作为一个初学者来说，这一章可能会过于提炼和抽象。但当真正自己要训练一个模型的时候，本章的指导性思维将很强烈地体现。 当你面对一个实际问题的时候，你应该如何选择你的模型、评判你的模型，这些并不只是经验论，而是实实在在有可以量化的方法。

一种训练集一种算法

经验误差和过拟合

这个情况较为常见 实例：手写算法识别 一共m张图片，其中a张图片判断错误。错误率E = a / m。精度accuracy = 1 - E。

评估方法

训练集、测试集和验证集 测试集来判断泛化能力，即对没见过的数据的预测能力

测试集的保留方法

- 留出法
  - 三七分、二八分，注意需要训练集和测试集独立同分布
  - 多次随机划分，训练多个模型，最后取平均值
- 交叉验证法
  - k折交叉验证，数据量比较大的时候就很慢
- 自助法

验证集

为什么要验证集，用来调参数

性能度量（有具体的公式）

- 性能度量
- 回归最常用的性能度量，均方误差
- 错误率和精度
- 查准率和查全率，很简单的定义被解释地很麻烦，代个数进去就行。以及为什么要用查全率和查准率。

比较检验

看起来似乎有了获取测试集*的评估方法和用于比较模型的性能度量之后，就能够通过不同模型在测试集上的性能表现来判断优劣了。但是！事实上，在机器学习中，模型比较并不是这样简单的比大小，而是要考虑更多。

注：指验证集，但无论是书中还是论文中，都使用测试集较多，明白两者的区别就可以了。

在模型比较中，主要有以下三个重要考虑：

1. 测试集上的性能只是泛化性能的近似，未必相同；
2. 测试集的选择对测试性能有很大影响，即使规模一致，但测试样例不同，结果也不同；

3. 一些机器学习算法有随机性，即便算法参数相同，在同一测试集上跑多次，结果也可能不同；

那么应该如何有效地进行模型比较呢？答案是采用**假设检验（hypothesis test）**。基于假设检验的结果，我们可以推断出，若在测试集上观察到模型A优于B，则是否A的泛化性能在统计意义上也优于B，以及做这个结论的把握有多大。

本小节首先介绍最基本的二项检验和t检验，然后再深入介绍其他几种比较检验方法。默认以错误率作为性能度量。

几个基础概念：

- **置信度**：表示有多大的把握认为假设是正确的。
- **显著度**：也称"显著性水平"，表示假设出错的概率。显著度越大，假设被拒绝的可能性越大。
- **自由度**：不被限制的样本数，也可以理解为能自由取值的样本数，记为 ν 或 df。

In [ ]: