Chapter 12. Classification with Support Vector Machines

Chapter 12. Classification with Support Vector Machines

12.1 Separating Hyperplanes

Consider a set of examples $x_n \in \mathbb{R}^D$ along with their corresponding (binary) labels $y_n \in \{+1, -1\}$.

Define the hyperplane that separates the two classes in our binary classification problem as

$$\{x \in \mathbb{R}^D : \langle w, x \rangle + b = 0\}, \tag{1}$$

where $w \in \mathbb{R}^D$ is a weight vector, x is input vector, $b \in \mathbb{R}$ is bias.

The examples with positive labels are on the positive side of the hyperplane

$$\langle w, x_n \rangle + b \geq 0 \text{ when } y_n = +1 \tag{2}$$

and the examples with negative labels are on the negative side

$$\langle w, x_n \rangle + b < 0 \text{ when } y_n = -1 \tag{3}$$

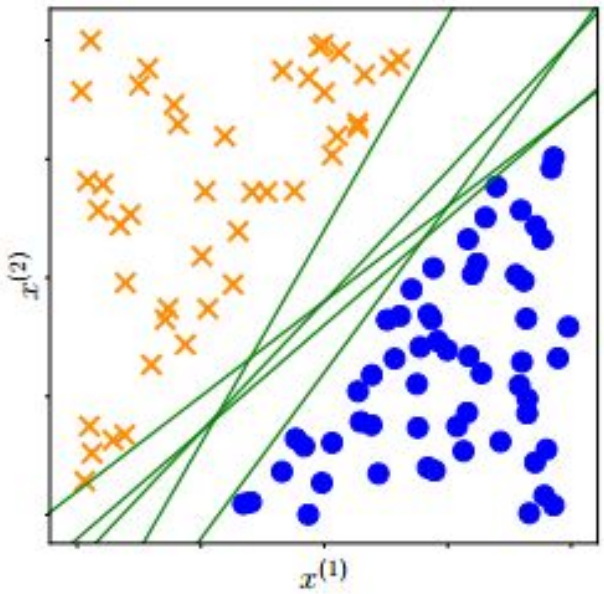
These conditions are equivalent to

$$y_n(\langle w, x_n \rangle + b) \geq 0. \tag{4}$$

12.2 Primal Support Vector Machine

For a dataset $(x_1, y_1), \ldots, (x_N, y_N)$ that is linearly separable, we have infinitely many candidate hyperplanes.

To find a unique solution, one idea is to choose the separating hyperplane that maximizes the margin between the positive and negative examples.



Concept of the margin

The margin is the distance of the separating hyperplane to the closest examples in the dataset, assuming that the dataset is linearly separable.

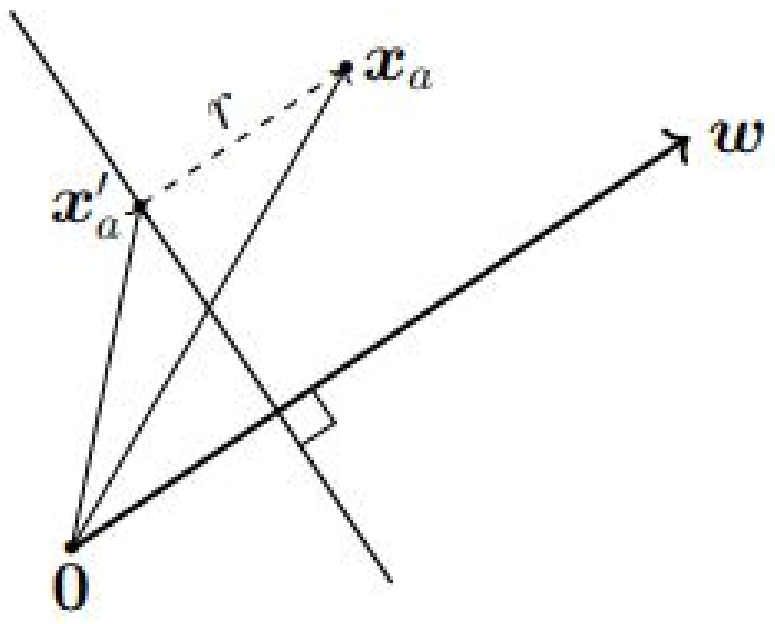
Consider the example x_a to be on the positive side of the hyperplane, i.e.,

$$\langle w, x_a \rangle + b > 0$$

The distance r of x_a from the hyperplane satisfies

$$x_a = x'_a + r\frac{w}{\|w\|}, \qquad (5)$$

where x'_a is the orthogonal projection of x_a onto the hyperplane. We choose x_a to be the point closest to the hyperplane, the distance r is the margin.

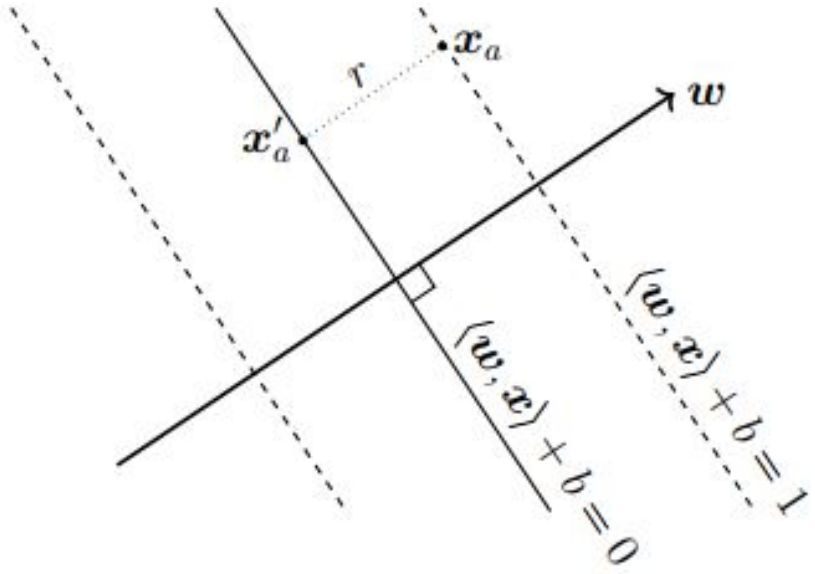


Combining of (2) and (3) into (4), we get

$$y_n(\langle w, x_n \rangle + b) \geq r \tag{6}$$

Assume that the parameter vector w is of unit, i.e., $\|w\| = 1$, we obtain the constrained problem

$$\begin{aligned} &\max_{w,b,r} r \\ &\text{subject to } y_n(\langle w, x_n \rangle + b) \geq r, \|w\| = 1, r > 0 \end{aligned} \tag{7}$$



Traditional Derivation of the Margin

We rescaled the axes, such that the example x_a lies exactly on the margin, i.e., $\langle w, x_a \rangle + b = 1$. We have

$$\langle w, x_a' \rangle + b = 0 \tag{8}$$

Substitute (5) into (8), we obtain

$$\langle w, x_a - r\frac{w}{\|w\|} \rangle + b = 0 \tag{9}$$

$$\Rightarrow \langle w, x_a \rangle + b - r\frac{\langle w, w \rangle}{\|w\|} = 0 \tag{10}$$

$$\Rightarrow r = \frac{1}{\|w\|}. \tag{11}$$

We want the positive and negative examples to be at least 1 away from the hyperplane, which yields the condition

$$y_n \langle w, x_n \rangle + b > 1. \tag{12}$$

Hence, we obtain the constrained optimization problem

$$\max_{w,b} \frac{1}{\|w\|} \tag{13}$$

$$\text{subject to } y_n \langle w, x_n \rangle + b > 1 \quad \text{for all } n = 1, \ldots, N. \tag{14}$$

Instead of maximizing the reciprocal of the norm as in (13), we often minimize the squared norm. This lead us to

$$\min_{w,b} \frac{1}{2} \|w\|^2 \tag{15}$$

$$\text{subject to } y_n \langle w, x_n \rangle + b > 1 \quad \text{for all } n = 1, \ldots, N. \tag{16}$$

Equation (15) is known as the hard margin SVM.

Why We Can Set the Margin to 1

Theorem

Maximizing the margin r, where we consider normalized weights

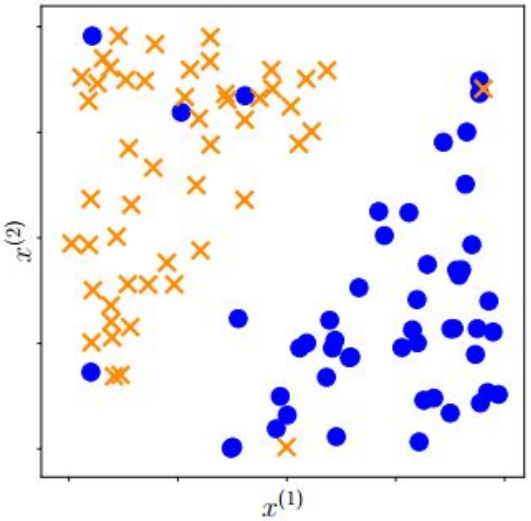
$$\max_{w,b,r} r$$
$$\text{subject to } y_n(\langle w, x_n \rangle + b) \geq r, \quad \|w\| = 1, \quad r > 0, \tag{17}$$

is equivalent to scaling the data, such that the margin is unity

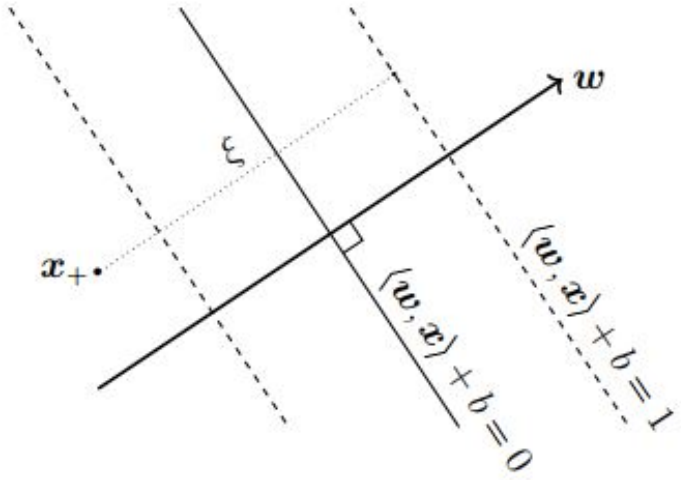
$$\min_{w,b} \|w\|^2$$
$$\text{subject to } y_n(\langle w, x_n \rangle + b) \geq 1. \tag{18}$$

Soft Margin SVM: Geometric View

When data is not linearly separable, we may wish to allow some examples to fall within the margin region, or even to be on the wrong side of the hyperplane.



The model that allows for some classification errors is called the soft margin SVM. For each label pair (x_n, y_n), we introduce a slack variable ξ_n: allows a particular example to be within the margin or even on the wrong side of the hyperplane, subtract ξ_n from the margin.



The soft margin SVM:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{n=1}^{N} \xi_n \tag{19}$$

$$\text{subject to } y_n(\langle w, x_n \rangle + b) \geq 1 - \xi_n \tag{20}$$

$$\xi_n \geq 0, \tag{21}$$

where $C > 0$ is called the regularization parameter, the margin term $\|w\|^2$ is called the regularizer.

Soft Margin SVM: Loss Function View

Consider the error between the output of a predictor $f(x_n) = \langle w, x_n \rangle + b$ and the label y_n. Define the hinge loss

$$l(t) = \max\{0, 1-t\} = \begin{cases} 0 & \text{if } t \geq 1 \\ 1-t & \text{if } t < 1 \end{cases} \quad \text{where } t = yf(x) = y(\langle w, x \rangle + b). \tag{22}$$

To minimize the total loss, while regularizing the objective with l_2-regularization, by using the hinge loss gives us the unconstrained optimization problem:

$$\min_{w,b} \underbrace{\frac{1}{2}\|w\|^2}_{\text{regularizer}} + \underbrace{C \sum_{n=1}^{N} \max\{0, 1 - y_n(\langle w, x_n \rangle + b)\}}_{\text{error term}}. \tag{23}$$

The problem in (23) can be solved with (sub-)gradient descent methods. The problems (23) and (19) are equivalent.

Convex Duality via Lagrange Multipliers Consider the primal soft margin SVM (19): we use $\alpha_n, \gamma_n \geq 0$ as the Lagrange multipliers corresponding to the constraints (20) and (21) respectively. Then the Lagrangian is given by

$$\begin{aligned}\mathcal{L}(w, b, \xi, \alpha, \gamma) =& \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n \\ & - \sum_{n=1}^{N}\alpha_n(y_n(\langle w, x_n\rangle + b) - 1 + \xi_n) - \sum_{n=1}^{N}\gamma_n\xi_n. \end{aligned} \tag{24}$$

We obtain

$$\frac{\partial\mathcal{L}}{\partial w} = w^T - \sum_{n=1}^{N}\alpha_n y_n x_n^T, \quad \frac{\partial\mathcal{L}}{\partial b} = -\sum_{n=1}^{N}\alpha_n y_n, \quad \frac{\partial\mathcal{L}}{\partial\xi_n} = C - \alpha_n - \gamma_n. \tag{25}$$

Find the maximum of Lagrange by setting these partial derivatives to zero:

$$w = \sum_{n=1}^{N}\alpha_n y_n x_n, \sum_{n=1}^{N}\alpha_n y_n = 0, C - \alpha_n - \gamma_n = 0. \tag{26}$$

Substitute the first expression in (26) into (24), we obtain the dual

$$\mathcal{D}(\xi,\alpha,\gamma) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle - \sum_{i=1}^{N} y_i \alpha_i \Big\langle \sum_{j=1}^{N} \alpha_j x_j, x_i \Big\rangle$$

$$+ C\sum_{i=1}^{N}\xi_i - b\sum_{i=1}^{N} y_i\alpha_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\gamma_i\xi_i \tag{27}$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle + \sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}(C-\alpha_i-\gamma_i)\xi_i \tag{28}$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle + \sum_{i=1}^{N}\alpha_i. \tag{29}$$

Moreover, we have that $\alpha_i \le C$. We obtain the dual SVM

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle - \sum_{i=1}^{N}\alpha_i$$

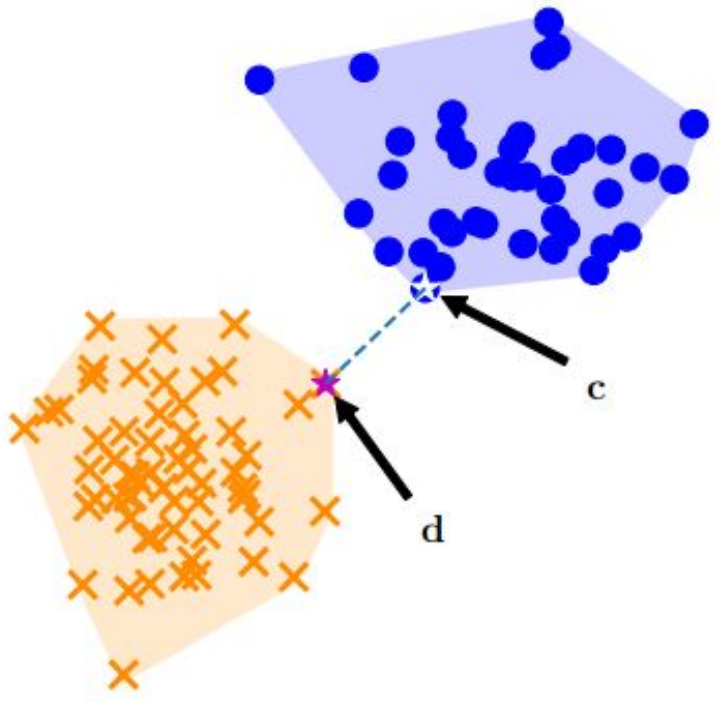
$$\text{subject to } \sum_{i=1}^{N} y_i\alpha_i = 0,\ 0 \le \alpha_i \le C \text{ for all } i = 1,\dots,N. \tag{30}$$

Dual SVM: Convex Hull View

The convex hull of a set of points $x_1, \ldots, x_k$

$$\text{conv}(X) = \left\{ \sum_{i=1}^{k} \alpha_i x_i : \sum_{i=1}^{k} \alpha_i = 1, \ \alpha_i \geq 0 \text{ for all } i = 1, \ldots, k \right\}. \tag{31}$$

Given the training data $(x_1, y_1), \ldots, (x_N, y_N)$. Pick a point c in the convex hull of the set of positive examples that is closest to the negative class distribution. Similarly, pick a point d in the convex hull of the set of negative examples that is closest to the positive class distribution.



Set

$$w := c - d. \tag{32}$$

Then the corresponding optimization problem is

$$\arg\min_{w} \|w\| = \arg\min \frac{1}{2} \|w\|^2 . \tag{33}$$

By the definition of c, d, we have

$$c = \sum_{n:y_n=+1} \alpha_n^+ x_n, \;\; d = \sum_{n:y_n=-1} \alpha_n^- x_n. \tag{34}$$

By substituting (32), (34) into (33), we obtain the objective

$$\min_{\alpha} \frac{1}{2} \left\| \sum_{n:y_n=+1} \alpha_n^+ x_n - \sum_{n:y_n=-1} \alpha_n^- x_n \right\|^2 , \tag{35}$$

where $\alpha = (\alpha^+, \alpha^-)$. Since $\sum_{n:y_n=+1} \alpha^+ = 1$, $\sum_{n:y_n=-1} \alpha^- = 1$, we get the constraint

$$\sum_{n=1}^{N} y_n \alpha_n = 0. \tag{36}$$

The objective function (35), the constraint (36) and the assumption $\alpha \geq 0$ give us a constraint optimization problem.

Consider the loss function view of the SVM (23). This is a convex unconstrained optimization problem, but the hinge loss (22) is not differentiable. We apply a subgradient approach for solving it. The subgradient g of the hinge loss is given by

$$g(t) = \begin{cases} -1 & \text{if } t < 1 \\ [-1, 0] & \text{if } t = 1. \\ 0 & \text{if } t > 1 \end{cases} \tag{37}$$

Using this subgradient, we can apply the optimization methods presented in Section 7.1.

To express the primal SVM in the standard form for quadratic programming, we rearrange the equation (19):

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{subject to } -y_n x_n^T w - y_n b - \xi_n \le -1,\ \xi_n \le 0 \text{ for all } n = 1,\dots,N. \tag{38}$$

Writing the variables w, b, x_n into a single vector, we obtain the matrix form of the soft margin SVM:

$$\min_{w,b,\xi} \frac{1}{2}\begin{bmatrix} w \\ b \\ \xi \end{bmatrix}^T \begin{bmatrix} I_D & 0_{D,N+1} \\ 0_{N+1,D} & 0_{N+1,N+1} \end{bmatrix}\begin{bmatrix} w \\ b \\ \xi \end{bmatrix} + \begin{bmatrix} 0_{D+1,1} & C1_{N,1} \end{bmatrix}^T \begin{bmatrix} w \\ b \\ \xi \end{bmatrix}$$

$$\text{subject to } \begin{bmatrix} -YX & -y & -I_N \\ 0_{N,D+1} & & -I_N \end{bmatrix}\begin{bmatrix} w \\ b \\ \xi \end{bmatrix} \le \begin{bmatrix} -1_{N,1} \\ 0_{N,1} \end{bmatrix}, \tag{39}$$

where $y = \begin{bmatrix} y_1 & \cdots, & y_N \end{bmatrix}^T$, $Y = \text{diag}(y) \in \mathbb{R}^{N\times N}$ and $X = \begin{bmatrix} x_1 & \cdots x_N \end{bmatrix}^T \in \mathbb{R}^{N\times D}$.

We express the dual SVM (30) in standard form. By setting $K = \left[\langle x_i, x_j \rangle\right] \in \mathbb{R}^{N \times N}$, the dual SVM can be written as

$$\min_{\alpha} \frac{1}{2}\alpha^T YKY\alpha - 1_{N,1}^T\alpha$$

$$\text{subject to} \begin{bmatrix} y^T \\ -y^T \\ -I_N \\ I_N \end{bmatrix} \alpha \leq \begin{bmatrix} 0_{N+2,1} \\ C1_{N,1} \end{bmatrix}. \tag{40}$$

Summary

We have studied:

- the idea of the margin and then extend to a classification error;
- two equivalent ways of formalizing the SVM: the geometric view (Section 12.2.4) and the loss function view;
- the dual version of the SVM using Lagrange multipliers;
- the SVM in terms of the convex hulls of the examples of each class;
- briefly describing kernels and how to numerically solve the nonlinear kernel-SVM optimization problem.