

第1章 统计学习方法概论

本章简要叙述统计学习方法的一些基本概念。这是对全书内容的概括，也是全书内容的基础。首先叙述统计学习的定义、研究对象与方法；然后叙述监督学习，这是本书的主要内容；接着提出统计学习方法的三要素：模型、策略和算法；介绍模型选择，包括正则化、交叉验证与学习的泛化能力；介绍生成模型与判别模型；最后介绍监督学习方法的应用：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之上的；(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。

2. 统计学习的对象

统计学习的对象是数据 (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

方法来加以处理。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。

在统计学习过程中，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进行预测与分析，特别是对未知新数据进行预测与分析。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)、非监督学习 (unsupervised learning)、半监督学习 (semi-supervised learning) 和强化学习 (reinforcement learning) 等组成。

本书主要讨论监督学习，这种情况下统计学习的方法可以概括如下：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。

本书以介绍统计学习方法为主，特别是监督学习方法，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 (statistical learning method)、统计学习理论 (statistical learning theory) 及统计学习应用 (application of statistical learning) 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 with 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 20 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、语音识别、图像识别、信息检索和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，也是计算机技术与开发的主要目标。近几十年来，人工智能等领域的研究表明，利用统计学习模仿人类智能的方法，虽有一定的局限性，但仍然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 监督学习

统计学习包括监督学习、非监督学习、半监督学习及强化学习。本书主要讨论监督学习问题。

监督学习 (supervised learning) 的任务是学习一个模型，使模型能够对任意给定的输入，对其相应的输出做出一个好的预测 (注意，这里的输入、输出是指某个系统的输入与输出，与学习的输入与输出不同)。计算机的基本操作就是给定一个输入产生一个输出，所以监督学习是极其重要的统计学习分支，也是统计学习中内容最丰富、应用最广泛的部分。

1.2.1 基本概念

1. 输入空间、特征空间与输出空间

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.

在监督学习过程中, 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、输出变量所取的值用小写字母表示, 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

$$x = (x^{(1)}, x^{(2)}, \dots, x^{(i)}, \dots, x^{(n)})^T$$

$x^{(i)}$ 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, 即

$$x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})^T$$

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

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

测试数据也由相应的输入与输出对组成. 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 具有联合概率分布的假设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 $y = f(x)$.

1.2.2 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 (prediction). 由于在这个过程中需要训练数据集, 而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.

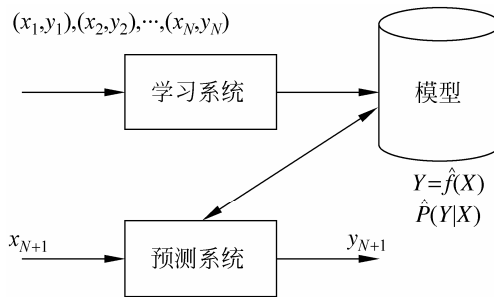


图 1.1 监督学习问题

首先给定一个训练数据集

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

其中 (x_i, y_i) , $i = 1, 2, \dots, N$, 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.

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 $y_{N+1} = \arg \max_{y_{N+1}} \hat{P}(y_{N+1} | x_{N+1})$ 或 $y_{N+1} = \hat{f}(x_{N+1})$ 给出相应的输出 y_{n+1} .

在学习过程中, 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. 学习系统通过不断的尝试, 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.

1.3 统计学习三要素

统计学习方法都是由模型、策略和算法构成的, 即统计学习方法由三要素构成, 可以简单地表示为

$$\text{方法} = \text{模型} + \text{策略} + \text{算法}$$

下面论述监督学习中的统计学习三要素. 非监督学习、强化学习也同样拥有这三要素. 可以说构建一种统计学习方法就是确定具体的统计学习三要素.

1.3.1 模型

统计学习首要考虑的问题是学习什么样的模型. 在监督学习过程中, 模型就是所要学习的条件概率分布或决策函数. 模型的假设空间 (hypothesis space) 包含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, 那么模型的假设空间就是所有这些线性函数构成的函数集合. 假设空间中的模型一般有无穷多个.

假设空间用 \mathcal{F} 表示. 假设空间可以定义为决策函数的集合

$$\mathcal{F} = \{f | Y = f(X)\} \quad (1.1)$$

其中, X 和 Y 是定义在输入空间 \mathcal{X} 和输出空间 \mathcal{Y} 上的变量. 这时 \mathcal{F} 通常是由一个参数向量决定的函数族:

$$\mathcal{F} = \{f | Y = f_{\theta}(X), \theta \in \mathbf{R}^n\} \quad (1.2)$$

参数向量 θ 取值于 n 维欧氏空间 \mathbf{R}^n , 称为参数空间 (parameter space).

假设空间也可以定义为条件概率的集合

$$\mathcal{F} = \{P | P(Y | X)\} \quad (1.3)$$

其中, X 和 Y 是定义在输入空间 \mathcal{X} 和输出空间 \mathcal{Y} 上的随机变量. 这时 \mathcal{F} 通常是由一个参数向量决定的条件概率分布族:

$$\mathcal{F} = \{P | P_{\theta}(Y | X), \theta \in \mathbf{R}^n\} \quad (1.4)$$

参数向量 θ 取值于 n 维欧氏空间 \mathbf{R}^n , 也称为参数空间.

本书中称由决策函数表示的模型为非概率模型, 由条件概率表示的模型为概率模型. 为了简便起见, 当论及模型时, 有时只用其中一种模型.

1.3.2 策略

有了模型的假设空间, 统计学习接着需要考虑的是按照什么样的准则学习或选择最优的模型. 统计学习的目标在于从假设空间中选取最优模型.

首先引入损失函数与风险函数的概念. 损失函数度量模型一次预测的好坏, 风险函数度量平均意义下模型预测的好坏.

1. 损失函数和风险函数

监督学习问题是在假设空间 \mathcal{F} 中选取模型 f 作为决策函数, 对于给定的输入 X , 由 $f(X)$ 给出相应的输出 Y , 这个输出的预测值 $f(X)$ 与真实值 Y 可能一致也可能不一致, 用一个损失函数 (loss function) 或代价函数 (cost function) 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 $L(Y, f(X))$.

统计学习常用的损失函数有以下几种:

(1) 0-1 损失函数 (0-1 loss function)

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases} \quad (1.5)$$

(2) 平方损失函数 (quadratic loss function)

$$L(Y, f(X)) = (Y - f(X))^2 \quad (1.6)$$

(3) 绝对损失函数 (absolute loss function)

$$L(Y, f(X)) = |Y - f(X)| \quad (1.7)$$

(4) 对数损失函数 (logarithmic loss function) 或对数似然损失函数 (log-likelihood loss function)

$$L(Y, P(Y | X)) = -\log P(Y | X) \quad (1.8)$$

损失函数值越小, 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

$$R_{\text{exp}}(f) = E_P[L(Y, f(X))] = \int_{X \times Y} L(y, f(x)) P(x, y) dx dy \quad (1.9)$$

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

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, $R_{\text{exp}}(f)$ 不能直接计算. 实际上, 如果知道联合分布 $P(X, Y)$, 可以从联合分布直接求出条件概率分布 $P(Y|X)$, 也就不需要学习了. 正因为不知道联合概率分布, 所以才需要进行学习. 这样一来, 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, 所以监督学习就成为一个病态问题 (ill-formed problem).

给定一个训练数据集

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

模型 $f(X)$ 关于训练数据集的平均损失称为经验风险 (empirical risk) 或经验损失 (empirical loss), 记作 R_{emp} :

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) \quad (1.10)$$

期望风险 $R_{\text{exp}}(f)$ 是模型关于联合分布的期望损失, 经验风险 $R_{\text{emp}}(f)$ 是模型关于训练样本集的平均损失. 根据大数定律, 当样本容量 N 趋于无穷时, 经验风险 $R_{\text{emp}}(f)$ 趋于期望风险 $R_{\text{exp}}(f)$. 所以一个很自然的想法是用经验风险估计期望风险. 但是, 由于现实中训练样本数目有限, 甚至很小, 所以用经验风险估计期望风险常常并不理想, 要对经验风险进行一定的矫正. 这就关系到监督学习的两个基本策略: 经验风险最小化和结构风险最小化.

2. 经验风险最小化与结构风险最小化

在假设空间、损失函数以及训练数据集确定的情况下, 经验风险函数式(1.10)就可以确定. 经验风险最小化 (empirical risk minimization, ERM) 的策略认为, 经验风险最小的模型是最优的模型. 根据这一策略, 按照经验风险最小化求最优模型就是求解最优化问题:

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) \quad (1.11)$$

其中, \mathcal{F} 是假设空间.

当样本容量足够大时, 经验风险最小化能保证有很好的学习效果, 在现实中

被广泛采用。比如，极大似然估计（maximum likelihood estimation）就是经验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的效果就未必很好，会产生后面将要叙述的“过拟合(over-fitting)”现象。

结构风险最小化（structural risk minimization, SRM）是为了防止过拟合而提出来的策略。结构风险最小化等价于正则化（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）。在假设空间、损失函数以及训练数据集确定的情况下，结构风险的定义是

$$R_{\text{sm}}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f) \quad (1.12)$$

其中 $J(f)$ 为模型的复杂度，是定义在假设空间 \mathcal{F} 上的泛函。模型 f 越复杂，复杂度 $J(f)$ 就越大；反之，模型 f 越简单，复杂度 $J(f)$ 就越小。也就是说，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与模型复杂度同时小。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（maximum posterior probability estimation, MAP）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。所以求最优模型，就是求解最优化问题：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f) \quad (1.13)$$

这样，监督学习问题就变成了经验风险或结构风险函数的最优化问题(1.11)和(1.13)。这时经验或结构风险函数是最优化的目标函数。

1.3.3 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，最后需要考虑用什么样的计算方法求解最优模型。

这时，统计学习问题归结为最优化问题，统计学习的算法成为求解最优化问题的算法。如果最优化问题有显式的解析解，这个最优化问题就比较简单。但通常解析解不存在，这就需要用数值计算的方法求解。如何保证找到全局最优解，并使求解的过程非常高效，就成为一个重要问题。统计学习可以利用已有的最优

化算法,有时也需要开发独自的最优化算法.

统计学习方法之间的不同,主要来自其模型、策略、算法的不同.确定了模型、策略、算法,统计学习的方法也就确定了.这也就是将其称为统计学习三要素的原因.

1.4 模型评估与模型选择

1.4.1 训练误差与测试误差

统计学习的目的是使学到的模型不仅对已知数据而且对未知数据都能有很好的预测能力.不同的学习方法会给出不同的模型.当损失函数给定时,基于损失函数的模型的训练误差(training error)和模型的测试误差(test error)就自然成为学习方法评估的标准.注意,统计学习方法具体采用的损失函数未必是评估时使用的损失函数.当然,让两者一致是比较理想的.

假设学习到的模型是 $Y = \hat{f}(X)$,训练误差是模型 $Y = \hat{f}(X)$ 关于训练数据集的平均损失:

$$R_{\text{emp}}(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i)) \quad (1.14)$$

其中 N 是训练样本容量.

测试误差是模型 $Y = \hat{f}(X)$ 关于测试数据集的平均损失:

$$e_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} L(y_i, \hat{f}(x_i)) \quad (1.15)$$

其中 N' 是测试样本容量.

例如,当损失函数是0-1损失时,测试误差就变成了常见的测试数据集上的误差率(error rate)

$$e_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} I(y_i \neq \hat{f}(x_i)) \quad (1.16)$$

这里 I 是指示函数(indicator function),即 $y \neq \hat{f}(x)$ 时为1,否则为0.

相应地,常见的测试数据集上的准确率(accuracy)为

$$r_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} I(y_i = \hat{f}(x_i)) \quad (1.17)$$

显然,

$$r_{\text{test}} + e_{\text{test}} = 1$$

训练误差的大小,对判断给定的问题是不是一个容易学习的问题是有意义的,但本质上不重要.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,是学习中的重要概念.显然,给定两种学习方法,测试误差小的方法具有更好的

预测能力,是更有效的方法.通常将学习方法对未知数据的预测能力称为泛化能力 (generalization ability), 这个问题将在 1.6 节继续论述.

1.4.2 过拟合与模型选择

当假设空间含有不同复杂度 (例如, 不同的参数个数) 的模型时, 就要面临模型选择 (model selection) 的问题. 我们希望选择或学习一个合适的模型. 如果在假设空间中存在“真”模型, 那么所选择的模型应该逼近真模型. 具体地, 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.

如果一味追求提高对训练数据的预测能力, 所选模型的复杂度则往往会比真模型更高. 这种现象称为过拟合 (over-fitting). 过拟合是指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. 可以说模型选择旨在避免过拟合并提高模型的预测能力.

下面, 以多项式函数拟合问题为例, 说明过拟合与模型选择. 这是一个回归问题.

例 1.1 假设给定一个训练数据集^①:

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

其中, $x_i \in \mathbf{R}$ 是输入 x 的观测值, $y_i \in \mathbf{R}$ 是相应的输出 y 的观测值, $i = 1, 2, \dots, N$. 多项式函数拟合的任务是假设给定数据由 M 次多项式函数生成, 选择最有可能产生这些数据的 M 次多项式函数, 即在 M 次多项式函数中选择一个对已知数据以及未知数据都有很好预测能力的函数.

假设给定如图 1.2 所示的 10 个数据点, 用 0~9 次多项式函数对数据进行拟合. 图中画出了需要用多项式函数曲线拟合的数据.

设 M 次多项式为

$$f_M(x, w) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

式中 x 是单变量输入, w_0, w_1, \dots, w_M 是 $M+1$ 个参数.

解决这一问题的方法可以是这样的. 首先确定模型的复杂度, 即确定多项式的次数; 然后在给定的模型复杂度下, 按照经验风险最小化的策略, 求解参数, 即多项式的系数, 具体地, 求以下经验风险最小化:

$$L(w) = \frac{1}{2} \sum_{i=1}^N (f(x_i, w) - y_i)^2 \quad (1.18)$$

^① 本例来自参考文献[3].

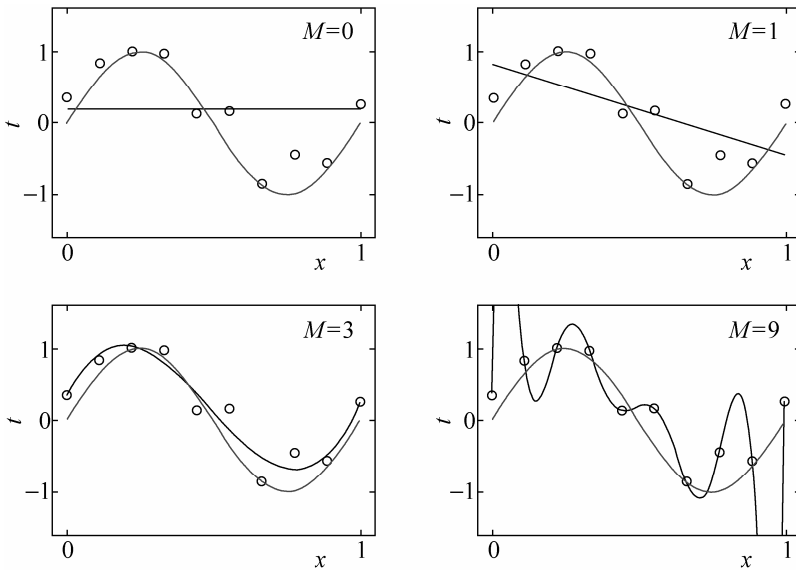


图 1.2 M 次多项式函数拟合问题的例子

这时，损失函数为平方损失，系数 $\frac{1}{2}$ 是为了计算方便。

这是一个简单的最优化问题。将模型与训练数据代入式(1.18)中，有

$$L(w) = \frac{1}{2} \sum_{i=1}^N \left(\sum_{j=0}^M w_j x_i^j - y_i \right)^2$$

对 w_j 求偏导数并令其为 0，可得

$$w_j = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^{j+1}}, \quad j = 0, 1, 2, \dots, M$$

于是求得拟合多项式系数 $w_0^*, w_1^*, \dots, w_M^*$ 。

图 1.2 给出了 $M=0$ ， $M=1$ ， $M=3$ 及 $M=9$ 时多项式函数拟合的情况。如果 $M=0$ ，多项式曲线是一个常数，数据拟合效果很差。如果 $M=1$ ，多项式曲线是一条直线，数据拟合效果也很差。相反，如果 $M=9$ ，多项式曲线通过每个数据点，训练误差为 0。从对给定训练数据拟合的角度来说，效果是最好的。但是，因为训练数据本身存在噪声，这种拟合曲线对未知数据的预测能力往往并不是最好的，在实际学习中并不可取。这时过拟合现象就会发生。这就是说，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。当 $M=3$ 时，多项式曲线对训练数据拟合效果足够好，模型也比较简单，是一个较好的选择。 ■

在多项式函数拟合中可以看到，随着多项式次数（模型复杂度）的增加，训练误差会减小，直至趋向于 0，但是测试误差却不如此，它会随着多项式次数（模型复杂度）的增加先减小而后增大。而最终的目的是使测试误差达到最小。这样，在多项式函数拟合中，就要选择合适的多项式次数，以达到这一目的。这一结论对一般的模型选择也是成立的。

图 1.3 描述了训练误差和测试误差与模型的复杂度之间的关系。当模型的复杂度增大时，训练误差会逐渐减小并趋向于 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。这样，在学习时就要防止过拟合，进行最优的模型选择，即选择复杂度适当的模型，以达到使测试误差最小的学习目的。下面介绍两种常用的模型选择方法：正则化与交叉验证。

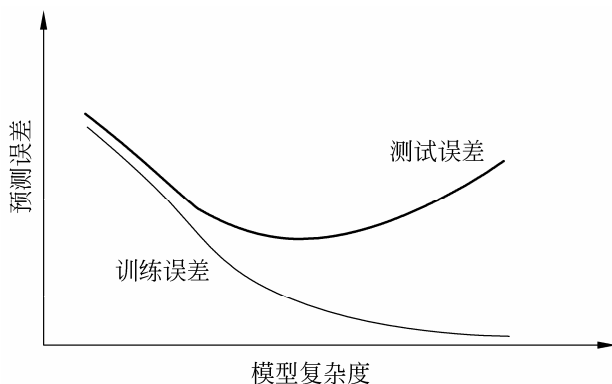


图 1.3 训练误差和测试误差与模型复杂度的关系

1.5 正则化与交叉验证

1.5.1 正则化

模型选择的典型方法是正则化（regularization）。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（regularizer）或罚项（penalty term）。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。

正则化一般具有如下形式：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f) \quad (1.19)$$

其中，第 1 项是经验风险，第 2 项是正则化项， $\lambda \geq 0$ 为调整两者之间关系的系数。

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

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

这里, $\|w\|$ 表示参数向量 w 的 L_2 范数.

正则化项也可以是参数向量的 L_1 范数:

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

这里, $\|w\|_1$ 表示参数向量 w 的 L_1 范数.

第1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2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.

正则化符合奥卡姆剃刀(Occam's razor)原理. 奥卡姆剃刀原理应用于模型选择时变为以下想法: 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, 也就是应该选择的模型. 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较大的先验概率, 简单的模型有较小的先验概率.

1.5.2 交叉验证

另一种常用的模型选择方法是交叉验证(cross validation).

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, 分别为训练集(training set)、验证集(validation set)和测试集(test set). 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. 在学习到的不同复杂度的模型中, 选择对验证集有最小预测误差的模型. 由于验证集有足够多的数据, 用它对模型进行选择也是有效的.

但是, 在许多实际应用中数据是不充足的. 为了选择好的模型, 可以采用交叉验证方法. 交叉验证的基本想法是重复地使用数据; 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.

1. 简单交叉验证

简单交叉验证方法是: 首先随机地将已给数据分为两部分, 一部分作为训练集, 另一部分作为测试集(例如, 70%的数据为训练集, 30%的数据为测试集); 然后用训练集在各种条件下(例如, 不同的参数个数)训练模型, 从而得到不同的模型; 在测试集上评价各个模型的测试误差, 选出测试误差最小的模型.

2. S 折交叉验证

应用最多的是 S 折交叉验证 (S -fold cross validation), 方法如下: 首先随机地将已给数据切分为 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.

3. 留一交叉验证

S 折交叉验证的特殊情形是 $S=N$, 称为留一交叉验证 (leave-one-out cross validation), 往往在数据缺乏的情况下使用. 这里, N 是给定数据集的容量.

1.6 泛化能力

1.6.1 泛化误差

学习方法的泛化能力 (generalization ability) 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. 现实中采用最多的办法是通过测试误差来评价学习方法的泛化能力. 但这种评价是依赖于测试数据集的. 因为测试数据集是有限的, 很有可能由此得到的评价结果是不可靠的. 统计学习理论试图从理论上对学习方法的泛化能力进行分析.

首先给出泛化误差的定义. 如果学到的模型是 \hat{f} , 那么用这个模型对未知数据预测的误差即为泛化误差 (generalization error)

$$R_{\text{exp}}(\hat{f}) = E_P[L(Y, \hat{f}(X))] = \int_{x \times y} L(y, \hat{f}(x)) P(x, y) dx dy \quad (1.20)$$

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. 事实上, 泛化误差就是所学习到的模型的期望风险.

1.6.2 泛化误差上界

学习方法的泛化能力分析往往是通过研究泛化误差的概率上界进行的, 简称为泛化误差上界 (generalization error bound). 具体来说, 就是通过比较两种学习方法的泛化误差上界的大小来比较它们的优劣. 泛化误差上界通常具有以下性质: 它是样本容量的函数, 当样本容量增加时, 泛化上界趋于 0; 它是假设空间容量 (capacity) 的函数, 假设空间容量越大, 模型就越难学, 泛化误差上界就越大.

下面给出一个简单的泛化误差上界的例子: 二类分类问题的泛化误差上界.

考虑二类分类问题. 已知训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 它是

从联合概率分布 $P(X, Y)$ 独立同分布产生的, $X \in \mathbf{R}^n$, $Y \in \{-1, +1\}$. 假设空间是函数的有限集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$, d 是函数个数. 设 f 是从 \mathcal{F} 中选取的函数. 损失函数是 0-1 损失. 关于 f 的期望风险和经验风险分别是

$$R(f) = E[L(Y, f(X))] \quad (1.21)$$

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) \quad (1.22)$$

经验风险最小化函数是

$$f_N = \arg \min_{f \in \mathcal{F}} \hat{R}(f) \quad (1.23)$$

人们更关心的是 f_N 的泛化能力

$$R(f_N) = E[L(Y, f_N(X))] \quad (1.24)$$

下面讨论从有限集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 中任意选出的函数 f 的泛化误差上界.

定理 1.1 (泛化误差上界) 对二类分类问题, 当假设空间是有限个函数的集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 时, 对任意一个函数 $f \in \mathcal{F}$, 至少以概率 $1 - \delta$, 以下不等式成立:

$$R(f) \leq \hat{R}(f) + \varepsilon(d, N, \delta) \quad (1.25)$$

其中,

$$\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2N} \left(\log d + \log \frac{1}{\delta} \right)} \quad (1.26)$$

不等式(1.25)左端 $R(f)$ 是泛化误差, 右端即为泛化误差上界. 在泛化误差上界中, 第 1 项是训练误差, 训练误差越小, 泛化误差也越小. 第 2 项 $\varepsilon(d, N, \delta)$ 是 N 的单调递减函数, 当 N 趋于无穷时趋于 0; 同时它也是 $\sqrt{\log d}$ 阶的函数, 假设空间 \mathcal{F} 包含的函数越多, 其值越大.

证明 在证明中要用到 Hoeffding 不等式, 先叙述如下.

设 $S_n = \sum_{i=1}^n X_i$ 是独立随机变量 X_1, X_2, \dots, X_n 之和, $X_i \in [a_i, b_i]$, 则对任意 $t > 0$, 以下不等式成立:

$$P(S_n - ES_n \geq t) \leq \exp \left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \quad (1.27)$$

$$P(ES_n - S_n \geq t) \leq \exp \left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \quad (1.28)$$

对任意函数 $f \in \mathcal{F}$, $\hat{R}(f)$ 是 N 个独立的随机变量 $L(Y, f(X))$ 的样本均值, $R(f)$

是随机变量 $L(Y, f(X))$ 的期望值. 如果损失函数取值于区间 $[0, 1]$, 即对所有 $i, [a_i, b_i] = [0, 1]$, 那么由 Hoeffding 不等式 (1.28) 不难得知, 对 $\varepsilon > 0$, 以下不等式成立:

$$P(R(f) - \hat{R}(f) \geq \varepsilon) \leq \exp(-2N\varepsilon^2)$$

由于 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 是一个有限集合, 故

$$\begin{aligned} P(\exists f \in \mathcal{F} : R(f) - \hat{R}(f) \geq \varepsilon) &= P\left(\bigcup_{f \in \mathcal{F}} \{R(f) - \hat{R}(f) \geq \varepsilon\}\right) \\ &\leq \sum_{f \in \mathcal{F}} P(R(f) - \hat{R}(f) \geq \varepsilon) \\ &\leq d \exp(-2N\varepsilon^2) \end{aligned}$$

或者等价的, 对任意 $f \in \mathcal{F}$, 有

$$P(R(f) - \hat{R}(f) < \varepsilon) \geq 1 - d \exp(-2N\varepsilon^2)$$

$$\text{令} \quad \delta = d \exp(-2N\varepsilon^2) \quad (1.29)$$

$$\text{则} \quad P(R(f) < \hat{R}(f) + \varepsilon) \geq 1 - \delta$$

即至少以概率 $1 - \delta$ 有 $R(f) < \hat{R}(f) + \varepsilon$, 其中 ε 由式 (1.29) 得到, 即为式 (1.26). ■

从泛化误差上界可知,

$$R(f_N) \leq \hat{R}(f_N) + \varepsilon(d, N, \delta)$$

其中, $\varepsilon(d, N, \delta)$ 由式 (1.26) 定义, f_N 由式 (1.23) 定义. 这就是说, 训练误差小的模型, 其泛化误差也会小.

以上讨论的只是假设空间包含有限个函数情况下的泛化误差上界, 对一般的假设空间要找到泛化误差界就没有这么简单, 这里不作介绍.

1.7 生成模型与判别模型

监督学习的任务就是学习一个模型, 应用这一模型, 对给定的输入预测相应的输出. 这个模型的一般形式为决策函数:

$$Y = f(X)$$

或者条件概率分布:

$$P(Y|X)$$

监督学习方法又可以分为生成方法 (generative approach) 和判别方法

(discriminative approach). 所学到的模型分别称为生成模型 (generative model) 和判别模型 (discriminative model).

生成方法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, 即生成模型:

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

这样的方法之所以称为生成方法, 是因为模型表示了给定输入 X 产生输出 Y 的生成关系. 典型的生成模型有: 朴素贝叶斯法和隐马尔可夫模型, 将在后面章节进行相关讲述.

判别方法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, 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、逻辑斯谛回归模型、最大熵模型、支持向量机、提升方法和条件随机场等, 将在后面章节讲述.

在监督学习中, 生成方法和判别方法各有优缺点, 适合于不同条件下的学习问题.

生成方法的特点: 生成方法可以还原出联合概率分布 $P(X, Y)$, 而判别方法则不能; 生成方法的学习收敛速度更快, 即当样本容量增加的时候, 学到的模型可以更快地收敛于真实模型; 当存在隐变量时, 仍可以用生成方法学习, 此时判别方法就不能用.

判别方法的特点: 判别方法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, 往往学习的准确率更高; 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.

1.8 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, 预测问题便成为分类问题. 这时, 输入变量 X 可以是离散的, 也可以是连续的. 监督学习从数据中学习一个分类模型或分类决策函数, 称为分类器 (classifier). 分类器对新的输入进行输出的预测 (prediction), 称为分类 (classification). 可能的输出称为类 (class). 分类的类别为多个时, 称为多类分类问题. 本书主要讨论二类分类问题.

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. 分类问题可用图 1.4 描述. 图中 $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ 是训练数据集, 学习系统由训练数据学习一个分类器 $P(Y|X)$ 或 $Y = f(X)$; 分类系

统通过学到的分类器 $P(Y|X)$ 或 $Y = f(X)$ 对于新的输入实例 x_{N+1} 进行分类, 即预测其输出的类标记 y_{N+1} .

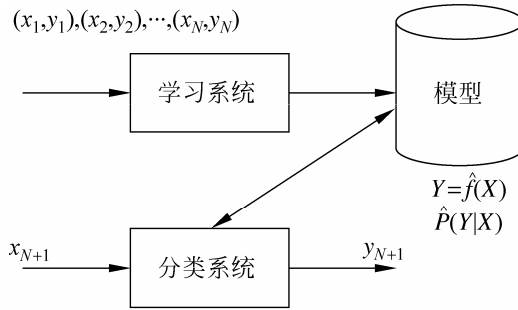


图 1.4 分类问题

评价分类器性能的指标一般是分类准确率 (accuracy), 其定义是: 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. 也就是损失函数是 0-1 损失时测试数据集上的准确率 (见公式 (1.17)).

对于二类分类问题常用的评价指标是精确率 (precision) 与召回率 (recall). 通常以关注的类为正类, 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.

精确率定义为

$$P = \frac{TP}{TP + FP} \quad (1.30)$$

召回率定义为

$$R = \frac{TP}{TP + FN} \quad (1.31)$$

此外, 还有 F_1 值, 是精确率和召回率的调和均值, 即

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R} \quad (1.32)$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (1.33)$$

精确率和召回率都高时, F_1 值也会高.

许多统计学习方法可以用于分类, 包括 k 近邻法、感知机、朴素贝叶斯法、决策树、决策列表、逻辑斯谛回归模型、支持向量机、提升方法、贝叶斯网络、神

经网络、Winnow 等。本书将讲述其中一些主要方法。

分类在于根据其特性将数据“分门别类”，所以在许多领域都有广泛的应用。例如，在银行业务中，可以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，可以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。

举一个分类应用的例子——文本分类(text classification)。这里的文本可以是新闻报道、网页、电子邮件、学术论文等。类别往往是关于文本内容的，例如政治、经济、体育等；也有关于文本特点的，如正面意见、反面意见；还可以根据应用确定，如垃圾邮件、非垃圾邮件等。文本分类是根据文本的特征将其划分到已有的类中。输入是文本的特征向量，输出是文本的类别。通常把文本中的单词定义为特征，每个单词对应一个特征。单词的特征可以是二值的，如果单词在文本中出现则取值是1，否则是0；也可以是多值的，表示单词在文本中出现的频率。直观地，如果“股票”“银行”“货币”这些词出现很多，这个文本可能属于经济类，如果“网球”“比赛”“运动员”这些词频繁出现，这个文本可能属于体育类。

1.9 标注问题

标注(tagging)也是一个监督学习问题。可以认为标注问题是分类问题的一个推广，标注问题又是更复杂的结构预测(structure prediction)问题的简单形式。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。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。注意，可能的标记个数是有限的，但其组合所成的标记序列的个数是依序列长度呈指数级增长的。

标注问题分为学习和标注两个过程(如图1.5所示)。首先给定一个训练数据集

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

这里， $x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})^T$ ， $i = 1, 2, \dots, N$ ，是输入观测序列， $y_i = (y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(n)})^T$ 是相应的输出标记序列， n 是序列的长度，对不同样本可以有不同的值。学习系统基于训练数据集构建一个模型，表示为条件概率分布：

$$P(Y^{(1)}, Y^{(2)}, \dots, Y^{(n)} | X^{(1)}, X^{(2)}, \dots, X^{(n)})$$

这里，每一个 $X^{(i)}$ ($i = 1, 2, \dots, n$)取值为所有可能的观测，每一个 $Y^{(i)}$ ($i = 1, 2, \dots, n$)取值为所有可能的标记，一般 $n \ll N$ 。标注系统按照学习得到的条件概率分布模

型, 对新的输入观测序列找到相应的输出标记序列. 具体地, 对一个观测序列 $x_{N+1} = (x_{N+1}^{(1)}, x_{N+1}^{(2)}, \dots, x_{N+1}^{(n)})^T$ 找到使条件概率 $P((y_{N+1}^{(1)}, y_{N+1}^{(2)}, \dots, y_{N+1}^{(n)})^T | (x_{N+1}^{(1)}, x_{N+1}^{(2)}, \dots, x_{N+1}^{(n)})^T)$ 最大的标记序列 $y_{N+1} = (y_{N+1}^{(1)}, y_{N+1}^{(2)}, \dots, y_{N+1}^{(n)})^T$.

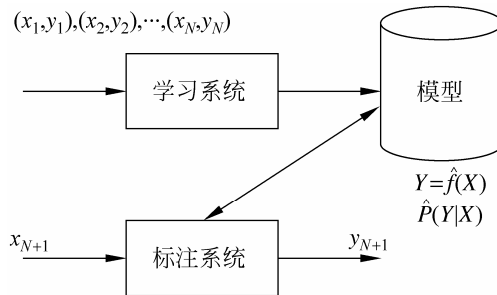


图 1.5 标注问题

评价标注模型的指标与评价分类模型的指标一样, 常用的有标注准确率、精确率和召回率. 其定义与分类模型相同.

标注常用的统计学习方法有: 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. 例如, 自然语言处理中的词性标注(part of speech tagging)就是一个典型的标注问题: 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.

举一个信息抽取的例子. 从英文文章中抽取基本名词短语 (base noun phrase). 为此, 要对文章进行标注. 英文单词是一个观测, 英文句子是一个观测序列, 标记表示名词短语的“开始”、“结束”或“其他”(分别以 B, E, O 表示), 标记序列表示英文句子中基本名词短语的所在位置. 信息抽取时, 将标记“开始”到标记“结束”的单词作为名词短语. 例如, 给出以下的观测序列, 即英文句子, 标注系统产生相应的标记序列, 即给出句子中的基本名词短语.

输入: At Microsoft Research, we have an insatiable curiosity and the desire to create new technology that will help define the computing experience.

输出: At/O Microsoft/B Research/E, we/O have/O an/O insatiable/B curiosity/E and/O the/O desire/BE to/O create/O new/B technology/E that/O will/O help/O define/O the/O computing/B experience/E.

1.10 回归问题

回归 (regression) 是监督学习的另一个重要问题. 回归用于预测输入变量 (自变量) 和输出变量 (因变量) 之间的关系, 特别是当输入变量的值发生变化时, 输

出变量的值随之发生的变化。回归模型正是表示从输入变量到输出变量之间映射的函数。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地预测未知数据（参照 1.4.2 节）。

回归问题分为学习和预测两个过程（如图 1.6 所示）。首先给定一个训练数据集：

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

这里， $x_i \in \mathbf{R}^n$ 是输入， $y \in \mathbf{R}$ 是对应的输出， $i = 1, 2, \dots, N$ 。学习系统基于训练数据构建一个模型，即函数 $Y = f(X)$ ；对新的输入 x_{N+1} ，预测系统根据学习的模型 $Y = f(X)$ 确定相应的输出 y_{N+1} 。

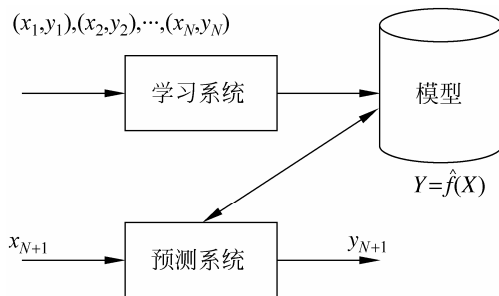


图 1.6 回归问题

回归问题按照输入变量的个数，分为一元回归和多元回归；按照输入变量和输出变量之间关系的类型即模型的类型，分为线性回归和非线性回归。

回归学习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（least squares）求解。

许多领域的任务都可以形式化为回归问题，比如，回归可以用于商务领域，作为市场趋势预测、产品质量管理、客户满意度调查、投资风险分析的工具。作为例子，简单介绍股价预测问题。假设知道某一公司在过去不同时间点（比如，每天）的市场上的股票价格（比如，股票平均价格），以及在各个时间点之前可能影响该公司股价的信息（比如，该公司前一周的营业额、利润）。目标是从过去的的数据学习一个模型，使它可以基于当前的信息预测该公司下一个时间点的股票价格。可以将这个问题作为回归问题解决。具体地，将影响股价的信息视为自变量（输入的特征），而将股价视为因变量（输出的值）。将过去的的数据作为训练数据，就可以学习一个回归模型，并对未来的股价进行预测。可以看出这是一个困难的预测问题，因为影响股价的因素非常多，我们未必能判断到哪些信息（输入的特征）有用并能得到这些信息。

本章概要

1. 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分析与预测的一门学科。统计学习包括监督学习、非监督学习、半监督学习和强化学习。

2. 统计学习方法三要素——模型、策略、算法，对理解统计学习方法起到提纲挈领的作用。

3. 本书主要讨论监督学习，监督学习可以概括如下：从给定有限的训练数据出发，假设数据是独立同分布的，而且假设模型属于某个假设空间，应用某一评价准则，从假设空间中选取一个最优的模型，使它对已给训练数据及未知测试数据在给定评价标准意义下有最准确的预测。

4. 统计学习中，进行模型选择或者说提高学习的泛化能力是一个重要问题。如果只考虑减少训练误差，就可能产生过拟合现象。模型选择的方法有正则化与交叉验证。学习方法泛化能力的分析是统计学习理论研究的重要课题。

5. 分类问题、标注问题和回归问题都是监督学习的重要问题。本书中介绍的统计学习方法包括感知机、 k 近邻法、朴素贝叶斯法、决策树、逻辑斯谛回归与最大熵模型、支持向量机、提升方法、EM 算法、隐马尔可夫模型和条件随机场。这些方法是主要的分类、标注以及回归方法。它们又可以归类为生成方法与判别方法。

继续阅读

关于统计学习方法一般介绍的书籍可以参阅文献[1~4]。

习 题

- 1.1 说明伯努利模型的极大似然估计以及贝叶斯估计中的统计学习方法三要素。伯努利模型是定义在取值为 0 与 1 的随机变量上的概率分布。假设观测到伯努利模型 n 次独立的数据生成结果，其中 k 次的结果为 1，这时可以用极大似然估计或贝叶斯估计来估计结果为 1 的概率。
- 1.2 通过经验风险最小化推导极大似然估计。证明模型是条件概率分布，当损失函数是对数损失函数时，经验风险最小化等价于极大似然估计。

参 考 文 献

- [1] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. 2001 (中译本: 统计学习基础——数据挖掘、推理与预测. 范明, 柴玉梅, 咎红英等译. 北京: 电子工业出版社, 2004)
- [2] Tom M Michelle. Machine Learning. McGraw-Hill Companies, Inc. 1997 (中译本: 机器学习. 北京: 机械工业出版社, 2003)
- [3] Bishop M. Pattern Recognition and Machine Learning. Springer, 2006
- [4] Brian Ripley. Pattern Recognition and Neural Networks. Cambridge University Press, 1996