۲ فصل دوم

رویکرد دسته‌بندی ماشین بردار پشتیبان

۲- ۱ مقدمه

در این فصل، به معرفی و فرمول‌بندی روش دسته‌بندی SVM می‌پردازیم. ابتدا آن را برای حالتی که داده‌های آموزش خطی-تفکیک‌پذیر هستند شرح می‌دهیم و سپس شیوه‌های توسعه آن را به حالتی که داده‌ها خطی تفکیک‌پذیر نیستند، بررسی می‌کنیم. همچنین، شیوه لحاظ کردن هسته و نیز روش‌های ارزیابی دقت را بیان می‌کنیم.

۲-۱ -۱ SVM با حاشیه سخت

در این بخش، به شرح روش SVM با فرض خطی-تفکیک‌پذیری داده‌ها می‌پردازیم.

تعریف ۱: داده‌های خطی-تفکیک‌پذیر

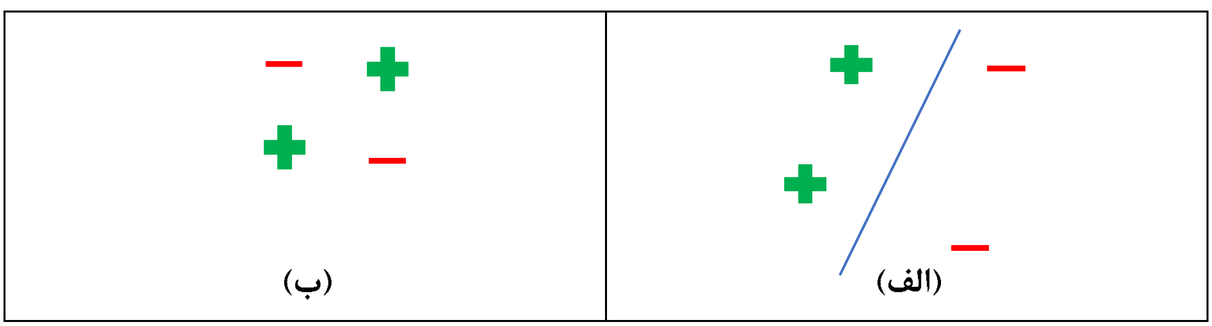
مجموعه داده‌های آموزشی D را خطی-تفکیک‌پذیر می‌نامیم هرگاه بتوان آنها را با یک ابرصفحه جدا کرد به طوری که همه داده‌های با برچسب $+1$ در یک سمت ابرصفحه و همه داده‌های با برچسب -1 در سمت دیگر آن باشند. به عبارت دیگر، مجموعه داده‌های آموزشی D را خطی-تفکیک‌پذیر می‌نامیم اگر $(w,b) \in (\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$ وجود داشته باشد به‌طوری که به‌ازای هر $i \in \{1,2,\dots,m\}$ شرایط زیر برقرار باشند:

$$\begin{cases} y_i = 1 \ \Rightarrow w^T x_i + b > 0 \\ y_i = -1 \ \Rightarrow w^T x_i + b < 0 \end{cases} \tag{۱-۲}$$

توجه کنید که شرط فوق را می‌توان به صورت زیر نیز بیان کرد:

$$y_i(w^T x_i + b) > 0 \tag{۲-۲}$$

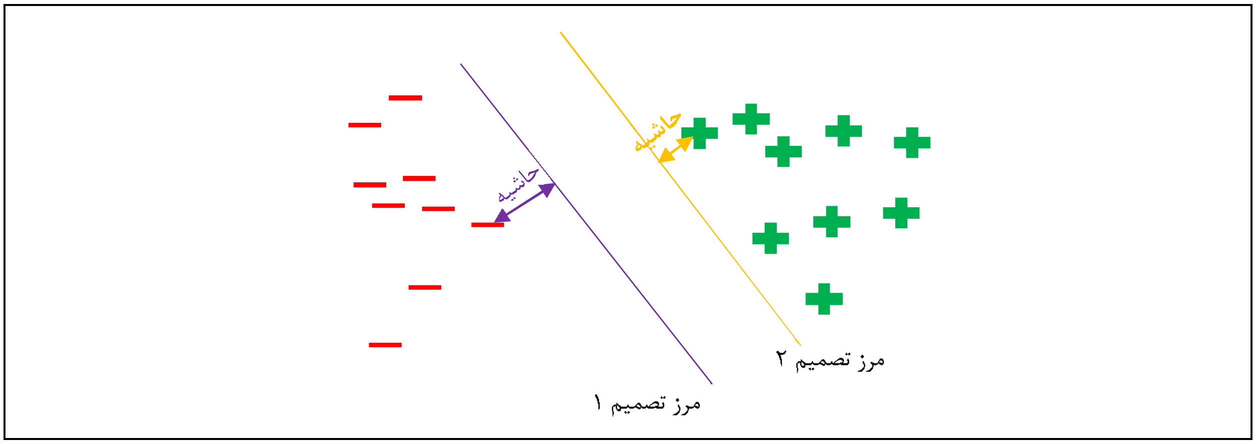
به عنوان مثال، شکل ۲-۱ مجموعه‌ای از داده‌ها را در حالت $n = 2$ و $m = 4$ نشان می‌دهد. همان‌طور که ملاحظه می‌شود، داده‌های قسمت الف خطی-تفکیک‌پذیرند در حالی که داده‌های قسمت ب، چنین نیستند.



شکل ۲-۱: خطی-تفکیک‌پذیری

شکل ۲-۲یک مجموعه از داده‌های خطی-تفکیک‌پذیر را در حالت $n = 2$ نشان می‌دهد. همان‌طور که مشاهده می‌کنید بی‌نهایت ابرصفحه وجود دارند که داده‌های دسته‌های $+1$ و -1 را از هم جدا می‌کنند. اما هدف یافتن

ابرصفحه‌ای است که ضمن آنکه داده‌های آموزشی را به خوبی تفکیک می‌کند، قدرت تعمیم[1] خوبی نیز داشته باشد بدین معنی که برای داده‌های جدید نیز بتواند تا حد خوبی دسته‌بندی را درست انجام دهد. لذا، در راستای رسیدن به تعمیم‌پذیری خوب، روش ماشین بردار پشتیبان، در صورتی که داده‌های آموزشی خطی-تفکیک‌پذیر باشند، از بین همه ابرصفحه‌هایی که داده‌های دسته‌های $+1$ و -1 را از هم جدا می‌کنند، ابرصفحه‌ای را انتخاب می‌کند که فاصله‌اش تا نزدیکترین داده بیشترین مقدار ممکن باشد. برای روشن شدن بحث، به تعریف مفاهیم مرز تصمیم[2] و حاشیه[3] می‌پردازیم.



شکل ۲-۲: مرز تصمیم و حاشیه

تعریف ۲: مرز تصمیم

به ابرصفحه $w^T x + b = 0$ که برای تفکیک داده‌های دو دسته شناسایی می‌شود، مرز تصمیم گفته می‌شود.

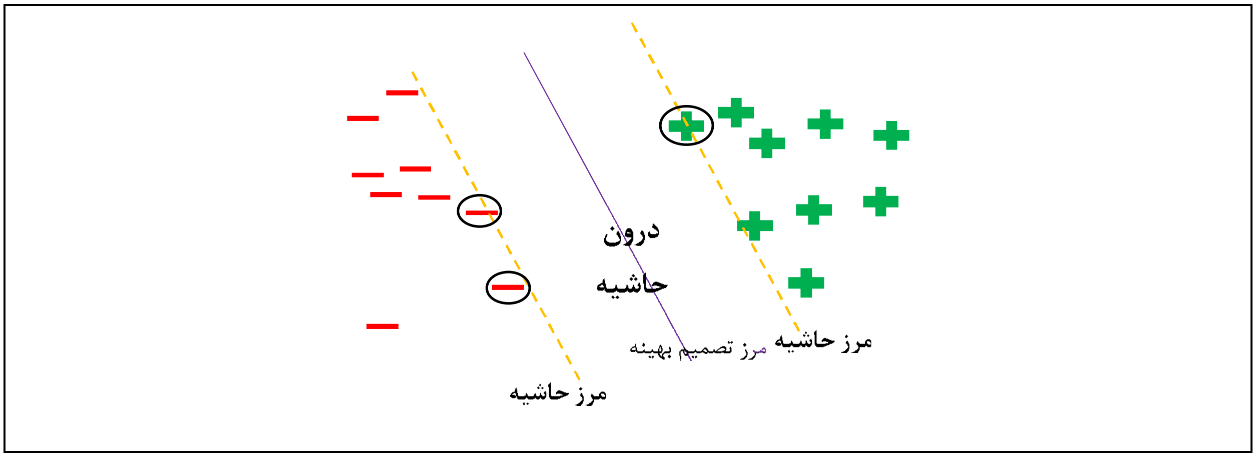
تعریف ۳: حاشیه متناظر با یک مرز تصمیم

فاصله یک مرز تصمیم تا نزدیکترین داده آموزش را حاشیه متناظر با آن مرز تصمیم می‌نامند.

درشکل ۲-۲ دو مرز تصمیم مختلف و حاشیه‌های متناظر با آنها نمایش داده شده است.

روش ماشین بردار پشتیبان در جستجوی مرز تصمیمی است که حاشیه آن بیشترین مقدار ممکن باشد. به عنوان مثال، شکل ۲-۳مرز تصمیم بهینه را برای یک مجموعه از داده‌ها نشان می‌دهد.

[1] Generalization
[2] Decision boundary
[3] Margin



شکل ۲-۳: مرز تصمیم بهینه و مرز حاشیه

قبل از ادامه بحث، لازم است به تعریف سه واژه که در مباحث بعدی از آنها استفاده خواهیم کرد، بپردازیم.

تعریف ۴: بردار پشتیبان

داده‌ای که کمترین فاصله را با مرز تصمیم بهینه (که از روش SVM به دست می‌آید) دارد، بردار پشتیبان[۱] نامیده می‌شود.

تعریف ۵: مرز حاشیه

خطوط موازی مرز تصمیم بهینه که فاصله‌شان تا مرز تصمیم بهینه برابر با مقدار حاشیه است، مرز حاشیه[۲] نام دارند.

مفاهیم فوق روی شکل ۲-۳نشان داده شده‌اند. نقاطی که دور آنها دایره کشیده شده است، بردارهای پشتیبان هستند و خط چین‌های نارنجی مرز حاشیه را نشان می‌دهند. لازم به ذکر است که در طول پایان‌نامه، ممکن است به نقاط روی مرز حاشیه[۳] و نقاط درون حاشیه[۴] اشاره کنیم. به عنوان مثال، درشکل ۲-۳ نقاط روی خطچین‌های نارنجی بیانگر نقاط روی مرز حاشیه و نقاط بین خطچین‌های نارنجی درون حاشیه را نشان می‌دهند.

با توجه به آنکه فاصله نقطه x_i از ابرصفحه $w^T x + b = 0$ با رابطه $\frac{|w^T x_i + b|}{||w||_2}$ محاسبه می‌شود، ابرصفحه با بیشترین حاشیه با حل مدل بهینه‌سازی زیر به دست می‌آید:

---

[۱] Support vector

[۲] Margin boundary

[۳] On the margin boundary

[۴] Inside the margin

مدل ۲-۱: مدل SVM

$$\max_{w\in\mathbb{R}^n\backslash\{0\},b\in\mathbb{R}}\ \min_{i=1,\dots,m}\frac{y_i(w^Tx_i+b)}{||w||_2}$$

اما چرا ابرصفحه‌ای که با حل مدل ۲-۱به دست می‌آید همه داده‌ها را به درستی دسته‌بندی می‌کند؟ در واقع، اگر ابرصفحه $\widetilde{w}^Tx+\tilde{b}=0$ به گونه‌ای باشد که مثلاً داده $\tilde{\imath}$ را به درستی دسته‌بندی نکند، آنگاه داریم $y_{\tilde{\imath}}\left(\widetilde{w}^Tx_{\tilde{\imath}}+\tilde{b}\right)<0$ و لذا، مقدار تابع هدف مدل ۲-۱ به ازای چنین ابرصفحه‌ای منفی خواهد شد. در حالی که باتوجه به آنکه داده‌ها خطی-تفکیک‌پذیر فرض شده‌اند، پس طبق رابطه(۲-۲)، همواره $(w,b)\in(\mathbb{R}^n-\{0\})\times\mathbb{R}$ وجود دارند که $y_i(w^Tx_i+b)>0$ و مقدار تابع هدف مدل ۲-۱ به ازای چنین ابرصفحه‌ای مثبت خواهد شد. بنابراین، هدف ماکزیمم‌سازی در مدل فوق سبب می‌شود که ابرصفحه‌هایی که داده‌ها را به درستی دسته‌بندی نمی‌کنند، انتخاب نشوند.

با اعمال تغییر متغیر زیر، مدل ۲-۱به طور معادل، به صورت مدل ۲-۲ بازنویسی می‌شود:

$$s=\quad\min_{i=1,\dots,m}y_i(w^Tx_i+b)$$

پس مدل ۲-۱ با مدل زیر معادل است:

مدل ۲-۲: بازنویسی مدل ۲-۱

$$\max_{w\in\mathbb{R}^n\backslash\{0\},b\in\mathbb{R}}\frac{s}{||w||_2}$$

$$s.t.$$

$$s\le y_i(w^Tx_i+b)\qquad\forall i=1,\dots,m$$

اما همان‌طور که قبلاً ذکر شد، اگر داده‌ها خطی-تفکیک‌پذیر باشند، در جواب بهین مسأله همواره داریم $s>0$. بنابراین می‌توان بدون از دست رفتن کلیت، طرفین قید $s\le y_i(w^Tx_i+b)$ را بر s تقسیم کرد (این کار صرفاً معادل با مقیاس‌گیری[۱] ضرایب w و b است به گونه‌ای که فاصله نزدیک‌ترین داده به ابرصفحه جداساز برابر با یک باشد). پس مدل ۲-۲را می‌توان به طور معادل به صورت زیر بازنویسی کرد:

[۱] Scaling

مدل ۲-۳: بازنویسی مدل ۲-۲

$$\max_{w\in\mathbb{R}^n\setminus\{0\},b\in\mathbb{R}} \frac{1}{\|w\|_2}$$
$$s.t.$$
$$y_i(w^T x_i + b) \geq 1 \quad \forall i = 1,\dots,m$$

از آنجا که تابع $\frac{1}{\|w\|_2}$ همواره مثبت است، با بازنویسی تابع هدف مدل ۲-۳ به صورت مینیمم‌سازی مدل زیر به دست می‌آید:

مدل ۲-۴: بازنویسی مدل ۲-۳

$$\min_{w\in\mathbb{R}^n\setminus\{0\},b\in\mathbb{R}} \|w\|_2$$
$$s.t.$$
$$y_i(w^T x_i + b) \geq 1 \ \forall i = 1,\dots,m$$

تذکر ۲-۱:

توجه کنید که مدل ۲-۲و مدل ۲-۴ تنها در صورتی معادل هستند که داده‌ها خطی-تفکیک‌پذیر باشند. اگر چنین نباشد، همواره به ازای هر w و b یک داده وجود دارد که $y_i(w^T x_i + b) < 0$ بنابراین مدل ۲-۴ نشدنی می‌شود در حالی که مدل ۲-۲ شدنی است چون ممکن است در جواب بهین مسأله مقدار s منفی شود در آن صورت داده‌ها دیگر خطی-تفکیک‌پذیر نیستند اما مدل ۲-۲ هم‌چنان جوابی را به‌دست می‌دهد. اما مدل ۲-۴ فرض می‌کند که فاصله نزدیک‌ترین داده به ابرصفحه جداساز برابر با یک است و چنین فرضی تنها در صورتی محدودکننده نیست که داده‌ها خطی تفکیک‌پذیر باشند.

تذکر ۲-۲:

در همه مدل‌هایی که در بالا ذکر شد می‌توان شرط $w \in \mathbb{R}^n\setminus\{0\}$ را حذف و w را متعلق به $\mathbb{R}^n$ در نظر گرفت.

چون تابع $\|w\|_2$ تابعی مثبت و تابع $\frac{1}{2}x^2$ در بازه $(0,\infty)$ مثبت و صعودی است، می‌توان تابع $\frac{1}{2}x^2$ را روی تابع $\|w\|_2$ اثر داد و به مسأله معادلی رسید. لذا، بدون از دست رفتن کلیت، از این پس مدل SVM را به صورت زیر در نظر می‌گیریم و به آن تحت عنوان SVM با حاشیه سخت (HMSVM) اشاره می‌کنیم.

مدل ۲-۵: HMSVM

$$\min_{w\in\mathbb{R}^n,b\in\mathbb{R}} \frac{1}{2}\|w\|_2^2$$
$$s.t.$$
$$y_i(w^T x_i + b) \geq 1 \ \ \forall i = 1, \dots, m$$

فرض کنید $w^T x + b = 0$ مرز تصمیم به دست آمده از مدل HMSVM باشد. در این صورت برچسب داده جدید $\hat{x}$، با رابطه $\hat{y} = sign(w^T x + b)$ برآورد می‌گردد.

مدل HMSVM یک مدل برنامه‌ریزی درجه دوم محدب است زیرا اولاً تابع $\|w\|_2$ تابعی محدب است (نرم‌ها توابع محدب هستند) همچنین، $\frac{1}{2}x^2$ روی $(0, \infty)$ صعودی است پس اگر تابع $\frac{1}{2}x^2$ را روی $\|w\|_2$ اثر دهیم، کماکان یک تابع محدب خواهیم داشت. بنابراین مسئله از نوع مینیمم سازی با تابع هدف محدب و قید مسئله نیز نسبت به w و b خطی است. هم‌چنین فاصله نسبی دوگانی[۱] در آن برابر با صفر است. برای فرمول‌بندی مسأله دوگان، تابع لاگرانژین را به صورت زیر تشکیل می‌دهیم:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i(w^T x_i + b)\right)$$

مسأله دوگان به صورت زیر است:

$$\max_{\alpha\geq 0} \min_{w,b} \mathcal{L}(w, b, \alpha)$$

با توجه به آنکه تابع $\mathcal{L}(w, b, \alpha)$ نسبت به (w, b) تابعی محدب است، جواب بهین مسأله $\min_{w,b} \mathcal{L}(w, b, \alpha)$ با حل دستگاه زیر حاصل خواهد شد:

$$\nabla_w \mathcal{L}(w, b, \alpha) = 0 \quad \Rightarrow \quad w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad (۲\text{-}۳)$$

$$\nabla_b \mathcal{L}(w, b, \alpha) = 0 \quad \Rightarrow \quad -\sum_{i=1}^{m} \alpha_i y_i = 0$$

با جایگذاری روابط فوق در $\mathcal{L}(w, b, \alpha)$ داریم:

[۱] Duality gap

$$\min_{w,b} \mathcal{L}(w,b,\alpha) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{m}\alpha_i y_i\, b = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

بنابراین دوگان مدل HMSVM که به آن تحت عنوان DHMSVM اشاره می‌کنیم به صورت زیر است که یک مسأله برنامه‌ریزی درجه دوم محدب است.

مدل ۲-۶: DHMSVM

$$\max \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t.}$$

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

$$\alpha_i \geq 0$$

با محاسبه جواب بهین مسأله دوگان (DHMSVM) می‌توان به راحتی جواب بهین مسأله اولیه (HMSVM) را استخراج کرد. برای روشن شدن بحث، فرض کنید (w^*, b^*) جواب بهین مسأله HMSVM (مدل ۲-۵) و α^* جواب بهین مسأله DHMSVM (مدل ۲-۶) باشد. با توجه به رابطه (۲-۳)، w^* به صورت زیر تعیین می‌گردد:

$$w^* = \sum_{i=1}^{m}\alpha_i^* y_i x_i$$

برای تعیین b^*، از مفهوم بردار بردار پشتیبان استفاده می‌کنیم. با توجه به شرایط مکمل زائد، داریم:

$$\alpha_i^*\left(1 - y_i\left(w^{*T}x_i + b^*\right)\right) = 0 \quad \forall i = 1, \dots, m$$

حال فرض کنید i' داده‌ای باشد که $\alpha_{i'}^* > 0$، پس $1 - y_{i'}\left(w^{*T}x_{i'} + b^*\right) = 0$ است و این به این معنا است که داده i'ام همان‌داده‌ای است که کوچکترین فاصله را با مرز تصمیمِ به دست آمده از حل مدل HMSVM دارد یا به عبارت بهتر، داده i'ام بردار پشتیبان است. بنابراین، برای تعیین مقدار b^* کافی است یک داده i' که

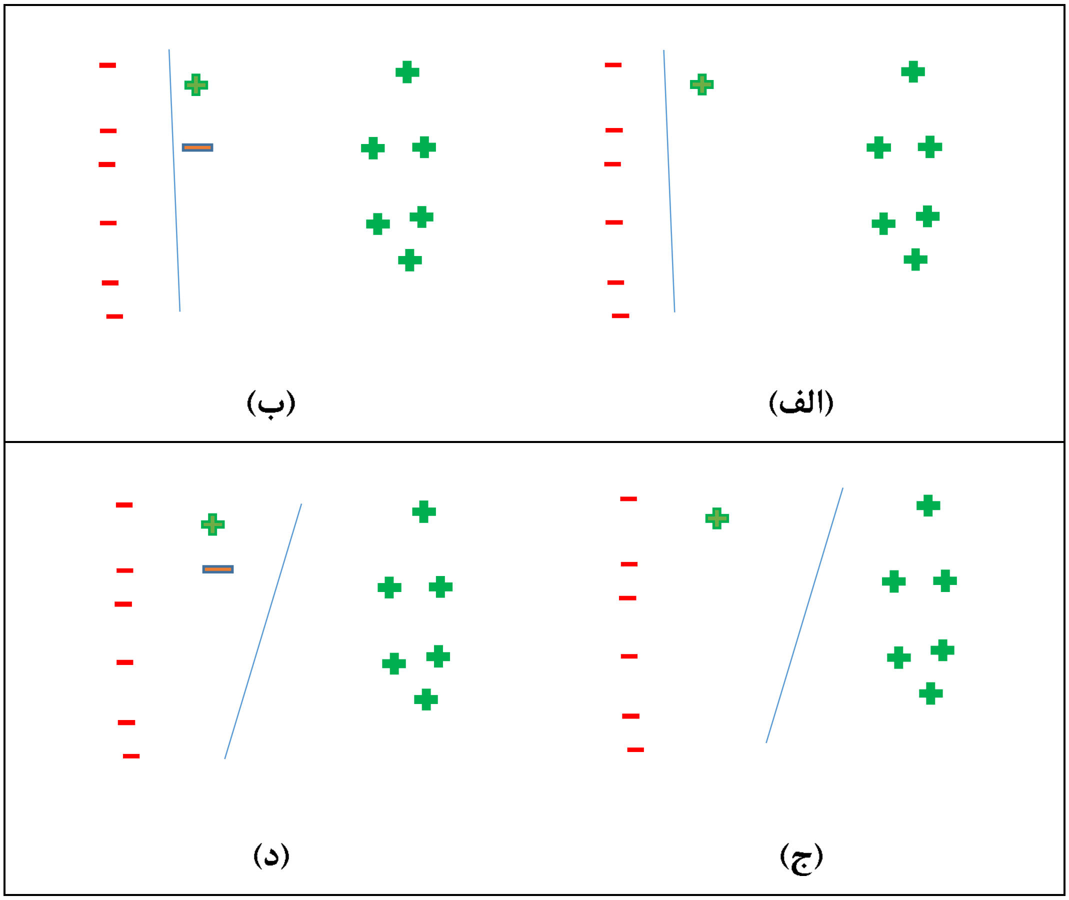
$\alpha^*_{i'} > 0$ را شناسایی کرد و سپس، باتوجه به آنکه $y_{i'}\left(w^{*T}x_{i'} + b^*\right) = 1$ و از اینکه $\frac{1}{y_{i'}} = y_{i'}$ نتیجه گرفت $b^* = y_{i'} - w^{*T}x_{i'}$.

همان‌طور که دیدیم هر دو مدل HMSVM و DHMSVM مسائل برنامه‌ریزی درجه دوم محدب هستند اما مزیت مسأله دوگان آن است که تعداد متغیرهای آن برابر با تعداد داده‌های آموزشی است و لذا، با افزایش تعداد ویژگی‌ها به خصوص وقتی که x را به فضای $\varphi(x)$ نگاشت می‌کنیم، تعداد متغیرهای آن افزایش نمی‌یابد. برای جزئیات بیشتر در خصوص نگاشت به فضای $\varphi(x)$ به بخش۲-۱ -۳ مراجعه نمایید.

۲-۱ -۲ SVM با حاشیه نرم

در بخش قبل، SVM با حاشیه سخت (HMSVM) و دوگان آن معرفی و فرمول‌بندی شد. دلیل استفاده از واژه «حاشیه سخت» در نامگذاری این مدل آن است که در آن باید ابرصفحه جداساز به گونه‌ای شناسایی شود که همه داده‌های آموزشی را به درستی دسته‌بندی کند. لذا، HMSVM وقتی که داده‌های آموزشی خطی-تفکیک-پذیر نیستند (یعنی امکان آن که با یک خط داده‌های اموزشی دسته 1+ را از داده‌های آموزشی دسته 1- به درستی تفکیک کنیم وجود ندارد)، قابل استفاده نیست. همچنین، حتی در مواقعی که داده‌های آموزشی خطی-تفکیک‌پذیرند، اگر ابرصفحه‌ای که از مدل HMSVM به دست می‌آید، حاشیه کوچکی داشته باشد، قدرت تعمیم‌پذیری خوبی نخواهد داشت. در چنین شرایطی، می‌توان از SVM با حاشیه نرم[۱] (SMSVM) استفاده نمود. برای روشن شدن بحث شکل ۲-۴را در نظر بگیرید. قسمت الف، داده‌های آموزشی خطی تفکیک‌پذیر را به همراه مرز تصمیم بهینه که از حل مدل HMSVM به دست می‌آید نشان می‌دهد. همان‌طور که ملاحظه می‌شود، این مرز تصمیم تحت تأثیر داده پرت دسته مثبت قرار گرفته و بیش از حد به داده‌های دسته منفی نزدیک شده است و بنابراین قدرت تعمیم‌پذیری خوبی ندارد به طوری که در قسمت ب، داده جدیدی که به رنگ قرمز است و متعلق به دست منفی است را اشتباهاً در دسته مثبت قرار می‌دهد (شکل ۲-۴-ب را ببینید). شکل ۲-۴ -ج نمایشی از مدل SVM با حاشیه نرم را به ازای داده‌های دسته الف، نشان می‌دهد که در آن به برخی داده‌ها اجازه داده می‌شود که اشتباه دسته‌بندی شوند اما در عوض حاشیه بزرگتری ایجاد خواهد شد که در مقایسه با قسمت الف، قدرت تعمیم‌پذیری بیشتری دارد و همان‌طور که در شکل ۲-۴-د دیده می‌شود، این بار داده جدیدی که به رنگ قرمز است و متعلق به دست منفی است را به درستی در دسته منفی قرار می‌دهد.

[۱] Soft-margin SVM



شکل ۲-۴: تاثیر حاشیه نرم در قدرت تعمیم‌پذیری

اگر در HMSVM امکان نقض قیود را فراهم و میزان نقض را در تابع هدف جریمه کنیم به مدل SMSVM می‌رسیم که فرمول‌بندی آن به صورت زیر است:

مدل ۲-۷: SMSVM

$$\min_{w,b,\xi} \|w\|_2^2 + C\sum_{i=1}^{m} \xi_i$$

$s.t.$

$$y_i(w^T x_i + b) + \xi_i \geq 1 \qquad \forall i = 1, \dots, m \qquad (۴\text{-}۲)$$

$$\xi_i \geq 0 \quad i = 1, \dots, m \qquad (۵\text{-}۲)$$

که در آن ξ_i میزان نقض قید $y_i(w^T x_i + b) \geq 1$ را نشان می‌دهد و پارامتر $C > 0$ برای جریمه نقض در نظر گرفته شده و به به عبارت $C\sum_{i=1}^{m} \xi_i$ تابع زیان هینگ[1] گفته می‌شود. بنابراین، به ازای مرز تصمیمی که از حل SMSVM به دست می‌آید، ممکن است برخی داده‌های آموزشی درست دسته‌بندی نشوند. یعنی داده i' وجود داشته باشد به طوری که

$$y_{i'} = +1 \ \ and \ \ w^T x_{i'} + b < 0$$

یا

$$y_{i'} = -1 \ \ and \ \ w^T x_{i'} + b > 0$$

یا ممکن است داده‌ها درست دسته‌بندی شوند اما درون حاشیه قرار گیرند. لذا، می‌توان گفت برای داده آموزشی i، شش حالت زیر امکان‌پذیر است:

حالت اول: $\xi_i = 0$

این حالت به این معنی است که داده iاُم و jاُم به درستی دسته‌بندی شده‌اند و نیز درون حاشیه قرار ندارند (داده iام و jاُم در شکل ۲-۵ را ببینید).

حالت دوم: $0 < \xi_i < 1$

این حالت به این معنی است که داده iاُم و jاُم به درستی دسته‌بندی شده‌اند اما درون حاشیه قرار دارند (داده iام و jاُم در شکل ۲-۶ را ببینید).

حالت سوم: $\xi_i = 1$

این حالت به این معنی است که داده iاُم و jاُم درون حاشیه و دقیقاً روی مرز تصمیم قرار دارند (داده iام و jاُم درشکل ۲-۷ را ببینید).

حالت چهارم: $1 < \xi_i < 2$

این حالت به این معنی است که داده iاُم و jاُم اشتباه دسته‌بندی شده‌اند اما درون حاشیه قرار دارند (داده iام و jاُم در شکل ۲-۸ را ببینید).

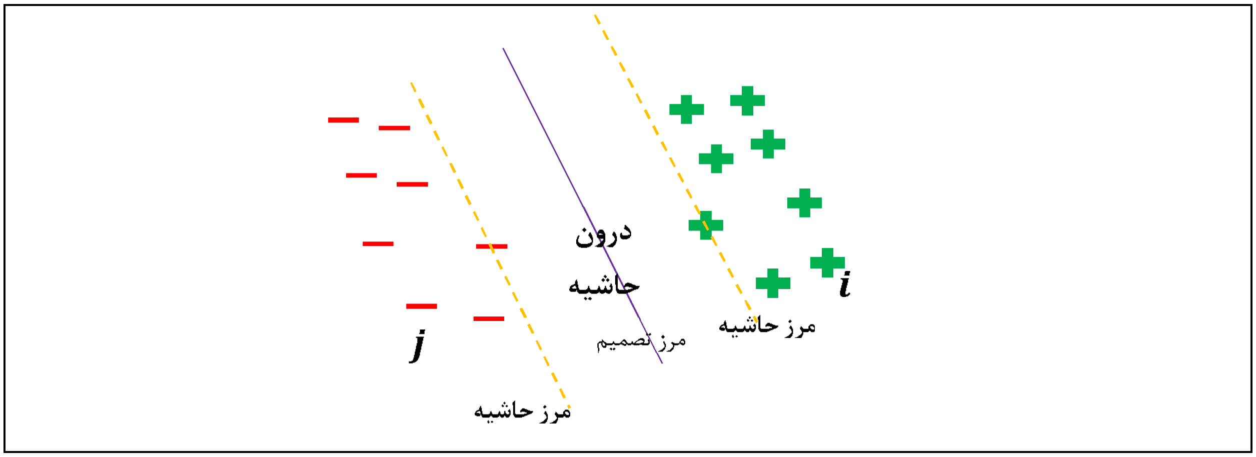
حالت پنجم: $\xi_i = 2$

این حالت به این معنی است که داده iاُم و jاُم اشتباه دسته‌بندی شده‌اند و دقیقاً روی مرز حاشیه قرار دارند (داده iام و jاُم درشکل ۲-۹را ببینید).

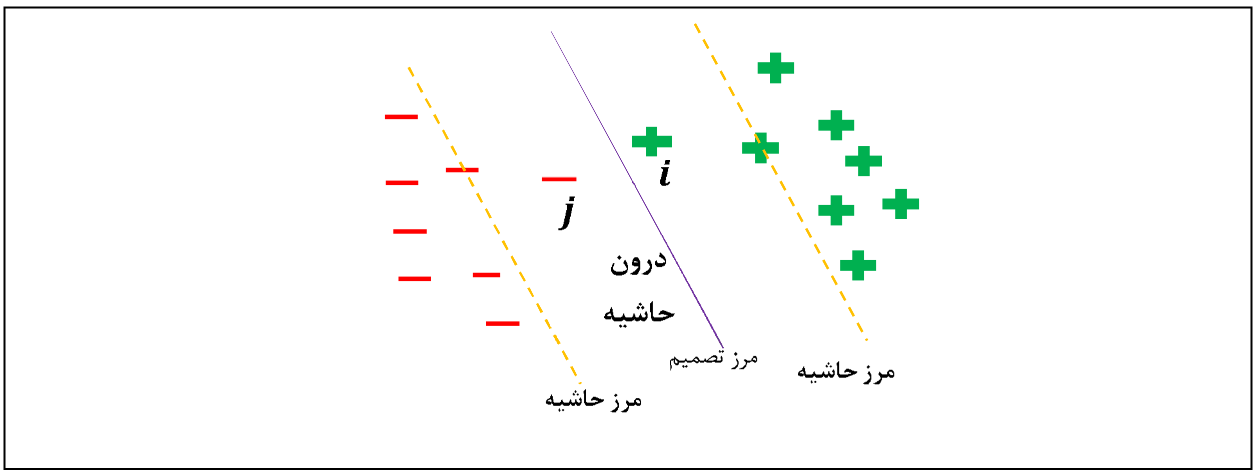
[1] Hing loss function

حالت ششم: $\xi_i > 2$

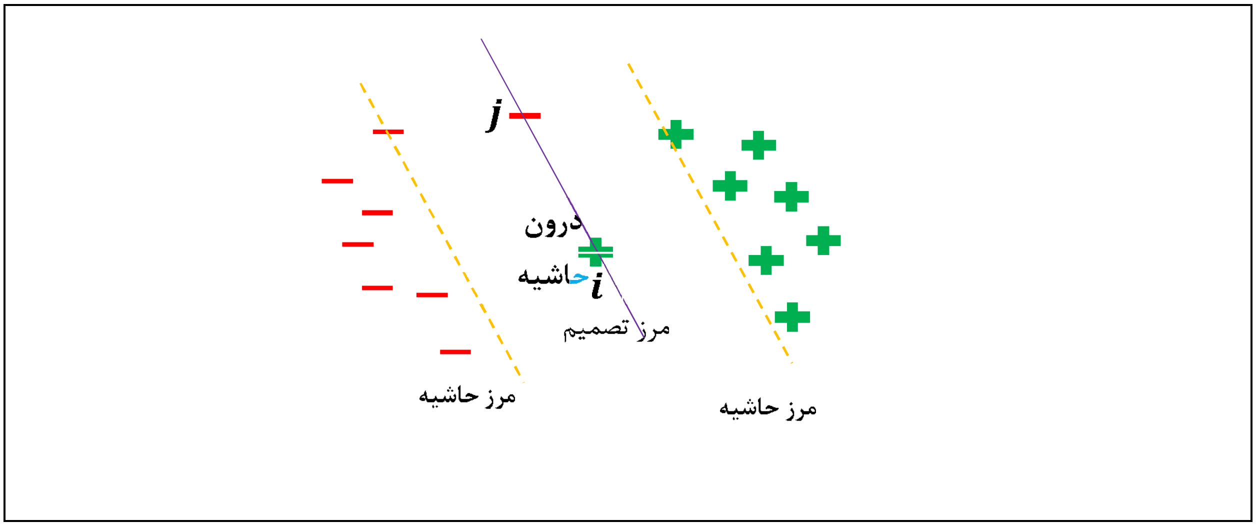
این حالت به این معنی است که داده iاُم و jاُم اشتباه دسته‌بندی شده اند و درون حاشیه و روی مرز حاشیه قرار ندارند (داده iاُم و jاُم در شکل ۲-۱۰را ببینید).



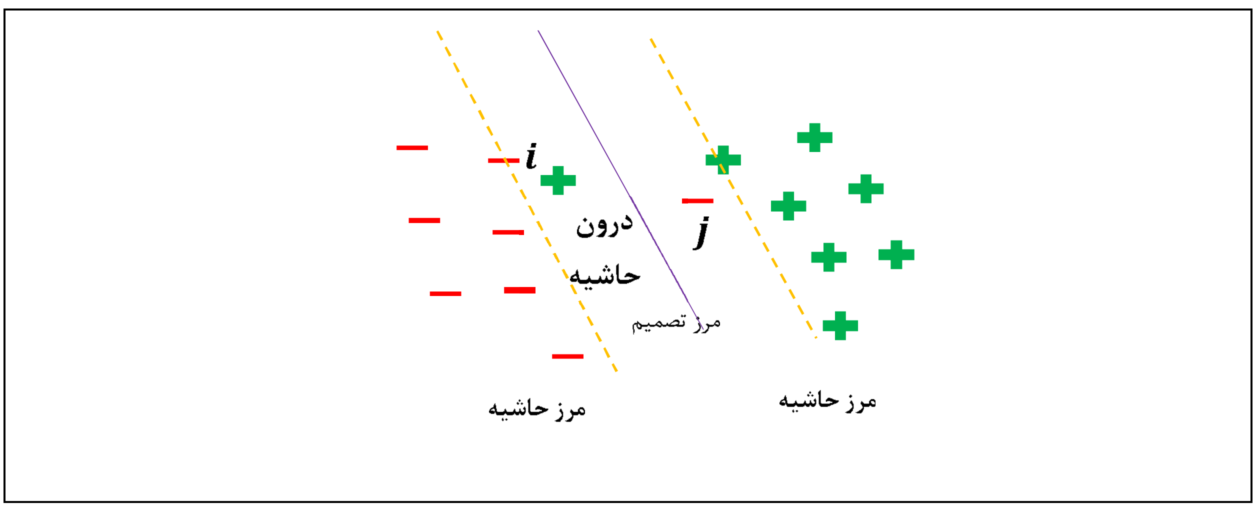
شکل ۲-۵: تابع زیان هینگ وقتی $\xi_i = 0$



شکل ۲-۶: تابع زیان هینگ وقتی $0 < \xi_i < 1$



شکل ۲-۷: تابع زیان هینگ وقتی $\xi_i = 1$



شکل ۲-۸: تابع زیان هینگ وقتی $1 < \xi_i < 2$