Lecture 9: Support Vector Machine

School of Mathematical Sciences, Xiamen University

1. Hyperplanes

- Formally, given a vector $\mathbf{w} \in \mathbb{R}^n$ with $\|\mathbf{w}\|_2 = 1$ and a scalar $s \in \mathbb{R}$, we define the hyperplane $\mathcal{H}(\mathbf{w}, s)$ determined by $\mathbf{w}$ and s as

$$\mathcal{H}(\mathbf{w}, s) = \{\mathbf{x} \in \mathbb{R} \mid \mathbf{w}^\top \mathbf{x} = s\}.$$

- Every vector $\mathbf{x} \in \mathbb{R}^n$ can be expressed as the sum of two orthogonal components:

$$\mathbf{x} = (\mathbf{w}^\top \mathbf{x})\mathbf{w} + (\mathbf{x} - (\mathbf{w}^\top \mathbf{x})\mathbf{w}) := t\mathbf{w} + \mathbf{x}^\perp$$

- If $\mathbf{y}$ is a point on the hyperplane $\mathcal{H}(\mathbf{w}, s)$, i.e.,

$$\mathbf{y} = s\mathbf{w} + \mathbf{y}^\perp, \quad \mathbf{w}^\top \mathbf{y}^\perp = 0,$$

then, we have

$$\|\mathbf{x} - \mathbf{y}\|_2^2 = \|(t - s)\mathbf{w} + (\mathbf{x}^\perp - \mathbf{y}^\perp)\|_2^2 = |t - s|^2 + \|\mathbf{x}^\perp - \mathbf{y}^\perp\|_2^2.$$

- Define the distance of $\mathbf{x}$ from the hyperplane $\mathcal{H}(\mathbf{w}, s)$ to be

$$\begin{aligned}\operatorname{dist}(\mathbf{x}, \mathcal{H}(\mathbf{w}, s)) &= \min\{\|\mathbf{x} - \mathbf{y}\|_2 \mid \mathbf{y} \in \mathcal{H}(\mathbf{w}, s)\} \\ &= |t - s|.\end{aligned}$$

  The minimum is attained for

$$\mathbf{y} = s\mathbf{w} + \mathbf{x}^{\perp},$$

  that is, when $\mathbf{y}$ is the orthogonal projection of $\mathbf{x}$ onto $\mathcal{H}(\mathbf{w}, s)$.
- A direct consequence is that the distance between two parallel hyperplanes $\mathcal{H}(\mathbf{w}, s_2)$ and $\mathcal{H}(\mathbf{w}, s_2)$, $s_1, s_2 \in \mathbb{R}$, is

$$\begin{aligned}&\operatorname{dist}(\mathcal{H}(\mathbf{w}, s_1), \mathcal{H}(\mathbf{w}, s_2)) \\ &= \{\min \|\mathbf{x} - \mathbf{y}\|_2 \mid \mathbf{x} \in \mathcal{H}(\mathbf{w}, s_1), \mathbf{y} \in \mathcal{H}(\mathbf{w}, s_2)\} \\ &= |s_1 - s_2|.\end{aligned}$$

2. Optimal separating hyperplanes

- Consider the set of annotated training data,

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, c^{(1)}), (\mathbf{x}^{(2)}, c^{(2)}), \cdots, (\mathbf{x}^{(p)}, c^{(p)})\},$$

  where $\mathbf{x}^{(j)} \in \mathbb{R}^n$ and $c^{(j)} = \pm 1$.

- Assume that the classes are linear separable, that is, there is a hyperplane $\mathcal{H}(\mathbf{w}, s)$, $\|\mathbf{w}\|_2 = 1$, such that the points in the two classes are on opposite sides of the hyperplane:

$$\mathbf{w}^\top \mathbf{x}^{(j)} - s < 0 \quad \text{if} \quad c^{(j)} = -1,$$
$$\mathbf{w}^\top \mathbf{x}^{(j)} - s > 0 \quad \text{if} \quad c^{(j)} = +1.$$

- The distances of the hyperplane from the two classes:

$$h_- = \min\{|\mathbf{w}^\top \mathbf{x}^{(j)} - s| \mid c^{(j)} = -1\} = \min\{s - \mathbf{w}^\top \mathbf{x}^{(j)} \mid c^{(j)} = -1\},$$
$$h_+ = \min\{|\mathbf{w}^\top \mathbf{x}^{(j)} - s| \mid c^{(j)} = +1\} = \min\{\mathbf{w}^\top \mathbf{x}^{(j)} - s \mid c^{(j)} = +1\}.$$

- Without loss of generality, we can translate the hyperplane along the axis defined by $\mathbf{w}$ until we find a value of s for which the hyperplane is equidistant from both classes, i.e., $h_- = h_+ = h$.
- After having adjusted the position of the hyperplane, we have

$$\mathbf{w}^\top \mathbf{x}^{(j)} - s \leq -h \quad \text{if} \quad c^{(j)} = -1,$$
$$\mathbf{w}^\top \mathbf{x}^{(j)} - s \geq +h \quad \text{if} \quad c^{(j)} = +1,$$

- Define $\mathcal{H}_\pm := \mathcal{H}(\mathbf{w}, s \pm h) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{w}^\top \mathbf{x} = s \pm h\}$. Dividing both sides of the equation

$$\mathbf{w}^\top \mathbf{x} = s \pm h$$

by h, we see that the condition to be satisfied by the vectors in $\mathbb{R}^n$ in the hyperplanes $\mathcal{H}_\pm$ can be reformulated as

$$\frac{1}{h}\mathbf{w}^\top \mathbf{x} - \frac{s}{h} = \pm 1.$$

- After performing the change of variables

$$\mathbf{q} = \frac{1}{h}\mathbf{w}, \quad b = \frac{s}{h}$$

and removing the condition that the normal vector defining a hyperplane must have a unit 2-norm, we can define the two hyperplanes as

$$\mathcal{H}_{\pm} := \mathcal{H}_{\pm}(\mathbf{q}, b) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{q}^\top \mathbf{x} - b = \pm 1\},$$

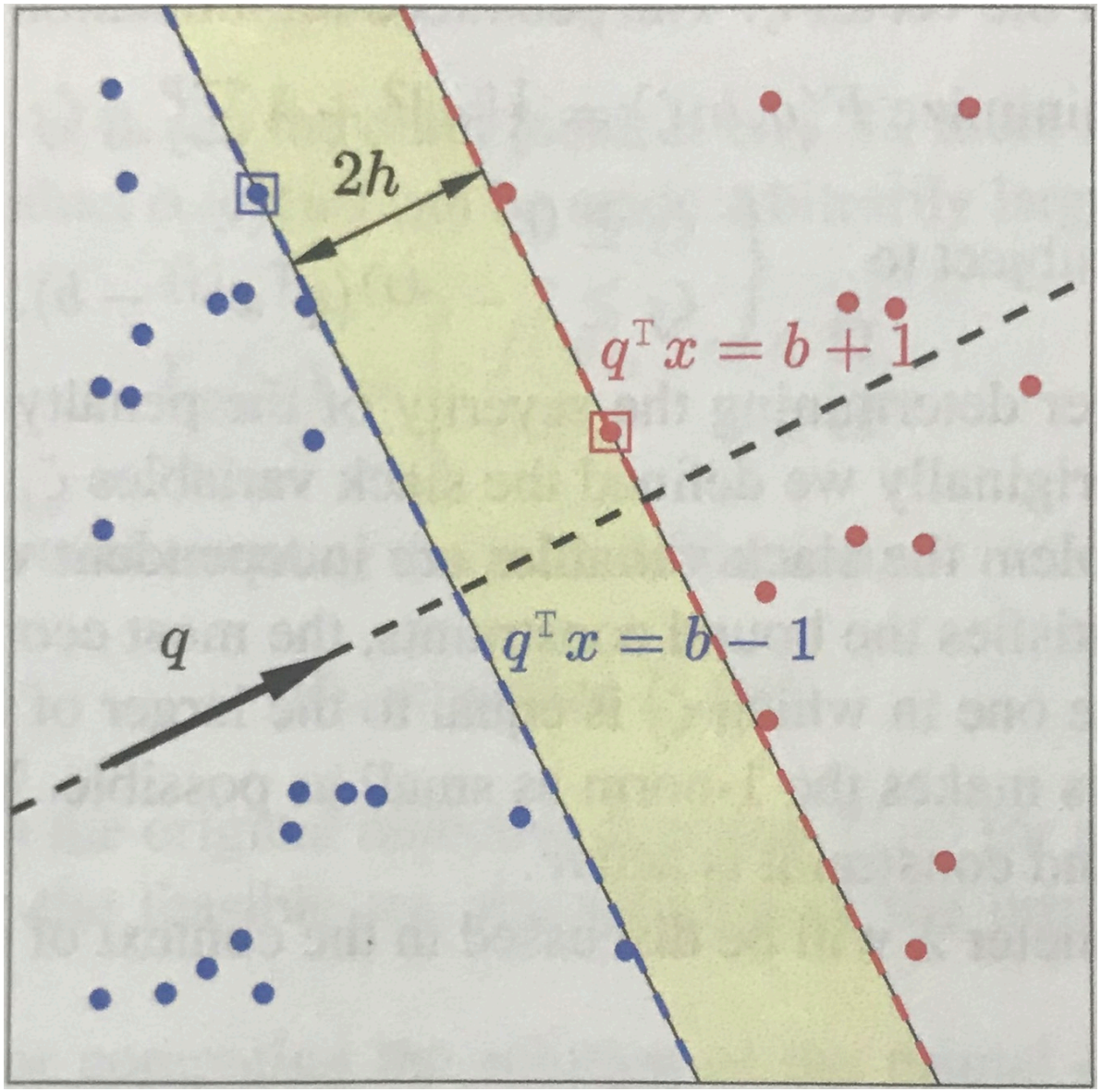
with the only requirement that $\mathbf{q}$ is a nonzero vector.

- The distance between $\mathcal{H}_-$ and $\mathcal{H}_+$ can be expressed as

$$\text{dist}(\mathcal{H}_+, \mathcal{H}_-) = \frac{2}{\|\mathbf{q}\|_2}, \quad \mathbf{q} \neq \mathbf{0}.$$

- The parallel hyperplanes $\mathcal{H}_-$ and $\mathcal{H}_+$ define a region of width $2h$ separating the two classes.

- Support vectors and the separating margin.



- We have

$$\mathbf{q}^\top \mathbf{x}^{(j)} - b \leq -1 \quad \text{if} \quad c^{(j)} = -1,$$
$$\mathbf{q}^\top \mathbf{x}^{(j)} - b \geq +1 \quad \text{if} \quad c^{(j)} = +1,$$

  or, more concisely,

$$c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) \geq 1, \quad 1 \leq j \leq p.$$

- The goal of the support vector machine algorithm is to find a vector $\mathbf{q}$ defining the pair of hyperplanes $\mathcal{H}_\pm$ that separate the two classes for which the separating margin is as wide as possible. This goal can be recast in the form of a constrained optimization problem:

$$\min_{\mathbf{q},b} \ \frac{1}{2}\|\mathbf{q}\|_2^2 \quad \text{subject to} \quad c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) \geq 1.$$

- Since, in general, there is no guarantee that the linear separability condition is satisfied, the hyperplane selection criterion needs to be modified to take into consideration the possibility that the linear separability condition may not hold.
- Let $\mathcal{H}_{\pm}(\mathbf{q}, b)$ be two parallel hyperplanes as above. Partition the data points into two groups by assigning $j \in \mathcal{J}$ if $(\mathbf{x}^{(j)}, c^{(j)})$ satisfies

$$c^{(j)}(\mathbf{q}^{\top}\mathbf{x}^{(j)} - b) \geq 1,$$

and to $j \in \mathcal{J}^c$ otherwise, where $\mathcal{J}^c$ is the set of indices between 1 and p that are not in $\mathcal{J}$. Thus, the indices in $\mathcal{J}^c$ correspond to data points that are not on the correct side of the margin.

- For each index j, define the slack variable ζ_j,

$$\zeta_j = \begin{cases} 1 - c^{(j)}(\mathbf{q}^{\top}\mathbf{x}^{(j)} - b) > 0 & \text{for} \quad j \in \mathcal{J}^c, \\ 0 & \text{for} \quad j \in \mathcal{J}, \end{cases}$$

or, compactly, $\zeta_j = \max\{1 - c^{(j)}(\mathbf{q}^{\top}\mathbf{x}^{(j)} - b), 0\},\ 1 \leq j \leq p.$

- With these auxiliary variables, searching for $\mathbf{q}$ and b so as to minimize the number of data points on the wrong side of the region defined by the hyperplanes is tantamount to looking for a vector $\boldsymbol{\zeta}$ as sparse as possible, that is, with as few nonzero entries as possible.
- The sparsity is promoted by adding to the objective function a penalty term for the growth in the 1-norm of the vector $\boldsymbol{\zeta}$. We have

$$\text{minimize } f(\mathbf{q}, b, \boldsymbol{\zeta}) = \frac{1}{2}\|\mathbf{q}\|_2^2 + \lambda \sum_{j=1}^{p} \zeta_j$$

$$\text{subject to } \begin{cases} \zeta_j \geq 0, \\ \zeta_j \geq 1 - c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b), \end{cases}$$

where $\lambda > 0$ is a parameter determining the severity of the penalty for increasing the 1-norm of $\boldsymbol{\zeta}$.

3. The primal-dual approach

- Given the objective function $f : \mathbb{R}^d \to \mathbb{R}$ and the vector valued function defining the constraints $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^k$, we want to

$$\text{minimize } f(\mathbf{w}) \quad \text{subject to } \; g_j(\mathbf{w}) \leq 0, \; 1 \leq j \leq k.$$

- Let $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ be the Lagrange function, defined as

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = f(\mathbf{w}) + \boldsymbol{\alpha}^\top \mathbf{g}(\mathbf{w}),$$

  where the vector $\boldsymbol{\alpha} \in \mathbb{R}^k_+$ is the vector of Lagrange multipliers.
- Let $\Omega \subset \mathbb{R}^d$ denote the feasible set of points in $\mathbb{R}^d$

$$\Omega = \{\mathbf{w} \in \mathbb{R}^d \mid g_j(\mathbf{w}) \leq 0, \; 1 \leq j \leq k\}.$$

- Define the primal function $\mathcal{L}_p(\mathbf{w})$ as

$$\mathcal{L}_p(\mathbf{w}) = \max_{\boldsymbol{\alpha}} \{\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \mid \alpha_j \geq 0, \; 1 \leq j \leq k\}.$$

- Observe that

$$\alpha_j g_j(\mathbf{w}) \leq 0, \quad \mathbf{w} \in \Omega,$$

with equality holding if $\alpha_j = 0$. If $\mathbf{w} \notin \Omega$, there is at least one index j such that $g_j(\mathbf{w}) > 0$, and the product $\alpha_j g_j(\mathbf{w})$ can be made arbitrarily large, and therefore

$$\mathcal{L}_p(\mathbf{w}) = \begin{cases} f(\mathbf{w}), & \mathbf{w} \in \Omega, \\ \infty, & \mathbf{w} \notin \Omega. \end{cases}$$

- We define the primal problem in terms of the primal function,

$$(\mathrm{P}): \quad \text{minimize} \quad \mathcal{L}_p(\mathbf{w}).$$

- Since $\mathcal{L}_p(\mathbf{w})$ coincides with the original objective function $f(\mathbf{w})$ for all vectors $\mathbf{w}$ in the feasible set and it diverges outside the feasible set, the solution of the primal problem is the desired constrained minimizer.

- Introduce the auxiliary dual function $\mathcal{L}_d(\boldsymbol{\alpha})$ in $\mathbb{R}^k$:

$$\mathcal{L}_d(\boldsymbol{\alpha}) = \min\{\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \mid \mathbf{w} \in \mathbb{R}^d\},$$

and define the corresponding dual problem

$$(\mathrm{D}): \quad \text{maximize} \quad \mathcal{L}_d(\boldsymbol{\alpha}) \quad \text{subject to} \quad \boldsymbol{\alpha} \geq \mathbf{0}.$$

- The primal and dual problems are not identical by default, because interchanging the order of minimization and maximization may change the problem.
- Since for every pair $(\mathbf{w}, \boldsymbol{\alpha}) \in \mathbb{R}^d \times \mathbb{R}^k_+$,

$$\mathcal{L}_d(\boldsymbol{\alpha}) \leq \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \leq \mathcal{L}_p(\mathbf{w}),$$

we have that

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \mathcal{L}_d(\boldsymbol{\alpha}) \leq \min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}_p(\mathbf{w}).$$

Theorem 1

Assume that the functions f and g_j are convex, $1 \leq j \leq k$, and that there exists at least one point $\mathbf{w} \in \mathbb{R}^d$ such that $g_j(\mathbf{w}) < 0$ for all j. Then there exists a vector $\mathbf{w}^\star \in \Omega$ that solves the primal problem, and coefficients $\alpha_j^\star \geq 0$, $1 \leq j \leq k$, that solve the dual problem, and

$$\max_{\alpha_j \geq 0} \mathcal{L}_d(\boldsymbol{\alpha}) = \mathcal{L}(\mathbf{w}^\star, \boldsymbol{\alpha}^\star) = \min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}_p(\mathbf{w}).$$

Moreover, the vectors $\mathbf{w}^\star$ and $\boldsymbol{\alpha}^\star$ satisfy the Karush–Kuhn–Tucker (KKT) conditions, i.e.,

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^\star, \boldsymbol{\alpha}^\star) = \mathbf{0}, \quad \mathbf{g}(\mathbf{w}^\star) \leq \mathbf{0}, \quad \boldsymbol{\alpha}^\star \geq \mathbf{0},$$

and (the complementarity condition)

$$(\boldsymbol{\alpha}^\star)^\top \mathbf{g}(\mathbf{w}^\star) = 0.$$

4. The SVM optimization problem

- For

$$(\mathbf{q}, b, \boldsymbol{\zeta}) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n,$$

consider the objective function

$$f(\mathbf{q}, b, \boldsymbol{\zeta}) = \frac{1}{2}\|\mathbf{q}\|_2^2 + \lambda \sum_{j=1}^{p} \zeta_j.$$

- Introduce the constraint functions g_j and g_{p+j}, $1 \leq j \leq p$,

$$g_j(\mathbf{q}, b, \boldsymbol{\zeta}) = -\zeta_j,$$

$$g_{p+j}(\mathbf{q}, b, \boldsymbol{\zeta}) = 1 - \zeta_j - c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b).$$

- The SVM optimization problem is

$$\text{minimize } f(\mathbf{q}, b, \boldsymbol{\zeta}) \quad \text{subject to } \; g_j(\mathbf{q}, b, \boldsymbol{\zeta}) \leq 0, \; 1 \leq j \leq 2p.$$

- The functions f and g_j are convex. The Lagrange function is

$$\begin{aligned}\mathcal{L}(\mathbf{q}, b, \boldsymbol{\zeta}, \boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{q}\|_2^2 + \lambda \sum_{j=1}^{p} \zeta_j - \sum_{j=1}^{p} \alpha_j \zeta_j \\ &\quad - \sum_{j=1}^{p} \alpha_{p+j}(\zeta_j + c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) - 1) \\ &= \frac{1}{2}\|\mathbf{q}\|_2^2 - \mathbf{q}^\top \left(\sum_{j=1}^{p} \alpha_{p+j} c^{(j)} \mathbf{x}^{(j)} \right) + b \sum_{j=1}^{p} \alpha_{p+j} c^{(j)} \\ &\quad - \sum_{j=1}^{p} \zeta_j (\alpha_j + \alpha_{p+j} - \lambda) + \sum_{j=1}^{p} \alpha_{p+j},\end{aligned}$$

where the coefficients

$$\alpha_j \geq 0, \quad 1 \leq j \leq 2p.$$

- The first-order optimality condition for the minimizer requires the gradient of $\mathcal{L}$ with respect to $\mathbf{q}, b$, and $\boldsymbol{\zeta}$ to vanish, yielding the system of equations

$$\nabla_{\mathbf{q}}\mathcal{L} = \mathbf{q} - \sum_{j=1}^{p} \alpha_{p+j} c^{(j)} \mathbf{x}^{(j)} = \mathbf{0},$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{j=1}^{p} \alpha_{p+j} c^{(j)} = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_j} = \lambda - \alpha_j - \alpha_{p+j} = 0, \quad 1 \leq j \leq p.$$

- Therefore, the Lagrange function at the maximizer of the dual problem simplifies to

$$\mathcal{L} = -\frac{1}{2} \left\| \sum_{j=1}^{p} \alpha_{p+j} c^{(j)} \mathbf{x}^{(j)} \right\|_2^2 + \sum_{j=1}^{p} \alpha_{p+j}.$$

- After performing the change of variables $\mathbf{y}^{(j)} = c^{(j)}\mathbf{x}^{(j)}$ and $\beta_j = \alpha_{p+j}$, we have

$$\mathcal{L} = -\frac{1}{2}\left\|\sum_{j=1}^{p}\beta_j\mathbf{y}^{(j)}\right\|_2^2 + \sum_{j=1}^{p}\beta_j = -\frac{1}{2}\boldsymbol{\beta}^\top\mathbf{G}\boldsymbol{\beta} + \mathbf{1}^\top\boldsymbol{\beta},$$

where $\mathbf{G} \in \mathbb{R}^{p\times p}$ is the matrix with entries

$$\mathbf{G}_{jk} = (\mathbf{y}^{(j)})^\top\mathbf{y}^{(k)}.$$

- In other words, to solve the SVM optimization problem it suffices to find a vector $\boldsymbol{\beta}$ that solves the following reduced dual problem:

$$\text{maximize } \left\{-\frac{1}{2}\boldsymbol{\beta}^\top\mathbf{G}\boldsymbol{\beta} + \mathbf{1}^\top\boldsymbol{\beta}\right\} \text{ subject to } \begin{cases} 0 \le \beta_j \le \lambda, \\ \displaystyle\sum_{j=1}^{p}\beta_j c^{(j)} = 0. \end{cases}$$

- By the complementarity condition, we have

$$\sum_{i=1}^{p} \alpha_j \zeta_j + \sum_{j=1}^{p} \beta_j (c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) + \zeta_j - 1) = 0.$$

If $\alpha_j > 0$, then $\zeta_j = 0$, and if $\beta_j > 0$, then

$$c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) + \zeta_j - 1 = 0.$$

- From $\alpha_j + \beta_j = \lambda$, we have that

$$0 < \beta_j < \lambda \Rightarrow \begin{cases} \zeta_j = 0, \\ c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) + \zeta_j - 1 = 0. \end{cases}$$

Therefore, we conclude that

$$0 < \beta_j < \lambda \Rightarrow c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) = 1,$$

which implies that $\mathbf{x}^{(j)}$ must belong to one of $\mathcal{H}_{\pm}(\mathbf{q}, b)$.

- If $\beta_j = 0$, then $\alpha_j = \lambda > 0$, and thus $\zeta_j = 0$. It follows from

$$1 - \zeta_j - c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) \leq 0$$

that

$$c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) \geq 1.$$

Similarly, $\beta_j = \lambda$ implies $\alpha_j = 0$ and

$$c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) = 1 - \zeta_j, \quad \zeta_j \geq 0.$$

- In summary,

$$\begin{aligned} \beta_j = 0 &\Rightarrow \zeta_j = 0,\ c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) \geq 1, \\ 0 < \beta_j < \lambda &\Rightarrow c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) = 1, \\ \beta_j = \lambda &\Rightarrow c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) = 1 - \zeta_j \leq 1. \end{aligned}$$

- $c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) > 1$ means that $\mathbf{x}^{(j)}$ is on the correct side of the separating hyperplanes, $c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) = 1$ means that $\mathbf{x}^{(j)}$ is a support vector, and $c^{(j)}(\mathbf{q}^\top \mathbf{x}^{(j)} - b) < 1$ means that $\mathbf{x}^{(j)}$ is on the wrong side.
- If β_j satisfies $0 < \beta_j < \lambda$, we have

$$b = \mathbf{q}^\top \mathbf{x}^{(j)} - c^{(j)}.$$

To make the determination of b more robust, one may compute the average of the estimates of b corresponding to all support vectors:

$$b = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{q}^\top \mathbf{x}^{(j_\ell)} - c^{(j_\ell)}),$$

where $j_1, \cdots, j_m$ are the indices of the support vectors.

Basic support vector machine (SVM) algorithm

1. Given a data set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(p)}\}$ with binary annotation, $c^{(j)} = \pm 1$, and $\lambda > 0$.
2. Set $\mathbf{y}^{(j)} = c^{(j)}\mathbf{x}^{(j)}$ and compute $\mathbf{G}_{jk} = (\mathbf{y}^{(j)})^\top \mathbf{y}^{(k)}$.
3. Find the vector $\boldsymbol{\beta}$ that solves the reduced dual problem.
4. Determine the vector $\mathbf{q}$ according to the formula
$$\mathbf{q} = \sum_{j=1}^{p} \beta_j \mathbf{y}^{(j)}.$$
5. Identify the support vectors with the indices $j_1, \cdots, j_m$ for which $0 < \beta_j < \lambda$. Compute b using
$$b = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{q}^\top \mathbf{x}^{(j_\ell)} - c^{(j_\ell)}).$$
6. An unlabeled vector $\mathbf{x} \in \mathbb{R}^n$ is assigned class label $c = \pm 1$ where $c = \text{sign}(\mathbf{q}^\top \mathbf{x} - b)$.

5. Solving the reduced dual problem

- We update a pair of coordinates in each step. Without loss of generality, let the coordinates be β_1 and β_2.
- Denote by β_j^c the current values of β_j and write the updated values of β_1 and β_2 as

$$\beta_1^+ = \beta_1^c + t, \quad \beta_2^+ = \beta_2^c + s.$$

- To satisfy the constraint, we require

$$c^{(1)}(\beta_1^c + t) + c^{(2)}(\beta_2^c + s) + \sum_{j=3}^{p} c^{(j)}\beta_j^c = 0,$$

which implies that

$$c^{(1)}t + c^{(2)}s = 0,$$

or

$$s = -\sigma t \quad \text{with} \quad \sigma = c^{(1)}/c^{(2)}.$$

- Let $\mathbf{v} = \begin{bmatrix} 1 & -\sigma & 0 & \cdots & 0 \end{bmatrix}^\top$. The objective function to be maximized can be expressed as a function of t of the form

$$\begin{aligned} f(t) &= -\frac{1}{2}(t\mathbf{v} + \boldsymbol{\beta}^c)^\top \mathbf{G}(t\mathbf{v} + \boldsymbol{\beta}^c) + \mathbf{1}^\top(t\mathbf{v} + \boldsymbol{\beta}^c) \\ &= -\frac{1}{2}t^2(\mathbf{v}^\top \mathbf{G}\mathbf{v}) + t\mathbf{v}^\top(\mathbf{1} - \mathbf{G}\boldsymbol{\beta}^c) + f(0). \end{aligned}$$

- The bound constraints require that

$$0 \leq \beta_1^c + t \leq \lambda, \quad 0 \leq \beta_2^c - \sigma t \leq \lambda.$$

This implies that t must satisfy $t_{\min} \leq t \leq t_{\max}$, with

$$t_{\min} = \begin{cases} \max\{-\beta_1^c, -\beta_2^c\} & \text{if} \quad \sigma = -1, \\ \max\{-\beta_1^c, \beta_2^c - \lambda\} & \text{if} \quad \sigma = 1, \end{cases}$$

$$t_{\max} = \begin{cases} \min\{\lambda - \beta_1^c, \lambda - \beta_2^c\} & \text{if} \quad \sigma = -1, \\ \min\{\lambda - \beta_1^c, \beta_2^c\} & \text{if} \quad \sigma = 1. \end{cases}$$

- If $\mathbf{v}^\top \mathbf{G}\mathbf{v} > 0$, then let

$$t^\star = \frac{\mathbf{v}^\top(\mathbf{1} - \mathbf{G}\boldsymbol{\beta}^c)}{\mathbf{v}^\top \mathbf{G}\mathbf{v}}.$$

The maximum of $f(t)$ is attained at

$$t^+ = \begin{cases} t_{\min} & \text{if} \quad t^\star < t_{\min}, \\ t^\star & \text{if} \quad t_{\min} \leq t^\star \leq t_{\max} \\ t_{\max} & \text{if} \quad t_{\max} < t^\star. \end{cases}$$

If $\mathbf{v}^\top \mathbf{G}\mathbf{v} = 0$, then $f(t)$ is linear, and the maximum is attained at one of the end points of the interval, that is

$$t^+ = \begin{cases} t_{\min} & \text{if} \quad \mathbf{v}^\top(\mathbf{1} - \mathbf{G}\boldsymbol{\beta}^c) < 0, \\ t_{\max} & \text{if} \quad \mathbf{v}^\top(\mathbf{1} - \mathbf{G}\boldsymbol{\beta}^c) > 0. \end{cases}$$

If $\mathbf{v}^\top \mathbf{G}\mathbf{v} = 0$ and $\mathbf{v}^\top(\mathbf{1} - \mathbf{G}\boldsymbol{\beta}^c) = 0$, then $f(t)$ is constant, and there is no need to update t.

Sequential minimal optimization (SMO) algorithm

1. Given: maximum number of iteration `maxit`, tolerance τ
   $c^{(j)} = \pm 1$, and $\lambda > 0$.
2. Initialize: $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, $\delta = \infty$, and $\ell = 0$.
3. **while** $\ell <$ `maxit` and $\delta > \tau$
   Set $\boldsymbol{\beta}^c = \boldsymbol{\beta}^{(\ell)}$.
   for $j = 1 : p$
   Select randomly $k \neq j$, $1 \leq k \leq p$.
   Find the optimal values for β_j^+ and β_k^+.
   Update $\beta_j^c = \beta_j^+$ and $\beta_k^c = \beta_k^+$.
   end for
   Set $\boldsymbol{\beta}^{(\ell+1)} = \boldsymbol{\beta}^c$.
   Compute $\delta = \dfrac{\|\boldsymbol{\beta}^{(\ell+1)} - \boldsymbol{\beta}^{(\ell)}\|_2}{\|\boldsymbol{\beta}^{(\ell)}\|_2}$, and set $\ell = \ell + 1$.
   end while

6. Generalization using kernel functions

- Definition: The function $\mathcal{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a symmetric positive definite kernel function if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \mathcal{K}(\mathbf{y}, \mathbf{x}),$$

and for all $m > 1$ and $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)} \in \mathbb{R}^n$, the matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ with entries

$$\mathbf{K}_{ij} = \mathcal{K}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), \quad 1 \leq i, j \leq m,$$

is positive semidefinite.

- Examples: $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$; $\mathcal{K}(\mathbf{x}, \mathbf{y}) = (\lambda + \mathbf{x}^\top \mathbf{y})^m$, $\lambda > 0$;

Gaussian kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{y}\|_2^2\right)$, $\sigma > 0$;

Laplace kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{\lambda}\|\mathbf{x} - \mathbf{y}\|_2\right)$, $\lambda > 0$.

- Mercer's Theorem:

  Let $\mathcal{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a symmetric positive definite kernel function. Then there are feature functions $\varphi_\ell : \mathbb{R}^n \to \mathbb{R}$ such that $\mathcal{K}$ can be expressed as

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \sum_{\ell=1}^{\infty} \varphi_\ell(\mathbf{x}) \varphi_\ell(\mathbf{y}).$$

- Let $\boldsymbol{\varphi}(\cdot) = \begin{bmatrix} \varphi_1(\cdot) & \cdots & \varphi_\ell(\cdot) & \cdots \end{bmatrix}^\top$.

  (1) Extend $\mathbf{x}^{(j)}$ to a higher-dimensional space via $\mathbf{z}^{(j)} = \boldsymbol{\varphi}(\mathbf{x}^{(j)})$

  (2) Apply the SVM algorithm to the data set $\{\mathbf{z}^{(j)}, c^{(j)}\}$ to find a separating hyperplane $\mathcal{H}(\mathbf{q}, b)$

  (3) Classify a vector $\mathbf{x} \in \mathbb{R}^n$ by checking on which side of $\mathcal{H}(\mathbf{q}, b)$ the vector $\mathbf{z} = \boldsymbol{\varphi}(\mathbf{x})$ lies.

- The idea of kernel SVM: We do not explicitly form $\boldsymbol{\varphi}$, and use the corresponding symmetric positive definite kernel function $\mathcal{K}(\mathbf{x}, \mathbf{y})$.

Kernel support vector machine (kSVM) algorithm

1. Given a data set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(p)}\}$ with binary annotation, $c^{(j)} = \pm 1$, an SPD kernel function $\mathcal{K}(\mathbf{x}, \mathbf{y})$ and $\lambda > 0$.
2. Compute the matrix $\mathbf{G}$ with entries $\mathbf{G}_{jk} = c^{(j)} c^{(k)} \mathcal{K}(\mathbf{x}^{(j)}, \mathbf{x}^{(k)})$.
3. Find the vector $\boldsymbol{\beta} \in \mathbb{R}^p$ that solves the reduced dual problem.
4. Define the vector $\mathbf{q}$ implicitly via the inner product

$$\mathbf{q}^\top \boldsymbol{\varphi}(\mathbf{x}) = \left(\sum_{j=1}^{p} \beta_j c^{(j)} \boldsymbol{\varphi}(\mathbf{x}^{(j)}) \right)^\top \boldsymbol{\varphi}(\mathbf{x}) = \sum_{j=1}^{p} \beta_j c^{(j)} \mathcal{K}(\mathbf{x}^{(j)}, \mathbf{x}).$$

5. Identify the support vectors with the indices $j_1, \cdots, j_m$ for which $0 < \beta_j < \lambda$. Compute b using

$$b = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{q}^\top \boldsymbol{\varphi}(\mathbf{x}^{(j_\ell)}) - c^{(j_\ell)}).$$

6. An unlabeled vector $\mathbf{x} \in \mathbb{R}^n$ is assigned class label $c = \pm 1$ where $c = \text{sign}(\mathbf{q}^\top \boldsymbol{\varphi}(\mathbf{x}) - b)$.