ماشین بردار پشتیبان

Support Vector Machines

ماشین بردار پشتیبان (Support Vector Machines)

کلاس‌های جداپذیر

- ✓ ماشین بردار پشتیبان (SVM) روشی دیگر برای طراحی طبقه‌بند خطی می‌باشد.
- ✓ برای مسئله دو کلاسه با جداپذیری خطی، یک راه حل یکتا برای ابرصفحه تصمیم خطی وجود ندارد.

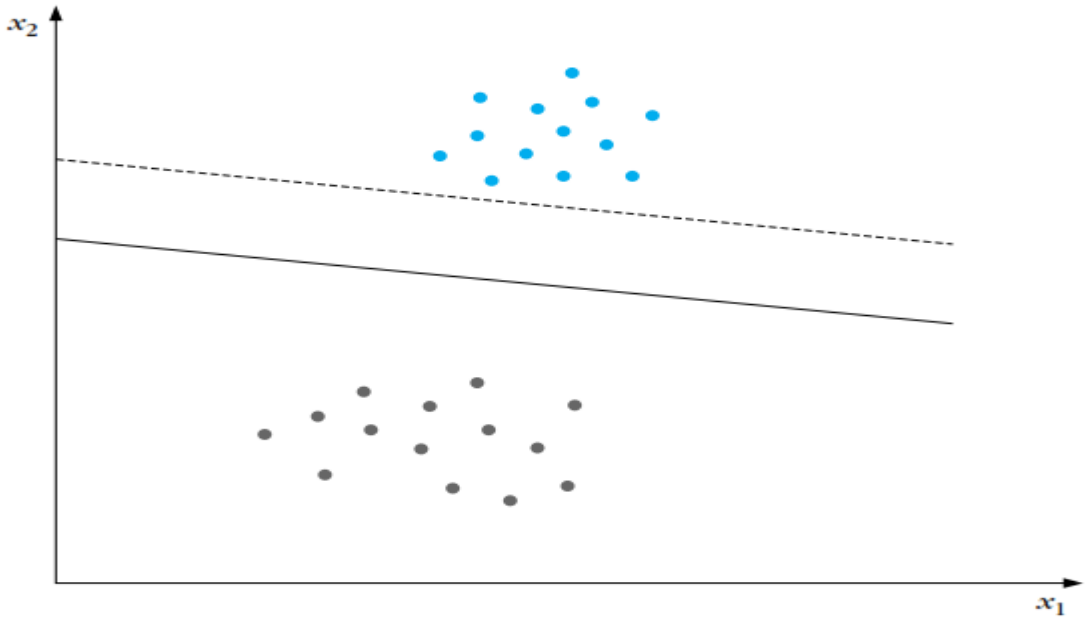


FIGURE 3.9

An example of a linearly separable two-class problem with two possible linear classifiers.

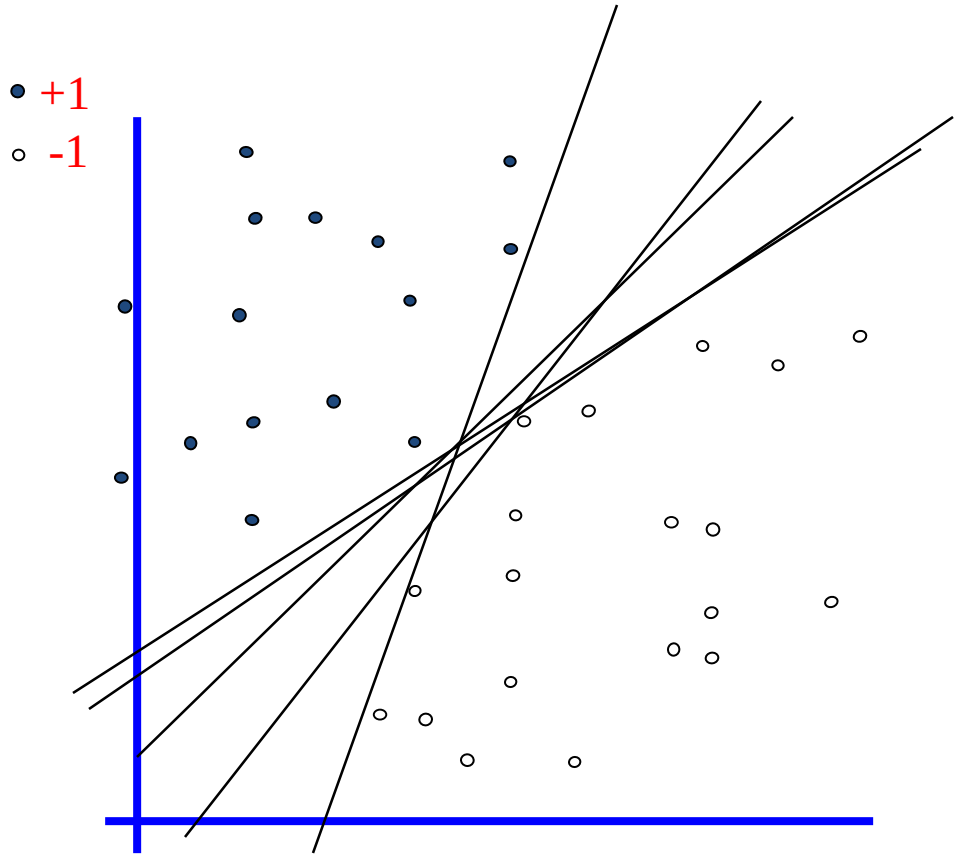
- SVMدر سال ۱۹۹۲ توسط Vapnik معرفی شده و بر پایه statistical learning theory بنا گردیده است.

- هدف این دسته الگوریتم ها تشخیص و متمایز کردن الگوهای پیچیده در داده هاست (از طریق خوشه بندی، دسته بندی، رتبه بندی، پاکسازی و غیره)

ایده اصلی SVM

- با فرض اینکه دسته ها بصورت خطی جداپذیر باشند، ابرصفحه هائی با حداکثر حاشیه (maximum margin) را بدست می آورد که دسته ها را جدا کنند.
- در مسایلی که داده ها بصورت خطی جداپذیر نباشند داده ها به فضای با ابعاد بیشتر نگاشت پیدا می کنند تا بتوان آن ها را در این فضای جدید بصورت خطی جدا نمود.

شما چگونه این داده‌ها را طبقه‌بندی می کنید؟

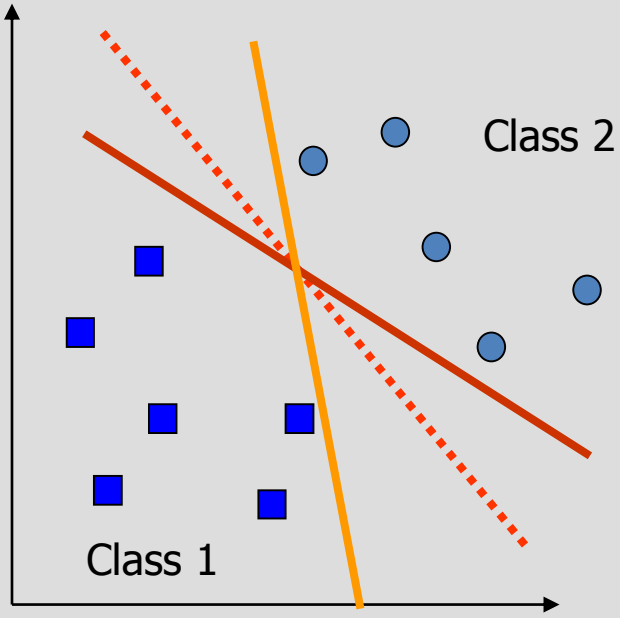


تمام این خطوط خوب هستند

اما، بهترین آنها کدام است؟

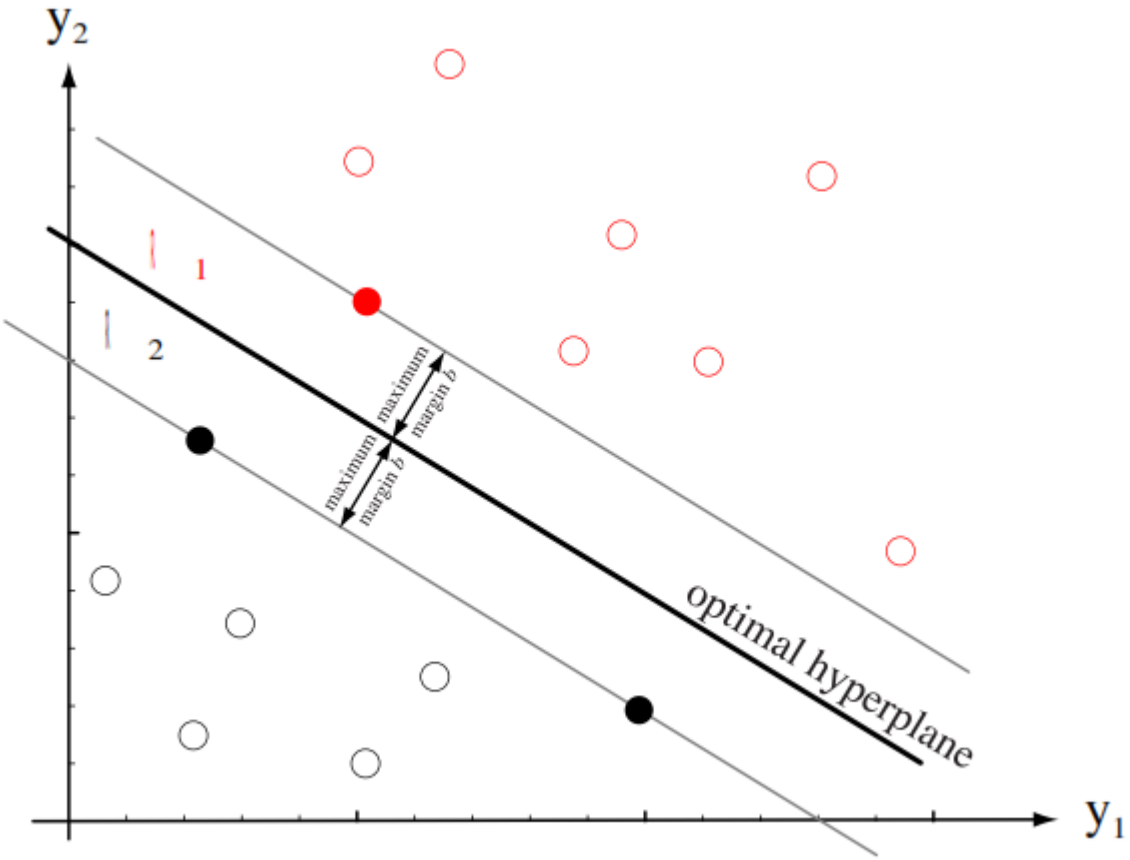
What is a good Decision Boundary?

- ✓ Consider a two-class, linearly separable classification problem
- ✓ Many decision boundaries!
- ✓ Are all decision boundaries equally good?

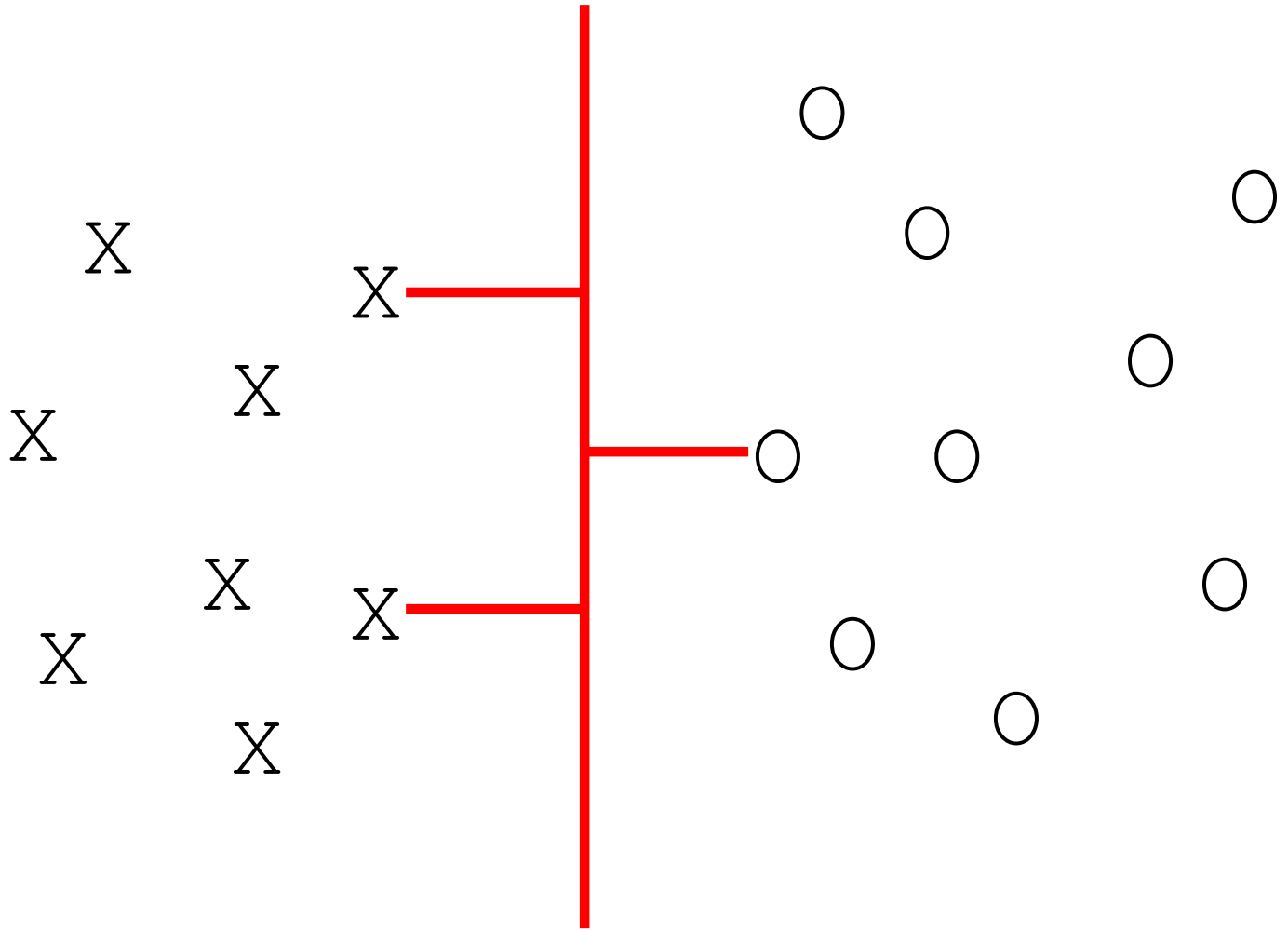


Linear Maximal Margin Classifier for Linearly Separable Data

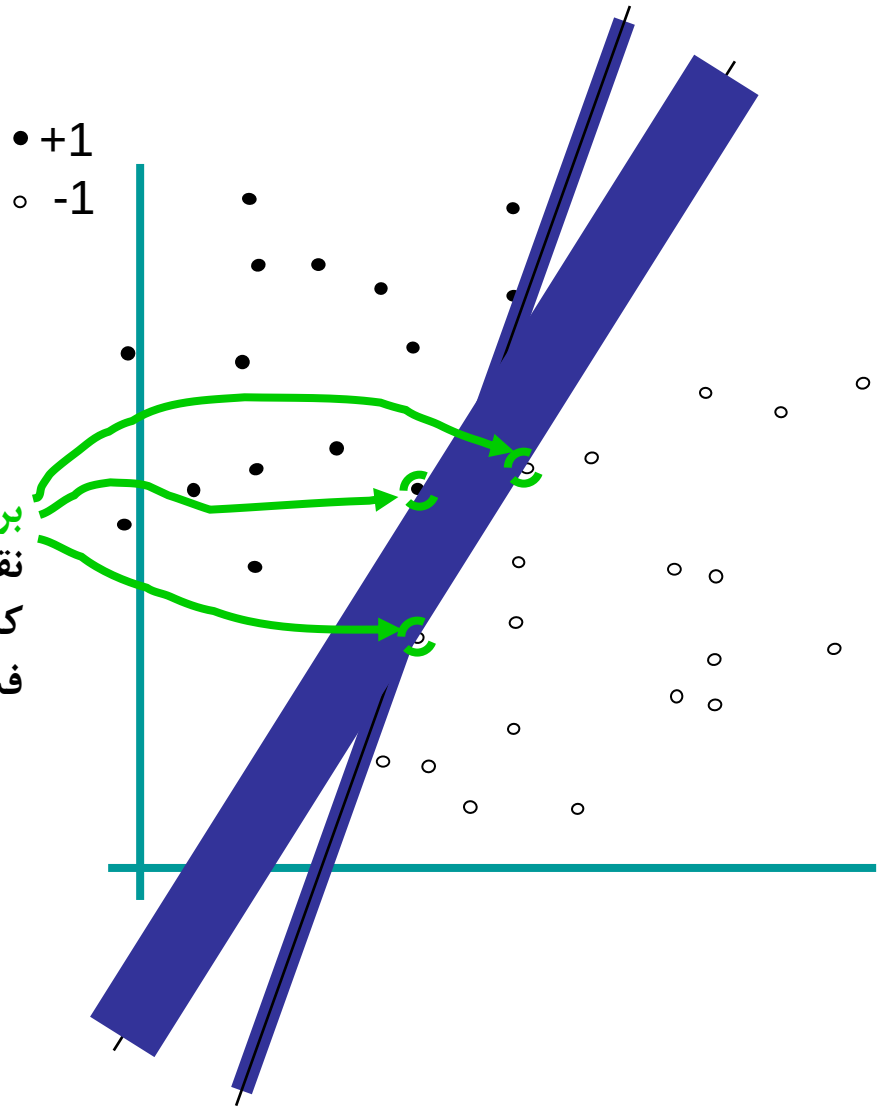
The decision boundary should be as far away from the data of both classes as possible



ماکزیمم کردن حاشیه



یک حاشیه از طبقه‌بند خطی را، بعنوان پهنایی که مرز بدون برخورد با نقاط داده می‌تواند بسط یابد، تعریف می‌کنیم.

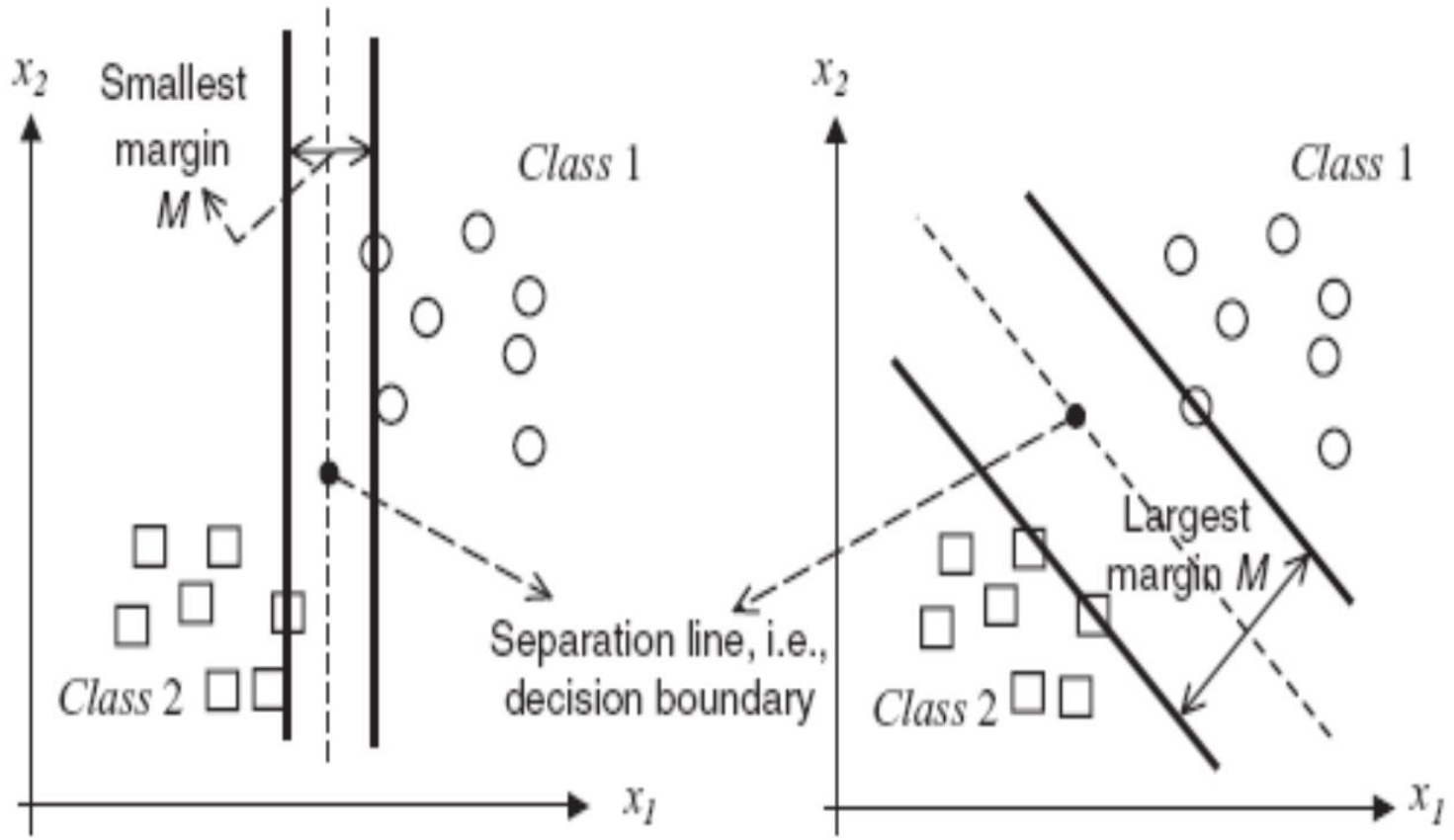


بردارهای پشتیبان نقاط داده‌ای هستند که حاشیه با آنها فشرده می‌شود.

طبقه‌بند خطی حداکثر حاشیه ساده‌ترین نوع از SVM موسوم به LSVM یا SVM خطی می‌باشد.

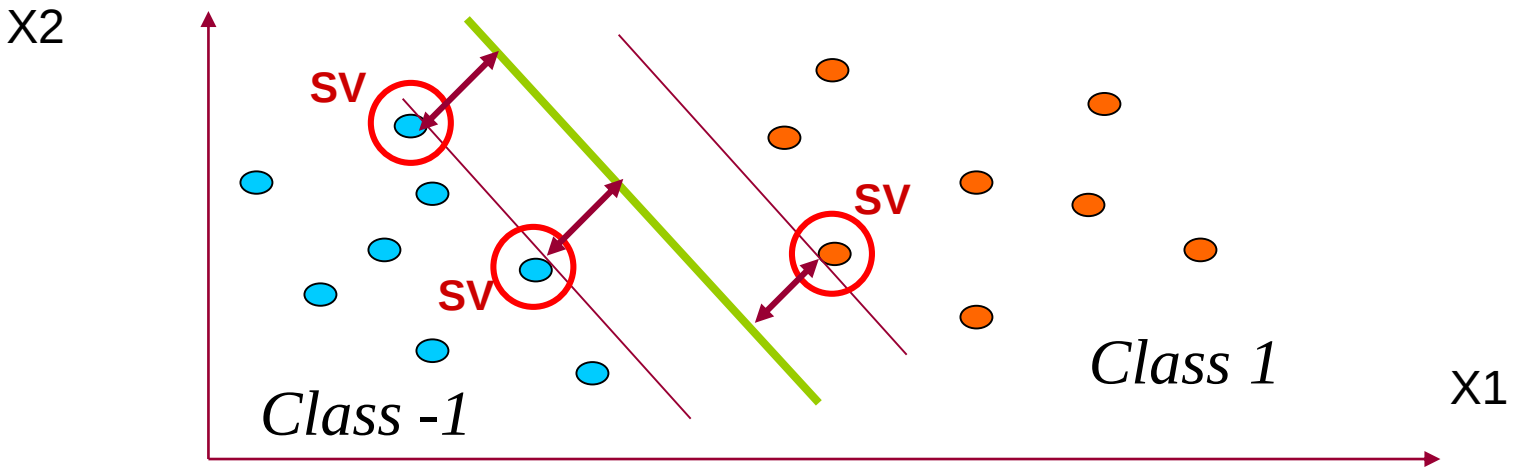
✓هدف طراحی یک ابرصفحه تصمیم خطی با حداکثر حاشیه نسبت به دو کلاس می‌باشد.

✓حداکثر حاشیه به جهت ابرصفحه وابسته است و هدف جستجوی جهتی با حداکثر حاشیه ممکن می‌باشد.



بردار پشتیبان

- نزدیکترین داده های آموزشی به ابر صفحه های جدا کننده بردار پشتیبان نامیده می شوند



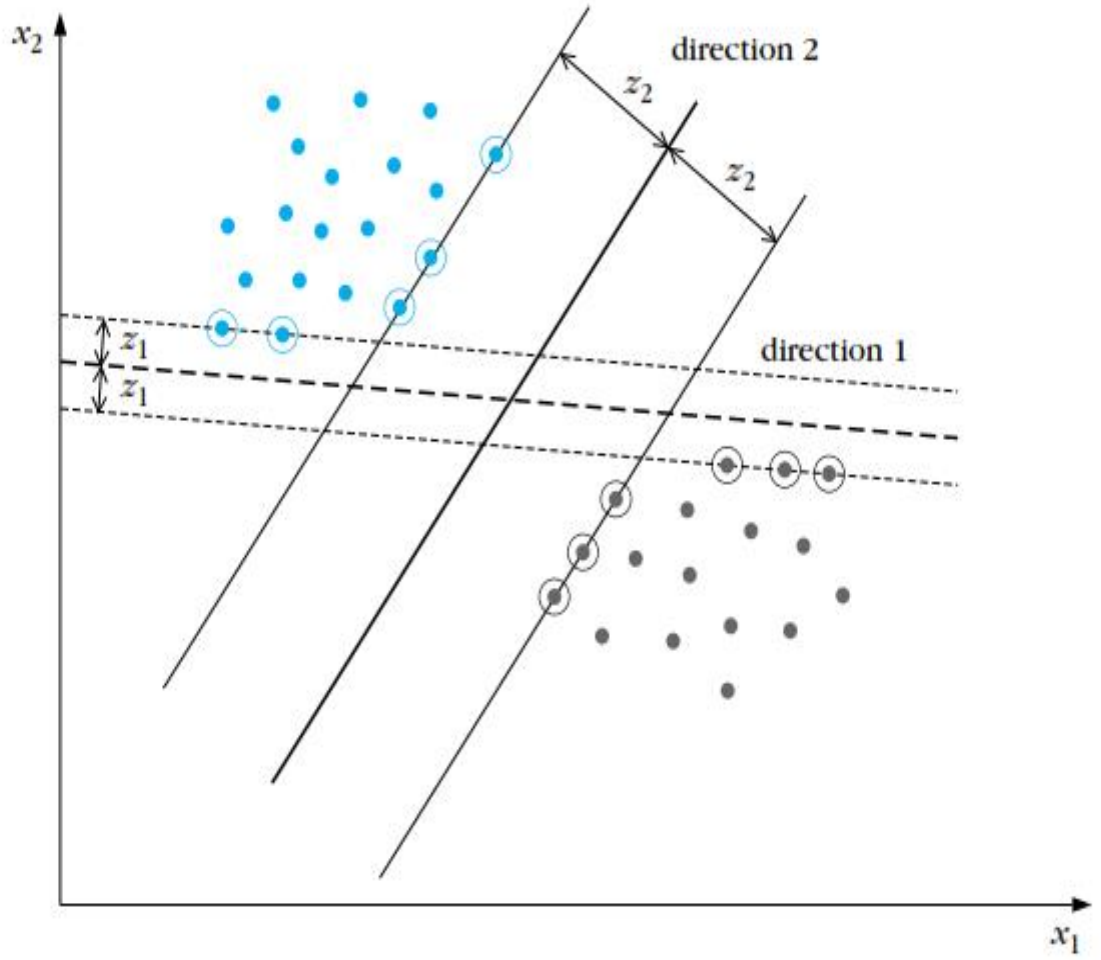
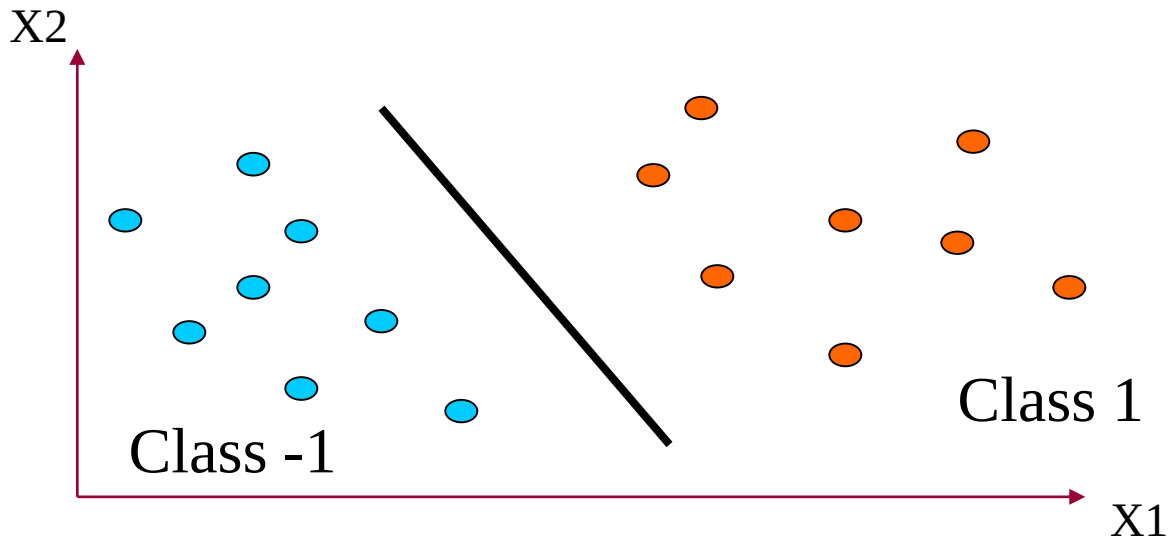


FIGURE 3.10

An example of a linearly separable two-class problem with two possible linear classifiers.

یافتن خط یا ابر صفحه جدا کننده



- هدف: پیدا کردن بهترین خط (ابر صفحه) که دو دسته را از هم جدا کند.
در حالت دو بعدی معادله این خط بصورت زیر است:

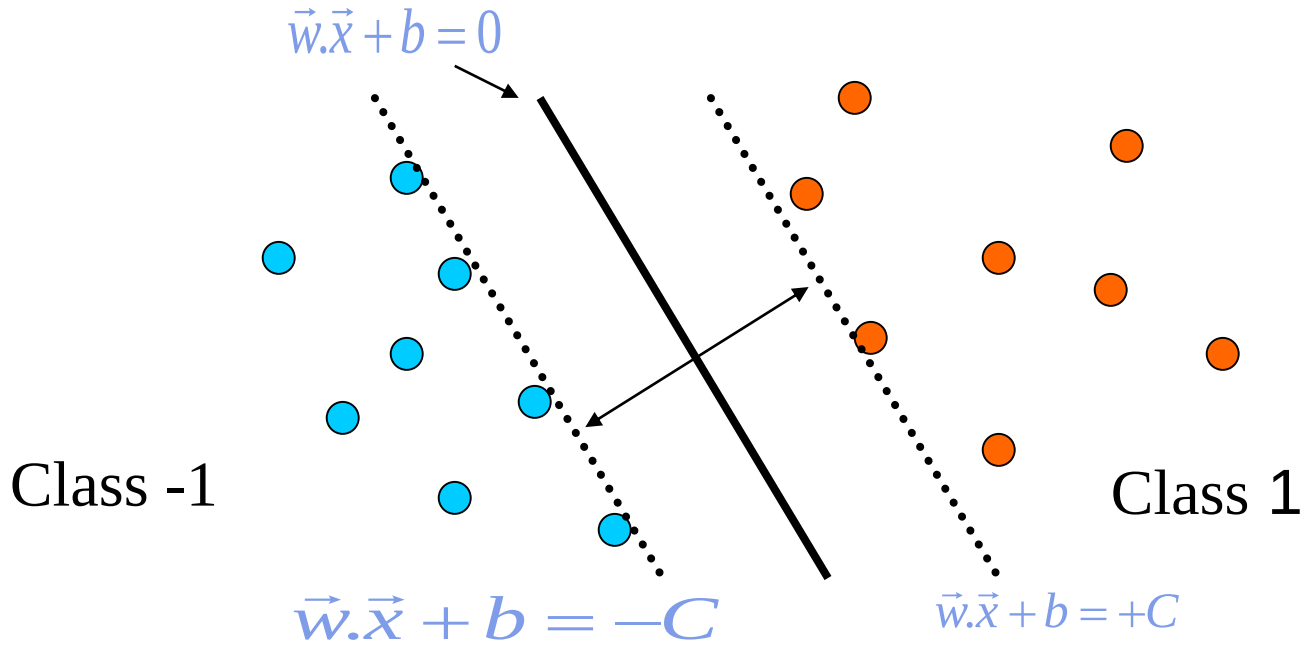
$$w_1 X_1 + w_2 X_2 + b = 0$$

- در حالت n بعدی خواهیم داشت:

$$\sum_{i=0}^{n} wi.xi + b = 0$$

ایده اصلی SVM برای جدا سازی دسته ها

- دو صفحه مرزی موازی با صفحه دسته بندی رسم کرده و آندو را آنقدر از هم دور می کنیم که به داده ها برخورد کنند.
- صفحه دسته بندی که بیشترین فاصله را از صفحات مرزی داشته باشد، بهترین جدا کننده خواهد بود.



- طبق قضیه ای در تئوری یادگیری اگر مثال های آموزشی بدرستی طبقه بندی شده باشند، از بین جداسازهای خطی، آن جداسازی که حاشیه داده های آموزشی را حداکثر می کند خطای تعمیم را حداقل خواهد کرد.

چرا حداکثر حاشیه؟

- به نظر میرسد که مطمئن ترین راه باشد.
- تئوری هائی برمبنای *Vapnik–Chervonenkis)VC dimension* (*dimension*) وجود دارد که مفید بودن آنرا اثبات می کند.
- بطور تجربی این روش خیلی خوب جواب داده است.

- در صورت استفاده مناسب از *SVM* این الگوریتم قدرت تعمیم خوبی خواهد داشت.

- علیرغم داشتن ابعاد زیاد (*high dimensionality*) از overfitting پرهیز می کند. این خاصیت ناشی از بهینه سازی(optimization) این الگوریتم است.

- فشرده سازی اطلاعات

بجای داده های آموزشی از بردارهای پشتیبان استفاده میکند.

- هر ابرصفحه با یک ضریب مقیاس تعیین می‌شود. برای مستقل‌نمودن نتیجه از این ضرایب، می‌توان بردار وزن و آستانه را طوری مقیاس نمود تا نزدیکترین نقاط در کلاس یک و دو دارای $g(x)$ بترتیب ۱ و ۱- باشند.

- فاصله هر نقطه تا ابرصفحه برابر است با:

$$z = \frac{|g(\boldsymbol{x})|}{\|\boldsymbol{w}\|}$$

1. Having a margin of $\frac{1}{\|\boldsymbol{w}\|} + \frac{1}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|}$
2. Requiring that

$$\boldsymbol{w}^T\boldsymbol{x} + w_0 \geq 1, \qquad \forall \boldsymbol{x} \in \omega_1$$
$$\boldsymbol{w}^T\boldsymbol{x} + w_0 \leq -1, \qquad \forall \boldsymbol{x} \in \omega_2$$

- برای هر بردار ویژگی x_i برچسب کلاس بصورت y_i ($+1$ for ω_1, -1 for ω_2.) تعریف‌می‌شود. کار ما محاسبه بردار وزن و آستانه با معیار زیر خواهدبود:

$$\text{minimize} \quad J(\boldsymbol{w}, w_0) \equiv \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) \geq 1, \quad i = 1, 2, \ldots, N$$

- حداقل‌کردن نُرم دو فوق منجر به بیشینه‌نمودن حاشیه می‌شود. این کار یک بهینه‌سازی غیرخطی با تعدادی قیود نامساوی خطی است.

- با شرایط Karush-Kuhn-Tucker (KKT) مسئله بالا حل می‌شود:

$$\frac{\partial}{\partial \boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0}$$

$$\frac{\partial}{\partial w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = 0$$

$$\lambda_i \geq 0, \quad i = 1, 2, \ldots, N$$

$$\lambda_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1] = 0, \quad i = 1, 2, \ldots, N$$

✓در روابط بالا، λ_i ضریب لاگرانژ برای تابع لاگرانژ است.

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} - \sum_{i=1}^{N} \lambda_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1]$$

✓نهایتا با ترکیب این روابط داریم:

$$\boldsymbol{w} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

- برای حل معادلات بالا با استفاده از دوگان لاگرانژی می‌توان بردار وزن را بدست آورد:

$$\text{maximize} \quad \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda})$$

$$\text{subject to} \quad \boldsymbol{w} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

$$\boldsymbol{\lambda} \geq \mathbf{0}$$

$$\max_{\boldsymbol{\lambda}} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \right)$$

$$\text{subject to} \quad \sum_{i=1}^{N} \lambda_i y_i = 0$$

$$\boldsymbol{\lambda} \geq \mathbf{0}$$

- باتوجه به ورود بردارهای آموزش بصورت جفتی در مسئله با ضرب داخلی، تابع هزینه بطور کامل به ابعاد فضای ورودی وابسته نیست.

- علیرغم یکتایی ابرصفحه‌ها، هیچ تضمینی برای یکتایی ضرایب‌لاگرانژ و به‌تبع آن، بردارهای وزن برحسب بردارهای پشتیبان وجود ندارد.

مثال:

مسئله دو کلاسه که شامل نقاط زیر می باشد را درنظر بگیرید

w_1: $[1, 1]^T, [1, -1]^T$

w_2: $[-1, 1]^T, [-1, -1]$

✓باتوجه به هندسه ساده مسئله وبه کمک روش SVM خط تصمیم بصورت $x_{1=0}$ بدست می‌آید. همچنین با کمی تامل روی شکل می توان دریافت که خط بهینه

$$g(x) = w_1x_1 + w_2x_2 + w_0$$

به ازای $w_2 = w_0 = 0$ و $w_1 = 1$ بدست آمده است. یعنی:

$$g(x) = x_1 = 0$$

بنابراین در این حالت، هر چهار نقطه بردار پشتیبان(SV) محسوب می‌شوند. با انتخاب هر جهت دیگر، حاشیه کمتر از ۱ خواهدبود.

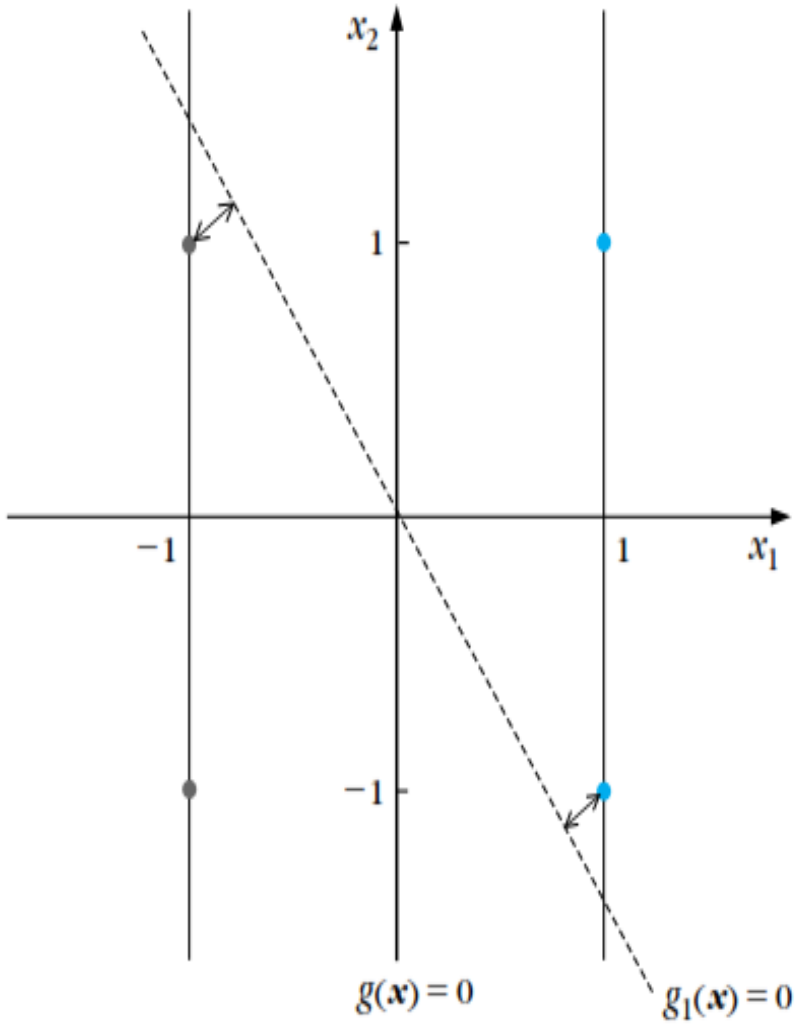


FIGURE 3.12

In this example all four points are support vectors. The margin associated with $g_1(\boldsymbol{x}) = 0$ is smaller compared to the margin defined by the optimal $g(\boldsymbol{x}) = 0$.

✓با نوشتن معادلات ریاضی از روی قیود KKT داریم:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) \geq 1$$

$$\begin{aligned} w_1 + w_2 + w_0 - 1 &\geq 0 \\ w_1 - w_2 + w_0 - 1 &\geq 0 \\ w_1 - w_2 - w_0 - 1 &\geq 0 \\ w_1 + w_2 - w_0 - 1 &\geq 0 \end{aligned}$$

توابع لاگرانژ متناظر بصورت زیر خواهند بود:

$$\begin{aligned} \mathcal{L}(w_2, w_1, w_0, \boldsymbol{\lambda}) = \frac{w_1^2 + w_2^2}{2} &- \lambda_1(w_1 + w_2 + w_0 - 1) \\ &- \lambda_2(w_1 - w_2 + w_0 - 1) \\ &- \lambda_3(w_1 - w_2 - w_0 - 1) \\ &- \lambda_4(w_1 + w_2 - w_0 - 1) \end{aligned}$$

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{w}^T \boldsymbol{w} - \sum_{i=1}^{N} \lambda_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1]$$

قیود KKT بصورت زیر می‌با شند:

$$\frac{\partial \mathcal{L}}{\partial w_1} = 0 \;\Rightarrow\; w_1 = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$$

$$\frac{\partial \mathcal{L}}{\partial w_2} = 0 \;\Rightarrow\; w_2 = \lambda_1 + \lambda_4 - \lambda_2 - \lambda_3$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \;\Rightarrow\; \lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 = 0$$

$$\lambda_1(w_1 + w_2 + w_0 - 1) = 0$$

$$\lambda_2(w_1 - w_2 + w_0 - 1) = 0$$

$$\lambda_3(w_1 - w_2 - w_0 - 1) = 0$$

$$\lambda_4(w_1 + w_2 - w_0 - 1) = 0$$

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$$

$$\frac{\partial}{\partial \boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0}$$

$$\frac{\partial}{\partial w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = 0$$

$$\lambda_i \geq 0, \quad i = 1, 2, \ldots, N$$

$$\lambda_i[y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1] = 0, \quad i = 1, 2, \ldots, N$$

✓باتوجه به دانستن یکتایی حل مسئله، با جایگزینی مقادیر $w_1 = 1, w_2 = w_0 = 0$ در معادلات خواهیم داشت:

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$$
$$\lambda_1 + \lambda_4 - \lambda_2 - \lambda_3 = 0$$
$$\lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 = 0$$

✓سه معادله با ۴ مجهول، منجر به بینهایت جواب برای ضرایب لاگرانژ می شود. با این وجود تمامی این جواب ها منجر به یک ابرصفحه جدا کننده منحصر بفرد می شوند.

Example 3.6

Figure 3.13 shows a set of training data points residing in the two-dimensional space and divided into two nonseparable classes. The full line in Figure 3.13a is the resulting hyperplane using Platt's algorithm and corresponds to the value $C = 0.2$. Dotted lines meet the conditions given in (3.82) and define the margin that separates the two classes, for those points with

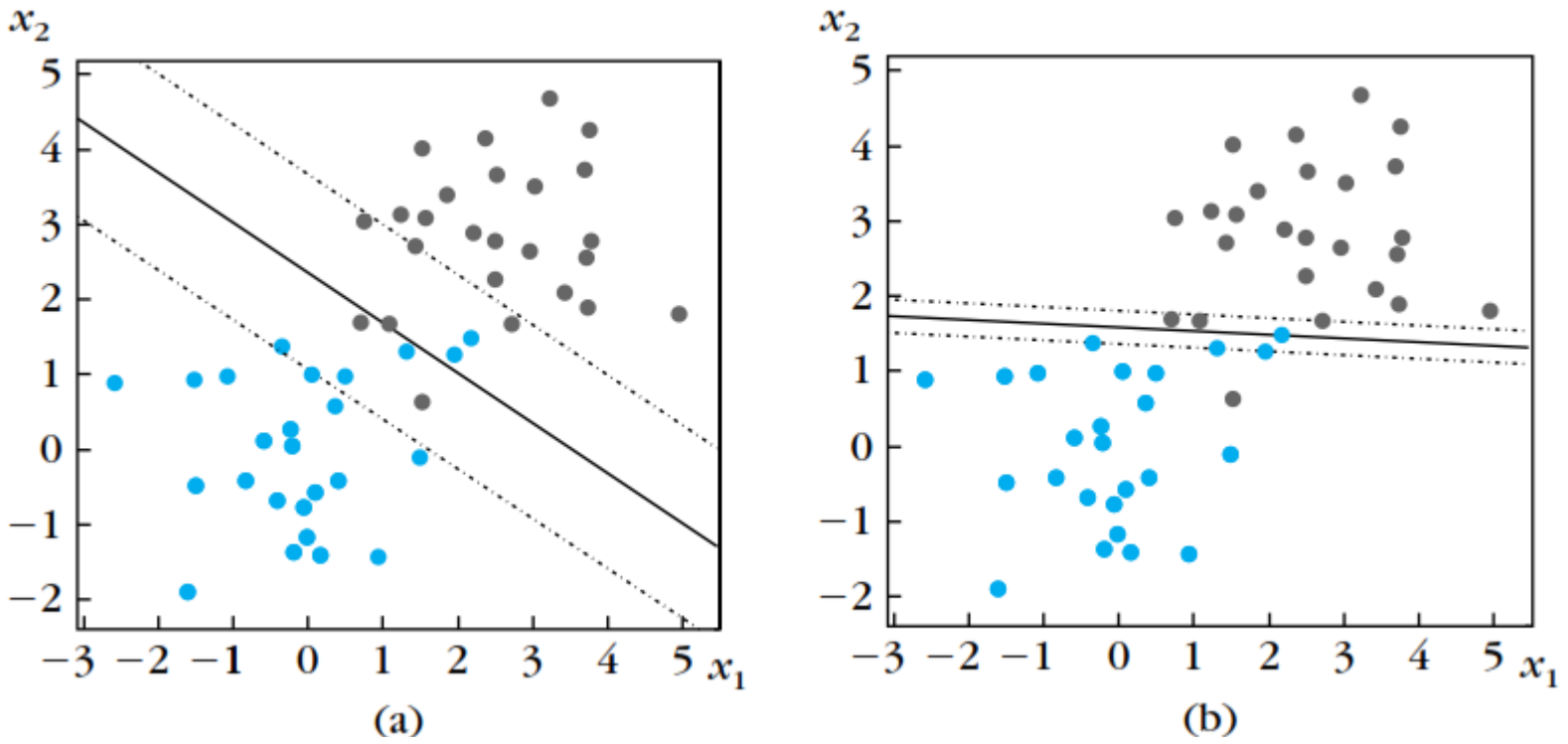
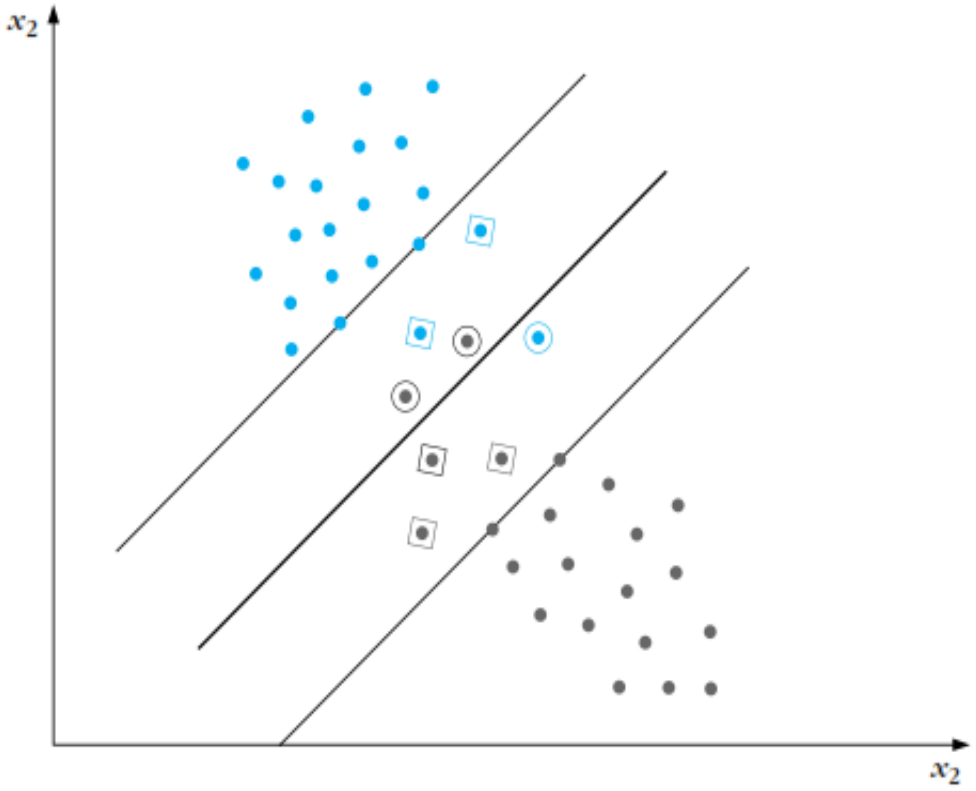


FIGURE 3.13

An example of two nonseparable classes and the resulting SVM linear classifier (full line) with the associated margin (dotted lines) for the values (a) $C = 0.2$ and (b) $C = 1000$. In the latter case, the location and direction of the classifier as well as the width of the margin have changed in order to include a smaller number of points inside the margin.

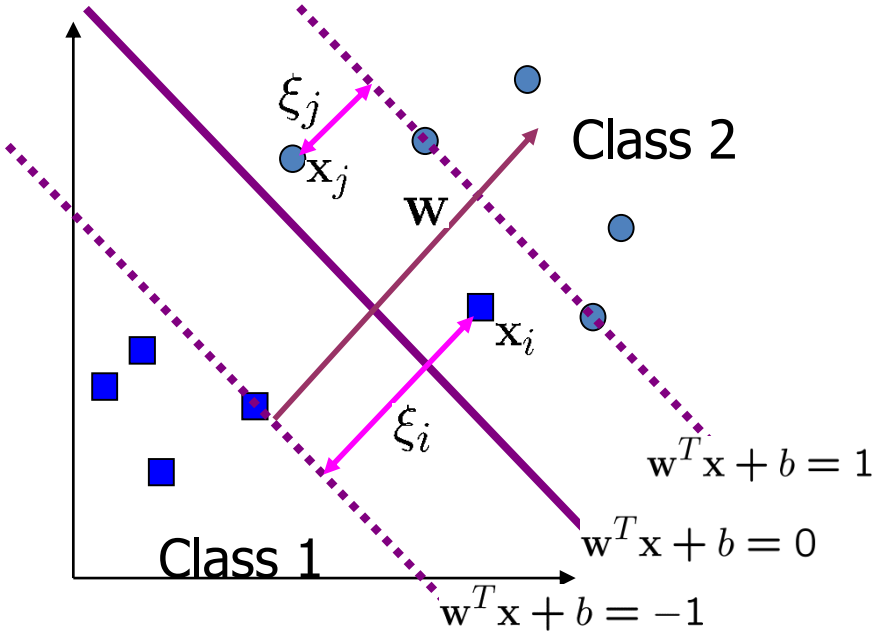
کلاس‌های تفکیک‌ناپذیر(داده هائی که بصورت خطی جدا پذیر نیستند)

یک فرض بسیار قوی در SVM این بود که داده ها بصورت خطی جداپذیر باشند.
در حالیکه در عمل در بسیاری مواقع این فرض برقرارنیست.



افزودن متغیر های *slack*

- یک راه حل این است که اندکی کوتاه آمده و مقداری خطا در دسته بندی را بپذیریم!
- این کار با معرفی متغیر ξ_i انجام می شود که نشانگر تعداد نمونه هائی است که توسط تابع $\boldsymbol{w}^T\boldsymbol{x}+b$ غلط ارزیابی می شوند.



افزودن متغیر های *slack*

- با معرفی متغیر ξ_i, $i=1, 2, \ldots, N$, محدودیت های قبلی ساده تر شده و رابطه

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) \geq 1, \quad i = 1, 2, \ldots, N$$

- بصورت زیر تغییر می کند:

$$y_i[\boldsymbol{w}^T \boldsymbol{x}_i + w_0] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N$$

- در حالت ایده آل همه این متغیر ها باید صفر باشند.

در اینصورت مسئله بهینه سازی تبدیل به مساله زیر می شود:

$$\text{minimize} \quad J(\boldsymbol{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad y_i[\boldsymbol{w}^T\boldsymbol{x}_i + w_0] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, N$$

که در آن $C > 0$ می باشد. جمله اضافه شده سعی دارد تا حد امکان همه متغیرهای $slack$ را کوچک نماید.

دوآل متناظر با این مساله به صورت زیر می باشد:

$$\text{maximize} \quad \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\mu})$$

$$\text{subject to} \quad \boldsymbol{w} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

$$C - \mu_i - \lambda_i = 0, \quad i = 1, 2, \ldots, N$$

$$\lambda_i \geq 0, \ \mu_i \geq 0, \quad i = 1, 2, \ldots, N$$