第1篇　监督学习

第 1 章　统计学习及监督学习概论

本书第 1 篇讲述监督学习方法。监督学习是从标注数据中学习模型的机器学习问题，是统计学习或机器学习的重要组成部分。

本章简要叙述统计学习及监督学习的一些基本概念。使读者对统计学习及监督学习有初步了解。

本章 1.1 节叙述统计学习或机器学习的定义、研究对象与方法；1.2 节叙述统计学习的分类，基本分类是监督学习、无监督学习、强化学习；1.3 节叙述统计学习方法的三要素：模型、策略和算法；1.4 节至 1.7 节相继介绍监督学习的几个重要概念，包括模型评估与模型选择、正则化与交叉验证、学习的泛化能力、生成模型与判别模型；最后 1.8 节介绍监督学习的应用：分类问题，标注问题与回归问题。

1.1　统 计 学 习

1. 统计学习的特点

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预测与分析的一门学科。统计学习也称为统计机器学习 (statistical machine learning）。

统计学习的主要特点是：(1) 统计学习以计算机及网络为平台，是建立在计算机及网络上的；(2) 统计学习以数据为研究对象，是数据驱动的学科；(3) 统计学习的目的是对数据进行预测与分析；(4) 统计学习以方法为中心，统计学习方法构建模型并应用模型进行预测与分析；(5) 统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科，并且在发展中逐步形成独自的理论体系与方法论。

赫尔伯特 • 西蒙（Herbert A. Simon）曾对“学习”给出以下定义：“如果一个系统能够通过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运用数据及统计方法提高系统性能的机器学习。现在，当人们提及机器学习时，往往是指统计机器学习。所以可以认为本书介绍的是机器学习方法。

2. 统计学习的对象

统计学习研究的对象是数据（data）。它从数据出发，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。作为统计学习的对象，数据是多样的，包括存在于计算机及网络上的各种数字、文字、图像、视频、音频数据以及它们的组合。

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。这里的同类数据是指具有某种共同性质的数据，例如英文文章、互联网网页、数据库中的数据等。由于它们具有统计规律性，所以可以用概率统计方法处理它们。比如，可以用随机变量描述数据中的特征，用概率分布描述数据的统计规律。在统计学习中，以变量或变量组表示数据。数据分为由连续变量和离散变量表示的类型。本书以讨论离散变量的方法为主。另外，本书只涉及利用数据构建模型及利用模型对数据进行分析与预测，对数据的观测和收集等问题不作讨论。

3. 统计学习的目的

统计学习用于对数据的预测与分析，特别是对未知新数据的预测与分析。对数据的预测可以使计算机更加智能化，或者说使计算机的某些性能得到提高；对数据的分析可以让人们获取新的知识，给人们带来新的发现。

对数据的预测与分析是通过构建概率统计模型实现的。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与分析，同时也要考虑尽可能地提高学习效率。

4. 统计学习的方法

统计学习的方法是基于数据构建概率统计模型从而对数据进行预测与分析。统计学习由监督学习（supervised learning）、无监督学习（unsupervised learning）和强化学习（reinforcement learning）等组成。

本书第 1 篇讲述监督学习，第 2 篇讲述无监督学习。可以说监督学习、无监督学习方法是最主要的统计学习方法。

统计学习方法可以概括如下：从给定的、有限的、用于学习的训练数据（training data）集合出发，假设数据是独立同分布产生的；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模型，使它对已知的训练数据及未知的测试数据（test data）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。这样，统计学习方法包括模型的假设空间、模型选择的准则以及模型学习的算法。称其为统计学习方法的三要素，简称为模型（model）、策略（strategy）和算法（algorithm）。

实现统计学习方法的步骤如下：

（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。

本书第 1 篇介绍监督学习方法，主要包括用于分类、标注与回归问题的方法。这些方法在自然语言处理、信息检索、文本数据挖掘等领域中有着极其广泛的应用。

5. 统计学习的研究

统计学习研究一般包括统计学习方法、统计学习理论及统计学习应用三个方面。统计学习方法的研究旨在开发新的学习方法；统计学习理论的研究在于探求统计学习方法的有效性与效率，以及统计学习的基本理论问题；统计学习应用的研究主要考虑将统计学习方法应用到实际问题中去，解决实际问题。

6. 统计学习的重要性

近二十年来，统计学习无论是在理论还是在应用方面都得到了巨大的发展，有许多重大突破，统计学习已被成功地应用到人工智能、模式识别、数据挖掘、自然语言处理、语音处理、计算视觉、信息检索、生物信息等许多计算机应用领域中，并且成为这些领域的核心技术。人们确信，统计学习将会在今后的科学发展和技术应用中发挥越来越大的作用。

统计学习学科在科学技术中的重要性主要体现在以下几个方面：

（1）统计学习是处理海量数据的有效方法。我们处于一个信息爆炸的时代，海量数据的处理与利用是人们必然的需求。现实中的数据不但规模大，而且常常具有不确定性，统计学习往往是处理这类数据最强有力的工具。

（2）统计学习是计算机智能化的有效手段。智能化是计算机发展的必然趋势，也是计算机技术研究与开发的主要目标。近几十年来，人工智能等领域的研究证明，利用统计学习模仿人类智能的方法，虽有一定的局限性，还是实现这一目标的最有效手段。

（3）统计学习是计算机科学发展的一个重要组成部分。可以认为计算机科学由三维组成：系统、计算、信息。统计学习主要属于信息这一维，并在其中起着核心作用。

1.2 统计学习的分类

统计学习或机器学习是一个范围宽阔、内容繁多、应用广泛的领域，并不存在（至少现在不存在）一个统一的理论体系涵盖所有内容。下面从几个角度对统计学习方法进行分类。

1.2.1 基本分类

统计学习或机器学习一般包括监督学习、无监督学习、强化学习。有时还包括半监督学习、主动学习。

1. 监督学习

监督学习（supervised learning）是指从标注数据中学习预测模型的机器学习问题。标注数据表示输入输出的对应关系，预测模型对给定的输入产生相应的输出。监督学习的本质是学习输入到输出的映射的统计规律。

（1）输入空间、特征空间和输出空间

在监督学习中，将输入与输出所有可能取值的集合分别称为输入空间（input space）与输出空间（output space）。输入与输出空间可以是有限元素的集合，也可以是整个欧氏空间。输入空间与输出空间可以是同一个空间，也可以是不同的空间；但通常输出空间远远小于输入空间。

每个具体的输入是一个实例（instance），通常由特征向量（feature vector）表示。这时，所有特征向量存在的空间称为特征空间（feature space）。特征空间的每一维对应于一个特征。有时假设输入空间与特征空间为相同的空间，对它们不予区分；有时假设输入空间与特征空间为不同的空间，将实例从输入空间映射到特征空间。模型实际上都是定义在特征空间上的。

在监督学习中，将输入与输出看作是定义在输入（特征）空间与输出空间上的随机变量的取值。输入输出变量用大写字母表示，习惯上输入变量写作 X，输出变量写作 Y。输入输出变量的取值用小写字母表示，输入变量的取值写作 x，输出变量的取值写作 y。变量可以是标量或向量，都用相同类型字母表示。除特别声明外，本书中向量均为列向量。输入实例 x 的特征向量记作

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

$x^{(i)}$ 表示 x 的第 i 个特征。注意 $x^{(i)}$ 与 x_i 不同，本书通常用 x_i 表示多个输入变量中的第 i 个变量，即

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

监督学习从训练数据（training data）集合中学习模型，对测试数据（test data）进行预测。训练数据由输入（或特征向量）与输出对组成，训练集通常表示为

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

测试数据也由输入与输出对组成。输入与输出对又称为样本（sample）或样本点。

输入变量 X 和输出变量 Y 有不同的类型，可以是连续的，也可以是离散的。人们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；输出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。

（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 $P(X,Y)$。$P(X,Y)$ 表示分布函数，或分布密度函数。注意在学习过程中，假定这一联合概率分布存在，但对学习系统来说，联合概率分布的具体定义是未知的。训练数据与测试数据被看作是依联合概率分布 $P(X,Y)$ 独立同分布产生的。统计学习假设数据存在一定的统计规律，X 和 Y 具有联合概率分布就是监督学习关于数据的基本假设。

（3）假设空间

监督学习的目的在于学习一个由输入到输出的映射，这一映射由模型来表示。换句话说，学习的目的就在于找到最好的这样的模型。模型属于由输入空间到输出空间的映射的集合，这个集合就是假设空间（hypothesis space）。假设空间的确定意味着学习的范围的确定。

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（decision function）$Y=f(X)$ 表示，随具体学习方法而定。对具体的输入进行相应的输出预测时，写作 $P(y|x)$ 或 $y=f(x)$。

（4）问题的形式化

监督学习利用训练数据集学习一个模型，再用模型对测试样本集进行预测。由于在这个过程中需要标注的训练数据集，而标注的训练数据集往往是人工给出的，所以称为监督学习。监督学习分为学习和预测两个过程，由学习系统与预测系统完成，可用图 1.1 来描述。

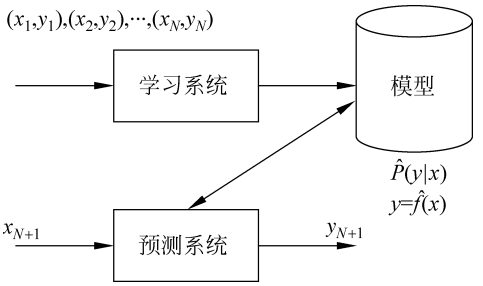

图 1.1 监督学习

首先给定一个训练数据集

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

其中 (x_i, y_i)，$i = 1, 2, \cdots, N$，称为样本或样本点。$x_i \in \mathcal{X} \subseteq \mathbf{R}^n$ 是输入的观测值，也称为输入或实例，$y_i \in \mathcal{Y}$ 是输出的观测值，也称为输出。

监督学习分为学习和预测两个过程，由学习系统与预测系统完成。在学习过程中，学习系统利用给定的训练数据集，通过学习（或训练）得到一个模型，表示为条件概率分布 $\hat{P}(Y|X)$ 或决策函数 $Y = \hat{f}(X)$。条件概率分布 $\hat{P}(Y|X)$ 或决策函数 $Y = \hat{f}(X)$ 描述输入与输出随机变量之间的映射关系。在预测过程中，预测系统对于给定的测试样本集中的输入 x_{N+1}，由模型 $y_{N+1} = \arg\max\limits_{y} \hat{P}(y|x_{N+1})$ 或 $y_{N+1} = \hat{f}(x_{N+1})$ 给出相应的输出 y_{N+1}。

在监督学习中，假设训练数据与测试数据是依联合概率分布 $P(X, Y)$ 独立同分布产生的。

学习系统（也就是学习算法）试图通过训练数据集中的样本 (x_i, y_i) 带来的信息学习模型。具体地说，对输入 x_i，一个具体的模型 $y = f(x)$ 可以产生一个输出 $f(x_i)$，而训练数据集中对应的输出是 y_i。如果这个模型有很好的预测能力，训练样本输出 y_i 和模型输出 $f(x_i)$ 之间的差就应该足够小。学习系统通过不断地尝试，选取最好的模型，以便对训练数据集有足够好的预测，同时对未知的测试数据集的预测也有尽可能好的推广。

2. 无监督学习

无监督学习[①]（unsupervised learning）是指从无标注数据中学习预测模型的机器学习问题。无标注数据是自然得到的数据，预测模型表示数据的类别、转换或概率。无监督学习的本质是学习数据中的统计规律或潜在结构。

模型的输入与输出的所有可能取值的集合分别称为输入空间与输出空间。输入空间与输出空间可以是有限元素集合，也可以是欧氏空间。每个输入是一个实例，由特征向量表示。每一个输出是对输入的分析结果，由输入的类别、转换或概率表示。模型可以实现对数据的聚类、降维或概率估计。

假设 $\mathcal{X}$ 是输入空间，$\mathcal{Z}$ 是隐式结构空间。要学习的模型可以表示为函数 $z = g(x)$，条件概率分布 $P(z|x)$，或者条件概率分布 $P(x|z)$ 的形式，其中 $x \in \mathcal{X}$ 是输入，$z \in \mathcal{Z}$ 是输出。包含所有可能的模型的集合称为假设空间。无监督学习旨在从假设空间中选出在给定评价标准下的最优模型。

无监督学习通常使用大量的无标注数据学习或训练，每一个样本是一个实例。训练数据表示为 $U = \{x_1, x_2, \cdots, x_N\}$，其中 x_i，$i = 1, 2, \cdots, N$，是样本。

无监督学习可以用于对已有数据的分析，也可以用于对未来数据的预测。分析时使用学习得到的模型，即函数 $z = \hat{g}(x)$，条件概率分布 $\hat{P}(z|x)$，或者条件概率分布 $\hat{P}(x|z)$。预测时，和监督学习有类似的流程。由学习系统与预测系统完成，如

① 也译作非监督学习。

图 1.2 所示。在学习过程中，学习系统从训练数据集学习，得到一个最优模型，表示为函数 $z = \hat{g}(x)$，条件概率分布 $\hat{P}(z|x)$ 或者条件概率分布 $\hat{P}(x|z)$。在预测过程中，预测系统对于给定的输入 x_{N+1}，由模型 $z_{N+1} = \hat{g}(x_{N+1})$ 或 $z_{N+1} = \arg\max_z \hat{P}(z|x_{N+1})$ 给出相应的输出 z_{N+1}，进行聚类或降维，或者由模型 $\hat{P}(x|z)$ 给出输入的概率 $\hat{P}(x_{N+1}|z_{N+1})$，进行概率估计。

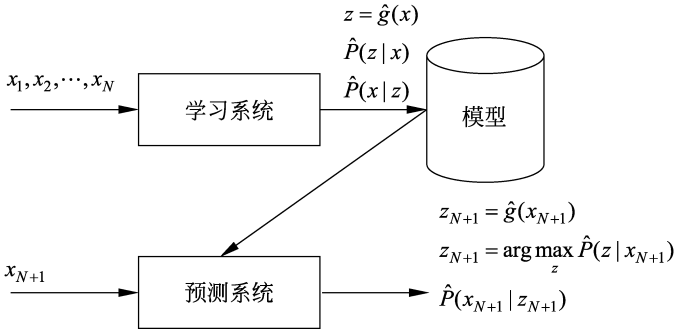

图 1.2 无监督学习

3. 强化学习

强化学习（reinforcement learning）是指智能系统在与环境的连续互动中学习最优行为策略的机器学习问题。假设智能系统与环境的互动基于马尔可夫决策过程（Markov decision process），智能系统能观测到的是与环境互动得到的数据序列。强化学习的本质是学习最优的序贯决策。

智能系统与环境的互动如图 1.3 所示。在每一步 t，智能系统从环境中观测到一个状态（state）s_t 与一个奖励（reward）r_t，采取一个动作（action）a_t。环境根据智能系统选择的动作，决定下一步 $t+1$ 的状态 s_{t+1} 与奖励 r_{t+1}。要学习的策略表示为给定的状态下采取的动作。智能系统的目标不是短期奖励的最大化，而是长期累积奖励的最大化。强化学习过程中，系统不断地试错（trial and error），以达到学习最优策略的目的。

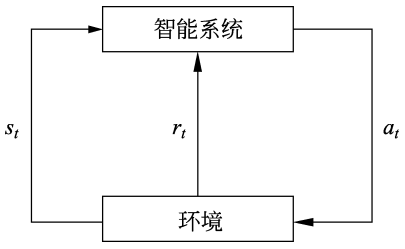

图 1.3 智能系统与环境的互动

强化学习的马尔可夫决策过程是状态、奖励、动作序列上的随机过程，由五元组 $\langle S, A, P, r, \gamma \rangle$ 组成。

- S 是有限状态（state）的集合
- A 是有限动作（action）的集合
- P 是状态转移概率（transition probability）函数:

$$P(s'|s,a) = P(s_{t+1} = s'|s_t = s, a_t = a)$$

- r 是奖励函数（reward function）: $r(s,a) = E(r_{t+1}|s_t = s, a_t = a)$
- γ 是衰减系数（discount factor）: $\gamma \in [0,1]$

马尔可夫决策过程具有马尔可夫性，下一个状态只依赖于前一个状态与动作，由状态转移概率函数 $P(s'|s,a)$ 表示。下一个奖励依赖于前一个状态与动作，由奖励函数 $r(s,a)$ 表示。

策略 π 定义为给定状态下动作的函数 $a = f(s)$ 或者条件概率分布 $P(a|s)$。给定一个策略 π，智能系统与环境互动的行为就已确定（或者是确定性的或者是随机性的）。

价值函数（value function）或状态价值函数（state value function）定义为策略 π 从某一个状态 s 开始的长期累积奖励的数学期望：

$$v_\pi(s) = E_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s] \tag{1.1}$$

动作价值函数（action value function）定义为策略 π 的从某一个状态 s 和动作 a 开始的长期累积奖励的数学期望：

$$q_\pi(s,a) = E_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s, a_t = a] \tag{1.2}$$

强化学习的目标就是在所有可能的策略中选出价值函数最大的策略 π^*，而在实际学习中往往从具体的策略出发，不断优化已有策略。这里 γ 表示未来的奖励会有衰减。

强化学习方法中有基于策略的（policy-based）、基于价值的（value-based），这两者属于无模型的（model-free）方法，还有有模型的（model-based）方法。

有模型的方法试图直接学习马尔可夫决策过程的模型，包括转移概率函数 $P(s'|s,a)$ 和奖励函数 $r(s,a)$。这样可以通过模型对环境的反馈进行预测，求出价值函数最大的策略 π^*。

无模型的、基于策略的方法不直接学习模型，而是试图求解最优策略 π^*，表示为函数 $a = f^*(s)$ 或者是条件概率分布 $P^*(a|s)$，这样也能达到在环境中做出最优决策的