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。

### 1.1.1.输入空间、特征空间、输出空间

在监督学习中，将输入和输出所有可能的取值集合分别称为输入空间和输出空间。

每个具体的输入，通常是一个实例(instance)，一般由特征向量(feature vector)表示。这时，所有特征向量存在的空间就称为特征空间。特征空间的每一维对应一个特征。

### 1.1.2.问题分类

输入变量X和输出变量Y可以是连续的，也可以是离散的。根据输入输出变量的不同类型，对预测任务给予不同的名称：

* 输入和输出变量均为连续变量的预测问题称为**回归**问题；
* 输入和输出变量均为离散序列的预测问题称为**标注**问题；
* 输入变量为连续变量，输出变量为离散序列的预测问题称为**分类**问题；

### 1.1.3.联合概率分布

监督学习假设输入与输出的随机变量X和Y遵循联合概率分布：P(X, Y)；P(X, Y)被称为分布函数，或分布密度函数。

统计学习假设数据之间存在一定的统计规律，X和Y具有联合概率分布的假设就是监督学习关于数据的基本假设。

### 1.1.4.假设空间

监督学习的目的在于学习一个输入到输出的映射，这一映射由**模型**来表示。换句话说，学习的目的就是找到最好的模型。

模型，属于输入空间到输出空间的映射的集合，这个集合称之为**假设空间**。假设空间的确定，意味着学习范围的确定。

监督学习的模型可以是概率模型或非概率模型，由条件概率分布P(Y|X)或决策函数Y=f(X)表示，随具体的学习方法而定。

## 1.2.统计学习三要素

统计学习的三要素可以简单的表示为：方法 = 模型 + 策略 + 算法；

### 1.2.1.模型

统计学习首要要考虑的问题是学习什么样的模型。

在监督学习过程中，模型就是要学习的条件概率分布或决策函数。

假设空间中包含了所有可能的条件概率分布或决策函数。例如，假设决策函数是输入变量的线性函数，那么模型的假设看哦那关键就是所有这些线性函数的集合。

假设空间中的模型一般有无穷多个。

### 1.2.2.策略

有了模型的假设空间，统计学习接下来要考虑的就是按照什么样的准则学习或选择最优的模型。

统计学习的目标就是从假设空间中选择最优模型。

#### 损失函数和风险函数

监督学习问题是在假设空间F中选择模型f作为决策函数，对于给定的输入X，由f(X)给出对应的输出Y，这个输出的预测值f(X)可能与实际的结果Y一致，也可能存在不同。这时候需要一个损失函数，用来度量预测错误的程度。

损失函数是f(X)和Y的非负实值函数，记作：L(Y, f(X))；

统计学习常用的几种损失函数有：

* 0-1损失函数：如果Y=f(X)，L(Y, f(X))=1；否则L(Y, f(X))=0；
* 平方损失函数：L(Y, f(X)) = (Y - f(X))2;
* 绝对损失函数：L(Y, f(X)) = |Y - f(X)|;
* 对数损失函数：L(Y, P(Y|X)) = -logP(Y|X);

损失函数的值越小，模型就越好。

由于模型的输入、输出(X, Y)是随机变量，遵循联合分布P(X, Y)，所以损失函数的期望是：Rexp(f) = Ep[L(Y, f(X))]；这是理论上模型f(X)对于联合分布P(X, Y)的平均意义下的损失，称为风险函数或期望损失。

学习的目标，就是选择期望风险最小的模型。由于联合分布P(X, Y)是未知的，因此期望损失无法直接计算得到。实际上，如果知道了联合分布P(X, Y)，那么也就不需要进行学习了。

所以一般来说，还会引出另一个概念：经验风险或经验损失。给定一个训练数据集，模型关于训练数据集的平均损失，称为经验风险。

理论上说，随着数据量的无限增加，经验风险会无限趋近于期望损失。但是现实中的数据量，往往较小，这就需要经常对经验风险进行矫正。这就关系到监督学习另外两个基本策略：经验风险最小化和结构风险最小化。

#### 经验风险最小化

在假设空间、损失函数以及训练数据集确定的情况下，经验风险函数就可以确定。经验风险最小化的策略认为，经验风险最小的模型就是最优的模型。按照这个策略，得到如下经验风险最小化模型：



其中,F是假设空间。

当样本空间的容量足够大时，经验风险最小化能够保证很好的学习效果。比如现实中使用广泛的极大似然估计，就是经验风险最小化的例子。但是，如果样本的数量很小，经验风险最小化的效果未必很好，很可能产生“过拟合”的现象。

#### 结构风险最小化

结构风险最小化是为了防止过拟合提出来的策略，等价于正则化。结构风险在经验风险的基础上，增加了表示模型复杂度的正则化项或罚项。在假设空间、损失函数和训练数据集确定了的情况下，结构风险的定义是：



其中J(f)是模型的复杂度，是定义在假设空间F上的泛函。模型f越复杂，复杂度就越大；反之就越小。λ是系数，非负数，用来权衡经验风险和模型复杂度。

结构风险小的模型一般对于训练数据以及未知的测试数据都有较好的预测。

### 1.2.3.算法

算法是指学习模型的具体计算方法。

统计学习基于训练数据集，根据学习策略，从假设空间中选择最优模型，最后需要考虑的，就是用什么样的计算方法求解最优模型。

这样，统计学习问题最终归结为最优化问题，统计学习的算法成为求解最优化问题的算法。