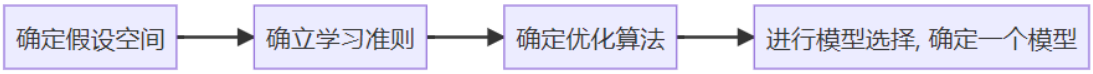
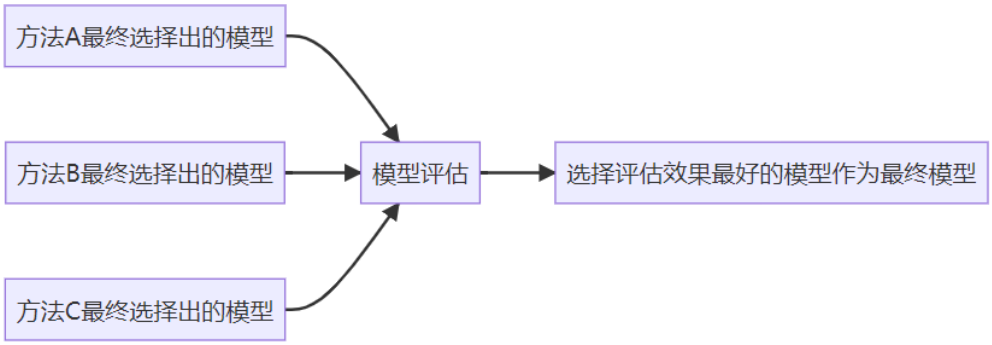
模型的选择与评估

模型选择

模型选择(model selection)有两层含义：

- 训练得到的模型可能不止一个，需要从中进行选择；
- 对于一个具体问题，采用不同的方法得到的模型，在这些模型训练结束后，我们需要决定使用哪一个。





对于每一种方法，都要经历三要素及内部的模型选择，再进行模型评估和最终选择。

评估方法

一些名词解释：

对于**分类模型**，假设M个样本中有n个样本分类错误：

- **错误率**（error rate）：分类错误的样本数占样本总数的比例

$$Err = \frac{n}{M}$$

- **精度**（accuracy）：分类正确的样本数占样本总数的比例

$$Acc = 1 - \frac{n}{M}$$

- **误差**（error）：学习器的实际预测输出 (**prediction**) 与样本的真实输出 (**true**)之间的差异
- **训练误差**（training error）/**经验误差**（empirical error）：学习器在训练集上的误差
- **泛化误差**（generalization error）：学习器在新样本上的误差
- 机器学习的**目标**：得到泛化误差小的学习器。但是实际能做的是**努力使经验误差最小化。**

通常，通过实验测试来对学习器的泛化误差进行评估。为此，需使用一个**测试集**（testing set）来测试学习器对新样本的判别能力，然后以测试集上的**测试误差**（testing error）作为泛化误差的近似。

评估方法即**通过适当的处理，从包含M个样例的数据集D中产生出训练集S和测试集T**

留出法（hold-out）

留出法（hold-out）直接将数据集D划分为两个**互斥**的集合，其中一个集合作为训练集S，另一个作为测试集T，满足：

$$S \cup T = D, \quad S \cap T = \varnothing$$

在 S 上训练出模型后，用 T 来评估其测试误差，作为对泛化误差的估计。

- 划分过程中需要采取均匀的分层取样（stratified sampling）
- 一般**单次留出法往往不可靠**，需要反复多次进行留出法产生不同的训练集与测试集，对**评估结果取均值**
- 常见的分法是训练集占总数据集的2/3~4/5

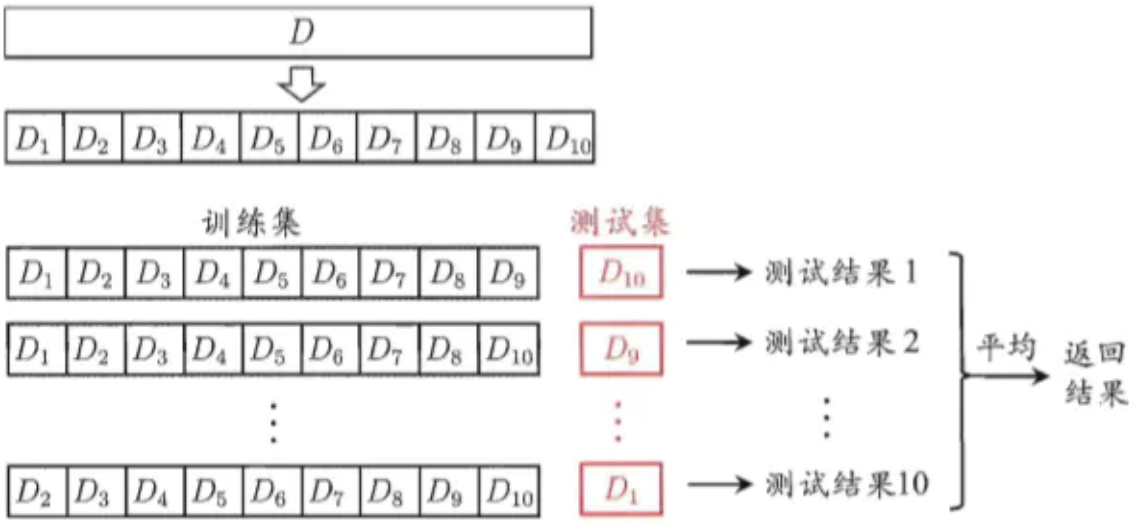
交叉验证法（cross-validation）

交叉验证法（cross validation）先将数据集D划分为k个大小相似的互斥子集：

$$D = D_1 \cup D_2 \cup \ldots \cup D_k$$
$$Di \cap Dj = \varnothing \quad i \neq j$$

每个子集都尽可能保持数据分布的一致性，即通过**分层采样**得到。每次用 k-1 个子集的并集作为训练集，余下的那个子集作为测试集。这样就可获得 k 组训练测试集，从而可进行 k 次训练和测试，最终返回的是这 k 个测试结果的均值。

- 通常把交叉验证法称为**k折交叉验证**（k-fold cross validation）。k最常用的取值是10，此时称为10折交叉验证
- **k折交叉验证通常也要随机使用不同的划分重复p次，最终会得到p×k次结果并取均值**



10 折交叉验证示意图

留一法（Leave-One-Out， LOO）

假定数据集D中包含m个样本，当k=m，则得到了交叉验证法的一个特例：留一法（Leave-One-Out，LOO）

- 在绝大多数情况下，留一法中被实际评估的模型与期望评估的用D训练出的模型很相似。因此，留一法的评估结果往往被认为**比较准确**
- 缺点：在数据集比较大时运算量无法接受

自助法（bootstrapping）

给定包含m个样本的数据集D，我们对它进行采样产生数据集D'。每次随机从D中挑选一个样本，将其拷贝放入D'，然后再将该样本放回初始数据集D中，使得该样本在下次采样时仍有可能被到。这个过程重复执行m次后，我们就得到了包含m个样本的数据集D'，即自助采样的结果。

这样做的合理性：有一部分D中的样本会在D'中多次出现，而也有一部分样本不会在D'中出现。m次采样时不被采到的概率为：

$$(1-\frac{1}{m})^m$$

取极限：

$$\lim_{x\to+\infty}(1-\frac{1}{m})^m=\frac{1}{e}\approx 0.368$$

即通过自助采样，初始数据集D中约有36.8%的样本未出现在采样数据集D中。于是可将D'用作训练集，D\D'用作测试集。

- 自助法在数据集较小、难以有效划分训练/测试集时很有用
- 在初始数据量足够时，留出法和交叉验证法更常用一些