模块评估与选择

泛化能力

评价一个模型好坏要看他的泛化能力好不好，但具体的“好”怎么定义有很多标准，这取决于我们的需求是什么

过拟合，欠拟合

泛化误差：在新的数据上的误差
经验误差：训练集上的误差
我们通过后者提高前者
然而并不是经验误差越小越好，会出现过度拟合，也就是把训练集中的特性当作了一般规律
泛化误差存在微笑曲线，一头是欠拟合，另一头是过拟合，要找到临界点至关重要
机器学习的算法就是在过拟合欠拟合之间找平衡

三大问题

1. 如何评估未来数据上的表现 评估方法
2. 在我们的需求上表现怎么样 性能度量
3. 统计学上表现怎么样 比较检验

评估方法

怎么获得测试集?
首先训练集，测试集要互斥，有三个方法

1. 留出法
   一部分测试，一部分训练
   确保分层采样，保证数据分布的一致
   测试集保持在0.2-0.33之间，越小泛化误差越不准，越大训练越少泛化误差可能更大
   要多次重修切割以抹除随机性

有可能会遗漏一些训练数据

2. k折训练
   把训练集分成k个小的训练包，每次留一个不同的训练包当测试
   极端版本是留一法，m99

3. 自助法
   每次挑选出来的训练数据还会被放回去，有多次抽到的可能

在尽力又保证训练量接近m100，又留有测试集（没有采到的数据大约=0.368）
改变了训练数据分布

调参，训练集

参数包括超参数（人为设置的次数）和模型参数（学习到的系数），在训练过程中调整超参数以优化
调参使用验证集，验证集来自于训练集，调参也是训练的一部分

性能度量

回归任务用均方误差
分类任务用错误率（或者精度，精度=1-错误率）
具体来说可以有混淆矩阵，tp，tn，fp，fn
进而得到查准率（tp/tp+fp），查全率（tp/tp+fn）（找到了几成正的样本）

比较检验

为了解决三个问题

1. 测试性能不等于泛化性能
2. 机器学习算法本身有随机性
3. 测试性能因测试集变化而变化

比较两个模型，我们使用

1. 交叉验证t检验
2. McNemar检验（基于列联表）