统计学习方法

《统计学习方法》一书的读书笔记。

# 1.概论

统计学习，statistical learning，是关于计算机基于数据，构建概率统计模型，并运用模型对数据进行分析和预测的一门学科。统计学习也称为统计机器学习，statistical machine learning。

统计学习关于数据的基本假设，是同类数据具有一定的统计规律性，这是统计学习的前提！

统计学习用于对数据进行预测和分析，特别是对未知数据进行预测和分析。这是通过构建概率统计模型来实现的。统计学习总的目标就是考虑学习什么样的模型和如何学习模型，以使模型能对数据进行准确的预测和分析，同时也要考虑尽可能的提高学习效率。

统计学习分为：监督学习(supervised learning)，非监督学习(unsupervised learning)，半监督学习(semi-supervised learning)，强化学习(reinforcement learning)等。

全书针对的是监督学习。

统计学习首先需要的，是一个训练数据集合(training data)，这个数据集合是预先准备好的、有限的数据集合，且假设数据是**独立同分布**的。

其次我们假设要学习的模型属于某个函数的集合，称为假设空间(hypothesis space)中；

这样，统计学习就可以概述为：应用某个评价准则(evaluation criterion)，从假设空间中选取一个最优的模型，使它对已知训练数据及未知测试数据(test data)在给定的评价标准下，都有最优的预测。最优模型的选取，由算法实现。

由此我们可以得到统计学习的三要素：模型(model)、策略(strategy)、算法(algorithm)。

实现统计学习的步骤如下：

1. 获取有效的训练数据集合；
2. 确定包含了所有的可能的模型的假设空间，也就是学习模型的集合；
3. 确定模型选择的准则，即学习的策略；
4. 实现求解最优模型的算法，即学习的算法；
5. 通过学习算法选择最优模型；
6. 利用学习的最优模型，对新数据进行预测和分析；

## 1.1.监督学习

监督学习的任务，是学习一个模型，使得模型可以对任意给定的输入，对其相应的输出做出一个好的预测。