模式识别与机器学习

Pattern Recognition & Machine Learning

第4讲 支持向量机

- 本讲学习目标

  - ✓理解大间隔原理
  - ✓掌握基本的支持向量机分类模型
  - ✓能够熟练运用拉格朗日对偶优化技术
  - ✓掌握数据线性不可分情形下的分类模型，以及核方法的建模原理
  - ✓理解支持向量机回归的原理
  - ✓了解支持向量机的模型扩展

目录

- 大间隔原理

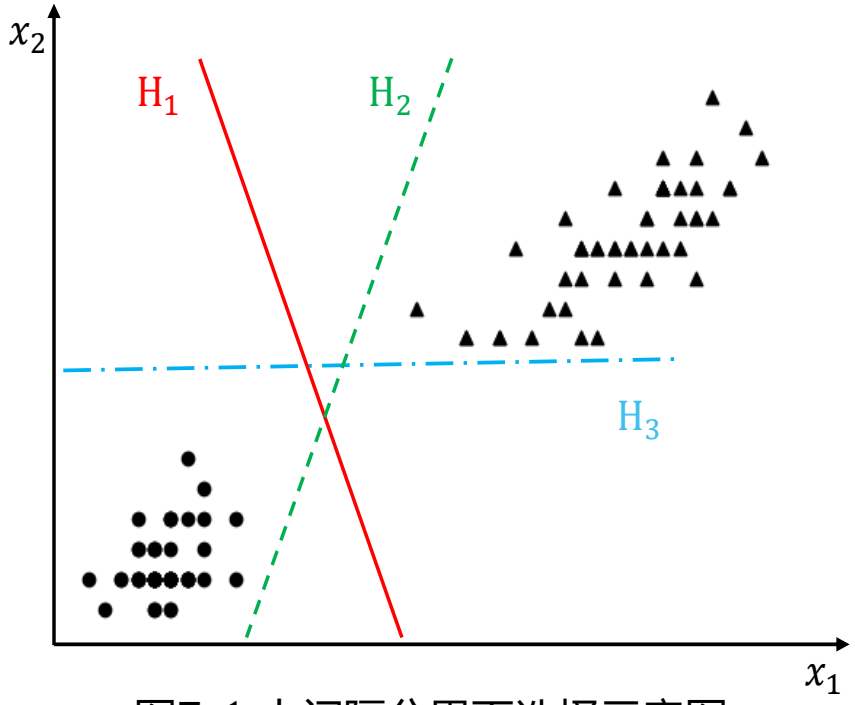


图7-1 大间隔分界面选择示意图
(H1是最大间隔分界面)

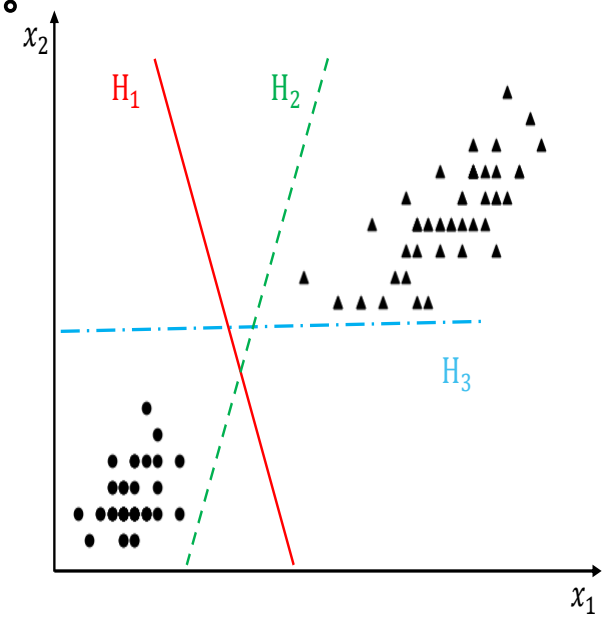
目录

- 基本分类模型

给定一个面向两类分类问题的线性可分训练集，其中包含N个样本$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$。通常，可令标签$y \in \{+1, -1\}$。
需要学习到一个分类超平面，设对应的参数化表示为

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = 0,$$

其中，$\mathbf{w}$是超平面的法向量，标量b是偏差参数。



训练完成得到参数以后

$$h(\mathbf{w}) = sign[\mathbf{w}^\top \mathbf{x} + b],$$

- 大间隔思想

任意一点x到超平面的距离为

$$d = \frac{|\mathbf{w}^{\top}\mathbf{x}+b|}{\|\mathbf{w}\|} = \frac{y(\mathbf{w}^{\top}\mathbf{x}+b)}{\|\mathbf{w}\|}.$$

考虑到偏差参数b的灵活性，可以认为距离d的大小与$\mathbf{w}$的长度无关，只与$\mathbf{w}$的方向有关系。

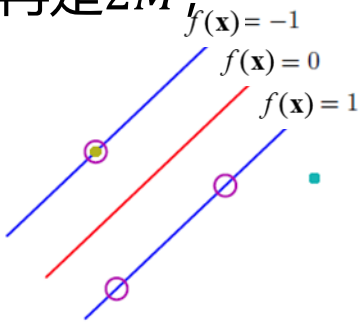
因此，我们可以固定参数向量 $\mathbf{w}$ 的长度为1，$\|\mathbf{w}\|=1$。定义两个超平面：$\mathbf{w}^{\top}\mathbf{x}+b=M$ 和 $\mathbf{w}^{\top}\mathbf{x}+b=-M$ ，之间的距离 $2M$ 被称为**间隔**（margin）

由于训练数据线性可分，我们希望能找到一个线性函数$f(\mathbf{x})$对所有的样本都满足$y_i f(\mathbf{x}_i) > 0$，而且可以确定这样的函数一定存在。那么，基于大间隔原理的分类模型的优化表达为

$$\begin{aligned} \max_{\mathbf{w},b} \quad & 2M \\ s.t. \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq M, \\ & \|\mathbf{w}\| = 1. \end{aligned}$$

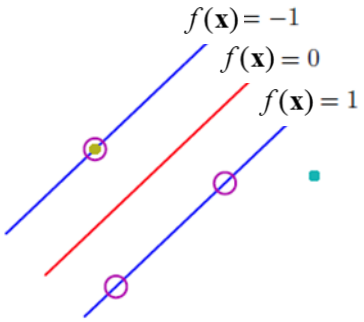
在上式中，两个超平面 $\mathbf{w}^\top \mathbf{x} + b = M$ 和 $\mathbf{w}^\top \mathbf{x} + b = -M$ 之间的距离$2M$被称为**间隔**（margin）。

若继续简化优化问题，令 $M = 1$，注意此时 $\|\mathbf{w}\| \neq 1$，间隔不再是$2M$，而是 $2/\|\mathbf{w}\|$，那么优化表达可以重新表示为



$$\begin{aligned} \max_{w,b} \quad & \frac{2}{\|\mathbf{w}\|} \\ s.t. \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1. \end{aligned} \quad \Rightarrow \quad \begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ s.t. \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \ (i = 1, 2, \ldots, N). \end{aligned}$$

- 大间隔思想（另一种理解）



任意一点x到超平面的距离为

$$d=\frac{|\mathbf{w}^{\top}\mathbf{x}+b|}{\|\mathbf{w}\|}=\frac{y(\mathbf{w}^{\top}\mathbf{x}+b)}{\|\mathbf{w}\|}.$$

最大化所有样本点到分界面最近的距离

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_{n}\left[y_n(\mathbf{w}^{\top}\mathbf{x}_n+b)\right]\right\}$$

考虑到偏差参数b的灵活性，可以认为距离d的大小与$\mathbf{w}$的长度无关，只与$\mathbf{w}$的方向有关系。令最近距离$=1/\|\mathbf{w}\|$，即$\min_{n}\left[y_n(\mathbf{w}^{\top}\mathbf{x}_n+b)\right]=1$ 那么其他数据需满足$y_n(\mathbf{w}^{\top}\mathbf{x}_n+b)\geq 1$。

那么最大间隔可以表示为 $\arg\max_{\mathbf{w},b}\frac{1}{\|\mathbf{w}\|}$

$$\text{s.t.},\ y_n(\mathbf{w}^{\top}\mathbf{x}_n+b)\geq 1$$

目录

- 拉格朗日对偶函数

假设带有约束的优化问题表示为

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \le 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p, \end{aligned}$$

定义原问题的最优解为x^*，目标的最优值为p^*，$p^* = f_0(x^*)$。

引入拉格朗日函数：

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

得到拉格朗日对偶函数:

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right).$$

$\text{argmax}_{\lambda,\nu}\ g(\lambda, \nu)$的最优解为$\lambda^*, \nu^*$，目标的最优值为$q^*$，$q^* = g(\lambda^*, \nu^*)$。

- 弱对偶

假设$\tilde{x}$ 是满足原问题的解，那么拉格朗日函数满足

$$L(\tilde{x},\lambda,\nu)=f_0(\tilde{x})+\sum_{i=1}^{m}\lambda_i f_i(\tilde{x})+\sum_{i=1}^{p}\nu_i h_i(\tilde{x})\le f_0(\tilde{x}).$$

且拉格朗日对偶函数满足：

$$g(\lambda,\nu)=\inf_{x\in\mathcal{D}}L(x,\lambda,\nu)\le L(\tilde{x},\lambda,\nu)\le f_0(\tilde{x}).$$

至此我们可以得到，对偶函数一定小于或等于原优化问题的最优值。

因此，在乘子变量非负的约束下，最大化对偶函数的值，可能会达到原优化问题的最优值。

定义对偶函数的解为λ^*,ν^* ，最优值为q^* ，可以得到$q^*\le p^*$

- 强对偶

什么条件下，可能会达到最优值=>一定会达到?

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star, \nu^\star) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \right) \\ &\le f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star) \\ &\le f_0(x^\star). \end{aligned}$$

第一行等式表示原问题的解与对偶解相同，
第二行等式是将最优对偶解代入对偶函数的定义式，
第三行不等式利用了下界的含义，
第四行利用了不等式约束条件。

① 所有的不等号都应该是等号（互补松弛）：

$$\sum_{i=1}^m \lambda_i^\star f_i(x^\star) = 0.$$

$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m.$$

② $f_0(x^*)$ 是 $f_0(x)+\sum_{i=1}^m \lambda_i^* f_i(x)+\sum_{i=1}^p \nu_i^* h_i(x)$ 的最小值

- KKT条件

什么条件下，可能会达到最优值=>一定会达到?

$$
\begin{aligned}
f_i(x^\star) &\leq 0, \quad i = 1, \ldots, m \\
h_i(x^\star) &= 0, \quad i = 1, \ldots, p \\
\lambda_i^\star &\geq 0, \quad i = 1, \ldots, m \\
\lambda_i^\star f_i(x^\star) &= 0, \quad i = 1, \ldots, m \\
\nabla f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^p \nu_i^\star \nabla h_i(x^\star) &= 0,
\end{aligned}
$$

① 不等式约束
② 等式约束
③ 拉格朗日乘子约束
④ 互补松弛
⑤ 最小值约束

- 最大最小理解

原问题：

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array}$$

等价描述（考虑只有不等式约束的简单情况）：

$$\min \sup_{\lambda \succeq 0} L(x, \lambda) = \sup_{\lambda \succeq 0} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right) = \begin{cases} f_0(x) & f_i(x) \le 0, \quad i = 1, \ldots, m \\ \infty & \text{otherwise.} \end{cases}$$

原问题是：

$$p^\star = \inf_x \sup_{\lambda \succeq 0} L(x, \lambda).$$

弱对偶性：

$$\sup_{\lambda \succeq 0} \inf_x L(x, \lambda) \le \inf_x \sup_{\lambda \succeq 0} L(x, \lambda),$$

（最小值里的最大值<=最大值里的最小值）

强对偶性：

$$\sup_{\lambda \succeq 0} \inf_x L(x, \lambda) = \inf_x \sup_{\lambda \succeq 0} L(x, \lambda).$$

（强对偶性满足的条件是KKT条件）

- SVM的拉格朗日对偶优化解

原问题：
$$\min_{w,b} \quad \frac{1}{2}\|\mathbf{w}\|^2$$
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \, (i = 1, 2, \ldots, N).$$

引入拉格朗日乘子向量$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_N]$，通过拉格朗日函数将约束条件融入到目标函数中，得到优化问题对应的拉格朗日函数为

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1),$$

其中各乘子变量$\alpha_i (i = 1, 2, \cdots, N)$均为非负值。

拉格朗日对偶函数 $g(\boldsymbol{\alpha}) = \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$

原问题可以转化为对偶问题： $\max_{\alpha_i \geq 0} g(\boldsymbol{\alpha}) = \max_{\alpha_i \geq 0} \min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}).$

需满足KKT条件：不等式约束，乘子约束，互补松弛，最小值约束

求解拉格朗日对偶优化问题：

$$\max_{\alpha_i \geq 0} \min_{\mathbf{w},b} L(\mathbf{w},b,\boldsymbol{\alpha}).$$

- 首先，固定$\boldsymbol{\alpha}$，关于$\mathbf{w}$和b最小化拉格朗日函数$L(\mathbf{w},b,\boldsymbol{\alpha})$。对$\mathbf{w}$和$b$求导，我们可以得出

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$

将上面两式带入函数$L(\mathbf{w},b,\boldsymbol{\alpha})$中，得到对偶函数为（第二个等式未使用）

$$\begin{aligned} g(\boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i + \sum_{i=1}^{N} \alpha_i \\ &= \sum_{i=1}^{N} \alpha_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j . \end{aligned}$$

- 其次，求解对偶优化函数，计算对偶变量**α**的最优解。根据对偶公式得到对偶优化问题的具体表示为

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\top}\mathbf{x}_j \\ s.t. \quad & \alpha_i \geq 0 \\ & \sum_{i=1}^{N}\alpha_i y_i = 0. \end{aligned}$$

- 最后，利用**α**的最优解得到参数**w**

$$\mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i$$

参数b可以通过如下互补松弛条件求得

$$\alpha_i (y_i(\mathbf{w}^{\top}\mathbf{x}_i + b) - 1) = 0, \quad i = 1, 2, ..., N.$$

支持向量：根据**w**的公式看出只有$\alpha_j > 0$的值对结果有用，对应的$\mathbf{x}_j, y_j$能够决定决策面的方向，这些样本点称为支持向量。因此，根据任一个支持向量$\mathbf{x}_j$，可得出

$$b = y_j - \mathbf{w}^{\top}\mathbf{x}_j.$$

至此，对于待分类样本$\mathbf{x}$，支持向量机分类器表示为

$$h(\mathbf{w}) = sign[\mathbf{w}^{\top}\mathbf{x} + b],$$

或

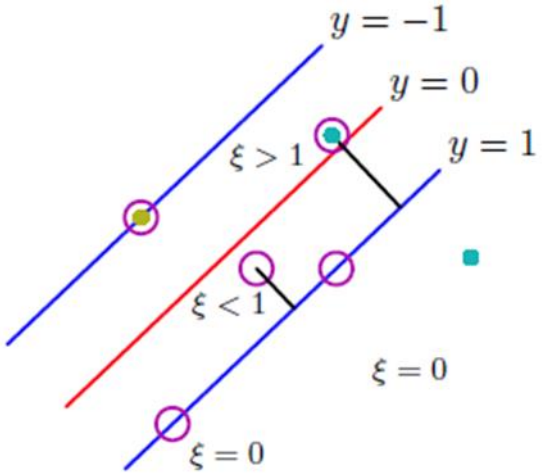
$$h(\mathbf{x}) = sign[\sum_{i=1}^{N} y_i a_i \mathbf{x}^{\top}\mathbf{x}_i + b],$$

其中运用了变换 $\mathbf{w}^{\top}\mathbf{x} = \sum_{i=1}^{N} y_i \alpha_i \mathbf{x}^{\top}\mathbf{x}_i$.

目录

- 松弛变量

基本分类模型的约束条件为 $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1.$
引入松弛变量后约束条件为 $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,$
其中松弛变量$\xi_i (i = 1,2,\cdots,N)$体现了样本点$\mathbf{x}_i$允许偏离原间隔的量，而且满足条件$\xi_i \geq 0 (i = 1,2,\cdots,N)$。



在基本分类模型基础上最小化铰链损失，得到用于线性不可分问题的线性支持向量机分类器的优化问题如下：

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0\,(i = 1,2,\ldots,N),$$

其中$\boldsymbol{\xi}$是由ξ_i构成的向量，折衷参数C用于控制目标函数中大间隔和经验损失两项之间的权重。

引入非负的乘子变量α_i和β_i，可得拉格朗日函数表达式如下

$$L(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\beta)=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^{\top}\mathbf{x}_i+b)-1+\xi_i)-\sum_{i=1}^{N}\beta_i\xi_i.$$

对上式关于w,b,ξ_i求导，可以得到

$$\frac{\partial L}{\partial \mathbf{w}}=\mathbf{0}\Rightarrow\mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i,$$

$$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{N}\alpha_i y_i=0,$$

$$\frac{\partial L}{\partial \xi_i}=0\Rightarrow C-\alpha_i-\beta_i=0.$$

将这些表达式代入拉格朗日函数，可以得到对偶优化问题是

$$\max_{\boldsymbol{\alpha}}\quad\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^{\top}\mathbf{x}_j$$

$$s.t.\quad 0\le\alpha_i\le C$$

$$\sum_{i=1}^{N}\alpha_i y_i=0\,.$$

参数b可以通过如下互补松弛条件求得

$$\alpha_i(y_i(\mathbf{w}^\top\mathbf{x}_i+b)-1+\xi_i)=0,\quad i=1,2,...,N$$
$$\beta_i\xi_i=0,\quad i=1,2,...,N,$$

其中，通过选择满足$0<\alpha_j<C$的支持向量$\mathbf{x}_j$，此使$\beta_j\neq 0$，则必有$\xi_j=0$，可得 $b=y_j-\mathbf{w}^\top\mathbf{x}_j$

最终的决策函数与基本分类模型的决策函数具有相同的表达式

$$h(\mathbf{w})=sign[\mathbf{w}^\top\mathbf{x}+b],$$

或

$$h(\mathbf{x})=sign[\sum_{i=1}^{N}y_i a_i\mathbf{x}^\top\mathbf{x}_i+b].$$

- 核方法

设x和z来自空间Γ（不一定是线性空间），满足下式的函数κ被称为核函数

$$\kappa(x,z)=\langle\phi(x),\phi(z)\rangle,$$

其中ϕ是从空间Γ到希尔伯特空间F的映射

$$\phi: x\in\Gamma\mapsto\phi(x)\in F,$$

空间F通常被称为特征空间。

对于支持向量机，设映射函数为$\phi(\mathbf{x})$，那么 $\kappa(\mathbf{x}_i,\mathbf{x}_j)=\phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_j)$
对偶问题的优化目标变为

$$\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\kappa(\mathbf{x}_i,\mathbf{x}_j).$$

无法仅使用参数 $\mathbf{w}$ 得到决策函数：$h(\mathbf{w})=sign[\mathbf{w}^\top\phi(\mathbf{x})+b],$
原来的 $\mathbf{w}^\top\mathbf{x}$ 变为 $\mathbf{w}^\top\phi(\mathbf{x})$：　$\mathbf{w}^\top\phi(\mathbf{x})=\sum_{i=1}^{N}\alpha_i y_i\kappa(\mathbf{x}_i,\mathbf{x}).$

常见的基本核函数：

- 线性核

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\top}\mathbf{x}_j.$$

- 多项式核

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^{\top}\mathbf{x}_j + 1)^d,$$

其中参数d是多项式次数。

- 高斯核

$$\mathbf{x}_i^{\top}\mathbf{x}_j = \exp\left\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right\},$$

也称为径向基函数核，其中参数σ和多项式核函数中的参数d一样，通过需要通过模型选择来确定具体取值。

- SVM的SMO优化算法

回顾对偶优化目标，如何优化$\boldsymbol{\alpha}$呢?

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\top}\mathbf{x}_j \\ s.t. \quad & \alpha_i \geq 0 \\ & \sum_{i=1}^{N}\alpha_i y_i = 0. \end{aligned}$$

- 选取一对需要更新的变量α_i和α_j[1]
- 固定α_i和α_j以外的参数，求解新的二次规划问题。

$$\min_{\alpha_i,\alpha_j} \quad \frac{1}{2}\begin{bmatrix}\alpha_i & \alpha_j\end{bmatrix}\begin{bmatrix}Q_{ii} & Q_{ij}\\ Q_{ij} & Q_{jj}\end{bmatrix}\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + (-\mathbf{e}_B + Q_{BN}\boldsymbol{\alpha}_N^k)^T\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + \text{constant}$$

$$\begin{aligned} \text{subject to} \quad & 0 \leq \alpha_i, \alpha_j \leq C, \\ & y_i\alpha_i + y_j\alpha_j = -\mathbf{y}_N^T\boldsymbol{\alpha}_N^k, \end{aligned}$$

[1] Fan R E, Chen P H, Lin C J. Working set selection using second order information for training support vector machines[J]. The Journal of Machine Learning Research, 2005, 6: 1889-1918.

- SVM的SMO优化算法
  - 求解新的二次规划问题。

$$\min_{\alpha_i,\alpha_j} \quad \frac{1}{2}\begin{bmatrix}\alpha_i & \alpha_j\end{bmatrix}\begin{bmatrix}Q_{ii} & Q_{ij}\\ Q_{ij} & Q_{jj}\end{bmatrix}\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + (-\mathbf{e}_B + Q_{BN}\boldsymbol{\alpha}_N^k)^T\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + \text{constant}$$

$$\text{subject to} \quad 0 \le \alpha_i, \alpha_j \le C,$$
$$y_i\alpha_i + y_j\alpha_j = -\mathbf{y}_N^T\boldsymbol{\alpha}_N^k,$$

  - 更新公式

$$a_{ij} \equiv K_{ii} + K_{jj} - 2K_{ij} > 0$$

$$b_{ij} \equiv -y_i\nabla f(\boldsymbol{\alpha}^k)_i + y_j\nabla f(\boldsymbol{\alpha}^k)_j > 0.$$

$$\alpha_i = \frac{b_{ij}}{a_{ij}},$$

$$\alpha_j = -\frac{b_{ij}}{a_{ij}}$$

建议查看LibSVM的代码，查看类Solver

目录

- 支持向量机回归

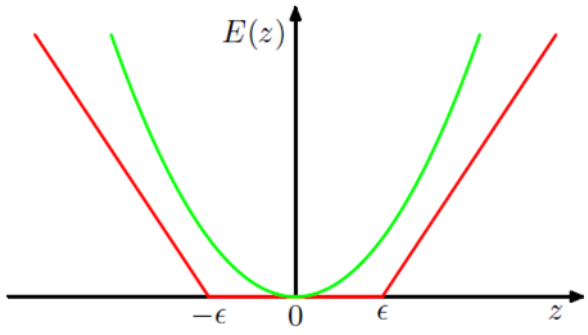
为了得到决策函数的稀疏表达，引入ϵ不敏感损失函数

$$|y-f(\mathbf{x})|_{\varepsilon}=\max\{0,|y-f(\mathbf{x})|-\varepsilon\},$$

其中$f(\mathbf{x})$是回归函数，参数$\epsilon \geq 0$。
由于ϵ不能设置得过大，所以不敏感损失的值往往不为零。
因此，支持向量回归的优化问题表示为

$$\min_{w,b} \quad \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}|y_i-f(\mathbf{x}_i)|_{\varepsilon},$$

其中非负参数C反应了函数复杂性与经验损失之间的折衷，对样本的预测输出函数为 $f(\mathbf{x})=\mathbf{w}^{\top}\mathbf{x}+b$ 。



理解：允许一定的误差，但误差不能过大

引入两组松弛变量ξ和ξ*，使得每个样本在间隔带两侧的松弛程度可以不同，则支持向量回归的优化函数等价为

$$\begin{aligned} \min_{w,b,\xi,\xi^*} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ s.t. \quad & y_i - \mathbf{w}^\top \mathbf{x}_i - b \le \epsilon + \xi_i \\ & \mathbf{w}^\top \mathbf{x}_i + b - y_i \le \epsilon + \xi_i^* \\ & \xi_i \ge 0 \\ & \xi_i^* \ge 0, \quad i = 1, 2, ..., N. \end{aligned}$$

对应的拉格朗日函数为

$$\begin{aligned} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \eta, \eta^*) = & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}(\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^{N} \alpha_i (\epsilon + \xi_i - y_i + \mathbf{w}^\top \mathbf{x}_i + b) \\ & - \sum_{i=1}^{N} \alpha_i^* (\epsilon + \xi_i + y_i - \mathbf{w}^\top \mathbf{x}_i - b), \end{aligned}$$

其中，$\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$是非负的乘子变量。

对拉格朗日函数关于$\mathrm{w}, b, \xi_i, \xi_i^*$进行求导，可以得到

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \mathbf{w} = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\mathbf{x}_i,$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0,$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \eta_i = 0,$$

$$\frac{\partial L}{\partial \xi_i^*} = 0 \Rightarrow C - \alpha_i^* - \eta_i^* = 0.$$

将上述结果带入拉格朗日函数中，得到对偶优化问题

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\mathbf{x}_i^\top \mathbf{x}_j + \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \\ s.t. \quad & \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ & \alpha_i, \alpha_i^* \in [0, C]. \end{aligned}$$

求得最优的α_i, α_i^*后，支持向量机回归的预测函数为

$$f(\mathbf{x}) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \mathbf{x}^\top \mathbf{x}_i + b.$$

其中，偏置项b通过如下互补松弛条件求得：

$$\alpha_i (\epsilon + \xi_i - y_i + \mathbf{w}^\top \mathbf{x}_i + b) = 0$$

$$\alpha_i^* (\epsilon + \xi_i + y_i - \mathbf{w}^\top \mathbf{x}_i - b) = 0$$

$$\eta_i \xi_i = 0$$

$$\eta_i^* \xi_i^* = 0$$

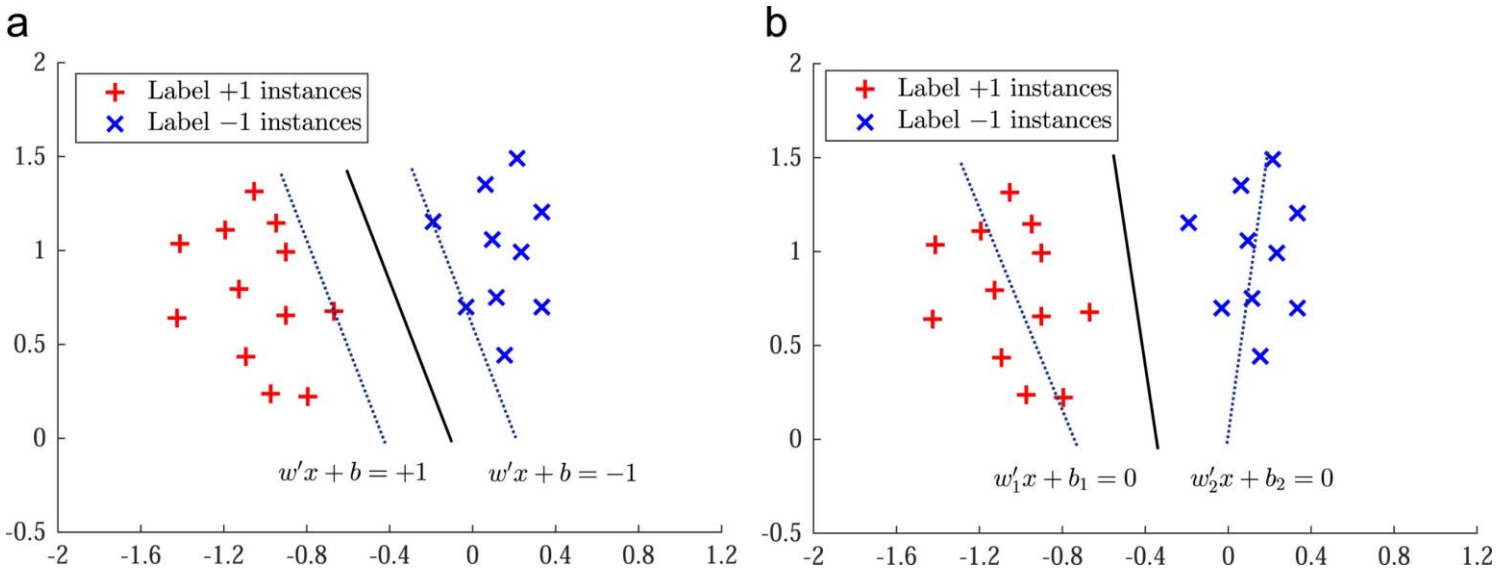
若$0 < \alpha_i < C$，那么$\eta_i \neq 0$，则必有$\xi_i = 0$，进而运用与α_i对应的样本输入和标签得出b的表达式为

$$b = y_i + \epsilon - \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \mathbf{x}^\top \mathbf{x}_i.$$

目录

- 模型扩展

双平面支持向量机的思想是使得一个超平面离一类样本近并且离另一类样本有一定的距离。与支持向量机求解单个二次优化问题不同，它求解两个二次优化问题，而且每个优化问题涉及的样本数量少于支持向量机中的样本数量。



1. Blumer A, Ehrenfeucht A, Haussler D, et al. Learnability and the Vapnik-Chervonenkis Dimension[J]. Journal of the ACM, 1989, 36(4): 929-965.
2. Boser B E, Guyon I M, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers[C]//Proceedings of the 5th Annual Workshop on Computational Learning Theory. New York: ACM, 1992: 144-152.
3. Shawe-Taylor J, Sun S. A Review of Optimization Methodologies in Support Vector Machines[J]. Neurocomputing, 2011, 74(17): 3609-3618.
4. Boyd S, Vandenberghe L. Convex Optimization[M]. Cambridge, UK: Cambridge University Press, 2004.
5. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis[M]. Cambridge, UK: Cambridge University Press, 2004.
6. Shawe-Taylor J, Sun S. Kernel Methods and Support Vector Machines[M]//Academic Press Library in Signal Processing: Chapter 16. Amsterdam: Elsevier, 2014: 857-881.
7. Vapnik V N. The Nature of Statistical Learning Theory[M]. Berlin: Springer, 1995.
8. Khemchandani R, Chandra S. Twin Support Vector Machines for Pattern Classification[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905-910.
9. Xie X, Sun S. PAC-Bayes Bounds for Twin Support Vector Machines[J]. Neurocomputing, 2017, 234(4): 137-143.
10. Sun S, Mao L, Dong Z, et al. Multiview Machine Learning[M]. Singapore: Springer, 2019.
11. https://www.csie.ntu.edu.tw/~cjlin/libsvm/