**周志华: 《机器学习》 ch1**

**1、概念**

**机器学习**：**让机器自动地去学习数据中学习，产生基于经验的模型**。机器学习主要研究学习算法。

**样本空间/属性空间/输入空间**：输入特征/属性张成的空间；

**特征向量**：一个样本或一个示例称为一个特征向量；

**学习/训练**：从数据中学习模型的过程，这个过程通过执行某个学习算法来说实现；

**样例**：拥有标记信息的示例或样本；

**标记空间/输出空间**：所有标记组成的空间；

研究的问题：

**有监督学习：**分类（二分类、多分类）、回归

**无监督学习：**聚类

**2、模型的泛化能力**

**泛化能力**：学得的模型，适用于新样本的能力，称为模型的泛化能力；具有强泛化能力的模型应该能够适用于整个样本空间；

我们希望训练集的样本对于样本空间具有很好的代表性，于是假设每个样本之间都是独立同分布的。因此，**如果我们训练集拥有的样本越多，相应地关于样本空间的信息也就越多，这样越有可能训练出泛化能力强的模型**。

**归纳与演绎：**

归纳：从特殊到一般，即从具体的事实归纳出一般的规律；

演绎：从一般到特殊，即从基本原理推到出具体情况；

机器学习实际上是从样例中学习，因此**机器学习是一个归纳的过程**。

**3、假设空间与归纳偏好**

**假设空间/函数空间：**

**机器学习中可能的函数或假设组成的空间。**模型是输入空间到输出空间之间的映射，而这样的映射的集合，组成了假设空间。**假设空间的确定意味着学习范围的确定。**

**归纳偏好**：

机器学习算法在学习过程中对**某种类型假设的偏好，称为归纳偏好**；我们的学习算法必须具有某种偏好，这样才能产出它认为正确的模型。

**归纳偏好可以看作学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或者价值观。**

**4、选择模型**

**一般性原则：**

**奥卡姆剃刀：**如果有多个假设与观察一致，那么选择最简单的那个；但是到底什么叫简单的模型，也很难评判；

**没有免费午餐原则(NFL)**：如果认为所有的问题同等重要（正类和负类），那么不同算法的期望性能是相同的；但是实际上，我们只关注自己想解决的问题，NFL告诉我们，如果要谈论算法的性能，要结合实际的问题和需求；

**5、机器学习与数据挖掘**

**机器学习与数据挖掘**：

数据挖掘是从海量的数据中发掘知识，数据库技术为数据挖掘提供数据管理技术，机器学习和统计学的研究为数据挖掘提供数据分析的技术；

**统计学主要通过机器学习对数据挖掘产生影响，而机器学习和数据库则是数据挖掘的两大支撑；**

**6、推荐的书**

1. **Mitchell 《Machine Learning》**
2. **Duda 《Pattern Classification》**
3. **李航 《统计机器学习》**

**周志华: 《机器学习》 ch2 模型评估与选择**

**1、经验误差与过拟合**

**错误率(error rate)**：分错样本占样本总数的比例

**准确率(accuracy)**：1 – error rate

**误差(error)**：学习器的预测输出与实际输出之间的差异

**训练/经验误差(training error)：**学习器在训练集上的误差

**泛化误差(generalization error)：**学习器在测试集上的误差

我们希望得到泛化误差小的模型。

**过拟合：**学习器将训练样本学的太好了，将一些样本本身的特点当做了所有潜在样本都会有的特征，这样模型的泛化能力会比较差。表现是模型在训练集上有很高的准确率，但是在测试集上的准确率比较低。

**欠拟合：**对训练样本的一般特性都没有学好。

**过拟合的可能原因：**

（1）训练样本抽取错误：训练样本太少，或者没有代表性；

（2）样本中的噪音数据干扰过大；

（3）模型复杂度太高

（4）训练次数过多

**过拟合解决办法：**

（1）增加训练样本数：采样获取更多的数据、数据增强；

（2）降低模型的复杂度，或者选用更简单的模型；

（3）减少模型训练的次数，或者提前结束训练；

（4）在目标函数中加入正则化项，设置适当的权重；

（5）对于树模型：剪枝

（6）对于神经网络：dropout、权重衰减、减少模型的层数

**过拟合只能缓解，或者减小其风险，无法彻底避免。**

**2、模型评估方法**

将训练数据划分为训练集和验证集，然后将模型在验证集上的误差作为模型泛化误差的估计。下面是几种常用的方法：

**（1）留出法**

按照一定的比例，对不同类别分别进行分层采样。常见的做法是将大约2/3-4/5的样本用于训练，剩余样本用于测试。

但是留出法有一个缺点，用于训练的样本偏少，评估结果可能不够稳定和准确。一般我们对数据进行多次划分，对分类结果取均值。

**（2）交叉验证法**

交叉验证先将数据集划分为k个大小相等的互斥子集，每个子集都通过分层采样获得。每次拿k-1个子集训练，剩下的一个子集做验证。因此可以进行k次试验，最后取k次试验的均值。这种方法也叫做k-folder。

**（3）留一法**

对于k folder，假设共有m个样本，如果取k=m，则得到特例-----留一法。

这种方法在数据量很大的情况下，性能比较差，费时。

**（4）自助法**

假设从包含m个样本的数据集中每次有放回抽取样本，则可以估计某个样本在m次抽样之后不被抽中的概率是：

当m趋近于无穷大时，极限是1/e，即0.368。每个样本永远不被抽中的概率为0.368。

自助法在数据量小时比较有用。

**一般都用留出法和交叉验证法。**

**3、关于调参**

除了要选择适当的模型，还要选择合适的模型参数，这就是通常说的“参数调节”(parameter tuning)。

学习算法的参数一般都是在实数范围内取值，因此对每种参数都进行训练显然是不现实的，常用的做法是，对每个参数选定一个范围和变化步长，只训练某些位置的参数的模型。

**在模型选择完，参数设置完毕之后，此时应该用所有数据重新训练一次模型，这个使用所有样本训练的模型，才是最终提交给用户的模型。**

**4、模型性能度量**

衡量模型泛化能力的标准。

**4.1 回归**

回归任务中，常用的性能度量就是均方误差（MSE），即误差平方和。

**4.2 分类**

**错误率：**错分样本的比例

**准确率：**正确分类样本的比例

**查准率(Precision, P)： 所有被分为正样本的样本中，真正是正样本的比例**

P = TP / (TP + FP)

**查全率(Recall, R)： 所有正样本，被正确分类的比例**

R = TP / (TP + FN)

**F1值(F1)： 查全率和查准率的调和平均**

2 / F1 = 1 / R + 1/P

**P-R曲线：**

按照我们根据学习器预测的结果，对样例进行排序，排在前的被学习器认为是最可能的正样本，排在后的最不可能是正样本。按照这个顺序，逐个选取样本作为正样本，则排在它前面的被认为都是正样本。

每次选取，都可以计算得到一组P、R，最后将所有P、R画成曲线。得到PR曲线。

比较模型性能，找平衡点（P=R），取值最大的效果更好。

**ROC曲线：**

类似于P-R曲线，ROC曲线是TPR和FPR之间的曲线，即正样本被正确分类与负样本被分为正样本之间的曲线。

**AUC：**

ROC曲线下的面积

**5、偏差与方差**

**偏差-方差分解**是一种理解模型泛化能力的方法。

**方差**：模型预测值的方差；度量模型的稳定性

**偏差**：模型期望预测值与真实值之间的平方误差；度量模型的拟合能力

**噪声**：样本标注与真实值之间的平方误差；刻画问题本身的复杂性

一般来说，**泛化误差 = 方差 + 偏差 + 噪声**

偏差-方差是存在矛盾的，被称为**偏差-方差窘境**

在模型训练初期，模型复杂度低，方差小，偏差大；此时**欠拟合**。

随着模型的训练，方差增大，偏差减小；

最后随着模型复杂度提升，方差变得非常大，而偏差很小，模型对样本的拟合能力很强，但是也拟合了噪声。模型的拟合能力上升，但是稳定性降低。此时**过拟合**