

支持向量机主要可分为三类：线性可分支持向量机、线性支持向量机、非线性支持向量机
1.线性可分支持向量机

线性可分支持向量机也成为硬间隔（Hard margin）支持向量机，即要求所有的样本都要被分类正确。

输入： $T=\{(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)\}$

其中， x_i 表示训练数据集第 i 个实例，是一个向量。

y_i 表示第 x_i 的类标记 $y_i = \{-1, +1\}$

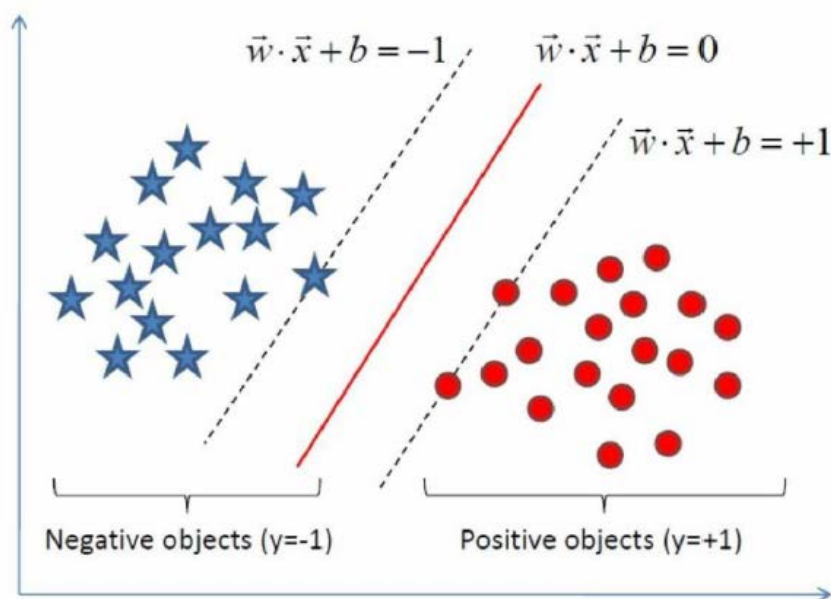
(x_i, y_i) 为样本点

给定线性可分数据集，通过间隔最大化得到分离超平面： $y(x) = \omega^t x + b$ 相应的分类决策函数为 $f(x) = \text{sign}(\omega^t x + b)$

即线性可分 SVM 的任务是不仅要找到把样本点正确分类的超平面，而且是要找到间隔最大的超平面。

此间隔的定义：设 C 和 D 为两不相交的凸集，则存在超平面 P ， P 可以将 C 和 D 分离，两集合间的距离定义为两集合中元素的最短距离， C, D 两集合最短距离的中垂线（多维空间即中垂面）即为超平面，任一集合到超平面的距离即为间隔。

由上可知，线性可分 SVM 即求距离超平面最近的点的间隔最大化。从而增强了 SVM 的泛化能力。



距离求解：假设 x_1, x_2 均为超平面上的点，即 x_1, x_2 两点满足 $w \cdot x_1 + b = 0$ And $w \cdot x_2 + b = 0$ 即， $w \cdot (x_1 - x_2) = 0$ $x_1 - x_2$ 可以看成是超平面上的向量，即 w 为超平面法向量。点 X 到超平面的距离等于， X 在法向量方向上的投影长度。则

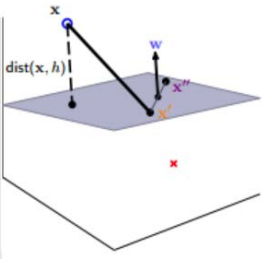
want: distance($\mathbf{x}, \mathbf{b}, \mathbf{w}$), with hyperplane $\mathbf{w}^T \mathbf{x}' + b = 0$

consider $\mathbf{x}', \mathbf{x}''$ on hyperplane

- 1 $\mathbf{w}^T \mathbf{x}' = -b, \mathbf{w}^T \mathbf{x}'' = -b$
- 2 $\mathbf{w} \perp$ hyperplane:

$$\begin{pmatrix} \mathbf{w}^T & \underbrace{(\mathbf{x}'' - \mathbf{x}')}_{\text{vector on hyperplane}} \end{pmatrix} = 0$$

- 3 distance = project $(\mathbf{x} - \mathbf{x}')$ to \perp hyperplane



$$\text{distance}(\mathbf{x}, \mathbf{b}, \mathbf{w}) = \left| \frac{\mathbf{w}^T}{\|\mathbf{w}\|} (\mathbf{x} - \mathbf{x}') \right| \stackrel{\textcircled{1}}{=} \frac{1}{\|\mathbf{w}\|} |\mathbf{w}^T \mathbf{x} + b|$$

则目标函数为 $\arg\max(\mathbf{w}, \mathbf{b}) \left\{ \frac{1}{\|\mathbf{w}\|} \min(i) [(\mathbf{w}^t \cdot \mathbf{x}_i + b)] \right\} \dots\dots\dots (1)$

超平面： $y(\mathbf{x}) = \omega^t \mathbf{x} + b$ $|y|$ 通过缩放 \mathbf{w}, b 使两类的值都满足 $|y| \geq 1 \dots\dots\dots (2)$

由 1, 2 两式可得目标函数转换为： $\arg\max(\mathbf{w}, \mathbf{b}) \left\{ \frac{1}{\|\mathbf{w}\|} \right\} \dots\dots\dots (3)$

$$\text{s.t. } y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b) \geq 1, i = 1, 2 \dots, n$$

将目标函数转换为 $\arg\min(\mathbf{w}, \mathbf{b}) \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\} \dots\dots\dots (4)$

$$\text{s.t. } y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b) \geq 1, i = 1, 2 \dots, n$$

根据新的目标函数可知，目标函数为 2 次式，约束条件为一次式，典型的二次规划问题。使用拉格朗日乘子法。构造拉格朗日函数

$$L(\lambda, \mathbf{w}, \mathbf{b}) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \lambda_i (1 - y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b)) \dots\dots\dots (5)$$

$$\text{s.t. } y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b) \geq 1, i = 1, 2 \dots, n$$

$$\text{s.t. } \lambda_i \geq 0 \dots\dots\dots (6)$$

因为 $\lambda_i \geq 0$ 且 $1 - y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b) \leq 0$ 则 $\sum_{i=1}^N \lambda_i (1 - y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b))$ 小于等于 0，所以 $\max(\lambda) L(\lambda, \mathbf{w}, \mathbf{b}) = \frac{1}{2} \|\mathbf{w}\|^2$ ，即原问题可以表示为 $\min(\mathbf{w}, \mathbf{b}) \max(\lambda) L(\lambda, \mathbf{w}, \mathbf{b})$ ，其对偶问题为 $\max(\lambda) \min(\mathbf{w}, \mathbf{b}) L(\lambda, \mathbf{w}, \mathbf{b})$ ，SVM 刚好满足这些条件。

先求 $\min(\mathbf{w}, \mathbf{b}) L(\lambda, \mathbf{w}, \mathbf{b})$ 即对 $L(\lambda, \mathbf{w}, \mathbf{b})$ 对 (\mathbf{w}, \mathbf{b}) 求导等于 0。

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} + \sum_{i=1}^N (-\lambda_i y_i \mathbf{x}_i) = 0 \text{ 即 } \mathbf{w} = \sum_{i=1}^N \lambda_i y_i \mathbf{x}_i \dots\dots\dots (7)$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^N \lambda_i y_i = 0 \quad \dots\dots\dots(8)$$

$$\text{将 (7) 代入 (5) 得到 } \max(\lambda) \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j x_i^T x_j \quad \dots\dots\dots(9)$$

对偶问题有最优解必然满足 KKT 条件

$$\text{s.t. } \lambda_i (1 - y_i (w^T \cdot x_i + b)) = 0 \quad \dots\dots\dots(10)$$

由 10 式可以看出当 $1 - y_i (w^T \cdot x_i + b) = 0$ 时, 即 $y_i (w^T \cdot x_i + b) = 1$ 即样本点在支持向量, 当 $y_i (w^T \cdot x_i + b) \neq 1$ 时, $\lambda_i = 0$, 即这个样本点是支持向量以外的点, 此时 9 式中
可以省去很多计算, 换句话说, w, b 只与 $\lambda_i > 0$ 有关。

根据上面式子求解所有 λ_i , 将 λ_i 代入 7 式即可以求 $w, f(x) = w^T \cdot x + b = \sum_{i=1}^N \lambda_i y_i x_i^T x + b$

$$\text{对于任意支持向量 } (x_s, t_s) \text{ 有 } \lambda_s > 0 \text{ 且 } t_s f(x_s) = 1 \quad \dots\dots\dots(11)$$

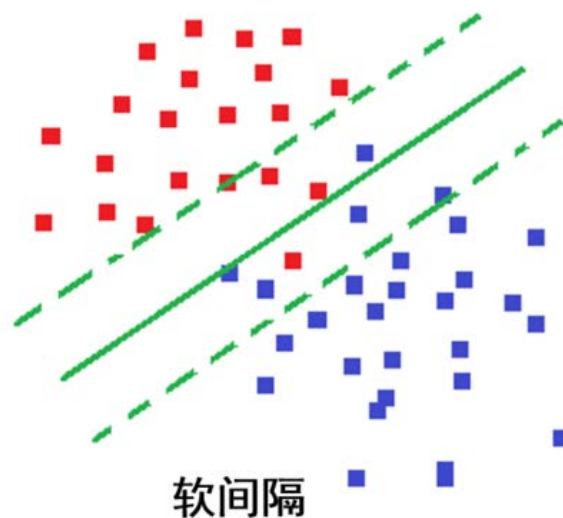
对于其他任意向量 (x_r, t_r) 有 $\lambda_r = 0$, 所以 $t_s (\sum_{i \in V} \lambda_i y_i x_i^T x_s + b) = 1$ 所以

$$b = \frac{1}{|S|} \sum (t_s - \sum_{i \in V} \lambda_i t_i x_i^T x_s) \quad \dots\dots\dots(12)$$

支持向量点的越少, 支持向量机的泛化性越好。

λ_i 怎么解? 对偶问题怎么求解?

线性支持向量机 (soft margin) 即可以允许部分分类存在错误, 但是给予损失项



同上目标函数变为在线性可分支持向量机的目标函数基础上加上损失项

$$\min(w, b) \frac{1}{2} \|w\|^2 + c \sum_{i=1}^N h(y_i (w^T x_i + b) - 1) \quad \dots\dots\dots(13)$$

其中 c 为惩罚系数, $h(z)$ 为损失函数

常见的 $h(z) \begin{cases} 1, & \text{if } z < 0 \\ 0, & \text{otherwise} \end{cases}$ 0/1 损失

指数损失 $h(z) = e^{-z}$

对率损失 $h(z) = \ln(1 + e^{-z})$

SVM 使用 hinge loss

Hinge 损失 $h(z) = \max(0, 1 - z)$ ，当分类正确时 z 大于等于 1，此时 $h(z) = 0$ ，即无损失，不惩罚，当分类错误是， z 小于 1，此时， $h(z) = 1 - z$ ，错误距离越大，则损失越大。

带有 hinge loss 的对应优化目标

$$\min(w, b) \frac{1}{2} ||w||^2 + c \sum_{i=1}^N \max(0, y_i(w^T x_i + b) - 1) \quad \dots\dots\dots(14)$$

引入松弛变量

$$\min(w, b) \frac{1}{2} ||w||^2 + c \sum_{i=1}^N \xi_i \quad \dots\dots\dots(15)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i \quad \dots\dots\dots(16)$$

$$\xi_i \geq 0, i = 1, \dots, N \quad \dots\dots\dots(17)$$

同样的方法得到对偶函数

$$\max(\lambda) \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j x_i^T x_j \quad \dots\dots\dots(18)$$

$$\text{s.t. } \sum_{i=1}^N \lambda_i y_i = 0$$

s.t. $0 \leq \lambda_i \leq c$ c 为惩罚系数，当 c 无穷大时，不允许有分类错误，即变为硬间隔。

同样怎么求解对偶问题？

SMO：通过把其他 λ 都固定住，仅仅优化某一对 λ ，这样就把 N 个 λ 优化变成优化两个 λ ，因为对偶问题最优解必定满足 KKT 条件，所以当所有 λ 满足 KKT 条件，则优化完成，得到最优解。

$\sum_{i=1}^N \lambda_i y_i = 0$ ，假设固定了 λ_1, λ_2 则原优化目标变为(省去常数项)：

$$\min(\lambda_1, \lambda_2) \frac{1}{2} K_{11} \lambda_1^2 + \frac{1}{2} K_{22} \lambda_2^2 + y_1 y_2 K_{12} \lambda_1 \lambda_2 + y_1 \lambda_1 \sum_{i=1}^N y_i \lambda_i K_{i1} + y_2 \lambda_2 \sum_{i=1}^N y_i \lambda_i K_{i2} - (\lambda_1 + \lambda_2) \quad \dots\dots\dots(19)$$

约束条件 s.t.

$$\lambda_1 y_1 + \lambda_2 y_2 = - \sum_{i=3}^n \lambda_i y_i = \xi \quad \dots\dots\dots(20)$$

$$0 \leq \lambda_i \leq c$$

由 20 式可得

$$\lambda_1 = (\xi - \lambda_2 y_2) y_1 \quad \dots\dots\dots(21)$$

将 21 代入到 19 可得关于 λ_2 的函数，函数求导等于 0，再将 $\xi = \lambda_1^{old} y_1 + \lambda_2^{old} y_2$ 代入

$$\text{可得 } \lambda_2^{new} = \lambda_2^{old} + \frac{y_2(E1-E2)}{\eta}$$

其中 $E1 = f(x_1) - y_1$

$$\eta = K_{11} + K_{22} - 2K_{12} = ||x_1 - x_2||^2$$

由于

$$0 \leq \lambda_i \leq c$$

$$\lambda_1 y_1 + \lambda_2 y_2 = - \sum_{i=3}^n \lambda_i y_i = \xi$$

所以约束共通过决定 λ_2 可行域（同种 $\lambda = \alpha$ ）

支持向量机主要可分为三类：线性可分支持向量机、线性支持向量机、非线性支持向量机

1.线性可分支持向量机

线性可分支持向量机也成为硬间隔（Hard margin）支持向量机，即要求所有的样本都要被分类正确。

输入： $T=\{(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)\}$

其中， x_i 表示训练数据集第 i 个实例，是一个向量。

y_i 表示第 x_i 的类标记 $y_i = \{-1, +1\}$

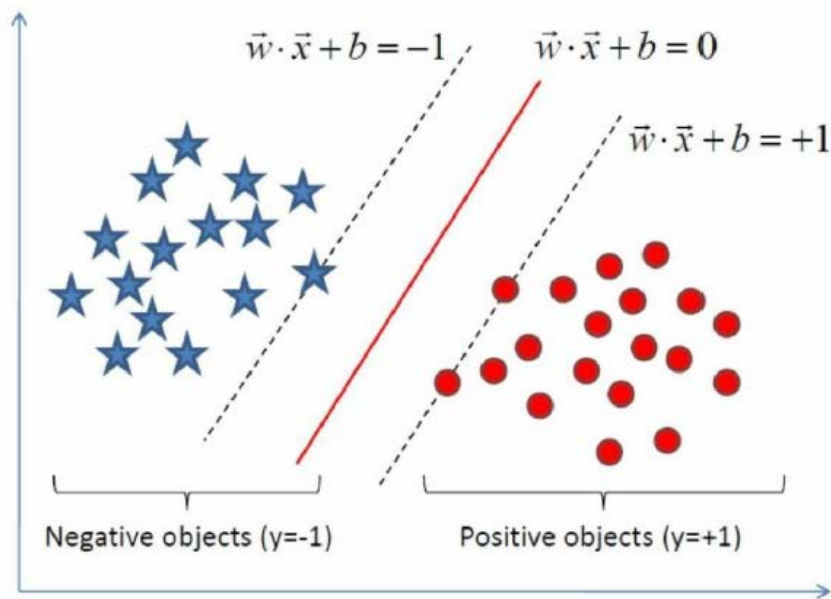
(x_i, y_i) 为样本点

给定线性可分数据集，通过间隔最大化得到分离超平面： $y(x) = \omega^t x + b$ 相应的分类决策函数为 $f(x) = \text{sign}(\omega^t x + b)$

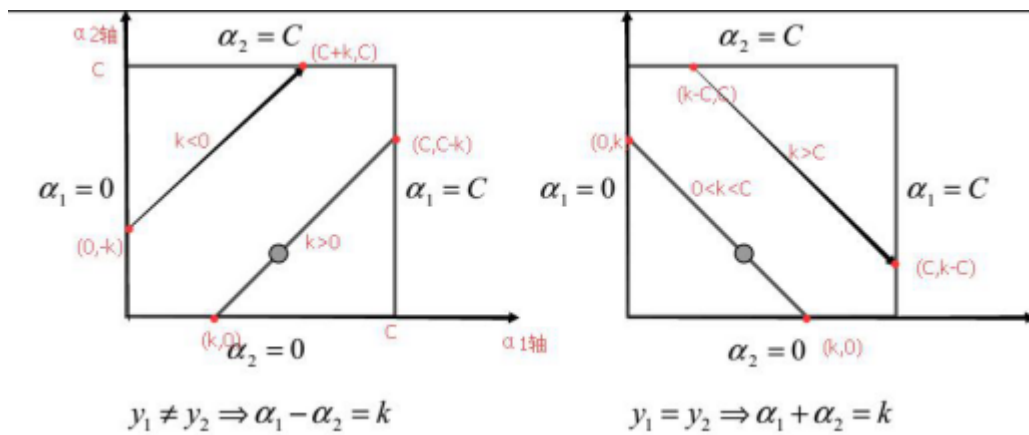
即线性可分 SVM 的任务是不仅要找到把样本点正确分类的超平面，而且是要找到间隔最大的超平面。

此时间隔的定义：设 C 和 D 为两不相交的凸集，则存在超平面 P ， P 可以将 C 和 D 分离，两集合间的距离定义为两集合中元素的最短距离， C, D 两集合最短距离的中垂线（多维空间即中垂面）即为超平面，任一集到超平面的距离即为间隔。

由上可知，线性可分 SVM 即求距离超平面最近的点的间隔最大化。从而增强了 SVM 的泛化能力。



距离求解：假设 x_1, x_2 均为超平面上的点，即 x_1, x_2 两点满足 $w \cdot x_1 + b = 0$ And $w \cdot x_2 + b = 0$ 即， $w \cdot (x_1 - x_2) = 0$ $x_1 - x_2$ 可以看成是超平面上的向量，即 w 为超平面法向量。点 X 到超平面的距离等于， X 在法向量方向上的投影长度。则



由于 y 的取值为 +1 或者 -1

当 y_1 与 y_2 同号时例如同为 1 时，即右边的图形

ξ 可能的几种取值

$\xi < 0$: α_2 为空值

$\xi = 0$: $\alpha_1 + \alpha_2 = 0$ 此时与正方形区域焦点为 $(0,0)$, $\alpha_2 = 0$

$0 < \xi < C$ 此时直线和正方形区域相交，即 α_2 的可行区间是 $[0, \xi]$ ，即 $[0, \alpha_1 + \alpha_2]$

当 $C < \xi < 2C$ 可行区间 $[\xi - C, C]$ 即 $[0, \alpha_1 + \alpha_2]$

当 $\xi > 2C$, 则 可行域为空集

综上 α_2 的可行域为 $[\max(0, \alpha_1 + \alpha_2 - C), \min(C, \alpha_1 + \alpha_2)]$ 即 $L = \max(0, \alpha_1 + \alpha_2 - C)$, $H = \min(C, \alpha_1 + \alpha_2)$

同理，当 y_1 与 y_2 异号时

α_2 的可行域为 $[\max(0, \alpha_1 - \alpha_2), \min(C, C - \alpha_1 + \alpha_2)]$ ，即 $L = \max(0, \alpha_1 - \alpha_2)$, $H = \min(C, C - \alpha_1 + \alpha_2)$

所以在更新 α_2 时，要先求出 α_2 的可行域，然后用之前的那个公式求出极值点 α_2 ，

然后看极值点 α_2 处的在不在可行域范围内，在的话就使用极值点处的 α_2 ，不在的

话就使用边界值 H 或者 L 更新

$$\alpha_2^{\text{new,clipped}} = \begin{cases} H & \text{if } \alpha_2^{\text{new}} \geq H; \\ \alpha_2^{\text{new}} & \text{if } L < \alpha_2^{\text{new}} < H; \\ L & \text{if } \alpha_2^{\text{new}} \leq L. \end{cases}$$

更新 α_2 后，可根据公式更新 α_1

如何选择 α_1 与 α_2

1. 寻找违反 KKT 条件的参数，因为最优解必满足 KKT 条件，当参数不满足 KKT 条件时就不是最优，所以进行优化。

(1) 首先遍历数据集，对每个不满足 KKT 条件的参数作为第一个待修改参数，作为外层循环

(2) 对整个数据集遍历完一遍之后，选择那些参数满足 $0 < \lambda < C$ 的子集，开始遍历，如果发现一个不满足 KKT 条件的作为第一个待修改参数，然后找到第二个待修改的参数，修改完之后，重新开始遍历这个子集

(3) 遍历完子集后，重新开始 (1) (2)，直到在 (1) (2) 执行时，没有任何修改就结束

2. 启发式寻找

找满足 $|E_1 - E_2|$ 最大

寻找一个随机位置满足 $0 < \alpha < C$ 的可以优化的参数进行修改

在整个数据集上寻找一个随机位置的可以优化的参数进行修改

都不行那就找下一个第一个参数