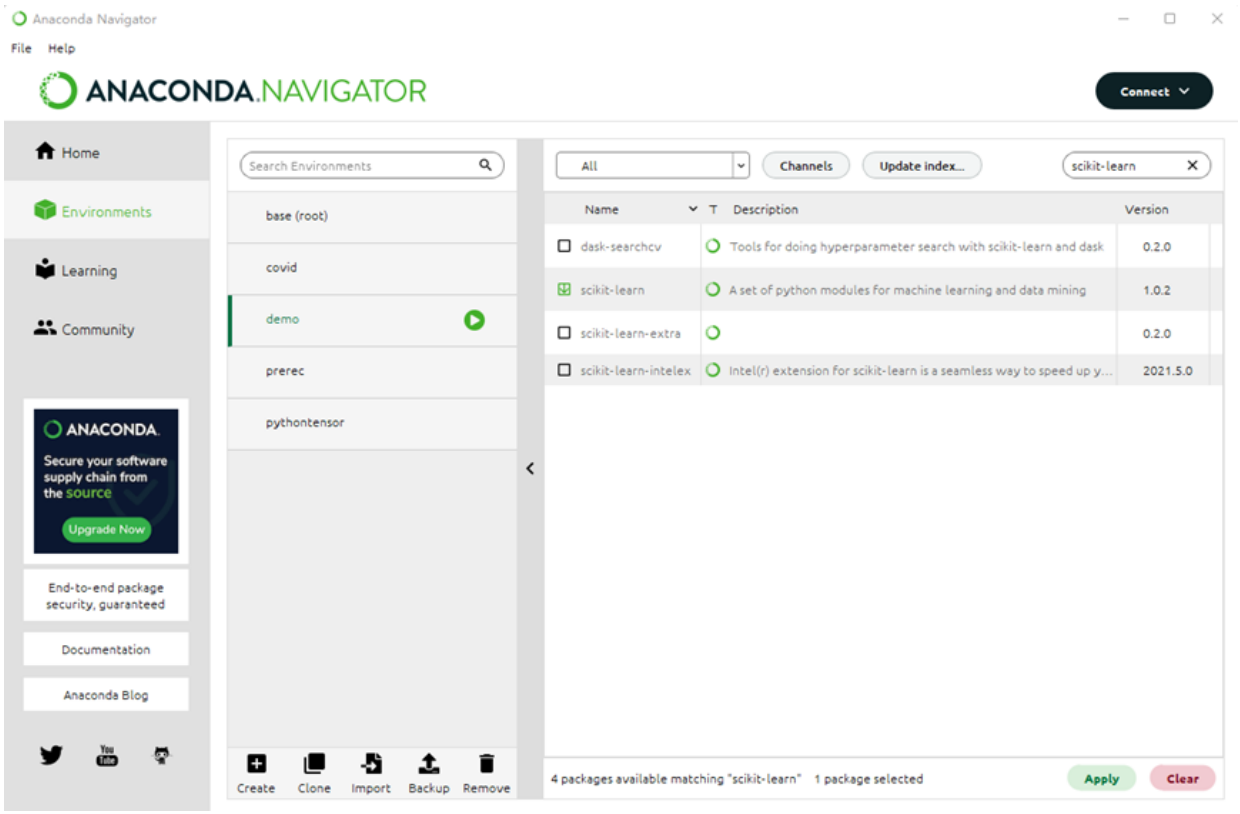
该课程主要为大家讲授如下的内容：

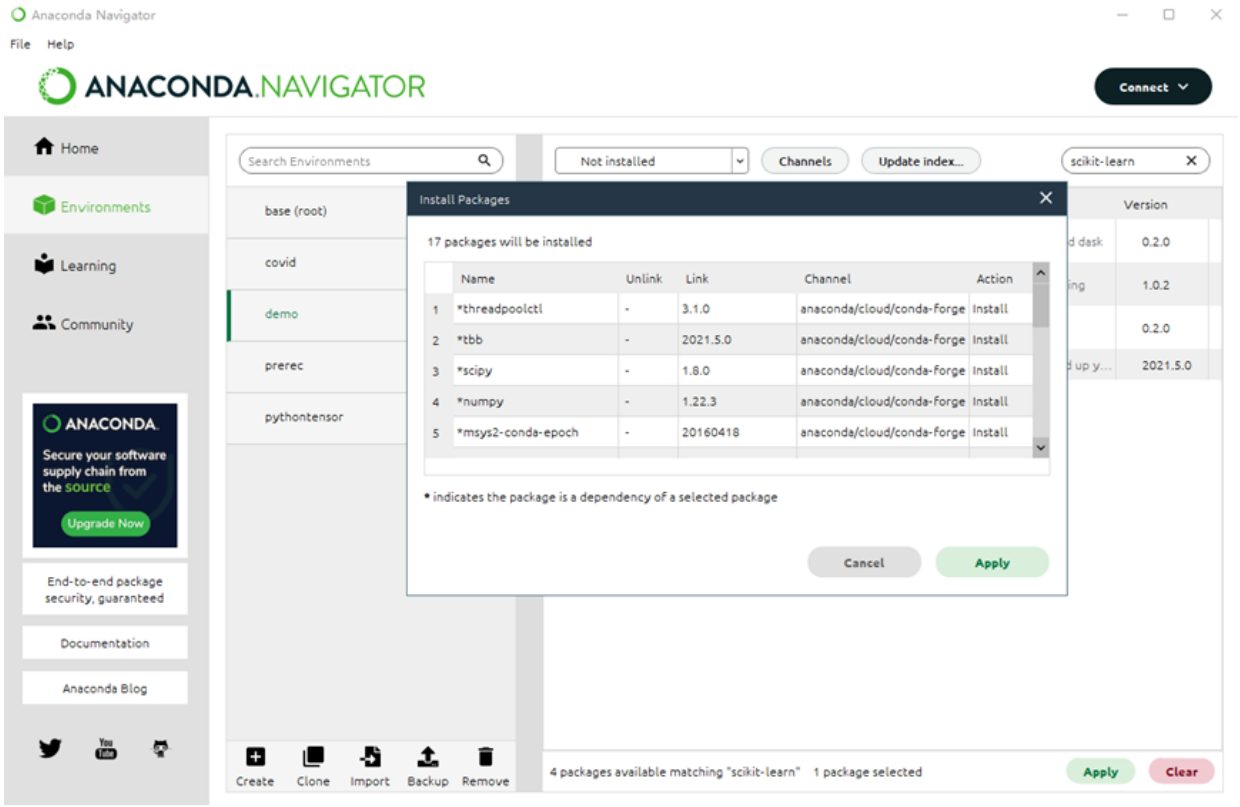
- sklearn简介
- sklearn的安装
- 数据集加载

1. Sklearn简介
   Sklearn，全称scikit_learn，是目前python上最流行的机器学习工具包，它提供了多种机器学习算法的实现，并且它的API简洁统一，很容易上手使用。但是sklearn并不支持在GPU上进行加速运算，所以更适合中小型、特别是数据量不大的项目。

   Sklearn的安装：
   通过命令行pip install –U scikit-learn安装sklearn。
   通过anaconda安装sklearn







2. 数据集加载
    1. 从sklearn内置的的数据集

       用leaders加载比较小的数据集；用fetch从网络上下载的比较大的数据。注意sklearn是基于NumPy开发的，所以我们加载进来参与计算的数据，是numpy数组格式。

       下面来详细讲解fetch方法的使用，首先是fetch的代码样式（以fetch_olivetti_faces为例）：

`data=fetch_olivetti_faces(data_home=None, shuffle=False, random_state=0)`

该方法中需要注意的参数有（以fetch_olivetti_faces为例）：

data_home：为数据集指定一个下载和缓存文件夹。

Shuffle：如果为 True，则对数据集的顺序进行随机排序，以避免对同一人的图像进行分组。

random_state：确定数据集随机排列的随机数生成

`Subset: {‘train’, ‘test’, ‘all‘}, default=’train’` 。选择要加载的数据集：“训练”用于训练集，“test”表示测试集，“all”表示两者，并带有随机排序。

2. 样本生成器

单标签生成器：随机生成散点以进行分类。

下面来详细讲解单标签生成器make_blobs方法的使用，首先是代码样式：

```
X, y = make_blobs(n_samples=50, centers=4,random_state=42)
```

该方法中需要注意的参数有：

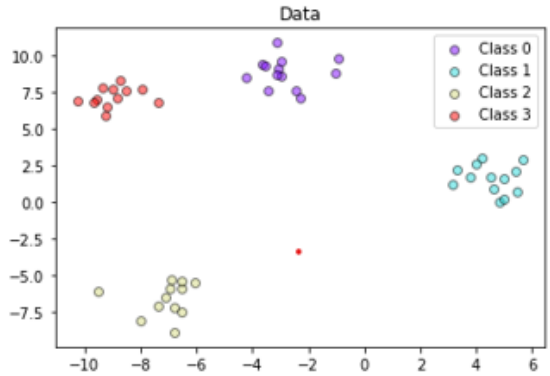
n_samples表示生成的样本个数

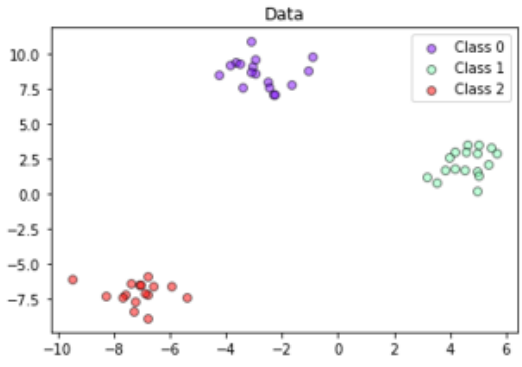
Centers表示生成样本中心数，也就是这些样本在几个分类里生成

random_state随机数种子，用于复现生成结果

返回的是特征组和变迁组

下面这两个图分别是4个样本中心数和3个样本中心数的图像





分类生成器：相对单标签生成器更复杂的生成器

下面来详细讲解分类生成器make_classification方法的使用，首先是代码样式：

```
X,y=make_classification(n_samples=100, n_features=20 , n_informative=2 , n_classes=2, random_state=0)
```

该方法中需要注意的参数有：

n_samples：样本个数。

n_features：特征总数。包括信息特征，冗余特征，重复特征和随机绘制的无用特征。
n_informative=2：信息特征数（和n_features 参数绑定出现）
n_redundant：冗余特征数。这些特征作为信息特征的随机线性组合生成。
n_repeated：重复特征数，从信息特征和冗余特征中随机抽取。
n_classes：分类问题的类（或标签）数。
random_state：随机数种子

线性回归生成器：针对线性回归设计的随机生成器
代码样式为：

```
X,y=make_regression(n_samples=100,n_features=100,n_informative=10, n_targets=1, random_state=None)
```

该方法中需要注意的参数有：
n_samples：样本个数。
n_features：特征总数。
n_informative=2：信息特征数
n_targets：回归目标的数量，即与样本关联的 y 输出向量的维度。
random_state：确定数据集随机排列的随机数生成

3. 从外部加载数据集
Sklearn提供一些简单的数据加载方式，用于加载固定格式的数据，包括简单的二维RGB图片，以及从openml.org网站上下载数据集。而我们更多使用的是通过其他的工具包从外部将数据加载成numpy格式。比如说常见的csv、json文件我们通过pandas载入，二进制文件通过Scipy载入，表格式文件通过numpy载入，等等。然而，这些工具包载入后的格式可能并不是numpy数组，比如pandas会加载成pandas.dataframe格式，需要进一步将他们转换成numpy数组。比如dataframe格式的文件需要通过values属性就可以提取出narray格式的数组。其他格式的转换遇到时可以自行搜索如何转换成numpy数组的格式。