20220525-机器学习

1.学习内容

1.1 机器学习

周志华西瓜书

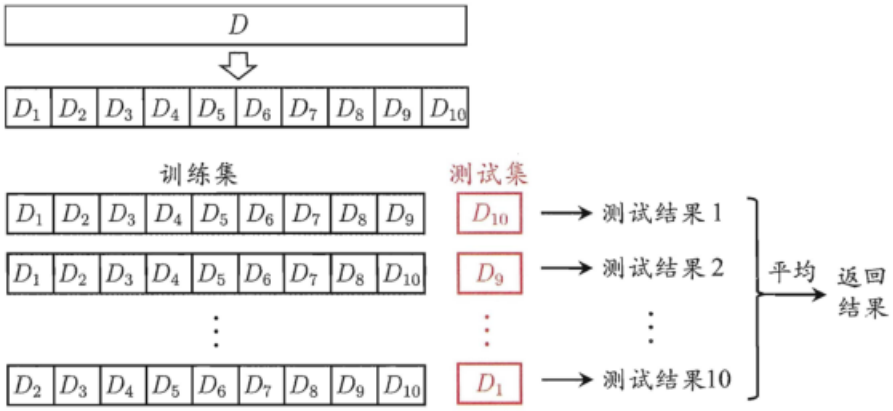
训练集、验证集和测试集

验证集与测试集看起来十分相似，主要区别在于应用阶段的不同。

- 一般首先将数据划分为训练数据和测试数据，之后再将训练数据划分为训练集和验证集，基于验证集上的性能来进行模型选择和调参
- 测试集主要用于对比不同算法的泛化性能。利用测试集上的判别效果来估计模型在实际使用时的泛化能力

交叉验证（K–折验证

交叉验证先将数据划分为K个大小相似的互斥子集（每个自己要尽可能保持数据分布的一致性，通过分层采样得到）。之后每次使用k–1个自己的并集作为训练集，余下的子集作为测试集。通过这种方式获得k组训练/测试集，从而可进行k次训练和测试。最终返回这k个测试结果的均值。



自助法

自助法依据的是一个极限：

$$\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^m \mapsto \frac{1}{e}\approx 0.368$$

给定包含m个样本的数据集，对其进行采样，每次取出一个样本，取出后把样本放回。这样重复执行m次后，得到一个包含m个样本的数据集，里面的元素可能会有重复。上面的极限表明，初始的数据集中36%左右的数据并未出现在采样数据集中，从而可以将新的数据集作为训练集，而将那些未出现的样本作为测试集。这样实际评估的模型与期望评估的模型都使用m个训练样本，但仍有约三分之一左右的数据未在训练集中出现的样本用于测试。

自助法在数据集较小，难以划分训练测试集时有一定作用。但它改变了初始数据集的分布，会引入估计偏差。

2.结果描述

今天看了周志华的机器学习的概述部分，其中一些内容还没有完全搞懂，不过目前感觉还好。明天继续，争取把神经网络跟SVM看完。