دانشگاه سمنان

فصل سوم
شناسایی الگو
طبقه‌بندی کننده‌های خطی، ماشین دسته‌بندی کننده بردار پشتیبان
Linear Classifiers
Support Vector Machines

محمدجواد فدائی‌اسلام

OPTIMIZATION FOR CONSTRAINED PROBLEMS, EQUALITY CONSTRAINTS

- همراه با تابع هزینه که باید بهینه‌سازی کنیم، مجموعه‌ای از محدودیت‌ها را داریم که باید به آنها پایبند باشیم.

$$\begin{aligned} &\text{minimize} && J(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) = 0, \quad i = 1, 2, \ldots, m \end{aligned}$$

- برای حل این مشکل، باید تابع لاگرانژ را به حداقل برسانیم.

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta})$$

- برای به حداقل رساندن تابع، گرادیان لاگرانژ باید صفر باشد.

$$\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{0}$$

Example Maximize $f(x_1, x_2) = x_1 x_2$ subject to $h(x_1, x_2) \equiv x_1 + 4x_2 = 16$.

Solution: Form the Lagrangian

$$L(x_1, x_2) = x_1 x_2 - \lambda (x_1 + 4x_2 - 16)$$

The first order conditions are

$$\frac{dL}{dx_1} = x_2 - \lambda = 0$$

$$\frac{dL}{dx_2} = x_1 - 4\lambda = 0$$

$$\frac{dL}{d\lambda} = x_1 + 4x_2 - 16 = 0$$

$$(x_1, x_2, \lambda) = (8, 2, 2)$$

EXCERCISE

Maximize $f(x, y, z) = xyz$ subject to

$$h_1(x, y, z) \equiv x^2 + y^2 = 1 \text{ and } h_2(x, y, z) \equiv x + z = 1.$$

OPTIMIZATION FOR CONSTRAINED PROBLEMS INEQUALITY CONSTRAINTS

مساله را می توان به صورت زیر مطرح کرد:

$$\begin{aligned} &\text{minimize} && J(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) \geq 0, \quad i = 1, 2, \ldots, m \end{aligned} \qquad \text{(C.29)}$$

هر یک از محدودیت‌ها یک ناحیه در R^l تعریف می‌کند. اشتراک تمام این نواحی، ناحیه ای که کمینه با محدودیت θ_*، در آن قرار گیرد را مشخص می‌کند. این به عنوان منطقه امکان‌پذیر و نقاط موجود در آن (راه حل های کاندید) به عنوان نقاط امکان‌پذیر شناخته می‌شود.

در نقطه کمینه باید شرایط KKT برقرار باشد.

KARUSH KUHN TUCKER (KKT) CONDITIONS

- اگر θ_* نقطه‌ای باشد که شروط مساله در آن ارضا شود، آنگاه بردار λ از ضرایب لاگرانژ وجود دارد که موارد زیر برای آنها صادق است:

$$(1) \quad \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) = \mathbf{0}$$

$$(2) \quad \lambda_i \geq 0, \quad i = 1, 2, \ldots, m$$

$$(3) \quad \lambda_i f_i(\boldsymbol{\theta}_*) = 0, \quad i = 1, 2, \ldots, m$$

LAGRANGIAN DUALITY

$$\begin{aligned} &\text{minimize} && J(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) \geq 0, \quad i = 1, 2, \ldots, m \end{aligned}$$

The Lagrangian function is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta}) \tag{C.35}$$

Let

$$\mathcal{L}^*(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.36}$$

However, since $\boldsymbol{\lambda} \geq \mathbf{0}$ and $f_i(\boldsymbol{\theta}) \geq 0$, the maximum value of the Lagrangian occurs if the summation in (C.35) is zero (either $\lambda_i = 0$ or $f_i(\boldsymbol{\theta}) = 0$ or both) and

$$\mathcal{L}^*(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) \tag{C.37}$$

Therefore our original problem is equivalent with

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.38}$$

As we already know, the dual problem of the above is

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.39}$$

Lagrangian Duality

Wolfe Dual Representation

A convex programming problem is equivalent to

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda})$$

$$\text{subject to } \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{0}$$

The last equation guarantees that $\boldsymbol{\theta}$ is a minimum of the Lagrangian.

مثال

Example C.1

Consider the quadratic problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} \\ \text{subject to} \quad & A\boldsymbol{\theta} \geq \boldsymbol{b} \end{aligned}$$

This is a convex programming problem; hence the Wolfe dual representation is valid:

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\lambda}^T(A\boldsymbol{\theta} - \boldsymbol{b}) \\ \text{subject to} \quad & \boldsymbol{\theta} - A^T\boldsymbol{\lambda} = \mathbf{0} \end{aligned}$$

For this example, the equality constraint has an analytic solution (this is not, however, always possible). Solving with respect to $\boldsymbol{\theta}$, we can eliminate it from the maximizing function and the resulting dual problem involves only the Lagrange multipliers,

$$\max_{\boldsymbol{\lambda}} \left\{ -\frac{1}{2}\boldsymbol{\lambda}^T AA^T\boldsymbol{\lambda} + \boldsymbol{\lambda}^T\boldsymbol{b} \right\}$$

$$\text{subject to} \quad \boldsymbol{\lambda} \geq \mathbf{0}$$

This is also a quadratic problem but the set of constraints is now simpler.

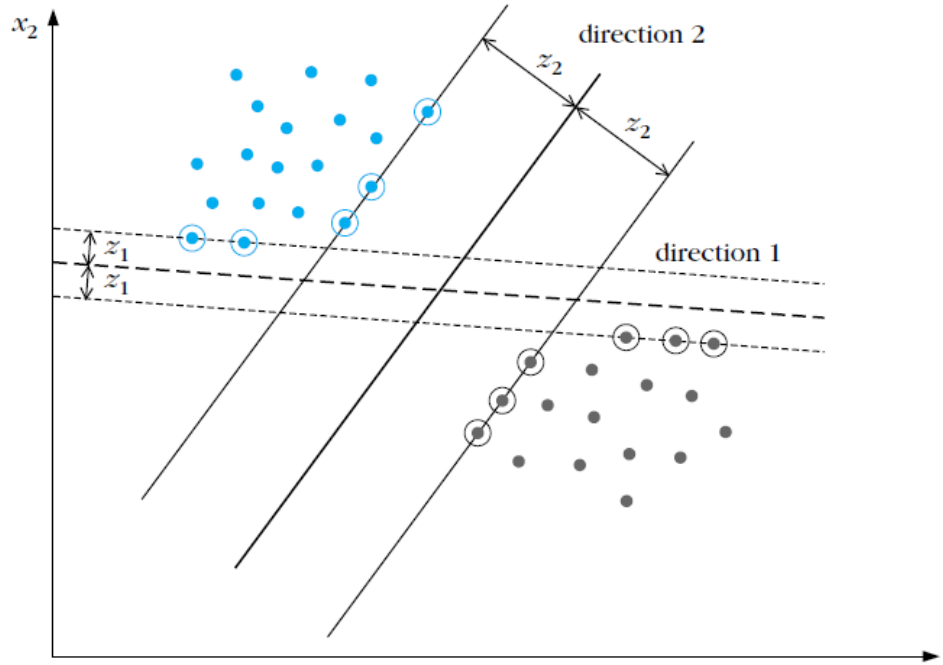
جداکننده خطی

The goal is to design a hyperplane

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = 0$$

that classifies correctly all the training vectors. As we have already discussed in Section 3.3, such a hyperplane is not unique. The perceptron algorithm may converge to any one of the possible solutions.

GENERALIZATION



linearly separable two-class problem with two possible linear classifiers

هدف، جستجوی خط جداکننده‌ای است که حداکثر حاشیه ممکن را دهد.

تعمیم، به قابلیت طبقه‌بندی کننده برای عملکرد رضایت‌بخش بر روی داده‌های آزمایش اشاره دارد.

SVM STEPS: SCALE

the distance of a point from a hyperplane is given by

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

We can now scale $\boldsymbol{w}, w_0$ so that the value of $g(\boldsymbol{x})$, at the nearest points in ω_1, ω_2 (circled in Figure 3.10), is equal to 1 for ω_1 and, thus, equal to -1 for ω_2. This is equivalent with

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

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

SVM STEPS: MINIMIZING THE NORM

Compute the parameters $\boldsymbol{w}, w_0$ of the hyperplane so that to:

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

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

Obviously, minimizing the norm makes the margin maximum. This is a nonlinear (quadratic) optimization task subject to a set of linear inequality constraints.

SVM Steps: KKT condition and Lagrangian Function

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

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

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

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

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

KKT Conditions

(1) $\dfrac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) = \mathbf{0}$

(2) $\lambda_i \geq 0, \quad i = 1, 2, \ldots, m$

(3) $\lambda_i f_i(\boldsymbol{\theta}_*) = 0, \quad i = 1, 2, \ldots, m$

SVM Steps: KKT condition and Lagrangian Function

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

$$\frac{\partial}{\partial \boldsymbol{w}}\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0} \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N}\lambda_i y_i \boldsymbol{x}_i$$

$$\frac{\partial}{\partial w_0}\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = 0 \Longrightarrow \sum_{i=1}^{N}\lambda_i y_i = 0$$

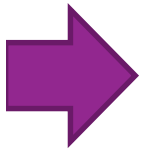
WOLFE DUAL

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

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

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

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



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

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

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

SUPPORT VECTORS

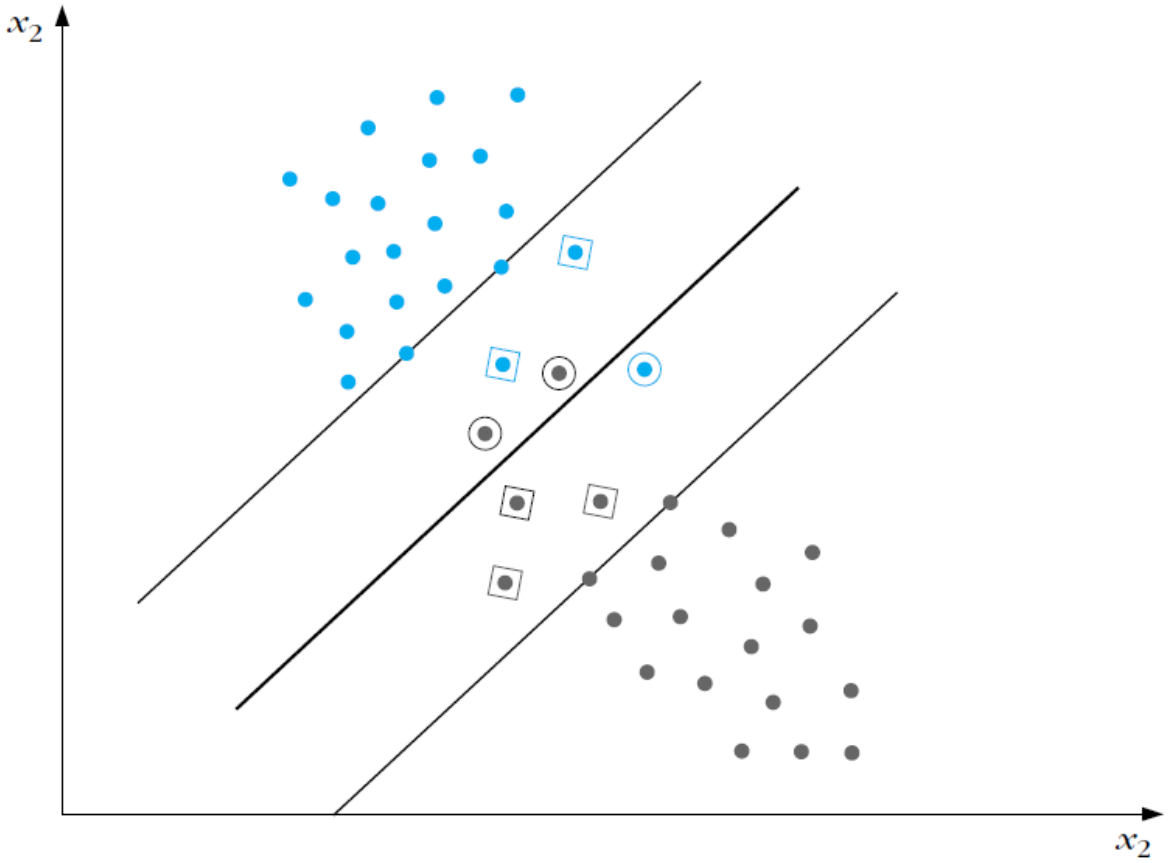
The Lagrange multipliers can be either zero or positive (Appendix C). Thus, the vector parameter $\boldsymbol{w}$ of the optimal solution is a linear combination of $N_s \leq N$ feature vectors that are associated with $\lambda_i \neq 0$. That is,

$$\boldsymbol{w} = \sum_{i=1}^{N_s} \lambda_i y_i \boldsymbol{x}_i \tag{3.81}$$

These are known as *support vectors* and the optimum hyperplane classifier as a *support vector machine* (SVM). As it is pointed out in Appendix C, a nonzero Lagrange multiplier corresponds to a so called active constraint. Hence, as the set of constraints in (3.77) suggests for $\lambda_i \neq 0$, *the support vectors lie on either of the two hyperplanes*, that is,

$$\boldsymbol{w}^T \boldsymbol{x} + w_0 = \pm 1 \tag{3.82}$$

Non-separable classes



THE TRAINING FEATURE VECTORS NOW BELONG TO ONE OF THE FOLLOWING THREE CATEGORIES

- Vectors that fall outside the band and are correctly classified. These vectors comply with the constraints in (3.73).
- Vectors falling inside the band and are correctly classified. These are the points placed in squares in Figure 3.11, and they satisfy the inequality

$$0 \leq y_i(\boldsymbol{w}^T \boldsymbol{x} + w_0) < 1$$

- Vectors that are misclassified. They are enclosed by circles and obey the inequality

$$y_i(\boldsymbol{w}^T \boldsymbol{x} + w_0) < 0$$

All three cases can be treated under a single type of constraints by introducing new set of variables, namely,

$$y_i[\boldsymbol{w}^T \boldsymbol{x} + w_0] \geq 1 - \xi_i$$

C-SVM

The first category of data corresponds to $\xi_i = 0$, the second to $0 < \xi_i \leq 1$, and the third to $\xi_i > 1$. The variables ξ_i are known as *slack variables*. The optimizing task becomes more involved, yet it falls under the same rationale as before. The goal now is to make the margin as large as possible but at the same time to keep the number of points with $\xi > 0$ as small as possible. In mathematical terms, this is equivalent to adopting to minimize the cost function

$$J(\boldsymbol{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} I(\xi_i) \tag{3.91}$$

where $\boldsymbol{\xi}$ is the vector of the parameters ξ_i and

Xi

$$I(\xi_i) = \begin{cases} 1 & \xi_i > 0 \\ 0 & \xi_i = 0 \end{cases} \tag{3.92}$$

C-SVM

The parameter C is a positive constant that controls the relative influence of the two competing terms. However, optimization of the above is difficult since it involves a discontinuous function $I(\cdot)$. As it is common in such cases, we choose to optimize a closely related cost function, and the goal becomes

$$\text{minimize} \quad J(\boldsymbol{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{3.93}$$

$$\text{subject to} \quad y_i[\boldsymbol{w}^T\boldsymbol{x}_i + w_0] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N \tag{3.94}$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, N \tag{3.95}$$

C-SVM

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i$$

$$- \sum_{i=1}^{N}\lambda_i[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 + \xi_i] \qquad (3.96)$$

The corresponding Karush–Kuhn–Tucker conditions are

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \mathbf{0} \quad \text{or} \quad \boldsymbol{w} = \sum_{i=1}^{N}\lambda_i y_i \boldsymbol{x}_i \qquad (3.97)$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \quad \text{or} \quad \sum_{i=1}^{N}\lambda_i y_i = 0 \qquad (3.98)$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \quad \text{or} \quad C - \mu_i - \lambda_i = 0, \quad i = 1, 2, \ldots, N \qquad (3.99)$$

$$\lambda_i[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 + \xi_i] = 0, \quad i = 1, 2, \ldots, N \qquad (3.100)$$

$$\mu_i\xi_i = 0, \quad i = 1, 2, \ldots, N \qquad (3.101)$$

$$\mu_i \geq 0, \quad \lambda_i \geq 0, \quad i = 1, 2, \ldots, N \qquad (3.102)$$

C-SVM

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

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

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

Wolfe dual representation now becomes

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

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

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

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

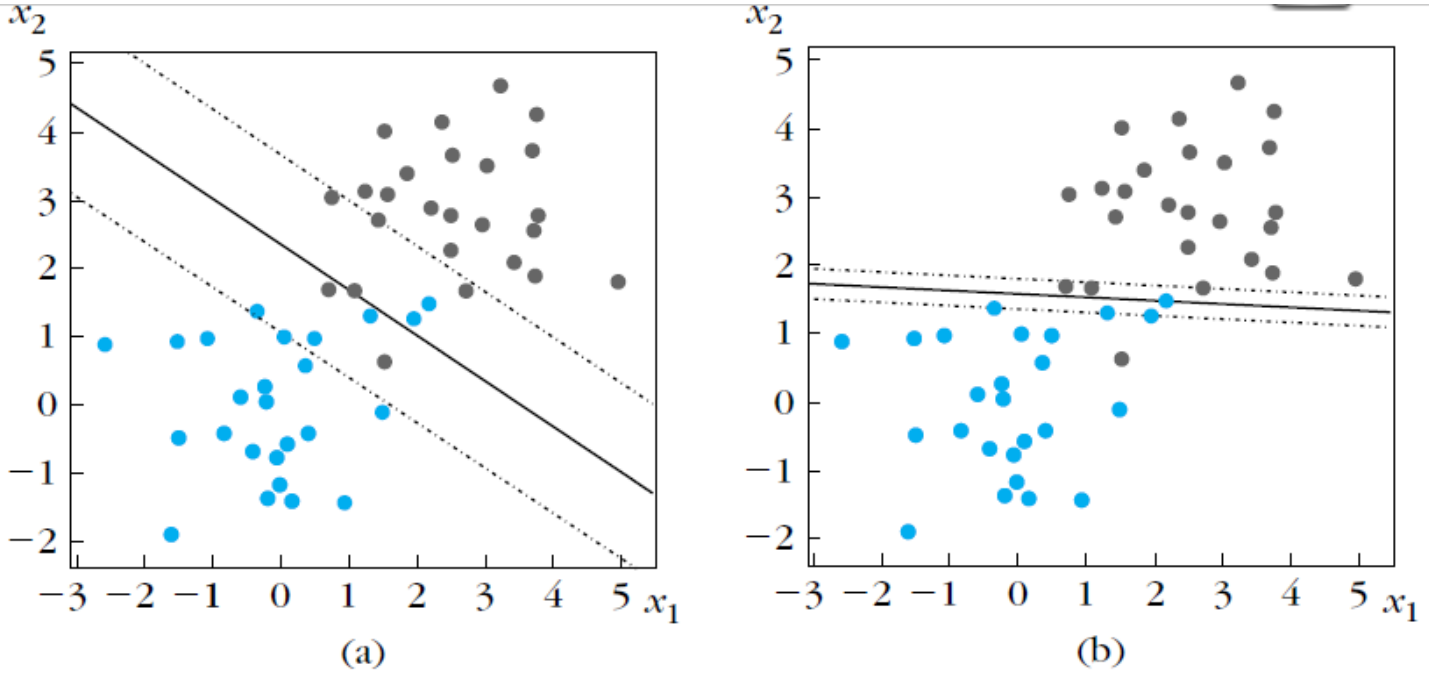
$$\lambda_i \geq 0, \ \mu_i \geq 0, \quad i = 1, 2, \ldots, N$$

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

$$\text{subject to} \quad 0 \leq \lambda_i \leq C, \quad i = 1, 2, \ldots, N$$

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

DIFFERENT C



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.

The width of the margin does not depend entirely on the data distribution, but is heavily affected by the choice of C.

This is the reason SVM classifiers, defined by are also known as *soft margin classifiers*.