3.5 作业

2023 年 9 月 18 日

作业-Numpy

运用 Numpy1 维、2 维数组相关的创建、切片、运算等方法解决以下问题：

1. 一维数组（0.5 分 *4 = 2 分）

- 创建一个 1 维数组，命名为 data01，包含 100，99，…，1，间隔为 1
- 对其进行从小到大排序
- 将数组 data01 变形成 4 行 25 列的 2 维数组
- 计算总和

[ ]:

2. 二维数组（0.5 分 *6=3 分）

- 创建一个大小为 3 行 4 列的 2 维随机数组，数组中的元素维随机数，命名为 data02
- 选取数组 data02 后 3 列，也就是第 2 列到第 4 列元素
- 求数组 data02 的元素中的最大值、最小值和总和。
- 求数组 data02 的元素中的每一列求最大值、最小值和总和。
- 将 data02 变形成 1 维数组。
- 将 data02 变成 4 行 3 列的 2 维数组

[ ]:

3. 最小最大标准化 (1 分)

最小最大标准化（Min-Max Normalization）是一种常用的数据标准化方法，用于将数据缩放到指定的范围内。该方法通过对数据进行线性变换，使其值域被缩放到指定的范围内，通常是 [0,1] 或 [-1,1]。

该方法的公式如下：

$$X_{norm} = (X - X_{min})/(X_{max} - X_{min})$$

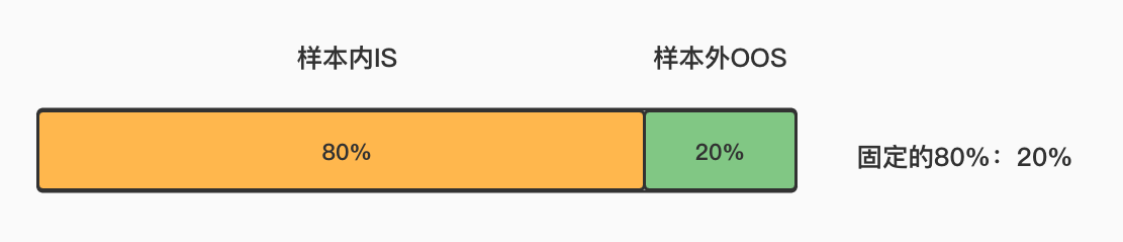
通过最小最大标准化，可以使得不同量纲的数据在同一标度下进行比较，且可以使得数据具有可比性。在机器学习、数据挖掘等领域中，常常需要对数据进行标准化，以提高算法的性能和效果。

应用上述方法，将序列 [16, 46, 33, 90, 64, 34, 91, 33, 24, 32, 30, 91, 85, 58, 77, 35, 1,34, 51, 36] 进行最小最大标准化操作，放缩到 [0,1]。

[ ]:

4. 数据集划分（2 分 *2=4 分）

在机器学习中，样本内和样本外是指数据集中的一部分样本，其中来说，样本内是指用于训练模型、调整模型参数、验证模型性能的数据；样本外是指未在训练过程中使用的数据，通常用于测试模型的性能和泛化能力。



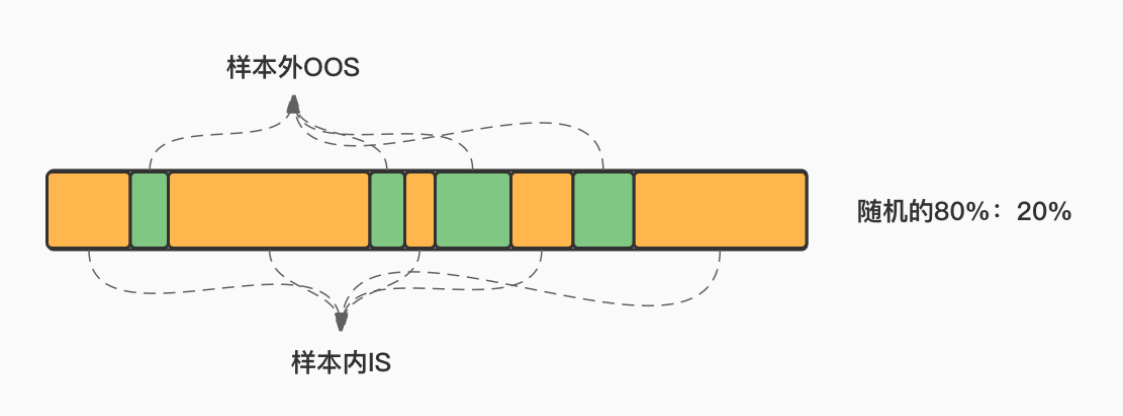
一般将 80% 的原始数据集的子样本集作为样本内 (In-sample, IS)，剩余 20% 作为样本外 (Out-of-sample, OOS)。样本总数为 500 个，一种简单的切割方法是直接指定前 400 个样本为 IS，后面为 OOS。

[48]:
```
X = np.array([0.36707772, 0.30442107, 0.09602864, 0.72140911, 0.30727176,
       0.30332833, 0.82704939, 0.0547752 , 0.91921485, 0.54352647,
       0.45263199, 0.1654269 , 0.42571673, 0.38253443, 0.11461271,
       0.56558215, 0.78277293, 0.82637154, 0.48652071, 0.49557472])
```

如果给定一个 X 序列，那么其样本内和样本外分别是：

[ ]:
```
X_IS =
X_OOS =
```

简单随机划分（Simple Random Sampling）：将原始数据集随机划分为训练集和测试集两部分，通常将数据集的 70% 到 80% 作为训练集，剩余部分作为测试集。



当我们并不了解原始数据集是不是被刻意排列了，最好的办法是使用随机抽样，即随机抽 80% 为 IS，剩余的 20% 为 OOS。

[ ]:
```
X_random_IS =
X_random_OOS =
```