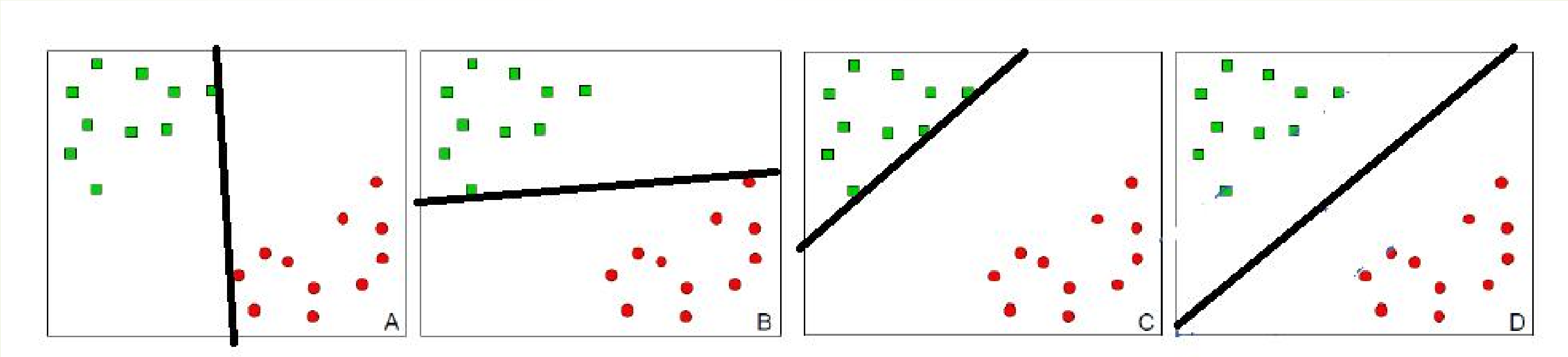
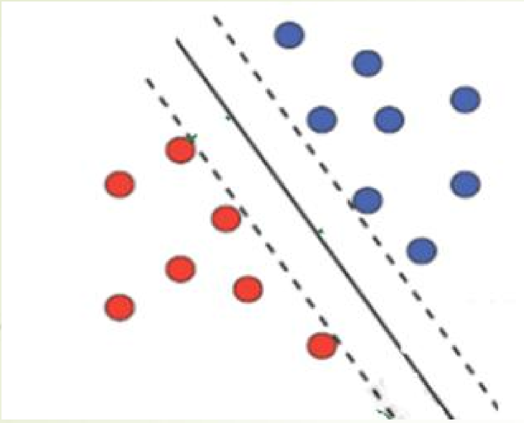
ماشین بردار پشتیبان (Support Vector Machine)

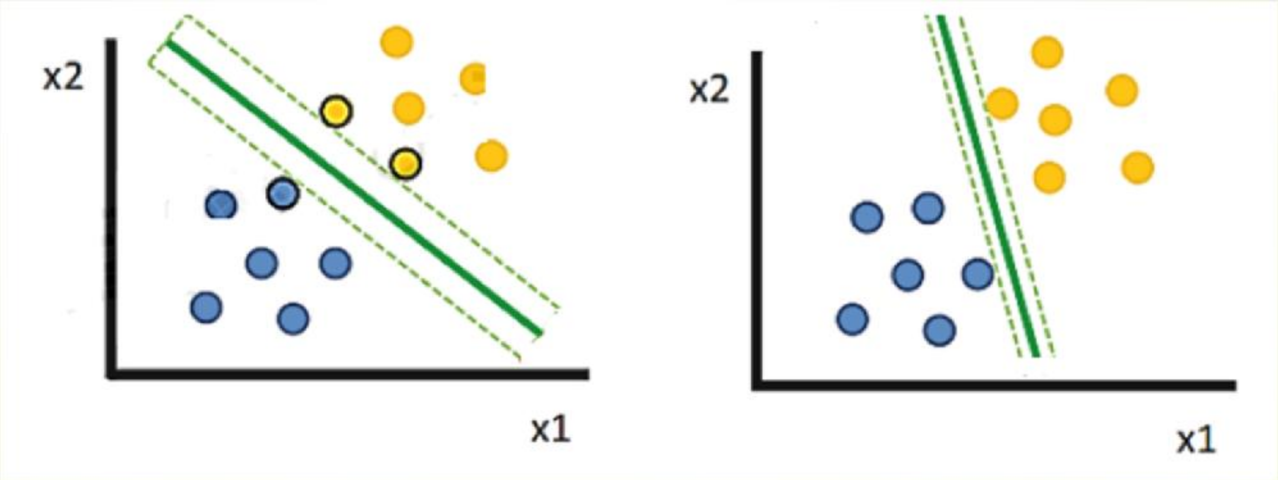
- ❖ یادگیری با سرپرست (SVM)
- ❖ در برخی از مسائل غیرخطی پیچیده کاربرد آن از رگرسیون لجستیک و شبکه عصبی بهتر است.
- ❖ کاربرد SVM عمدتاً در مواردی است که داده ها تفکیک پذیر خطی نباشند.
- ❖ SVM فضای مسئله را تغییر می دهد که نمونه های جدید تفکیک پذیر خطی شوند.
- ❖ در SVM دسته ها طوری تفکیک می شوند که فاصله بین آنها حداکثر شود



- SVM خط راستی که بیشترین حاشیه را با داده دسته ها دارد.
- فاصله خط جدا کننده تا تا نمونه دو دسته برابر است.



- به SVMدسته بندی با حاشیه پهن نیز می گویند چون خطای تعمیم را کاهش می دهد.



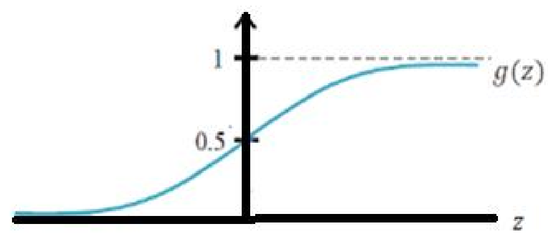
طراحی تابع هزینه SVM

- SVM با اصلاحات در الگوریتم لجستیک حاصل می شود.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \qquad 0 \le h_\theta(x) \le 1$$

$$g(z) = \frac{1}{1+e^{-z}}$$

$$\begin{aligned} &if \;\; h_\theta(x) \ge 0.5 \quad : y = 1 \\ &if \;\; h_\theta(x) < 0.5 \quad : y = 0 \end{aligned}$$



Sigmoid function

$$\begin{aligned} &if \;\; \theta^T x \ge 0 \; : y = 1 \\ &if \;\; \theta^T x < 0 \; : y = 0 \end{aligned}$$

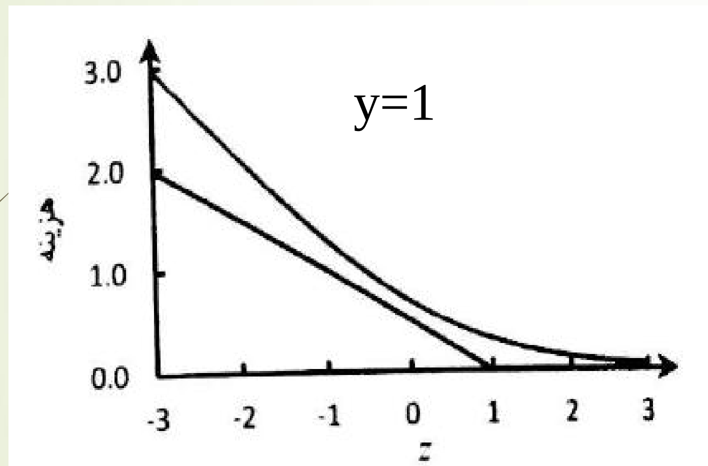
تابع هزینه

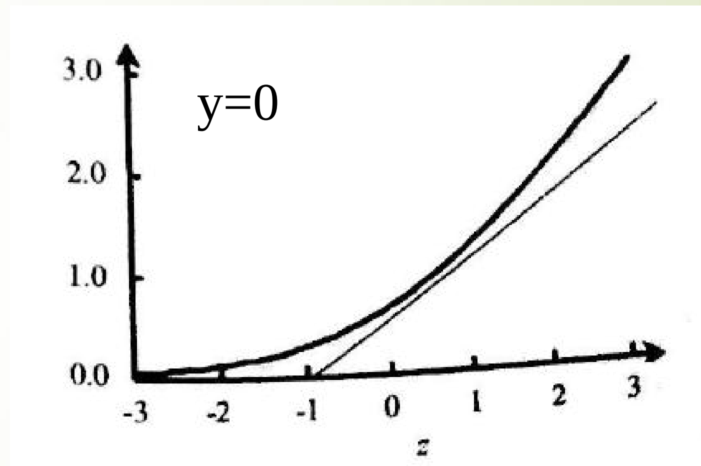
$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & y = 1 \\ -\log(1 - h_\theta(x)) & y = 0 \end{cases}$$

$$Cost(h_\theta(x), y) = -y\log(h_\theta(x)) - (1 - y)\log(1 - h_\theta(x))$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(x^{(i)})) - (1 - y^{(i)})\log(1 - h_\theta(x^{(i)}))\right]$$

$$J(\theta) = -[\frac{1}{m}\sum_{i=1}^{m} y\log\left(\frac{1}{1+e^{\theta^T x}}\right) + (1-y)\log(1-\log\left(1-\frac{1}{1+e^{\theta^T x}}\right)]$$



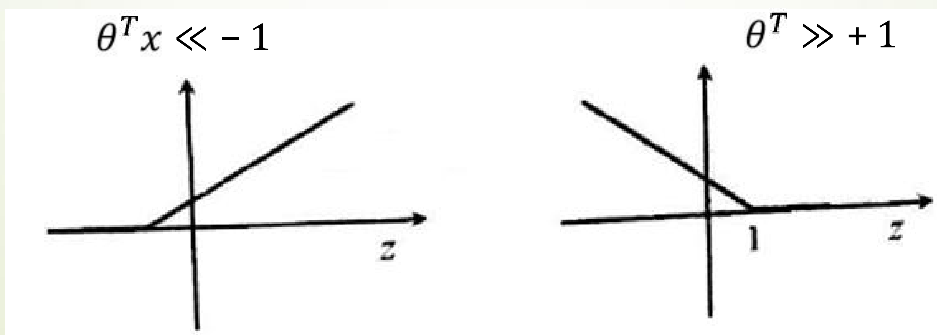


ایده ساخت SVM این است که بجای تابع هزینه لگاریتمی از این خطوط به عنوان تابع هزینه استفاده شود.

$$Min\, J(\theta) = \min([\frac{-1}{m}\sum_{i=1}^{m} y \log\left(\frac{1}{1+ e^{\theta^T x}}\right) + (1 - y)\log(1 - \log\left(1 - \frac{1}{1+ e^{\theta^T x}}\right)] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta^2)$$

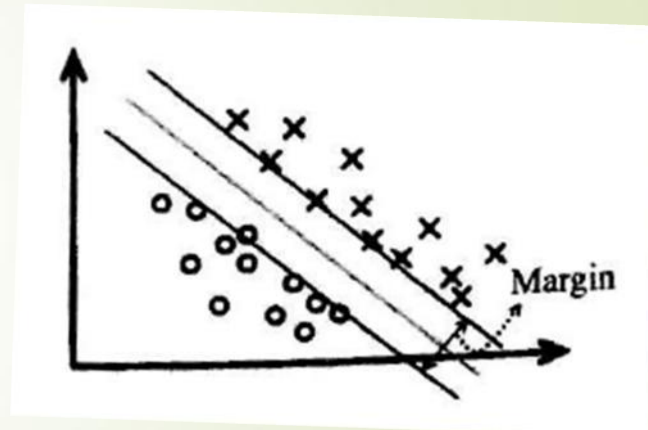
$$\min(\sum_{i=1}^{m} y \log\left(\frac{1}{1+ e^{\theta^T x}}\right) + (1 - y)\log(1 - \log\left(1 - \frac{1}{1+ e^{\theta^T x}}\right) + \frac{\lambda}{2}\sum_{j=1}^{n} \theta^2)$$

$$min_\theta\, (A + \lambda B) \qquad min_\theta\, (CA + B) \qquad \mathrm{C} = \frac{1}{\lambda}$$



- برای تعیین مرز تفکیک تا جایی که امکان دارد مقدار تابع هزینه باید حداقل شود.
- اگر مقدار C کم شود A به سمت صفر میل کرده و فقط باید مقدار B حداقل شود.

$$min_{\theta}\ (CA + B) \qquad C = \frac{1}{\lambda}$$



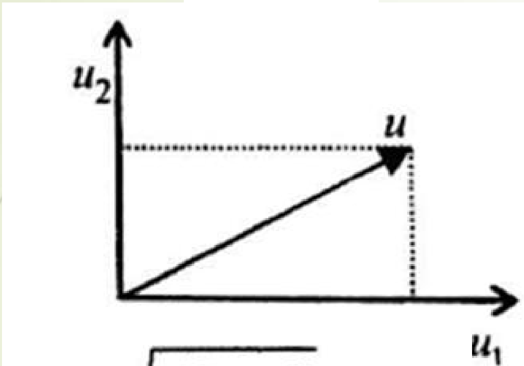
$$\min B = \min \frac{\lambda}{2} \sum_{j=1}^{n} \theta^2 = \min \frac{1}{2} ||\theta||^2$$

$$s.t. \quad \theta^T x^i \geq 1 \quad if\ \ y^i = 1$$

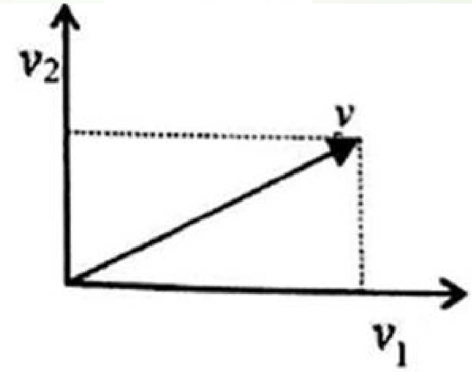
$$\theta^T x^i \leq -1 \quad if\ y^i = 0$$

بردارها

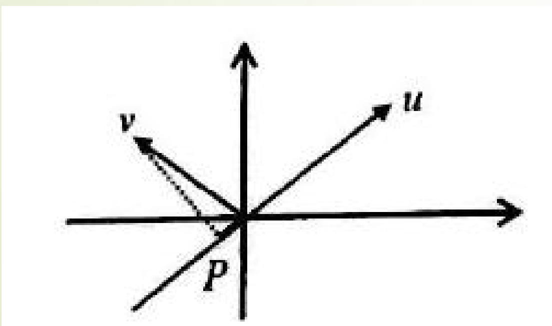
$$u = \begin{bmatrix} u_۱ \\ u_۲ \end{bmatrix} \qquad v = \begin{bmatrix} v_۱ \\ v_۲ \end{bmatrix}$$

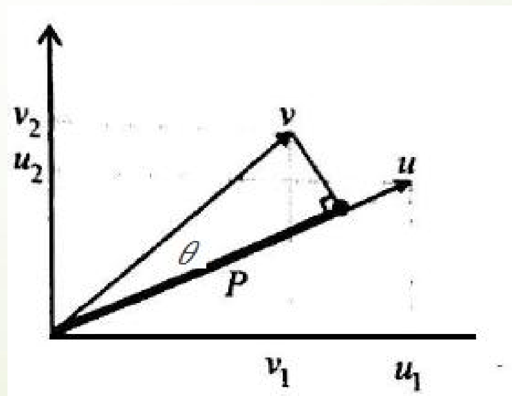


$$\|u\| = \sqrt{u_۱^۲ + u_۲^۲} \in \mathbb{R}$$



$$u^T v = [u_۱ \quad u_۲]\begin{bmatrix} v_۱ \\ v_۲ \end{bmatrix} = u_۱ v_۱ + u_۲ v_۲$$

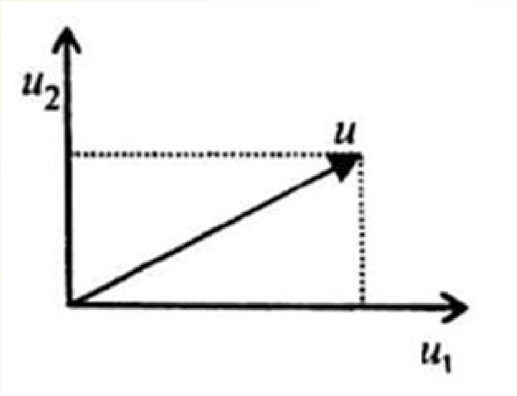




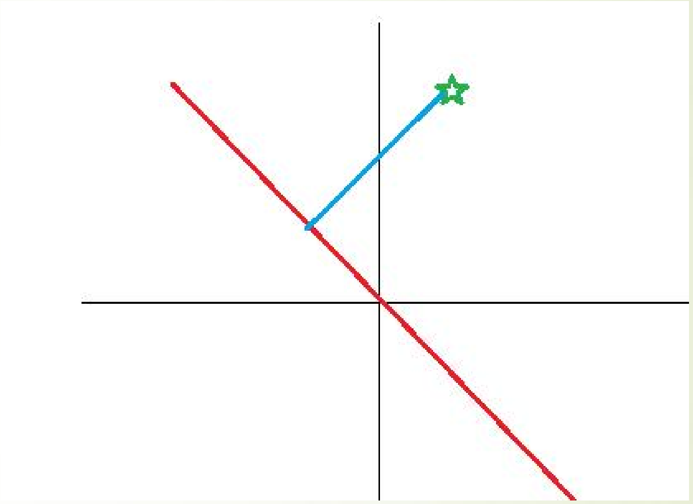
$$Cos(\theta) = \frac{||P||}{||V||}$$

$$\big||P|\big| = \big||V|\big| \cos(\theta) = \frac{\boldsymbol{U.V}}{||U||}$$

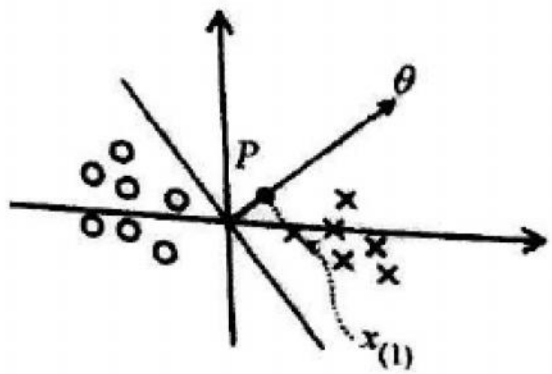
جهت بردار و فاصله نقطه از خط

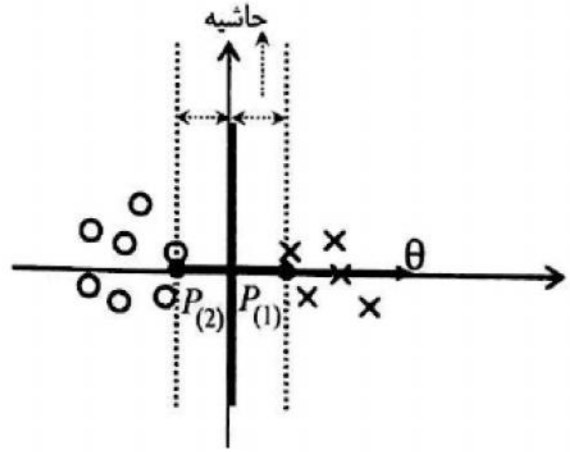


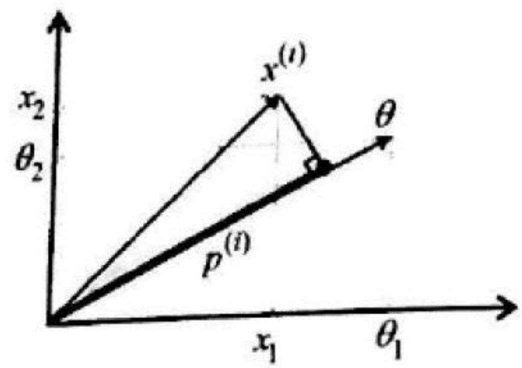
$$w = \left(\frac{u_1}{||u||}, \frac{u_2}{||u||}\right)$$



$$\frac{\theta^T X_A}{||\theta||}$$







$$\min \frac{۱}{۲}\sum_{j=۱}^{n}\theta_j^{۲} = \frac{۱}{۲}\left(\theta_۱^{۲}+\theta_۲^{۲}\right) = \frac{۱}{۲}\left(\sqrt{\theta_۱^{۲}+\theta_۲^{۲}}\right)^{۲}$$

$$\theta^T x^{(i)} = P^{(i)}.\theta = \theta_۱ x_۱ + \theta_۲ x_۲$$

ضرایب لاگرانژ

برای پیدا کردن مینیمم یا ماکزیمم یک تابع با توجه به محدودیت‌ها، از لاگرانژ استفاده می‌کنیم.

optimization problem

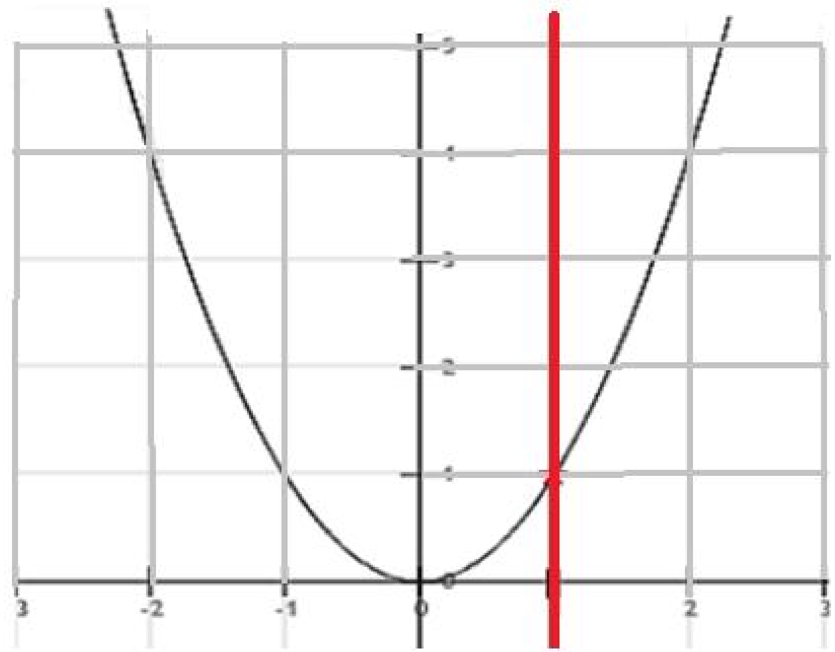
$$\min \; f(x, y)$$
$$\text{subject to } g(x, y) = 0.$$

$$\boxed{\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y)}$$

$$\nabla_{x,y,\lambda} \mathcal{L}(x, y, \lambda) = 0$$

$$L(x, \lambda) = f(x) - \sum_i \lambda_i g_i(x)$$
$$\nabla L(x, \lambda) = 0$$

مثال



$$f(x) = x^2$$

$$g(x) : x = 1$$

مثال

$$f(x, y) = x^2 + y^2$$

$$\begin{cases} g_1(x, y) = x + 1 = 0 \\ g_2(x, y) = y + 1 = 0 \end{cases}$$

اگر شروط نامساوی باشند؟

$$\min \quad f(x, y)$$

$$g(x) \geq 0 \;\Rightarrow \lambda \geq 0$$

$$g(x) \leq 0 \;\Rightarrow \lambda \leq 0$$

مثال:

$$f(x, y) = x^3 + y^2$$

$$g(x, y) = x^2 - 1 \geq 0$$

$$f(x, y) = x^3 + y^3$$

$$g_1(x, y) = x^2 - 1 \geq 0$$

$$g_2(x, y) = y^2 - 1 \geq 0$$

بهینه سازی SVM

$$\min \frac{1}{2}||\theta||^2$$

$$s.t. \quad \theta^T x^i \geq 1 \quad if\ y^i = 1$$
$$\theta^T x^i \leq -1 \quad if\ y^i = 0$$

$$s.t. \quad \theta^T x^i \geq 1 \quad if\ y^i = 1$$
$$\theta^T x^i \leq -1 \quad if\ y^i = -1$$

$$s.t. \quad y^i(\theta^T x^i) \geq 1 = y^i(\theta^T x^i) - 1 \geq 0 = y^i(\theta_0 + x_1^i\theta_1 + \cdots + x_n^i\theta_n) - 1 \geq 0$$

$$L(\theta, \theta_0, \lambda) = \frac{1}{2}||\theta^2|| - \sum_{i=1}^{m} \lambda_i(y^i(\theta^T x^i) - 1) \qquad \lambda_i \geq 0$$

بهینه سازی SVM

$$L(\theta,\theta_0,\lambda) = \frac{1}{2}||\theta^2|| - \sum_{i=1}^{m} \lambda_i(y^i(\theta^T x^i) + \sum_{i=1}^{m} \lambda_i \qquad \lambda_i \geq 0$$

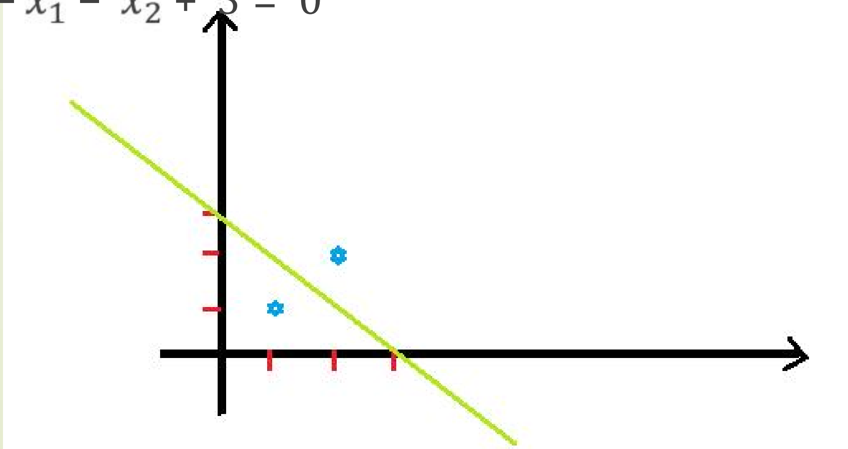
مثال:

داده ها:

(1,1)→+1

(2,2)→ -1

$-x_1 - x_2 + 3 = 0$



دوگان لاگرانژ

$$L(\theta, \theta_0, \lambda) = \frac{1}{2}||\theta^2|| - \sum_{i=1}^{m} \lambda_i(y^i(\theta^T x^i) + \sum_{i=1}^{m} \lambda_i \qquad \lambda_i \geq 0$$

$$\boldsymbol{\theta} = \sum_{i=1}^{m} \lambda_i y^i x^i \qquad , \qquad \sum_{i=1}^{m} \lambda_i y^i = 0$$

$$L_d = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i . x_j)$$

$$L_d = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i . x_j)$$

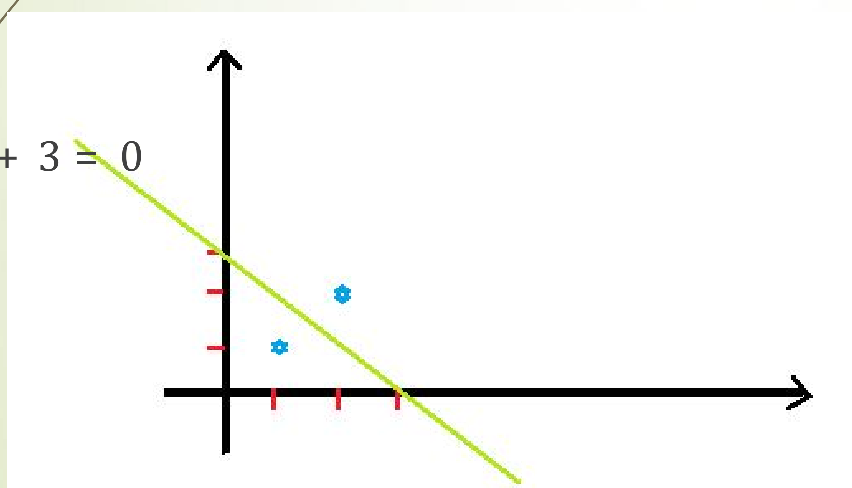
مثال:

داده ها:

(1,1)→+1

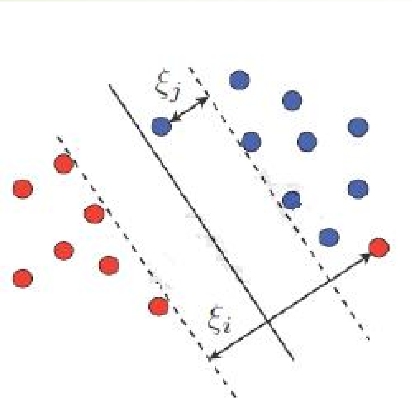
(2,2)→ -1

$- x_1 - x_2 + 3 = 0$



ماشین بردار پشتیبان با حاشیه نرم

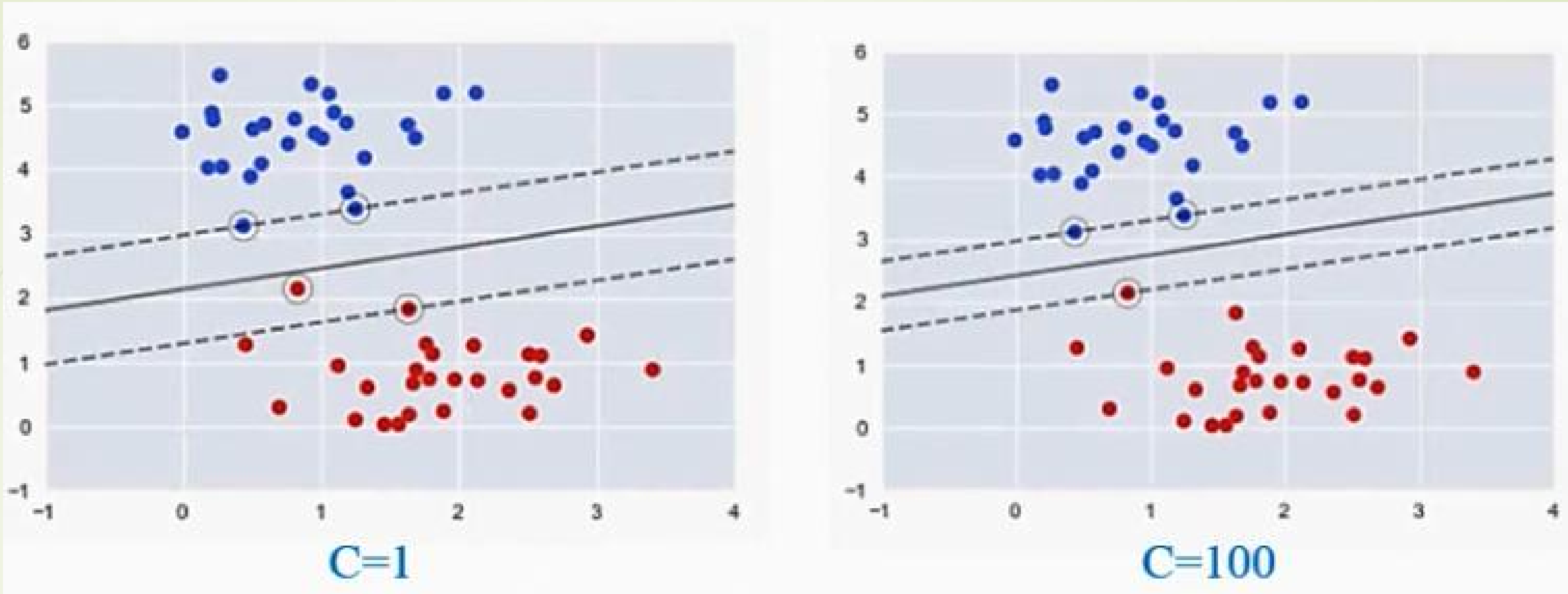
- در SVM قبلی داده ها کاملاً به دو دسته تفکیک می شوند.
- در SVM نرم تعدادی از داده ها می توانند مرزها را رعایت نکنند.



$$\min \frac{1}{2}||\theta||^2 + C\sum_{i=1}^{m} \varepsilon_i$$

$$y^i(\theta^T x^i) \geq 1 - \varepsilon_i$$

$$\varepsilon_i \geq 0$$



$$\min \frac{1}{2}||\theta||^2 + C\sum_{i=1}^{m} \varepsilon_i$$

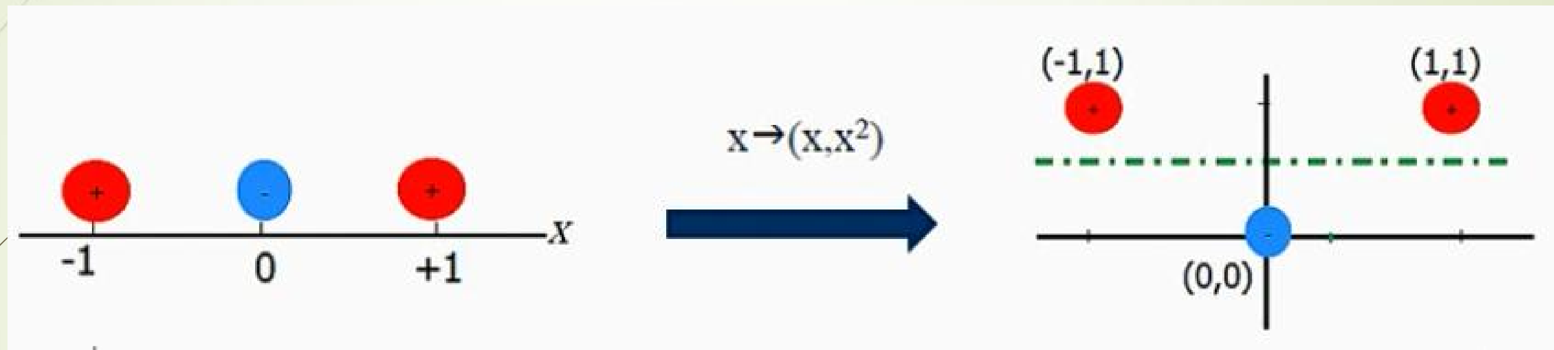
$$S.t. \quad y^i(\theta^T x^i) \geq 1 - \varepsilon_i$$

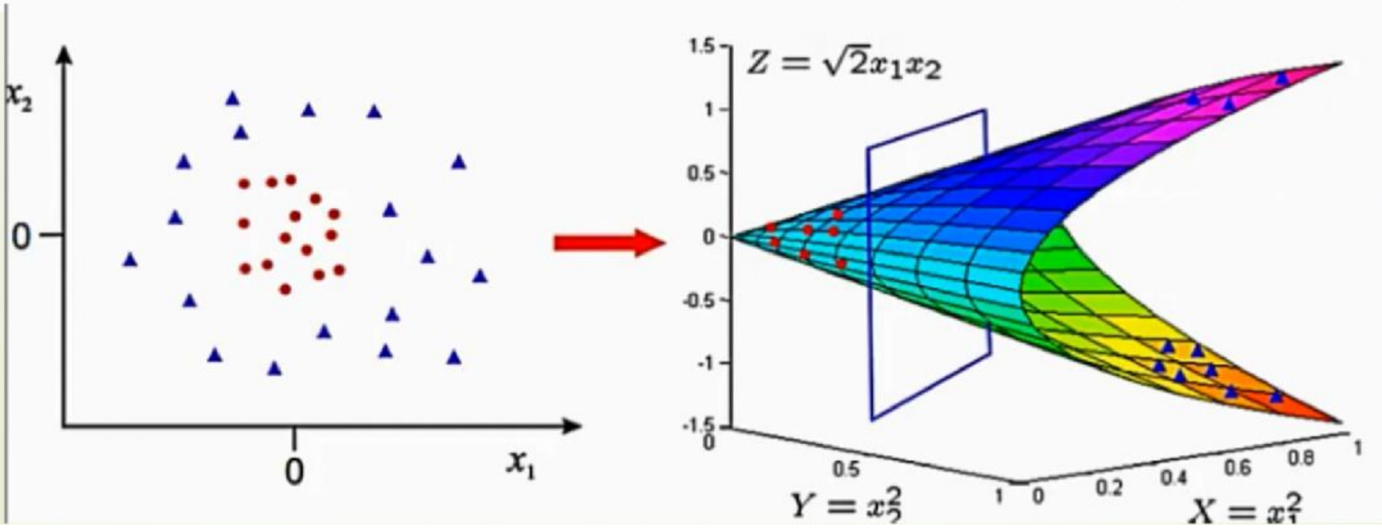
$$\varepsilon_i \geq 0$$

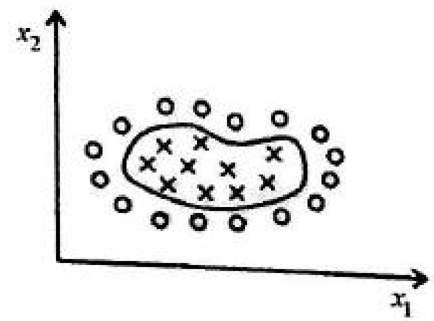
$$L(\theta, \theta_0, \lambda, \varepsilon) = \frac{1}{2}||\theta^2|| + C\sum_{i=1}^{m} \varepsilon_i - \sum_{i=1}^{m} \lambda_i(y^i(\theta^T x^i) - 1 + \varepsilon_i) - \sum_{i=1}^{m} \mu_i \varepsilon_i \qquad \lambda_i, \varepsilon_i \geq 0$$

کرنل (Kernel)

- در واقعیت داده ها به صورت خطی تفکیک پذیر نیستند.
- نگاشت داده ها از یک فضا با ابعاد کمتر به یک فضا با ابعاد بیشتر







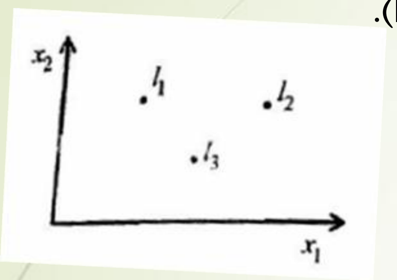
$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \cdots \geq 0$$

$$h_\theta(x) = \begin{cases} 1 & \text{if} \quad \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \cdots \geq 0 \\ 0 & \text{if} \quad \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \cdots < 0 \end{cases}$$

$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \cdots$$

$$f_1 = x_1 \;, f_2 = x_2 \;, f_3 = x_1 x_2 \;, f_4 = x_1^2, f_5 = x_2^2, \cdots$$

- می توان ویژگی های f را به صورت دیگر نیز تعیین کرد؟
  - براساس شباهت با یکسری نقاط مشخص شده در فضای داده ویژگی های جدید را بدست آورد.
  - فرض داده ها به جای دو ویژگی (X_1, X_2) با سه ویژگی جایگزین شوند (l_1, l_2, l_3).
  - F_1 میزان مشابهت داده ها با شاخص l_1
  - F_2 میزان مشابهت داده ها با شاخص l_2
  - F_3 میزان مشابهت داده ها با شاخص l_3



- از فرمول های مختلف می توان برای سنجش شباهت نمونه ها استفاده کرد که به آن کرنل گویند.

$$k(x_i, x_j) = (x_i . x_j + c)^d \qquad c > 0, d \in N$$

$$\boxed{K(\mathbf{x}, \mathbf{x'}) = \exp\left(-\frac{\|\mathbf{x'} - \mathbf{x}\|^2}{2\sigma^2}\right)}$$

$$\boxed{K(\mathbf{x}, \mathbf{x'}) = \tanh\left(a(\mathbf{x} \cdot \mathbf{x'}) + b\right)} \qquad a, b > 0$$

کرنل گوسی (Gaussian Kernel)

- استفاده از فاصله اقلیدسی در تابع کرنل گوسی
  - برای هر نمونه مانند X:

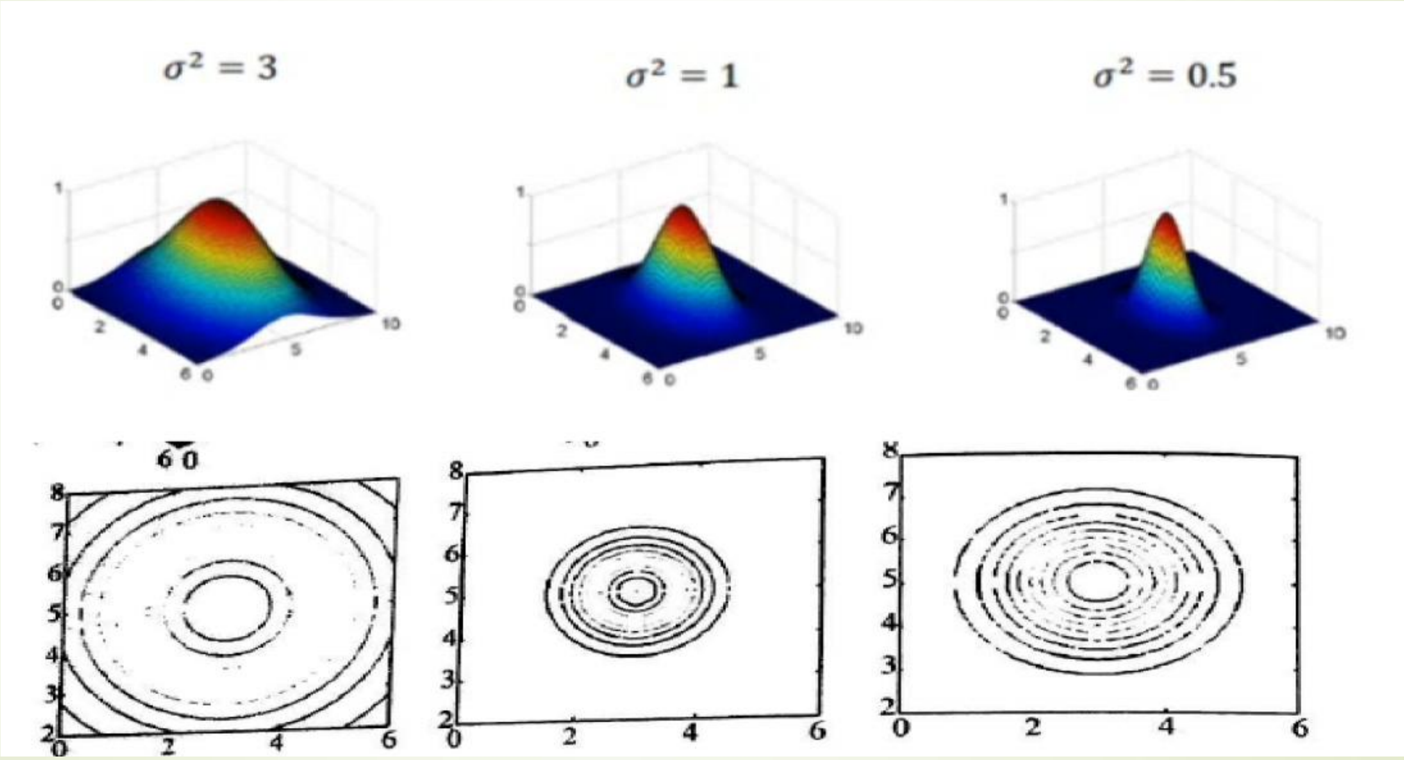
$$f_1 = sim\,(x, l^1) = \exp\left(- \frac{||x - l^1||^2}{2\sigma^2}\right) = e^{\left(-\frac{||x-l^1||^2}{2\sigma^2}\right)}$$

$$f_2 = sim\,(x, l^2) = \exp\left(- \frac{||x - l^2||^2}{2\sigma^2}\right) = e^{\left(-\frac{||x-l^2||^2}{2\sigma^2}\right)}$$

$$f_3 = sim\,(x, l^3) = \exp\left(- \frac{||x - l^3||^2}{2\sigma^2}\right) = e^{\left(-\frac{||x-l^3||^2}{2\sigma^2}\right)}$$

- اگر x به l نزدیک باشد f متناظر تقریباً یک می شود و اگر دور باشد تقریباً صفر خواهد شد.

- هر چه σ کوچکتر باشد بایاس کمتر، واریانس بیشتر و خروجی سریعتر کاهش می یابد.
هر چه σ بزرگتر باشد بایاس بیشتر، واریانس کمتر و خروجی کندتر کاهش می یابد.



تعیین مراکز کرنل

- تعیین m نقطه شاخص دلخواه
- در نظر گرفتن هر داده به عنوان یک نقطه شاخص

$$x^1 = l^1, x^2 = l^2, \ldots, x^m = l^m$$

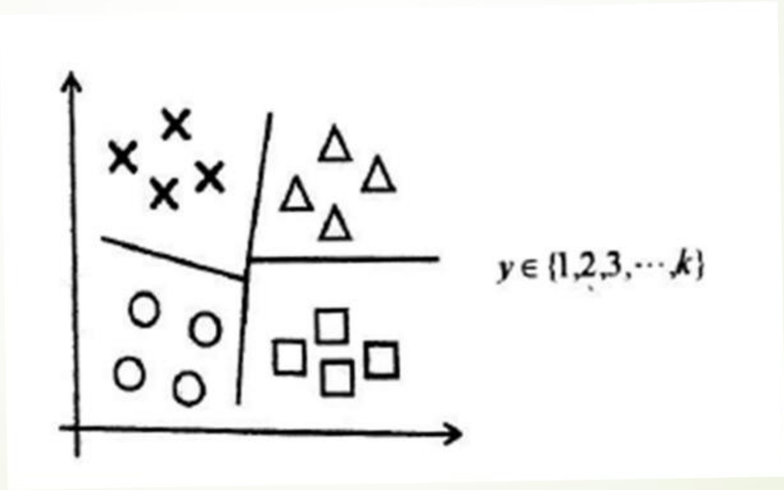
$$f_j^i = sim\left(x^i, l^j\right) = \exp\left(-\frac{||x^i - l^j||^2}{2\sigma^2}\right) = e^{\left(-\frac{||x^i - l^j||^2}{2\sigma^2}\right)}$$

<u>مثال:</u>

(2, 3)→+1, (1, 1)→+1, (2, 2)→ -1

دسته بندی چند دسته ای

- روش یک دسته در مقابله بقیه



مقایسه SVM و الگوریتم لجستیک

- اگر ویژگی ها زیادتر از تعداد داده ها باشد (n>m) می توان از الگوریتم لجستیک یا SVM بدون کرنل استفاده کرد (ایمیل ها).
- اگر ویژگی ها کم و تعداد داده ها متوسط باشد استفاده از SVM با کرنل گوسی توصیه می شود.
- اگر ویژگی ها کم و تعداد داده ها زیاد باشد می توان از لجستیک یا SVM بدون کرنل استفاده کرد (کرنل بار محاسبات را زیاد می کند).
- در موارد فوق می توان از شبکه عصبی استفاده کرد فقط آموزش آن کندتر است.
- SVM محدب است و همیشه جواب بهینه و یا نزدیک بهینه تولید می کند.