Support Vector machines

Arthur Gretton

November 30, 2018

1 Outline

- Review of convex optimization
- Support vector classification. C-SV and ν-SV machines

2 Review of convex optimization

This review covers the material from [1, Sections 5.1-5.5]. We begin with definitions of convex functions and convex sets.

2.1 Convex sets, convex functions

Definition 1 (Convex set)**.** A set C is convex if for all $x_1, x_2 \in C$ and any $0 \leq \theta \leq 1$ we have $\theta x_1 + (1-\theta)x_2 \in C$, i.e. every point on the line between x_1 and x_2 lies in C. See Figure 2.1.

In other words, every point in the set can be seen from any other point in the set, along a straight line that never leaves the set.

We next introduce the notion of a convex function.

Definition 2 (Convex function)**.** A function f is convex if its domain $\mathrm{dom} f$ is a convex set and if $\forall x, y \in \mathrm{dom} f$, and any $0 \leq \theta \leq 1$,

$$f\left(\theta x + (1-\theta)y\right) \leq \theta f(x) + (1-\theta)f(y).$$

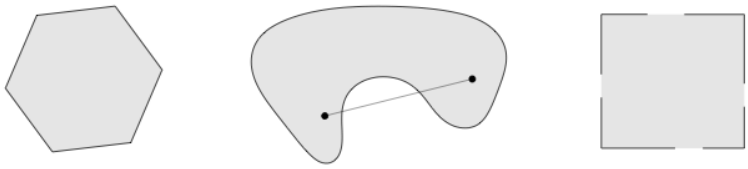

Figure 2.1: Examples of convex and non-convex sets (taken from [1, Fig. 2.2]). The first set is convex, the last two are not.

Figure 2.2: Convex function (taken from [1, Fig. 3.1])

The function is strictly convex if the inequality is strict for $x \neq y$. See Figure 2.2.

2.2 The Lagrangian

We now consider an optimization problem on $x \in \mathbb{R}^n$,

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, m \\ & h_i(x) = 0 & i = 1, \ldots p. \end{array} \tag{2.1}$$

We define by p^* the optimal value of (2.1), and by $\mathcal{D} := \bigcap_{i=0}^m \text{dom} f_i \cap \bigcap_{i=1}^p \text{dom} h_i$, where we require the domain[1] $\mathcal{D}$ to be nonempty.

The **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with problem (2.1) is written

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x),$$

and has domain $\text{dom} L := \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$. The vectors λ and ν are called **lagrange multipliers** or **dual variables**. The **Lagrange dual function** (or just "dual function") is written

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu).$$

If this is unbounded below, then the dual is $-\infty$. The domain of g, $\text{dom}(g)$, is the set of values (λ, μ) such that $g > -\infty$. The dual function is a pointwise infimum of affine[2] functions of (λ, ν), hence it is concave in (λ, ν) [1, p. 83].

When[3] $\lambda \succeq 0$, then for all ν we have

$$g(\lambda, \nu) \leq p^*. \tag{2.2}$$

[1] The domain is the set on which a function is well defined. Eg the domain of $\log x$ is $\mathbb{R}^{++}$, the strictly positive real numbers [1, p. 639].

[2] A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is affine if it takes the form $f(x) = Ax + b$.

[3] The notation $a \succeq b$ for vectors a, b means that $a_i \geq b_i$ for all i.

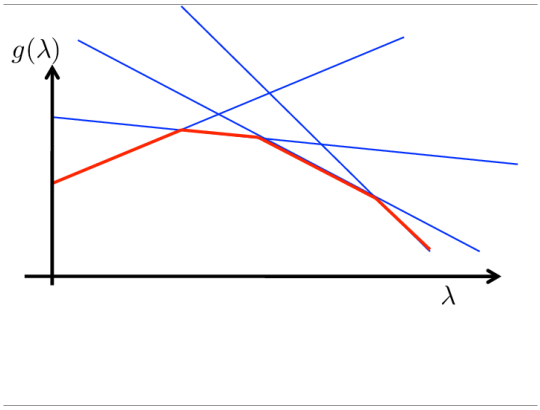

Figure 2.3: Example: Lagrangian with one inequality constraint, $L(x, \lambda) = f_0(x) + \lambda f_1(x)$, where x here can take one of four values for ease of illustration. The infimum of the resulting set of four affine functions is concave in λ.

See Figure (2.4) for an illustration on a toy problem with a single inequality constraint. A **dual feasible** pair (λ, ν) is a pair for which $\lambda \succeq 0$ and $(\lambda, \nu) \in \mathrm{dom}(g)$.

Proof. (of eq. (2.2)) Assume $\tilde{x}$ is feasible for the optimization, i.e. $f_i(\tilde{x}) \leq 0$, $h_i(\tilde{x}) = 0$, $\tilde{x} \in \mathcal{D}$, $\lambda \succeq 0$. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \leq 0$$

and so

$$\begin{aligned} g(\lambda, \nu) &:= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \\ &\leq f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \\ &\leq f_0(\tilde{x}). \end{aligned}$$

This holds for every feasible $\tilde{x}$, hence (2.2) holds. □

We now give a lower bound interpretation. Idealy we would write the problem (2.1) as the unconstrained problem

$$\text{minimize } f_0(x) + \sum_{i=1}^{m} I_-\left(f_i(x)\right) + \sum_{i=1}^{p} I_0\left(h_i(x)\right),$$

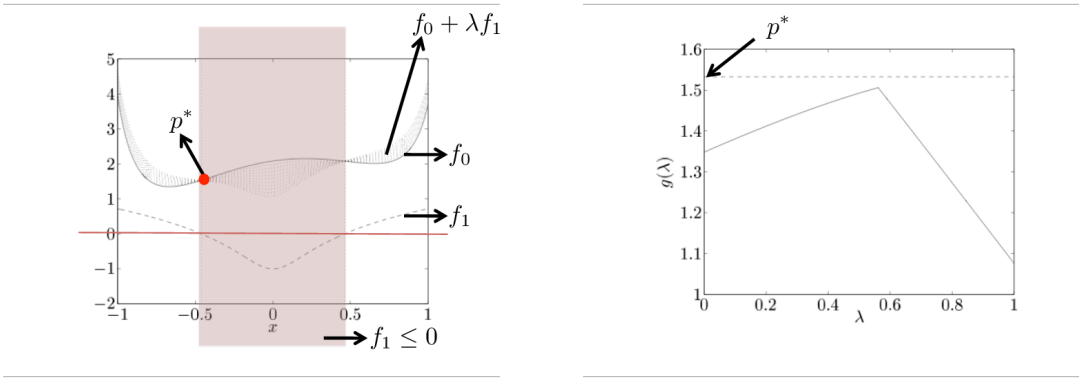

Figure 2.4: Illustration of the dual function for a simple problem with one inequality constraint (from [1, Figs. 5.1 and 5.2]). In the right hand plot, the dashed line corresponds to the optimum p^* of the original problem, and the solid line corresponds to the dual as a function of λ. Note that the dual as a function of λ is concave.

where

$$I_-(u) = \begin{cases} 0 & u \leq 0 \\ \infty & u > 0 \end{cases}$$

and $I_0(u)$ is the indicator of 0. This would then give an infinite penalization when a constraint is violated. Instead of these sharp indicator functions (which are hard to optimize), we replace the constraints with a set of soft linear constraints, as shown in Figure 2.5. It is now clear why λ must be positive for the inequality constraint: a negative λ would not yield a lower bound. Note also that as well as being penalized for $f_i > 0$, the linear lower bounds reward us for achieving $f_i < 0$.

2.3 The dual problem

The dual problem attempts to find the best lower bound $g(\lambda, \nu)$ on the optimal solution p^* of (2.1). This results in the **Lagrange dual problem**

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{2.3}$$

We use **dual feasible** to describe (λ, ν) with $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$. The solutions to the dual problem are written (λ^*, ν^*), and are called **dual optimal**. Note that (2.3) is a convex optimization problem, since the function being maximized is concave and the constraint set is convex. We denote by d^* the optimal value of the dual problem. The property of **weak duality** always holds:

$$d^* \leq p^*.$$

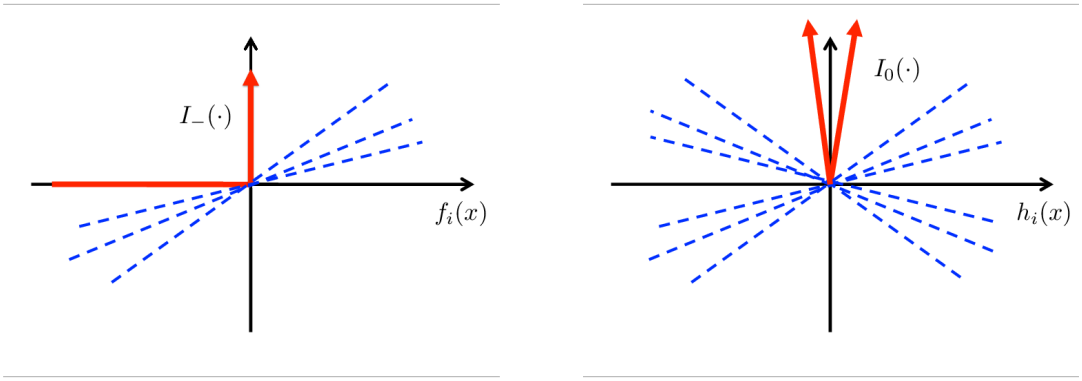

Figure 2.5: Linear lower bounds on indicator functions. Blue functions represent linear lower bounds for different slopes λ and ν, for the inequality and equality constraints, respectively.

The difference $p^* - d^*$ is called the **optimal duality gap**. If the duality gap is zero, then **strong duality** holds:

$$d^* = p^*.$$

Conditions under which strong duality holds are called **constraint qualifications**. As an important case: strong duality holds if the primal problem is convex,[4] i.e. of the form

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, n \\ & Ax = b & \end{array} \tag{2.4}$$

for convex $f_0, \ldots, f_m$, *and* if **Slater's condition** holds: there exists some *strictly* feasible point[5] $\tilde{x} \in \text{relint}(\mathcal{D})$ such that

$$f_i(\tilde{x}) < 0 \quad i = 1, \ldots, m \quad A\tilde{x} = b.$$

A weaker version of Slater's condition is sufficient for strong convexity when some of the constraint functions $f_1, \ldots, f_k$ are affine (note the inequality constraints are no longer strict):

$$f_i(\tilde{x}) \leq 0 \quad i = 1, \ldots, k \quad f_i(\tilde{x}) < 0 \quad i = k+1, \ldots, m \quad A\tilde{x} = b.$$

A proof of this result is given in [1, Section 5.3.2].

[4] Strong duality can also hold for non-convex problems: see e.g. [1, p. 229].

[5] We denote by $\text{relint}(\mathcal{D})$ the relative interior of the set $\mathcal{D}$. This looks like the interior of the set, but is non-empty even when the set is a subspace of a larger space. See [1, Section 2.1.3] for the formal defintion.

2.4 A saddle point/game characterization of weak and strong duality

In this section, we ignore equality constraints for ease of discussion. We write the solution to the primal problem as an optimization

$$\begin{aligned}\sup_{\lambda \succeq 0} L(x,\lambda) &= \sup_{\lambda \succeq 0} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \right) \\ &= \begin{cases} f_0(x) & f_i(x) \leq 0,\ i = 1,\ldots,m \\ \infty & \text{otherwise.} \end{cases}\end{aligned}$$

In other words, we recover the primal problem when the inequality constraint holds, and get infinity otherwise. We can therefore write

$$p^* = \inf_x \sup_{\lambda \succeq 0} L(x,\lambda).$$

We already know

$$d^* = \sup_{\lambda \succeq 0} \inf_x L(x,\lambda).$$

Weak duality therefore corresponds to the **max-min inequality**:

$$\sup_{\lambda \succeq 0} \inf_x L(x,\lambda) \leq \inf_x \sup_{\lambda \succeq 0} L(x,\lambda). \tag{2.5}$$

which holds for general functions, and not just $L(x,\lambda)$. Strong duality occurs at a saddle point, and the inequality becomes an equality.

There is also a **game interpretation**: $L(x,\lambda)$ is a sum that must be paid by the person ajusting x to the person adjusting λ. On the right hand side of (2.5), player x plays first. Knowing that player 2 (λ) will maximize their return, player 1 (x) chooses their setting to give player 2 the worst possible options over all λ. The max-min inequality says that whoever plays second has the advantage.

2.5 Optimality conditions

If the primal is equal to the dual, we can make some interesting observations about the duality constraints. Denote by x^* the optimum solution of the original problem (the minimum of f_0 under its constraints), and by (λ^*, ν^*) the solutions to the dual. Then

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &\underset{\text{(a)}}{=} \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \\ &\underset{\text{(b)}}{\leq} f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \\ &\leq f_0(x^*), \end{aligned}$$

where in (a) we use the definition of g, in (b) we use that $\inf_{x\in\mathcal{D}}$ of the expression in the parentheses is necessarily no greater than its value at x^*, and the last line we use that at (x^*, λ^*, ν^*) we have $\lambda^* \succeq 0$, $f_i(x^*) \leq 0$, and $h_i(x^*) = 0$. From this chain of reasoning, it follows that

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0, \tag{2.6}$$

which is the condition of **complementary slackness**. This means

$$\begin{aligned} \lambda_i^* > 0 \quad &\Longrightarrow \quad f_i(x^*) = 0, \\ f_i(x^*) < 0 \quad &\Longrightarrow \quad \lambda_i^* = 0. \end{aligned}$$

Consider now the case where the functions f_i, h_i are differentiable, and the duality gap is zero. Since x^* minimizes $L(x, \lambda^*, \nu^*)$, the derivative at x^* should be zero,

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0.$$

We now gather the various conditions for optimality we have discussed. The **KKT conditions** for the primal and dual variables (x, λ, ν) are

$$\begin{aligned} f_i(x) &\leq 0,\ i = 1, \ldots, m \\ h_i(x) &= 0,\ i = 1, \ldots, p \\ \lambda_i &\geq 0,\ i = 1, \ldots, m \\ \lambda_i f_i(x) &= 0,\ i = 1, \ldots, m \\ \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) &= 0 \end{aligned}$$

If a convex optimization problem with differentiable objective and constraint functions satisfies Slater's conditions, then the KKT conditions are necessary and sufficient for global optimality.

3 The representer theorem

This description comes from Lecture 8 of Peter Bartlett's course on Statistical Learning Theory.

We are given a set of paired observations $(x_1, y_1), \ldots (x_n, y_n)$ (the setting could be regression or classification). We consider problems of a very general type: we want to find the function f in the RKHS $\mathcal{H}$ which satisfies

$$J(f^*) = \min_{f\in\mathcal{H}} J(f), \tag{3.1}$$

where

$$J(f) = L_y(f(x_1), \ldots, f(x_n)) + \Omega\left(\|f\|_{\mathcal{H}}^2\right),$$

Ω is non-decreasing, and y is the vector of y_i. Examples of loss functions might be

- Classification: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n \mathbb{I}_{y_i f(x_i) \leq 0}$ (the number of points for which the sign of y disagrees with that of the prediction $f(x)$),
- Regression: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n (y_i - f(x_i))^2$, the sum of squared errors (eg. when $\Omega\left(\|f\|_{\mathcal{H}}^2\right) = \|f\|_{\mathcal{H}}^2$, we are back to the standard ridge regression setting).

The representer theorem states that a solution to 3.1 takes the form

$$f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot).$$

If Ω is strictly increasing, all solutions have this form.

Proof: We write as f_s the projection of f onto the subspace

$$\operatorname{span}\{k(x_i, \cdot) : 1 \leq i \leq n\}, \tag{3.2}$$

such that

$$f = f_s + f_\perp,$$

where $f_s = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$. Consider first the regularizer term. Since

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2,$$

then

$$\Omega\left(\|f\|_{\mathcal{H}}^2\right) \geq \Omega\left(\|f_s\|_{\mathcal{H}}^2\right),$$

so this term is minimized for $f = f_s$. Next, consider the individual terms $f(x_i)$ in the loss. These satisfy

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_s + f_\perp, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_s, k(x_i, \cdot)\rangle_{\mathcal{H}},$$

so

$$L_y(f(x_1), \ldots, f(x_n)) = L_y(f_s(x_1), \ldots, f_s(x_n)).$$

Hence the loss $L(\ldots)$ only depends on the component of f in the subspace 3.2, and the regularizer $\Omega(\ldots)$ is minimized when $f = f_s$. If Ω is strictly non-decreasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is required at the minimum, otherwise this may be one of several minima.

4 Support vector classification

4.1 The linearly separable case

We consider problem of classifying two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.

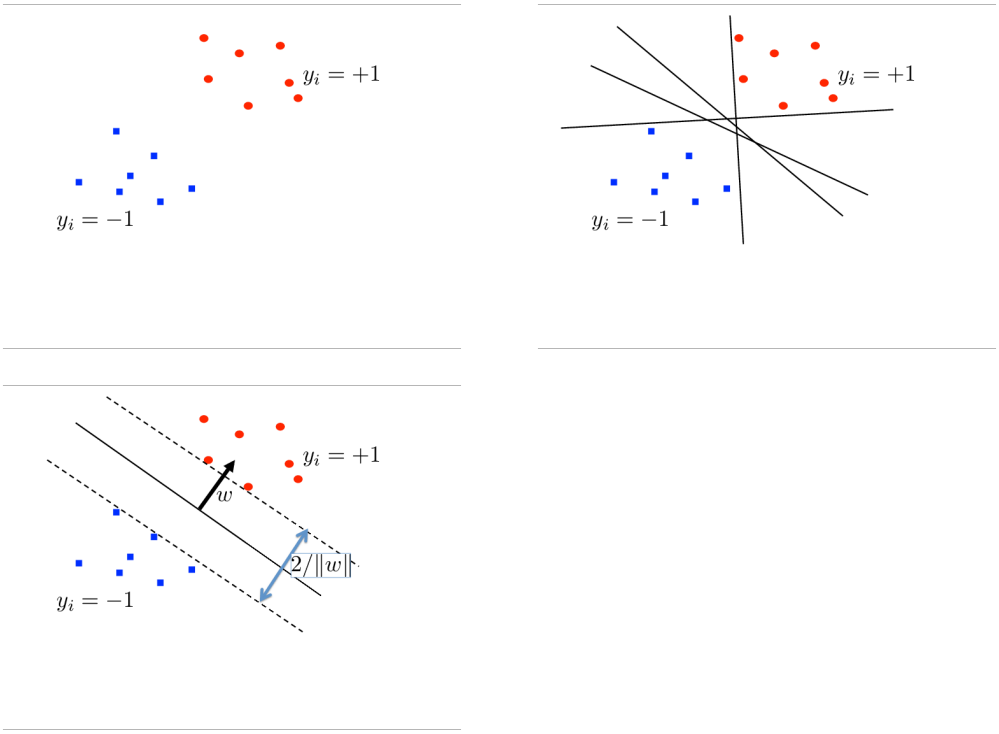

Figure 4.1: The linearly separable case. There are many linear separating hyperplanes, but only one max. margin separating hyperplane.

This is illustrated in Figure (4.1). As can be seen, there are infinitely many possible hyperplanes that solve this problem: the question is then: which one to choose? We choose the one which has the largest margin: it is the largest possible distance from both classes, and the *smallest* distance from each class to the separating hyperplane is called the margin.

This problem can be expressed as follows:[6]

$$\max_{w,b} (\text{margin}) = \max_{w,b} \left(\frac{2}{\|w\|} \right) \tag{4.2}$$

[6]It's easy to see why the equation below is the margin (the distance between the positive and negative classes): consider two points x_+ and x_- of opposite label, located on the margins. The width of the margin, d_m, is the difference $x_+ - x_-$ projected onto the unit vector in the direction w, or

$$d_m = (x_+ - x_-)^\top \frac{w}{\|w\|} \tag{4.1}$$

Subtracting the two equations in the constraints (4.3) from each other, we get

$$w^\top (x_+ - x_-) = 2.$$

Substituting this into (4.1) proves the result.

subject to

$$\begin{cases} \min\left(w^\top x_i + b\right) = 1 & i \,:\, y_i = +1, \\ \max\left(w^\top x_i + b\right) = -1 & i \,:\, y_i = -1. \end{cases} \tag{4.3}$$

The resulting classifier is

$$y = \text{sign}(w^\top x + b),$$

where sign takes value $+1$ for a positive argument, and -1 for a negative argument (its value at zero is not important, since for non-pathological cases we will not need to evaluate it there). We can rewrite to obtain

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{or} \quad \min_{w,b} \|w\|^2$$

subject to

$$y_i(w^\top x_i + b) \geq 1. \tag{4.4}$$

4.2 When no linear separator exists (or we want a larger margin)

If the classes are not linearly separable, we may wish to allow a certain number of errors in the classifier (points on the wrong side of the decision boudary). Ideally, we would optimise

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \mathbb{I}[y_i \left(w^\top x_i + b\right) < 0] \right),$$

where C controls the tradeoff between maximum margin and loss, and $\mathbb{I}(A) = 1$ if A holds true, and 0 otherwise (the factor of $1/2$ is to simplify the algebra later, and is not important: we can adjust C accordingly). This is a combinatorical optimization problem, which would be very expensive to solve. Instead, we replace the indicator function with a convex upper bound,

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \theta\left(y_i \left(w^\top x_i + b\right)\right) \right).$$

We use the hinge loss,

$$\theta(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha & 1-\alpha > 0 \\ 0 & \text{otherwise.} \end{cases}$$

although obviously other choices are possible (e.g. a quadratic upper bound). See Figure 4.2.

Substituting in the hinge loss, we get

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \theta\left(y_i \left(w^\top x_i + b\right)\right) \right).$$

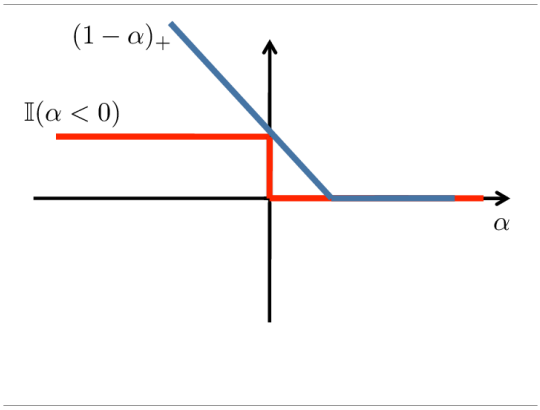

Figure 4.2: The hinge loss is an upper bound on the step loss.

or equivalently the constrained problem

$$\min_{w,b,\xi} \left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right) \tag{4.5}$$

subject to[7]

$$\xi_i \geq 0 \qquad y_i \left(w^\top x_i + b\right) \geq 1 - \xi_i$$

(compare with (4.4)). See Figure 4.3.

Now let's write the Lagrangian for this problem, and solve it.

$$L(w,b,\xi,\alpha,\lambda) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i) \tag{4.6}$$

with dual variable constraints

$$\alpha_i \geq 0, \qquad \lambda_i \geq 0.$$

We minimize wrt the primal variables w, b, and ξ.

Derivative wrt w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \qquad w = \sum_{i=1}^{n}\alpha_i y_i x_i. \tag{4.7}$$

Derivative wrt b:

$$\frac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0. \tag{4.8}$$

[7] To see this, we can write it as $\xi_i \geq 1 - y_i\left(w^\top x_i + b\right)$. Thus either $\xi_i = 0$, and $y_i\left(w^\top x_i + b\right) \geq 1$ as before, or $\xi_i > 0$, in which case to minimize (4.5), we'd use the smallest possible ξ_i satisfying the inequality, and we'd have $\xi_i = 1 - y_i\left(w^\top x_i + b\right)$.

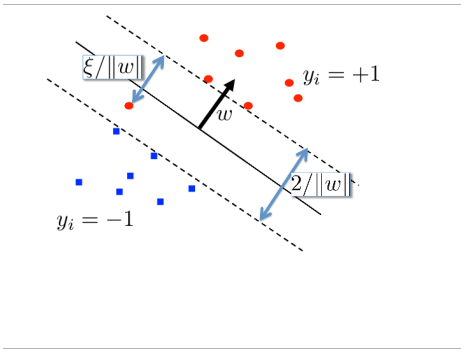

Figure 4.3: The nonseparable case. Note the red point which is a distance $\xi/\|w\|$ from the margin.

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i. \tag{4.9}$$

We can replace the final constraint by noting $\lambda_i \geq 0$, hence

$$\alpha_i \leq C.$$

Before writing the dual, we look at what these conditions imply about the scalars α_i that define the solution (4.7).

Non-margin SVs: $\alpha_i = C$:
Remember complementary slackness:

1. We immediately have $1 - \xi_i = y_i \left(w^\top x_i + b\right)$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, hence it is possible that $\xi_i > 0$.

Margin SVs: $0 < \alpha_i < C$:

1. We again have $1 - \xi_i = y_i \left(w^\top x_i + b\right)$
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Non-SVs: $\alpha_i = 0$

1. This time we have: $y_i \left(w^\top x_i + b\right) > 1 - \xi_i$
2. From $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

This means that the solution is sparse: all the points which are not either on the margin, or "margin errors", contribute nothing to the solution. In other words, only those points on the decision boundary, or which are margin errors,

contribute. Furthermore, the influence of the non-margin SVs is bounded, since their weight cannot exceed C: thus, severe outliers will not overwhelm the solution.

We now write the dual function, by substituting equations (4.7), (4.8), and (4.9) into (4.6), to get

$$\begin{aligned} g(\alpha,\lambda) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i) \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j - b\underbrace{\sum_{i=1}^{m}\alpha_i y_i}_{0} \\ &\qquad + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\underbrace{(C-\alpha_i)}_{\lambda_i}\xi_i \\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j. \end{aligned}$$

Thus, our goal is to maximize the dual,

$$g(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j,$$

subject to the constraints

$$0 \le \alpha_i \le C, \quad \sum_{i=1}^{n} y_i\alpha_i = 0.$$

So far we have defined the solution for w, but not for the offset b. This is simple to compute: for the margin SVs, we have $1 = y_i\left(w^\top x_i + b\right)$. Thus, we can obtain b from any of these, or take an average for greater numerical stability.

4.3 Kernelized version

We can straightforwardly define a maximum margin classifier in feature space. We write the original hinge loss formlation (ignoring the offset b for simplicity):

$$\min_w \left(\frac{1}{2}\|w\|_{\mathcal{H}}^2 + C\sum_{i=1}^{n}\left(1 - y_i\left\langle w, k(x_i,\cdot)\right\rangle_{\mathcal{H}}\right)_+\right)$$

for the RKHS $\mathcal{H}$ with kernel $k(x,\cdot)$. When we kernelize, we use the result of the representer theorem,

$$w = \sum_{i=1}^{n}\beta_i k(x_i,\cdot). \tag{4.10}$$

In this case, maximizing the margin is equivalent to minimizing $\|w\|_{\mathcal{H}}^2$: as we have seen, for many RKHSs (e.g. the RKHS corresponding to a Gaussian kernel), this corresponds to enforcing smoothness.

Substituting (4.10) and introducing the ξ_i variables, get

$$\min_{w,b} \left(\frac{1}{2}\beta^\top K\beta + C\sum_{i=1}^{n} \xi_i \right) \tag{4.11}$$

where the matrix K has i,jth entry $K_{ij} = k(x_i, x_j)$, subject to

$$\xi_i \geq 0 \qquad y_i \sum_{j=1}^{n} \beta_j k(x_i, x_j) \geq 1 - \xi_i$$

Thus, the primal variables w are replaced with β. The problem remains convex since K is positive definite. With some calculation (exercise!), the dual becomes

$$g(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j k(x_i, x_j),$$

subject to the constraints

$$0 \leq \alpha_i \leq C,$$

and the decision function takes the form

$$w = \sum_{i=1}^{n} y_i \alpha_i k(x, \cdot).$$

4.4 The ν-SVM

It can be hard to interpret C. Therefore we modify the formulation to get a more intuitive parameter. Again, we drop b for simplicity. Solve

$$\min_{w,\rho,\xi} \left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i \right)$$

subject to

$$\begin{aligned} \rho &\geq 0 \\ \xi_i &\geq 0 \\ y_i w^\top x_i &\geq \rho - \xi_i, \end{aligned}$$

where we see that we now optimize the margin width ρ. Thus, rather than choosing C, we now choose ν; the meaning of the latter will become clear shortly.

The Lagrangian is

$$\frac{1}{2}\|w\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i - \nu\rho + \sum_{i=1}^{n} \alpha_i \left(\rho - y_i w^\top x_i - \xi_i\right) + \sum_{i=1}^{n} \beta_i(-\xi_i) + \gamma(-\rho)$$

for $\alpha_i \geq 0$, $\beta_i \geq 0$, and $\gamma \geq 0$. Differentiating wrt each of the primal variables w, ξ, ρ, and setting to zero, we get

$$\begin{aligned} w &= \sum_{i=1}^{n} \alpha_i y_i x_i \\ \alpha_i + \beta_i &= \frac{1}{n} \qquad (4.12) \\ \nu &= \sum_{i=1}^{n} \alpha_i - \gamma \qquad (4.13) \end{aligned}$$

From $\beta_i \geq 0$, equation (4.12) implies

$$0 \leq \alpha_i \leq n^{-1}.$$

From $\gamma \geq 0$ and (4.13), we get

$$\nu \leq \sum_{i=1}^{n} \alpha_i.$$

Let's now look at the complementary slackness conditions.
Assume $\rho > 0$ at the global solution, hence $\gamma = 0$, and (4.13) becomes

$$\sum_{i=1}^{n} \alpha_i = \nu. \qquad (4.14)$$

1. Case of $\xi_i > 0$: then complementary slackness states $\beta_i = 0$, hence from (4.12) we have $\alpha_i = n^{-1}$ for these points. Denote this set as $N(\alpha)$. Then

   $$\sum_{i \in N(\alpha)} \frac{1}{n} = \sum_{i \in N(\alpha)} \alpha_i \leq \sum_{i=1}^{n} \alpha_i = \nu,$$

   so

   $$\frac{|N(\alpha)|}{n} \leq \nu,$$

   and ν is an upper bound on the number of non-margin SVs.

2. Case of $\xi_i = 0$. Then $\alpha_i < n^{-1}$. Denote by $M(\alpha)$ the set of points $n^{-1} > \alpha_i > 0$. Then from (4.14),

   $$\nu = \sum_{i=1}^{n} \alpha_i = \sum_{i \in N(\alpha)} \frac{1}{n} + \sum_{i \in M(\alpha)} \alpha_i \leq \sum_{i \in M(\alpha) \cup N(\alpha)} \frac{1}{n},$$

   thus

   $$\nu \leq \frac{|N(\alpha)| + |M(\alpha)|}{n},$$

   and ν is a lower bound on the number of support vectors with non-zero weight (both on the margin, and "margin errors").

Substituting into the Lagrangian, we get

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \rho\nu - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n}\alpha_i\rho - \sum_{i=1}^{n}\alpha_i\xi_i \\
&\qquad - \sum_{i=1}^{n}\left(\frac{1}{n} - \alpha_i\right)\xi_i - \rho\left(\sum_{i=1}^{n}\alpha_i - \nu\right) \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j
\end{aligned}$$

Thus, we must maximize

$$g(\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j,$$

subject to

$$\sum_{i=1}^{n}\alpha_i \geq \nu \qquad 0 \leq \alpha_i \leq \frac{1}{n}.$$

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004.