Chap1:

1. 输入变量与输出变量均为连续变量的预测问题称为回归问题；

输入变量为有限个离散变量的预测问题称为分类问题；

输入变量与输出变量均为变量序列的预测问题称为标注问题。

1. 统计学习方法由三要素构成：

方法=模型+策略+算法

1. 结构风险最小化是为了防止过拟合而提出的策略，结构风险最小化等价于正则化。
2. 如果一味追求提高对训练数据的预测能力，所选模型的复杂度往往会比真模型更高，这种现象称为过拟合。过拟合是指学习时选择的模型所包含的参数过多，以至于出现这一模型对已知数据预测得很好，但对未知数据预测得很差的现象。模型选择旨在避免过拟合并提高模型的预测能力。
3. 在多项式函数拟合中可以看到，随着多项式次数的增加，训练误差会减小，直至趋向于0，但测试误差却并不是如此，它会随着多项式次数的增加先减小后增大，而最终的目的是使测试误差达到最小。
4. 在学习的时候，我们要防止过拟合，进行最优的模型选择，即选择复杂度适当的模型，以达到测试误差最小的学习目的。两种常用的模型选择方法：正则化和交叉验证。
5. 正则化与交叉验证：

正则化是结构风险最小化策略的实现，是在经验风险上加一个正则化项或者罚项。正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化值就越大。

正则化的作用是选择经验风险与模型复杂度同时较小的模型。

另外一种常见的模型选择方法是交叉验证：

如果样本数据充足，进行模型选择的一种简单方法是随机将数据集切分成三部分，分为训练集、验证集和测试集。训练集用来训练模型，验证集用于模型的选择，而测试集用于最终对学习方法的评估。

为了选择较好的模型，可以采用交叉验证方法。交叉验证的基本思想是重复地使用数据，把给定的数据进行切分，将切分的数据集组合为训练集与测试集，在此基础上反复地进行训练，测试以及模型选择。

简单交叉验证/ S折交叉验证 / 留一交叉验证

1. 泛化能力

学习方法的泛化能力是指由该方法学习到的模型对未知数据的预测能力，是学习方法本质上重要的性质。泛化误差反映了学习方法的泛化能力，如果一种方法学习的模型比另一种方法学习的模型具有更小的泛化误差，那么这种方法就更有效。

1. 生成模型与判别模型：监督学习方法可以分为生成方法和判别方法，所学到的模型称为生成模型和判别模型。

生成方法由数据学习联合概率分布，然后求出条件概率分布作为预测的模型，即生成模型。典型的生成模型有：朴素贝叶斯方法和隐马尔可夫模型。

判别方法由数据直接学习决策函数或者条件概率分布作为预测的模型，即判别模型。判别方法关心的是对给定的输入X，应该预测什么的输出Y。典型的判别模型包括：k近邻法、感知机、决策树、逻辑斯谛回归模型、最大熵模型、支持向量机、提升方法和条件随机场等。

1. 分类问题：分类是监督学习的一个核心问题，在监督学习中，当输出变量Y取有限个离散值时，预测问题便成为了分类问题。

对于二类分类问题常用的评价指标是精确率与召回率。通常关注的类为正类，其他类为负类，分类器在测试数据集上的预测或正确或不正确，4种情况出现的总数分别记作：

TP -----将正类预测为正类数

FN -----将正类预测为负类数

FP -----将负类预测为正类数

TN -----将负类预测为负类数

分类问题的应用：

在银行业务中，构建一个客户分类模型，对客户按照贷款风险的大小进行分类；

在网络安全领域，利用日志数据的分类对非法入侵进行检测；

在图像处理中，分类可以用来检测图像中是否有人脸出现；

在手写识别中，分类可以用于识别手写的数字；

在互联网搜索中，网页的分类可以帮助网页的抓取、索引与排序。

1. 标注问题：

标注问题的输入是一个观测序列，输出是一个标记序列或者状态序列。标注问题的目标在于学习一个模型，使它能够对观测序列给出标记序列作为预测。

标注常用的统计学习方法：隐马尔可夫模型、条件随机场。

标注问题在信息抽取、自然语言处理等领域被广泛应用。是这些领域的基本问题。

1. 回归问题：

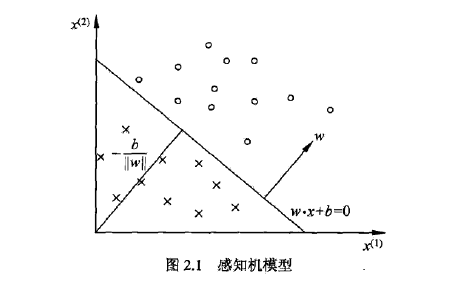
回归用于预测输入变量和输出变量之间的关系。回归问题的学习等价于函数拟合，选择一条函数曲线使其很好地拟合已知数据且很好的预测未知数据。回归问题分为学习和预测两个过程，首先给定一个训练数据集，学习系统基于训练数据构建一个模型，预测系统根据学习的模型确定相应的输出。

回归问题最常用的损失函数是平方损失函数，在此情况下，回归问题可以由著名的最小二乘法求解。

举例：股价预测问题（对未来的股价进行预测）

Chap 2: 感知机

1. 感知机是二类分类的线性分类模型，其输入为实例的特征向量，输出为实例的类别。取+1和-1二值。感知机对应于输入空间中将实例划分为正负两类的分离超平面，属于判别模型。感知机学习旨在求出将训练数据进行线性划分的分离超平面。感知机是神经网络和支持向量机的基础。
2. 感知机模型：由输入空间到输出空间的函数称为感知机。感知机为一种线性分类模型，属于判别模型。线性方程w\*x + b = 0，对于特征空间中的一个超平面S，其中w为超平面的法向量，b为超平面的截距，这个超平面将特征空间划分为两个部分。位于两个部分的点（特征向量）分别被分为正、负两类，因此，超平面S称为分离超平面。



1. 感知机学习策略

感知机学习的目标是求得一个能够将训练集正实例点和负实例点完全正确分开的分离超平面。为了找出这样的超平面，即确定感知机模型参数w,b,需要确定一个学习策略，即定义损失函数并将损失函数极小化。感知机学习问题转化为求解损失函数式的最优化问题，最优化的方法是随机梯度下降法。

Chap3: k近邻法

1. 定义：k近邻算法简单、直观：给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最近邻的k个实例，这k个实例的多数属于某个类，就把该输入实例分为这个类。
2. K近邻模型：模型由三个基本要素组成----距离度量、k值的选择和分类决策规定决定。K近邻算法中，当训练集、距离度量、k值以及分类决策规定确定后，对于任何一个新的输入实例，它所属的类唯一被确定。
3. K近邻法的实现：kd树。实现k近邻法时，主要考虑的问题是如何对训练数据进行快速k近邻搜索，这点在特征空间的维数大及训练数据容量大时尤其必要。Kd树是二叉树，表示对k维空间的一个划分。构造kd树相当于不断地用垂直于坐标轴的超平面将k维空间切分，构成一系列的k维超矩形区域。
4. 本章概要：k近邻算法是基本且简单的分类与回归方法。K近邻法的基本做法是：对给定的训练实例点和输入实例点，首先确定输入实例点的k个最近邻训练实例点，然后利用这k个训练实例点的类的多数来预测输入实例点的类。

K近邻法中，当训练集、距离度量、k值及分类决策规则确定后，其结果唯一确定。

K近邻三要素：距离度量、k值的选择和分类决策规则。常用的距离度量是欧式距离及更一般的Lp距离。K值小时，k近邻模型更复杂；k值大时，k近邻模型更简单。K值的选择反映了对近似误差与估计误差之间的权衡。通常由交叉验证选择最优的k。

Chap4: 朴素贝叶斯法

1. 朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件假设学习输入/输出的联合概率分布，然后基于此模型，对于给定的输入x, 利用贝叶斯定理求出后验概率最大的输出y.
2. 朴素贝叶斯法是典型的生成学习方法。生成学习方法由训练数据学习联合概率分布P(X,Y),然后求得后验概率分布P(Y|X). 具体的说，利用训练数据学习P(X|Y)和P(Y)的估计，得到联合概率分布：

P(X,Y) = P(Y)P(X|Y)

概率估计方法可以是极大似然估计或者贝叶斯估计。

Chap5: 决策树

1. 决策树是一种基本的分类与回归方法。本章主要讨论用于分类的决策树。决策树呈树形结构，在分类问题中，表示基于特征对实例进行分类的过程。

决策树的学习通常包括3个步骤：特征选择、决策树的生成和决策树的修剪。

1. 模型与学习：

分类决策树模型是一种描述对实例进行分类的树形结构。决策树由节点和有向边组成。节点有两种类型：内部节点和叶节点。内部节点表示一个特征或者属性，叶节点表示一个类。

由决策树的根结点到叶节点的每一条路径构成一条规则：路径上的内部结点的特征对应着规则的条件，而叶节点的类对应着规则的结论。决策树的路径或其对应着的if-then规则集合具有一个重要的性质：互斥并且完备。也就是说，每一个实例都被一条路径或一条规则所覆盖。而且只被一条路径或一条规则所覆盖。

决策树学习的损失函数通常是正则化的极大似然函数。决策树学习的策略是以损失函数为目标函数的最小化。当损失函数确定以后，学习问题就变为在损失函数意义下选择最优决策树的问题。

决策树可能对训练数据有很好的分类能力，但是对未知的测试数据却未必有很好的分类能力，即可能发生过拟合现象。我们需要对已生成的树自上而下的进行剪枝，将树变得更简单，从而使它具有更好的泛化能力。具体地，就是去掉过于细分的叶节点，使其退回到父节点，甚至更高的结点，然后将父节点或更高的结点改为新的叶结点。

决策树学习常用的算法有ID3、C4,5与CART, 下面结合这些算法分别叙述决策树学习的特征选择、决策树的生成与剪枝过程。

1. 特征选择

特征选择的准则是信息增益或者信息增益比。特征选择是决定用哪个特征来划分特征空间。

1. 特征增益

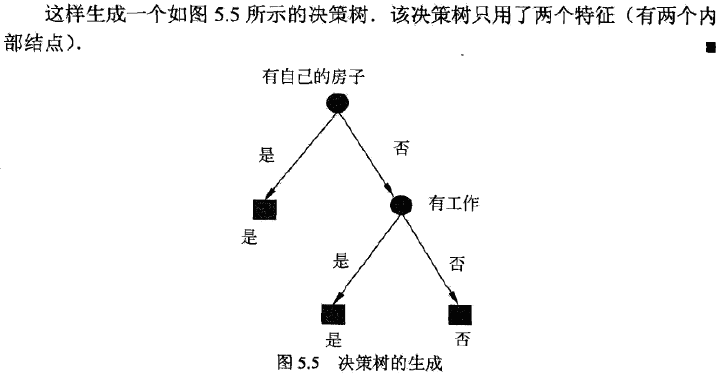
在信息论与概率统计中，熵是表示随机变量不确定性的度量。熵越大，随机变量的不确定性就越大。信息增益表示得知特征X的信息而使得类Y的信息的不确定性减少的程度。

根据信息增益准则的特征选择方法是：对训练数据集D，计算其每个特征的信息增益，并比较它们的大小，选择信息增益最大的特征。

信息增益值的大小是相对于训练数据集而言的，并没有绝对的意义。在分类问题困难时，也就是说在训练数据集的经验熵大的时候，信息增益值会偏大。反之，信息增益值会偏小。信息增益比可以对这一问题进行校正。这是特征选择的另一个准则。

1. 决策树的生成

ID3算法: 核心是在决策树的各个节点上应用信息增益准则选择特征，递归地构建决策树。ID3相当于用极大似然法进行概率模型的选择。



ID3算法只有树的生成，所以该算法生成的树容易产生过拟合。

C4,5算法: C4,5算法对ID3算法进行了改进。C4,5在生成的过程中，用信息增益比来选择特征。

1. 决策树的剪枝