2.8 支持向量机
（Support Vector Machines, SVM）

CSDN学院

2017年11月

▶ 大纲

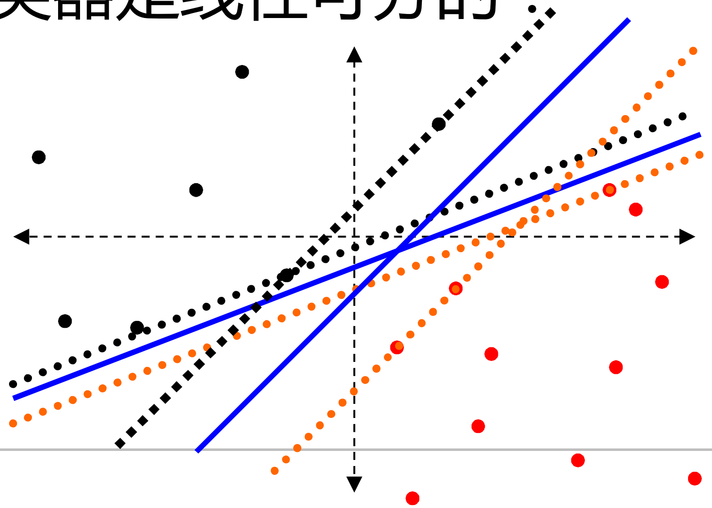
- SVM基本原理
- 带松弛因子的SVM：C-SVM
- 核方法
- 支持向量回归
- Scikit learn中的SVM实现
- 案例分析

SVM基本原理

- 最大间隔原则
- 对偶表示(Dual Representation)
- KKT条件

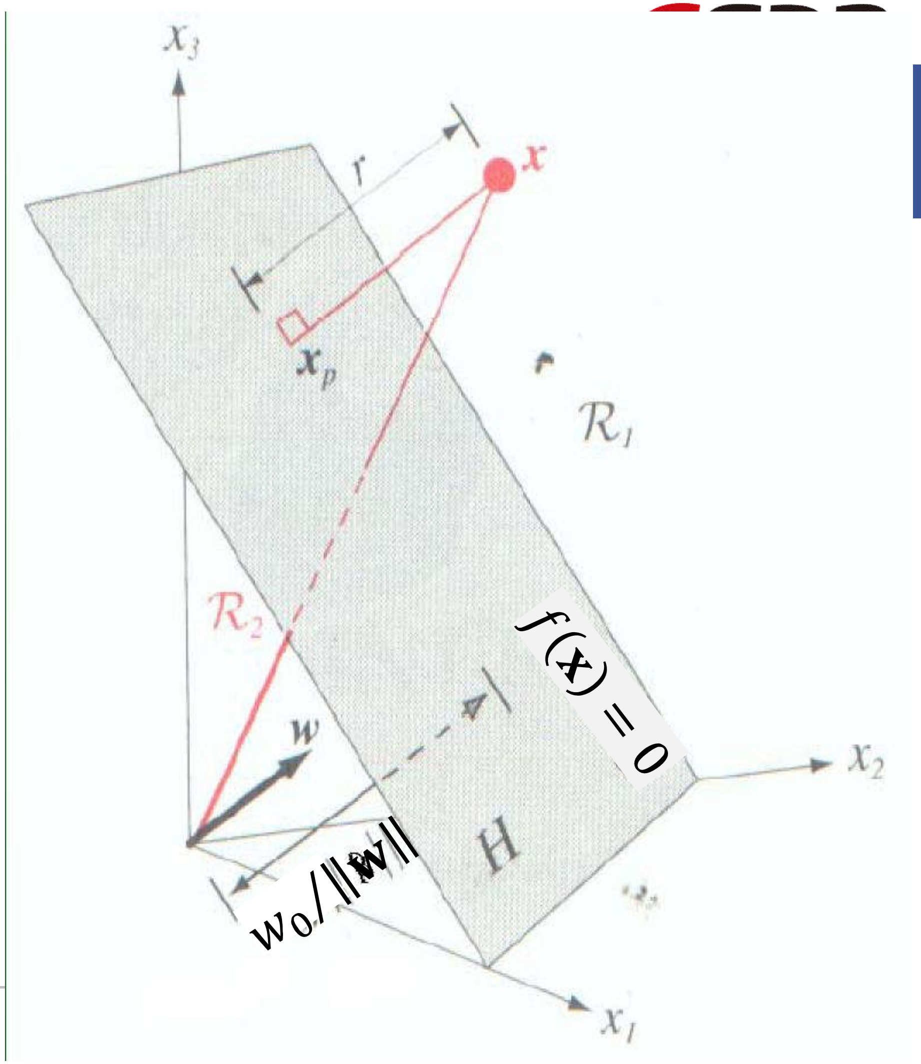
▶ SVM as 最大间隔分类器

- 最大间隔原则：最大化两个类最近点之间的距离
  - 这个距离被称为间隔(*margin*)。
  - 边缘上的点被称为支持向量(*support vectors*)。
- 我们先假设分类器是线性可分的：



▶ 间隔

- 线性分类面：$f(\mathbf{x})=\mathbf{w}^{\mathrm{T}}\mathbf{x}+w_0$
- 则有 $\mathbf{x}=\mathbf{x}_{\mathrm{p}}+r\dfrac{\mathbf{w}}{\|\mathbf{w}\|}$
- 其中**x**到分类面的距离r



▶ 间隔

- 代入得到 $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \mathbf{w}^{\mathrm{T}}\left(x_{\mathrm{p}} + r\dfrac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0$

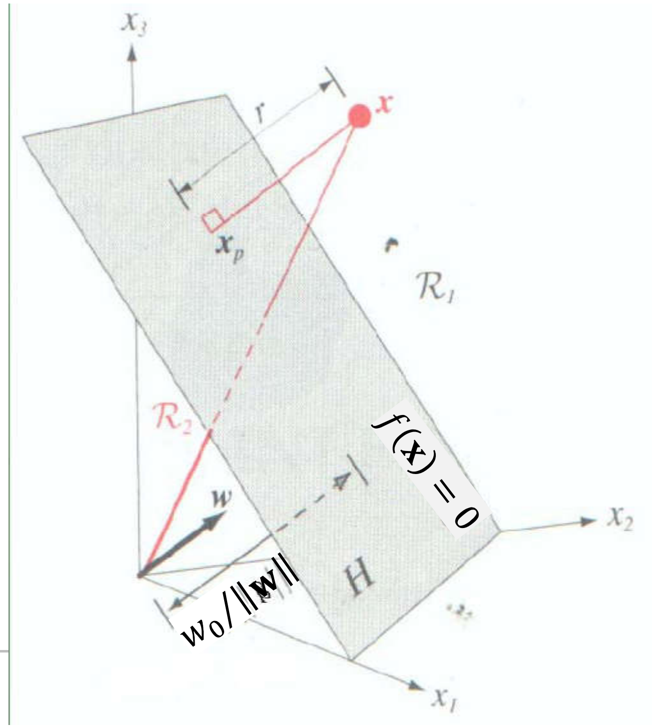
$$\boxed{\begin{array}{l} f(\mathbf{x}_{\mathrm{p}}) = \mathbf{w}^{\mathrm{T}} x_{\mathrm{p}} + w_0 = 0 \\ \mathbf{w}^{\mathrm{T}}\mathbf{w} = \|\mathbf{w}\|^2 \end{array}}$$

$$= \mathbf{w}^{\mathrm{T}} x_{\mathrm{p}} + r\frac{\mathbf{w}^{\mathrm{T}}\mathbf{w}}{\|\mathbf{w}\|} + w_0$$

$$\Rightarrow r = \frac{f(\mathbf{x})}{\|\mathbf{w}\|}$$

- 当**x**=0时，原点到分类面的距离

$$r_0 = \frac{f(\mathbf{0})}{\|\mathbf{w}\|} = \frac{w_0}{\|\mathbf{w}\|}$$



▶ 线性判别函数

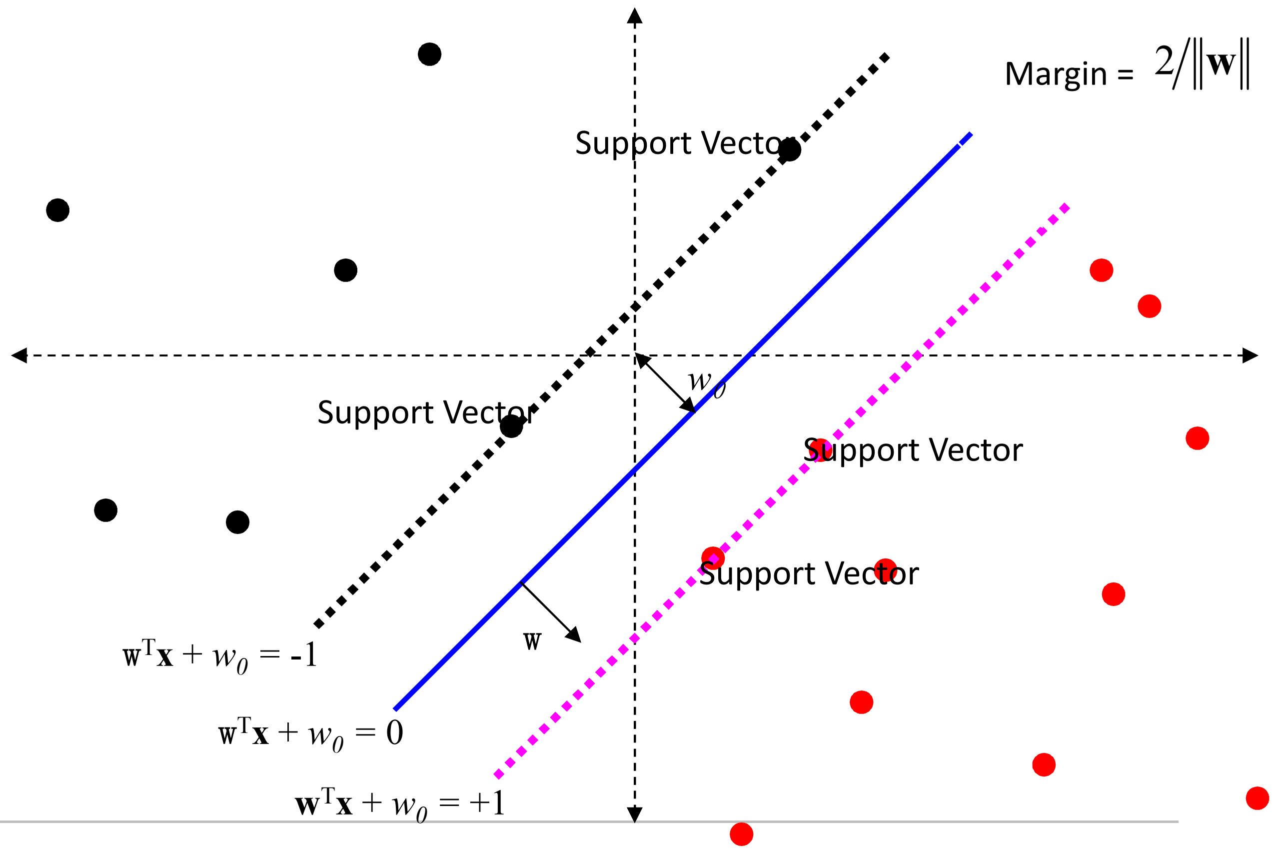
- 线性判别函数利用一个超平面把特征空间分隔成两个区域。
- 超平面的方向由法向量**w**确定，它的位置由阈值w_0确定。
- 判别函数$f(\mathbf{x})$正比于**x**点到超平面的代数距离（带正负号）
  - 当**x**点在超平面的正侧时，$f(\mathbf{x}) > 0$；
  - 当**x**点在超平面的负侧时，$f(\mathbf{x}) < 0$

$$\boxed{r = \frac{f(\mathbf{x})}{\|\mathbf{w}\|}}$$

  - **x**点到超平面的距离 $ry_i = \dfrac{y_i f(\mathbf{x})}{\|\mathbf{w}\|}$ 可视为对**x**判别的"置信度"

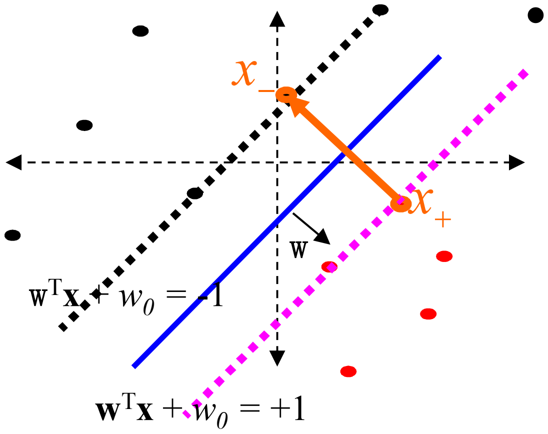
$$y_i \in \{1, -1\}$$

SVM 符号表示



Margin = $2/\|\mathbf{w}\|$
Support Vector
Support Vector
Support Vector
Support Vector
w_0
w
$\mathrm{w}^T\mathbf{x} + w_0 = -1$
$\mathrm{w}^T\mathbf{x} + w_0 = 0$
$\mathbf{w}^T\mathbf{x} + w_0 = +1$

▶ 间隔计算



- 间隔 $= |\mathbf{x}_+ - \mathbf{x}_-|$

$$= \lambda \|\mathbf{w}\| \qquad (\because \mathbf{x}_+ = \mathbf{x}_- + \lambda \mathbf{w})$$

$$= \frac{2}{\|\mathbf{w}\|^2}\|\mathbf{w}\| \qquad \left(\because \lambda = \frac{2}{\|\mathbf{w}\|^2}\right)$$

$$= \frac{2}{\|\mathbf{w}\|}$$

$$\left.\begin{array}{l} w_0 + \mathbf{w}^T x_+ = 1 \\ w_0 + \mathbf{w}^T x_- = -1 \end{array}\right\} \Rightarrow \mathbf{w}^T(\mathbf{x}_+ - \mathbf{x}_-) = 2$$

$$\Rightarrow \lambda \mathbf{w}^T \mathbf{w} = 2 \Rightarrow \lambda = \frac{2}{\|\mathbf{w}\|^2}$$

SVM：最大间隔

- 最大化间隔的超平面为

$$\max_{w_0,\mathbf{w}} \frac{2}{\|\mathbf{w}\|}, \quad \text{subject to} \quad y_i\left(w_0+\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\geq 1, \quad \forall i$$

- 等价于

$$\min_{w_0,\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2, \quad \text{subject to} \quad y_i\left(w_0+\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\geq 1, \quad \forall i$$

  - 二次规划问题(目标函数为二次函数，约束为线性约束)
  - 变量数为D+1，约束项的数目为N

▶ SVM基本原理

- 最大间隔原则
- 对偶表示(Dual Representation)
- KKT条件

► 对偶表示

- 凸优化理论告诉我们可以将该优化问题等价地写成其对偶形式(*dual formulation*)。
- 定义拉格朗日函数

$$L(\boldsymbol{\alpha}, w_0, \mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i\left(y_i\left(w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) - 1\right), \quad \alpha_i \geq 0$$

- 求使得目标 $L(\boldsymbol{\alpha}, w_0, \mathbf{w})$ 最小的对w_0和$\mathbf{w}$：

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

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i y_i = 0$$

▶ 对偶表示

CSDN 不止于代码

- 将 $w_0, \mathbf{w}$ 从 $L(\boldsymbol{\alpha}, w_0, \mathbf{w})$ 消去，得到对偶表示

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

$$L(\boldsymbol{\alpha}, w_0, \mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i\left(y_i\left(w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) - 1\right)$$

$$= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \sum_{i=1}^{N}\alpha_i y_i w_0 + \sum_{i=1}^{N}\alpha_i$$

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

$$\mathbf{w}^T\mathbf{w} = \left(\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i\right)^T\left(\sum_{j=1}^{N}\alpha_j y_j \mathbf{x}_j\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$\sum_{i=1}^{N}\alpha_i y_i \mathbf{w}^T\mathbf{x}_i = \sum_{i=1}^{N}\alpha_i y_i\left(\sum_{j=1}^{N}\alpha_j y_j \mathbf{x}_j\right)^T\mathbf{x}_i = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

▶ 对偶表示

- 解对偶问题：寻找 $\{\alpha_i\}_{i=1}^{N}$ 最大化目标函数

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

  – 满足限制

$$\begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \end{cases}$$

  – 仍然是一个QP问题：变量数为N，约束项的数目为$(N+1)$
    - 当N较大时，对偶问题的复杂度可能比原问题更高，
    - 但对偶问题可利用kernel trick与核方法结合
    - 对此问题有高效的求解算法：SMO（Sequential Minimal Optimization）

▶ 优化算法

- 最流行的SVM训练算法： SMO (*sequential minimal optimization*)
  - 坐标下降法
  - 在SVM中，因为 $\Delta L_{w_0} = 0 \Rightarrow \boldsymbol{\alpha}^* y\mathbf{1} = 0$ ，所以不能单独改变一个 α_i ，而是每次每次选取一对 α_i, α_j 做优化

SVM模型

- 求解出 α_i 后，再求出 $\mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i$ 和 w_0 ，可得到判别函数为

$$f(\mathbf{x}) = w_0 + \mathbf{w}^{\mathrm{T}} x = w_0 + \sum_i \alpha_i y_i \mathbf{x}_i^T \mathbf{x} = w_0 + \sum_i \alpha_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle$$

- 对一个新的点x进行预测，分类器为

$$\hat{y} = \operatorname{sgn}\left(f(\mathbf{x})\right)$$

SVM基本原理

- 最大间隔原则
- 对偶表示(Dual Representation)
- KKT条件

▶ 对偶性(Duality)

- 原问题：$P = \min_{\mathbf{x}} f_0(\mathbf{x})$

$$s.t. f_i(\mathbf{x}) \le 0, \quad 1 \le i \le N$$
$$h_j(\mathbf{x}) = 0, \quad 1 \le j \le M$$

- 拉格朗日函数：$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x}) + \sum_j \mu_i h_i(\mathbf{x})$

- 对偶问题：$D = \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$

$$s.t. \lambda_i \ge 0$$

▶ 对偶性

- 拉格朗日对偶通常是凹的（即使原问题非凸），可能更容易优化求解

- 弱对偶性：$P \geq D$
  - 总是成立

- 强对偶性：$P = D$
  - 并不总是成立
  - 对凸问题通常成立
  - 对SVM QP成立

Karush-Kuhn-Tucker (KKT) Conditions

- 如果强对偶条件成立，则对最优的 $\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*$ ，必须满足下述KKT条件

- 原问题的可行域： $f_i(\mathbf{x}^*) \le 0,\ h_j(\mathbf{x}^*) = 0$
- 对偶问题的可行域：$\boldsymbol{\lambda}^* \ge 0$
- 平稳条件： $\Delta_x L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0$
- 互补松弛条件： $\lambda_i^* f_i(\mathbf{x}^*) = 0$

- 如果 $\mathbf{x}^+, \boldsymbol{\lambda}^+, \boldsymbol{\mu}^+$ 满足凸问题的KKT条件，则其是最优的。

SVM – Duality

不止于代码

- 原问题：$P = \min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w}$

  $s.t.\ \ y_i\left(w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) \geq 1$

- 拉格朗日函数: $L(\mathbf{x},\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_i \alpha_i\left(y_i\left(w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) - 1\right)$

- 对偶问题：$D = \max_{\alpha}\ \left(\mathbf{1}^T\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{yKy}\right), where\ K_{ij} = \left\langle x_i, x_j \right\rangle$

  $s.t.\ \alpha_i \geq 0$

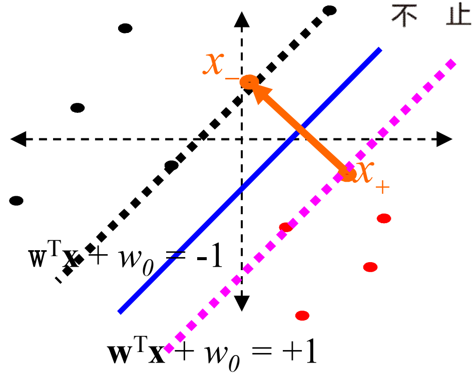
SVM – KKT Conditions

- 拉格朗日函数: $L(\mathbf{x},\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \boldsymbol{\alpha}^T\left(\mathbf{y}\left(\mathbf{x}^T\mathbf{w} + w_0\mathbf{1}\right) - 1\right)$
- 对偶问题的可行域： $\alpha_i^* \ge 0$
- 原问题的可行域： $y_i\left(w_0^* + \mathbf{w}^{*T}\mathbf{x}_i\right) - 1 \ge 0$
- 互补松弛条件： $\alpha_i^*\left[y_i\left(w_0^* + \mathbf{w}^{*T}\mathbf{x}_i\right) - 1\right] = 0$
- 平稳条件： $\Delta L_{\mathbf{w}} = 0 \Rightarrow \mathbf{w}^* = \mathbf{x}y\boldsymbol{\alpha}^*$

  $\Delta L_{w_0} = 0 \Rightarrow \boldsymbol{\alpha}^* y\mathbf{1} = 0$

► α的稀疏性

$$\boxed{\alpha_i^*\left[y_i\left(w_0^*+\mathbf{w}^{*T}\mathbf{x}_i\right)-1\right]=0}$$



- 根据KKT条件，对每个点

$$\alpha_i^*=0 \quad or \quad y_i\left(w_0^*+\mathbf{w}^{*T}x_i\right)=1$$

- 当 $\alpha_i^*=0$ 时，该点在决策函数

$$f(\mathbf{x})=w_0+\mathbf{w}^{\mathrm{T}}\mathbf{x}=w_0+\sum_i \alpha_i y_i\langle\mathbf{x},\mathbf{x}_i\rangle$$

- 中不起作用

- 其他点称为支持向量，满足 $y_i\left(w_0^*+\mathbf{w}^{*T}\mathbf{x}_i\right)=1$
- 对应位于最大间隔超平面上的点
  - 模型训练好后，大多数点可以抛掉，只需保留支持向量

► w_0的计算

- 由于支持向量满足 $y_i\left(w_0+\mathbf{w}^T\mathbf{x}_i\right)=1$
- 将 $f(\mathbf{x})=w_0+\mathbf{w}^{\mathrm{T}}\mathbf{x}=\sum_i\alpha_i y_i\langle\mathbf{x},\mathbf{x}_i\rangle$
- 代入，得到

$$y_i\left[\sum_{m\in\mathcal{S}}\alpha_m y_m\langle\mathbf{x}_i,\mathbf{x}_m\rangle+w_0\right]=1$$

  – 用任意一个支持向量即可求得w_0
- 为了得到更稳定的解，两边同乘以y_i，$y_i^2=1$
- 并对所有的支持向量求平均，得到

$$w_0=\frac{1}{N_{\mathcal{S}}}\sum_{m\in\mathcal{S}}\left[y_i-\sum_{m\in\mathcal{S}}\alpha_m y_m\langle\mathbf{x}_i,\mathbf{x}_m\rangle\right]$$

▶ 小结

- SVM基本原理
  - 最大间隔原则
  - 对偶表示(Dual Representation)
  - KKT条件

不 止 于 代 码

THANK YOU

AI100