第 2 章　统计学习理论基本知识

统计学习理论是一种专门基于小样本的统计理论，它为研究有限样本情况下的统计模式识别和更广泛的机器学习问题建立了一个较好的理论框架，同时也发展了一种新的模式识别方法——支持向量机，能够较好地解决小样本学习问题。

2.1　统计学习理论的核心内容

机器学习的目的是根据给定的已知训练样本求取对系统输入和输出之间的依赖关系的估计，使它能够对未知输出作出尽可能准确的预测。机器学习问题可以形式化地表示为：已知变量 y 与输入 x 之间存在一定的未知依赖关系，即存在一个未知的联合概率 $F(x,y)$，机器学习就是根据 n 个独立同分布观测样本

$$(x_1,y_1),(x_2,y_2),...,(x_n,y_n) \tag{2.1}$$

在一组函数 $\{f(x,w)\}$ 中求一个最优的函数 $f(x,w_0)$，使预测的期望风险最小。

$$R(w)=\int L(y,f(x,w))dF(x,y) \tag{2.2}$$

其中，$\{f(x,w)\}$ 为预测函数集，$w\in\Omega$ 为函数的广义参数，故 $\{f(x,w)\}$ 可以表示任何函数集；$L(y,f(x,w))$ 为由于用 $f(x,w)$ 对 y 进行预测而造成的损失。

要使式（2.2）定义的期望风险最小化，必须依赖关于联合概率 $F(x,y)$ 的信息，但在实际的机器学习问题中，我们只能利用已知样本（2.1）的信息，因此期望风险无法直接计算和最小化。

根据概率论中大数定律定理的思想，人们自然想到用算术平均代替式（2.2）的数学期望，于是定义了

$$R_{emp}(w)=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i,w)) \tag{2.3}$$

来逼近式（2.2）定义的期望风险。由于 $R_{emp}(w)$ 是用已知的训练样本（即经验数

据）定义的，因此称作经验风险。用对参数 w 求经验风险 $R_{emp}(w)$ 的最小值代替求期望风险 $R(w)$ 的最小值就是所谓的经验风险最小化（Empirical Risk Minimization，ERM）原则。

仔细研究经验风险最小化原则和机器学习问题中的期望风险最小化要求可以发现，从期望风险到经验风险最小化并没有可靠的理论依据，只是直观上合理的想当然做法。但是，经验风险最小化作为解决模式识别等机器学习问题的基本思想仍在相当长的时间内统治了这一领域的几乎所有研究，人们多年来一直将大部分注意力集中到如何更好地求取最小经验风险上。与此相反，统计学习理论则对用经验风险最小化原则解决期望风险最小化问题的前提是什么、当这些前提不成立时经验风险最小化方法的性能如何，以及是否可以找到更合理的原则等基本问题进行了深入的研究。

统计学习理论被认为是目前针对小样本统计估计和预测学习的最佳理论，它从理论上较系统地研究了经验风险最小化原则成立的条件、有限样本下经验风险与期望风险的关系、如何利用这些理论找到新的学习原则和方法等问题。其主要内容包括以下四个方面[8]：

（1）经验风险最小化原则下统计学习一致性的条件。

（2）在这些条件下关于统计学习方法推广性的界的结论。

（3）在这些界的基础上建立的小样本归纳推理原则。

（4）实现这些新原则的实际方法（算法）。

2.1.1 学习过程一致性的条件

学习过程一致性是统计学习理论的基础，也是与传统渐进统计学的基本联系。学习过程一致性就是指当训练样本的数目趋于无穷大时，经验风险的最优值能够收敛到真实风险的最优值。只有满足一致性条件，才能保证经验风险最小化原则下得到的最优解在样本无穷大时趋近于使用期望风险最小的最优结果[8]。

定义 2.1 记 $f(x,w^*)$ 为在式（2.1）的 n 个独立同分布样本下，在函数集中使经验风险取最小的预测函数，由它带来的损失函数为 $L(y,f(x,w^*))$，相应的最小

经验风险值为 $R_{emp}(w^*)$。记 $R(w^*)$ 为在 $L(y,f(x,w^*))$ 函数下的式（2.2）所取得的真实风险值（期望风险）。当下面两式成立时称这个经验风险最小化学习过程是一致的：

$$R(w^*) \underset{n\to\infty}{\to} R(w_0) \tag{2.4}$$

$$R_{emp}(w^*) \underset{n\to\infty}{\to} R(w_0) \tag{2.5}$$

其中，$R(w_0)=\inf_w R(w)$ 为实际可能的最小风险，即式（2.2）的下确界或最小值。

现在的关键问题是保证经验风险最小化方法一致性的条件，这个条件针对函数集的一般特性和概率测度。对于前面的一致性的定义存在一种特殊的情况：预测函数集中包含某个特殊函数，它使定义中的条件得到满足；而如果从函数集中去掉这个函数，这些条件就不能得到满足。为了保证一致性不是由于函数集中的个别函数导致的而产生了所谓非平凡一致性的概念，即要求定义中的条件对预测函数集的所有子集都成立。后面说到的一致性指的就是非平凡一致性。

下面的定理给出了保证经验风险最小化方法一致性的条件，由于该定理在统计学习理论中的重要地位，该定理被称为学习理论的关键定理[8]。

定理 2.1 对于有界的损失函数，经验风险最小化学习一致性的充分必要条件是经验风险在如下意义上一致地收敛于真实风险：

$$\lim_{n\to\infty} P[\sup_w(R(w)-R_{emp}(w))>\varepsilon]=0,\quad \forall\varepsilon>0 \tag{2.6}$$

其中，P 为概率，$R_{emp}(w)$ 和 $R(w)$ 分别为在 n 个样本下的经验风险和对同一个 w 的真实风险。

该定理把学习一致性的问题转化为式（2.6）的一致收敛问题，但是并没有给出哪种函数集能够满足这个充分必要条件，因此，统计学习理论定义了衡量函数集性能的一些指标，其中最重要的指标是 VC 维。

2.1.2 VC 维

模式识别问题中的 VC 维的直观定义是：如果一个指示函数集存在 h 个样本能够被函数集中的函数按所有可能的 $2h$ 种形式分开，则称函数集能够把 h 个样本

打散，函数集的 VC 维就是它能打散的最大样本数目 h，即如果存在 h 个样本的样本集能够被函数集打散，而不存在有 $h+1$ 个样本的样本集能被函数集打散，则函数集的 VC 维就是 h。若对任意数目的样本都有函数能将其打散，则函数集的 VC 维是无穷大。有界实函数的 VC 维可以通过用一定的阈值将其转化成指示函数来定义。VC 维反映了函数集的学习能力，VC 维越大则学习机器越复杂（容量越大）。遗憾的是，目前尚没有通用的关于任意函数集 VC 维计算的理论，只对一些特殊的函数集知道其 VC 维。对于一些比较复杂的学习机器（如神经网络），其 VC 维除了与函数集（神经网结构）有关外，还受学习算法等的影响，其确定更加困难[30]。

根据文献[7]和[31]，经验风险最小化学习过程一致的充分必要条件是函数集的 VC 维有限，且这时收敛速度较快。

2.1.3 推广性的界

前面关于一致收敛和收敛速度的条件在理论上有重要意义，但在实践中无法直接应用。统计学习理论系统地研究了对于各种类型的函数集的经验风险和实际风险之间的关系，即推广性的界[31]。

关于两类分类问题，其结论是：对指示函数集中的所有函数（包括使经验风险最小的函数），经验风险 $R_{emp}(w)$ 和期望风险 $R(w)$ 之间以至少 $1-\eta$ 的概率满足如下关系[32]：

$$R(w) \leqslant R_{emp}(w) + \sqrt{\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}} \tag{2.7}$$

其中，h 是函数集的 VC 维，n 是样本数。

这一结论从理论上说明了学习机器的实际风险是由两部分组成的：一是经验风险（训练误差）；二是置信范围，它与学习机器的 VC 维和训练样本数有关。可以简单地表示为

$$R(w) \leqslant R_{emp}(w) + \Phi(\frac{n}{h}) \tag{2.8}$$

进一步分析可以发现，当 n/h 较小时（比如小于 20，此时说样本数较少），

置信范围 Φ 较大，用经验风险近似期望风险就有较大的误差，用经验风险最小化得到的最优解可能具有较差的推广性；如果样本数较多，n/h 较大，则置信范围就会很小，经验风险最小化的最优解就接近实际的最优解。另一方面，对于一个特定的问题，其样本数 n 是固定的，此时学习机器的 VC 维越高（即复杂性越高），则置信范围就越大，导致期望风险与经验风险之间可能的差别就越大。因此，在有限训练样本下，学习机器的 VC 维越高（复杂性越高）则置信范围越大，导致期望风险与经验风险之间可能的差别越大，这就是为什么会出现过学习现象的原因。机器学习过程不但要使经验风险最小，还要使 VC 维尽量小以缩小置信范围，这样才能取得较小的期望风险，即对未来样本有较好的推广性。

2.1.4 结构风险最小化

前面讨论了在样本数较多的情况下可以用经验风险最小化的最优值来估计实际的最优值，但是当样本数较少时，这个估计是不准确的。因为这时要同时最小化经验风险和置信范围，即在经验风险最小化的同时设法控制学习机器的 VC 维数。实际上，在传统的学习机器中，选择学习模型和算法的过程就是优化置信范围的过程，如果选择的模型比较适合现有的训练样本（相当于 n/h 的值适当），则可以得到比较好的结果，但是这种选择往往依赖先验知识和经验。

由于有式（2.8）的理论依据，统计学习理论提供了一种在小样本情况下，使 $R_{emp}(w)$ 极小化的同时控制 VC 维（模型复杂性）的方法，即对于给定的有限样本选择最佳模型复杂性的方法，该方法描述如下：

首先把函数集 $S=\{f(x,w),w\in\Omega\}$ 分解为一个函数子集序列，使其具有一种嵌套结构

$$S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots \subset S \tag{2.9}$$

其中 $S_k=\{f(x,w),w\in\Omega_k\}$ 为函数集的子集（元素），其 VC 维数 h_k 为有限。在此结构中，嵌套子集按其复杂性（即 VC 维数大小）的顺序排列：

$$h_1 \leqslant h_2 \leqslant \cdots \leqslant h_k \leqslant \cdots \tag{2.10}$$

这样在同一个子集中，置信范围是相同的。在每个子集中寻找最小经验风险，通常它随着子集复杂度的增加而减小。选择最小经验风险与置信范围之和最小的子

集就可以达到期望风险的最小，这个子集中使经验风险最小的函数就是要求的最优函数，这种思想称为结构风险最小化原则（Structural Risk Minimization，SRM）。如图 2.1 所示给出了结构风险最小化的示意。

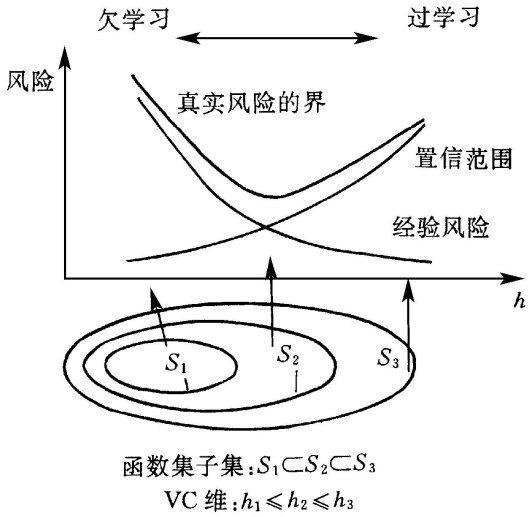

图 2.1　结构风险最小化示意图

SRM 原理实际上提供了一种对于给定的样本数据，在近似精度和模型近似函数复杂性之间折中的定量方法，即在近似函数集的结构中找出一个最佳子集，使实际风险确保上界达到极小。然而上述一般算法的计算量太大，不是一种在实际应用中可行的算法。在实际应用中可行的是以下两种算法。

算法 1：

（1）在函数集中，根据所给样本数据集的大小 n 和其他先验知识选定一个子集 S_{k0}，使其置信区间足够小。

（2）在 S_{k0} 中求解经验风险的极小化问题。

这个算法实质上是通过选择适当的近似函数结构使置信区间保持不变，然后进行经验风险极小化。这是在神经网络中应用的方法。

算法 2：

（1）找出一种特殊的函数集，其结构中每一个子集 S_k 的经验风险都相同（等

于零或一个非常小的数）。

（2）求出使置信区间最小的一个子集，则该子集的期望风险为极小。

这个算法就是在保持经验风险不变的条件下使置信区间极小。支持向量机用的就是这种算法。

2.2 支持向量分类机

支持向量机是在实际问题中具体实现统计学习理论的算法。支持向量机是统计学习理论中最新、最实用的内容，其核心内容于 1992～1995 年间提出[6,9-10]，目前在国内外的机器学习领域得到广泛的重视，并且还在不断地发展。

2.2.1 线性支持向量分类机

考虑线性可分的情况[33-35]，给定有 l 个样本 $\{x_i, y_i\}_{i=1}^{l}$ 的训练集合，其中第 i 个输入数据 $x_i \in R^n$ 且第 i 个输出数据 $y_i \in \{-1,+1\}$ 是类标。定义判别函数

$$f(x) = < w \cdot x > + b = 0 \tag{2.11}$$

这个判别函数是 n 维矢量空间中的一个超平面，简称为分界面，其中 $<\cdot>$ 是矢量的内积。为了使超平面（2.11）能将 $y_i = +1$ 和 $y_i = -1$ 的两类样本正确地区分，应选择适当的 w 和 b 使样本 $x_i (i = 1,...,l)$ 满足下列条件：

$$\begin{cases} < w \cdot x_i > + b \geqslant +1 & y_i = +1 \\ < w \cdot x_i > + b \leqslant -1 & y_i = -1 \end{cases} \tag{2.12}$$

这里式（2.12）可以改写成更紧凑的形式：

$$y_i[< w \cdot x_i > + b] \geqslant +1 \qquad i = 1,...,l \tag{2.13}$$

任意一个样本点 x_i 到分界面（2.11）的距离为

$$d_i = \frac{|f(x_i)|}{\|w\|} \tag{2.14}$$

若存在一个 τ，对任意样本都有：

$$\frac{y_i f(x_i)}{\|w\|} \geqslant \tau \qquad i = 1,...,l \tag{2.15}$$

则称 τ 为判别函数（2.11）的余量，它表示样本点与分界面之间的最小距离。余量越大，基于该分界面的分类推广能力越好，但是对同一组分类样本可以做出许多分界面。

从式（2.15）可以看出，余量越大则 $\|w\|$ 越小，因此求最优分界面的问题可以表述为下列二次优化问题：

对于给定的训练样本 $\{x_i, y_i\}_{i=1}^{l}$，求使下列二次泛函取极小值的 w 和 b

$$\min \frac{1}{2}\|w\|^2 \tag{2.16}$$

约束条件为

$$y_i[< w \cdot x_i > +b] \geqslant +1 \qquad i = 1,...,l \tag{2.17}$$

对于这样一个二次规划问题，通常转换成与其对应的 Lagrange 对偶问题来求解，该问题对应的 Lagrange 函数为：

$$L(\alpha, w, b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i [y_i (< w \cdot x_i > +b) - 1] \tag{2.18}$$

其中 α_i（$\alpha_i \geqslant 0$）为 Lagrange 乘子。Lagrange 对偶问题为:

$$\max_{\alpha} \min_{w,b} L(\alpha, w, b) \tag{2.19}$$

利用 Kuhn-Tucker 条件

$$\frac{\partial L}{\partial w} = 0$$

$$\frac{\partial L}{\partial b} = 0$$

得到

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{2.20}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.21}$$

将式（2.20）代入式（2.18），并利用式（2.21）得到原二次规划问题的对偶优化问题：

对给定的训练样本 $\{x_i, y_i\}_{i=1}^{l}$，求使下列二次函数取极大值的 α_i

$$\max_{\alpha} L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j <x_i\cdot x_j> \tag{2.22}$$

约束条件为

$$\sum_{i=1}^{l} y_i\alpha_i=0 \qquad \alpha_i \geqslant 0,\ i=1,2,...,l \tag{2.23}$$

式（2.22）中只需要计算输入矢量的内积，约束条件也很简单，因此该对偶问题比原问题简单得多，较容易用标准的二次规划方法求解。设式（2.22）和式（2.23）的解为 $\alpha_i=\alpha_i^*$（$i=1,2,...l$），最优分界面的参数 w^*、b^* 与 α_i^* 有如下关系：

$$w^*=\sum_{i=1}^{l}\alpha_i^* y_i x_i \tag{2.24}$$

$$\alpha_i^*[y_i(<w^*\cdot x_i>+b)-1]=0 \qquad i=1,2,...l \tag{2.25}$$

由式（2.25）可以看出，仅当

$$[y_i(<w^*\cdot x_i>+b)-1]=0 \tag{2.26}$$

即约束条件（2.23）取等号时，α_i^* 才能取非零的数值，这时它们所对应的样本中的输入矢量 $\tilde{x}_i$ 称为支持向量（$i=1,2,...s$）。因此式（2.24）中的求和只需对支持向量进行

$$w^*=\sum_{i=1}^{s}\alpha_i^* y_i\tilde{x}_i \tag{2.27}$$

其中 s 为支持向量数，通常 $s<<l$。式（2.27）表明，最优分界面的权重参数向量 w^* 可以表示为支持向量的线性组合。最优分界面的阈值 b^* 可以由式（2.26）求得

$$b^*=-\frac{1}{2}[<w\cdot\tilde{x}(+1)>+<w\cdot\tilde{x}(-1)>] \tag{2.28}$$

其中 $\tilde{x}(+1)$ 表示某一个属于第 1 类（$y=+1$）的支持向量，$\tilde{x}(-1)$ 表示某一个属于第 2 类（$y=-1$）的支持向量。分界面方程也完全由支持向量确定

$$f(x)=\sum_{i=1}^{s}\alpha_i^* y_i <x\cdot\tilde{x}_i>+b \tag{2.29}$$

2.2.2 非线性支持向量分类机

如果样本不是线性可分的情形，则采用在输入空间构造最优分界面的方法求得的解不能使经验误差等于零，因此这种方法常常由于经验误差过大而失去意义。线性不可分问题有可能通过非线性变换转化为高维空间中的线性可分问题[36]，事实上，从输入空间映射到高维特征空间会增加线性可分的可能性。这大体上还能从相应的 VC 维看出。在 N 维空间 R^N 中，线性函数的 VC 维是 $N+1$，这表明存在着 $N+1$ 个点

$$x_i \in R^N \quad i=1,2,...,N+1 \tag{2.30}$$

使得对任意的 $y_i \in \{-1,+1\}$（$i=1,2,...,N+1$），由 $N+1$ 个样本点组成的训练集

$$\{(x_i, y_i),\ i=1,2,...,N+1\} \tag{2.31}$$

总是线性可分的。所以从直观上看，维数 N 越高，训练集线性可分的可能性越大。

因此，对于线性不可分问题，首先采用一个非线性映射 φ 把输入向量映射到一个高维特征空间，然后在高维特征空间构造最优分界面进行线性分类，映回到原空间之后就成了输入空间的非线性分类。

设用非线性变换函数

$$z_j = \varphi_j(x) \qquad j=1,2,...m \tag{2.32}$$

将输入空间中的向量 x 变换为 m 维特征空间中的向量 z（$m>n$），即

$$z = \varphi(x) = (\varphi_1(x), \varphi_2(x), ..., \varphi_m(x)) \tag{2.33}$$

要在特征空间中构造最优分界面，只涉及到计算两个向量的内积

$$< z^i \cdot z^j >= \sum_{k=1}^{m} \varphi_k(x_i)\varphi_k(x_j) = \varphi(x_i)^T \varphi(x_j) = < \varphi(x_i) \cdot \varphi(x_j) > \tag{2.34}$$

由于特征空间维数很高，直接利用公式（2.32）计算内积会很困难，因此支持向量机采用核函数 $\psi(x_i, x_j)$ 代替高维空间中的内积运算 $<\varphi(x_i)\cdot\varphi(x_j)>$。统计理论表明，只要对称函数 $\psi(x_i, x_j)$ 满足 Mercer 条件就可以作为核函数。关于 Mercer 条件有下面的定理[8]：

定理 2.2 对于任意的对称函数 $\psi(x,y)$，其某个特征空间中的内积运算的充分必要条件是，对于任意的不恒为零的 $h(x)$ 且 $\int h^2(x)dx < \infty$，有

$$\iint \psi(x,y)h(x)h(y)dxdy > 0 \tag{2.35}$$

目前常用的核函数有以下几种：

（1）线性核

$$\psi(x,y) = x^T y$$

（2）多项式核

$$\psi(x,y) = (x^T y + 1)^k$$

（3）高斯径向基核

$$\psi(x,y) = \exp\{-\|x-y\|_2^2 / \sigma^2\}$$

（4）Fourier 核

$$\psi(x,y) = \frac{\sin(N+\frac{1}{2})(x-y)}{\sin(\frac{1}{2}(x-y))}$$

（5）B 样条核

$$\psi(x,y) = B_{2N+1}(x-y)$$

根据以上分析，对于线性不可分情形，就是在高维特征空间中构造样本 $[\varphi(x_i), y_i]_{i=1}^l$ 的最优分界面，其中 $x_i \in R^n$，$\varphi(x_i) \in R^m$，$y_i \in \{-1,+1\}$。与 2.2.1 节中的内容相似，设分界面方程为

$$f(x) = < w \cdot \varphi(x) > + b = 0 \tag{2.36}$$

求最优分界面的问题可以表述为下列二次优化问题：

对于给定的训练样本 $\{x_i, y_i\}_{i=1}^l$ 和映射函数 φ，求使下列二次泛函取极小值的 w 和 b

$$\min \frac{1}{2}\|w\|^2 \tag{2.37}$$

约束条件为

$$y_i[< w \cdot \varphi(x_i) > + b] \geqslant 1 \quad i = 1,...,l \tag{2.38}$$

该问题的对偶优化问题为：对给定的训练样本 $\{x_i, y_i\}_{i=1}^l$ 和核函数 $\psi(x_i, x_j)$，求使下列二次函数取极大值的 α_i

$$\max_{\alpha} L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \psi(x_i,x_j) \tag{2.39}$$

约束条件为

$$\sum_{i=1}^{l} y_i\alpha_i=0 \qquad \alpha_i \geqslant 0\ ,\ \ i=1,2,...,l \tag{2.40}$$

上面的讨论是针对高维空间中样本是线性可分的情形，这时选择了能够正确划分高维空间中样本的超平面。现在讨论在高维空间中样本是线性不可分的情形，因为不存在这样的超平面，如果坚持用超平面进行划分，那么必须软化对间隔的要求，即允许有不满足约束条件 $y_i[<w\cdot\varphi(x_i)>+b]\geqslant+1$ 的样本点存在。通过引入松弛变量

$$\xi_i \geqslant 0 \qquad i=1,2,...,l \tag{2.41}$$

可得软化了的约束条件

$$y_i[<w\cdot\varphi(x_i)>+b]\geqslant 1-\xi_i \qquad i=1,2,...,l \tag{2.42}$$

显然，当 ξ_i 充分大时，高维空间中的样本点总可以满足上述约束条件，但是应该设法避免 ξ_i 取太大的值。为此我们在目标函数里对其进行惩罚，比如可以在目标函数中加入含有 $\sum_i \xi_i$ 的项。这样优化问题（2.37）和（2.38）就变成如下形式：

对于给定的训练样本 $\{x_i,y_i\}_{i=1}^{l}$ 和映射函数 φ，求使得下列二次泛函取极小值的 w 和 b

$$\min_{w,e} J(w,e)=\frac{1}{2}\|w\|^2+\gamma\sum_{i=1}^{l}\xi_i \tag{2.43}$$

约束条件为

$$y_i[<w\cdot\varphi(x_i)>+b]\geqslant 1-\xi_i \qquad i=1,...,l \tag{2.44}$$

其中 $\gamma>0$ 是一个惩罚参数。

上述问题的对偶问题为：对给定的训练样本 $\{x_i,y_i\}_{i=1}^{l}$、核函数 $\psi(x_i,x_j)$ 和惩罚参数 γ，求使下列二次函数取极大值的 α_i

$$\max_{\alpha} L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \psi(x_i,x_j) \tag{2.45}$$

约束条件为

$$\sum_{i=1}^{l} y_i \alpha_i = 0 \quad 0 \leqslant \alpha_i \leqslant \gamma, \ i = 1,2,...,l \tag{2.46}$$

2.3 支持向量回归机

回归问题与分类问题相似，已知一个包含 l 个样本 $\{x_i, y_i\}_{i=1}^{l}$ 的训练集合，其中第 i 个输入数据 $x_i \in R^n$ 且第 i 个输出数据 $y_i \in R$。这里的 y_i 并不限定取+1 或-1，而可以取任意实数。寻找一个实值函数 $f(x)$，以便用 $y = f(x)$ 来推断任一 x 所对应的 y 值，也就是要寻找一个实值函数 $f(x)$，使得该函数能够表达 y 对 x 的依赖关系。

2.3.1 线性支持向量回归机

与分类问题类似，我们先考虑可以用一个线性函数来近似训练集合中的 y 对 x 的依赖关系。设给定的训练样本集合为

$$S = \{(x_i, y_i) \mid x_i \in R^n, y_i \in R, \ i = 1,2,...,l\}$$

定义 2.2[38] 样本集 S 是 ε-线性近似的，如果存在一个超平面 $f(x) = < w \cdot x > + b$，其中 $w \in R$，$b \in R$，下面的式子成立

$$|y_i - f(x_i)| \leqslant \varepsilon \quad i = 1,2,...l$$

用 d_i 表示点 $(x_i, y_i) \in S$ 到超平面 $f(x)$ 的距离，则有

$$d_i = \frac{|< w \cdot x > + b - y_i|}{\sqrt{1 + \|w\|^2}}$$

因为 S 集合是 ε-线性近似的，所以有

$$|< w \cdot x > + b - y_i| \leqslant \varepsilon \quad i = 1,2,...,l$$

这样可以得到

$$\frac{|< w \cdot x > + b - y_i|}{\sqrt{1 + \|w\|^2}} \leqslant \frac{\varepsilon}{\sqrt{1 + \|w\|^2}} \quad i = 1,2,...l$$

于是有

$$d_i \leqslant \frac{\varepsilon}{\sqrt{1+\|w\|^2}} \quad i=1,2,...l$$

上式表明，$\frac{\varepsilon}{\sqrt{1+\|w\|^2}}$ 是 S 中的点到超平面的距离的上界。

定义 2.3[38] ε-线性近似集 S 的最优近似超平面是通过最大化 S 中的点到超平面的距离上界而得到的超平面。

由这个定义能够得出最优近似超平面是通过最大化 $\frac{\varepsilon}{\sqrt{1+\|w\|^2}}$ 得到的，即最小化 $\sqrt{1+\|w\|^2}$。因此，只要最小化 $\|w\|^2$ 就可以得到最优近似超平面。于是线性回归问题就转化为求下面公式的优化问题

$$\min \frac{1}{2}\|w\|^2 \tag{2.47}$$

约束条件为

$$\left|< w\cdot x_i > +b-y_i\right| \leqslant \varepsilon \quad i=1,2,...l \tag{2.48}$$

这是一个二次规划问题，我们通常不直接求解，而是求解它的 Lagrange 对偶问题。为此引入 Lagrange 函数：

$$L(w,b,\alpha,\alpha^*)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i(\varepsilon-y_i+< w\cdot x_i > +b)-\sum_{i=1}^{l}\alpha_i^*(\varepsilon+y_i-< w\cdot x_i > -b) \tag{2.49}$$

其中 $\alpha_i,\alpha_i^* \geqslant 0$，$i=1,2,...,l$

Lagrange 对偶问题为

$$\max_{\alpha,\alpha^*}\min_{w,b} L(w,b,\alpha,\alpha^*)$$

利用 Kuhn-Tucker 条件，函数 L 的极值应满足条件

$$\frac{\partial L}{\partial w}=0,\quad \frac{\partial L}{\partial b}=0$$

于是得到

$$w=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i \tag{2.50}$$

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \tag{2.51}$$

把式（2.50）和式（2.51）代入到式（2.49）中，得到原优化问题的对偶形式为

$$\min_{\alpha,\alpha^*}\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)<x_i\cdot x_j>+\varepsilon\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i) \tag{2.52}$$

约束条件为

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \tag{2.53}$$

$$\alpha_i^*,\alpha_i\geqslant 0,\quad i=1,2,...l \tag{2.54}$$

通过求解该对偶问题得到线性回归函数。

2.3.2 非线性支持向量回归机

非线性回归同分类情况相似，首先使用一个非线性映射 φ 把原始数据 x_i 映射到一个高维空间，然后在高维空间进行线性回归。在优化过程中涉及到高维空间中的内积运算，用一个核函数 $\psi(x_i,x_j)$ 代替内积 $<\varphi(x_i)\cdot\varphi(x_j)>$ 来实现非线性回归。因此，求非线性回归函数的问题转化为求下面的优化问题

$$\min\frac{1}{2}\|w\|^2 \tag{2.55}$$

约束条件为

$$\left|<w\cdot\varphi(x_i)>+b-y_i\right|\leqslant\varepsilon\qquad i=1,2,...l \tag{2.56}$$

同线性回归类似，该问题的 Lagrange 对偶问题为

$$\min_{\alpha,\alpha^*}\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\psi(x_i,x_j)+\varepsilon\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i) \tag{2.57}$$

约束条件为

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \tag{2.58}$$

$$\alpha_i^*,\alpha_i\geqslant 0,\quad i=1,2,...l \tag{2.59}$$

另外，考虑到可能存在误差引入两个松弛变量：

$$\xi_i,\xi_i^*\geqslant 0,\quad i=1,2,...,l$$

这时优化方程为

$$\min \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2.60}$$

约束条件为

$$< w \cdot \varphi(x_i) > + b - y_i \leqslant \xi_i^* + \varepsilon \qquad i = 1,2,...l \tag{2.61}$$

$$y_i - < w \cdot \varphi(x_i) > - b \leqslant \xi_i + \varepsilon \qquad i = 1,2,...l \tag{2.62}$$

$$\xi_i, \xi_i^* \geqslant 0, \quad i = 1,2,...,l \tag{2.63}$$

为了求解这个二次规划问题，引入 Lagrange 函数

$$L(w,b,\alpha,\alpha^*) = \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}\alpha_i(\xi_i + \varepsilon - y_i + < w \cdot \varphi(x_i) > + b)$$

$$-\sum_{i=1}^{l}\alpha_i^*(\xi_i^* + \varepsilon + y_i - < w \cdot \phi(x_i) > - b) - \sum_{i=1}^{l}\eta_i(\xi_i + \xi_i^*) \tag{2.64}$$

其中 $\alpha_i, \alpha_i^* \geqslant 0$，$i = 1,2,...,l$

函数 L 的极值应满足条件：

$$\frac{\partial L}{\partial w} = 0, \quad \frac{\partial L}{\partial b} = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0, \quad \frac{\partial L}{\partial \xi_i^*} = 0$$

于是得到

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i \tag{2.65}$$

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \tag{2.66}$$

$$\gamma - \alpha_i - \eta_i = 0 \qquad i = 1,2,...,l \tag{2.67}$$

$$\gamma - \alpha_i^* - \eta_i^* = 0 \qquad i = 1,2,...,l \tag{2.68}$$

把式（2.65）～式（2.68）代入式（2.64）中，得到 Lagrange 对偶问题为

$$\min_{\alpha,\alpha^*} \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\psi(x_i, x_j) + \varepsilon \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i) \tag{2.69}$$

约束条件为

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \tag{2.70}$$

$$0 \leqslant \alpha_i^*, \alpha \leqslant \gamma\ ,\quad i = 1,2,...l \tag{2.71}$$

2.4 本章小结

本章首先介绍了统计学习理论的核心内容，在此基础上按照从线性到非线性的顺序详细介绍了与支持向量机分类和支持向量机回归对应的优化问题及其 Lagrange 对偶问题，这两种问题是本文以下工作的基础。