《统计学习方法》(李航)
归纳总结一
统计学习方法概论

西安交通大学

联系方式：williamyi96@gmail.com

易凯

2017 年 1 月 21 日

目录

1 统计学习方法概论

1.1 统计学习

统计学习的特点 统计学习是关于计算机基于数据构建概率统计模型并且运用模型对数据进行分析预测的学科，其也被称之为统计机器学习。其具有对于数据的强依赖性。

统计学习的对象 数据

统计学习的目的 考虑学习怎样的模型以及怎样学习模型来提高模型对于数据预测的准确率，同时提高学习的效率。

统计学习的方法 监督学习、非监督学习、半监督学习、强化学习。

实现统计学习方法得到一般步骤

1. 得到一个有限的训练数据集合；
2. 确定包含所有可能模型的假设空间，即学习模型的集合；
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. 通过学习方法选择最优模型；
6. 通过学习的最优模型对新数据进行预测或者分析。

统计学习的研究

1. 统计学习方法。开发新的学习方法；
2. 统计学习理论。探究统计学习方法的有效性和效率以及统计学习基本理论；
3. 统计学习应用。将统计学习方法应用到实际问题之中，解决实际问题。

1.2 监督学习

监督学习基础认识 学习一个模型，使模型能够对任意给定的输出，对应相应的输出做出一个好的预测 (注意其好坏的标准已经进行了限定)。

基本概念

1. 输入空间、特征空间、输出空间: 输入空间是指所有输入量的所有可能取值的集合，由于往往是包含许多特征的列向量，因此输入空间也被称之为特征空间。

输出空间是所有可能得到的输出值的集合。

2. 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)。其中 P(X,Y) 表示分布函数或者是分布密度函数。

3. 假设空间： 假设空间是所有可能的模型的集合。

问题的形式化 问题的形式化过程如图所示，主义预测系统可以是分类问题、回归问题或者是标签问题 (关于三者的区别将在后面讲解)

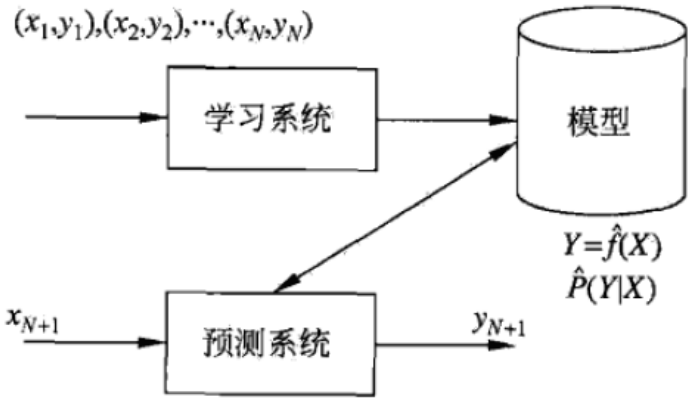

图 1.1 监督学习问题

1.3 统计学习三要素

方法 = 模型 + 策略 + 算法

模型 模型就是在统计学习方法中联系输入空间与输出空间的 f。在这里指明的是模型的假设空间。

策略 策略是指应该按照怎样的规则进行学习，并且选择最优的模型。

1. 损失函数和风险函数

a). 0-1 损失函数 (0-1 loss function)

$$L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases}$$

b). 平方损失函数 (quadratic loss function)

$L(Y, f(X)) = (Y - f(X))^2$

c). 绝对损失函数 (absolute loss function)

$L(Y, f(X)) = |Y - f(X)|$

d). 对数损失函数 (logarithmic loss function) 或者对数似然损失函数 (log-likelihood loss function)

$L(Y, P(Y|X)) == logP(Y|X)$

风险函数 (risk function) 或期望损失 (expected loss)

$R_{exp}(f) = E_p[L(Y, f(X))] = \int_{X \times Y} L(y, f(x))P(x, y)dxdy$

经验风险 (empirical risk) 或经验损失 (empirical loss)

$R_emp(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$

经验风险最小化和结构风险最小化

监督学习的两个基本策略就是经验风险最小化和结构风险最小化。

其中，经验风险的最小化就是求解模型的最优化问题：

$min_{f \in F} \sum_{i=1}^{N} L(y_i, f(x_i))$

但是由于在经验风险最小化的过程中，如果数据量过少，那么可能产生过拟合的现象。解决这种问题的最好方法就是使用正则化，也便是我们提到的结构风险最小化 (structural risk minimization, SRM)。

结构风险为 $R_{srm}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$

其中 J(f) 为模型的复杂度，复杂度表示的是对复杂模型的惩罚；而其系数则是用以权衡经验风险和模型复杂度。

算法 算法是指学习模型中的具体计算方法

1.4 模型评估与模型选择

训练误差与测试误差 训练误差与测试误差的本质是数据集上的实际值与模型计算出来的理论值之间的误差函数之和的平均值。不同的是训练误差是针对训练数据而言的，而测试误差是针对测试数据集而言的。

过拟合与模型选择 如果一味得追求提高对训练数据的预测能力，那么所选用的模型往往较为复杂，其一般比真实的模型的复杂度要高，这种现象被称之为过拟合现象。

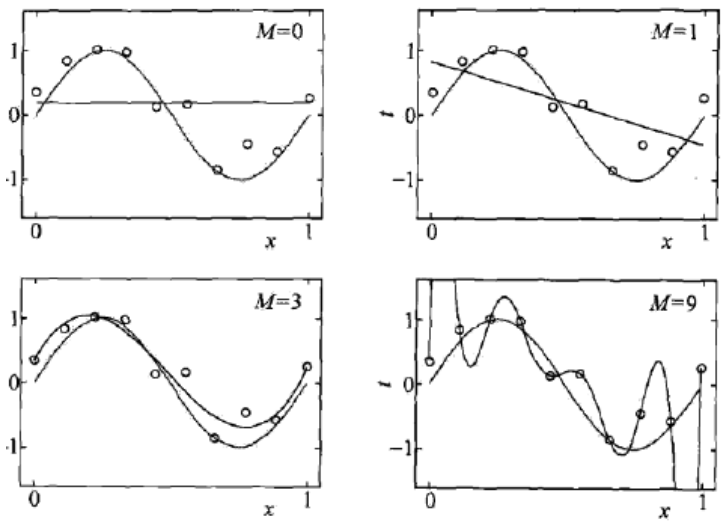

图 1.2 M 次多项式函数拟合问题的例子

为防止过拟合现象的发生，因此我们需要选用复杂度适当的模型，模型选择的两种常用方法就是正则化和交叉验证。

1.5 正则化与交叉验证

正则化 前面也提到了，正则化就是在原先的经验损失之后加上一项，其被称之为正则化项 (regularizer) 或者是罚项 (penalty term)。

在回归问题的平方损失函数中，损失函数的表达式为：

$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i;w) - y_i)^2 + \frac{\lambda}{2}|w|^2$

(其中 |w| 表示参数向量 w 的 L 范数)

交叉验证 在数据集充足的情况之下，可以将数据分为训练集、验证集和测试集三个部分。

在数据集不足的情况下，为了得到较好的模型，一般采用交叉验证的方法。

交叉验证分为三种形式：

1. 简单交叉验证 73 开。然后用不同的参数个数来训练模型得到不同模型，从中选出误差率最小的模型。

2. S 折交叉验证 将数据分成随机的 S 个子集，其中仅留一个自己进行测试，然后反复上述过程多次，选出误差率最小的模型。

3. 留一交叉验证 S 折交叉验证的特例，其中 S=N。

1.6 泛化能力

学习方法的泛化能力 (generalization ability) 是指该方法学习到的模型对未知数据的预测能力。

其中泛化误差为：

$R_{exp} = E_p[L(Y, \widehat{f}(x))] = \int_{X\times Y} L(Y, \widehat{f}(x))P(x,y)dxdy$,

其中学到的模型为 $\widehat{f}$。

1.7 生成模型与判别模型

生成模型 (generative model) $P(Y|X) = \frac{P(X,Y)}{P(X)}$

判别模型 (discriminative model) 判别模型是直接进行 P(Y|X) 的求解

生成方法的优点

1. 可以还原联合概率分布 P(X,Y)
2. 学习收敛速度更快，样本容量增加时，更容易收敛到真实模型
3. 存在隐变量时仍可以使用

判别方法的优点

1. 学习的准确率往往较高，直接预测；
2. 可以对数据进行各种程度的抽象，简化学习

1.8 分类、回归、标注问题

分类问题、回归问题和标准问题解决的思路都是相同的，参照的都是图 1.1，其中分类模型是特征空间是连续的，而输出空间是离散的。回归问题是特征空间和输出空间都是连续的。而标准问题时特征空间和输出空间都是离散的。