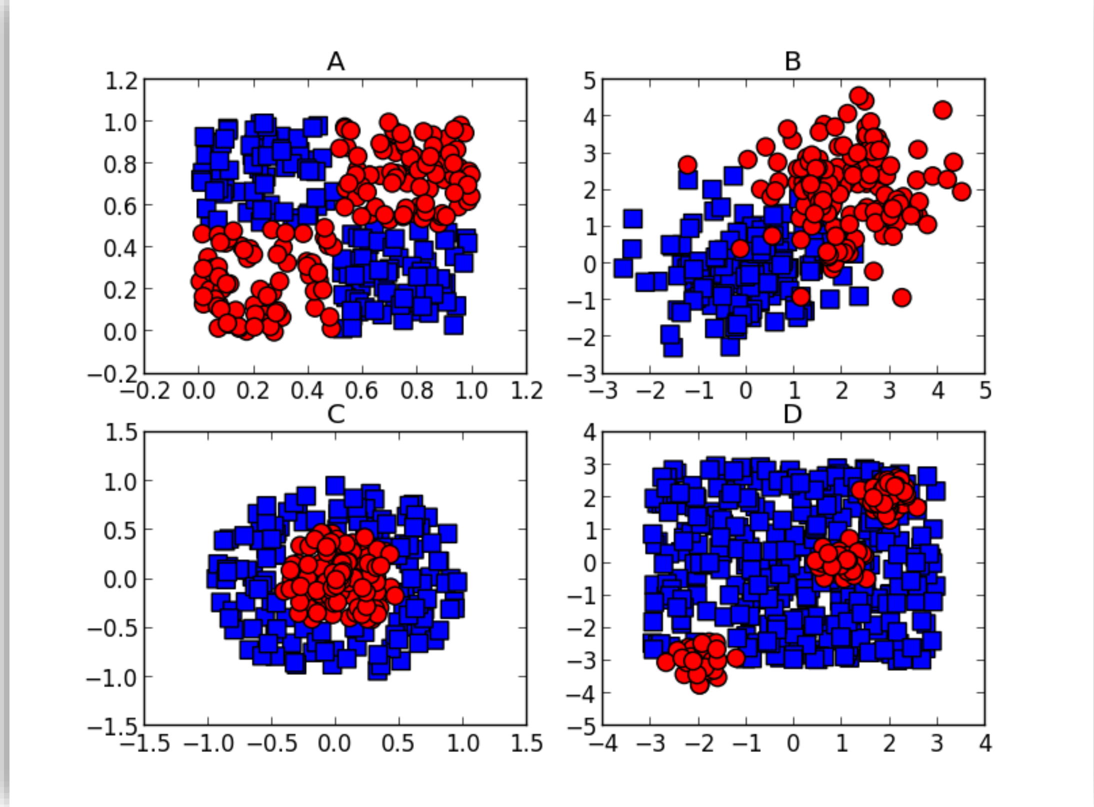
# فصل دهم : فهرست مطالب

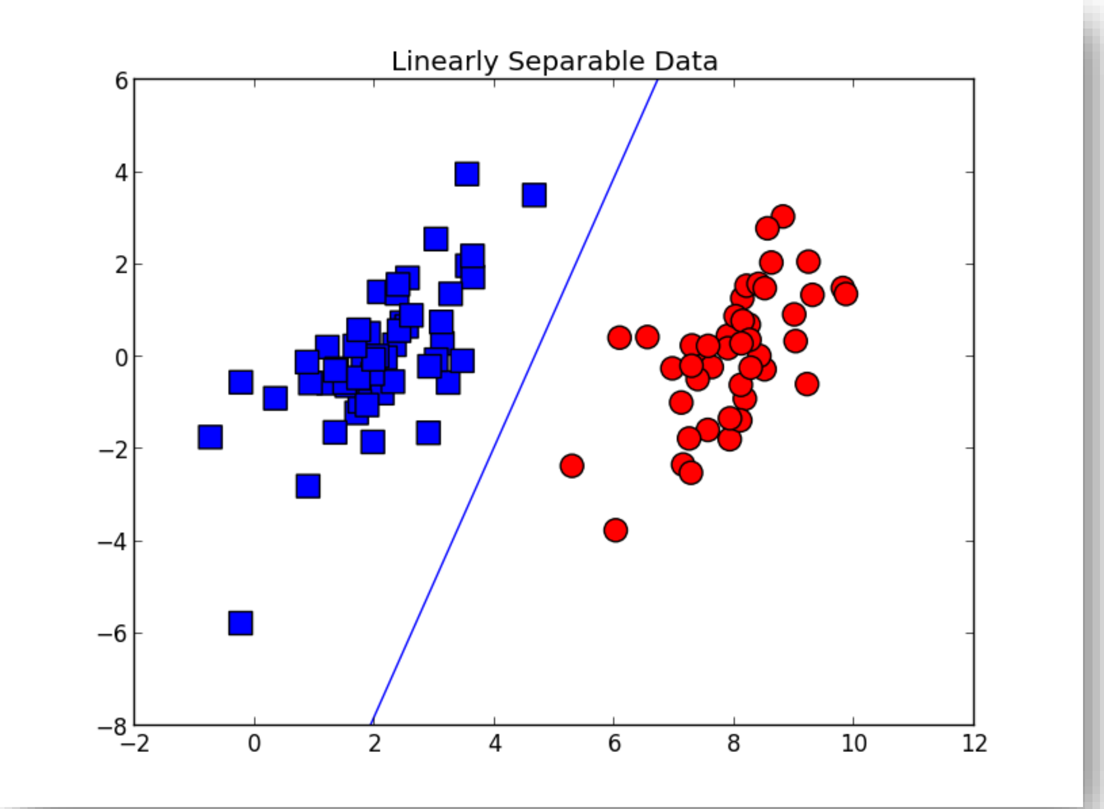
# فصل دهم : معرفی

- ماشین‌های بردار پشتیبان [وپنیک، ۱۹۹۲]
  - یکی از پرطرفدارترین الگوریتم‌های یادگیری ماشین!
  - جداسازی بهتر داده‌ها نسبت به سایر روش‌های یادگیری ماشین (مسائل کلاس‌بندی)!
  - استفاده از آن نسبتاً آسان است!
  - استفاده از ترفند کرنل:
    - کلاس‌بندی، رگرسیون، تخمین توزیع، کلاس‌بندی تک کلاسی و ...

# فصل دهم : داده های تفکیک ناپذیر خطی

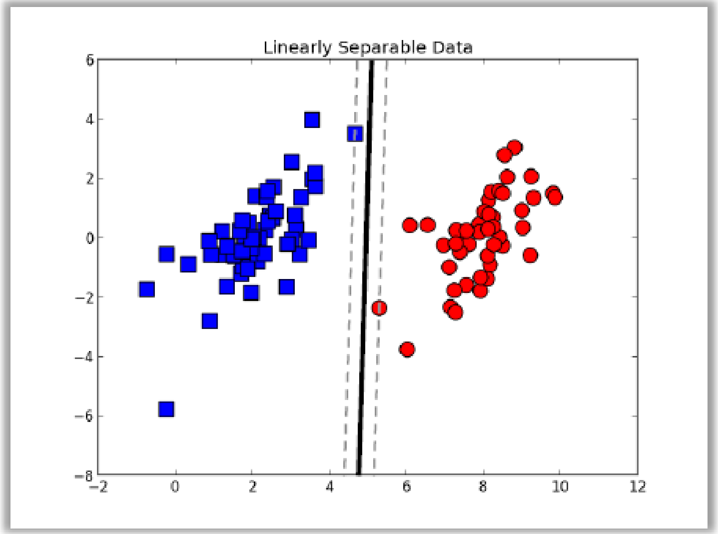


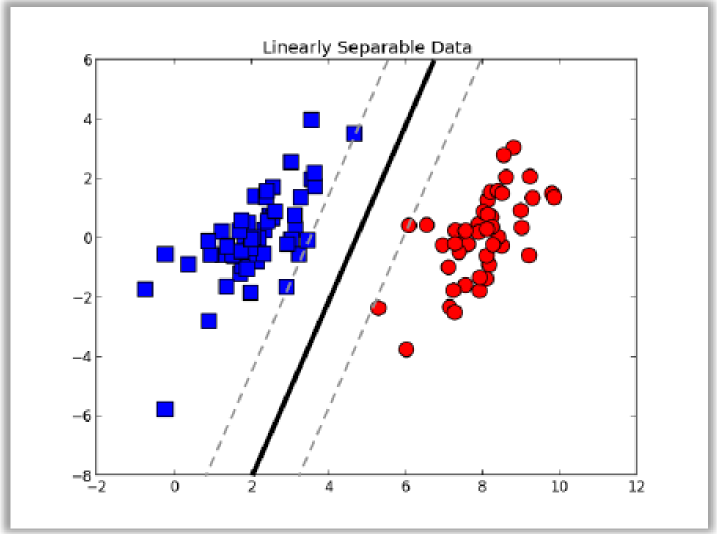
# فصل دهم : داده های تفکیک پذیر خطی



# فصل دهم : مرز تصمیم گیری بهینه

- س. کدام مرز تصمیم‌گیری بهتر است؟

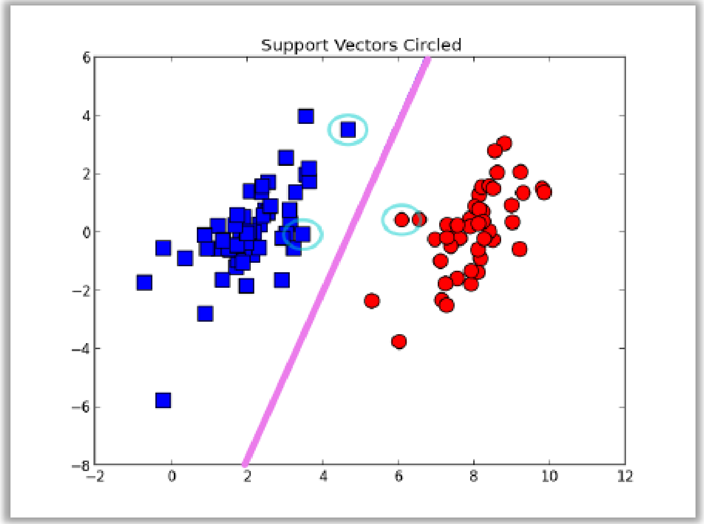



- راه‌حل بیشترین حاشیه. بیشترین پایداری در برابر تخریب داده‌ها. [افزایش قابلیت تعمیم]

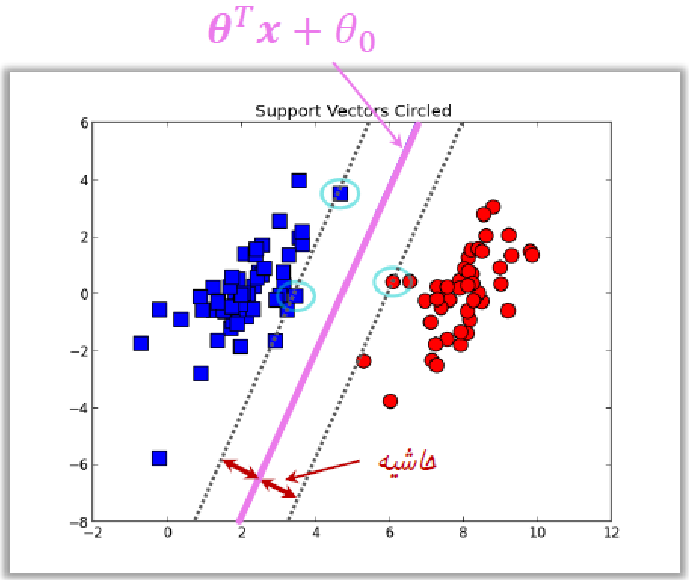
# فصل دهم : بردارهای پشتیبان

- بردار پشتیبان. نزدیک‌ترین نقاط به مرز تصمیم‌گیری.
- هدف. بیشینه کردن فاصله بردارهای پشتیبان از مرز تصمیم‌گیری.



- حاشیه. فاصله بردارهای پشتیبان تا مرز تصمیم‌گیری.

- هدف. بیشینه کردن فاصله بردارهای پشتیبان از مرز تصمیم‌گیری.




$$\frac{|\boldsymbol{\theta}^T\boldsymbol{x} + \theta_0|}{\|\boldsymbol{\theta}\|} \geq \rho$$

# فصل دهم : مرز تصمیم گیری بهینه(نمادها)

- نمونه‌های آموزشی.

$$X = (\boldsymbol{x}^t, y^t), \qquad y^t = \begin{cases} +1 & if\ x^t \in C_1 \\ -1 & if\ x^t \in C_2 \end{cases}$$

- هدف. یافتن بردار $\theta$ و مقدار $\theta_0$ به طوری که:

$$\begin{aligned} \boldsymbol{\theta}^T\boldsymbol{x}^t + \theta_0 &\geq +1 \quad \text{for} \quad y^t = +1 \\ \boldsymbol{\theta}^T\boldsymbol{x}^t + \theta_0 &\leq -1 \quad \text{for} \quad y^t = -1 \end{aligned} \Rightarrow \quad y^t(\boldsymbol{\theta}^T\boldsymbol{x}^t + \theta_0) \geq +1$$

# فصل دهم : تابع هدف

- هدف. بیشینه کردن فاصله بردارهای پشتیبان از مرز تصمیم‌گیری.
  - فاصله داده $x$ از مرز تصمیم‌گیری:

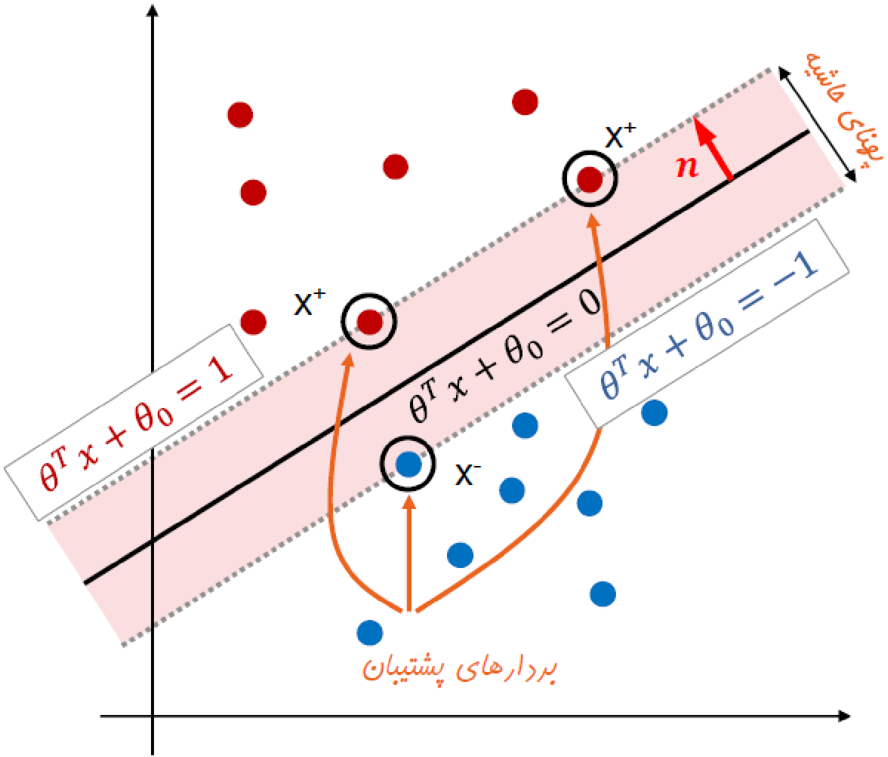
$$\frac{|\boldsymbol{\theta}^T \boldsymbol{x} + \theta_0|}{\|\boldsymbol{\theta}\|} \geq \rho \quad \Rightarrow \quad |\boldsymbol{\theta}^T \boldsymbol{x} + \theta_0| \geq \rho \|\boldsymbol{\theta}\|$$

  - این معادله بی‌نهایت جواب دارد؛ اما با قرار دادن $\rho\|\boldsymbol{\theta}\| = 1$، خواهیم داشت:

$$\rho\|\boldsymbol{\theta}\| = 1 \Rightarrow \rho = \frac{1}{\|\boldsymbol{\theta}\|}$$

- هدف. برای بیشینه کردن حاشیه، می‌توان اندازه بردار $\boldsymbol{\theta}$ را کمینه نمود.
- محدودیت‌ها. مرز تصمیم‌گیری باید داده‌های دو کلاس را به درستی از یکدیگر تفکیک کند.

# فصل دهم : تابع هدف




- می‌دانیم:

$$\theta^T x^+ + \theta_0 = +1$$

$$\theta^T x^- + \theta_0 = -1$$

- بنابراین:

$$M = (x^+ - x^-) \cdot n$$

$$= (x^+ - x^-) \cdot \frac{\theta}{\|\theta\|} = \frac{2}{\|\theta\|}$$

# فصل دهم : تابع هدف(بیان رسمی)

- تابع هدف.

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

$$\text{s.t.} \quad (\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) \geq +1 \quad if\ y^t = +1$$

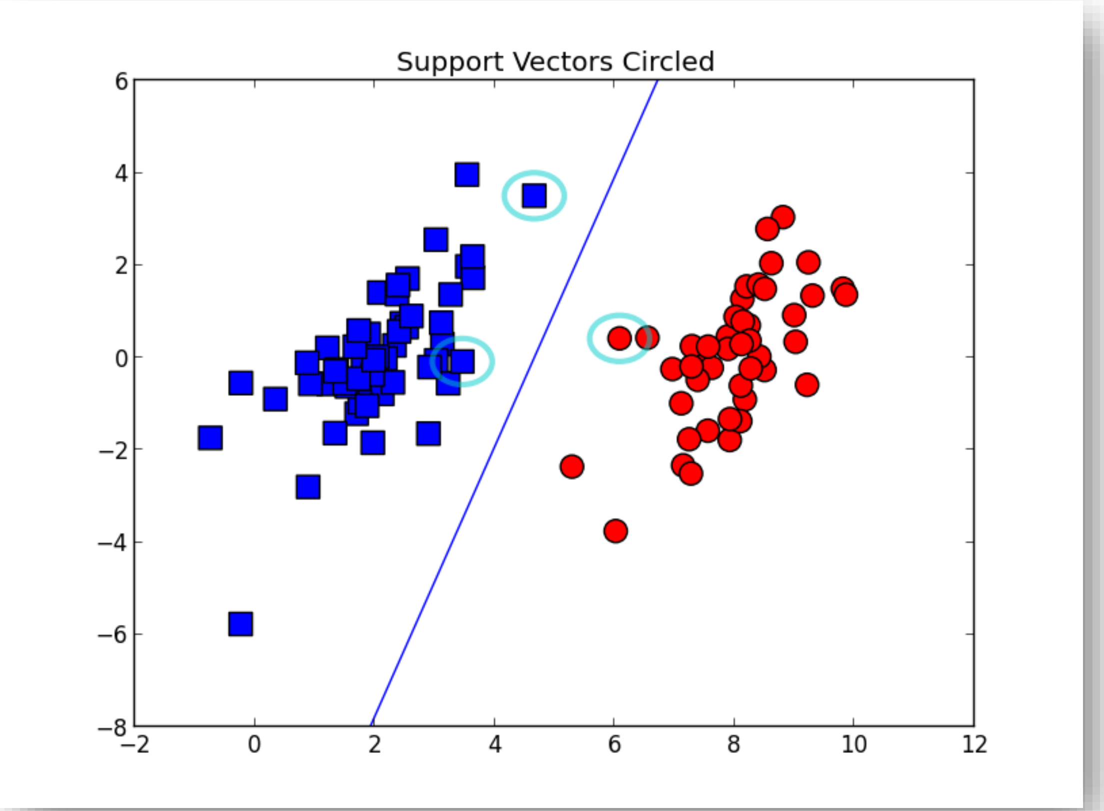
$$(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) \leq -1 \quad if\ y^t = -1$$

- ساده‌سازی.

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

$$\text{s.t.} \quad y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) \geq +1$$

# فصل دهم : تابع هدف



# فصل دهم : تابع هدف(بیان رسمی)

- تابع هدف.

$$\min \frac{1}{2}\|\boldsymbol{\theta}\|^2$$
$$\text{s.t.} \quad y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) \geq +1$$

بهینه‌سازی محدب

- حل مسئله با استفاده از ضرایب لاگرانژ.

$$L_p = \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{t=1}^{m} \alpha^t [y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) - 1]$$
$$= \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{t=1}^{m} \alpha^t y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) + \sum_{t=1}^{m} \alpha^t$$

جوزف لویی لاگرانژ

۱۸۱۳ - ۱۷۳۶

# فصل دهم : تابع هدف(بیان رسمی)

$$L_p = \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{t=1}^{m} \alpha^t [y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) - 1]$$

$$= \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{t=1}^{m} \alpha^t y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) + \sum_{t=1}^{m} \alpha^t$$

$$\frac{\partial L_p}{\partial \boldsymbol{\theta}} = 0 \Rightarrow \boldsymbol{\theta} = \sum_{t=1}^{m} \alpha^t y^t \boldsymbol{x}^t$$

مرز تصمیم‌گیری یک ترکیب خطی از داده‌های آموزشی

$$\frac{\partial L_p}{\partial \theta_0} = 0 \Rightarrow \sum_{t=1}^{m} \alpha^t y^t = 0$$

# فصل دهم : تابع هدف(بیان رسمی)

$$L_d = \frac{1}{2}(\boldsymbol{\theta}^T\boldsymbol{\theta}) - \boldsymbol{\theta}^T \sum_{t=1}^{m} \alpha^t y^t \boldsymbol{x}^t - \theta_0 \sum_{t=1}^{m} \alpha^t y^t + \sum_{t=1}^{m} \alpha^t$$

$$= -\frac{1}{2}(\boldsymbol{\theta}^T\boldsymbol{\theta}) + \sum_{t=1}^{m} \alpha^t$$

$$= -\frac{1}{2}\sum_{t=1}^{m}\sum_{s=1}^{m} \alpha^t \alpha^s y^t y^s (\boldsymbol{x}^t)^T \boldsymbol{x}^s + \sum_{t=1}^{m} \alpha^t$$

الگوریتم
بهینه‌سازی ترتیبی مینیمال
پلَت (۱۹۹۹)

subject to $\sum_{t=1}^{m} \alpha^t y^t = 0$ and $\alpha^t \geq 0\ \forall t$

- مقدار بسیاری از آلفاها برابر با صفر است و تنها تعداد اندکی دارای مقدار بزرگ‌تر از صفر هستند؛
- $x$ هایی که به ازای آنها مقدار آلفا بزرگ‌تر از صفر است، همان بردارهای پشتیبان هستند.

# فصل دهم : تابع هدف(شکل ساده شده)

$$L_d = -\frac{1}{2}\sum_{t=1}^{m}\sum_{s=1}^{m}\alpha^t\alpha^s y^t y^s (\boldsymbol{x}^t)^T\boldsymbol{x}^s + \sum_{t=1}^{m}\alpha^t$$

$$= -\frac{1}{2}\alpha^T Q\alpha + e^T\alpha$$

الگوریتم
بهینه‌سازی ترتیبی مینیمال
پلَت (۱۹۹۹)

$Q_{ts} = y^t y^s (x^t)^T x^s,\qquad e = [1 \quad 1 \quad \dots \quad 1]^T \in \mathbb{R}^m$

subject to $\sum_{t=1}^{m}\alpha^t y^t = 0$ and $\alpha^t \geq 0\ \forall t$

- مقدار بسیاری از آلفاها برابر با صفر است و تنها تعداد اندکی دارای مقدار بزرگ‌تر از صفر هستند؛
- $x$ هایی که به ازای آنها مقدار آلفا بزرگ‌تر از صفر است، همان بردارهای پشتیبان هستند.

# فصل دهم : داده های تفکیک ناپذیر خطی(حاشیه نرم)

- س. اگر داده‌ها به صورت خطی تفکیک‌پذیر نباشند چه می‌شود؟
- حاشیه نرم. اجازه دادن اندکی خطا در جداسازی!

$$y^t(\boldsymbol{\theta}^T\boldsymbol{x}^t + \theta_0) \geq 1 - \varepsilon^t$$

- خطای نرم.

$$soft\ error = \sum_{t=1}^{m} \varepsilon^t$$

- تابع هدف جدید.

ضریب جریمه

$$\min \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{m}\varepsilon^t$$

$$\text{s.t.} \quad y^t(\boldsymbol{\theta}^T\boldsymbol{x}^t + \theta_0) \geq 1 - \varepsilon^t$$

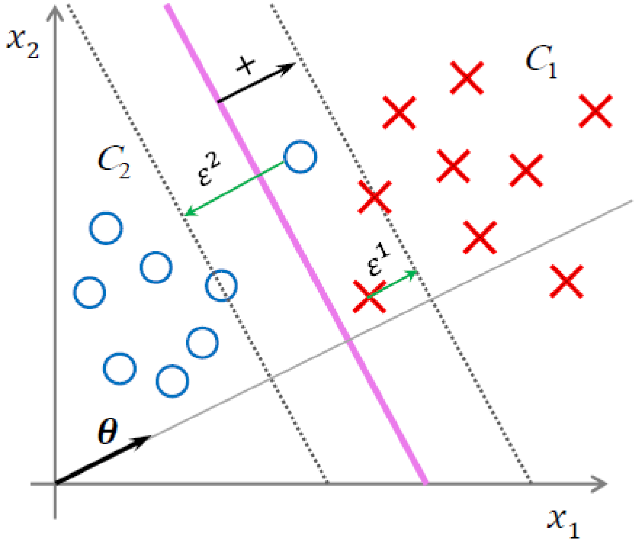
$$\varepsilon^t \geq 0$$

# فصل دهم : داده های تفکیک ناپذیر خطی(حاشیه نرم )

$$\min \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{m} \varepsilon^t$$

$$\text{s.t.} \quad y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) \geq 1 - \varepsilon^t$$

$$\varepsilon^t \geq 0$$




ضرایب لاگرانژ

$$L_p = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{m} \varepsilon^t - \sum_{t=1}^{m} \alpha^t[y^t(\boldsymbol{\theta}^T \boldsymbol{x}^t + \theta_0) - 1 + \varepsilon^t] - \sum_{t=1}^{m} \mu^t \varepsilon^t$$

# فصل دهم : داده های تفکیک ناپذیر خطی(حاشیه نرم )

$$L_p = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{m}\varepsilon^t - \sum_{t=1}^{m}\alpha^t[y^t(\boldsymbol{\theta}^T\boldsymbol{x}^t + \theta_0) - 1 + \varepsilon^t] - \sum_{t=1}^{m}\mu^t\varepsilon^t$$

$$\frac{\partial L_p}{\partial \boldsymbol{\theta}} = 0 \Rightarrow \boldsymbol{\theta} = \sum_{t=1}^{m}\alpha^t y^t \boldsymbol{x}^t$$

$$\frac{\partial L_p}{\partial \theta_0} = 0 \Rightarrow \sum_{t=1}^{m}\alpha^t y^t = 0$$

$$\frac{\partial L_p}{\partial \varepsilon^t} = 0 \Rightarrow C - \alpha^t - \mu^t = 0 \Rightarrow 0 \le \alpha^t \le C$$

# فصل دهم : تابع هدف(دوگان)

$$L_d = -\frac{1}{2}\sum_{t=1}^{m}\sum_{s=1}^{m}\alpha^t\alpha^s y^t y^s (\boldsymbol{x}^t)^T \boldsymbol{x}^s + \sum_{t=1}^{m}\alpha^t$$
$$= -\frac{1}{2}\alpha^T Q\alpha + e^T\alpha$$

الگوریتم
بهینه‌سازی ترتیبی مینیمال
پلَت (۱۹۹۹)

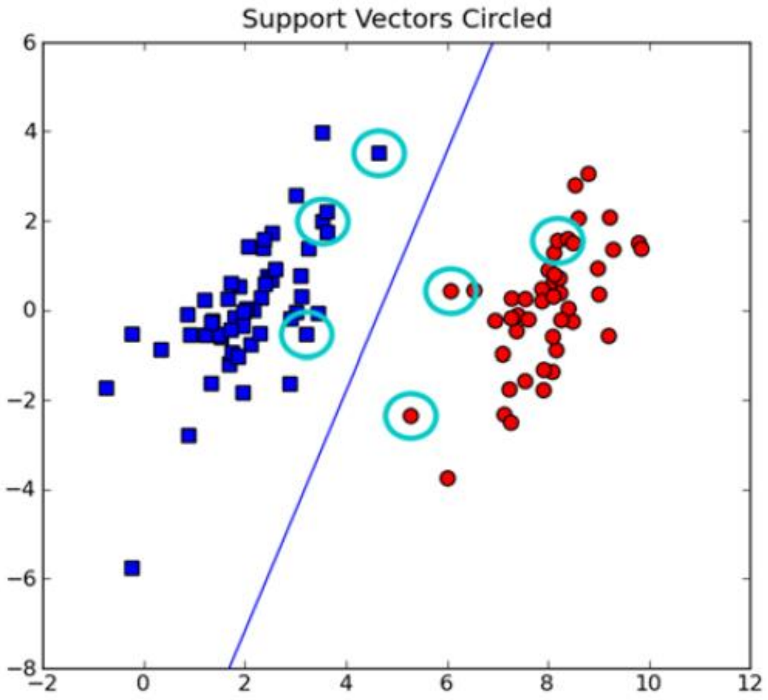
subject to $\sum_{t=1}^{m}\alpha^t y^t = 0$ and $0 \le \alpha^t \le C \ \forall t$

- تخمین خطا بر اساس تعداد بردارهای پشتیبان. [وپنیک، ۱۹۹۵]
  - مقدار بسیاری از آلفاها برابر با صفر است و تنها تعداد اندکی دارای مقدار بزرگ‌تر از صفر هستند؛
  - $x$ هایی که به ازای آنها مقدار آلفا بزرگ‌تر از صفر است، همان بردارهای پشتیبان هستند.

$$E_m[P(error)] \le \frac{E_m[\ \#of\ support\ vectors]}{m}$$

# فصل دهم : راه حل حاشیه نرم

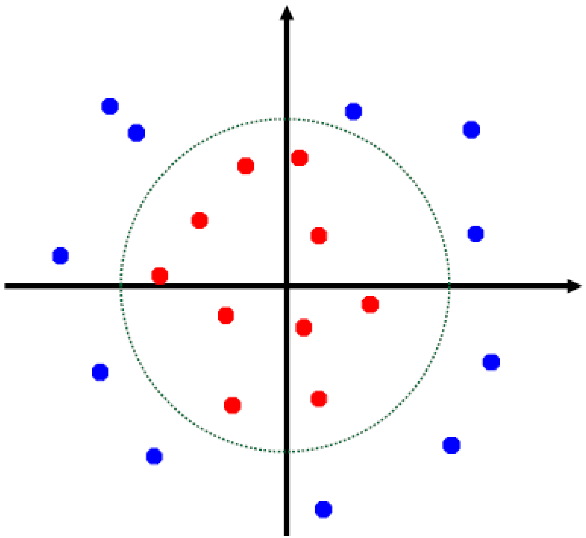
- بردارهای پشتیبان.



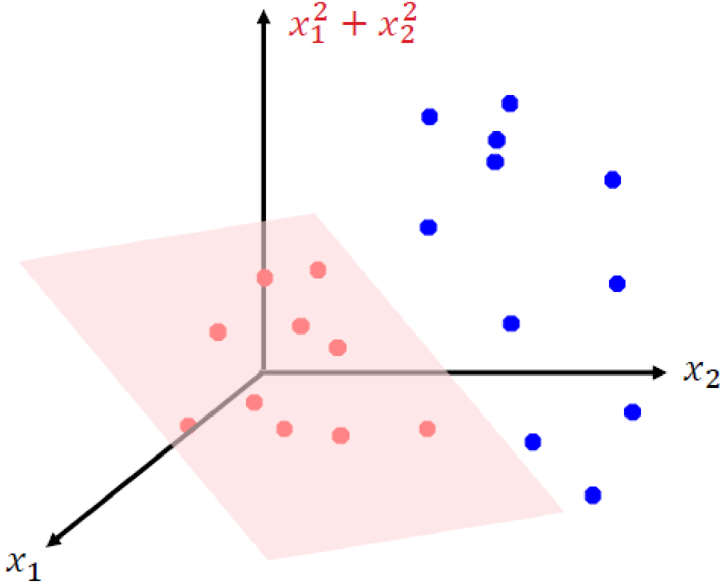
# ترفند کرنل

# فصل دهم : توابع کرنل

- ایده. نگاشت مسئله به یک فضای ویژگی جدید با استفاده از تبدیلات غیرخطی.
  - استفاده از یک مدل خطی در فضای جدید به منظور کلاس‌بندی داده‌ها.
  - مدل خطی در فضای جدید متناظر با یک مدل غیرخطی در فضای اصلی است.

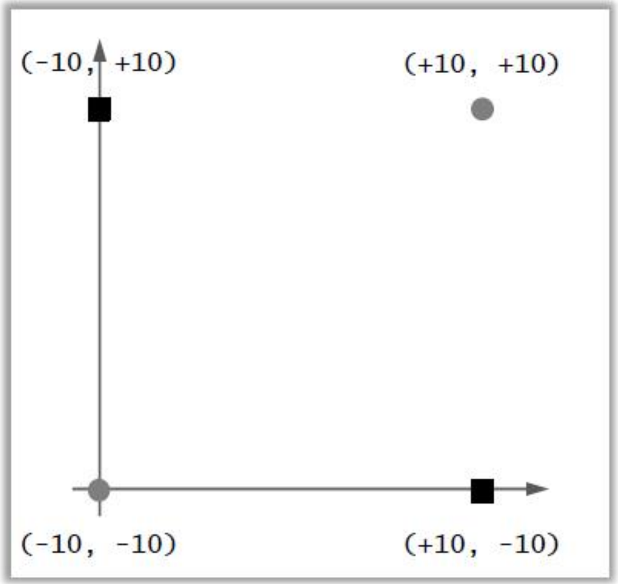


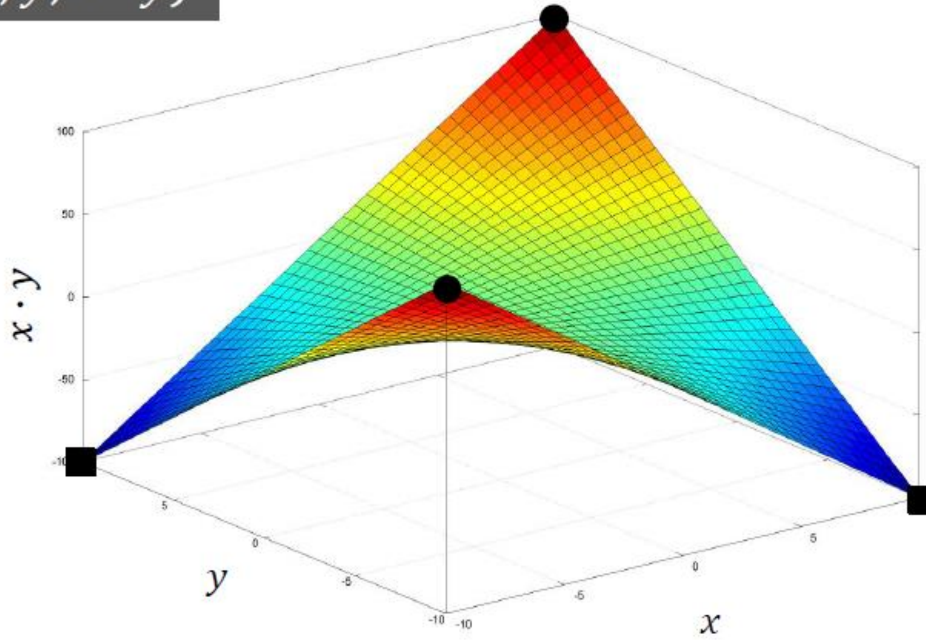
$\Phi:\ \mathbf{x} \rightarrow \varphi(\mathbf{x})$



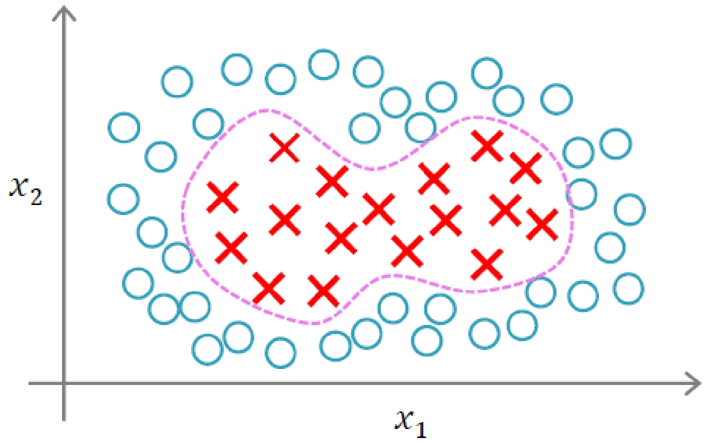
# فصل دهم : مثال(مسئله XOR )

$$(x, y) \rightarrow (x, y, x \cdot y)$$



# فصل دهم : مرزهای تصمیم گیری غیرخطی




پیش‌بینی. $y = 1$ اگر:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \cdots \geq 0$$

ویژگی‌ها.

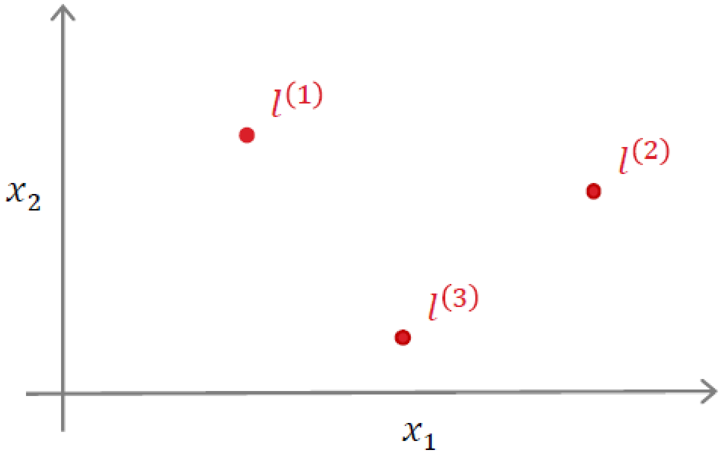
$f_1 = x_1, \quad f_2 = x_2, \quad f_3 = x_1^2, \quad f_4 = x_2^2, \quad f_5 = x_1 x_2, \quad \cdots$

$$h_\theta(f) = \theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \theta_4 f_4 + \theta_5 f_5 + \cdots$$ ← مرز تصمیم‌گیری خطی

س. آیا روش بهتری برای انتخاب ویژگی‌های جدید $f_1$ ، $f_2$ و ... وجود دارد؟

# فصل دهم : کرنل

□ ایده. با داشتن $x$، مجموعه جدید ویژگی‌ها را بر اساس شباهت آن با نقاط راهنمای $l^{(1)}$، $l^{(2)}$ و $l^{(3)}$ انتخاب کن.




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

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

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

کرنل (کرنل گوسی)

□ تابع کرنل. معیاری به منظور محاسبه شباهت میان داده‌های $x$ و $y$

# فصل دهم : کرنل ها به عنوان معیار شباهت

- تابع کرنل.

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

- حالت اول. $x \approx l^{(i)}$

$$f_i \approx \exp\left(-\frac{0}{2\sigma^2}\right) = \exp(0) = 1$$

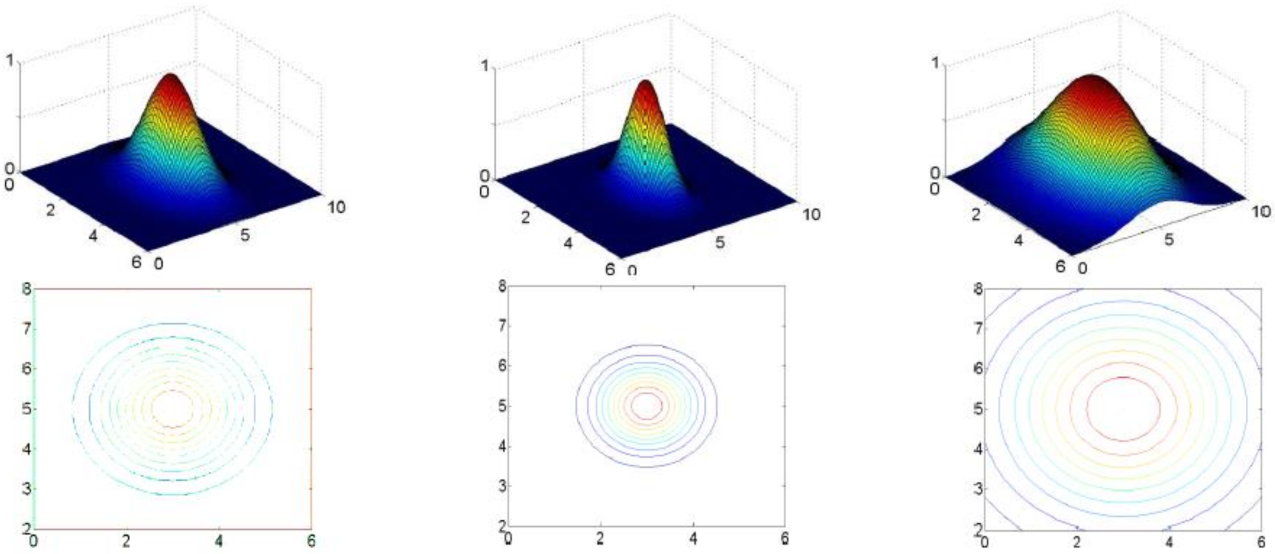
- حالت دوم. $x$ بسیار دور از $l^{(i)}$

$$f_i \approx \exp\left(-\frac{\infty}{2\sigma^2}\right) = \exp(-\infty) = 0$$

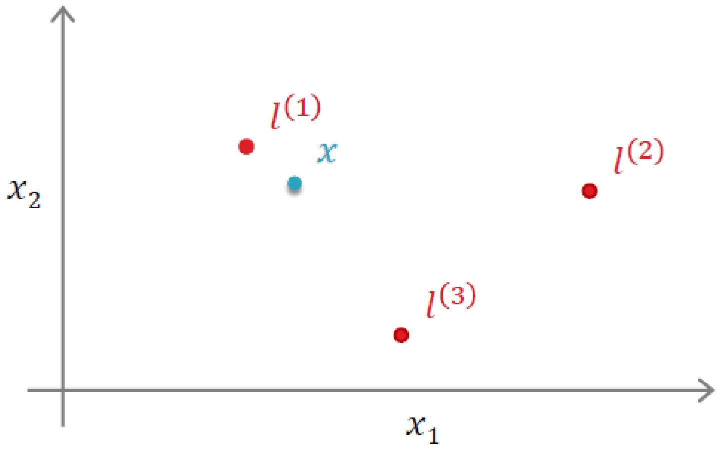
- مثال.

$$l^{(1)} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

$\sigma^2 = 1$ $\quad$ $\sigma^2 = 0.5$ $\quad$ $\sigma^2 = 3$



# فصل دهم : کرنل ها به عنوان معیار شباهت




پیش‌بینی. $y = 1$ اگر:

$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$$

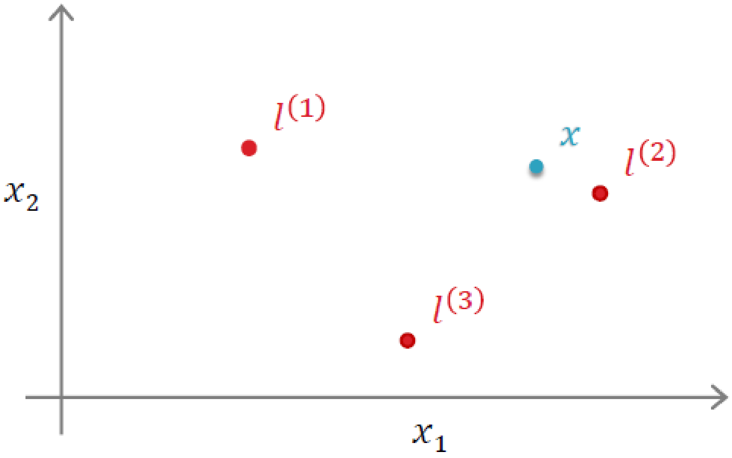
$-0.5$ $\quad 1.0$ $\quad 1.0$ $\quad 0.0$

$$f_1 \approx 1, f_2 \approx f_3 \approx 0$$

$$h_\theta(f) \approx -0.5 + (1.0)(1.0) + (1.0)(0.0) + (0.0)(0.0) = 0.5 \geq 0 \Rightarrow y = 1$$

# فصل دهم : کرنل ها به عنوان معیار شباهت

پیش‌بینی. $y = 1$ اگر:



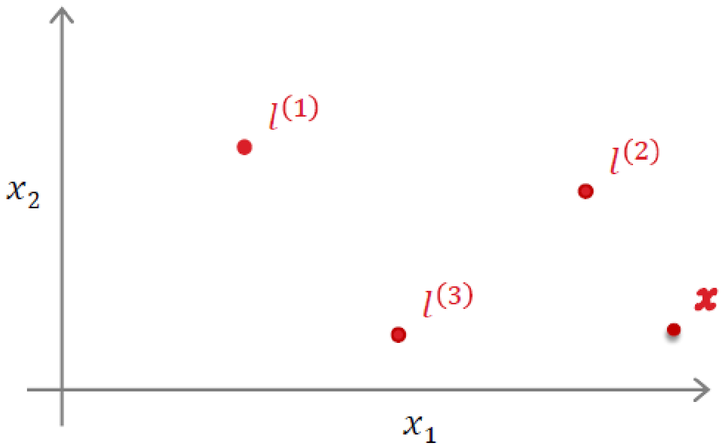

$$\underset{-0.5}{\theta_0} + \underset{1.0}{\theta_1} f_1 + \underset{1.0}{\theta_2} f_2 + \underset{0.0}{\theta_3} f_3 \geq 0$$

$$f_1 \approx f_3 \approx 0, f_2 \approx 1$$

$$h_\theta(f) \approx -0.5 + (1.0)(0.0) + (1.0)(1.0) + (0.0)(0.0) = 0.5 \geq 0 \Rightarrow y = 1$$

# فصل دهم : کرنل ها به عنوان معیار شباهت

پیش‌بینی. $y = 1$ اگر:

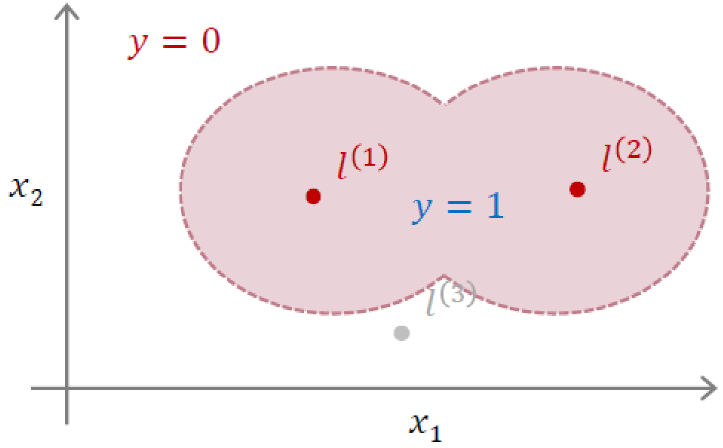



$$\underset{-0.5}{\theta_0} + \underset{1.0}{\theta_1} f_1 + \underset{1.0}{\theta_2} f_2 + \underset{0.0}{\theta_3} f_3 \geq 0$$

$$f_1 \approx f_3 \approx f_2 \approx 0$$

$$h_\theta(f) \approx -0.5 + (1.0)(0.0) + (1.0)(0.0) + (0.0)(0.0) = -0.5 \leq 0 \Rightarrow y = 0$$

# فصل دهم : کرنل ها به عنوان معیار شباهت

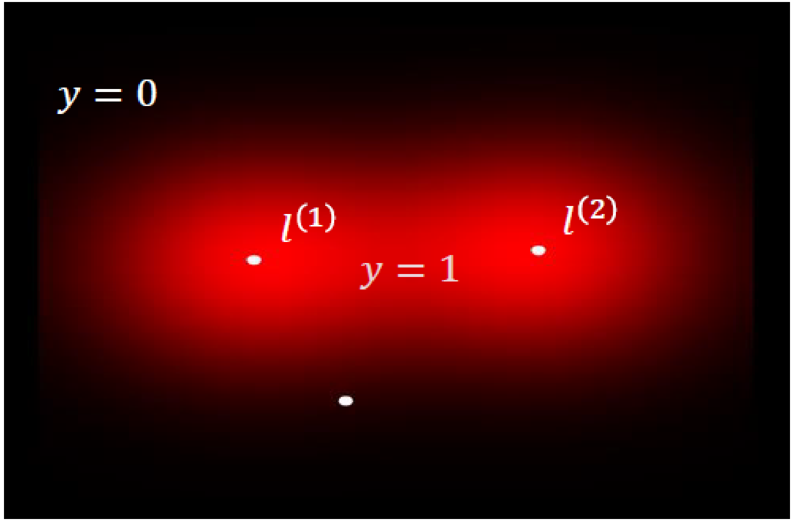



پیش‌بینی. $y = 1$ اگر:

$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$$

$-0.5$ $\quad 1.0$ $\quad 1.0$ $\quad 0.0$

- مرز تصمیم‌گیری. نقاط نزدیک به $l^{(1)}$ و $l^{(2)}$ را در کلاس ۱ و سایر نقاط را به کلاس صفر کلاس‌بندی می‌کند.




پیش‌بینی. $y = 1$ اگر:

$$\underset{-0.5}{\theta_0} + \underset{1.0}{\theta_1} f_1 + \underset{1.0}{\theta_2} f_2 + \underset{0.0}{\theta_3} f_3 \geq 0$$

- مرز تصمیم‌گیری. نقاط نزدیک به $l^{(1)}$ و $l^{(2)}$ را در کلاس ۱ و سایر نقاط را به کلاس صفر کلاس‌بندی می‌کند.

# فصل دهم : جزییات باقیمانده

- س. الگوریتم یادگیری نقاط راهنما را چگونه به صورت خودکار انتخاب می‌کند؟
- س. مقدار مناسب برای پارامترهای تابع کرنل چگونه تعیین می‌شوند؟
- س. آیا انواع دیگری از کرنل‌ها وجود دارد؟

# فصل دهم : انتخاب نقاط راهنما

- س. الگوریتم یادگیری نقاط راهنما را چگونه به صورت خودکار انتخاب می‌کند؟
  - به ازای هر نمونه در مجموعه آموزشی، یک نقطه راهنما مساوی با آن نمونه انتخاب می‌شود.

# فصل دهم : نگاشت ویژگی ها

- مجموعه آموزشی.

$$\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\cdots,(x^{(m)},y^{(m)})\}$$

- نقاط راهنما.

$$l^{(1)}=x^{(1)},\ l^{(2)}=x^{(2)},\cdots,\ l^{(m)}=x^{(m)}$$

- نگاشت فضای ویژگی.

$$x=\begin{bmatrix} x_0=1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \xrightarrow{\text{نگاشت فضای ویژگی}} f=\begin{bmatrix} f_0=1 \\ f_1=K(x,x^{(1)}) \\ f_2=K(x,x^{(2)}) \\ \vdots \\ f_m=K(x,x^{(m)}) \end{bmatrix}$$

# فصل دهم : ترفند کرنل

- تابع کرنل. پیش‌پردازش داده $x$ با استفاده از توابع کرنل:

$$\mathbf{z} = \varphi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \dots, \varphi_k(\mathbf{x}))$$

ممکن است بینهایت باشد!!!

$$g(\mathbf{z}) = \boldsymbol{\theta}^T \mathbf{z} + \theta_0$$

$$g(\mathbf{x}) = \boldsymbol{\theta}^T \varphi(\mathbf{x}) + \theta_0$$

- راه‌حل SVM.

$$\boldsymbol{\theta} = \sum_{t=1}^{m} \alpha^t y^t \mathbf{z}^t = \sum_{t=1}^{m} \alpha^t y^t \varphi(\mathbf{x}^t)$$

- کلاس‌بندی داده جدید.

$$g(\mathbf{x}) = \boldsymbol{\theta}^T \varphi(\mathbf{x}) + \theta_0 = \left(\sum_{t=1}^{m} \alpha^t y^t \varphi(\mathbf{x}^t)^T\right) \varphi(\mathbf{x}) + \theta_0 = \left(\sum_{t=1}^{m} \alpha^t y^t \varphi(\mathbf{x}^t)^T \varphi(\mathbf{x})\right) + \theta_0$$

$$g(\mathbf{x}) = \left(\sum_{t=1}^{m} \alpha^t y^t k(\mathbf{x}^t, \mathbf{x})\right) + \theta_0$$

مرز تصمیم‌گیری

داده آموزشی

داده جدید

$$L_p = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{m} \varepsilon^t$$

s.t. $$y^t \boldsymbol{\theta}^T \varphi(\boldsymbol{x}^t) \geq 1 - \varepsilon^t$$

$$\varepsilon^t \geq 0$$

$$L_p = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{m} \varepsilon^t - \sum_{t=1}^{m} \alpha^t [y^t \boldsymbol{\theta}^T \varphi(\boldsymbol{x}^t) - 1 + \varepsilon^t] - \sum_{t=1}^{m} \mu^t \varepsilon^t$$

ضرایب لاگرانژ

ضرایب لاگرانژ

# فصل دهم : توابع کرنل(مسئله اصلی)

$$L_p = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{m}\varepsilon^t - \sum_{t=1}^{m}\alpha^t[y^t\boldsymbol{\theta}^T\varphi(\boldsymbol{x}^t) - 1 + \varepsilon^t] - \sum_{t=1}^{m}\mu^t\varepsilon^t$$

$$\frac{\partial L_p}{\partial \boldsymbol{\theta}} = 0 \Rightarrow \boldsymbol{\theta} = \sum_{t=1}^{m}\alpha^t y^t \varphi(\boldsymbol{x}^t)$$

$$\frac{\partial L_p}{\partial \varepsilon^t} = 0 \Rightarrow C - \alpha^t - \mu^t = 0 \Rightarrow 0 \le \alpha^t \le C$$

# فصل دهم : توابع کرنل(مسئله دوگان)

$$L_d = -\frac{1}{2}\sum_{t=1}^{m}\sum_{s=1}^{m}\alpha^t\alpha^s y^t y^s \varphi(\boldsymbol{x}^t)^T\varphi(\boldsymbol{x}^s) + \sum_{t=1}^{m}\alpha^t$$

subject to $\sum_{t=1}^{m}\alpha^t y^t = 0$ and $0 \le \alpha^t \le C\ \forall t$

- ایده ماشین‌های کرنل. [ترفند کرنل]
  - جایگزینی حاصل‌ضرب داخلی توابع پایه با یک تابع کرنل به صورت $K(\boldsymbol{x}^t,\boldsymbol{x}^s)$

$$L_d = -\frac{1}{2}\sum_{t=1}^{m}\sum_{s=1}^{m}\alpha^t\alpha^s y^t y^s K(\boldsymbol{x}^t,\boldsymbol{x}^s) + \sum_{t=1}^{m}\alpha^t$$

ماتریس گرَم: یک ماتریس متقارن و مثبت معین (برای تفکیک‌پذیری خطی)

# فصل دهم : توابع کرنل(کرنل چند جمله ای)

- کرنل چند جمله‌ای. یک چندجمله‌ای از درجه‌ی $q$.

$$K(x^t, x) = (x^T x^t + 1)^q$$

- مثال. $[q = 2, d = 2]$

$$\begin{aligned} K(x, y) &= (x^T y + 1)^2 \\ &= (x_1 y_1 + x_2 y_2 + 1)^2 \\ &= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2 \end{aligned}$$

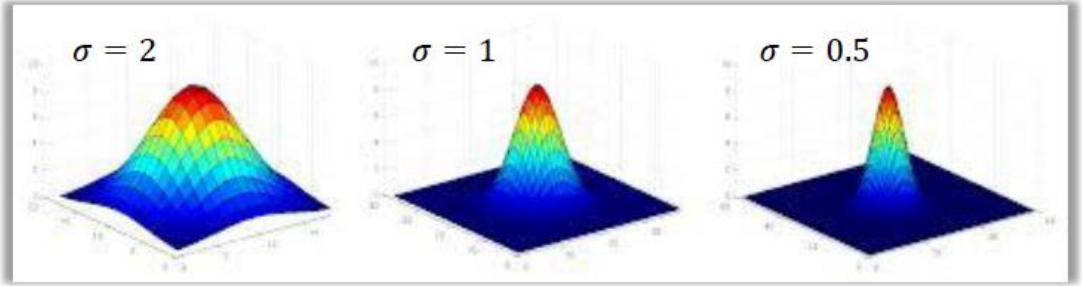
۳ ضرب، ۲ جمع

$$\varphi(x) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2]^T$$

۶ ضرب، ۵ جمع

$$\varphi(y) = [1, \sqrt{2}y_1, \sqrt{2}y_2, \sqrt{2}y_1 y_2, y_1^2, y_2^2]^T$$

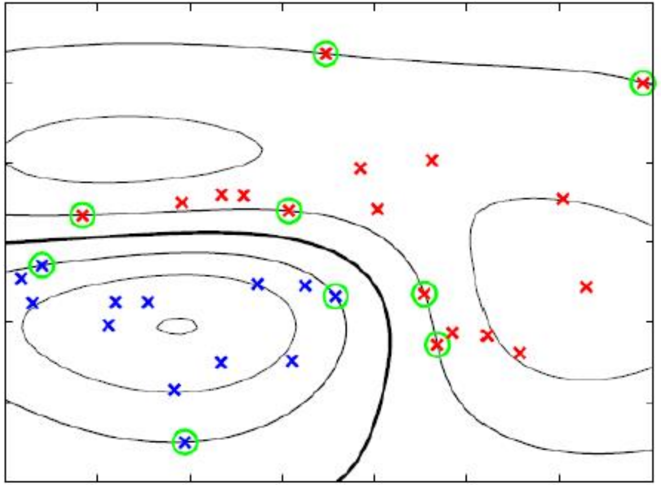
# فصل دهم : تابع کرنل گوسی



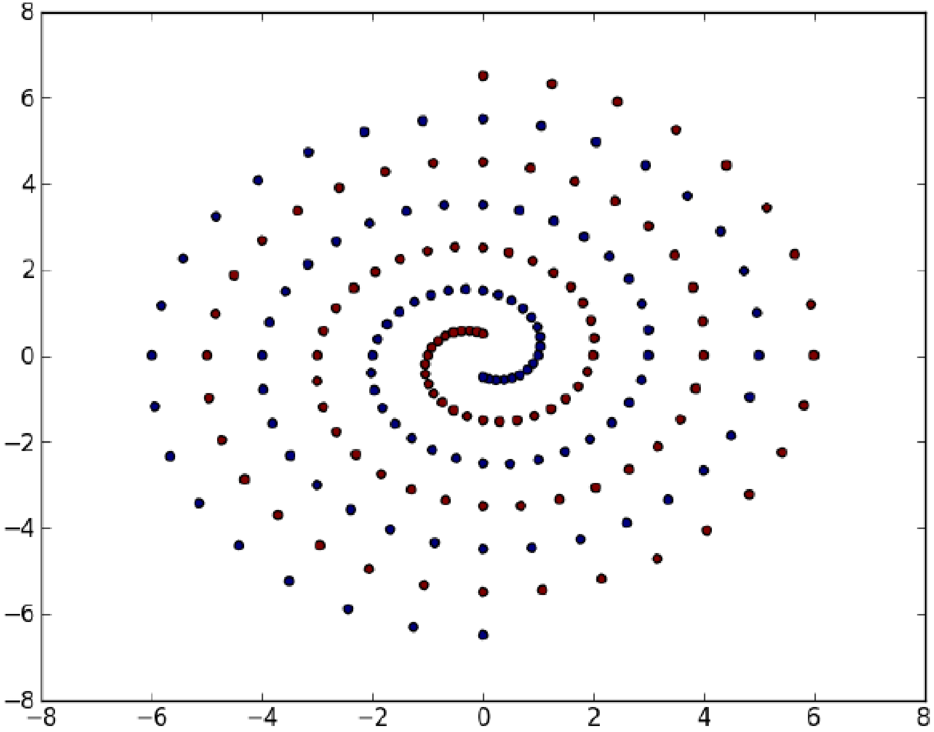

- تابع کرنل گوسی.

$$K(\boldsymbol{x}^t, \boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x}^t - \boldsymbol{x}\|^2}{2\sigma^2}\right)$$

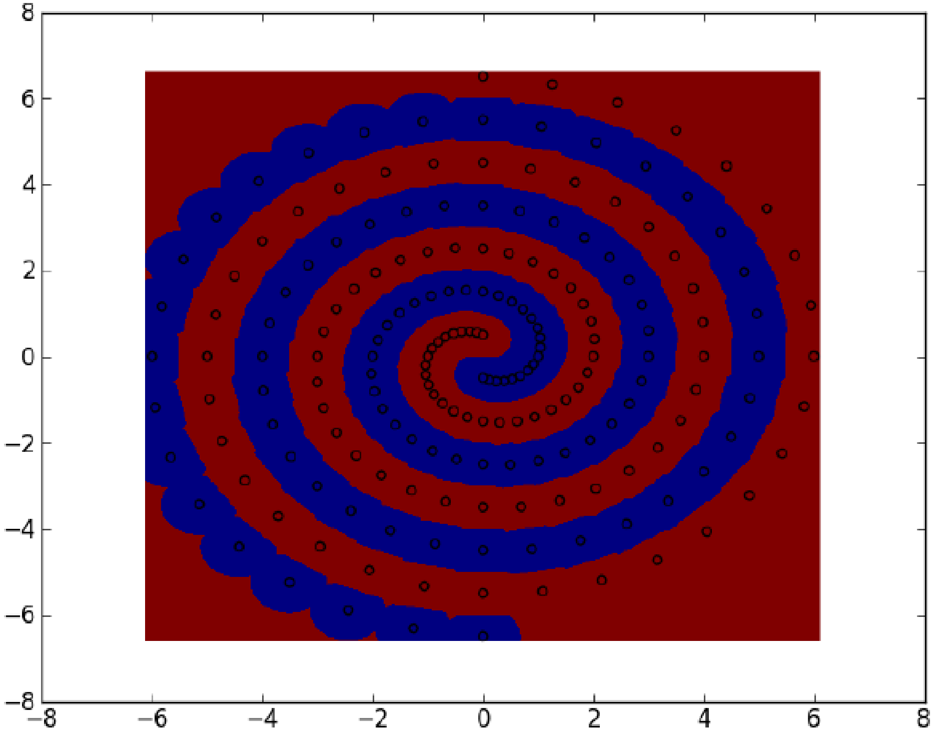
- یافتن یک مقدار مناسب برای $\sigma$.
  - با استفاده از مجموعه اعتبارسنجی [انتخاب مدل]
  - مقادیر بزرگ‌تر: مرز تصمیم‌گیری هموارتر



# فصل دهم : مثال(تابع کرنل گوسی)



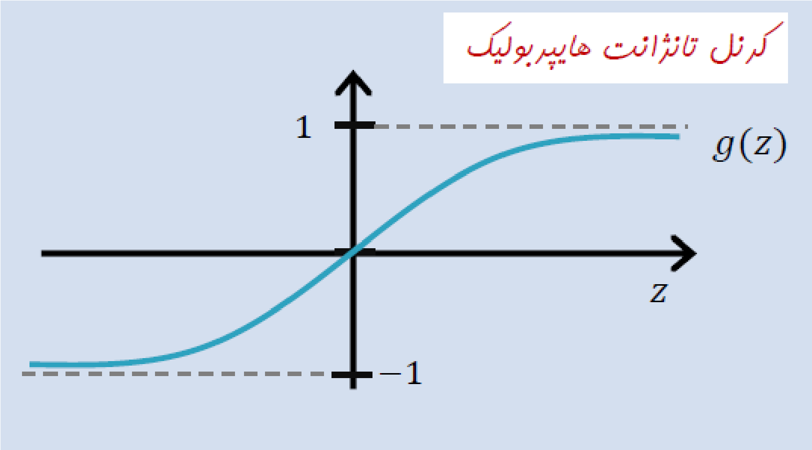
# فصل دهم : مثال(تابع کرنل گوسی)



# فصل دهم : توابع کرنل(انواع دیگر)

- تابع کرنل. معیاری به منظور محاسبه شباهت میان داده‌های $x$ و $y$

- انواع دیگر.
  - کرنل تانژانت هایپربولیک
  - کرنل رشته‌ای
  - کرنل درختی
  - کرنل گرافی
  - ...




$$K(x^t, x) = \tanh(2x^T x^t + 1)$$

# فصل دهم : پارامترهای SVM

- س. مقدار مناسب برای پارامترهای تابع کرنل چگونه تعیین می‌شوند؟

- پارامتر $C$.
  - مقادیر کوچک‌تر: بایاس بیشتر، واریانس کمتر
  - مقادیر بزرگ‌تر: بایاس کمتر، واریانس بیشتر

- پارامتر $\sigma$.
  - مقادیر کوچک‌تر: بایاس کمتر، واریانس بیشتر
  - مقادیر بزرگ‌تر: بایاس بیشتر، واریانس کمتر

تعیین مقادیر هر دو پارامتر:
جستجوی شبکه‌بندی

# فصل دهم : کلاس بندی چندکلاسی

- روش اول. یکی در برابر همه [روش ترجیحی]
  - آموزش: آموزش $k$ ماشین بردار پشتیبان یکی به ازای هر کلاس
  - آزمایش: محاسبه $g_i(x)$ به ازای $0 \leq i \leq k$ و انتخاب بزرگ‌ترین مقدار

$$\arg\max_i g_i(x)$$

- روش دوم. جداسازی دو به دو
  - آموزش $k(k-1)/2$ ماشین بردار پشتیبان به طوری که $g_{ij}(x)$ نمونه های دو کلاس $C_i$ و $C_j$ را از هم جدا می‌کند.
  - ساده‌تر و سریع‌تر
- روش سوم. حل یک مسئله بهینه‌سازی چند کلاسی

$$\min \frac{1}{2}\sum_{i=1}^{k}\|\boldsymbol{\theta}_{\boldsymbol{i}}\|^2 + C\sum_i\sum_t \varepsilon_i^t$$

$$\text{s.t.} \quad \boldsymbol{\theta}_{z^t}\boldsymbol{x}^t + \theta_{z^t 0} \geq \boldsymbol{\theta}_i\boldsymbol{x}^t + \theta_{i0} + 2 - \varepsilon_i^t, \forall i \neq z^t$$

# فصل دهم : راهنمای استفاده از SVM

- پیاده‌سازی. استفاده از بسته‌های نرم‌افزاری موجود مانند $\text{SVM}^{\text{light}}$ و LIBSVM

- تعیین تابع کرنل.
  - کرنل خطی (عدم استفاده از کرنل): وقتی که $n \gg m$
  - گوسی، چندجمله‌ای، رشته‌ای و …

- تعیین مقدار پارامترها. جستجوی شبکه‌بندی
  - انتخاب مقدار برای پارامتر $C$
  - انتخاب مقدار برای پارامترهای تابع کرنل (مانند $\sigma$)

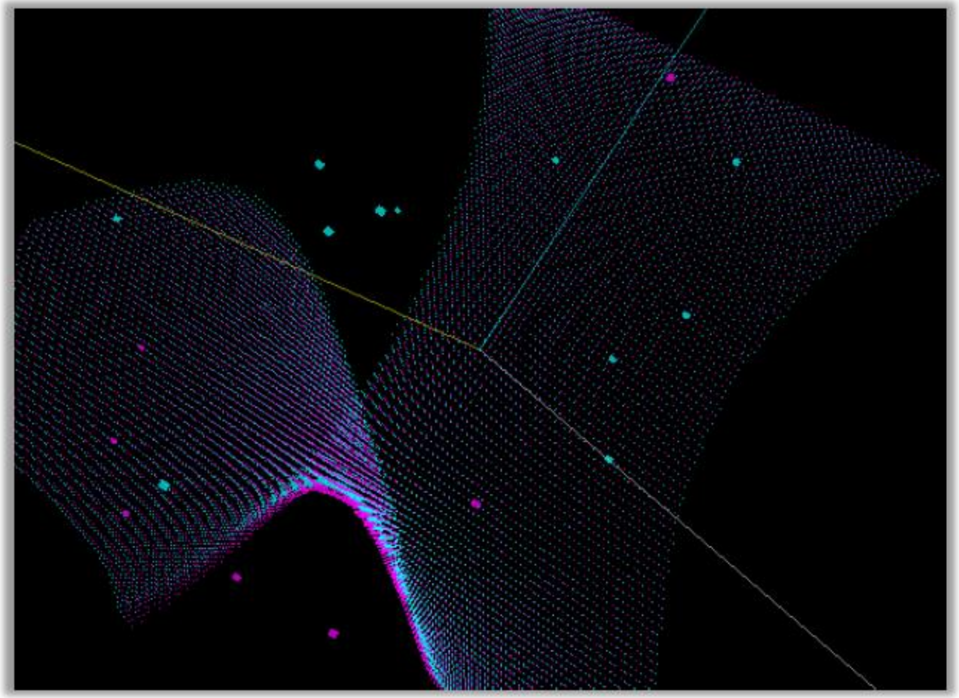
| | $\sigma$ | | | | | |
|---|---|---|---|---|---|---|
| $C$ | err | .01 | .1 | 1 | 10 | 100 |
| | .01 | | | | | |
| | .1 | | | | | |
| | 1 | | | | | |
| | 10 | | | | | |
| | 100 | | | | | |

# فصل دهم : SVM ، رگرسیون لجستیک یا شبکه عصبی؟

- حالت ۱. $[n \gg m]$
  - مثال: تشخیص هرزنامه (۱۰۰۰ نمونه آموزشی، ۵۰۰۰۰ ویژگی)
  - رگرسیون لجستیک یا SVM خطی
- حالت ۲. [تعداد ویژگی‌ها کم، تعداد نمونه‌های آموزشی زیاد]
  - SVM با کرنل گوسی
- توجه. شبکه‌های عصبی در تمامی حالت‌های فوق قابل استفاده هستند، اما ممکن است به زمان بیشتری برای آموزش نیاز داشته باشند.

# فصل دهم : اجرای نمایشی

- SVM toy 3D. [http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/svmtoy3d/]



# فصل دهم : مطالب بیشتر در مورد کرنل ها

- س. چگونه می‌دانیم استفاده از کرنل‌ها در جداسازی داده‌ها به ما کمک می‌کند؟
- در فضای $n$–بعدی، هر مجموعه از $n$ بردار مستقل، به صورت خطی تفکیک‌پذیر هستند.
- اگر ماتریس $K$ یک ماتریس مثبت معین باشد، آنگاه داده‌ها به صورت خطی تفکیک‌پذیر هستند.

- قضیه. ماتریس $K$ یک ماتریس مثبت معین است، زیرا $K = L^T L$
  - ستون $i$ در ماتریس $L$ برابر است با بردار $\phi(x^{(i)})$

- اثبات. بردار غیر صفر $v$ را در نظر بگیرید. در این صورت:

$$v^T K v = v^T L^T L v = (Lv)^T (Lv) = w^T w = \|w\|^2 \geq 0$$

- و چون $L$ و $v$ هر دو مخالف صفر هستند، بردار $w$ نیز مخالف صفر است. یعنی:

$$\|w\|^2 > 0 \Rightarrow v^T K v > 0 \Rightarrow K \text{ is positive definite}$$

# با تشکر از توجه شما