ماشین های بردار پشتیبان

Support Vector Machines

مجدد در نظر می‌گیریم که فقط دو کلاس و دسته وجود دارد $y_i \in \{+1, -1\}$

معادله‌ی ابرصفحه (Hyperplane)

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = w_1x_1 + w_2x_2 + \cdots + w_dx_d + b$$

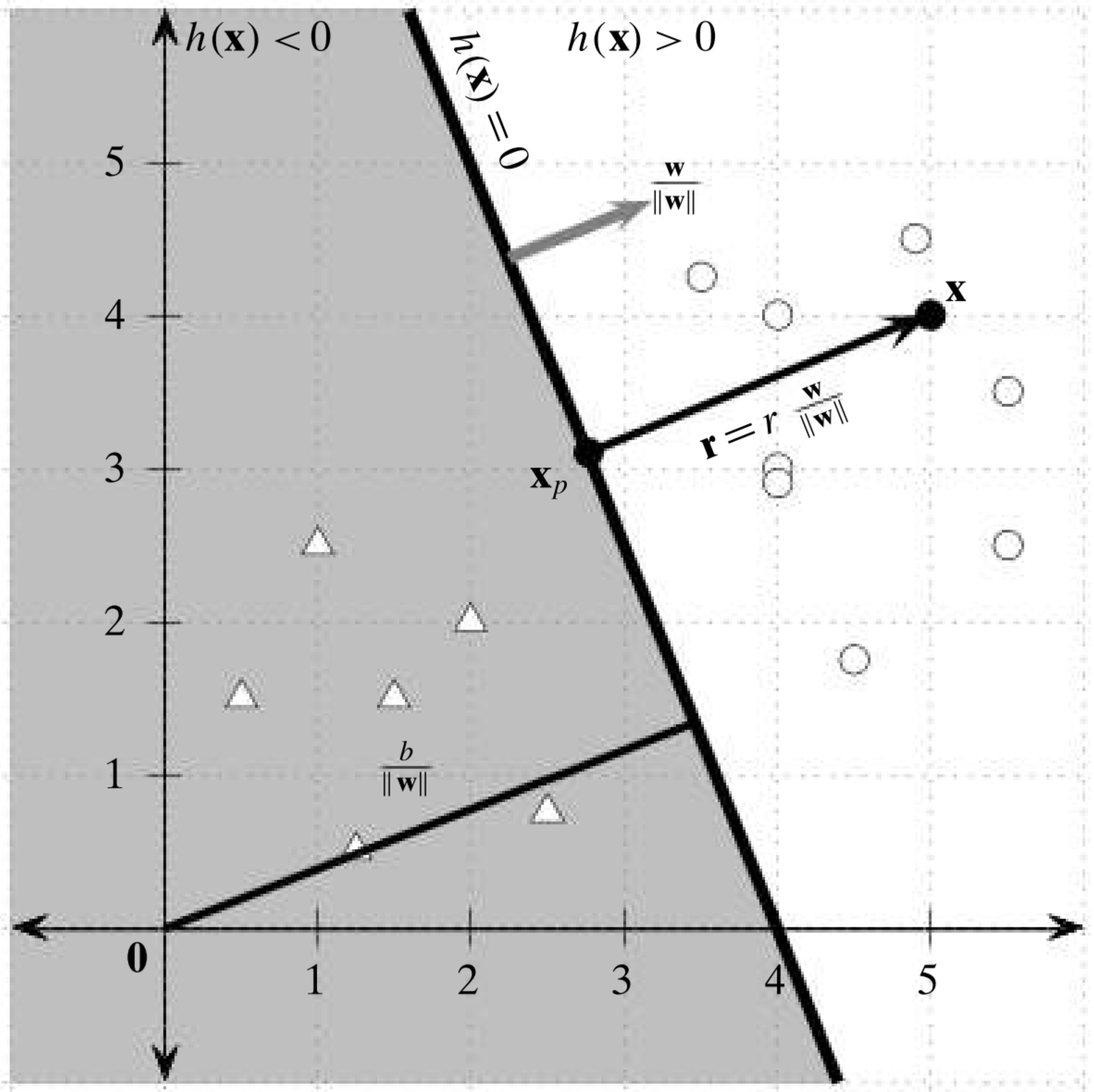
$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$$

بردار $\boldsymbol{w}$ دارای d بعد است و عمود بر ابرصفحه می‌باشد.
b یک عدد است که پیشقدر (Bias) می‌نامیم.

ابرصفحه فضای d بعدی را به دو نیم‌فضا تقسیم می‌کند.

اگر یک مجموعه داده را بتوان با یک ابرصفحه به دو نیم‌فضا تقسیم کرد که در هر نیم‌فضا فقط داده‌های یک کلاس یا دسته باشند، آن مجموعه داده را تفکیک‌پذیر خطی (Linearly Separable) می‌نامیم.

$$y = \begin{cases} +1 & \text{if } h(\mathbf{x}) > 0 \\ -1 & \text{if } h(\mathbf{x}) < 0 \end{cases}$$



فاصله‌ی یک نقطه از یک ابرصفحه:

سایه‌ی نقطه‌ی $\boldsymbol{x}$ روی ابرصفحه را با $\boldsymbol{x}_p$ نمایش می‌دهیم.

$$\boldsymbol{x} = \boldsymbol{x}_p + \boldsymbol{r} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

$$\begin{aligned} h(\mathbf{x}) &= h\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) \\ &= \mathbf{w}^T\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + b \\ &= \underbrace{\mathbf{w}^T\mathbf{x}_p + b}_{h(\mathbf{x}_p)} + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} \\ &= \underbrace{h(\mathbf{x}_p)}_{0} + r\,\|\mathbf{w}\| \\ &= r\,\|\mathbf{w}\| \end{aligned}$$

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

برای اینکه δ همواره مثبت باشد باید y برچسب کلاس را در r ضرب کرد.

$$\boxed{\delta = y\,r = \frac{y\,h(\mathbf{x})}{\|\mathbf{w}\|}}$$

برای مبدا مختصات $\boldsymbol{x} = 0$ فاصله از رابطه‌ی زیر محاسبه می‌شود.

$$r = \frac{h(\mathbf{0})}{\|\mathbf{w}\|} = \frac{\mathbf{w}^T\mathbf{0} + b}{\|\mathbf{w}\|} = \frac{b}{\|\mathbf{w}\|}$$

فاصله‌ی هر نقطه (بردار) $\boldsymbol{x}_i$ از ابرصفحه‌ی $h(\boldsymbol{x})$

$$\delta_i = \frac{y_i\, h(\mathbf{x}_i)}{\|\mathbf{w}\|} = \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|}$$

حاشیه‌ی (Margin) یک دسته‌بندی‌گر خطی فاصله‌ی نزدیک‌ترین نقطه یا بردار از ابرصفحه می‌باشد.

$$\boxed{\delta^* = \min_{\mathbf{x}_i} \left\{ \frac{y_i\,(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|} \right\}}$$

بردار پشتیبان (Support Vector) $\boldsymbol{x}^*$ نقطه‌ای است که دقیقا در فاصله‌ی حاشیه‌ی یک دسته‌بندی‌گر قرار دارد.

$$\delta^* = \frac{y^*(\mathbf{w}^T\mathbf{x}^* + b)}{\|\mathbf{w}\|}$$

با توجه به اینکه ضرب یک عدد s در معادله‌ی یک ابرصفحه آن را تغییر نمی‌دهد، با انتخاب یک s خاص، می‌توان یک ابرصفحه‌ی کانونیک (Canonical Hyperplane) تعریف نمود که فاصله‌ی مطلق (Absolute Distance) آن از بردار پشتیبان ۱ باشد.

$$s\,h(\mathbf{x}) = s\,\mathbf{w}^T\mathbf{x} + s\,b = (s\mathbf{w})^T\mathbf{x} + (sb) = 0$$

$$sy^*(\mathbf{w}^T\mathbf{x}^* + b) = 1 \qquad s = \frac{1}{y^*(\mathbf{w}^T\mathbf{x}^* + b)} = \frac{1}{y^*h(\mathbf{x}^*)}$$

برای ابرصفحه‌ی کانونیک حاشیه می‌شود: $\delta^* = \frac{y^*h(\mathbf{x}^*)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$

برای هر بردار پشتیبان داریم: $y^*h(\boldsymbol{x}^*) = 1$
و در حالت کلی برای هر نقطه داریم:

$$\boxed{y_i\,(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \text{ for all points } \mathbf{x}_i \in \mathbf{D}}$$

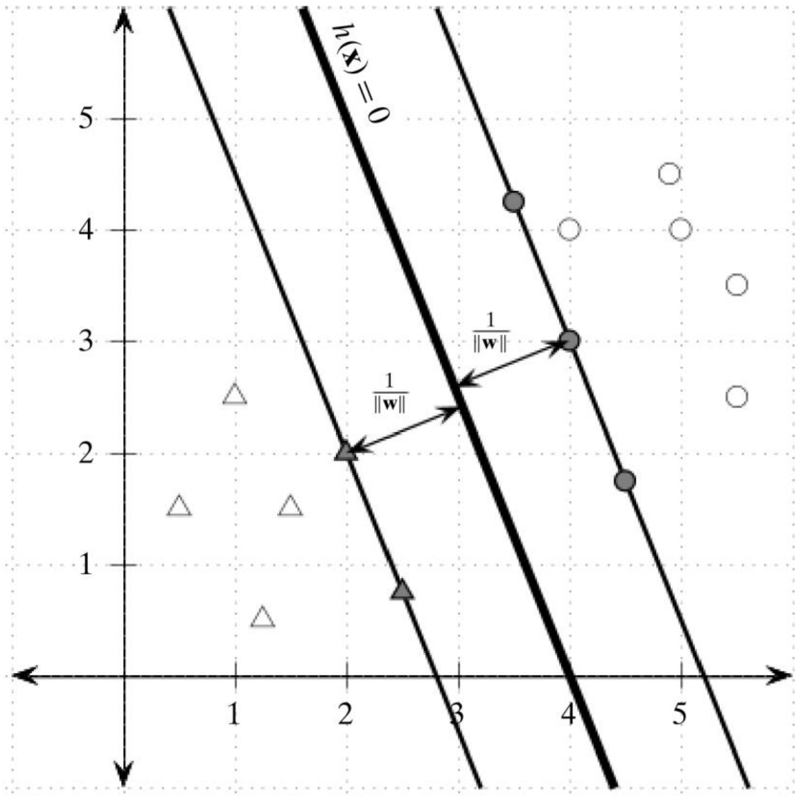


Figure 21.2. Margin of a separating hyperplane: $\frac{1}{\|\mathbf{w}\|}$ is the margin, and the shaded points are the support vectors.

a.fahim@ut.ac.ir

$$h^* = \arg\max_h \left\{ \delta_h^* \right\} = \arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \right\}$$

Objective Function: $\min_{\mathbf{w},b} \left\{ \frac{\|\mathbf{w}\|^2}{2} \right\}$

Linear Constraints: $y_i \, (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \;\; \forall \mathbf{x}_i \in \mathbf{D}$

شرایط کاروش-کوهن-تاکر (Karush–Kuhn–Tucker (KKT

$$\alpha_i \left(y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1 \right) = 0$$

$$\text{and } \alpha_i \geq 0$$

$$\min L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1 \right)$$

$$\frac{\partial}{\partial \mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{n} \alpha_i \, y_i \mathbf{x}_i = \mathbf{0} \quad \text{or} \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial}{\partial b} L = \sum_{i=1}^{n} \alpha_i y_i = 0$$

لاگرانژی دوگان (Dual Lagrangian)

$$L_{dual} = \frac{1}{2}\mathbf{w}^T \mathbf{w} - \mathbf{w}^T \underbrace{\left(\sum_{i=1}^{n} \alpha_i \, y_i \, \mathbf{x}_i \right)}_{\mathbf{w}} - b \underbrace{\sum_{i=1}^{n} \alpha_i \, y_i}_{0} + \sum_{i=1}^{n} \alpha_i$$

$$= -\frac{1}{2}\mathbf{w}^T \mathbf{w} + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i \, y_j \mathbf{x}_i^T \mathbf{x}_j$$

Objective Function: $\max_{\boldsymbol{\alpha}} \; L_{dual} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$

Linear Constraints: $\alpha_i \geq 0, \; \forall i \in \mathbf{D}, \;$ and $\sum_{i=1}^{n} \alpha_i \, y_i = 0$

لاگرانژی دوگان محدب است و با روش‌های استاندارد قابل بهینه‌سازی برای بدست آورد α_i بهینه است.

بردار وزن و پیش‌قدر(Weight Vector and Bias)

بر مبنای شرایط KKT رابط زیر را بدست می‌آوریم.

$$\alpha_i \left(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \right) = 0$$

(1) $\alpha_i = 0$, or

(2) $y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 = 0$, which implies $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$

یک نقطه یا بردار پشتیبان است و یا $\alpha_i = 0$

بعد از بدست آوردن α_iها می‌توان بردارهای وزن $\boldsymbol{w}$ و پیش‌قدر b را محاسبه کرد.

$$\boxed{\mathbf{w} = \sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i}$$

$$\boxed{b_i = \frac{1}{y_i} - \mathbf{w}^T\mathbf{x}_i = y_i - \mathbf{w}^T\mathbf{x}_i}$$

$$\boxed{b = \underset{\alpha_i > 0}{\text{avg}} \{b_i\}}$$

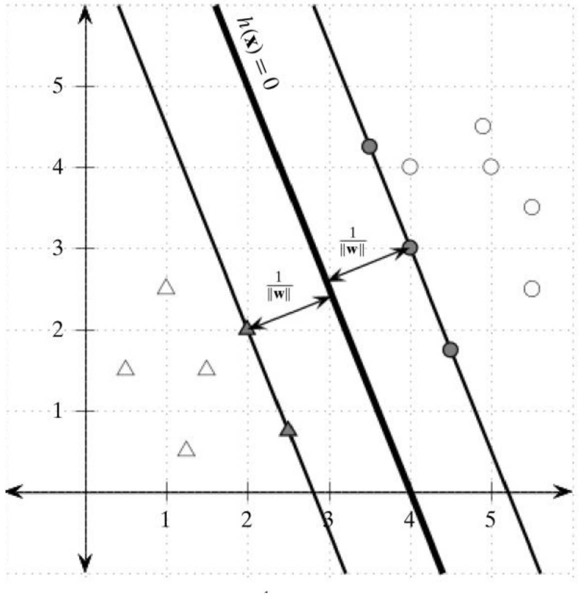
SVM Classifier

$$\boxed{\hat{y} = \text{sign}(h(\mathbf{z})) = \text{sign}(\mathbf{w}^T\mathbf{z} + b)}$$

حالت تفکیک‌پذیر و خطی (SVM: Linear and Separable Case)

مثال: برای داده‌های جدول SVM Classifier را بدست آورید.

$\mathbf{x}_i^T$	x_{i1}	x_{i2}	y_i
$\mathbf{x}_1^T$	3.5	4.25	+1
$\mathbf{x}_2^T$	4	3	+1
$\mathbf{x}_3^T$	4	4	+1
$\mathbf{x}_4^T$	4.5	1.75	+1
$\mathbf{x}_5^T$	4.9	4.5	+1
$\mathbf{x}_6^T$	5	4	+1
$\mathbf{x}_7^T$	5.5	2.5	+1
$\mathbf{x}_8^T$	5.5	3.5	+1
$\mathbf{x}_9^T$	0.5	1.5	−1
$\mathbf{x}_{10}^T$	1	2.5	−1
$\mathbf{x}_{11}^T$	1.25	0.5	−1
$\mathbf{x}_{12}^T$	1.5	1.5	−1
$\mathbf{x}_{13}^T$	2	2	−1
$\mathbf{x}_{14}^T$	2.5	0.75	−1



با کمینه کردن لاگرانژی دوگان همه‌ی α_iها صفر می‌شوند بجز برای نقاط این جدول.

$\mathbf{x}_i^T$	x_{i1}	x_{i2}	y_i	α_i
$\mathbf{x}_1^T$	3.5	4.25	+1	0.0437
$\mathbf{x}_2^T$	4	3	+1	0.2162
$\mathbf{x}_4^T$	4.5	1.75	+1	0.1427
$\mathbf{x}_{13}^T$	2	2	−1	0.3589
$\mathbf{x}_{14}^T$	2.5	0.75	−1	0.0437

برای محاسبه‌ی $\mathbf{w}$ و b فقط از α_iهای مخالف صفر (بردارهای پشتیبان) استفاده می‌شود.

$$\mathbf{w} = \sum_{\alpha_i>0} \alpha_i y_i \mathbf{x}_i$$
$$= 0.0437\begin{pmatrix}3.5\\4.25\end{pmatrix} + 0.2162\begin{pmatrix}4\\3\end{pmatrix} + 0.1427\begin{pmatrix}4.5\\1.75\end{pmatrix} - 0.3589\begin{pmatrix}2\\2\end{pmatrix} - 0.0437\begin{pmatrix}2.5\\0.75\end{pmatrix}$$
$$= \begin{pmatrix}0.833\\0.334\end{pmatrix}$$

برای محاسبه‌ی b بین b_iهای محاسبه شده از بردارهای پشتیبان، میانگین گرفته می‌شود.

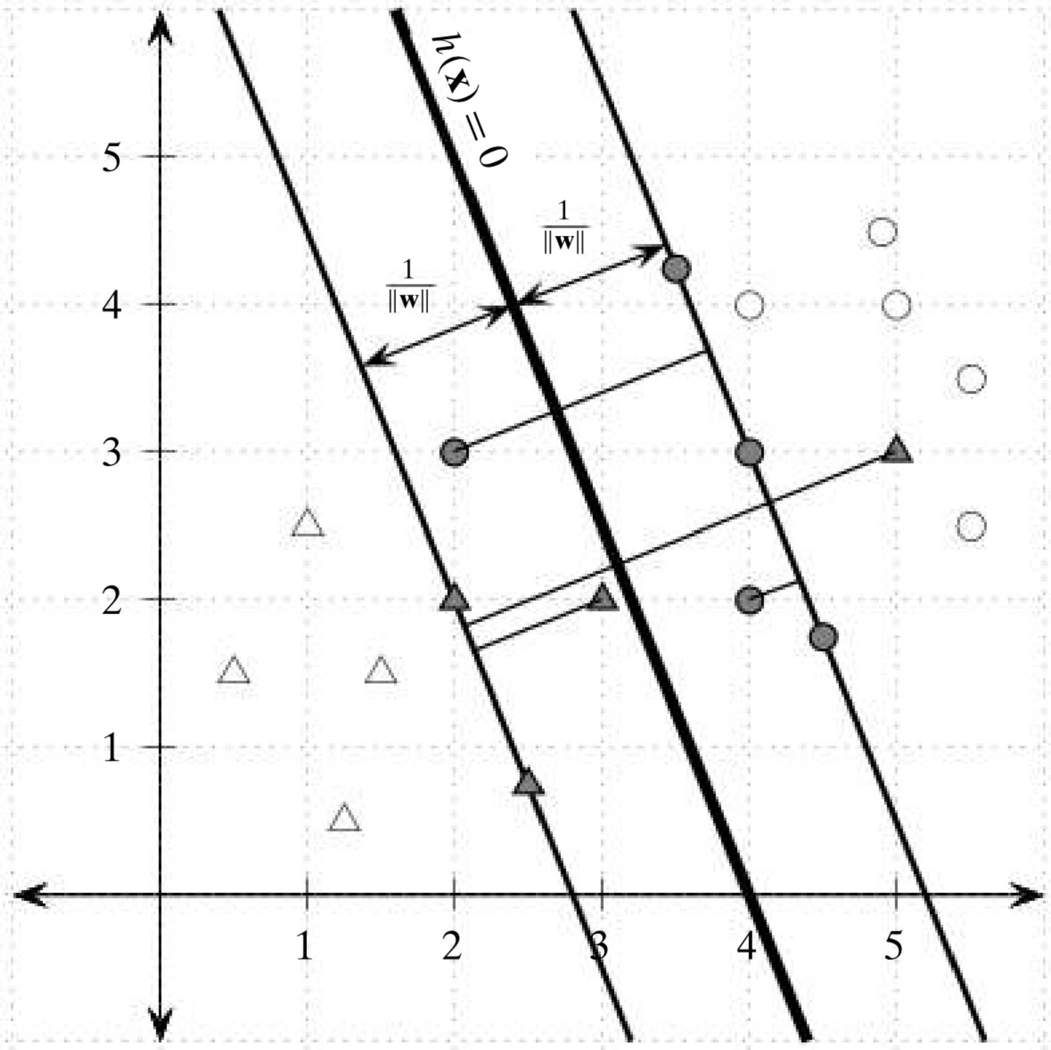
$\mathbf{x}_i$	$\mathbf{w}^T\mathbf{x}_i$	$b_i = y_i - \mathbf{w}^T\mathbf{x}_i$
$\mathbf{x}_1$	4.332	−3.332
$\mathbf{x}_2$	4.331	−3.331
$\mathbf{x}_4$	4.331	−3.331
$\mathbf{x}_{13}$	2.333	−3.333
$\mathbf{x}_{14}$	2.332	−3.332
$b = \text{avg}\{b_i\}$		−3.332

با داشتن $\mathbf{w}$ و b ابرصفحه‌ی دسته‌بندی کننده محاسبه می‌شود. علامت این معادله بازای هر نقطه، کلاس آن نقطه را مشخص می‌کند.

$$h(\mathbf{x}) = \begin{pmatrix}0.833\\0.334\end{pmatrix}^T \mathbf{x} - 3.332 = 0$$

در حالت‌های خطی و تفکیک ناپذیر، نقاطی در مرز دو کلاس در هم نفوذ می‌کنند و عملا تعریف ابرصفحه‌ای که دو کلاس را کاملا از هم جدا کند امکان‌پذیر نیست. پس در تعریف حاشیه، اجازه می‌دهیم نقاطی درون آن نفوذ کنند. برای مدیریت این نقاط میانی، متغیرهای شل!! ξ_i (Slack Variables) را در محاسبات وارد می‌کنیم تا $y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)$ بتواند مقادیر کمتر از 1 را برای نقاط درون حاشیه بگیرد.

$$y_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1-\xi_i$$



جمع متغیر ξ_iها که میزان انحراف از جدایش کلاس‌ها است باید کمینه باشد.
پس در تابع هزینه به شکل $\sum_{i=1}^{n}(\xi_i)^k$ وارد می‌شود.
استفاده از دو مقدار $k=1$ هزینه‌ی لولا!! (Hinge Loss) و $k=2$ هزینه‌ی درجه دو (Quadratic Loss) بسیار مرسوم است.

Objective Function: $\min_{\mathbf{w},b,\xi_i}\left\{\frac{\|\mathbf{w}\|^2}{2}+C\sum_{i=1}^{n}(\xi_i)^k\right\}$

Linear Constraints: $y_i\ (\mathbf{w}^T\mathbf{x}_i+b)\geq 1-\xi_i,\ \ \forall\mathbf{x}_i\in\mathbf{D}$

$\xi_i\geq 0\ \ \forall\mathbf{x}_i\in\mathbf{D}$

تابع زیان هینج یا لولا!! (Hinge Loss)

شرایط KKT در پاسخ بهینه

$$\alpha_i\left(y_i(\mathbf{w}^T\mathbf{x}_i+b)-1+\xi_i\right)=0 \text{ with } \alpha_i\geq 0$$

$$\beta_i(\xi_i-0)=0 \text{ with } \beta_i\geq 0$$

$$L=\frac{1}{2}\|\mathbf{w}\|^2+C\sum \xi_i-\sum \alpha_i\left(y_i(\mathbf{w}^T\mathbf{x}_i+b)-1+\xi_i\right)-\sum_{i=1}^{n}\beta_i\xi_i$$

$$\frac{\partial}{\partial \mathbf{w}}L=\mathbf{w}-\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i=\mathbf{0} \quad \text{or} \quad \mathbf{w}=\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i$$

$$\frac{\partial}{\partial b}L=\sum_{i=1}^{n}\alpha_i y_i=0$$

$$\frac{\partial}{\partial \xi_i}L=C-\alpha_i-\beta_i=0 \quad \text{or} \quad \beta_i=C-\alpha_i$$

از سه شرط $\alpha_i \geq 0$ و $\beta_i \geq 0$ و $\alpha_i+\beta_i=C$ نتیجه می‌گیریم: $0\geq \alpha_i \geq C$

$$L_{dual}=\frac{1}{2}\mathbf{w}^T\mathbf{w}-\mathbf{w}^T\underbrace{\left(\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i\right)}_{\mathbf{w}}-b\underbrace{\sum_{i=1}^{n}\alpha_i y_i}_{0}+\sum_{i=1}^{n}\alpha_i+\sum_{i=1}^{n}\underbrace{(C-\alpha_i-\beta_i)}_{0}\xi_i$$

$$=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j$$

مقدار α_i ها از بیشینه کردن لاگرانژی دوگان محاسبه می‌شوند. پاسخ‌ها باید در شرط‌های KKT نیز صدق کنند. تنها برای بردارهای پشتیبان $\alpha_i > 0$ بدست می‌آید. بردارهای پشتیبان شامل نقاط روی خط حاشیه با $\xi_i=0$ و درون حاشیه با $\xi_i>0$ می‌شوند.

$$\textbf{Objective Function: } \max_{\boldsymbol{\alpha}} \; L_{dual}=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j$$

$$\textbf{Linear Constraints: } 0\leq\alpha_i\leq C,\ \forall i\in\mathbf{D} \text{ and } \sum_{i=1}^{n}\alpha_i y_i=0$$

بردار وزن و پیش‌قدر

$$\mathbf{w}=\sum_{\alpha_i>0}\alpha_i y_i\mathbf{x}_i$$

از شرط KKT می‌دانیم $\beta_i \geq 0$ و $\beta_i(\xi_i-0)=0$ است. که ایجاب می‌کند یا $\xi_i=0$ (بردارهای پشتیبان روی حاشیه) و یا $\alpha_i=C$ $(\beta_i=0)$ برای $\xi_i>0$ (بردارهای پشتیبان درون حاشیه) باشد.

$$\alpha_i\left(y_i(\mathbf{w}^T\mathbf{x}_i+b_i)-1\right)=0$$

$$y_i(\mathbf{w}^T\mathbf{x}_i+b_i)=1$$

$$b_i=\frac{1}{y_i}-\mathbf{w}^T\mathbf{x}_i=y_i-\mathbf{w}^T\mathbf{x}_i$$

$$\hat{y}=\text{sign}(h(\mathbf{z}))=\text{sign}(\mathbf{w}^T\mathbf{z}+b)$$

مثال: نقاط $\boldsymbol{x}_{15}$ تا $\boldsymbol{x}_{18}$ که در حاشیه قرار دارند به مجموعه داده‌ها اضافه می‌شود.

$\mathbf{x}_i$	x_{i1}	x_{i2}	y_i
$\mathbf{x}_{15}^T$	4	2	$+1$
$\mathbf{x}_{16}^T$	2	3	$+1$
$\mathbf{x}_{17}^T$	3	2	-1
$\mathbf{x}_{18}^T$	5	3	-1

مقدار $k = 1$ و $C = 1$ انتخاب می‌کنیم و لاگرانژین دوگان ($\mathcal{L}_{dual}$) را حل می‌کنیم.

$\mathbf{x}_i^T$	x_{i1}	x_{i2}	y_i
$\mathbf{x}_1^T$	3.5	4.25	$+1$
$\mathbf{x}_2^T$	4	3	$+1$
$\mathbf{x}_3^T$	4	4	$+1$
$\mathbf{x}_4^T$	4.5	1.75	$+1$
$\mathbf{x}_5^T$	4.9	4.5	$+1$
$\mathbf{x}_6^T$	5	4	$+1$
$\mathbf{x}_7^T$	5.5	2.5	$+1$
$\mathbf{x}_8^T$	5.5	3.5	$+1$
$\mathbf{x}_9^T$	0.5	1.5	-1
$\mathbf{x}_{10}^T$	1	2.5	-1
$\mathbf{x}_{11}^T$	1.25	0.5	-1
$\mathbf{x}_{12}^T$	1.5	1.5	-1
$\mathbf{x}_{13}^T$	2	2	-1
$\mathbf{x}_{14}^T$	2.5	0.75	-1

برای 9 نقطه (بردار پشتیبان) $\alpha_i \neq 0$ است و برای الباقی نقاط $\alpha_i = 0$.

5 نقطه‌ی بردار پشتیبان روی حاشیه هستند ($\xi_i = 0$) و
4 بردار پشتیبان دیگر درون حاشیه با $\alpha_i = C = 1$ ($\xi_i > 0$) می‌باشند.

$\mathbf{x}_i$	x_{i1}	x_{i2}	y_i	α_i
$\mathbf{x}_1$	3.5	4.25	$+1$	0.0271
$\mathbf{x}_2$	4	3	$+1$	0.2162
$\mathbf{x}_4$	4.5	1.75	$+1$	0.9928
$\mathbf{x}_{13}$	2	2	-1	0.9928
$\mathbf{x}_{14}$	2.5	0.75	-1	0.2434
$\mathbf{x}_{15}$	4	2	$+1$	1
$\mathbf{x}_{16}$	2	3	$+1$	1
$\mathbf{x}_{17}$	3	2	-1	1
$\mathbf{x}_{18}$	5	3	-1	1

برای محاسبه $\boldsymbol{w}$ از هر 9 نقطه بردار پشتیبان استفاده می‌کنیم.

$$\begin{aligned}\mathbf{w} &= \sum_{\alpha_i>0} \alpha_i y_i \mathbf{x}_i \\ &= 0.0271\begin{pmatrix}3.5\\4.25\end{pmatrix} + 0.2162\begin{pmatrix}4\\3\end{pmatrix} + 0.9928\begin{pmatrix}4.5\\1.75\end{pmatrix} - 0.9928\begin{pmatrix}2\\2\end{pmatrix} \\ &\quad - 0.2434\begin{pmatrix}2.5\\0.75\end{pmatrix} + \begin{pmatrix}4\\2\end{pmatrix} + \begin{pmatrix}2\\3\end{pmatrix} - \begin{pmatrix}3\\2\end{pmatrix} - \begin{pmatrix}5\\3\end{pmatrix} \\ &= \begin{pmatrix}0.834\\0.333\end{pmatrix}\end{aligned}$$

برای محاسبه‌ی پیش‌قدر b از بردارهای پشتیبان درون حاشیه ($\xi_i \neq 0$) که $\alpha_i = C = 1$ است، استفاده نمی‌کنیم.

$\mathbf{x}_i$	$\mathbf{w}^T\mathbf{x}_i$	$b_i = y_i - \mathbf{w}^T\mathbf{x}_i$
$\mathbf{x}_1$	4.334	-3.334
$\mathbf{x}_2$	4.334	-3.334
$\mathbf{x}_4$	4.334	-3.334
$\mathbf{x}_{13}$	2.334	-3.334
$\mathbf{x}_{14}$	2.334	-3.334
$b = \mathrm{avg}\{b_i\}$		-3.334

برای تمام نقاطی که بردار پشتیبان نیستند و یا بردارهای پشتیبان روی حاشیه $\xi_i = 0$ و برای بردارهای پشتیبان درون حاشیه مقدار ξ_i را می‌توانیم محاسبه کنیم.

$$\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

برای نقاط اشتباه کلاس‌بندی شده $\xi_i > 1$ (نقاط $\boldsymbol{x}_{16}$ و $\boldsymbol{x}_{18}$) و برای نقاطی که درست تشخیص داده شده‌اند اما درون حاشیه قرار دارند $0 < \xi_i < 1$ است.

$\mathbf{x}_i$	$\mathbf{w}^T\mathbf{x}_i$	$\mathbf{w}^T\mathbf{x}_i + b$	$\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)$
$\mathbf{x}_{15}$	4.001	0.667	0.333
$\mathbf{x}_{16}$	2.667	-0.667	1.667
$\mathbf{x}_{17}$	3.167	-0.167	0.833
$\mathbf{x}_{18}$	5.168	1.834	2.834

تابع زیان درجه‌ی دو (Quadratic Loss)

$$\textbf{Objective Function: } \min_{\mathbf{w},b,\xi_i} \left\{ \frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{n} \xi_i^2 \right\}$$

$$\textbf{Linear Constraints: } y_i\ (\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \ \forall \mathbf{x}_i \in \mathbf{D}$$

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\left(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i\right)$$

با مشتق گرفتن نسبت به $\boldsymbol{w}$ و b و ξ_i و مساوی صفر قرار دادن آن‌ها این روابط را بدست می‌آوریم.

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

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

$$\xi_i = \frac{1}{2C}\alpha_i$$

با اعمال این روابط در لاگرانژی اصلی، لاگرانژی دوگان بدست می‌آید و با بیشینه کردن آن مقدار ضرایب لاگرانژ α_i محاسبه می‌شود.

$$L_{dual} = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \frac{1}{4C}\sum_{i=1}^{n}\alpha_i^2$$

$$= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left(\mathbf{x}_i^T\mathbf{x}_j + \frac{1}{2C}\delta_{ij}\right)$$

$$\max_{\boldsymbol{\alpha}}\ L_{dual} = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left(\mathbf{x}_i^T\mathbf{x}_j + \frac{1}{2C}\delta_{ij}\right)$$

$$\text{subject to the constraints } \alpha_i \geq 0, \forall i \in \mathbf{D}, \text{ and } \sum_{i=1}^{n}\alpha_i y_i = 0$$

مقدار $\boldsymbol{w}$ و b را با استفاده از بردارهای پشتیبان محاسبه می‌کنیم.

$$\mathbf{w} = \sum_{\alpha_i > 0}\alpha_i y_i \mathbf{x}_i$$

$$b = \operatorname*{avg}_{\alpha_i > 0}\left\{y_i - \mathbf{w}^T\mathbf{x}_i\right\}$$

مثالی از نگاشت با تابع تبدیل:

$$\phi(\boldsymbol{x}) = \left(\sqrt{2}\boldsymbol{x}_1, \sqrt{2}\boldsymbol{x}_2, \boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \sqrt{2}\boldsymbol{x}_1\boldsymbol{x}_2\right)^T$$

$$, \qquad K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$$

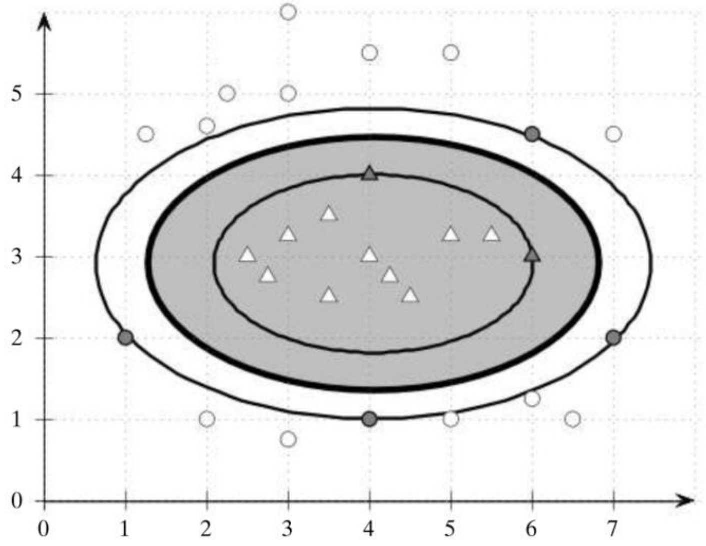


Figure 21.4. Nonlinear SVM: shaded points are the support vectors.

با نگاشت مجموعه داده‌ها به فضای ویژگی‌ها (Feature Space)، مجموعه‌داده‌ی جدید D_ϕ بدست می‌آید.

Objective Function: $\min_{\mathbf{w},b,\xi_i} \left\{ \frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{n}(\xi_i)^k \right\}$

Linear Constraints: $y_i\,(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i$, and $\xi_i \geq 0, \;\; \forall \mathbf{x}_i \in \mathbf{D}$

هزینه‌ی هینج (Hinge Loss):

$$\max_{\boldsymbol{\alpha}} L_{dual} = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$$

$$= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad 0 \leq \alpha_i \leq C, \qquad \sum_{i}^{n}\alpha_i y_i = 0$$

نقاط نگاشت شده در فضای ویژگی‌ها بصورت کرنل $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ در لاگرانژی دوگان ظاهر می‌شوند.

هزینه‌ی درجه دو (Quadratic Loss):

توان دو در تابع عینیِ هزینه منجر به جمله‌ای اضافه از جنس دلتای کرونکر می‌شود و برای رسیدگی به آن می‌توان تابع کرنل جدیدی تعریف کرد.

$$K_q(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j + \frac{1}{2C}\delta_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2C}\delta_{ij}$$

$$\max_{\boldsymbol{\alpha}} L_{dual} = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K_q(\mathbf{x}_i, \mathbf{x}_j)$$

بردار وزن و پیش‌قدر

تابع نگاشت $\phi(\boldsymbol{x}_i)$ در بردار وزن $\boldsymbol{w}$ مشاهده می‌شوند.

$$\mathbf{w}=\sum_{\alpha_i>0}\alpha_i y_i \phi(\mathbf{x}_i)$$

$$b=\underset{0<\alpha_i<C}{\operatorname{avg}}\{b_i\}=\underset{0<\alpha_i<C}{\operatorname{avg}}\left\{y_i-\mathbf{w}^T\phi(\mathbf{x}_i)\right\}$$

در محاسبه‌ی b بردار $\boldsymbol{w}$ ظاهر می‌شود که با جایگذاری آن، در b فقط کرنل $\boldsymbol{K}$ ظاهر می‌شود.

$$b_i=y_i-\sum_{\alpha_j>0}\alpha_j y_j \phi(\mathbf{x}_j)^T\phi(\mathbf{x}_i)$$

$$=y_i-\sum_{\alpha_j>0}\alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i)$$

وقتی بردار w و پیش‌قدر b را در معادله‌ی دسته‌بندی‌گر جاگذاری می‌کنیم، تابع نگاشت از بین می‌رود و مجدد فقط تابع کرنل در آن ظاهر می‌شود.

$$\hat{y}=\operatorname{sign}(\mathbf{w}^T\phi(\mathbf{z})+b)=\operatorname{sign}\left(\sum_{\alpha_i>0}\alpha_i y_i \phi(\mathbf{x}_i)^T\phi(\mathbf{z})+b\right)$$

$$=\boxed{\operatorname{sign}\left(\sum_{\alpha_i>0}\alpha_i y_i K(\mathbf{x}_i,\mathbf{z})+b\right)}$$

مثال: برای کلاس‌بندی داده‌های نشان داده شده در شکل، کرنل چندجمله‌ای درجه‌ی 2 ناهمگن و نیز $C=4$ انتخاب می‌کنیم.

$$K(\mathbf{x}_i,\mathbf{x}_j)=\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)=(1+\mathbf{x}_i^T\mathbf{x}_j)^2$$

$$\phi\left(\mathbf{x}=(x_1,x_2)^T\right)=\left(1,\sqrt{2}x_1,\sqrt{2}x_2,x_1^2,x_2^2,\sqrt{2}x_1x_2\right)^T$$

بطور مثال برای نقطه‌ی $\boldsymbol{x}_1=(1,2)^T$

$$\phi(\mathbf{x}_i)=\left(1,\sqrt{2}\cdot 1,\sqrt{2}\cdot 2,1^2,2^2,\sqrt{2}\cdot 1\cdot 2\right)^T$$

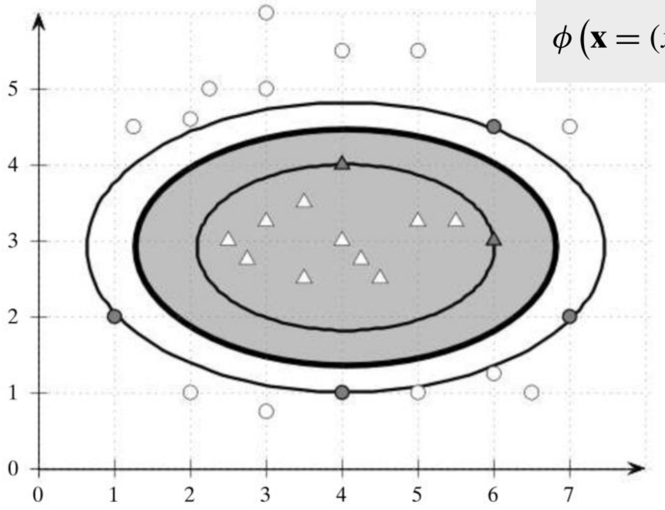


Figure 21.4. Nonlinear SVM: shaded points are the support vectors.

$\mathbf{x}_i$	$(x_{i1},x_{i2})^T$	$\phi(\mathbf{x}_i)$	y_i	α_i
$\mathbf{x}_1$	$(1,2)^T$	$(1,1.41,2.83,1,4,2.83)^T$	$+1$	0.6198
$\mathbf{x}_2$	$(4,1)^T$	$(1,5.66,1.41,16,1,5.66)^T$	$+1$	2.069
$\mathbf{x}_3$	$(6,4.5)^T$	$(1,8.49,6.36,36,20.25,38.18)^T$	$+1$	3.803
$\mathbf{x}_4$	$(7,2)^T$	$(1,9.90,2.83,49,4,19.80)^T$	$+1$	0.3182
$\mathbf{x}_5$	$(4,4)^T$	$(1,5.66,5.66,16,16,15.91)^T$	-1	2.9598
$\mathbf{x}_6$	$(6,3)^T$	$(1,8.49,4.24,36,9,25.46)^T$	-1	3.8502

$$\mathbf{w}=\sum_{\alpha_i>0}\alpha_i y_i\phi(\mathbf{x}_i)=(0,-1.413,-3.298,0.256,0.82,-0.018)^T$$

$$b=-8.841$$

برای سادگی تغییر متغیر به متغیرهای افزوده (Augmented Variables) $\tilde{\boldsymbol{x}}$ میدهیم که پیش‌قدر b را در بردار $\boldsymbol{w}$ ادغام می‌کند.

$$\tilde{\mathbf{x}}_i = (x_{i1}, \ldots, x_{id}, 1)^T \qquad \tilde{\mathbf{w}} = (w_1, \ldots, w_d, b)^T$$

$$h(\tilde{\mathbf{x}}) : \tilde{\mathbf{w}}^T \tilde{\mathbf{x}} = 0$$

$$h(\tilde{\mathbf{x}}) : w_1 x_1 + \cdots + w_d x_d + b = 0$$

مجموعه جدیدی از قیود بوجود می‌آید.

$$y_i \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i \geq 1 - \xi_i$$

Objective Function: $\min\limits_{\tilde{\mathbf{w}}, \xi_i} \left\{ \frac{\|\tilde{\mathbf{w}}\|^2}{2} + C \sum_{i=1}^{n} (\xi_i)^k \right\}$

Linear Constraints: $y_i \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i \geq 1 - \xi_i$ and $\xi_i \geq 0, \ \forall i = 1, 2, \ldots, n$

$$L = \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(y_i \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i - 1 + \xi_i \right) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$\frac{\partial}{\partial \tilde{\mathbf{w}}} L = \tilde{\mathbf{w}} - \sum_{i=1}^{n} \alpha_i y_i \tilde{\mathbf{x}}_i = \mathbf{0} \quad \text{or} \quad \tilde{\mathbf{w}} = \sum_{i=1}^{n} \alpha_i y_i \tilde{\mathbf{x}}_i$$

$$\frac{\partial}{\partial \xi_i} L = C - \alpha_i - \beta_i = 0 \quad \text{or} \quad \beta_i = C - \alpha_i$$

$$L_{dual} = \frac{1}{2} \tilde{\mathbf{w}}^T \tilde{\mathbf{w}} - \tilde{\mathbf{w}}^T \underbrace{\left(\sum_{i=1}^{n} \alpha_i y_i \tilde{\mathbf{x}}_i \right)}_{\tilde{\mathbf{w}}} - \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \underbrace{(C - \alpha_i - \beta_i)}_{0} \xi_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_j$$

برای تعمیم به مساله‌ی غیرخطی می‌توانیم تابع نگاشت و کرنل افزوده را تعریف نمود.

$$\tilde{\phi}(\mathbf{x}_i)^T = \left(\phi(\mathbf{x}_i)^T \ \ 1 \right)$$

$$\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = \tilde{\phi}(\mathbf{x}_i)^T \tilde{\phi}(\mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) + 1 = K(\mathbf{x}_i, \mathbf{x}_j) + 1$$

Objective Function: $\max\limits_{\boldsymbol{\alpha}} \ J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \tilde{K}(\mathbf{x}_i, \mathbf{x}_j)$

Linear Constraints: $0 \leq \alpha_i \leq C, \ \forall i = 1, 2, \ldots, n$

از نتایج نگاشت نقاط به فضای $\mathbb{R}^{d+1}$ این است که چون جمله‌ی پیش‌قدر بطور مجزا دیگر وجود ندارد، نیازی به اعمال قید $\sum_{i=1}^{n} \alpha_i y_i = 0$ در لاگرانژی دوگان نیست.
و از سوی دیگر اعمال قید $\alpha_i \in [0, C]$ برای هزینه‌ی هینج و یا $\alpha_i \geq 0$ برای هزینه‌ی درجه‌ی دو بسیار ساده خواهد بود.

Dual Solution: Stochastic Gradient Ascent

$$J(\alpha_k) = \alpha_k - \frac{1}{2}\alpha_k^2 y_k^2 \tilde{K}(\mathbf{x}_k, \mathbf{x}_k) - \alpha_k y_k \sum_{\substack{i=1 \\ i \neq k}}^{n} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{x}_k)$$

$$\frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_k} = \frac{\partial J(\alpha_k)}{\partial \alpha_k} = 1 - y_k \left(\sum_{i=1}^{n} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{x}_k) \right)$$

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \eta_t \nabla J(\boldsymbol{\alpha}_t)$$

$$\alpha_k = \alpha_k + \eta_k \frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_k} = \alpha_k + \eta_k \left(1 - y_k \sum_{i=1}^{n} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{x}_k) \right)$$

البته حین محاسبه، پیوسته باید چک کنیم که $\alpha_k \in [0, C]$ می‌باشد و در صورت خروج به بازه برمی‌گردانیم.

Algorithm 21.1: Dual SVM Algorithm: Stochastic Gradient Ascent

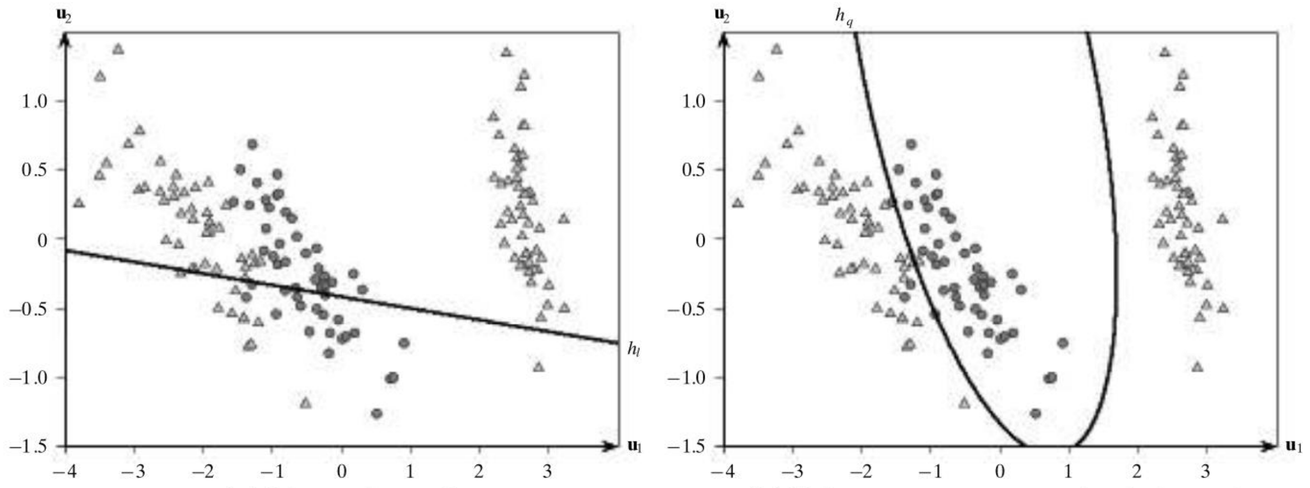
SVM-DUAL ($\mathbf{D}, K, loss, C, \epsilon$):
- **if** $loss$ = hinge **then**
  - $\mathbf{K} \leftarrow \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1,\dots,n}$ // kernel matrix, hinge loss
- **else if** $loss$ = quadratic **then**
  - $\mathbf{K} \leftarrow \{K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2C}\delta_{ij}\}_{i,j=1,\dots,n}$ // kernel matrix, quadratic loss
- $\tilde{\mathbf{K}} \leftarrow \mathbf{K} + 1$ // augmented kernel matrix
- **for** $k = 1, \dots, n$ **do** $\eta_k \leftarrow 1 / \tilde{K}(\mathbf{x}_k, \mathbf{x}_k)$ // set step size
- $t \leftarrow 0$
- $\boldsymbol{\alpha}_0 \leftarrow (0, \dots, 0)^T$
- **repeat**
  - $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}_t$
  - **for** $k = 1$ *to* n **do**
    - // update kth component of $\boldsymbol{\alpha}$
    - $\alpha_k \leftarrow \alpha_k + \eta_k \left(1 - y_k \sum_{i=1}^{n} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{x}_k)\right)$
    - **if** $\alpha_k < 0$ **then** $\alpha_k \leftarrow 0$
    - **if** $loss$ = hinge **and** $\alpha_k > C$ **then** $\alpha_k \leftarrow C$
  - $\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}$
  - $t \leftarrow t + 1$
- **until** $\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1}\| \leq \epsilon$

$$\hat{y} = \text{sign}\Big(h(\tilde{\phi}(\mathbf{z}))\Big) = \text{sign}\Big(\tilde{\mathbf{w}}^T \tilde{\phi}(\mathbf{z})\Big) = \text{sign}\left(\sum_{\alpha_i > 0} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{z}) \right)$$

linear kernel $\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j + 1$

inhomogeneous quadratic kernel $\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = (c + \mathbf{x}_i^T \mathbf{x}_j)^2 + 1$, with $c = 1$

gaussian kernel $\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\} + 1$



(a) Linear kernel (b) Inhomogeneous quadratic kernel (c) Gaussian kernel ($\sigma^2 = 0.05$) (d) Gaussian kernel ($\sigma^2 = 1$)

Figure 21.6. SVM dual algorithm: linear, inhomogeneous quadratic, and gaussian kernels.

یک روش برای انتخاب، کنترل پایداری در فرآیند بهینه‌یابی است. لذا اندازه‌ی گام η_k را باید به نحوی انتخاب کنیم که مشتقات نسبت به α_k به صفر میل کند تا فرآیند بهینه‌یابی پایدار بماند.

$$\eta_k = \frac{1}{\tilde{K}(\mathbf{x}_k, \mathbf{x}_k)}$$

$$\frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_k} = \left(1 - y_k \sum_{i \neq k} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{x}_k)\right) - y_k \alpha_k y_k \tilde{K}(\mathbf{x}_k, \mathbf{x}_k)$$

$$\frac{\partial J(\boldsymbol{\alpha})}{\partial a_k} = \left(1 - \frac{1}{\tilde{K}(\mathbf{x}_k, \mathbf{x}_k)} \tilde{K}(\mathbf{x}_k, \mathbf{x}_k)\right)\left(1 - y_k \sum_{i=1}^{n} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{x}_k)\right) = 0$$

$$\hat{y} = \mathrm{sign}\Big(h(\tilde{\phi}(\mathbf{z}))\Big) = \mathrm{sign}\Big(\tilde{\mathbf{w}}^T \tilde{\phi}(\mathbf{z})\Big) = \mathrm{sign}\left(\sum_{\alpha_i > 0} \alpha_i y_i \tilde{K}(\mathbf{x}_i, \mathbf{z})\right)$$