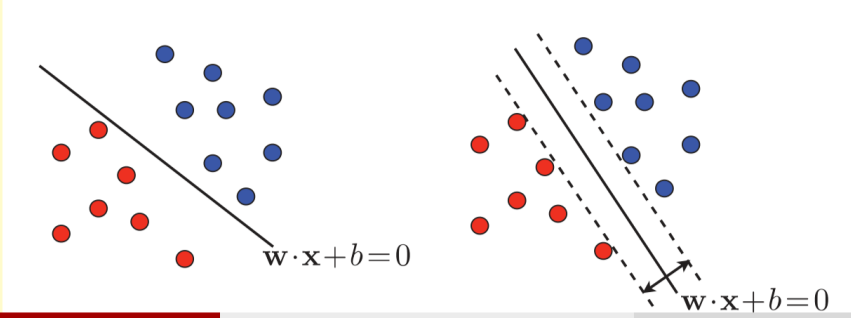
H1 第二章 支持向量机与核算法

H2 2.1 线性分类

示例



基本符号

- $\mathcal{X} \in \mathbb{R}^d$： 输入空间
- $\mathcal{Y} = \{-1, +1\}$： 输出空间
- $f : \mathcal{X} \mapsto \mathcal{Y}$： 未知的目标函数
- $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ 其中 $\y_i = f(\boldsymbol{x}_i)$， $\{\boldsymbol{x}_i\}_{i=1}^n$ 从输入空间 X 中依据某个分布 D 采集得到： 训练集合
- $\mathcal{H} = \{\boldsymbol{x} \mapsto sig(\boldsymbol{w}^T \boldsymbol{x} + b) : \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$： 假说集合
- $R_S(h) = \frac{1}{n} \sum_{i=1}^{n} [h(\boldsymbol{x}_i) \neq f(\boldsymbol{x}_i)], h \in \mathcal{H}$： 经验误差
- $R_D(h) = Pr_{\boldsymbol{X} \sim D}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})], h \in \mathcal{H}$： 泛化误差

分类问题

二分类问题形式化定义:

给定训练样本集合$S = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$，其中$\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathcal{X}^n$独立同分布，$y_i = f(\boldsymbol{x}_i) \in \mathcal{Y} (\forall i = 1, \cdots, n)$。二分类问题的目标是基于数据$S$，从假说集合$\mathcal{H}$中选择一个假说$h$，以使得期望误差

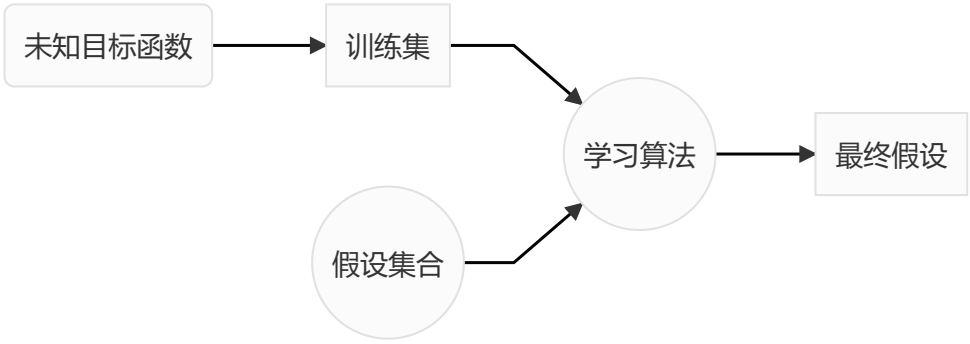
$$E_{\boldsymbol{x} \sim D}[h(\boldsymbol{x} \neq f(\boldsymbol{x}))] \tag{1}$$

最小

其中

$$\begin{aligned} E_{\boldsymbol{x} \sim D}[h(\boldsymbol{x} \neq f(\boldsymbol{x}))] &= 1 \cdot Pr_{\boldsymbol{X} \sim D}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) + 0 \cdot Pr_{\boldsymbol{X} \sim D}(h(\boldsymbol{x}) = f(\boldsymbol{x})) \quad (2) \\ &= Pr_{\boldsymbol{X} \sim D}(h(\boldsymbol{x}) \sim f(\boldsymbol{x})) \quad (3) \end{aligned}$$

学习一个分类器



假说集（线性）

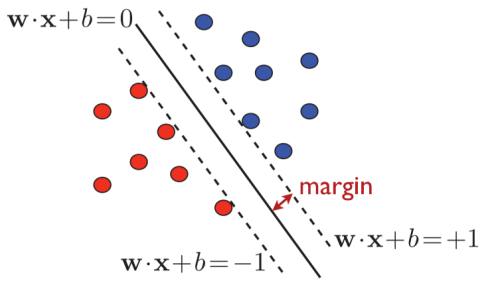
$$\mathcal{H} = \{\boldsymbol{x} \mapsto sign(\boldsymbol{w}^T \boldsymbol{x} + b) : \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\} \tag{4}$$

学习算法

$\mathcal{A}$：支持向量机（Support Vector Machines, SVMs）

2.2 支持向量机

线性分类算法



(a) 支持向量机

- 考虑超平面$\boldsymbol{w}^T\boldsymbol{x}+b=0$
- 给定$a>0$，要求该超平面满足
  - 对于正样本（即$y_i=1$）

$$\boldsymbol{w}^T\boldsymbol{x}_i+b\geq a \tag{5}$$

  - 对于负样本（即$y_i=-1$）

$$\boldsymbol{w}^T\boldsymbol{x}_i+b\leq a \tag{6}$$

  - 即$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b\geq a)\forall i$

定义（几何间隔）

样本$\boldsymbol{x}_i$到超平面$\boldsymbol{w}^T\boldsymbol{x}+b=0$的几何距离：$\frac{|\boldsymbol{w}^T\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|}=\frac{y_i(\boldsymbol{w}_i\boldsymbol{x}_i+b)}{\|\boldsymbol{w}\|}$

样本集S到超平面$\boldsymbol{w}^T\boldsymbol{x}_i+b=0$的几何距离$\rho$被定义为：

$$\rho=\min_{(\boldsymbol{x}_i,y_i)\in S}\frac{y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)}{\|\boldsymbol{w}\|}=\frac{a}{\|\boldsymbol{w}\|} \tag{7}$$

算法推导

间隔最大化

> 支持向量机学习的基本想法是求解能够正确划分训练集并且几何间隔最大的分离超平面，对线性可分的训练集而言，线性可分分离超平面有无穷个（等价于感知机），但是几何间隔最大的分离超平面是唯一的。这里的间隔最大化又称为硬间隔最大化。
>
> 间隔最大化的直观解释是：对训练数据集找到几个间隔最大的超平面意味着以充分大的确信度对训练集进行分类。也就是说，不仅将正负实例分开，而且对最难的实例（离超平面最近的点）也能够又足够大的确信度将它们分开。[1]

最大间隔分离超平面

考虑如何求一个几何间隔最大的分离超平面，即最大间隔分离超平面，其可以表示为以下的约束最优化问题：

$$\max_{\boldsymbol{w},b}\rho=\frac{a}{\|\boldsymbol{w}\|} \tag{8}$$

$$s.t.\quad y_i(\boldsymbol{w}_i^T\boldsymbol{x}_i+b)\geq a,\forall i \tag{9}$$

即希望最大化超平面$(\boldsymbol{w},b)$关于训练数据集的几何间隔ρ，约束条件表示的是超平面$(\boldsymbol{w}_i,b)$关于每个训练样本点的几何间隔至少是a。

函数间隔a的取值并不影响最优化问题的解。事实上，假设将$\boldsymbol{w}$和b按比例改变为$\lambda\boldsymbol{w}$和λb，函数间隔的这一改变对上面最优化问题的不等式约束以及对目标函数的优化也没有影响。于是取$a=1$，令$\hat{\boldsymbol{w}}=\frac{\boldsymbol{w}}{a},\hat{b}=\frac{b}{a}$，则优化目标等价于

$$\max_{\hat{\boldsymbol{w}},\hat{b}} \frac{1}{\|\hat{\boldsymbol{w}}\|} \tag{10}$$

$$s.t. \quad y_i(\hat{\boldsymbol{w}}^T\boldsymbol{x}_i+\hat{b})\geq 1, \forall i \tag{11}$$

注意到最大化$\frac{1}{\|\hat{\boldsymbol{w}}\|}$与最小化$\frac{1}{2}\|\hat{\boldsymbol{w}}\|^2$等价，故优化问题等价于

$$\min_{\hat{\boldsymbol{w}},\hat{b}} \frac{1}{2}\|\hat{\boldsymbol{w}}\|^2 \tag{12}$$

$$s.t. \quad y_i(\hat{\boldsymbol{w}}^T\boldsymbol{x}_i+\hat{b})\geq 1, \forall i \tag{13}$$

这是一个凸二次规划问题。

> 凸优化问题指的是约束最优化问题
>
> $$\min_{\boldsymbol{w}} f(\boldsymbol{w}) \tag{14}$$
>
> $$s.t. \quad g_i(\boldsymbol{w})\leq 0, \quad i=1,2,\cdots,k \tag{15}$$
>
> $$h_j(\boldsymbol{w})=0, i=1,2,\cdots,l \tag{16}$$
>
> 其中，目标函数$f(\boldsymbol{w})$和约束函数$g_i(\boldsymbol{w})$都是$\boldsymbol{R}^n$上的连续可微凸函数，约束函数$h_i(\boldsymbol{w})$是$\boldsymbol{R}^n$上的仿射函数。[1][2]

H5 最大间隔分离超平面的存在唯一性

> 定理[1]：若训练集T线性可分，则可将训练集中的样本点完全正确分开的最大间隔分离超平面存在且唯一
>
> 证明：
>
> - 存在性
>
> 由于训练数据集线性可分，则最优化问题一定存在可行解[3]。又由于目标函数有下界0，故最优化问题必有解，记作$(\boldsymbol{w}^*,b^*)$。由于训练数据集中既有正类又有负类，则$(\boldsymbol{w},b)=(0,b)$不是最优化的可行解（不满足约束条件），因而最优解$(\boldsymbol{w}^*,b^*)$必满足$\boldsymbol{w}^*\neq 0$。因此得知分离超平面的存在性。
>
> - 唯一性
>
> 首先证明最优化问题解中$\boldsymbol{w}^*$的唯一性。假设存在两个最优解$(\boldsymbol{w}_1^*,b_1^*)$与$(\boldsymbol{w}_2^*,b_2^*)$，显然$\|\boldsymbol{w}_1^*\|=\|\boldsymbol{w}_2^*\|=c$，其中$c$是一个常数。令$\boldsymbol{w}=\frac{\boldsymbol{w}_1^*+\boldsymbol{w}_2^*}{2},b=\frac{b_1^*+b_2^*}{2}$，则$(\boldsymbol{w},b)$为算法的一个可行解，从而
>
> $$c\leq\|\boldsymbol{w}\|\leq\frac{1}{2}\|\boldsymbol{w}_1^*\|+\frac{1}{2}\|\boldsymbol{w}_2^*\|=c \tag{17}$$
>
> 这表明$\|\boldsymbol{w}\|=\frac{1}{2}\|\boldsymbol{w}_1^*\|+\frac{1}{2}\|\boldsymbol{w}_2^*\|$，从而$\boldsymbol{w}_1^*=\lambda\boldsymbol{w}_2^*$，其中$|\lambda|=1$，若$\lambda=-1$，则$\boldsymbol{w}$不是最优化问题的可行解，矛盾，因此$\lambda=1$，即

$$\boldsymbol{w}_1^* = \boldsymbol{w}_2^* \tag{18}$$

由此可以把两个最优解分别写成$(\boldsymbol{w}, b_1^*)$与$(\boldsymbol{w}, b_2^*)$，下面证明$b_1^* = b_2^*$

设x_1', x_2'分别是集合$\{x_i | y_i = +1\}$中分别对应于$(\boldsymbol{w}, b_1^*)$与$(\boldsymbol{w}, b_2^*)$使得问题不等式中等号成立的点，x_1'', x_2''分别是是集合$\{x_i | y_i = -1\}$中分别对应于$(\boldsymbol{w}, b_1^*)$与$(\boldsymbol{w}, b_2^*)$使得问题不等式中等号成立的点，则由
$b_1* = \frac{1}{2}(\boldsymbol{w}^* \cdot x_1' + \boldsymbol{w}^* \cdot x_1'')$, $b_2* = \frac{1}{2}(\boldsymbol{w}^* \cdot x_2' + \boldsymbol{w}^* \cdot x_2'')$可得

$$b_1^* - b_2^* = \frac{1}{2}[\boldsymbol{w}^*(x_1'' - x_2'') + \boldsymbol{w}^*(x_1'' - x_2'')] \tag{19}$$

$$\because \tag{20}$$

$$\boldsymbol{w}^* \cdot x_2' + b_1^* \geq q = \boldsymbol{w}^* \cdot x_1' + b_1^* \tag{21}$$

$$\boldsymbol{w}^* \cdot x_1' + b_2^* \geq q = \boldsymbol{w}^* \cdot x_2' + b_2^* \tag{22}$$

故$\boldsymbol{w}^*(x_1' - x_2') = 0$，同理$\boldsymbol{w}^*(x_1'' - x_2'') = 0$，故$b_1^* = b2^*$

则两个最优解相同，故唯一性得证

- 验证

  由训练集线性可分得知：分离超平面可以将训练数据集中两类点完全正确分开

H4 求解算法

H5 对偶问题

由附录可知，可以将原始问题转化为对偶问题

首先列出Lagrange函数

$$L(\hat{\boldsymbol{w}}, \hat{b}; \boldsymbol{\alpha}) = \frac{1}{2}\|\hat{\boldsymbol{w}}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\hat{\boldsymbol{w}}^T \boldsymbol{x}_i + \hat{b}) - 1) \tag{23}$$

其中$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_n]^T, \alpha_i \geq 0 \; \forall i$为Lagrange乘子

对应的KKT条件为

$$\begin{cases} \nabla_{\hat{\boldsymbol{w}}} L = \hat{\boldsymbol{w}} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0 & \Rightarrow \hat{\boldsymbol{w}} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \\ \nabla_{\hat{b}} L = -\sum_{i=1}^{n} \alpha_i y_i = 0 & \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \forall i \; \alpha_i (y_i(\hat{\boldsymbol{w}}^T \boldsymbol{x}_i + \hat{b}) - 1) = 0 & \Rightarrow \alpha_i \vee y_i(\hat{\boldsymbol{w}}^T \boldsymbol{x}_i + \hat{b}) = 1 \end{cases} \tag{24}$$

将KKT条件代入可得

$$\min_{\hat{\boldsymbol{w}}, \hat{b}} L(\hat{\boldsymbol{w}}, \hat{b}; \boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^T \boldsymbol{x}_j) + \sum_{i=1}^{n} \alpha_i \tag{25}$$

则对偶问题为

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i^T\boldsymbol{x}_j) - \sum_{i=1}^{n}\alpha_i \tag{26}$$

$$s.t.\ \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{27}$$

$$\alpha_i \geq 0, \forall i \tag{28}$$

这是一个标准的带约束二次规划问题

H5 解的形式

定理：设$\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_l^*)^T$是对偶最优化问题26, 27, 28的解，则存在下标j，使得$\alpha_j^* > 0$，并可按下式求得原始最优化问题的解$\boldsymbol{w}^*, b^*$

$$\boldsymbol{w}^* = \sum_{i=1}^{n}\alpha_i^* y_i \boldsymbol{x}_i \tag{29}$$

$$b^* = y_j - \sum_{i=1}^{n}\alpha_i^* y_i(\boldsymbol{x}_i^T \cdot \boldsymbol{x}_j) \tag{30}$$

证明

由???可知

$$\boldsymbol{w}^* = \sum_{i}\alpha_i^* y_i \boldsymbol{x}_i \tag{31}$$

其中至少由一个$\alpha_j > 0$（用反证法，由???可知$\boldsymbol{w}^* = \boldsymbol{0}$，矛盾），由此$j$有

$$y_j((\boldsymbol{w}^*)^T \cdot \boldsymbol{x}_j + b^*) - 1 = 0 \tag{32}$$

将29代入32并由$y_j^2 = 1$可得

$$b^* = y_j - \sum_{i=1}^{n}\alpha_i^* y_i(\boldsymbol{x}_i^T \cdot \boldsymbol{x}_j) \tag{33}$$

由此定理可知，分离超平面可写成

$$\sum_{i=1}^{n}\alpha_i^* y_i(\boldsymbol{x}^T \cdot \boldsymbol{x}_i) + b^* = 0 \tag{34}$$

分类决策函数可写为

$$f(x) = sign(\sum_{i=1}^{n}(\boldsymbol{w}^*)^T\boldsymbol{x} + b^*) = sign(\sum_{i=1}^{n}\alpha_i^* y_i(\boldsymbol{x}^T \cdot \boldsymbol{x}_i) + b^*) \tag{35}$$

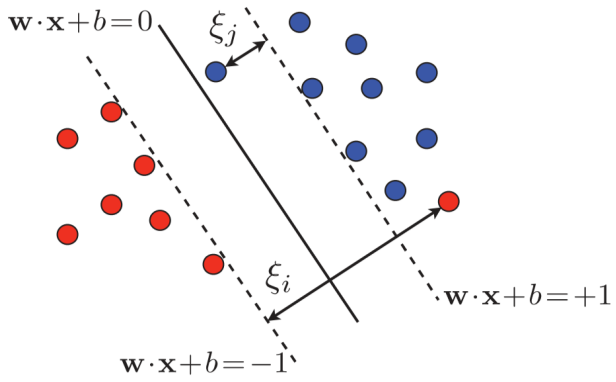
H5 说明

- 支持向量（Support Vectors）：对应于 $\alpha_i \geq 0$ 的样本
- 样本总是成对出现的

H3 线性不可分

在绝大多数应用中，训练数据并非线性可分，即对任意超平面$\boldsymbol{w}^T\boldsymbol{X} + b$，存在$\boldsymbol{x}_i \in S$，使得

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_I + b) \ngeq 1 \tag{36}$$



- 超平面错误分类 $\boldsymbol{x}_i$
- 超平面正确分类 $\boldsymbol{x}_j$，但间隔小于1

H4 非线性可分——优化目标

考虑如下两个相互矛盾的因素

- 间距最大化
- 训练误差最小化

为了将支持向量机算法拓展到可以适应不可分问题，需要修改硬间隔最大化，使其成为软间隔最大化，即对每个样本点引进一个松弛变量$\xi_i \geq 0$，使函数间隔加上松弛变量大于等于1。这压根，约束条件变为

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \tag{37}$$

同时，对每个松弛变量，支付一个代价C

则新的学习问题变为

$$\min_{\hat{\boldsymbol{w}},\hat{b},\boldsymbol{\xi}} \frac{1}{2}\|\hat{\boldsymbol{w}}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{38}$$

$$s.t. \quad y_I(\hat{\boldsymbol{w}}^T\boldsymbol{x}_i + \hat{b}) \geq 1 - \xi + i \tag{39}$$

$$\xi_i \geq 0, \forall i \tag{40}$$

- 允许训练过程中有误差
- C 为正则化参数--调节模型复杂度与训练误差

H4 Lagerange函数

$$L(\hat{\boldsymbol{w}}, \hat{b}, \boldsymbol{\xi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\hat{\boldsymbol{w}}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(\hat{\boldsymbol{w}}^T\boldsymbol{x}_i + \hat{b}) - 1 + \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i \tag{41}$$

其中$\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]^T, \boldsymbol{\beta} = [\beta_1, \cdots, \beta_n]^T$为Lagrange乘子

其KKT条件为

$$\begin{cases} \nabla_{\hat{\boldsymbol{w}}} L = \hat{\boldsymbol{w}} - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i = 0 & \Rightarrow \hat{\boldsymbol{w}} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i \\ \nabla_{\hat{b}} L = -\sum_{i=1}^{n}\alpha_i y_i = 0 & \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \nabla_{\boldsymbol{\xi}_i} L = C - \alpha_i - \beta_i & \Rightarrow \alpha_i + \beta_i = C \\ \forall i \quad \alpha_i(y_i(\hat{\boldsymbol{w}}^T\boldsymbol{x}_i + \hat{b}) - 1 + \xi_i) = 0 & \Rightarrow \alpha_i \vee y_i(\hat{\boldsymbol{w}}^T\boldsymbol{x}_i + \hat{b}) = 1 - \xi_i \\ \forall i \beta_i\xi_i = 0 & \Rightarrow \beta_i = 0 \vee \xi_i = 0 \end{cases} \tag{42}$$

学习的对偶算法

将???代入38可得

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i^T\boldsymbol{x}_j) - \sum_{i=1}^{n}\alpha_i \tag{43}$$

$$s.t. \quad \sum_{i=1}^{n}\alpha_i y_i = 0, C \geq \alpha_i \geq 0, \forall i \tag{44}$$

解的形式

定理：设$\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots, \alpha_n^*)^T$是对偶问题43, 44的一个解，若存在$\boldsymbol{\alpha}^*$的一个分量$\alpha_j^*, 0 < \alpha_j^* < C$，则原始问题的解$\boldsymbol{w}^*, b^*$可按下式求得：

$$\boldsymbol{w}^* = \sum_{i=1}^{n}\alpha_i^* y_i \boldsymbol{x}_i \tag{45}$$

$$b^* = y_i - \sum_{i=1}^{n} y_i\alpha_i^*(\boldsymbol{x}_i^T\boldsymbol{x}_j) \tag{46}$$

证明

由KKT条件可知

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i \tag{47}$$

若存在$\boldsymbol{\alpha}^*$的一个分量$\alpha_j^*, 0 < \alpha_j^* < C$

则

$$\begin{gathered} y_j((\boldsymbol{w}^*)^T\boldsymbol{x}_j + \hat{b}) = 1 - \xi_j \\ \beta_j^* > C - \alpha_j^* \end{gathered} \tag{48}$$

则

$$\xi_j = 0 \tag{49}$$

故

$$y_j((\boldsymbol{w}^*)^T\boldsymbol{x}_j + \hat{b}) = 1 \tag{50}$$

证毕

由此定理可知，分离超平面可写成

$$\sum_{i=1}^{n}\alpha_i^* y_i(\boldsymbol{x}^T\boldsymbol{x}_i) + b^* = 0 \tag{51}$$

分类决策函数可写成

$$f(x) = sign(\sum_{i=1}^{n}\alpha_i^* y_i(\boldsymbol{x}^T\boldsymbol{x}_i) + b^*) \tag{52}$$

参数选择

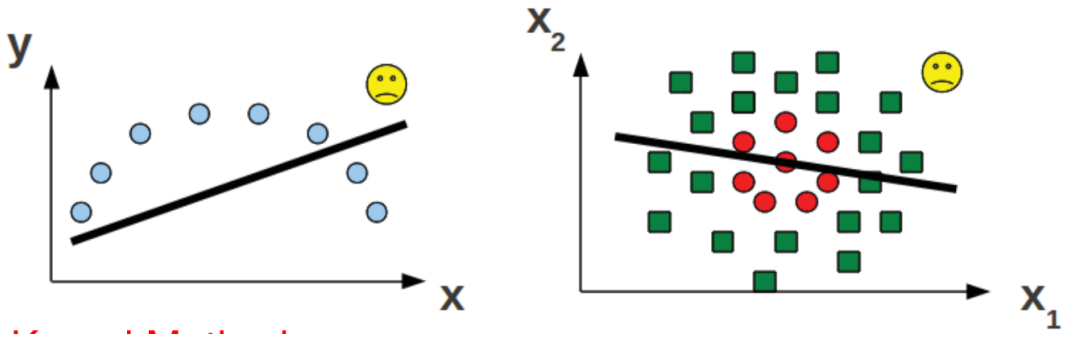
留一法与k折验证

2.3 总结

- 支持向量机推导
- 拉格朗日对偶式求解
- 参数选择

线性支持向量机优缺点

- 优点：
  - 算法简单、只管、可理解性强
- 缺点：表达能力有限，无法捕获数据间的非线性模式
  - 输入--输出不再是线性关系
  - 类与类之间不能通过线性边界来划分



附录 拉格朗日对偶性[1]

在约束最优化问题中，常利用拉格朗日对偶性（Lagrange duality）将原始问题转化为对偶问题，通过解对偶问题而得到原始问题得解。

原始问题

假设$f(x), c_i(x), h_j(x)$是定义在$\boldsymbol{R}^n$上得连续可微函数，考虑约束最优化问题

$$\min_{x \in \boldsymbol{R}^n} f(\boldsymbol{x}) \tag{53}$$

$$s.t. \quad c_i(\boldsymbol{x}) \leq 0, i = 1, 2, \cdots, k \tag{54}$$

$$h_j(\boldsymbol{x}) = 0, j = 1, 2, \cdots, l \tag{55}$$

称此约束最优化问题为原始最优化问题或原始问题

首先，引进广义拉格朗日函数（generalized Lagrange function）

$$L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{x}) + \sum_{i=1}^{k} \alpha_i c_i(\boldsymbol{x}) + \sum_{j=1}^{l} \beta_j h_j(\boldsymbol{x}) \tag{56}$$

这里$\boldsymbol{x} = [x^{(1)}, x^{(2)}, \cdots, x^{(n)}]^T \in \boldsymbol{R}^n$，$\alpha_i, \beta_j$是拉格朗日乘子，$\alpha_i \geq 0$。考虑$\boldsymbol{x}$得函数

$$\theta_P(\boldsymbol{x}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \alpha_i \geq 0} L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{57}$$

其中下标P表示原始问题。

如果$\boldsymbol{x}$满足原始问题的约束条件，则$\theta_P(\boldsymbol{x})=f(\boldsymbol{x})$，而若$\boldsymbol{x}$不满足原始问题约束，则可取对应的$\alpha_i=+\infty$，则$\theta_P(\boldsymbol{x})=+\infty$，或取$\beta_j$满足$\beta_j h_j\to+\infty$，其余$\alpha_i=\beta_j=0$，则$\theta_P(\boldsymbol{x})=+\infty$

所以如果考虑极小化问题

$$\min_{\boldsymbol{x}}\theta_P(\boldsymbol{x})=\min_{\boldsymbol{x}}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{58}$$

它是与原始优化问题等价的。58被称为广义拉格朗日函数的极小极大问题。为了方便，定义原始问题的最优值

$$p^*=\min_{\boldsymbol{x}}\theta_P(\boldsymbol{x}) \tag{59}$$

称为原始问题的值。

对偶问题

定义

$$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{60}$$

再考虑极大化$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$，即

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0}\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0}\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{61}$$

61称为广义朗格朗日函数的极大极小问题，可以将其表示为约束最优化问题：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{62}$$

$$s.t.\quad \alpha_i\geq 0, i=1,2,\cdots,k \tag{63}$$

称为原始问题的对偶问题。定义对偶问题的最优值

$$d^*=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0}\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{64}$$

为对偶问题的值

原始问题和对偶问题的关系

定理C.1：若原始问题和对偶问题都有最优值，则

$$d^*=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\leq\min_{\boldsymbol{x}}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})=p^* \tag{65}$$

证明

由57与62可知，$\forall\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{x}$，均有

$$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\leq L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\leq\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})=\theta_P(\boldsymbol{x}) \tag{66}$$

即

$$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})\leq\theta_P(\boldsymbol{x}) \tag{67}$$

由于原始问题和对偶问题均有最优值，故

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})\leq\min_{\boldsymbol{x}}\theta_P(\boldsymbol{x}) \tag{68}$$

即

$$d^* = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \leq \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) = p^* \tag{69}$$

推论C.1：设$\boldsymbol{x}^*$和$\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$分别是原始问题和对偶问题的可行解，且$d^* = p^*$，则$\boldsymbol{x}^*$和$\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$分别是原始问题和对偶问题的最优解

定理C.2：考虑原始问题和对偶问题，假设函数$f(x)$和$c_i(x)$是凸函数，$h_j(x)$是仿射函数；且假设不等式$c_i(x)$是严格可行的，即$\exists \boldsymbol{x}, s.t. \forall c_i(\boldsymbol{x}) < 0$，则存在$\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$，使$\boldsymbol{x}^*$是原始问题的解，$\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$是对偶问题的解，且

$$p^* = d^* = L(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) \tag{70}$$

定理C.3：对原始问题和对偶问题，假设$f(\boldsymbol{x})$和$c_i(\boldsymbol{x})$是凸函数，$h_j(\boldsymbol{x})$是仿射函数，并且不等式约束$c_i(\boldsymbol{x})$是严格可行的，则则$\boldsymbol{x}^*$和$\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$分别是原始问题和对偶问题的解的充要条件是$\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$满足下面的Karush-Kuhn0Tucher（KKT）条件

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) = 0 \tag{71}$$
$$\alpha_i^* c_i(\boldsymbol{x}^*) = 0, i = 1,2,\cdots,k \tag{72}$$
$$c_i(\boldsymbol{x}^*) \leq 0, i = 1,2,\cdots,k \tag{73}$$
$$\alpha_i^* \geq 0, i = 1,2,\cdots.k \tag{74}$$
$$h_j(\boldsymbol{x}^*) = 0, j = 1,2,\cdots,l \tag{75}$$

特别的，72称为KKT条件的对偶互补条件。由此条件可知：若$\alpha_i^* > 0$，则$c_i(\boldsymbol{x}^*) = 0$

---

1. 李航. 统计学习方法 第2版[M]. 清华大学出版社, 2019. ↩ ↩ ↩ ↩

2. 称$f(x)$为仿射函数，如果它满足$f(x) = a \cdot x + b, c \in \boldsymbol{R}^n, b \in \boldsymbol{R}, x \in \boldsymbol{R}^n$ ↩

3. 可行解：满足所有约束条件的解 ↩