# 机器学习算法系列（16）：统计学习概论

## 一、统计学习

### 1.1 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运用模型对数据进行预测与分析的一门学科。赫尔伯特·西蒙曾对学习定义为：“如果一个系统能够执行某个过程改进它的性能，这就是学习。”按照这一观点，统计学习就是计算机系统通过数据及统计方法提高系统性能的机器学习。

### 1.2 对象

统计学习的对象是数据，从数据出发，提取数据的特征，抽象出数据的模型，发现数据中的知识，又回到对数据的分析与预测中去。数据包括存在于计算机及网络上的各种数字、文字、图像、视频、音频及它们的组合。统计学习对于数据的基本假设是同类数据具有一定###的统计规律性。

### 1.3 目的

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与分析，同时也要尽可能地提高学习效率。

### 1.4 方法

统计学习由监督学习（supervised learning）、非监督学习（unsupervised learning）、半监督学习（semi-supervised learning）、强化学习（reinforcement learning）等组成

- 监督学习：从给定的、有限的、用于学习的训练数据（training data）集合出发，假设数据独立同分布，并且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），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在给定的评价准则下有最优的预测；最优模型的选取由算法实现。
- 三要素：模型的假设空间（模型）、模型选择的准则（策略）、模型学习的算法（算法）
- 步骤：

- 1）得到一个有限的训练数据集合
- 2）确定包含所有可能的模型的假设空间，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）利用学习的最优模型对新数据进行预测或分析

# 二、监督学习

## 2.1 定义

监督学习的任务是学习一个模型，使模型能够对任意给定的输入，对其相应的输出做一个好的预测。它从训练数据（training data）集合中学习模型，对测试数据（test data）进行预测。

## 2.2 基本概念

## 2.2.1 输入空间、特征空间与输出空间

- 输入空间与输出空间：输出与输出所有可能值的集合。通常输出空间远远小于输入空间
- 特征空间：所有特征向量存在的空间。特征空间的每一维对应于一个特征。模型都定义在特征空间上。
- 输入实例$x$的特征向量：

$$x = \left(x^{(1)}, x^{(2)}, \cdots, x^{(i)}, \cdots, x^{(n)}\right)^T$$

其中$x^{(i)}$表示$x$的第$i$个特征，$x_i$表示多个输入向量的第$i$个，即

$$x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^T$$

- 训练数据和测试数据由输入输出对（即样本）组成，通常表示为：

$$T = \left\{\left(x_1, y_1\right), \left(x_2, y_2\right), \cdots, \left(x_N, y_N\right)\right\}$$

- 回归问题：输入变量与输出变量均为连续变量的预测问题。
- 分类问题：输出变量为有限个离散变量的预测问题。
- 标注问题：输入变量与输出变量均为变量序列的预测问题。

## 2.2.2 联合概率分布

统计学习假设数据存在一定的统计规律，监督学习的基本假设为$X$和$Y$具有联合概率分布的假设，我们把训练数据与测试数据看作是依联合概率分布$P(X, Y)$独立同分布产生的。

### 2.2.3 假设空间

监督学习的目的在于找到由输入到输出的映射模型集合中最好的一个。这个集合即假设空间。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。

## 三、统计学习三要素

### 3.1 模型

模型就是所要学习的条件概率分布或决策函数。模型的假设空间包含所有可能的条件概率分布（概率模型）或决策函数（非概率模型）。假设空间用F表示，它是由一个参数向量决定的决策函数族：

$$\mathrm{F} = \left\{ f \mid Y = f_\theta(X), \theta \in R^n \right\}$$

参数向量$\theta$取值于$n$维欧式空间$R^n$，称为参数空间。
也可以是一个参数向量决定的条件概率分布族：

$$\mathrm{F} = \left\{ P \mid P_\theta(Y|X), \theta \in R^n \right\}$$

### 3.2 策略

有了模型的假设空间，接下来需要考虑按照什么样的准则学习或选择最优的模型。

### 3.2.1 损失函数和风险函数

**损失函数（loss function）**度量一次预测的好坏。损失函数越小，模型就越好常用的损失函数有：

- 0-1损失函数（0-1 loss function）：

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

- 平方损失函数（quadratic loss function）：

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

- 绝对损失函数（absolute loss function）：

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

- 对数损失函数（logarithmic loss function）：

$$L = -\log P(Y|X)$$

**风险函数（risk function）或期望损失（expected loss）**是理论上模型$f(X)$关于联合分布$P(X, Y)$的平均意义下的损失。

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

我们学习的目标就是选择期望风险最小的模型。由于联合分布$P(X, Y)$未知，风险函数不能直接计算。这样，一方面根据期望风险最小学习模型要用到联合分布，另一方面联合分布又是未知的，所以监督学习就沦为病态问题。
但我们可以计算训练数据集的平均损失，即**经验风险（empirical risk）或经验损失（empirical loss）**：

$$\mathrm{R}_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right)$$

根据大数定理，当样本容量$N$趋于无穷时，经验风险趋于期望风险。自然而然想到可以使用经验风险来估计期望风险。但现实中训练样本数目很小，这种估计往往不理想，需要矫正，以此引出经验风险最小化和结构风险最小化。

## 3.2.2 经验风险最小化和结构风险最小化

**经验风险最小化（empirical risk minimization）**认为经验风险最小的模型是最优的模型，即求解最优化问题：

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

当样本容量足够大的时候，经验风险最小化学习效果良好。比如极大似然估计，当模型是条件概率分布，损失函数是对数损失函数时，经验风险最小化就等价于极大似然估计。

但是当样本容量很小时，经验风险最小化学习会产生过拟合（over-fitting）的现象。这就引出了**结构风险最小化**，它等价于正则化（regularization）。结构风险在经验风险上加上表示模型复杂度的正则化项（regularizer）或罚项（penalty term），它的定义为：

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right) + \lambda J(f)$$

其中$J(f)$为模型的复杂度，模型$f$越复杂，复杂度$J(f)$就越大；反之，模型越简单，复杂度$J(f)$就越小，即复杂度表示了对复杂模型的惩罚。$\lambda \geq 0$是系数，用以权衡经验风险和模型复杂度。结构风险小需要经验风险和模型复杂度同时小。结构风险小的模型往往对训练数据以及未知的测试数据都有较好的预测。比如贝叶斯估计中的最大后验概率估计就是结构风险最小化的一个例子。当模型是条件概率分布、损失函数是对数损失函数、模型复杂度由模型的先验概率表示时，结构风险最小化就等价于最大后验概率估计。
结构风险最小化的策略认为结构风险最小的模型是最优的模型，求解最优模型即求解最优化问题：

$$\min_{f \in \mathrm{F}} \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right) + \lambda J(f)$$

这样，监督学习问题变成了经验风险或结构风险函数的最优化问题。

## 3.3 算法

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，最后需要考虑用什么样的计算方法求解最优化。如何用数值计算求解，如何保证找到全局最优化，并使求解过程高效，是一个重要的问题。

# 四、模型评估与模型选择

## 4.1 训练误差与测试误差

训练误差（training error）是模型关于训练数据集的平均损失：

$$R_{emp}\left(\hat{f}\right) = \frac{1}{N_1}\sum_{i=1}^{N_1} L\left(y_i, \hat{f}\left(x_i\right)\right)$$

测试误差(test error)是模型关于测试数据集的平均损失：

$$R_{emp}\left(\hat{f}\right) = \frac{1}{N_2}\sum_{i=1}^{N_2} L\left(y_i, \hat{f}\left(x_i\right)\right)$$

测试误差反映了学习方法对未知的测试数据集的预测能力，即泛化能力。

## 4.2 过拟合与模型选择

我们希望选择或学习一个合适的模型。若在空间中存在“真模型”，那我们所选择的模型要与真模型的参数个数相同，所选择的模型的参数向量与真模型的参数向量相近。

过拟合指的是我们以为追求提高模型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。即学习时选择的模型所包含的参数过多，以致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，模型选择时，不仅要考虑对已知数据的预测能力，而且还要考虑对未知数据的预测能力。下图描述了训练误差和测试误差与模型的复杂度之间的关系：

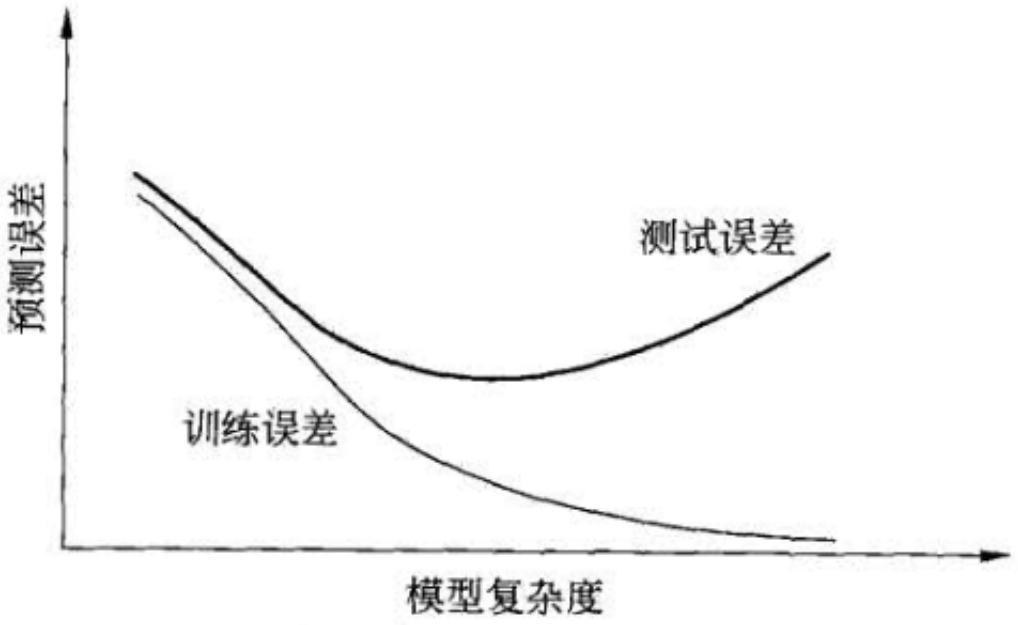


图 1.3　训练误差和测试误差与模型复杂度的关系

当模型复杂度增大时，训练误差会逐渐减小并趋于0；而测试误差会先减小，达到最小值后又增大。当选择的模型复杂度过大时，过拟合现象就会发生。所以要选择复杂度适当的模型，已达到测试误差最小的目的。以此引出正则化与交叉验证。

# 五、正则化与交叉验证

## 5.1 正则化

### 5.1.1 定义

模型选择的典型方法是正则化（regularzation）。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。比如，正则化项可以是模型参数向量的范数。它的一般形式如下：

$$\min_{f\in \mathrm{F}} \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right) + \lambda J(f)$$

第一项是经验风险，第二项是正则化项，$\lambda \geq 0$为调整两者之间关系的系数。

## 5.1.2 不同形式

正则化项可以取不同的形式。例如，回归问题中，损失函数是平方误差，正则化项可以是参数向量的$L_2$范数:

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

也可以是参数向量的$L_1$范数：

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\left(f\left(x_i; w\right) - y_i\right)^2 + \lambda||w||_1$$

第一项的经验风险较小的模型可能较复杂（有多个非零参数），这时第二项的模型复杂度会较大。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。

## 5.1.3 奥卡姆剃刀

正则化符合奥卡姆剃刀原理，应用于模型选择时变为：在所有可能选择的模型中，能够很好地解释已知数据并且十分简单才是最好的模型。从贝叶斯估计的角度来看，正则化项对应于模型的先验概率。可以假设复杂的模型有很小的先验概率，简单的模型有较大的先验概率。

# 5.2 交叉验证

## 5.2.1 定义

如果给定的样本数据充足，进行模型选择的一种简单方法是随机地将数据集切分成三部分，分别是训练集（training set）用来训练模型、验证集（validation set）用于模型的选择、测试集（test set）用于最终对学习方法的评估，最终选择对验证集有最小预测误差的模型。

但是实际应用中数据不充足，所以我们采用交叉验证，它的基本思想是重复的使用数据，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、测试和

模型选择。

### 5.2.2 方法

- 简单交叉验证：首先随机地将已给数据分为两个部分，一部分作为训练集（70%），另一部分作为测试集（30%）；然后用训练集在各种条件下（如不同的参数个数）训练模型，从而得到不同的模型；在测试机上评价各个模型的测试误差，选出测试误差最小的模型。
- S折交叉验证（S-fold cross validation）：应用最广泛。首先随即将已给数据切分为S个互不相交的大小相同的子集；然后利用S-1个子集的数据训练模型，利用余下的子集测试模型；将这一过程对可能的S种选择重复进行；最后选出S次评测中平均测试误差最小的模型。
- 留一交叉验证leave-one-out cross validation）：S折交叉验证的特殊情形是S=N（N为给定数据集的容量），往往在数据缺乏的情况下使用

# 六、泛化能力

## 6.1 泛化误差

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。现实中常常通过测试误差来评价学习方法的泛化能力，但因为测试数据及有限，评价结果不一定可靠。

理论上，通过泛化误差来反映学习方法的泛化能力，泛化误差即用学习到的模型对未知数据预测的误差：

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

泛化误差越小，模型效果就好。泛化误差就是所学习到的模型的期望风险。

# 七、生成模型与判别模型

## 7.1 判别模型

判别模型由数据直接学习决策函数$f(X)$或者条件概率分布$P(Y|X)$作为预测的模型。它关心的是对给定的输入$X$，应该预测什么样的输出$Y$。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、最大熵模型、支持向量机、提升方法、条件随机场。

判别方法的特点：

- 直接学习的是条件概率$P(Y|X)$或决策函数$f(X)$，直接面对预测，往往学习的准确率很高；
- 由于直接学习$P(Y|X)$或$f(X)$，可以对数据进行各种程度上的抽象、定义特征并使用特征，因

此可以简化学习问题。

## 7.2 生成模型

生成模型由数据学习联合概率分布$P(X, Y)$，然后求出条件概率分布$P(Y|X)$作为预测的模型：

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

因为模型表示了给定输入$X$产生输出$Y$的生成关系，所以被称为生成模型。典型的生成模型有：朴素贝叶斯、隐马尔科夫模型

生成方法的特点：

- 生成方法可以还原出联合概率分布$P(X, Y)$，而判别方法不能；
- 生成方法的学习收敛速度快，即当样本容量增加时，学到的模型可以很快收敛于真实模型；
- 当存在隐变量时，仍可以用生成方法学习，此时判别方法不能用。

# 八、分类问题

## 8.1 定义

在监督学习中，当输出变量$Y$取有限个离散值时，预测问题便成为分类问题。它从数据中学习一个分类模型或分类决策函数，即学习一个分类器，然后对新的输入进行输出的预测，即进行分类。分为多类分类和二类分类问题。

## 8.2 学习过程

如图所示，分类问题包括学习和分类两个过程。在学习过程中，根据已知的训练数据集利用有效的学习方法学习一个分类器；在分类过程中，利用学习的分类器对新的输入实例进行分类。

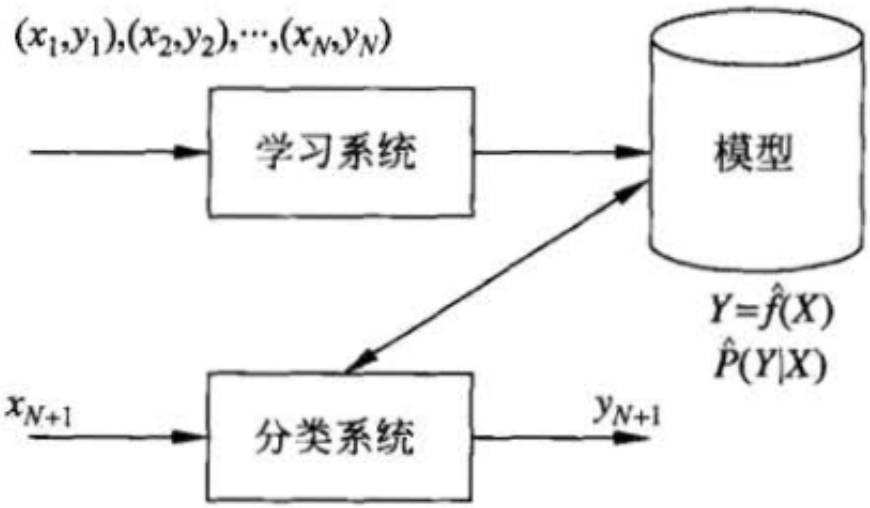



图 1.4　分类问题

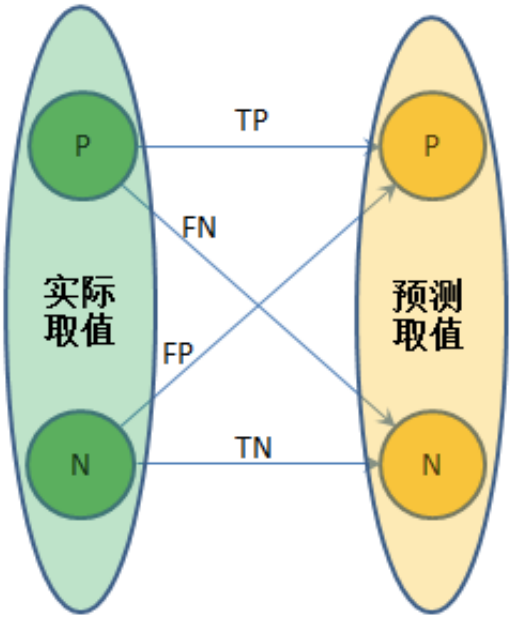
## 8.3 分类准确率

## 8.3.1 定义

评价分类器性能的指标一般是分类准确率（accuracy）,即对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。也即损失函数是0-1损失时测试数据集上的准确率。

## 8.3.2 常用指标

通常将关注的类为正类，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（正类预测为正类数）、FN（正类预测为负类书）、FP（负类预测为正类数）、TN（负类预测为负类数）



- 精确率：

$$P = \frac{TP}{TP + FP}$$

- 召回率：

$$R = \frac{TP}{TP + FN}$$

- $F_1$值，是精确率和召回率的调和均值，即

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

  精确率与召回率都很高时，$F_1$值也会很高。

### 8.3.3 分类方法与应用

- 常见分类统计方法：K近邻、感知机、朴素贝叶斯、决策树、决策列表、逻辑斯谛回归、支持向量机、提升、贝叶斯网络、神经网络Winnow
- 应用：银行业务构建客户分类模型，对客户按照贷款风险大小分类；网络非法入侵检测；人脸是否出现的检测；网页分类；文本分类等。

## 九、标注问题

### 9.1 定义

标注问题的输入是一个观测序列，输出是一个标记序列或状态序列。它的目的在于学习一个模型，使它能够对观测序列给出标记序列作为预测。标注问题分为学习和标注两个过程：

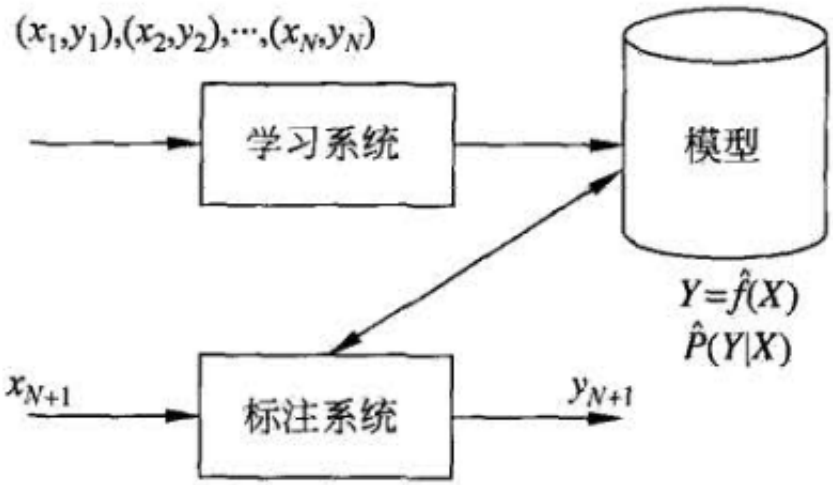



图 1.5　标注问题

## 9.2 应用

标注常用的统计学习方法有：隐马尔科夫模型、条件随机场

它在信息抽取、自然语言处理领域被广泛应用。

- 自然语言处理的词性标注：给定一个由单词组成的句子，对这个句子中的每一个单词进行词性标注，即对一个单词序列预测其对应的词性标记序列。

# 十、回归问题

## 10.1 定义

回归模型表示输入变量和输出变量之间映射的函数，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地预测未知数据。可分为一元回归和多元回归，线性回归和非线性回归。它最常用的损失函数为平方损失函数，可以用最小二乘法求解。回归问题分为学习和标注两个过程：

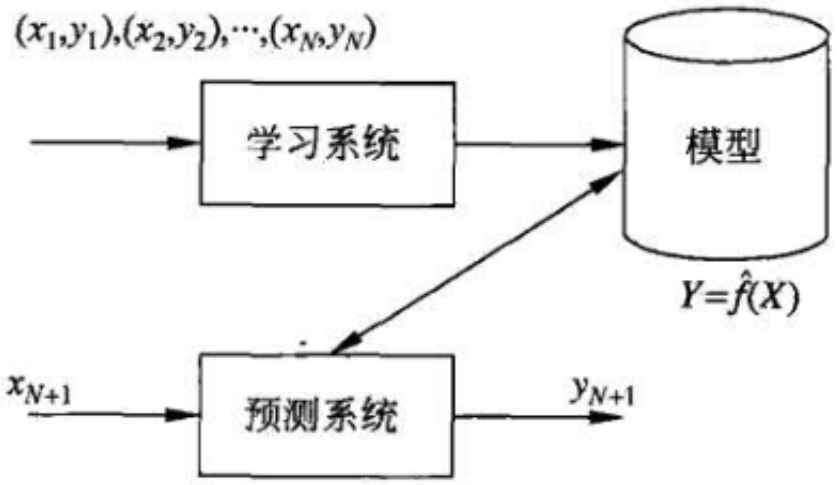



图 1.6　回归问题

## 10.2 应用

股价预测：将影响股价的信息视作自变量，将股价视为因变量，将过去的数据作为训练数据，学习一个回归模型，并对未来的股价进行预测。