机器学习
支持向量机

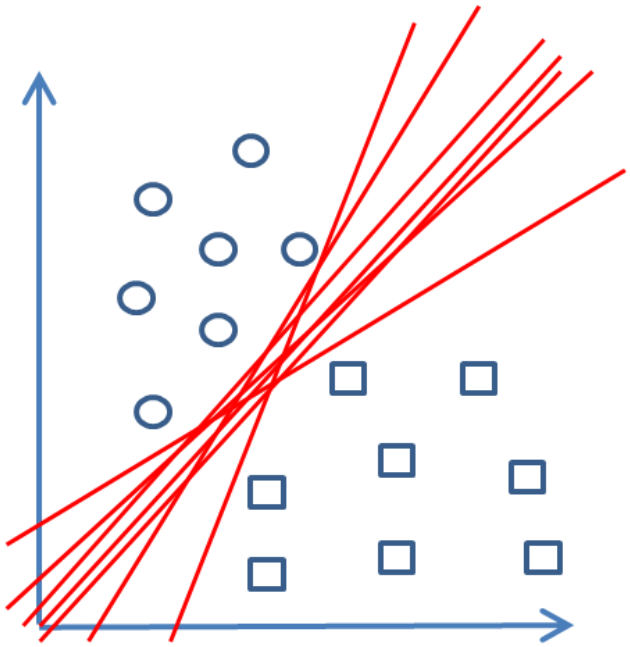
夏睿
文本挖掘组
南京理工大学
rxia@njust.edu.cn

概述

- 最大间隔线性分类器
- 对偶优化
- Soft-margin SVM
- 核函数
- *序列最小优化
- SVM工具包的使用

最大间隔线性分类器

回忆之前的线性分类

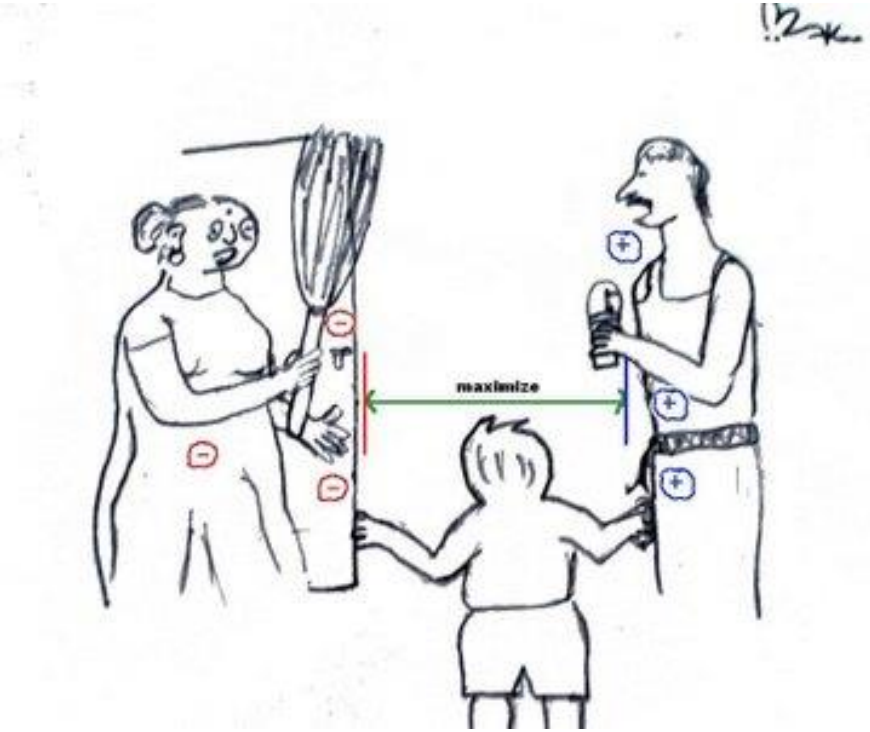


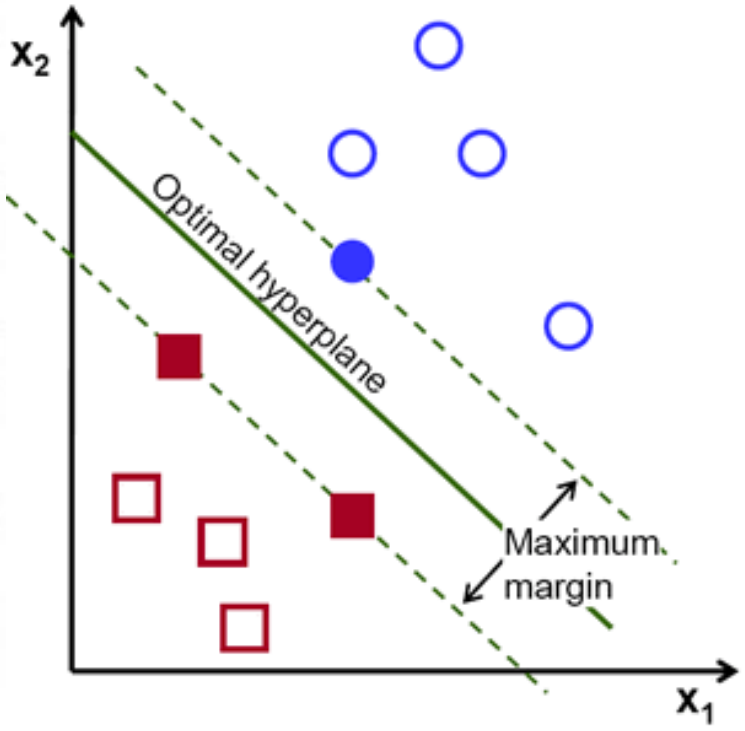
- 感知机标准
- 交叉熵标准(逻辑回归)
- 最小均方(LMS) 标准
- …

哪个线性超平面更好?

选择哪个学习标准?

最大间隔准则





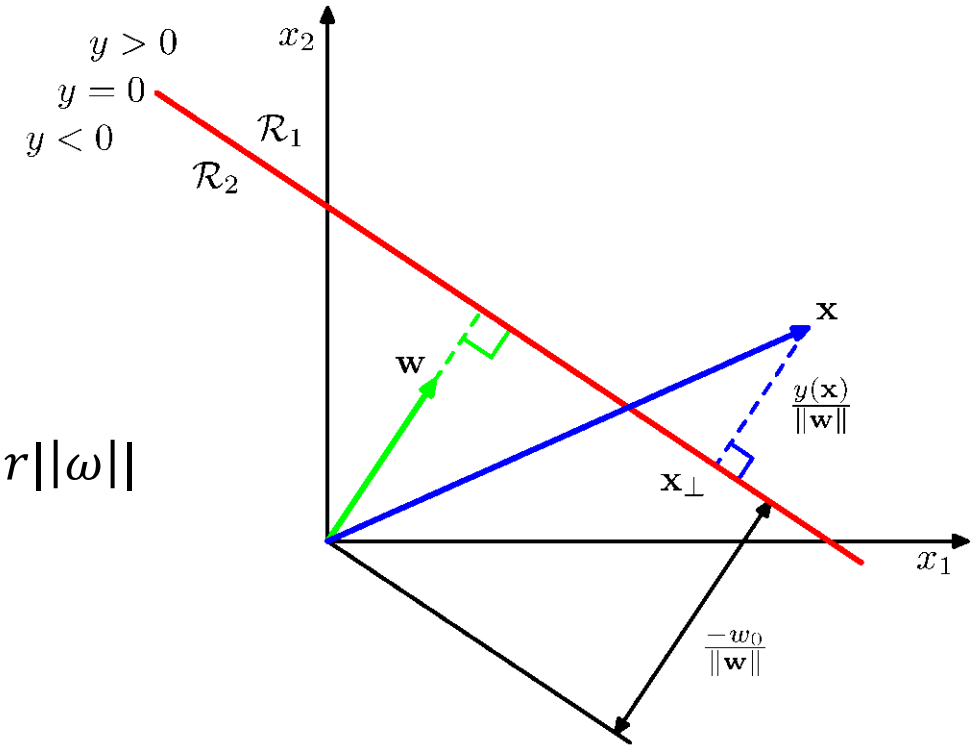
点到超平面的距离

- 线性模型 $y(x) = \omega^T x + b$
- 超平面 $\omega^T x + b = 0$
- 距离(正面)

$$x_+ = x_\perp + r\frac{\omega}{||\omega||}$$

$$b + \omega^T x_+ = b + \omega^T x_\perp + r\frac{\omega^T\omega}{||\omega||} = r||\omega||$$

$$r = \frac{\omega^T x_+ + b}{||\omega||}$$



几何距离 & 函数距离

- 距离 (负面)

$$x_{\perp} = x\_ + r\frac{\omega}{||\omega||} \quad \Rightarrow \quad r = -\frac{\omega^T x\_ + b}{||\omega||}$$

- 几何距离(统一表示)

$$r^{(i)} = y^{(i)}\left(\frac{\omega^T x^{(i)} + b}{||\omega||}\right)$$

$$where \;\; y^{(i)} = \{+1, -1\} denotes \;\; the \;\; class \;\; label \;\; of \;\; x^{(i)}$$

- 函数距离

$$\hat{r}^{(i)} = y^{(i)}\left(\omega^T x^{(i)} + b\right) = ||\omega|| r^{(i)}$$

参数缩放

- 通过比例因子缩放函数

$$\omega_s = c\omega \qquad\qquad b_s = cb$$

- 几何间隔：独立于比例因子

$$r_s^{(i)} = y^{(i)}\left(\frac{\omega_s^T x^{(i)} + b_s}{||\omega_s||}\right) = y^{(i)}\left(\frac{c\omega^T x^{(i)} + cb}{||c\omega||}\right) = y^{(i)}\left(\frac{\omega^T x^{(i)} + b}{||\omega||}\right) = r^{(i)}$$

- 函数间隔: 和比例因子成正比

$$\hat{r}_s^{(i)} = y^{(i)}\left(\omega_s^T x^{(i)} + b_s\right) = y^{(i)}\left(c\omega^T x^{(i)} + cb\right) = cy^{(i)}\left(\omega^T x^{(i)} + b\right) = c\hat{r}^{(i)}$$

最大间隔准则

- 公式 1

$$\max_{\omega,b} \quad \gamma$$

$$s.t. \quad y^{(i)}\left(\frac{\omega^T x^{(i)} + b}{||\omega||}\right) \geq \gamma, i = 1, \dots, m$$

$$I.e., \quad \gamma = \min_{i=1,\dots,m} \gamma^{(i)}$$

最大间隔准则

- 公式 2

$$\max_{\hat{\gamma},\omega,b} \quad \frac{\hat{\gamma}}{||\omega||}$$

$$s.t. \quad y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq \hat{\gamma}, i = 1, \ldots, m$$

$$I.e. \quad \hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}$$

最大间隔准则

- 比例限制

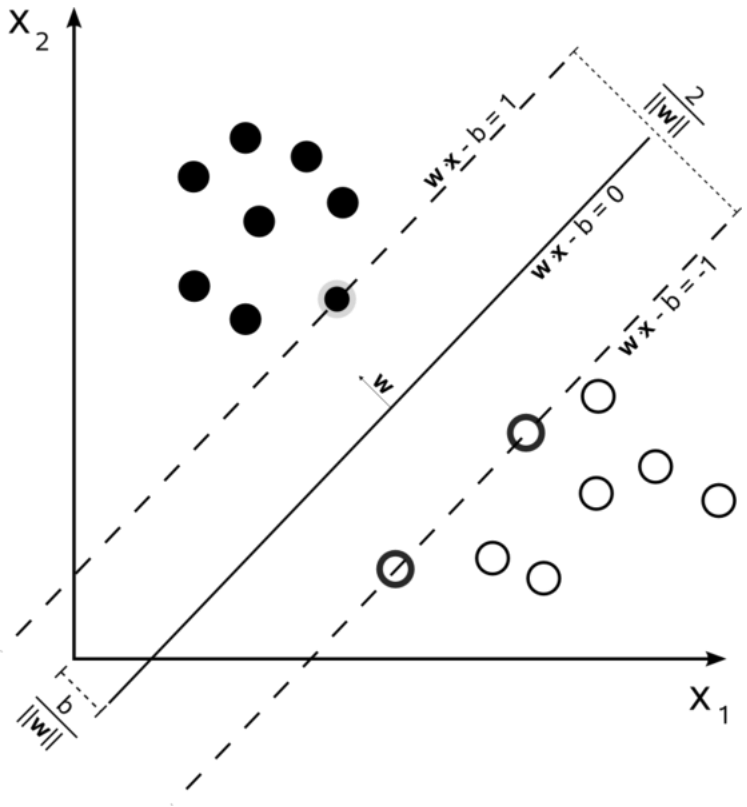
$$\hat{\gamma} = \min_{i=1,\dots,m} \hat{\gamma}^{(i)} = 1$$

$$\gamma = \min_{i=1,\dots,m} \gamma^{(i)} = \frac{1}{||\omega||}$$

$$I.e., scaling\ \omega\ and\ b, let\ ||\omega|| = \frac{1}{r}$$

- 在这个限制下

$$y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq \hat{\gamma} = 1$$



最大间隔准则

- 公式 3

$$\max_{\omega,b} \quad \frac{1}{||\omega||}$$

$$s.t. \;\; y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq 1, i = 1, \ldots, m$$

$$\min_{\omega,b} \quad \frac{1}{2}||\omega||^2$$

$$s.t. \;\; y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq 1, i = 1, \ldots, m$$

对偶优化

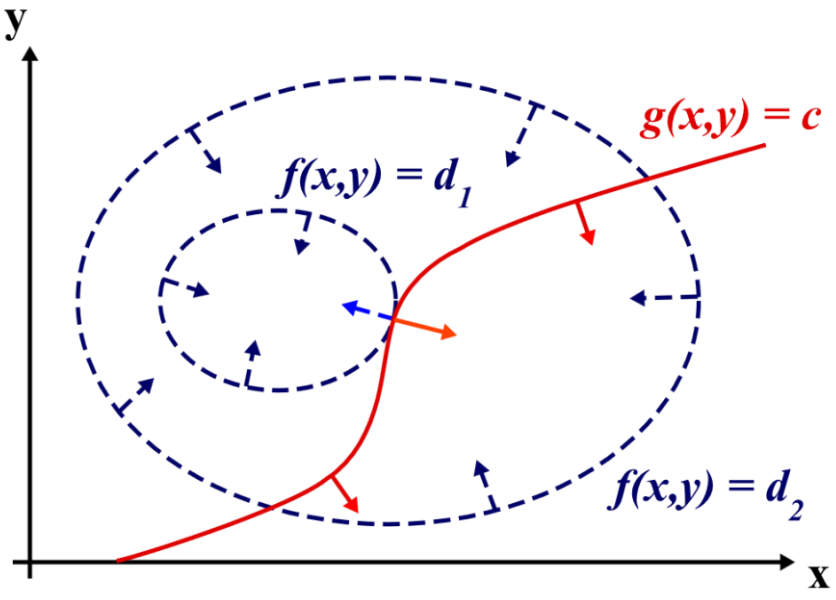
拉格朗日函数

- 在相同约束条件下

$$max \quad f(x)$$
$$s.t.\ g(x) = 0$$

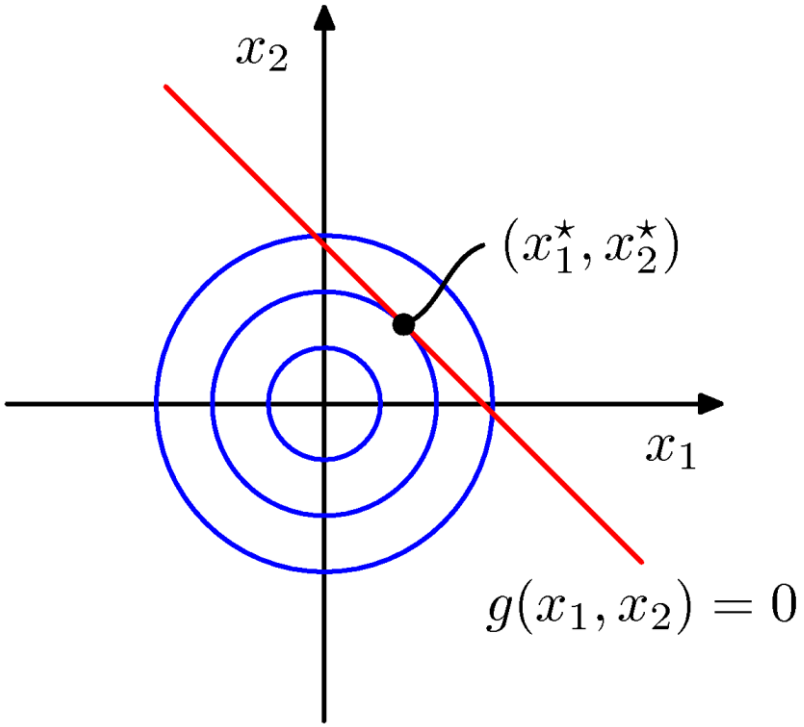
$$L(x, \lambda) = f(x) + \lambda\, g(x)$$

$$\begin{cases} \dfrac{\partial L}{\partial x} = 0 \\ g(x) = 0 \end{cases}$$



$$\nabla f + \lambda \nabla g = 0$$

例子



$$\max \ f(x_1, x_2) = 1 - x_1^2 - x_2^2$$
$$s.t. \ g(x_1, x_2) = x_1 + x_2 - 1 = 0$$

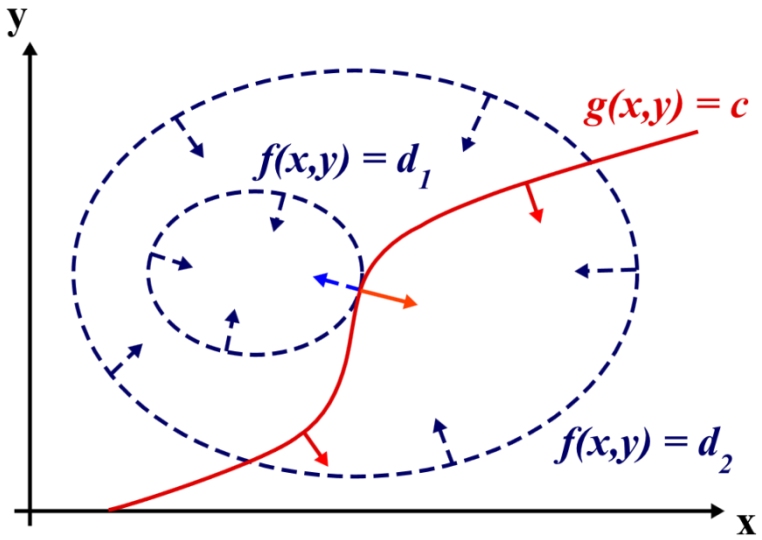
$$L(x_1, x_2, \lambda) = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$$

$$\begin{cases} \dfrac{\partial L}{\partial x_1} = -2x_1 + \lambda = 0 \\ \dfrac{\partial L}{\partial x_2} = -2x_2 + \lambda = 0 \\ g(x_1, x_2) = x_1 + x_2 - 1 \end{cases}$$

拉格朗日函数

- 在不等式约束下

$$max \quad f(x)$$
$$s.t.\ g(x) \geq 0$$



$$L(x,\lambda) = f(x) + \lambda g(x)$$

激活

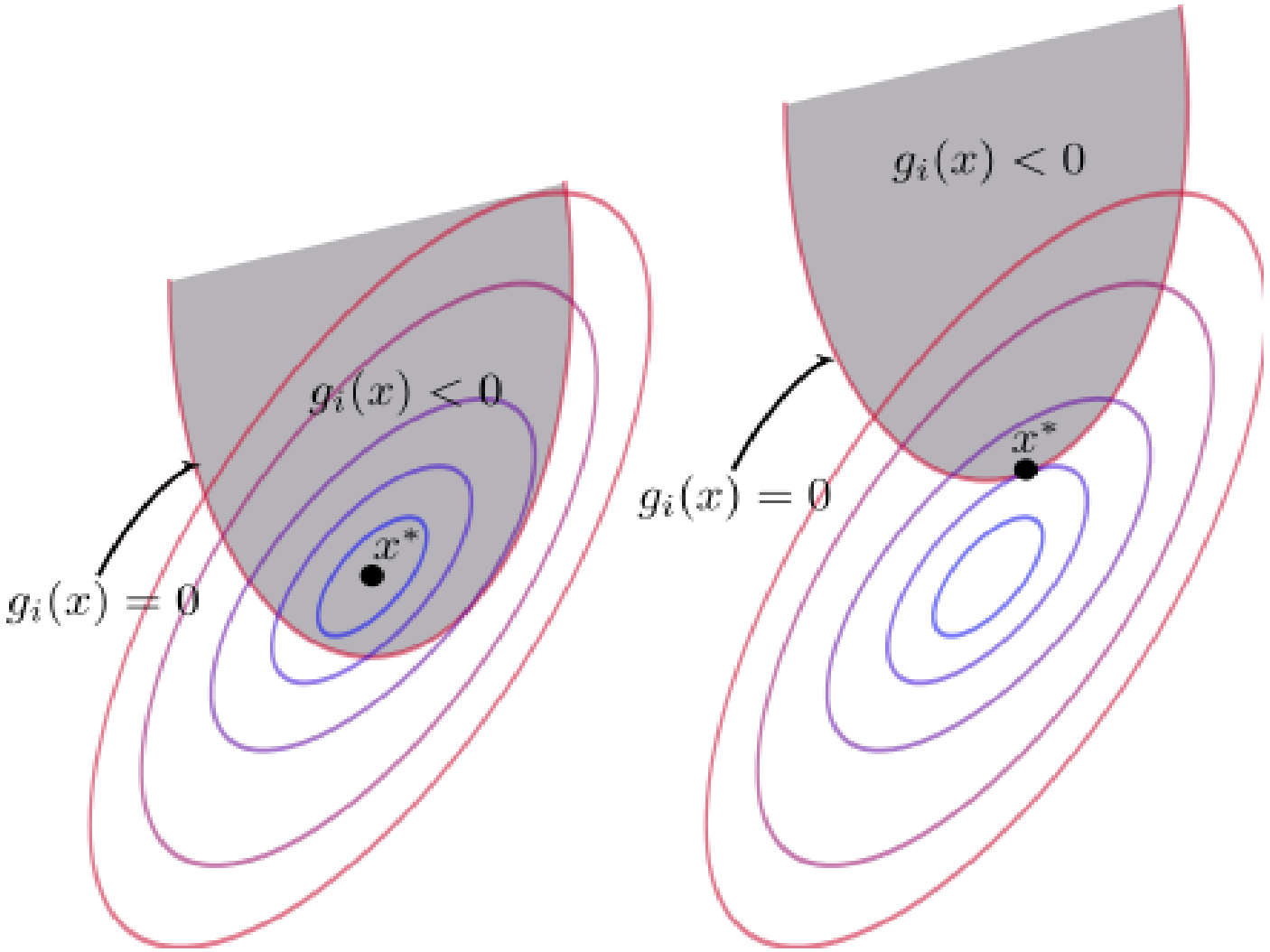
$$\begin{cases} \dfrac{\partial L}{\partial x} = 0 \\ g(x) = 0 \\ \lambda > 0 \end{cases}$$

非激活

$$\begin{cases} \dfrac{\partial L}{\partial x} = 0 \\ g(x) > 0 \\ \lambda = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \dfrac{\partial L}{\partial x} = 0 \\ g(x) \geq 0 \\ \lambda \geq 0 \\ \lambda g(x) = 0 \end{cases}$$

图解



拉格朗日乘数

- 在多重等式和不等式约束情况下

$$max \quad f(x)$$
$$s.t. \begin{cases} h_i(x) = 0 \\ g_i(x) \geq 0 \end{cases} \Rightarrow L(x, \lambda) = f(x) + \sum_i \lambda_i h_i(x) + \sum_j \mu_j g(x)$$

$$\begin{cases} \dfrac{\partial L}{\partial x} = 0 & \text{固定值} \\ h_i(x) = 0 & \text{原始可行性} \\ g_j(x) \geq 0 & \\ \mu_j(x) \geq 0 & \text{对偶可行性} \\ \mu_j g_j(x) = 0 & \text{互补条件} \end{cases}$$

Karush–Kuhn–Tucker (KKT) Conditions

广义拉格朗日和二元性

- 原始优化问题

$$\min_{\omega} f(\omega)$$
$$s.t.\ \ g_i(\omega) \leq 0, i = 1, \dots, k$$
$$h_j(\omega) = 0, j = 1, \dots, l$$

- 广义拉格朗日

$$L(\omega, \alpha, \beta) = f(\omega) + \sum_{i=1}^{k} \alpha_i g_i(\omega) + \sum_{j=1}^{l} \beta_j h_j(\omega)$$

拉格朗日最小最大值

$$\max_{\alpha,\beta:\alpha_i\geq 0} L(\omega,\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0}\left(f(\omega)+\sum_{i=1}^{k}\alpha_i g_i(\omega)+\sum_{j=1}^{l}\beta_j h_j(\omega)\right)$$

$$= \begin{cases} f(\omega) & if\ g_i(\omega)\leq 0, h_j(\omega)=0 \\ \infty & otherwise \end{cases}$$

$$\min_{\omega}\max_{\alpha,\beta;\alpha_i\geq 0} L(\omega,\alpha,\beta) = \min_{\omega}\max_{\alpha,\beta;\alpha_i\geq 0} L(\omega,\alpha,\beta) = \min_{\omega} f(\omega)$$

$$s.t.\ \ g_i(\omega)\leq 0 \qquad s.t.\ \ g_i(\omega)\leq 0$$
$$h_j(\omega)=0 \qquad h_j(\omega)=0$$

NUSTM

原问题 & 对偶问题

- 原问题(拉格朗日最小最大值)

$$\min_{\omega} \max_{\alpha,\beta;\alpha_i \geq 0} L(\omega, \alpha, \beta)$$

- 对偶问题(拉格朗日最小最大值)

$$\max_{\alpha,\beta;\alpha_i \geq 0} \min_{\omega} L(\omega, \alpha, \beta)$$

- 最大-最小 vs. 最小-最大

$$\max_{\alpha,\beta;\alpha_i \geq 0} \min_{\omega} L(\omega, \alpha, \beta) \leq \min_{\omega} \max_{\alpha,\beta;\alpha_i \geq 0} L(\omega, \alpha, \beta)$$

何时取等号？

两个问题的等价

- 取等号当：
  - f 和g_i's 是凸函数, h_i's 是affine;
  - g_i (严格) 可行: 这意味着存在一些w 使得 $g_i(w)<0$.

- 原和对偶问题的等同性

$$\min_{\omega} \quad f(\omega)$$

$$s.t. \quad \begin{matrix} g_i(\omega) \le 0 \\ h_i(\omega) = 0 \end{matrix} \quad = \quad \min_{\omega} \max_{\alpha,\beta;\alpha_i \ge 0} L(\omega,\alpha,\beta) \quad = \quad \max_{\alpha,\beta;\alpha_i \ge 0} \min_{\omega} L(\omega,\alpha,\beta)$$

原问题

对偶问题

Karush-Kuhn-Kucker (KKT) Conditions

- 此外，原问题和对偶问题的解决方案满足KTT条件：

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,n && \text{固定值}\\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,l && \text{原始可行性}\\
\alpha_i^* g_i(w^*) &= 0, \quad i=1,\ldots,k && \text{互补条件}\\
g_i(w^*) &\le 0, \quad i=1,\ldots,k && \text{原始可行性}\\
\alpha^* &\ge 0, \quad i=1,\ldots,k && \text{对偶可行性}
\end{aligned}
$$

足够和必要的条件

SVM的拉格朗日函数

- SVM的优化问题

$$\min_{\omega,b} \quad \frac{1}{2}||\omega||^2$$

$$s.t. \quad y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq 1, i = 1, \dots, m$$

$$g_i(\omega) = -\ y^{(i)}\left(\omega^T x^{(i)} + b\right) + 1 \leq 0, i = 1, \dots, m$$

- 拉格朗日函数

$$L(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 - \sum_{i=1}^{m} \alpha_i \left(y^{(i)}\left(\omega^T x^{(i)} + b\right) - 1\right)$$

拉格朗日最小化

- 以拉格朗日的导数为例

$$\frac{\partial L(\omega, b, \alpha)}{\partial \omega} = \omega - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \Rightarrow \omega = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial L(\omega, b, \alpha)}{\partial b} = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- 插回拉格朗日

$$L(\omega, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left(x^{(i)}\right)^T x^{(j)} - b \sum_{i=1}^{m} \alpha_i y^{(i)}$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left(x^{(i)}\right)^T x^{(j)}$$

SVM的对偶问题

- 对偶问题

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$s.t. \quad \alpha_i \geq 0, i = 1, \ldots, m$$

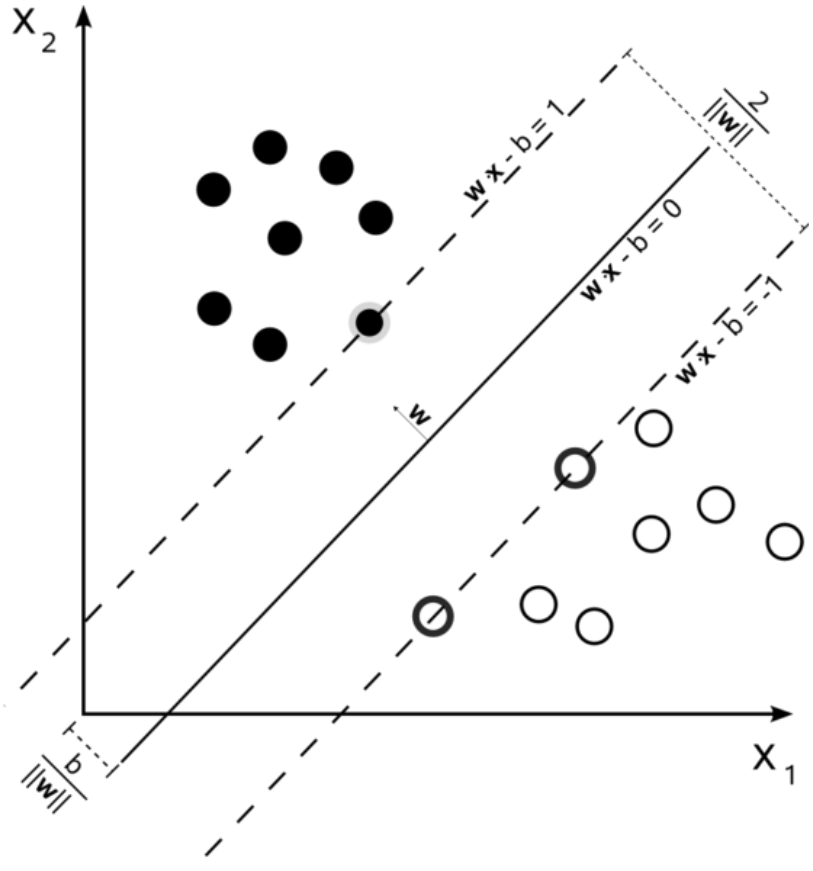
$$\sum_{i=1}^{m} \alpha_i y^{(j)} = 0$$

保证KKT条件得到满足。

为什么 “Support Vector”?

- 决策函数

$$f(x) = (\omega^*)^T x + b$$
$$= \left(\sum_{i=1}^{N} \alpha_i^* y^{(i)} x^{(i)} \right)^T x + b^*$$
$$= \sum_{i=1}^{N} \alpha_i^* y^{(i)} \left(x^{(i)} \right)^T x + b^*$$



- KKT 条件

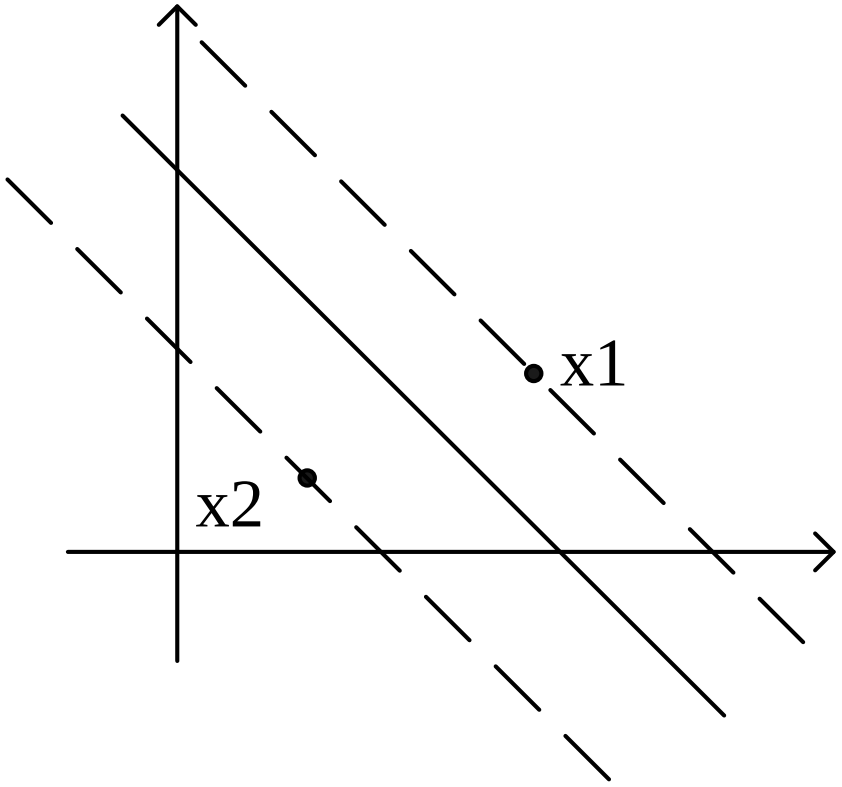
$$\begin{cases} \alpha_i^* g_i(\omega^*) & = 0 \\ g_i(\omega^*) & \leq 0 \\ \alpha_i^* & \geq 0 \end{cases}$$

$$where \;\; g_i(\omega) = -y^{(i)} \left(\omega^T x^{(i)} + b \right) + 1$$

偏差值

x_1 is a positive support vector

is a negative support vector



$$\begin{cases} (\omega^*)^T x_1 + b = 1 \\ (\omega^*)^T x_2 + b = -1 \end{cases}$$

$$b^* = -\frac{(\omega^*)^T x_1 + (\omega^*)^T x_2}{2}$$

$$= -\frac{max_{i:y^{(i)}=-1}(\omega^*)^T x_i + min_{i:y^{(i)}=1}(\omega^*)^T x_i}{2}$$

一个遗留的问题

- 决策函数

怎样计算 alpha?

$$f(x) = (\omega^*)^T x + b = \sum_{i=1}^{m} \alpha_i^* y^{(i)} \left(x^{(i)}\right)^T x + b^*$$

$$where\ \ b^* = -\frac{max_{i:y^{(i)}=-1}(\omega^*)^T x_i + min_{i:y^{(i)}=1}(\omega^*)^T x_i}{2}$$

- SVM的对偶问题

$$\max_{\alpha}\ \ W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$s.t.\ \ \alpha_i \geq 0, i = 1, \ldots, m$$

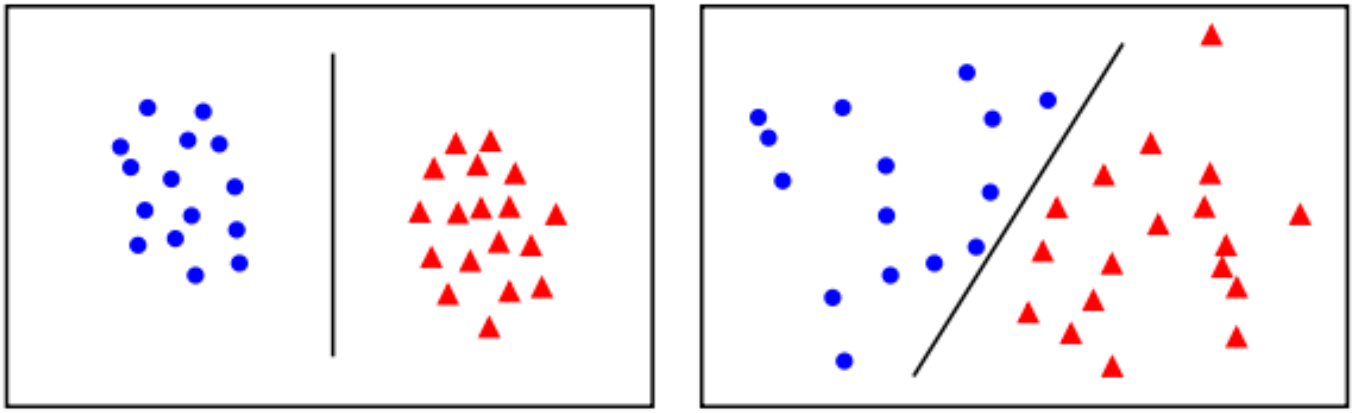
$$\sum_{i=1}^{m} \alpha_i y^{(j)} = 0$$

如何解决对偶优化问题?

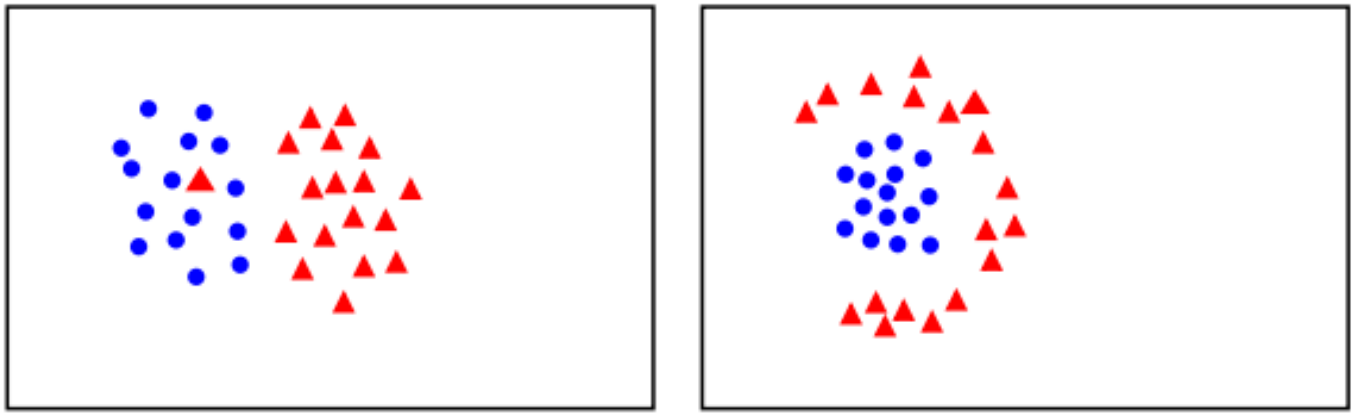
Soft-margin SVM

线性不可分离的情况

线性可分离



线性不可分离



Soft Margin 准则

$$\min_{\omega,b} \quad \frac{1}{2}||\omega||^2$$

$$s.t. \;\; y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq 1, i = 1, \dots, m$$

最大间隔

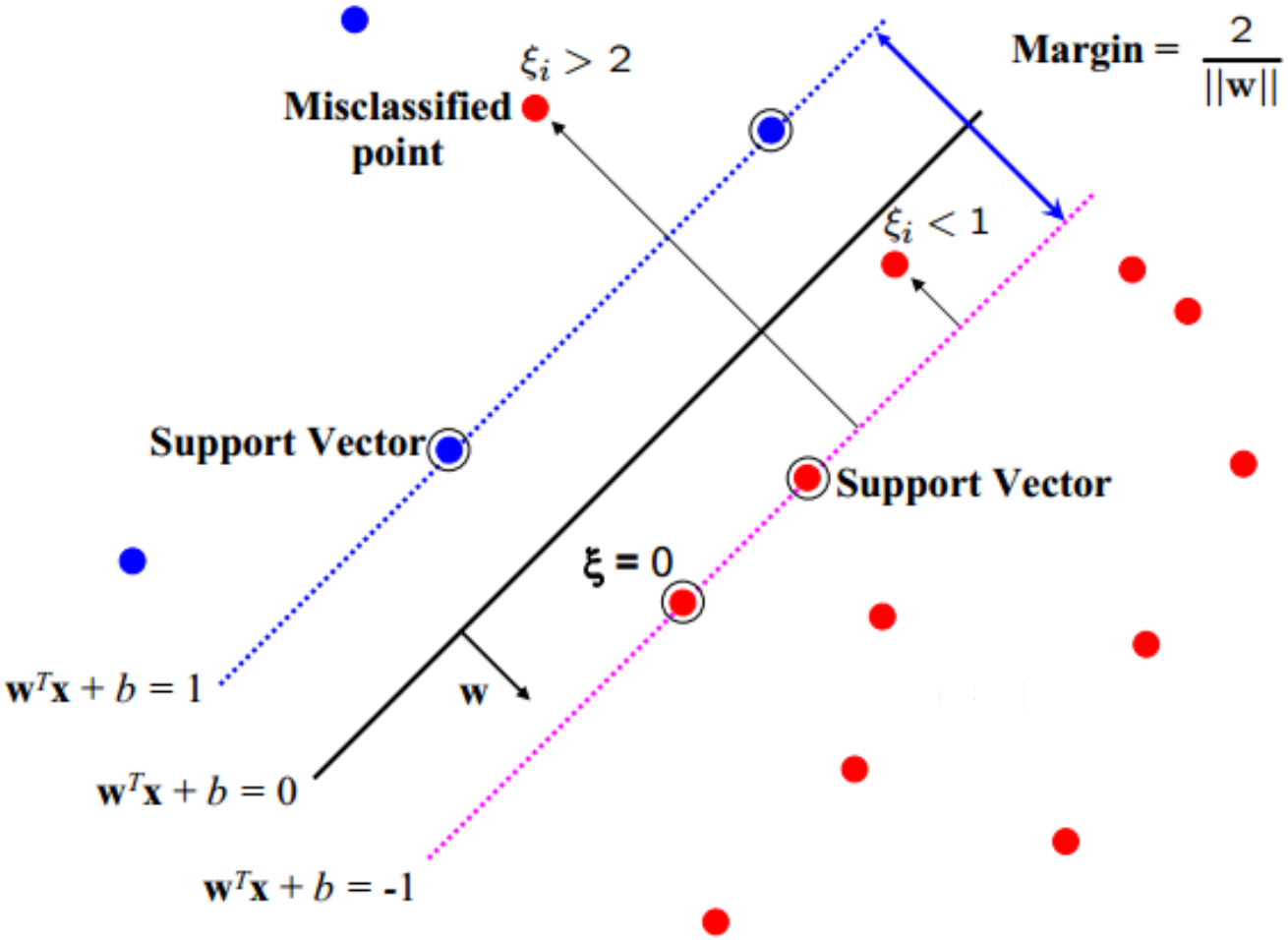
$$\min_{\omega,b,\varepsilon} \quad \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{m} \varepsilon_i$$

$$s.t. \;\; y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq 1 - \varepsilon_i, i = 1, \dots, m$$

$$\varepsilon_i \geq 0, i = 1, \dots, m$$

Soft margin

Three Types of Slacks



Lagrangian for Soft-margin SVM

- 回忆原问题和对偶问题的等价性

$$\min_{\omega} f(\omega) \quad s.t.\ g_i(\omega) \le 0,\ h_i(\omega) = 0 \quad = \quad \min_{\omega} \max_{\alpha,\beta;\alpha_i \ge 0} L(\omega,\alpha,\beta) \quad = \quad \max_{\alpha,\beta;\alpha_i \ge 0} \min_{\omega} L(\omega,\alpha,\beta)$$

原问题

对偶问题

- Lagrangian 形式

$$L(\omega, b, \varepsilon, \alpha, \gamma) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\varepsilon_i - \sum_{i=1}^{m}\alpha_i\left(y^{(i)}\left(\left(x^{(i)}\right)^T\omega + b\right) - 1 + \varepsilon_i\right) - \sum_{i=1}^{m}\gamma_i\varepsilon_i$$

软间隔 SVM 的对偶问题

- 梯度

$$\frac{\partial L}{\partial \omega} = 0 \Rightarrow \omega = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0 \Rightarrow C = \alpha_i + \gamma_i \Rightarrow \alpha_i \leq C$$

- 插回拉格朗日

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$s.t. \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

最大间隔SVM vs. 软间隔SVM

- 最大间隔 SVM

$$\max_{\alpha} \; W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$s.t. \quad \alpha_i \geq 0, i = 1, \dots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- 软间隔 SVM

$$\max_{\alpha} \; W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$s.t. \quad 0 \leq \alpha_i \leq \boxed{C}, i = 1, \dots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

KKT互补条件

- 两个KKT互补条件

$$\alpha_i\left(y^{(i)}\left(\omega^T x^{(i)}+b\right)-1+\varepsilon_i\right)=0$$

$$\gamma_i\varepsilon_i=(C-\alpha_i)\,\varepsilon_i=0 \qquad y^{(i)}\left(\omega^T x^{(i)}+b\right)\geq 1-\varepsilon_i$$

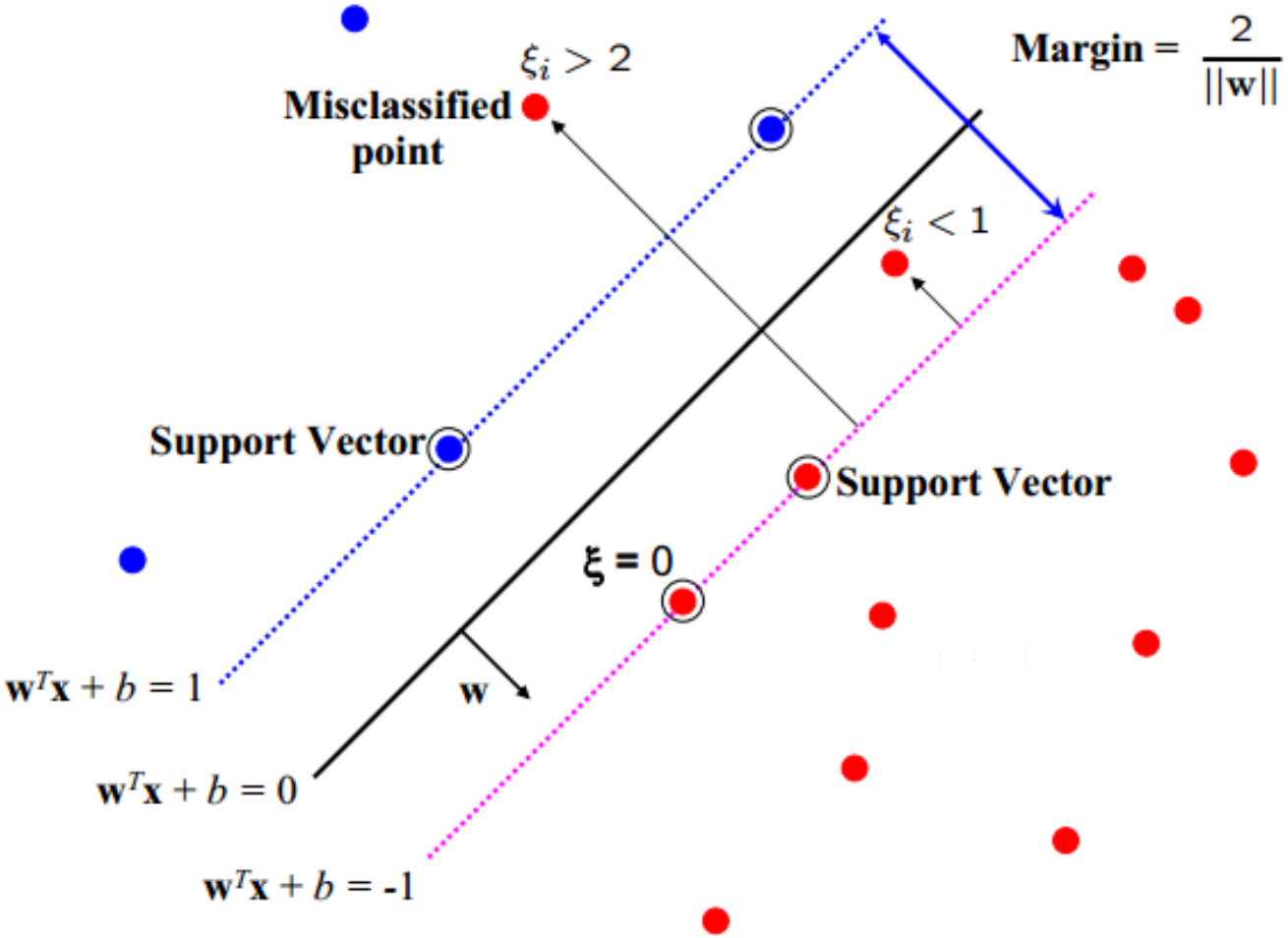
- 一些有用的结论

$$\alpha_i=0\Rightarrow\ \varepsilon_i=0\Rightarrow y^{(i)}\left(\omega^T x^{(i)}+b\right)\geq 1$$

$$\alpha_i=\mathrm{C}\Rightarrow y^{(i)}\left(\omega^T x^{(i)}+b\right)-1+\varepsilon_i\Rightarrow y^{(i)}\left(\omega^T x^{(i)}+b\right)\leq 1$$

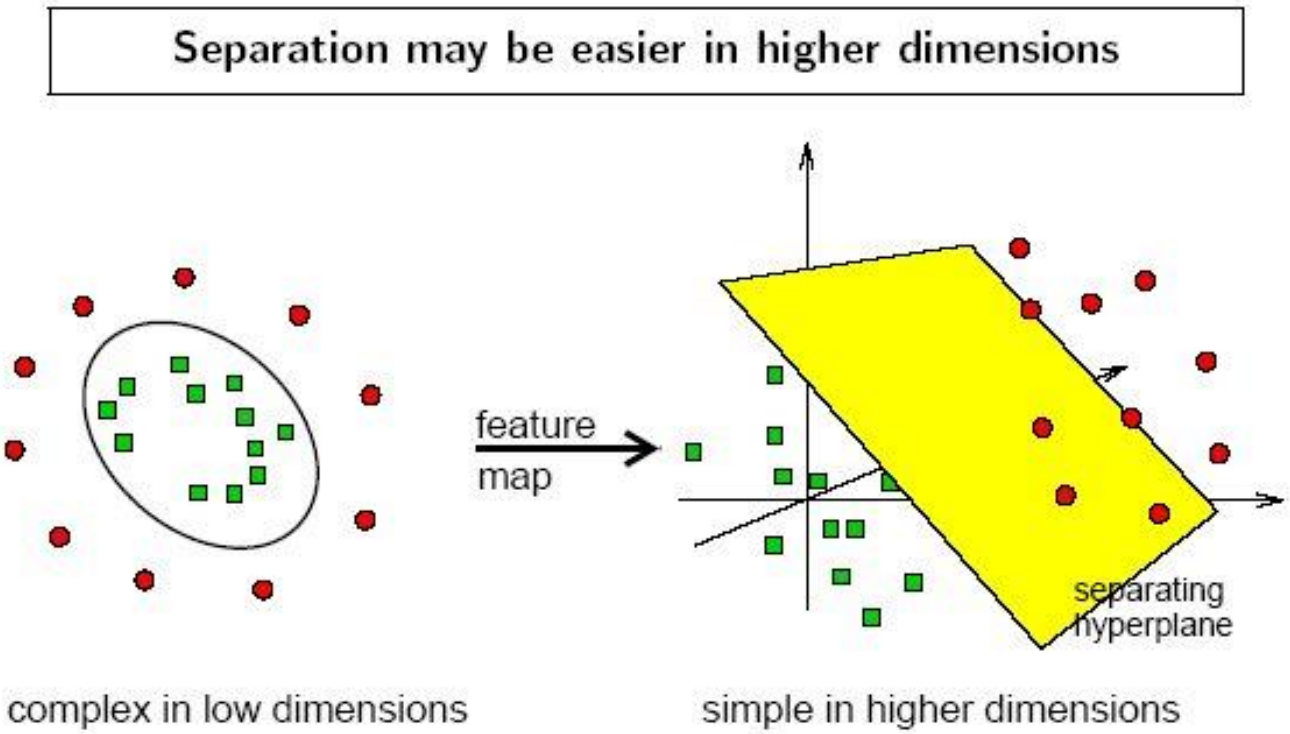
$$0<\alpha_i<\mathrm{C}\ \Rightarrow\ \begin{cases}\varepsilon_i=0\\ y^{(i)}\left(\omega^T x^{(i)}+b\right)-1+\varepsilon_i=0\end{cases}\Rightarrow y^{(i)}\left(\omega^T x^{(i)}+b\right)=1$$

Slacks and Support Vectors



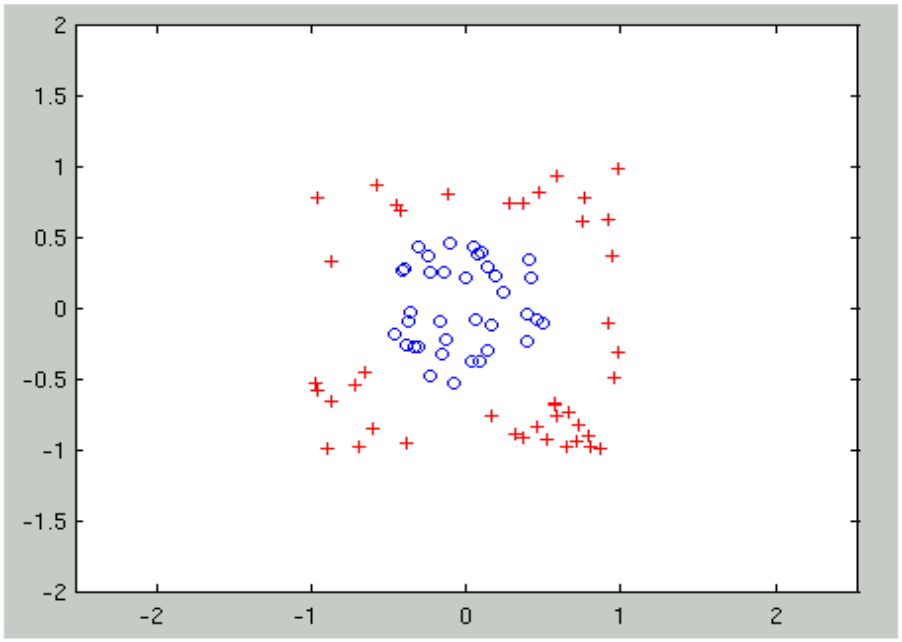
核函数

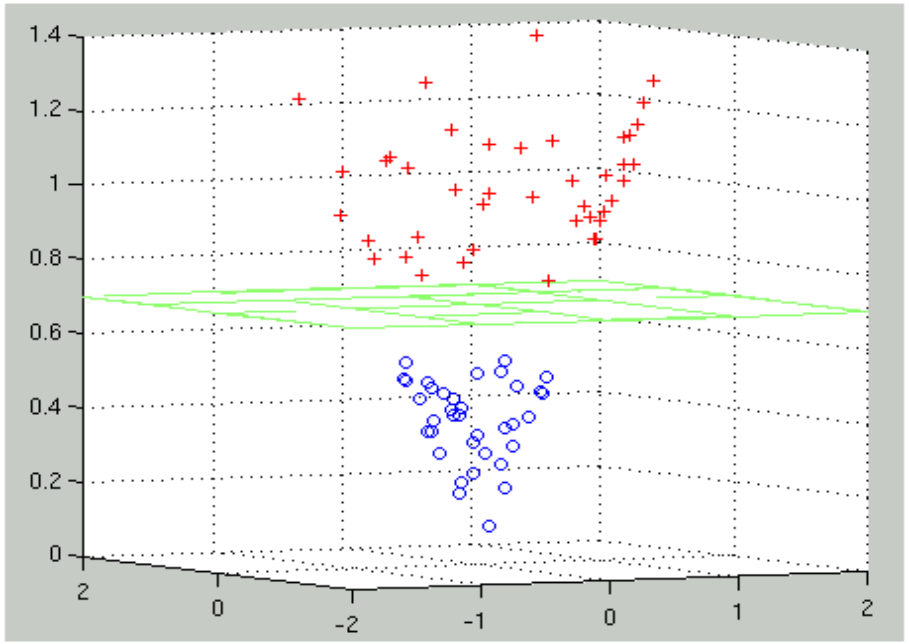
低维-不可分离的高维分离



从低维到更高维

- 特征空间映射：从低维不可分离到高维可分离





$$(x_1; x_2) \Rightarrow (x_1; x_2; \sqrt{x_1^2 + x_2^2})$$

核函数

- 定义: 高特征空间乘积

$$K(x,z) = \emptyset(x)^T \emptyset(z)$$

- 例子

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \Rightarrow \emptyset(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ \sqrt{2c}x_3 \end{bmatrix} \quad \Leftrightarrow \quad K(x,z) = (x^Tz + c)^2$$

高维特征空间的SVM

- 决策函数

$$f(x)=\sum_{i=1}^{m}\alpha_i y^{(i)}\langle x_i,x\rangle+b$$

$$f(x)=\sum_{i=1}^{m}\alpha_i y^{(i)}\langle \emptyset(x_i),\emptyset(x)\rangle+b$$
$$=\sum_{i=1}^{m}\alpha_i y^{(i)}K(x_i,x)+b$$

- 训练过程

$$\max_{\alpha}\ W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)},x^{(j)}\rangle$$

$$s.t.\ \ \alpha_i\geq 0,i=1,\dots,m$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

$$\max_{\alpha}\ W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j\langle \emptyset(x^{(i)}),\emptyset(x^{(j)})\rangle$$
$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j K\langle x^{(i)},x^{(j)}\rangle$$

$$s.t.\ \ \alpha_i\geq 0,i=1,\dots,m$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

SVM中的核函数

- SVM中的核函数
  - 有时很难知道准确的投影函数，但相对来说比较容易知道内核的功能
  - 在 SVM中, 特征向量的所有计算都是乘积的形式
  - 因此，我们只需要SVM中使用的核函数，而不需要知道精确的投影函数.

Mercer 条件

- 核矩阵
  - 对于任何有限集合点 $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$
  - 核矩阵元素 $K_{ij} = K\left(x^{(i)}, x^{(j)}\right)$
- 有效内核满足
  - 对称 $K_{ij} = K_{ji}$
  - 正半定 $\forall z \in R^n : z^T K z \geq 0$
- Mercer定理

Theorem (Mercer). Let $K : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ be given. Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{x^{(1)}, \ldots, x^{(m)}\}$, $(m < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

常用核函数

- 线性核函数

$$K(x_1, x_2) = x_1^T x_2$$

- 多项式核函数

$$K(x_1, x_2) = (x_1^T x_2 + c)^d$$

- 高斯核函数

$$K(x_1, x_2) = \exp(-\frac{|x_1 - x_2|^2}{2\delta^2})$$

- Sigmoid kernel, pyramid kernel, string kernel, tree kernel…

Kernel SVM

- 训练集

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)})$$

$$s.t. \quad \alpha_i \geq 0, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- 决策

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x_i, x) + b$$

Soft-margin Kernel SVM

- 训练

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)})$$

$$s.t. \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- 决策

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x_i, x) + b$$

序列最小优化算法

坐标上升法

- 考虑一个无约束的优化问题

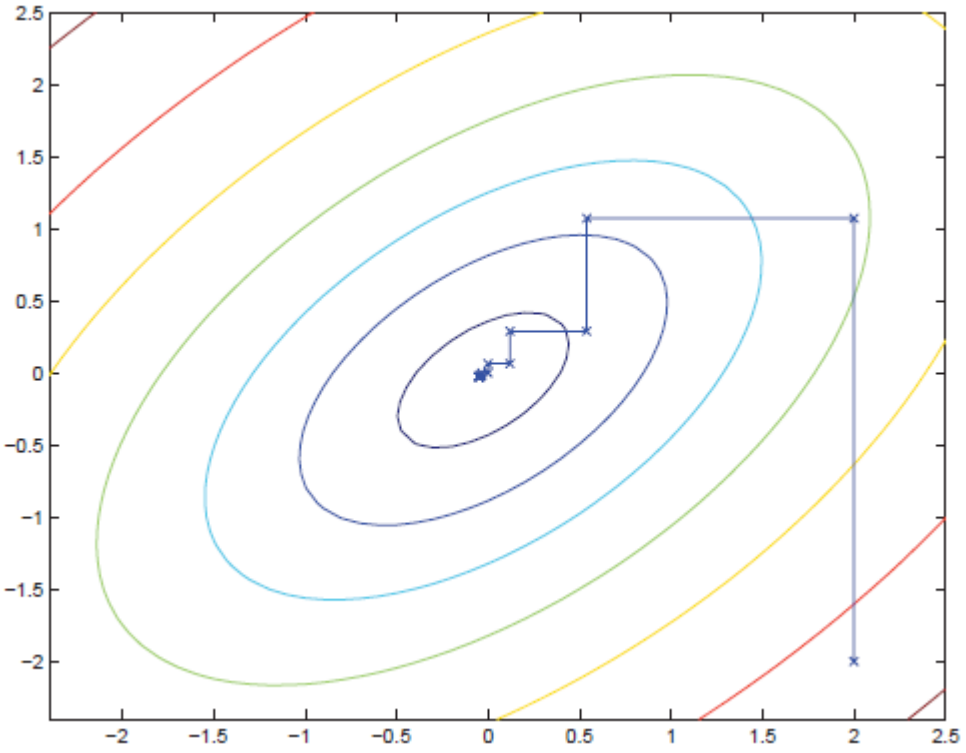
$$\max_x f(x_1, x_2, \dots, x_m)$$

- 坐标上升算法

$$\begin{aligned}
&Loop\ until\ convergence: \{ \\
&\quad For\ i = 1, \dots, m \{ \\
&\quad\quad x_i := arg \max_{\hat{x}_i} f(x_1, \dots, x_{i-1}, \hat{x}_i, x_{i+1}, \dots, x_m) \\
&\quad \} \\
&\}
\end{aligned}$$

坐标上升法

- 例子



$$f(x_1, x_2) = 6x_1x_2 - 5x_1^2 - 5x_2^2$$

回顾SVM的对偶问题

- 对偶优化问题

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)})$$

$$s.t. \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- KKT 条件

$$\alpha_i = 0 \Rightarrow y^{(i)}(\omega^T x^{(i)} + b) \geq 1$$

$$\alpha_i = C \Rightarrow y^{(i)}(\omega^T x^{(i)} + b) \leq 1$$

$$0 < \alpha_i < C \Rightarrow y^{(i)}\left(\omega^T x^{(i)} + b\right) = 1$$

SVM的坐标上升

- 每次选择两个坐标进行优化

$$\max_{\alpha} \quad W(\alpha_1, \alpha_2)$$

$$s.t. \quad 0 \le \alpha_i \le C, i = 1,2$$

$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^{m} \alpha_i y^{(i)} = \gamma$$

- 选择哪两个坐标?

SMO算法

$$W(\alpha_1,\alpha_2)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_jK_{ij}\alpha_i\alpha_j$$

$$=\alpha_1+\alpha_2+\sum_{i=3}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{2}(\sum_{j=1}^{2}y_iy_j\alpha_i\alpha_jK_{ij}+\sum_{j=3}^{m}y_iy_j\alpha_i\alpha_j\,K_{ij})$$

$$-\frac{1}{2}\sum_{i=3}^{m}(\sum_{j=1}^{2}y_iy_jK_{ij}\alpha_i\alpha_j+\sum_{j=3}^{m}y_iy_j\alpha_i\alpha_j\,K_{ij})$$

$$=\alpha_1+\alpha_2+\sum_{i=3}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}y_iy_jK_{ij}\alpha_i\alpha_j-\sum_{i=1}^{2}\sum_{j=3}^{m}y_iy_j\alpha_i\alpha_j\,K_{ij}-\frac{1}{2}\sum_{i=3}^{m}\sum_{j=3}^{m}y_iy_j\alpha_i\alpha_j\,K_{ij}$$

$$=\alpha_1+\alpha_2-\frac{1}{2}K_{11}\alpha_1^2-\frac{1}{2}K_{22}\alpha_2^2-y_1y_2K_{12}\alpha_1\alpha_2-y_1\alpha_1\sum_{j=3}^{m}y_j\alpha_jK_{1j}-y_2\alpha_2\sum_{j=3}^{m}y_j\alpha_jK_{2j}$$

$$+\sum_{i=3}^{m}\alpha_i-\frac{1}{2}\sum_{i=3}^{m}\sum_{j=3}^{m}y_iy_j\alpha_i\alpha_j\,K_{ij}$$

$$=\alpha_1+\alpha_2-\frac{1}{2}K_{11}\alpha_1^2-\frac{1}{2}K_{22}\alpha_2^2-y_1y_2K_{12}\alpha_1\alpha_2-y_1\alpha_1v_1-y_2\alpha_2v_2+constant$$

where

$$K_{ij}=K(x_i,x_j)\qquad v_i=\sum_{j=3}^{m}y_j\alpha_jK_{ij}$$

SMO 算法

- 使用平等约束的变量消除

$$\alpha_1 y^{(1)} = -\sum_{i=2}^{m} \alpha_i y^{(i)} \Rightarrow \alpha_1 = -y^{(1)} \sum_{i=3}^{m} \alpha_i y^{(i)} = \gamma - y^{(1)} y^{(2)} \alpha_2$$

$$W(\alpha_2) = \gamma - s\alpha_2 + \alpha_2 - \frac{1}{2} K_{11}(\gamma - s\alpha_2)^2 - \frac{1}{2} K_{22} {\alpha_2}^2 - sK_{12}(\gamma - s\alpha_2)\alpha_2 - y_1(\gamma - s\alpha_2)v_1 - y_2\alpha_2 v_2 + constant$$

- 通过梯度等于零进行优化

$$\frac{\partial W(\alpha_2)}{\partial \alpha_2} = -s + 1 + sK_{11}\gamma - K_{11}\alpha_2 - K_{22}\alpha_2 - s\gamma K_{12} + 2K_{12}\alpha_2 + y_2 v_1 - y_2 v_2 = 0$$

$$\alpha_2^{new} = \frac{y_2(y_2 - y_1 + y_1\gamma(K_{11} - K_{12}) + v_1 - v_2)}{K_{11} + K_{22} - 2K_{12}}$$

SMO 更新

- 利用

$$v_i = \sum_{j=3}^{m} y_j \alpha_j K_{ij} \qquad \gamma = \alpha_1^{old} + s\alpha_2^{old}$$

- 最终由有

$$\alpha_2^{new} = \frac{y_2(y_2 - y_1 + y_1\gamma(K_{11} - K_{12}) + v_1 - v_2)}{K_{11} + K_{22} - 2K_{12}}$$

$$\alpha_2^{new} = \alpha_2^{old} + \frac{y_2(E_1 - E_2)}{\eta}$$

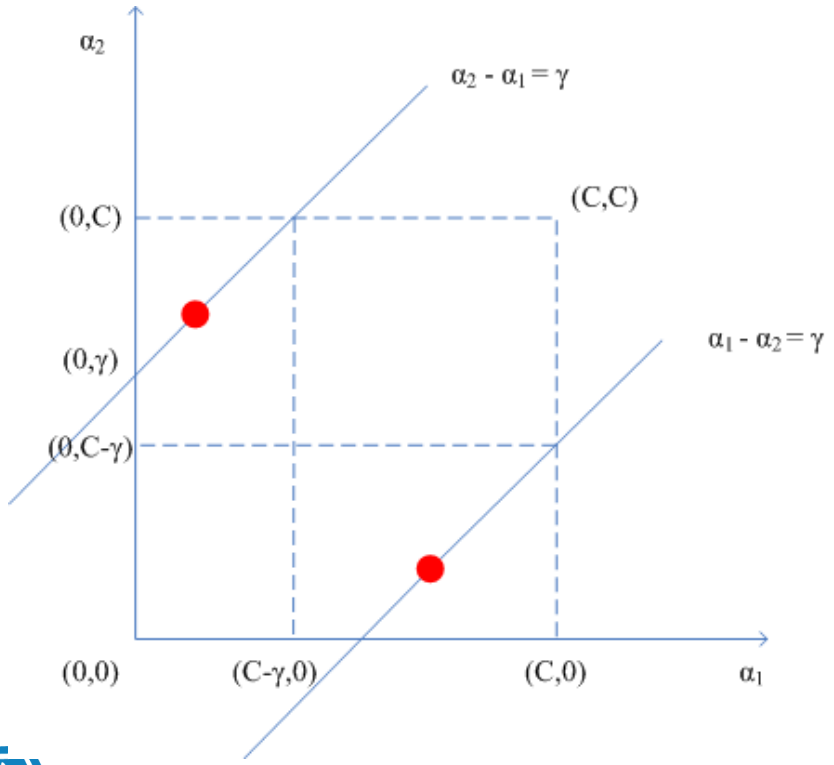
$$E_i = f(x_i) - y_i \qquad \eta = K_{11} + K_{22} - 2K_{12} = ||\emptyset(x_1) - \emptyset(x_2)||^2$$

增加不平等约束

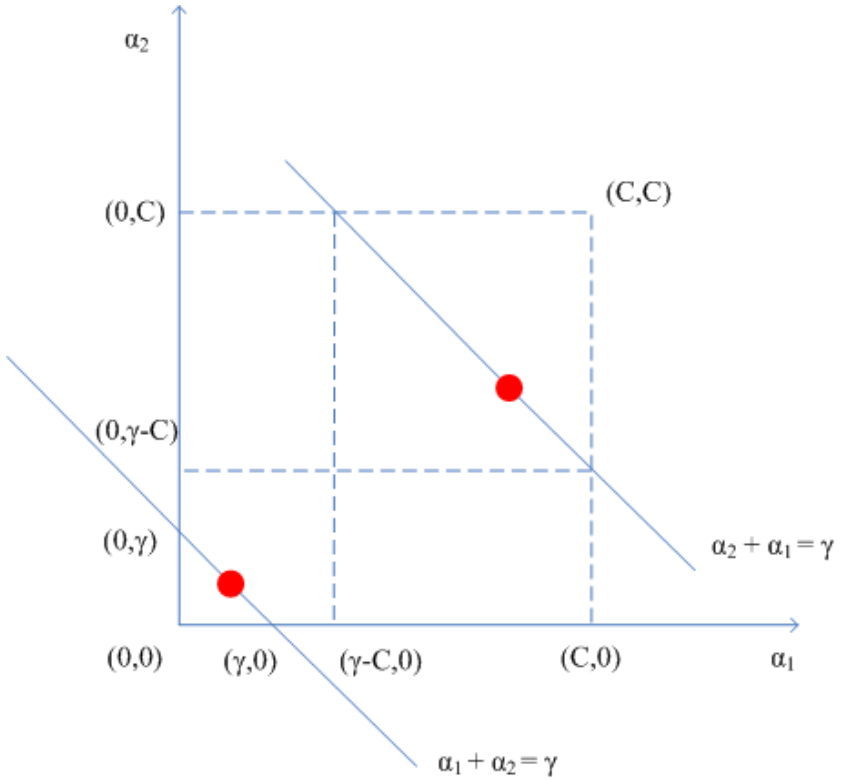
- 平等约束
- 不平等约束

$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^{m} \alpha_i y^{(i)} = \gamma$$

$$0 \le \alpha_i \le C, i = 1,2$$



$y_1 y_2 = -1$



$y_1 y_2 = 1$

两个乘数的最终更新

- 如果 $y_1 y_2 = -1$

$$\begin{cases} L = \max\{0, \alpha_2^{old} - \alpha_1^{old}\} \\ H = \min\{C, C + \alpha_2^{old} - \alpha_1^{old}\} \end{cases}$$

- 如果 $y_1 y_2 = 1$

$$\begin{cases} L = \max\{0, \alpha_1^{old} + \alpha_2^{old} - C\} \\ H = \min\{C, \alpha_1^{old} + \alpha_2^{old}\} \end{cases}$$

- 最终更新

$$\alpha_2^{new,clipped} = \begin{cases} L & \alpha_2^{new} \le L \\ \alpha_2^{new} & L < \alpha_2^{new} < H \\ H & \alpha_2^{new} \ge H \end{cases}$$

$$\alpha_1^{new} = \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new})$$

启发式选择两个乘数

- 首先选择一个违反KKT条件的拉格朗日乘数（Osuna理论）

$$\alpha_i = 0 \ \ and \ \ y^{(i)}\left(\omega^T x^{(i)} + b\right) < 1$$

$$0 \le \alpha_i \le C \ \ and \ \ y^{(i)}(\omega^T x^{(i)} + b) \ne 1$$

$$\alpha_i = C \ \ and \ \ y^{(i)}\left(\omega^T x^{(i)} + b\right) > 1$$

- 其次选择最大化的拉格朗日乘数|E1-E2|

$$E_i = f(x_i) - y_i$$

$$|E_1 - E_2| = |f(x_1) - y_1 - (f(x_2) - y_2)|$$

Bias的更新

- 选择b使KKT条件成立（当alpha不在范围内时）

$$b_1 = E_1 + y_1(\alpha_1^{new} - \alpha_1^{old})K_{11} + y_2(\alpha_2^{new,clipped} - \alpha_2^{old})K_{12} + b^{old}$$

$$b_2 = E_2 + y_1(\alpha_1^{new} - \alpha_1^{old})K_{12} + y_2(\alpha_2^{new,clipped} - \alpha_2^{old})K_{22} + b^{old}$$

- 更新b

If α_1^{new} is not at the bounds (0 or C), $b^{new} = b_1$

If $\alpha_2^{new,clipped}$ is not at the bounds, $b^{new} = b_2$

If α_1^{new} and $\alpha_2^{new,clipped}$ are both not at the bounds, $b^{new} = b_1 = b_2$

If α_1^{new} and $\alpha_2^{new,clipped}$ are both at the bounds, $b^{new} = \frac{b_1 + b_2}{2}$

收敛条件

- 在线性内核的情况下更新权重

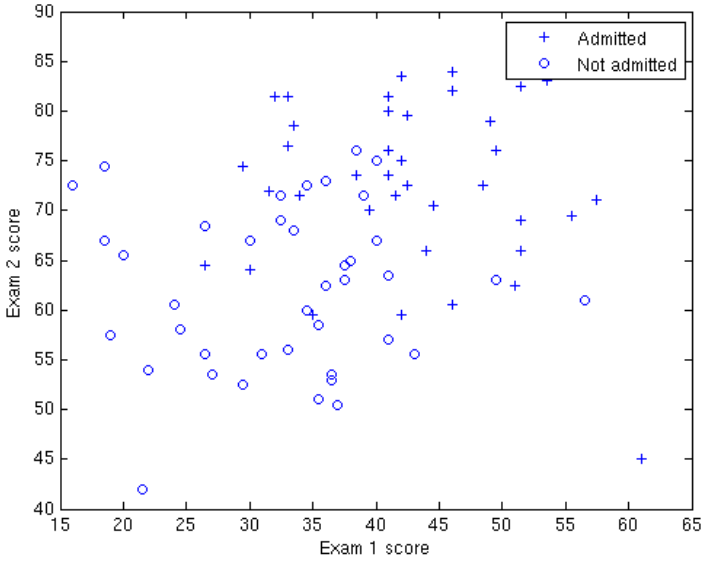
$$\frac{\partial L(\omega, b, \alpha)}{\partial \omega} = \omega - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \Rightarrow \omega = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\omega^{new} = \omega^{old} + y_1(\alpha_1^{new} - \alpha_1^{old})x_1 + y_2\,(\alpha_2^{new} - \alpha_2^{old})x_2$$

- 当所有拉格朗日乘数满足KKT条件（在用户定义的公差内）时，问题已经解决。

练习：Logistic Regression vs. ANN vs. SVM

- 给出训练数据：



http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=DeepLearning&doc=exercises/ex4/ex4.html

- 编码实现Logistic回归模型（基于SGD），并对结果进行5倍交叉验证；
- 使用 Tensorflow实现三层前向神经网络 （ANN），并对结果进行5倍交叉验证；
- 利用LibSVM实现软间隔SVM解决上述问题，调节参数C和核函数，并将结果与Logistic回归、三层前馈ANN进行比较。

NUSTM

Questions?