第四章　支持向量机

4.1　简 介

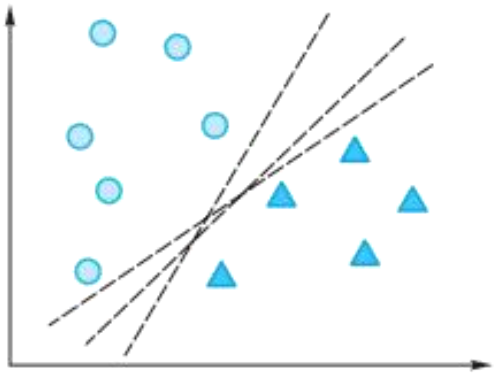
4.1 简 介

- 对给定样本数据集进行分类处理通常有多种办法, 如前面章节中介绍的GLM 或者 McGLM 等, 而支持向量机 (support vector machine, SVM) 作为众多分类器中的一种, 却有着其他分类器没有的优点, 下面我们就来介绍一下 SVM.
- 支持向量机是一种基于统计学习理论的有监督新型学习机, 是由苏联教授 Vladimir N.Vapnik 等[58] 于 1963 年在统计学习理论基础上提出的. 支持向量机理论最初用来解决两类线性可分问题, 之后逐步推广到多分类、非线性等问题. 与传统学习方法不同, 支持向量机是结构风险最小化方法的近似实现, 是借助最优化方法来解决机器学习分类问题的新工具. 在处理线性可分问题时, 其“对样本依赖小”等优点使其成为众多分类器当中的佼佼者.

4.2　SVM 算法

4.2.1 SVM 的基本内容

- 最初SVM被提出时, 是用来解决二分类问题的. 其主要内容是：对于给定的包含两个类别的样本数据集 $D=\{(\boldsymbol{X}_1, Y_1),(\boldsymbol{X}_2, Y_2), \cdots,(\boldsymbol{X}_n, \mathrm{Y}_n)\}, \boldsymbol{X}_i \in \mathbf{R}^p, Y_i \in \{-1, 1\}$, 确定一个超平面, 对数据集进行二分类处理, 如图 4.1.



  - 其分类判别式为

$$f(\boldsymbol{X})=\operatorname{sign}\left(\boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{X}+b\right).$$

- 可以看到, 不止有一个超平面能够把样本数据正确地分为两类, SVM 算法是在这众多超平面中挑选出鲁棒性最强的超平面 $\boldsymbol{\omega}^{*\mathrm{T}} \boldsymbol{X}+b^*$ 来作为最终的分类器.
- SVM 可分为三种不同的类别：线性可分 SVM、线性 SVM、非线性 SVM.下面分别对这三种类型的 SVM 进行介绍.

4.2.2 线性可分 SVM

- 若至少存在一个超平面, 可以将两类数据完全分开, 此时的 SVM 被称为线性可分 SVM. 线性可分 SVM 的原理就是要达到硬间隔最大化, 又被称作最大间隔分类器. 在样本点可以被超平面分为两类的基础上, 线性可分 SVM的目标是找到一个超平面, 使得每个样本点到该超平面的距离都足够大. 下面对算法理论进行详细阐述. 对于给定样本数据集 $D=\{(\boldsymbol{X}_1, \boldsymbol{Y}_1),(\boldsymbol{X}_2, \boldsymbol{Y}_2),\cdots,(\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$, $\boldsymbol{X}_i \in \boldsymbol{R}_p$, $Y_i \in \{-1, 1\}$. 在样本空间中, 划分超平面用如下线性方程来描述:

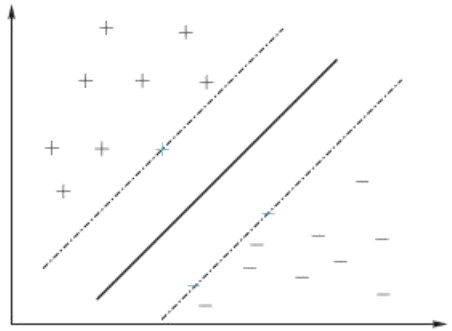
$$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X}+b=0,$$

  - 其中 $\boldsymbol{\omega}=\left(\omega_1,\omega_2,\cdots,\omega_p\right)^{\mathrm{T}}$ 为法向量, b 为位移项. 显然, 超平面可以由法向量 $\boldsymbol{\omega}$ 和位移 b 所确定, 记为 $(\boldsymbol{\omega}, b)$. 对于 $(\boldsymbol{X}_i, Y_i) \in \mathrm{D}$, 其到超平面 $(\boldsymbol{\omega}, b)$ 的距离可表示为

$$r_i=\frac{\left|\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X}_i+b\right|}{\|\boldsymbol{\omega}\|}. \tag{4.2.2}$$

4.2.2 线性可分 SVM

- 所有 r_i $(i = 1, 2, \cdots, n)$ 的最小值称为间隔, 可表示为式 (4.2.3). 而这些和超平面 $(\boldsymbol{\omega}, b)$ 距离最小的样本点称作支持向量 (如图 4.2 所示)



$$\text{margin} = \min_{1 \leqslant i \leqslant n} r_i . \tag{4.2.3}$$

- 为了使这个超平面更具鲁棒性, SVM 算法的目标是通过调整参数 $\boldsymbol{\omega}$ 和b, 找到以最大间隔把两类样本分开的超平面, 也称之为最大间隔超平面. 首先, 我们希望两类样本分别分割在该超平面的两侧; 其次, 两侧距离超平面最近的样本点到超平面的距离被最大化. 即

$$\max_{\omega, b} \quad \min_{1 \leqslant i \leqslant n} r_i . \tag{4.2.4}$$

4.2.2 线性可分 SVM

- 假设样本点可以被超平面分为两个类别, 显然对于 $(\boldsymbol{X}_i, Y_i) \in D$, 满足下面约束条件:

$$\begin{cases} \omega^{\mathrm{T}} \boldsymbol{X}_i + b > 0, & Y_i = 1, \\ \omega^{\mathrm{T}} \boldsymbol{X}_i + b < 0, & Y_i = 1-. \end{cases} \tag{4.2.4}$$

  - 可将其化简为

$$Y_i\left(\omega^{\mathrm{T}} \boldsymbol{X}_i + b\right) > 0. \tag{4.2.5}$$

  - 这一约束条件意味着所有样本点被超平面正确分类. 下面希望从能将样本数据集正确分类的超平面里选取特殊的一个超平面, 使得两类样本数据集到这一超平面的最短距离最大化, 即间隔最大化:

$$\max_{\omega, b} \min_{1 \leqslant i \leqslant n} r_i, \quad \text{s.t.} \quad Y_i\left(\omega^{\mathrm{T}} \boldsymbol{X}_i + b\right) > 0. \tag{4.2.6}$$

  - 在该约束条件下, 可以将 r_i 转化为

$$r_i = \frac{Y_i\left(\omega^{\mathrm{T}} \boldsymbol{X}_i + b\right)}{\|\omega\|}.$$

4.2.2 线性可分 SVM

- 优化目标函数转化为

$$\max_{\omega,b} \min_{1\leqslant i\leqslant n} r_i \frac{Y_i\left(\omega^{\mathrm{T}}\boldsymbol{X}_i+b\right)}{\|\omega\|}.$$

- 为了便于简化计算, 对超平面 (ω, b) 进行放缩变换, 使得 $\left|\omega^{\mathrm{T}}\boldsymbol{X}_i+b\right|\geq 1$, 即 $Y_i\left|\omega^{\mathrm{T}}\boldsymbol{X}_i+b\right|\geq 1.$

因此, 优化问题 (4.2.6) 转换为下面形式:

$$\max_{\omega,b}\frac{1}{\|\omega\|},\quad \text{s.t. } Y_i\left(\omega^{\mathrm{T}}\boldsymbol{X}_i+b\right)\geq 1\quad i=1,2,\cdots,n. \tag{4.2.7}$$

- 注意到目标函数中的 $\|\omega\|$ 在分母上, 所以可以将求最大值的约束优化问题转换成求最小值的凸二次规划问题:

$$\max_{\omega,b}\frac{1}{2}\|\omega\|^2,\quad \text{s.t. } Y_i\left(\omega^{\mathrm{T}}\boldsymbol{X}_i+b\right)\geq 1\quad i=1,2,\cdots,n. \tag{4.2.8}$$

(注: 此处的 $\frac{1}{2}$ 并不影响最终的计算结果.)

4.2.2 线性可分 SVM

- 上式即为线性可分 SVM 的最优化问题. 上述优化问题为不等式约束的优化问题, 可利用拉格朗日乘子法将原问题转化为对偶问题. 此外, 这样做可以更自然地引入核函数, 进而推广到非线性的分类问题. 对于更一般化的约束优化问题来说, 对偶问题可以将非凸问题转化为凸优化问题.
- 首先, 引入拉格朗日乘子向量 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_n)^{\mathrm{T}}$, 并写出拉格朗日函数

$$L(\omega, b, \boldsymbol{\lambda}) \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{n} \lambda_i \left(Y_i\left(\omega^{\mathrm{T}} \boldsymbol{X}_i + b\right) - 1\right), \qquad \lambda_i \geq 0. \tag{4.2.9}$$

- 此时, 可以得到原问题 (4.2.8) 等价于一个极小化极大问题:

$$\min_{\omega, b} \max_{\boldsymbol{\lambda}} L(\omega, b, \boldsymbol{\lambda}), \tag{4.2.10}$$

  - 进而, 可定义原问题 (4.2.8) 的对偶问题:

$$\max_{\boldsymbol{\lambda}} \min_{\omega, b} L(\omega, b, \boldsymbol{\lambda}). \tag{4.2.11}$$

4.2.2 线性可分 SVM

- 根据凸优化理论, 可以知道, 当原问题的目标函数和不等式约束函数是凸函数时, 在不等式约束函数严格可行情况下, 原问题最优解 $(\boldsymbol{\omega}^*, b^*)$ 与其对偶问题的最优解 $\boldsymbol{\lambda}^*$ 满足下面等式

$$L(\boldsymbol{\omega}^*, b^*, \boldsymbol{\lambda}^*) = \min_{\boldsymbol{\omega}, b} \max_{\boldsymbol{\lambda}} L(\boldsymbol{\omega}, b, \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \min_{\boldsymbol{\omega}, b} L(\boldsymbol{\omega}, b, \boldsymbol{\lambda}). \tag{4.2.12}$$

- 显然, 这里目标函数 $\frac{1}{2}\|\boldsymbol{\omega}\|^2$ 及其约束函数 $1 - Y_i(\boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{X}_i + b)$ $(i = 1, 2, \cdots, n)$ 是凸函数, 那么上述等式在 SVM 中仍然成立. 此时, 可以对原问题的对偶问题求最优解, 进而可以求出原问题的最优解.

- 根据式 (4.2.11) , 首先对 $\boldsymbol{\omega}$ 和 b 求极小值. 令 $L(\boldsymbol{\omega}, b, \boldsymbol{\lambda})$ 对 $\boldsymbol{\omega}$ 和 b 的偏导数为 0, 可以得到

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i Y_i$$

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = 0 \Rightarrow \boldsymbol{\omega} = \sum_{i=1}^{n} \lambda_i Y_i \boldsymbol{X}_i .$$

4.2.2 线性可分 SVM

- 将所得的关系代入 L($\boldsymbol{\omega}$, b, $\boldsymbol{\lambda}$), 式 (4.2.10) 转化为

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \boldsymbol{X}_i^T \boldsymbol{X}_j ,$$
$$\text{s.t.} \sum_{i=1}^{n} \lambda_i Y_i , \quad \lambda_i \geq 0. \tag{4.2.13}$$

- 式 (4.2.13) 是一个含等式约束的优化问题. 我们将其转化为求最小值问题, 上式等价于

$$\min_{\lambda} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \boldsymbol{X}_i^T \boldsymbol{X}_j - \sum_{i=1}^{n} \lambda_i ,$$
$$\text{s.t.} \sum_{i=1}^{n} \lambda_i Y_i , \quad \lambda_i \geq 0. \tag{4.2.14}$$

4.2.2 线性可分 SVM

- 可以发现, 上述对偶问题是一个二次规划问题, 可以采用序列最小优化(sequential minimal optimization, SMO) 算法求解. 序列最小优化算法, 其核心思想非常简单：每次只优化一个参数, 其他参数先固定住, 仅求当前这个优化参数的极值.
- 为提高求解效率, 其求解过程每次选择两个变量进行优化, 其他变量固定, 具体而言, 算法流程可描述为
  - (1) 选取两个待优化变量;
  - (2) 固定其他变量的值, 求解优化问题 (4.2.14)(直接令目标函数关于待优化变量梯度等于零即可);
  - (3) 不断重复第 (1)(2) 两个步骤, 直到算法收敛.
- SMO 算法一般选择违反 KKT 条件 (Karush-Kuhn-Tucker Conditions)最严重的样本所对应的变量作为第一个优化变量, 第二个变量的选择应当使得目标函数有足够大的下降. 一种启发式的选择方式为, 首先寻找与第一个变量所对应样本间隔最大的样本, 然后将该样本所对应的变量设置为第二个变量.
- 使用 SMO 算法可以求得对偶问题的最优解 $\boldsymbol{\lambda}=\left(\lambda_1^*,\lambda_2^*,\cdots,\lambda_n^*\right)^{\mathrm{T}}$. 根据式 (4.2.12), 对偶问题的最优解也是原问题的最优的拉格朗日乘子向量, 接下来把最优的拉格朗日乘子向量代入原问题求解即可.

4.2.2 线性可分 SVM

- 为了简化计算, 我们引入 KKT 条件. 已知 KKT 条件是判断等式约束和不等式约束的优化问题的必要条件. 对于只含不等式约束的优化问题来说,如果目标函数和约束函数是凸函数, 且不等式约束函数是可以满足的, 那么KKT 条件是这一不等式约束的优化问题的充要条件. 在 SVM 中, 原问题显然满足上述要求, 那么可以得到原问题的充要条件 (KKT 条件), 即

$$
\begin{cases}
\dfrac{\partial L}{\partial b}=0 \Rightarrow \sum\limits_{i=1}^{n} \lambda_i Y_i=0, \\
\dfrac{\partial L}{\partial \omega}=0 \Rightarrow \omega=\sum\limits_{i=1}^{n} \lambda_i Y_i \boldsymbol{X}_i, \\
\lambda_i \geq 0, \quad i=1,2,\cdots,n, \\
Y_i\left(\omega^{\mathrm{T}} \boldsymbol{X}_i+b\right)-1 \geq 0, \quad i=1,2,\cdots,n, \\
\lambda_i\left(Y_i\left(\omega^{\mathrm{T}} \boldsymbol{X}_i+b\right)-1\right)=0, \quad i=1,2,\cdots,n.
\end{cases}
$$

  - 从而求解原问题等价于求解上述 KKT 条件, 其中 λ_i 是拉格朗日乘数.

4.2.2 线性可分 SVM

- 分析 KKT 条件我们还能发现更多关于样本的性质. 对任意训练样本$(\boldsymbol{X}_i, Y_i)$ 总有 $\lambda_i = 0$ 或 $Y_i\left(\omega^{\mathrm{T}}\boldsymbol{X}_i + b\right) - 1 = 0$. 若 $\lambda_i = 0$, 则所对应的样本点不会出现在系数 $\boldsymbol{\omega}$ 中, 即该样本点不影响最终模型; 若 $\lambda_i > 0$, 则必有则所对应的样本点位于最大间隔边界上, 是一个支持向量. 这显示出支持向量机的一个重要性质：训练完成后, 大部分的训练样本都不需要保留, 最终模型仅与支持向量有关.

- 将对偶问题的最优解 $\boldsymbol{\lambda}^* = \left(\lambda_1^*, \lambda_2^*, \cdots, \lambda_n^*\right)^{\mathrm{T}}$ 代入上述 KKT 条件等式,可以得到 $\boldsymbol{\omega}^*$,

$$\omega^* = \sum_{i=1}^{n} \lambda_i^* Y_i \boldsymbol{X}_i . \tag{4.2.15}$$

  - 为了得到最优决策平面, 还需求解 b^* . 注意到对任意支持向量 $(\boldsymbol{X}_k, Y_k)$, 都有 $\lambda_k^* > 0$ 且 $Y_i\left(\omega^{\mathrm{T}}\boldsymbol{X}_i + b\right) - 1 = 0$; 对任意非支持向量 $(\boldsymbol{X}_l, Y_l)$, 都有 $\lambda_l^* = 0$.

- 那么此时我们可以得到：

$$\omega^* = \sum_{k \in K} \lambda_k^* Y_k \boldsymbol{X}_k , \tag{4.2.16}$$

  - $K = \left\{ i \middle| \lambda_i^* > 0,\ \ i = 1, 2, \cdots, n \right\}$ 为所有支持向量的下标集.

4.2.2 线性可分 SVM

- 将 ω^* 代入 $Y_s\left(\omega^{\mathrm{T}}\boldsymbol{X}_s+b\right)-1=0$，其中 $(\boldsymbol{X}_s, Y_s)$ 是一支持向量，进而可以得到

$$Y_s\left(\sum_{k\in K}\lambda_k^* Y_k\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_s+b\right)=1.$$

  - 由于 $Y_s^2=1$，等式两边同时乘 Y_s 可以解得 b^*. 理论上可选取任意支持向量通过上式解得 b^*，但在现实任务中，常采用取所有支持向量的平均值的做法,即,

$$b^*=\frac{1}{|K|}\sum_{s\in K}\left(Y_s-\sum_{k\in K}\lambda_k^* Y_k\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_s\right).$$

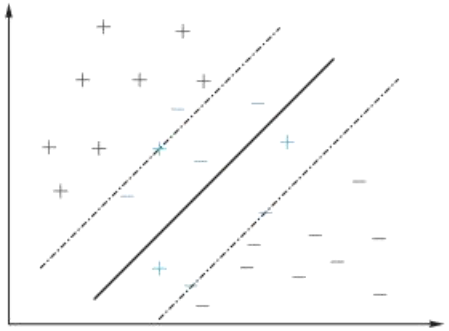
  - 最终得到最优决策平面为

$$\omega^{*\mathrm{T}}X+b^*=\sum_{i\in K}\lambda_i^* Y_i\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}+\frac{1}{|K|}\sum_{s\in K}\left(Y_s-\sum_{k\in K}\lambda_k^* Y_k\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_s\right).$$

  - 观察上述式子可以发现，最优决策平面并不需要用所有的样本来计算，而只需要用到支持向量. 此外，最优决策平面与 $\boldsymbol{X}_i^{\mathrm{T}}X$ 有关，这为将数据映射到高维空间，引入核函数做好了铺垫.

4.2.3 软间隔与线性 SVM

- 在实际问题当中常存在一些“噪声点”(如图 4.3 所示), 采用线性可分SVM 解决此类问题的过程中试图把这些“噪声点”也正确分类, 往往影响模型的泛化能力, 所以应当对模型放松要求, 允许其在分类时出一点错误, 这样会使模型的鲁棒性更好, 这也是线性 SVM 的基本出发点：即划分超平面虽然不能完全分开所有样本, 但是可以使绝大多数样本正确被分类, 其目标是软间隔最大化.



- 假设划分超平面为

$$f(\boldsymbol{X})=\omega_1 X_1+\omega_2 X_2+\cdots+\omega_p X_p+b=\omega^{\mathrm{T}}\boldsymbol{X}+b=0, \tag{4.2.17}$$

4.2.3 软间隔与线性 SVM

- 由于存在样本点不能满足式 (4.2.7) 中的约束条件, 使得该约束不成立, 因此在线性 SVM 中, 对每一个样本引入松弛因子 ζ_i, 此时约束条件被放宽为

$$Y_i f(\boldsymbol{X}_i) = Y_i\left(\omega^{\mathrm{T}}\boldsymbol{X}_i + b\right) \geq 1 - \zeta_i, \quad i = 1,2,\cdots,n.$$

- 上述约束条件允许样本位于间隔区域内, 也允许出现错误分类. 为了使在最大化间隔的同时不满足约束的点尽可能少, 线性 SVM 优化问题可写成如下形式:

$$\begin{aligned} &\min_{\omega,b,\zeta} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\zeta_i, \\ &\text{s.t.} \quad Y_i\left(\omega^T\boldsymbol{X}_i + b\right) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, i = 1,2,\cdots,n. \end{aligned} \tag{4.2.18}$$

- 其中 C 是可调节控制参数, 在最小化目标函数处加了 $C\sum_{i=1}^{n}\zeta_i$, 相当于施加了一个惩罚项. 当 C 值越大时, 对 ζ_i 惩罚越大; 反之, C 值越小, 对 ζ_i 惩罚越小.

4.2.3 软间隔与线性 SVM

- 最优化问题 (4.2.18) 是一个凸二次规划问题, 可以通过拉格朗日乘子法进行求解, 拉格朗日函数为

$$L(\omega,b,\boldsymbol{\zeta},\boldsymbol{\lambda},\boldsymbol{\beta})=\frac{\|\omega\|^2}{2}+C\sum_{i=1}^{n}\zeta_i-\sum_{i=1}^{n}\lambda_i\left(Y_i\left(\omega^T\boldsymbol{X}_i+b\right)-1+\zeta_i\right)-\sum_{i=1}^{n}\beta_i\zeta_i, \tag{4.2.19}$$

  - 其中 $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_n)^{\mathrm{T}}$, $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_n)^{\mathrm{T}}$. 分别对 $\boldsymbol{\omega}$, b, ζ_i 求导数, 并令值为零, 可以得到

$$\begin{aligned}&\omega=\sum_{i=1}^{n}\lambda_iY_i\boldsymbol{X}_i,\\&\sum_{i=1}^{n}\lambda_iY_i=0,\quad C-\lambda_i-\beta_i=0.\end{aligned} \tag{4.2.20}$$

  - 将上面结果代入拉格朗日函数 (4.2.19), 可以得到如下对偶问题:

$$\begin{aligned}&\max_{\lambda}\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_jY_iY_j\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j+\sum_{i=1}^{n}\lambda_i,\\&\text{s.t.}\ \sum_{i=1}^{n}\lambda_iY_i=0,\qquad 0\le\lambda_i\le C,\qquad \mathrm{i}=1,2,\cdots,n.\end{aligned} \tag{4.2.21}$$

4.2.3 软间隔与线性 SVM

- 对偶问题 (4.2.21) 等价于

$$\begin{aligned} &\min_{\lambda} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}\lambda_i\lambda_j Y_i Y_j \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j - \sum_{i=1}^{n}\lambda_i, \\ &\text{s.t.} \ \sum_{i=1}^{n}\lambda_i Y_i = 0, \quad 0 \le \lambda_i \le C, \quad \mathrm{i} = 1,2,\cdots,n. \end{aligned} \tag{4.2.22}$$

- 求解优化问题 (4.2.22), 得到最优解 $\boldsymbol{\lambda}^* = \left(\lambda_1^*, \lambda_2^*, \cdots, \lambda_n^*\right)^{\mathrm{T}}$, 代入式 (4.2.20),即可得到 $\boldsymbol{\omega}^*$. 由于原始问题是凸二次规划问题, 其解满足 KKT 条件, 即

$$\beta_i \zeta_i = 0, \quad i = 1,2,\cdots,n, \tag{4.2.23}$$

$$\lambda_i\left(Y_i\left(\omega^{\mathrm{T}}\boldsymbol{X}_i + b\right) - 1 + \zeta_i\right) = 0, \quad i = 1,2,\cdots,n. \tag{4.2.24}$$

- 在最优解 $\boldsymbol{\lambda}^*$ 中, 取分量 λ_k^* 满足 $0 < \lambda_k^* < C$, 由式 $C - \lambda_k^* - \beta_k^* = 0$ 得 $0 < \beta_k^* < C$, 再由式 (4.2.23) 和 (4.2.24) 易计算得到 b^* , 进而得到划分超平面

4.2.3 软间隔与线性 SVM

- 由 ω 的表达式可知, 当 $\lambda_i^* = 0$ 时, 该样本不会对最终的模型产生影响,当 $0 < \lambda_i^* \leqslant C$ 时, 该样本是支持向量, 包括以下几种类型:
  - (1) 若 $0 < \lambda_i^* < C$, 则 $\beta_i > 0$, 进而由式 (4.2.23) 可以得到 $\zeta_i = 0$, 此时 $\boldsymbol{X}_i$ 位于最大间隔边界上;
  - (2) 若 $\lambda_i^* = C, 0 < \zeta_i < 1$, 则由式 (4.2.24) 知, $Y_i\left(\omega^{*\mathrm{T}}\boldsymbol{X}_i + b^*\right) = 1 - \zeta_i > 0$.那么, 此时 Xi 分类正确, 并且位于划分超平面和间隔边界之间;
  - (3) 若 $\lambda_i^* = C, \zeta_i = 1$, 则由式 (4.2.24) 知, $Y_i\left(\omega^{*\mathrm{T}}\boldsymbol{X}_i + b^*\right) = 1 - \zeta_i = 0$.此时 $\boldsymbol{X}_i$ 位于划分超平面上;
  - (4) 若 $\lambda_i^* = C, \zeta_i > 1$, 则由式 (4.2.24) 知, $Y_i\left(\omega^{*\mathrm{T}}\boldsymbol{X}_i + b^*\right) = 1 - \zeta_i < 0$.此时 $\boldsymbol{X}_i$ 被分类错误.

4.2.4 核函数与非线性 SVM

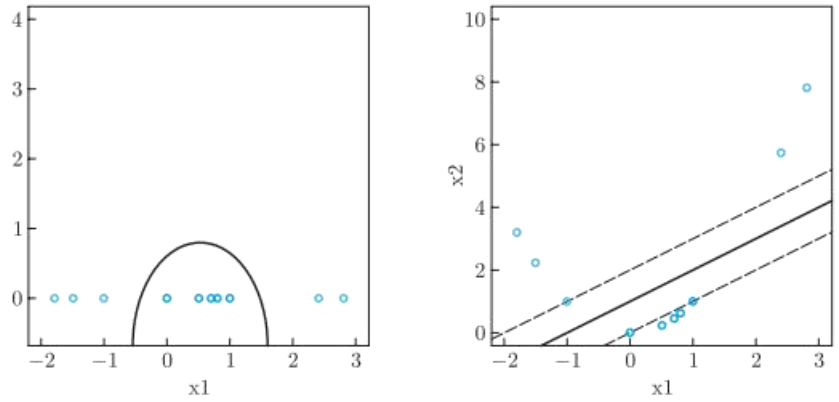
- 一般地, 如果包含两个类别的样本集之间存在线性边界, 那么建立线性SVM 可以得到较好的分类效果. 但是一些情况下, 如果问题本身不是线性可分的, 即边界是非线性的, 那么线性 SVM 往往效果不佳, 此时需建立非线性SVM 对样本数据进行非线性分类.
- 首先看一个简单的例子, 如图 4.4(a) 所示, 一维空间中的样本点属于两个类别, 分别用不同颜色表示不同类别. 在这种情况下, 无法用一个点 (即一维空间的超平面) 来将不同类别的样本分开, 但是可用一条复杂的曲线将它们分为两个类别. 显然, 分类边界不是线性的.
- 为了解决上述非线性可分问题, 尝试将自变量的二次项添加到超平面中,即此时的划分面是

$$\left\{X : f\left(X\right)=\omega_1 X+\omega_2 X^2+b=0\right\}. \tag{4.2.25}$$

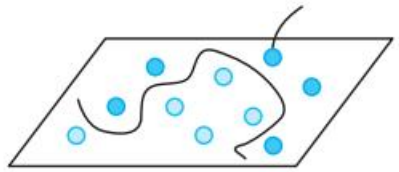
  - 式 (4.2.25) 中, 可以将 X 看作一个变量 X_1, X_2 看作另一个变量 X_2, 此时变成二维空间的线性问题. 如图 4.4 (b), 在构造出的二维空间中, 样本点可以用一个线性超平面 (即图中的实线) 分为两个类别.

4.2.4 核函数与非线性 SVM

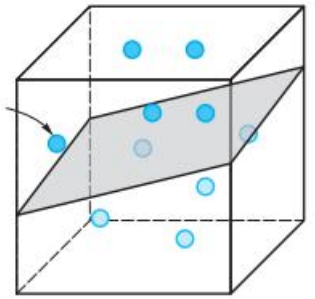
- 因此, 在二维空间中, 线性SVM 是有效的. 同理, 可将上述问题扩展到 p 维空间中, 此外, 还可使用不同类型的多项式, 如三次、四次甚至是更高阶, 以及交叉项构造特征空间, 进而在新构造的特征空间中建立线性超平面.



▶ 图 4.4
非线性 SVM 图例



▶ 图 4.5
线性不可分的数据



▶ 图 4.6
高维空间线性可分

- 总的来说, 对于线性不可分的数据 (如图 4.5), 可对样本进行变换, 将其映射到高维特征空间, 使得样本集在高维空间中线性可分, 如图 4.6 所示.

4.2.4 核函数与非线性 SVM

- 如果可以找到合适的特征空间, 便可以将原问题通过特征空间中的线性超平面进行划分. 然而实际问题中, 一方面, 构造合适的特征空间是非常困难的; 另一方面, 特征空间的构造方法可能并不唯一, 如果处理不当, 将会得到维数较大的特征空间, 此时的计算量将变得复杂. 因此有必要寻找合适的构造特征空间的方法, 使得在新的特征空间中能有效求解得到线性超平面.

- 核方法通过核函数 (kernel) 构造特征空间, 使得在新的特征空间中能有效求解线性超平面. 在介绍核函数之前, 有必要先对内积的概念进行介绍.

- 两个样本 $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^{\mathrm{T}}$ 和 $\boldsymbol{X}_k = (X_{k1}, \cdots, X_{kp})^{\mathrm{T}}$ 的内积定义为

$$\left\langle \boldsymbol{X}_i, \boldsymbol{X}_k \right\rangle = \sum_{j=1}^{p} X_{ij} X_{kj}. \tag{4.2.26}$$

4.2.4 核函数与非线性 SVM

- 通过前几小节的证明可以发现, 线性支持向量机的对偶问题可以描述成内积的形式, 即,

$$\begin{aligned} &\min_{\lambda} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j \left\langle \boldsymbol{X}_i \boldsymbol{X}_j \right\rangle - \sum_{i=1}^{n}\lambda_i, \\ &\text{s.t.} \quad \sum_{i=1}^{n}\lambda_i Y_i = 0, \qquad 0 \le \lambda_i \le C, \qquad \mathrm{i} = 1,2,\cdots,n. \end{aligned} \tag{4.2.27}$$

- 通过计算可得到对偶问题的拉格朗日乘子 λ_i, 根据 KKT 条件, 可以计算最优决策平面并将其描述为内积的形式:

$$f(\boldsymbol{X}) = \sum_{i=1}^{n}\lambda_i Y \left\langle \boldsymbol{X}_i \boldsymbol{X}_j \right\rangle + b. \tag{4.2.28}$$

- 上式中有n个参数 $\lambda_i, i = 1, 2, \cdots, n$,每个训练样本对应一个参数. 为了得到$f(\mathbf{X})$, 需要计算$\mathbf{X}$与每个训练样本$\boldsymbol{X}_i$之间的内积, 但事实证明, 有且仅有支持向量所对应的 λi 是非零的. 若用 S 表示支持向量样本点指标的集合, 那么式 (4.2.28) 可以改写成

$$f(\boldsymbol{X}) = \sum_{i=1}^{}\lambda_i Y_i \left\langle \boldsymbol{X}, \boldsymbol{X}_j \right\rangle + b. \tag{4.2.29}$$

4.2.4 核函数与非线性 SVM

- 式 (4.2.29) 的求和项比式 (4.2.28) 少得多. 总而言之, 我们仅需知道内积便可以计算 $f(\boldsymbol{X})$.

- 对于非线性可分问题, 我们的初衷是找到一个映射 ψ, 将样本点 $\boldsymbol{X}_i$ 映射为新的高维特征空间的特征向量 $\psi(\boldsymbol{X}_i)$, 进而通过其对偶问题求解相应拉格朗日乘子. 最后, 考虑原问题的 KKT 条件, 求解最优决策平面.

- 在上述步骤中, 可以发现在原空间内, 对偶问题及原问题的 KKT 条件求解以内积的形式出现. 自然的, 在新的特征空间内, 只需知道〈$\psi(\boldsymbol{X}_i)$, $\psi(\boldsymbol{X}_j)$〉便可以在新的特征空间中求解原问题的对偶问题, 然后得到最优决策平面.高维空间中内积的计算较为复杂, 可通过核技巧直接定义核函数 $K(\boldsymbol{X}_i, \boldsymbol{X}_j)$=〈$\psi(\boldsymbol{X}_i)$, $\psi(\boldsymbol{X}_j)$〉来解决复杂计算问题, 通俗来说, 就是将求解映射 ψ 的问题转化为选择合适核函数 $K(\boldsymbol{X}_i, \boldsymbol{X}_j)$ 的问题.

- 对于非线性可分问题, 考虑核技巧后, 此时 $f(\boldsymbol{X})$ 变为

$$f(\boldsymbol{X})=\sum_{i=1}\lambda_i Y_i K\left\langle \boldsymbol{X}, \boldsymbol{X}_j \right\rangle + b, \tag{4.2.30}$$

- 其中$K(\boldsymbol{X}_i, \boldsymbol{X}_j)$ 被称为核函数.

4.2.4 核函数与非线性 SVM

- 非线性 SVM 的处理方法是构造核函数以代替高维空间中的内积计算,在高维特征空间中解决原始空间中线性不可分的问题. 具体来说, 在线性不可分的情况下, 首先选择一个合适的核函数, 核函数的作用是避免在高维空间中进行内积计算, 然后在高维空间中执行线性可分的 SVM, 最终计算出最优的划分超平面, 从而达到把低维空间中线性不可分的数据集进行分类的目的.
- 核函数的选择至关重要, 将影响 SVM 对数据集的分类效果. 实际问题中, 应当根据问题本身特征选择合适的核函数. 表 4.1 给出了常用的几种核函数.

名称	表达式	参数
线性核	$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j$	
多项式核	$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j)^d$	$d \geqslant 0$ 为多项式的次数
高斯核	$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp\left(-\frcd{\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2}{2\sigma^2}\right)$	$\sigma>0$ 为高斯的带宽 (width)
拉普拉斯核	$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp\left(-\frac{\|\boldsymbol{X}_i - \boldsymbol{X}_j\|}{\sigma}\right)$	$\sigma>0$
Sigmoid 核	$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \tanh\left(\beta \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_j + \theta\right)$	tanh 为双曲正切函数,$\beta > 0, \theta > 0$

4.2.4 核函数与非线性 SVM

- 注意, 关于核函数的选择一直以来都是支持向量机研究的热点, 通常情况下, 高斯核函数是使用最多的. 此外, 采用核函数而不是直接将样本集映射到特征空间的优势在于, 无须明确指明映射函数, 并且可以避免计算高维空间中的内积, 大大降低计算量.

4.3　SVM 与逻辑斯谛回归的关系

4.3 SVM 与逻辑斯谛回归的关系

- 本节讨论 SVM 与逻辑斯谛回归的关系. 首先, 为了建立支持向量机分类器

$$f(\boldsymbol{X}) = b + \omega_1 X_1 + \cdots + \omega_p X_p,$$

  - 最优化问题 (4.2.18) 采用软间隔最大化的学习策略, 进而得到分隔超平面以及决策函数. 若将松弛变量 ζ_i 取为如下形式:

$$\zeta_i = \begin{cases} 1 - Y_i f(\boldsymbol{X}_i), & Y_i f(\boldsymbol{X}_i) < 1, \\ 0, & Y_i f(\boldsymbol{X}_i) \geq 1, \end{cases}$$

  - 即

$$\zeta_i = \max\{0, 1 - Y_i f(\boldsymbol{X}_i)\}. \tag{4.3.1}$$

  - 则问题 (4.2.18) 等价于如下优化问题:

$$\min_{\omega, b} \gamma \|\boldsymbol{\omega}\|^2 + \sum_{i=1}^{n} \max\{0, 1 - Y_i f(\boldsymbol{X}_i)\}. \tag{4.3.2}$$

4.3 SVM 与逻辑斯谛回归的关系

- 其中 γ 为调节控制参数, $\gamma\|\omega\|^2$ 是岭回归的惩罚项, 这一项需要根据偏差的关系来确定. 下面阐述问题 (4.3.2) 与 (4.2.18) 的等价性:

- 由 ζ_i 的定义, 显然问题 (4.2.18) 中的两个约束条件均成立. 又因为问题(4.3.2) 可写为

$$\min_{\omega,b} \gamma\|\boldsymbol{\omega}\|^2 + \sum_{i=1}^{n} \zeta_i, \tag{4.3.3}$$

  - 所以, 如果取 $\gamma = \dfrac{1}{2C}$, 那么可转化为

$$\min_{\omega,b} \frac{1}{C}\left(\frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n} \zeta_i\right),$$

  - 问题 (4.3.2) 可转化为 (4.2.18).

- 反之, 当 (4.2.18) 中的 ζ_i 取为 (4.3.1) 时, 可转化为 (4.3.2). 因此 (4.3.2)与 (4.2.18) 是等价的.

4.3 SVM 与逻辑斯谛回归的关系

- (4.2.18) 是等价的.令 $P(\boldsymbol{\omega})=\|\boldsymbol{\omega}\|^2$, $L(Y_i f(\boldsymbol{X}_i))=\max\{0,1-Y_i f(\boldsymbol{X}_i)\}$, 问题 (4.3.2) 可转化为如下的“损失函数 + 惩罚”的形式:

$$L(Yf(\boldsymbol{X}))=\max\{0,1-Yf(\boldsymbol{X})\}$$

  - 的损失函数称为 hinge **损失函数** (铰链损失函数). 这类损失函数的特点是,对于边界外且分类正确的样本点, 损失为零; 对于边界上的样本点以及分类错误的样本点, 损失是线性的.

- 对于逻辑斯谛回归, 加入 L_2 正则化项目后, 其优化函数变为: 使用的损失函数为

$$L(\boldsymbol{\omega})=-\frac{1}{n}\sum_{i=1}^{n}[Y_i\log(h_\omega(\boldsymbol{X}_i))+(1-Y_i)\log(1-h_\omega(\boldsymbol{X}_i))]+\gamma\|\boldsymbol{\omega}\|^2,$$

  - 其中, n 是样本数量, Y_i 是第 i 个样本的真实标签 (0 或 1), $h_\omega(X_i)=\frac{1}{1+e^{-\omega TX_i}}$ 是第 i 个样本的预测概率 (由 Sigmoid 函数得到).

4.3 SVM 与逻辑斯谛回归的关系

- SVM 与逻辑斯谛回归的目标是都减少 “错误率”. SVM 是寻找最优划分超平面降低错误率, 其损失函数为 hinge 损失; 逻辑斯谛回归通过最大化样本属于其真实类别的概率来降低错误率, 其损失函数为负对数似然. 两者的正则化项都是 L_2 正则.
- 因此, SVM 和逻辑斯谛回归的结果通常是非常接近的, 对于一个给定的问题, 该如何选择是使用 SVM 还是逻辑斯谛回归呢？这个问题更多时候需要根据实际数据选择合适的算法, 当然也有一些特殊情形：(1) 当类别区分度较高时, 可以选择 SVM; (2) 如果想要得到估计的概率, 那么就需要选择逻辑斯谛回归; (3) 对于决策边界是非线性的情况, 核函数的 SVM 方法应用更加广泛.

4.4　支持向量回归

4.4 支持向量回归

- 本节将介绍如何将支持向量机扩展到回归问题中, 称为支持向量回归(support vector regression, SVR), 其思想与分类问题类似, 不同之处在于支持向量回归的目的是要寻找一个超平面, 在距离超平面 ε 范围内尽可能地包含最多的样本点. 传统回归方法通过计算预测值与真实值差别计算损失, 只要两者不一致, 就会产生损失; 而支持向量回归容许预测值与真实值之间存在 ε 的差别, 仅当差别值大于 ε 时才会产生损失
- 因此, 可引入如下ε **不敏感损失函数**:

$$L_{\varepsilon}(r)=\begin{cases}0, & |r|\geq\varepsilon,\\ |r|-\varepsilon, & |r|>\varepsilon.\end{cases}$$

  - 此时, SVR 问题的优化目标函数可表示为

$$\min_{\omega,b}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{n}l_{\varepsilon}\left(f\left(X_i\right)-Y_i\right), \tag{4.4.1}$$

  - 其中, $\|\boldsymbol{\omega}\|$ 表示权重的向量范数, C 是正则化参数, ε 是预定义间隔.

4.4 支持向量回归

- 引入两个松弛变量 ζ_i 和 $\hat{\xi}_i$，分别表示上下两侧的松弛程度，(4.4.1) 可变化为

$$\begin{aligned}
&\min_{\omega,b}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{n}\left(\zeta_i+\hat{\xi}_i\right),\\
&\text{s.t.}\quad f\left(\boldsymbol{X}_i\right)-Y_i\le\varepsilon+\zeta_i,\\
&Y_i-f\left(\boldsymbol{X}_i\right)\le\varepsilon+\hat{\xi}_i,\\
&\zeta_i\ge 0,\qquad \hat{\xi}_i\ge 0,\qquad i=1,2,3,\cdots,n.
\end{aligned}\tag{4.4.2}$$

  - 利用拉格朗日乘子法，推导上述优化问题的对偶问题. 首先写出拉格朗日函数，

$$\begin{aligned}
&L\left(\boldsymbol{\omega},b,\boldsymbol{\zeta},\hat{\boldsymbol{\xi}},\boldsymbol{\mu},\hat{\boldsymbol{\mu}},\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}\right)\\
&=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{n}\left(\zeta_i+\hat{\xi}_i\right)-\sum_{i=1}^{n}\left(\mu_i\zeta_i+\hat{\mu}_i\hat{\xi}_i\right)+\\
&\sum_{i=1}^{n}\alpha_i\left(f\left(\boldsymbol{X}_i\right)-Y_i-\varepsilon-\zeta_i\right)+\sum_{i=1}^{n}\hat{\alpha}_i\left(Y_i-f\left(\boldsymbol{X}_i\right)-\varepsilon-\zeta_i\right),
\end{aligned}\tag{4.4.3}$$

4.4 支持向量回归

- 其中 $\boldsymbol{\mu},\hat{\boldsymbol{\mu}},\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}\geq 0.$. 分别对 ω,b,ζ_i和$\hat{\zeta}_i$ 求偏导并令其为零得到,

$$\frac{\partial L}{\partial \boldsymbol{\omega}}=\boldsymbol{\omega}-\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)\boldsymbol{X}_i=0, \tag{4.4.4}$$

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)=0, \tag{4.4.5}$$

$$\frac{\partial L}{\partial \zeta_i}=C-\mu_i-\alpha_i=0, \tag{4.4.6}$$

$$\frac{\partial L}{\partial \zeta_i}=C-\hat{\mu}_i-\hat{\alpha}_i=0. \tag{4.4.7}$$

- 将上述等式代入式 (4.4.3) 可得,

$$Q(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}})=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j-\varepsilon\sum_{i=1}^{n}(\alpha_i+\hat{\alpha}_i)+\sum_{i=1}^{n}Y_i(\hat{\alpha}_i-\alpha_i).$$

4.4 支持向量回归

- 那么, 可以得到对偶问题,

$$\min_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} Q(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}), \quad \text{s.t.} \quad \sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)=0,\ \boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}\in[0,C]. \tag{4.4.8}$$

- KKT 条件为

$$\begin{aligned}
\alpha_i\left(f(X_i)-Y_i-\varepsilon-\zeta_i\right)&=0,\\
\hat{\alpha}_i\left(Y_i-f(X_i)-\varepsilon-\zeta_i\right)&=0,\\
(C-\alpha_i)\zeta_i&=0,\\
(C-\hat{\alpha}_i)\hat{\zeta}_i&=0,\\
\alpha_i\hat{\alpha}_i&=0,\\
\zeta_i\hat{\zeta}_i&=0.
\end{aligned} \tag{4.4.9}$$

4.4 支持向量回归

- 接下来, 利用 SMO 方法求解上述对偶问题, 得到 α_i和$\hat{\alpha}_i$ 的值, 并由等式(4.4.4) 可得 SVR 的解为

$$f(\boldsymbol{X})=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}+b. \tag{4.4.10}$$

- 显然, SVR 的支持向量为所有使得 $\hat{\alpha}_i-\alpha_i\neq 0$ 的样本, 这些样本位于间隔外面, 故 SVR 解依然具有稀疏性.

- 那么, 如何求解b呢? 事实上, 得到所有α_i 的值后, 任取$0<\alpha_j<\mathrm{C}$, 根据 KKT 条件必然有 $\zeta_j=0$, 此时可以得到,

$$b=Y_j+\varepsilon-\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j. \tag{4.4.11}$$

- 在实际求解中, 一般选取多个满足 $0<\alpha_j<\mathrm{C}$ 的样本, 求解得到多个 b 后求平均值, 作为最终 b 的取值.

4.5　SVM 实践

实践代码