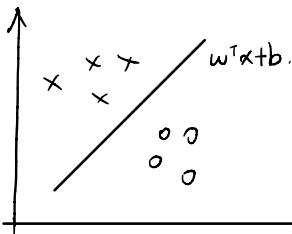


支持向量机 (SVM)

SVM有三宝：间隔，对偶，核技巧。

hard-margin SVM . soft-margin SVM Kernel SVM



模型 $f(w) = \text{sign}(w^T x + b) \rightarrow \text{判别模型}$.

很明显，有很多条直线都可以正确分离样本，但若超平面距离样本点过近，则它的泛化性能会变得很差。

因此，我们需要找到一个超平面，使得样本点到超平面的最小距离足够大。

$$y = w^T x + b$$
$$w^T x + b - y = 0$$

一. hard-margin SVM (最大间隔分类器)

设有N个样本点 $\{(x_i, y_i)\}_{i=1}^N$ $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$

即求 $\max \text{margin}(w, b)$

因为SVM为分类器，超平面实现将数据的正、负例分隔开，因此有：

$$\text{s.t. } \begin{cases} w^T x_i + b > 0, & y_i = +1 \\ w^T x_i + b < 0, & y_i = -1 \end{cases} \quad \text{等价为 } y_i(w^T x_i + b) > 0 \quad i=1, \dots, N.$$

间隔 $\text{margin}(w, b)$ 定义为点到超平面的距离：

$$\text{margin}(w, b) = \min_{x_i} \text{distance}(w, b) = \min_{x_i} \frac{|w^T x_i + b|}{\|w\|}$$

$$\Rightarrow \begin{cases} \max_{w, b} \min_{x_i} \frac{1}{\|w\|} |w^T x_i + b| \\ \text{s.t. } y_i(w^T x_i + b) > 0, \end{cases}$$

其中，由于 $y_i = +1$ 或 -1 , y_i 与 $w^T x_i + b$ 同号，因此

$|w^T x_i + b| \Leftrightarrow y_i(w^T x_i + b)$, 又 $\min y_i$ 与 x_i 相关, 与 w 无关, 因此上式可写为:

$$\max_{w,b} \frac{1}{\|w\|} \min_{x_i} y_i(w^T x_i + b)$$

观察约束条件: $y_i(w^T x_i + b) > 0$, 则 $\exists \gamma > 0$, 使得 $\min_{x_i} y_i(w^T x_i + b) = \gamma$.

由于对同一个超平面, $w^T x_i + b$ 与 $2w^T x_i + 2b$ 等价, 表示同一个超平面, 因此我们可以固定一个 $\|w\|$ 的值, 使得 $\min y_i(w^T x_i + b) = 1$.

$$\Rightarrow \begin{cases} \max_{w,b} \frac{1}{\|w\|} \\ \text{s.t. } \begin{cases} \min_{x_i} y_i(w^T x_i + b) = 1 \\ y_i(w^T x_i + b) \geq 1 \end{cases} \end{cases} \Rightarrow \begin{cases} \min_{w,b} \frac{1}{2} w^T w \\ \text{s.t. } y_i(w^T x_i + b) \geq 1, \text{ for } \forall i = 1, \dots, N. \end{cases}$$

这是一个带 N 个约束的凸优化问题.

二、优化问题的求解

带约束

$$\begin{cases} \min_{w,b} \frac{1}{2} w^T w \\ \text{s.t. } y_i(w^T x_i + b) \geq 1 \Leftrightarrow 1 - y_i(w^T x_i + b) \leq 0. \end{cases}$$

该优化问题可用拉格朗日乘子法来求解, 构建拉格朗日函数:

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_{i=1}^N \lambda_i (1 - y_i(w^T x_i + b))$$

则

$$\begin{cases} \min_{w,b} \max_{\lambda} L(w,b,\lambda) \\ \text{s.t. } \lambda_i \geq 0. \end{cases}$$

对 $\min_{w,b} \max_{\lambda} L(w,b,\lambda)$ 的理解:

(对样本 (x_i, y_i) , 若 $1 - y_i(w^T x_i + b) > 0$, 则 $\max_{\lambda} L = \frac{1}{2} w^T w + \infty = \infty$)
 (若 $1 - y_i(w^T x_i + b) \leq 0$, 则 $\max_{\lambda} L = \frac{1}{2} w^T w + 0 = \frac{1}{2} w^T w$)
 $\therefore \min_{w,b} \max_{\lambda} L(w,b,\lambda) = \min_{w,b} (\infty, \frac{1}{2} w^T w) = \min_{w,b} \frac{1}{2} w^T w.$

即如果 $1 - y_i(w^T x_i + b) > 0$ 不符合约束条件时, $\max_{\lambda} L = \infty$, 无论 w, b 取何值都是无穷大; 当符合约束条件时 $\max_{\lambda} L = \frac{1}{2} w^T w$, 而无论 w, b 取何值, $\frac{1}{2} w^T w$ 都比 ∞ 小. $\therefore \min_{w,b} \max_{\lambda} L = \min_{w,b} \frac{1}{2} w^T w$. 不符合约束条件的情况被去掉了.

凤尾 > 鸡头

等价为对偶问题为：

$$\begin{cases} \max_{\lambda} \min_{w,b} L(w,b,\lambda) \\ \text{s.t. } \lambda_i \geq 0. \end{cases}$$

$$\begin{array}{l} \min_{w,b} \max_{\lambda} L \geq \max_{w,b} \min_{\lambda} L \\ \geq \text{弱对偶关系} \\ = \text{强对偶关系} \end{array}$$

由于该问题为凸优化问题，所以 $\min_{w,b} \max_{\lambda} L = \max_{\lambda} \min_{w,b} L$ ，为强对偶关系，因此等价。

模型求解：

$\min_{w,b} L(w,b,\lambda)$ ，此时 w, b 无任何约束。

(1) 对 b 求偏导

$$\begin{aligned} \frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \left[\sum_{i=1}^N \lambda_i - \sum_{i=1}^N \lambda_i y_i (w^T x_i + b) \right] \\ &= \frac{\partial}{\partial b} \left[- \sum_{i=1}^N \lambda_i y_i b \right] = \sum_{i=1}^N \lambda_i y_i = 0 \end{aligned}$$

将其代入 $L(w,b,\lambda)$ ：

$$\begin{aligned} L(w,b,\lambda) &= \frac{1}{2} w^T w + \sum_{i=1}^N \lambda_i - \sum_{i=1}^N \lambda_i y_i (w^T x_i + b) \\ &= \frac{1}{2} w^T w + \sum_{i=1}^N \lambda_i - \sum_{i=1}^N \lambda_i y_i w^T x_i - \sum_{i=1}^N \lambda_i y_i b \rightarrow 0 \\ &= \frac{1}{2} w^T w + \sum_{i=1}^N \lambda_i - \sum_{i=1}^N \lambda_i y_i w^T x_i \end{aligned}$$

(2) 对 w 求偏导

$$\frac{\partial L}{\partial w} = \frac{1}{2} \cdot 2w - \sum_{i=1}^N \lambda_i y_i x_i \triangleq 0 \Rightarrow w = \boxed{\sum_{i=1}^N \lambda_i y_i x_i}^{w^*}$$

将其代入 $L(w,b,\lambda)$ ：

$$\begin{aligned} \min_b L(w,b,\lambda) &= \frac{1}{2} \left(\sum_{i=1}^N \lambda_i y_i x_i \right)^T \left(\sum_{j=1}^N \lambda_j y_j x_j \right) - \sum_{i=1}^N \lambda_i y_i \left(\sum_{j=1}^N \lambda_j y_j x_j \right)^T x_i + \sum_{i=1}^N \lambda_i \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^N \lambda_i \end{aligned}$$

因此该优化问题相当于：

$$\begin{cases} \max_{\lambda} -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^N \lambda_i \\ \text{s.t. } \lambda_i \geq 0, \sum_{i=1}^N \lambda_i y_i = 0 \end{cases}$$

$$\text{BP} \quad \begin{cases} \min_{\lambda} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_{i=1}^N \lambda_i \\ \text{s.t. } \lambda_i \geq 0, \sum_{i=1}^N \lambda_i y_i = 0. \end{cases}$$

三. KKT 条件.

$$\left\{ \begin{array}{l} \frac{\partial L}{\partial w} = 0 \quad \frac{\partial L}{\partial b} = 0 \quad \frac{\partial L}{\partial \lambda} = 0 \\ \lambda_i(1 - y_i(w^T x_i + b)) = 0 \\ \lambda_i \geq 0 \\ 1 - y_i(w^T x_i + b) \leq 0 \end{array} \right.$$

松弛互补条件.

原问题, 对偶问题满足强对偶关系
 \Leftrightarrow 满足 KKT 条件.

\star 松弛互补条件: λ_i 和 $1 - y_i(w^T x_i + b)$ 总有一个为 0. 也就是说只有支撑向量对应的 x_i 才可能有值 ($\lambda_i \neq 0$), 而其它不在 $w^T x + b = 1$ 和 $w^T x + b = -1$ 上的样本点, 对应的不一定为 0, 该性质可以用来求 b^*

$\exists (x_k, y_k)$ 使得 $1 - y_k(w^T x_k + b) = 0$.

$$R) \quad y_k(w^T x_k + b) = 1$$

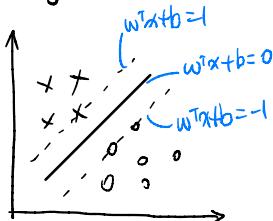
$$y_k^2(w^T x_k + b) = y_k$$

$$b^* = y_k - w^T x_k = y_k - \sum_{i=1}^N \lambda_i y_i x_i^T x_k.$$

$$B) \quad w^* = \sum_{i=1}^N \lambda_i y_i x_i \quad b^* = y_k - \sum_{i=1}^N \lambda_i y_i x_i^T x_k.$$

可以看出: w^* 是 data 的线性组合.

soft-margin SVM.



训练数据通常不是理想的线性可分, 有时甚至是线性不可分的数据. 此外, 对于一些噪声数据, 我们应该允许一点分类错误.

因此, 目标函数可调整为:

$$\min \frac{1}{2} w^T w + \Delta loss$$

① 使用误分类点的个数作为 loss

$$\text{loss} = \sum_{i=1}^N I\{y_i(w^T x_i + b) < 1\}$$

I 为指示函数

$$\text{令 } z = y_i(w^T x_i + b)$$

$$\text{loss}_{0/1} = \begin{cases} 1, & z < 1 \\ 0, & \text{otherwise} \end{cases}$$

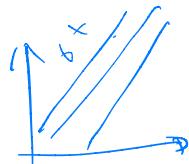
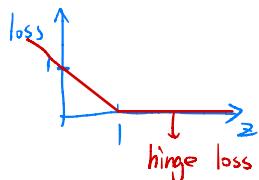
不连续，不可导。

② loss：距离 \rightarrow hinge loss

$$\begin{cases} \text{如果 } y_i(w^T x_i + b) \geq 1, \text{ loss} = 0 \\ \text{如果 } y_i(w^T x_i + b) < 1, \text{ loss} = 1 - y_i(w^T x_i + b) \end{cases}$$

$$\Rightarrow \text{loss} = \max\{0, 1 - y_i(w^T x_i + b)\}$$

$$\text{loss} = \max\{0, 1 - z\}$$



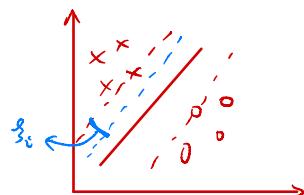
优化目标变为

$$\begin{cases} \min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^N \max\{0, 1 - y_i(w^T x_i + b)\} \\ \text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i \end{cases}$$

- 和 soft-margin 并不写成 max 的形式，

$$\text{写 } \lambda \xi_i = 1 - y_i(w^T x_i + b), \xi_i \geq 0$$

$$\boxed{\begin{cases} \min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \\ \text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}}$$



若有一个样本分错，则会给予 $1 - y_i(\mathbf{w}^T \mathbf{x}_i - b)$ 的惩罚。C 代表惩罚系数，当 C 取无穷大时，有一个点分错就会带来巨大的惩罚，因此模型就会使所有点都满足限制条件，即当 C 为无穷大时，模型又变为硬间隔。

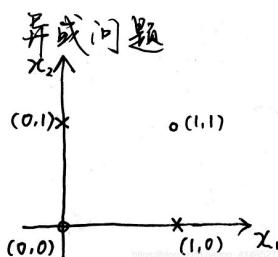
Kernel Method

Kernel Function

非线性带来高维转换（从模型角度）
对偶表示带来内积（从优化角度）

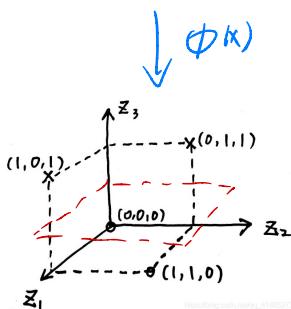
线性可分	一点点错误	严格非线性	$\phi(\mathbf{x})$ 为非线性转换
PLA	Pocket Algorithm	$\phi(\mathbf{x}) + PLA$	
hard-margin SVM	Soft-margin SVM	$\phi(\mathbf{x}) + hard-margin SVM$	

有时线性可分的数据夹杂一点噪音，可以通过改进算法来分类，比如感知机的口袋算法和 Soft-margin SVM，但有时数据是完全非线性可分的，如异或问题：



我们可以将数据映射到高维空间后实现线性可分。对于异或问题，我们可以通过寻找一个映射 ϕ 将低维空间中的数据 \mathbf{x} 映射成高维空间中的 \mathbf{z} 来实现数据的线性可分。

$$\text{假设 } \mathbf{x} = (x_1, x_2) \xrightarrow{\phi(\mathbf{x})} \mathbf{z} = (x_1, x_2, (x_1 - x_2)^2)$$



Cover Theorem：高维空间比低维空间更容易线性可分。

Primal Problem

$$\begin{cases} \min_{w,b} \frac{1}{2} w^T w \\ \text{s.t. } y_i(w^T x_i + b) \geq 1 \end{cases}$$

Dual Problem

$$\begin{cases} \max_{\lambda} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j [x_i^T x_j] - \sum_{i=1}^N \lambda_i \\ \text{s.t. } \sum_{i=1}^N \lambda_i y_i = 0 \end{cases}$$

$\phi(x)^T \phi(x_j)$

在线性可分问题中对偶问题会让我们求出 x_i 的内积，但如果问题是非线性可分的，当我们将其映射到高维空间时，我们就需要求其映射后向量的内积，对于现实问题来说，往往高维空间维度非常大甚至是无限维的，计算内积的计算量会很大，但我们关心的是内积之后的值，而不想计算 $\phi(x)$ 和 $\phi(x_j)$ 。

Kernel Function 的引入恰恰用于解决该问题。

Kernel Function:

$$K(x, x') = \phi(x)^T \phi(x') = \langle \phi(x), \phi(x') \rangle$$

定义：对 $\forall x, x' \in X$, $\exists \phi: x \rightarrow z$, 使得 $K(x, x') = \phi(x)^T \phi(x')$, 则称 $K(x, x')$ 是一个核函数。

比如 $K(x, x') = \exp(-\frac{(x-x')^2}{2\sigma^2})$, 我们直接将 x, x' 代入即可求出 $\phi(x)^T \phi(x')$
核函数实际上蕴含了非线性转换以及非线性转换上的内积。

正定核两个定义

Hilbert 空间

定义1. $K: X \times X \mapsto \mathbb{R}$, $\phi \in \mathcal{H}$, 使得 $K(x, z) = \langle \phi(x), \phi(z) \rangle$, 则称 $K(x, z)$ 为正定核函数。

定义2. $K: X \times X \mapsto \mathbb{R}$, $\forall x, z \in X$, 有 $K(x, z)$, 若 $K(x, z)$ 满足如下两条性质：

① 对称性

② 正定性

则称 $K(x, z)$ 为正定核函数

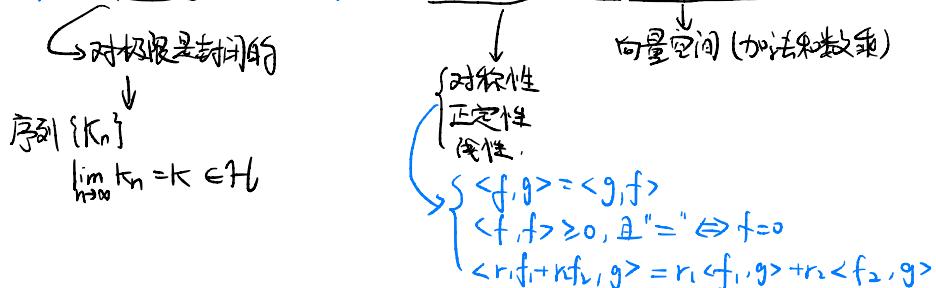
① 对称性 $\Leftrightarrow K(x, z) = K(z, x)$

② 正定性 \Leftrightarrow 选取 N 个元素, $x_1, x_2, \dots, x_N \in X$, 对应的 Gram 矩阵是半正定的。

$$K = [K(x_i, x_j)]$$

要证： $K(x, z) = \phi(x)^T \phi(z) \Leftrightarrow \text{Gram matrix 半正定}$

Hilbert space : 完备的，可能是无限维的被赋予内积的向量空间



必要性证明 (\Rightarrow)

已知 $K(x, z) = \phi(x)^T \phi(z)$. 说明 Gram matrix 半正定，且 $K(x, z)$ 对称

证： $K(x, z) = \langle \phi(x), \phi(z) \rangle$

对称性 $K(z, x) = \langle \phi(z), \phi(x) \rangle$

又内积具有对称性质, 即 $\langle \phi(x), \phi(z) \rangle = \langle \phi(z), \phi(x) \rangle$

$\therefore K(x, z) = K(z, x)$, 即 $K(x, z)$ 满足对称性.

正定性 故记 Gram matrix: $K = [K(x_i, x_j)]_{N \times N}$ 半正定, 即证: $\forall \alpha \in \mathbb{R}^N, \alpha^T K \alpha \geq 0$,

$$\alpha^T K \alpha = (\alpha_1, \alpha_2, \dots, \alpha_N) \begin{pmatrix} K_{11} & \cdots & K_{1N} \\ \vdots & \ddots & \vdots \\ K_{N1} & \cdots & K_{NN} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix}$$

$$= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j K_{ij}$$

$$= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle$$

$$= \sum_{i=1}^N \alpha_i \langle \phi(x_i)^T \phi(x_i) \rangle$$

$$= \left[\sum_{i=1}^N \alpha_i \phi(x_i) \right]^T \cdot \sum_{j=1}^N \alpha_j \phi(x_j)$$

$$= \langle \sum_{i=1}^N \alpha_i \phi(x_i), \sum_{j=1}^N \alpha_j \phi(x_j) \rangle$$

$$= \left\| \sum_{i=1}^N \alpha_i \phi(x_i) \right\|^2 \geq 0 \quad \because \text{大是半正定的}$$

充分性(\Leftarrow):

对 K 进行特征分解, 对于对称矩阵 $A = V \Lambda V^T$, 那么令 $\Phi(x) = \sqrt{\lambda_i} V_i$.
其中 V_i 是特征向量, 于是就构造了 $K(x, z) = \sqrt{\lambda_i} V_i^T V_j$.