25 Oct 2021

CS492(F) - Computational Learning Theory

Chap 5. Support Vector Machine

1. Overview

(in short, SVM)

(1) Support Vector Machine is one of the most elegant machine learning algorithms, which performs well in practice and comes with a strong and somewhat surprising theoretical guarantee. Studying it will help you to improve your toolset for developing or selecting a learning algorithm. It will also help you to see how the learning theory that we studied can be used to analyse realistic learning algorithms.

(2) SVM is a learning algo. for classification. Given a labelled sample $S = ((x_1, y_1), \dots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ with $\mathcal{Y} = \{-1, 1\}$, it computes a linear classifier h, i.e., a function from $\mathcal{X}$ to $\mathcal{Y}$ that has the following form:

$$h(x) = \operatorname{sgn}(f(x)), \quad f(x) = \langle w, x \rangle + b \quad \text{for some } w \in \mathbb{R}^n, b \in \mathbb{R}$$

of course, here we are assuming that $\mathcal{X} \subseteq \mathbb{R}^n$ so that the inner product is well-defined.

(3) The key idea behind SVM is to maximize so called margin when computing h. We will soon explain what this means. One important consequence of margin maximization is that we get generalization bound independent of the dimension n of the input x. ($\mathcal{X} \subseteq \mathbb{R}^n$).

Contrast this with the generalisation bound for the set of hyperplanes. $\mathcal{H}_{PL} = \{ x \mapsto \text{sgn}(\langle w, x \rangle + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R} \}$ that we derived in Chap 3. We recall the bound below:

① $\text{VCDim}(\mathcal{H}_{PL}) = n+2$ .... See Example 3.12 in the textbook.

② By Cor 3.19, we have the following:

For all $D \in P(X \times Y)$, $m \in \mathbb{N}$, $\delta > 0$,

$$\mathbb{P}_{S \sim D^m}\left[\forall h \in \mathcal{H}_{PL}.\ R(h) \le \hat{R}_S(h) + \sqrt{\frac{2(n+1)\log\frac{em}{n+1}}{m}} + \sqrt{\frac{\log 1/\delta}{2m}} \right] \ge 1 - \delta.$$

③ Note that the bound becomes loose as n increases.

④ The independence on the input dimension n indicates that SVM might work well when n is very large or even infinite. Such very large or infinite n arises when we use a large number of features to express inputs to our classifier h (or its linear function f).

⑤ Another important and appealing property of SVM is that (the dual version of) the algorithm and its result (i.e., learnt classifier) both process inputs x only via inner products. This lets SVM be combined nicely with kernel method, which we will study after this chapter. This combination enables SVM to work on inputs of infinite dimension.

2 Support Vector Machine ... Separable case.

(1) We explain SVM in a simpler setting first where labelled inputs can be separated by a hyperplane.

(2) Assumption: about the distribution $D \in Pr(\mathcal{X} \times \mathcal{Y})$.

$\exists w, b$ s.t.

$$\mathbb{P}_{(x,y)\sim D}\left[(\langle w, x\rangle + b)\, y > 0 \right] = 1.$$

Intuitively, this means that $f(x) = \langle w, x\rangle + b$ classifies every example from D correctly. We will also assume that $\mathcal{X} = \mathbb{R}^n$.

(3) SVM algorithm: (primal form).

minimization of the size of w.

Input: $S = ((x_1, y_1), \dots, (x_m, y_m)) \sim D^m$

$$(w^*, b^*) \overset{\text{def}}{=} \operatorname*{argmin}_{\substack{w \in \mathbb{R}^n \\ b \in \mathbb{R}}} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i(\langle w, x_i\rangle + b) \geq 1 \quad \text{for all } i \in [m].$$

$$f^*(x) \overset{\text{def}}{=} \langle w^*, x\rangle + b^* \;;\; h^*(x) \overset{\text{def}}{=} \operatorname{sgn}(f^*(x))$$

1, not 0.

return h^*

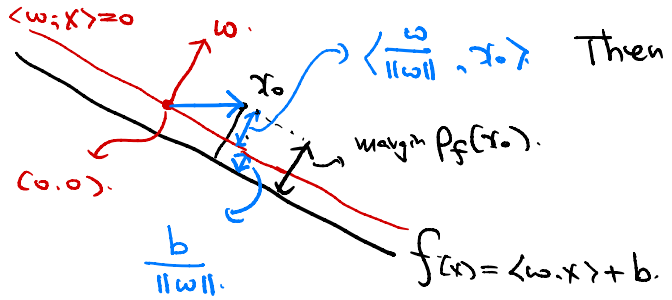
(4) By our assumption, the constraint of the SVM optimization is satisfiable (with probability 1). Note that b influences only the constraint, not the optimization objective. Furthermore, it can be eliminated in the constraint: when $J = \{i \mid y_i = +1\}$,

$$\left(\exists b.\ \forall i \in [m].\ y_i(\langle w, x_i\rangle + b) \geq 1\right) \Leftrightarrow \forall j \in J\ \forall i \in [m] \setminus J.\ \ -1 - \langle w, x_i\rangle \geq 1 - \langle w, x_j\rangle.$$

If we do so, we get an optimization problem over w only whose objective is strictly convex and whose constraint is the conjunction of linear inequalities. Since the constraint is satisfiable, there is a unique sol. w^*.

(5) To understand the meaning of the optimization of SVM, we define the geometric margin $\rho_f(x_0)$ of x_0 to a hyperplane $f(x) = \langle x, w \rangle + b$.

It is the distance of x_0 to f.



Then, $\rho_f(x_0) = \dfrac{|\langle w, x_0 \rangle + b|}{\|w\|}$

For a sample $S_x = (x_1, \dots, x_m)$, the geometric margin ρ_f over S_x is $\min_{i \in [m]} \rho_f(x_i)$.

(6) The next derivation shows that the SVM computes a hyperplane with the maximum geometric margin ρ_f over S_x.

Let $\rho_{w,b}(x_0) = \rho^{(x_0)}_{x \mapsto \langle w, x \rangle + b}$

$$\sup_{w,b} \min_{i \in [m]} \rho_{w,b}(x_i)$$
$$\forall i \; y_i (\langle w, x_i \rangle + b) > 0$$

$$= \sup_{w,b} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|}$$
$$\forall i. \; y_i (\langle w, x_i \rangle + b) > 0.$$

$$= \sup_{w,b} \min_{i \in [m]} \frac{y_i (\langle w, x_i \rangle + b)}{\|w\|}$$
$$\forall i. \; y_i (\langle w, x_i \rangle + b) > 0.$$

$$= \sup_{w,b} \min_{i \in [m]} \frac{y_i (\langle w, x_i \rangle + b)}{\|w\|} = \sup_{w,b} \frac{1}{\|w\|}$$
$$\min_{i \in [m]} y_i (\langle w, x_i \rangle + b) = 1 \qquad \min_{i \in [m]} y_i (\langle w, x_i \rangle + b) = 1$$

$$= \boxed{\sup_{w,b} \frac{1}{\|w\|}}$$
$$\boxed{\forall i \in [m], \; y_i (\langle w, x_i \rangle + b) \geq 1}$$

(7) If the maximum of the above supremum exists, it also solves the following optimization problem:

1
$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{subject to} \quad y_i(\langle w, x_i\rangle + b) \geq 1 \quad \text{for all } i \in [m]$$

SVM

Also, if this optimization problem has a solution, so does the supremum in the previous page. By the explanation in (4), there exists a unique w such that for some b, the SVM optimization problem is solved by (w, b).

(8) Dual optimization problem and support vectors.

① The SVM optimization problem is convex, and can be dualized via Lagrangian L and Lagrange variables $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, one for each conjunct $y_i(\langle w, x_i\rangle + b) \geq 1$, with $\alpha_i \geq 0$.

Lagrange variable · RHS of $\geq$

$$L(w, b, \alpha) \overset{\text{def}}{=} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\langle w, x_i\rangle + b))$$

↑ each conjunct: $0 \geq 1 - y_i(\langle w, x_i\rangle + b)$

Then,

$$\min_{w,b} \frac{1}{2}\|w\|^2 = \inf_{w,b} \sup_{\alpha} L(w, b, \alpha)$$
$$\text{s.t. } \forall i \in [m].\ y_i(\langle w, x_i\rangle + b) \geq 1 \qquad \text{s.t. } \forall i \in [m].\ \alpha_i \geq 0.$$

② If we swap inf and sup, we get something potentially smaller:

2
$$\sup_{\alpha} \inf_{w,b} L(w, b, \alpha) \leq \inf_{w,b} \sup_{\alpha} L(w, b, \alpha)$$
$$\forall i \in [m].\ \alpha_i \geq 0 \qquad \forall i \in [m].\ \alpha_i \geq 0$$

But for this specific SVM optimization (i.e., 1 above), the equality holds. Furthermore, the so called KKT condition applies in this case, which implies that it suffices to consider

(w, b, α)'s satisfying the following conditions:

c.i) $\nabla_w \mathcal{L}(w,b,\alpha) = 0$ $\wedge$ c.ii) $\nabla_b \mathcal{L}(w,b,\alpha) = 0$

$\wedge$ c.iii) $\forall i \in [m].\ \alpha_i (1 - y_i(\langle w, x_i \rangle + b)) = 0.$

Note:

c.i) $\nabla_w \mathcal{L}(w,b,\alpha) = 0 \Rightarrow w + \sum_{i=1}^{m} -\alpha_i y_i x_i = 0.$

$\therefore w = \sum_{i=1}^{m} \alpha_i y_i x_i.$

c.ii) $\nabla_b \mathcal{L}(w,b,\alpha) = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0.$

c.iii) $\forall i \in [m].\ \alpha_i [1 - y_i(\langle w, x_i \rangle + b)] = 0 \Rightarrow \forall i \in [m].\ \alpha_i = 0$ or

$y_i(\langle w, x_i \rangle + b) = 1.$

When these conditions hold,

$$\begin{aligned}\mathcal{L}(w,b,\alpha) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\langle w, x_i \rangle + b)) \\ &= \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i y_i \alpha_j y_j \langle x_j, x_i \rangle \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i y_i \alpha_j y_j \langle x_i, x_j \rangle.\end{aligned}$$

Thus, the dual, equivalent optimization problem is:

3 $$\sup_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle.$$
$$\text{s.t.}\quad \alpha_i \geq 0 \ \wedge \ \sum_{i=1}^{m} \alpha_i y_i = 0$$

(3) In typical use cases, there exists α that attains the supremum in 3 and has some $\alpha_{i_0} \neq 0$. Assume such an α.

Using it, we can define w and b:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i, \qquad b = y_{i_0} - \langle w, x_{i_0} \rangle \text{ for } i_0 \text{ with } \alpha_{i_0} \neq 0.$$

Thus, the hyperplane built by SVM is $= y_{i_0} - \sum \alpha_i y_i \langle x_i, x_{i_0} \rangle$

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x \rangle - \langle w, x_{i_0} \rangle + y_{i_0} = \sum_{i=1}^{m} \alpha_i y_i (\langle x_i, x \rangle - \langle x_i, x_{i_0} \rangle) + y_{i_0}.$$

④ Note that only those x_i with $\alpha_i \neq 0$ contribute to the definition of f. Such x_i's are called support vectors. Another important fact is that x, x_i are used in f only through their inner products.

⑤ Finally, we can derive the geometric margin of f as follows:

$$\alpha_i y_i b = \alpha_i y_i \left(y_i - \sum_{j=1}^{m} \alpha_j y_j \langle x_i, x_j \rangle \right) \quad \text{by the first and third conditions of KKT.}$$

$$\therefore \sum_{i=1}^{m} \alpha_i y_i b = \sum_{i=1}^{m} \alpha_i - \|w\|^2$$

$$\therefore 0 = \sum_{i=1}^{m} \alpha_i - \|w\|^2 \quad \text{by the second condition of KKT.}$$

$$\therefore \rho^2 = \frac{1}{\|w\|^2} = \frac{1}{\sum_{i=1}^{m} \alpha_i}.$$

3. Support Vector Machine : Non-separable Case.

(1) In practice, often $S \cup D^{neg}$ cannot be separated by a hyper-plane. How should we extend SVM to handle this non-separable case?

One answer that we will look at is to use slack variables. For each conjunct $y_i(\langle w, x_i \rangle + b) \geq 1$ in the original SVM optimisation [1], we introduce a slack variable $\xi_i \geq 0$, which measures the amount of violation of this conjunct, and replace the conjunct by

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$$

and include ξ_i as a part of optimisation objective.

(2) Thus, the SVM for the non-sep. case works as follows:

Input $S = ((x_1, y_1), \dots, (x_m, y_m)) \sim D^m$.

$$(w^*, b^*, \xi^*) \overset{\text{def}}{=} \underset{w,b,\xi}{\arg\min} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i^p$$

subject to $y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$

$\xi_i \geq 0$

for all $i \in [m]$.

hyperparameters.

$p \geq 1$.

new part.

penalize incorrect or small-margin predictions.

4

$$f^*(x) \overset{\text{def}}{=} \langle w^*, x\rangle + b^*$$

$$h^*(x) \overset{\text{def}}{=} \operatorname{sgn}(f^*(x))$$

return h^*.

① The new $C\sum_{i=1}^{m} \xi_i^p$ corresponds to empirical risk, and the old $\frac{1}{2}\|w\|^2$ corresponds to regularizer or measure of model complexity. We didn't have something like $C\sum_{i=1}^{m} \xi_i^p$ in the separable case, because in a sense we can achieve zero empirical risk at that time.

② Note that the minimization forces

$$\xi_i = \max(0, 1 - y_i(\langle w, x_i\rangle + b))$$

If we let $\Phi_p(u) = \max(0, (1+u)^p)$, then the optimization from above is equivalent to:

$$\underset{w,b}{\arg\min} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \Phi_p(-y_i(\langle w, x_i\rangle + b))$$

which is closely related to what we looked at in Sec 4.7, since Φ_p is convex and non-decreasing. In fact, Φ_1 is precisely the Hinge loss in Sec 4.7.

(3) The new optimisation problem can also be dualised via Lagrangian and Lagrange variables. In this case, we can also switch inf and sup, and the KKT conditions can be imposed to transform the optimization problem and analyse its solution. In so doing, we can characterize support vectors as well.

① Lagrange variables ... α_i for $0 \geq 1-\xi_i - y_i(\langle w, x_i\rangle + b)$

β_i for $0 \geq -\xi_i$

② Lagrangian L.

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left(1-\xi_i - y_i(\langle w, x_i\rangle + b)\right) + \sum_{i=1}^{m}\beta_i(-\xi_i)$$

③ KKT conditions.

$\nabla_w L = 0 \Rightarrow w = \sum_{i=1}^{m}\alpha_i y_i x_i$.

$\nabla_b L = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i = 0$.

$\nabla_{\xi_i} L = 0 \Rightarrow C - \alpha_i - \beta_i = 0$ (i.e., $\alpha_i + \beta_i = C$)

$\forall i.\ \alpha_i(1-\xi_i - y_i(\langle w, x_i\rangle + b)) = 0 \Rightarrow \forall i.\ \alpha_i = 0$ or $y_i(\langle w, x_i\rangle + b) = 1-\xi_i$

$\forall i\ \ \beta_i(-\xi_i) = 0 \Rightarrow \forall i\ \ \beta_i = 0$ or $\xi_i = 0$.

④ By what we explained,

optimization of SVM $= \inf_{w,b,\xi}\sup_{\alpha,\beta} L(w, b, \xi, \alpha, \beta)$ s.t. $\alpha_i \geq 0$ and $\beta_i \geq 0$ for all $i \in [m]$

$= \sup_{\alpha,\beta}\inf_{w,b,\xi} L(w, b, \xi, \alpha, \beta)$ for all i s.t. $\alpha_i \geq 0$, $\beta_i \geq 0$, $C = \alpha_i + \beta_i$, $\sum_{i=1}^{m}\alpha_i y_i = 0$, $w = \sum_{i=1}^{m}\alpha_i y_i x_i$.

I

We simplify [5] by rewriting $\mathcal{L}$ and $\beta_i \geq 0$:

$$\beta_i \geq 0 \quad \dashrightarrow \quad \alpha_i \leq C.$$

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \sum_{i,j=1}^{m} \frac{1}{2} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle.$$

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

Thus, [5] can be simplified as follows.

[6]
$$\sup_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$
$$\text{subj. to} \quad 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \quad \text{for all } i \in [m].$$

Note that the only difference wrt. the objective for the sep. case is the presence of the upper bound C. By the way, this upper bound ensures that the constraint of [6] defines a compact set, so that the optimization objective (as a continuous function on α) can be maximized by some α. So, we can replace sup by max.

(5) If a maximizing α of [6] has some α_{i_0} with $0 < \alpha_{i_0} < C$, we can derive w and b as before:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad b = y_{i_0} - \sum_{j=1}^{m} \alpha_j y_j \langle x_j, x \rangle.$$

Support vectors are, however, slightly different. x_i with $\alpha_i \neq 0$ is still called support vector, but $\alpha_i \neq 0$ doesn't imply that

$$y_i (\langle w, x_i \rangle + b) = 1$$

Instead, it implies that $y_i (\langle w, x_i \rangle + b) = 1 - \xi_i$.

The traditional support vectors are x_i's with $\alpha_i \neq 0$ and $C - \alpha_i = \beta_i \neq 0$.

4. Margin Theory

(1) As we said in the beginning of this chapter, maximizing the margin ensures that SVM has a generalization bound independent of the input dimension. Specifically, we have the following bound:

for all $r, r' > 0$, $D \in Pr(X \times Y)$, $m \in \mathbb{N}$, and $\delta > 0$,

if $\mathbb{P}_{(x,y)\sim D}[\|x\| \le r] = 1$ and $\mathcal{H} = \{x \mapsto \mathrm{sgn}(\langle w, x\rangle) \mid w \in \mathbb{R}^n \text{ and } \|w\| \le \Lambda\}$

then (margin ρ; related to slack variable ξ_i)

$$\mathbb{P}_{S\sim D^m}\left[\forall h_w \in \mathcal{H}.\ \forall \rho \in (0, r'].\ R(h_w) \le \frac{1}{m}\sum_{i=1}^{m} \max\left(0, 1 - \frac{y_i(\langle w, x_i\rangle)}{\rho}\right) + 4\sqrt{\frac{r^2\Lambda^2}{\rho^2 m}} + \sqrt{\frac{\log\log_2(2r'/\rho)}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \ge 1-\delta.$$

large margin ρ makes these small.

※ bound not exactly for the formula used by SVM, but for closely related one

Contrast it with the VC-dimension bound for the hypo. set with hyperplanes $\mathcal{H}_{PL}$ (which is very closely related to $\mathcal{H}$ from above).

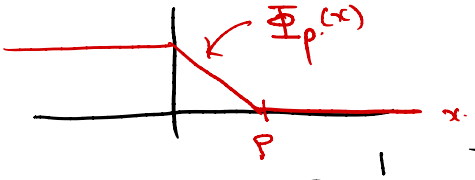
$$\mathbb{P}_{S\sim D^m}\left[\forall h \in \mathcal{H}_{PL}.\ R(h) \le \hat{R}_S(h) + \sqrt{\frac{2(n+1)\log\frac{em}{n+1}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta.$$

dependence on n, the dimension of the inputs.

The SVM margin bound says that with high probability, the generalization error of any h is bounded by the empirical slack-variable-based error plus terms related to margin.

(2) We will now discuss margin theory, which shows the SVM margin bound in the previous page.

Def 5.5 Let $\rho > 0$.



① $\Phi_\rho : \mathbb{R} \to \mathbb{R}_+ = [0,\infty)$

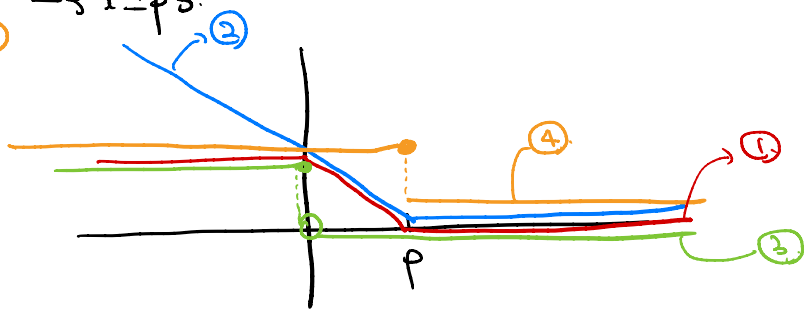
$$\Phi_\rho(x) \overset{\text{def}}{=} \min(1, \max(1 - \frac{x}{\rho}, 0)) = \begin{cases} 1 & \text{if } x \le 0 \\ 1 - \frac{x}{\rho} & \text{if } 0 \le x \le \rho \\ 0 & \text{if } \rho \le x \end{cases}$$

② The ρ-margin loss function is $L_\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ defined by

$$L_\rho(y, y') \overset{\text{def}}{=} \Phi_\rho(yy').$$

① ② ③

Compare Φ_ρ (for ρ) with the hinge loss and the functions $x \mapsto \mathbb{1}_{\{x \le 0\}}$ and $x \mapsto \mathbb{1}_{\{x \le \rho\}}$.



Only ② is convex.

Def 5.6 $S = ((x_1, y_1), \dots, (x_m, y_m))$ ... sample $f : \mathcal{X} \to \mathbb{R}$.

The empirical ρ-margin loss of f with respect to S is (or just empirical margin loss.)

$$\widehat{R}_{S,\rho}(f) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m L_\rho(f(x_i), y_i) = \frac{1}{m}\sum_{i=1}^m \Phi_\rho(y_i f(x_i)).$$

→ almost = in practice.

Note that

$$\widehat{R}_S(\operatorname{sgn} \circ f) \le \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{\{f(x_i) y_i \le 0\}} \le \widehat{R}_{S,\rho}(f) \le \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{\{f(x_i) y_i \le \rho\}}.$$

(3) Let $\mathcal{F} \subseteq [\mathcal{X} \to \underset{\text{measurable}}{\mathbb{R}}]$ We consider two induced sets of functions from $\mathcal{X}\times\mathcal{Y}$ to $\mathbb{R}$:

$$\mathcal{G}_0 = \{ z=(x,y) \longmapsto yf(x) \mid f \in \mathcal{F} \}.$$

$$\mathcal{G} = \{ \underbrace{\Phi_p \circ g_0}_{} \mid g_0 \in \mathcal{G}_0 \}$$

$$= (x,y) \longmapsto \Phi_p(yf(x)) \quad \text{where } g_0(x,y) = yf(x)$$

$$= (x,y) \longmapsto L_p(f(x),y)$$

Note that the range of each $g \in \mathcal{G}$ is $[0,1]$. Thus, we can apply Thm 3.3., and get the following:

$\forall D \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})\ \forall m \in \mathbb{N}\ \forall \delta > 0.$

$$\mathbb{P}_{S\sim D^m}\left[\forall g \in \mathcal{G}.\ \mathbb{E}_{(x,y)\sim D}[g(x,y)] \le \frac{1}{m}\sum_{i=1}^{m} g(x_i,y_i) + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \right] \ge 1-\delta.$$

and

$$\mathbb{P}_{S\sim D^m}\left[\forall g \in \mathcal{G}.\ \mathbb{E}_{(x,y)\sim D}[g(x,y)] \le \frac{1}{m}\sum_{i=1}^{m} g(x_i,y_i) + 2\hat{\mathcal{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \right] \ge 1-\delta.$$

But the sum on the RHS is $\hat{R}_{S,p}(f)$ where $g(x,y) = L_p(f(x),y)$.

Also, the expectation on the LHS is larger than or equal to the risk of $(\operatorname{sgn}\circ f)$:

$$R(\operatorname{sgn}\circ f) = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}\{\operatorname{sgn}(f(x)) \neq y\}]$$

$$\le \mathbb{E}_{(x,y)\sim D}[\mathbb{1}\{f(x)y \le 0\}] \le \mathbb{E}_{(x,y)\sim D}[\Phi_p(f(x)y)]$$

$L_p(f(x),y) = g(x,y)$

hypothesis built from f.

By Talagrand's lemma that we will explain soon,

$$\hat{\mathcal{R}}_S(\mathcal{G}) \le \frac{1}{p}\hat{\mathcal{R}}_S(\mathcal{G}_0) \quad \text{and} \quad \mathcal{R}_m(\mathcal{G}) \le \frac{1}{p}\mathcal{R}_m(\mathcal{G}_0).$$

Furthermore,

$$\hat{\mathcal{R}}_S(\mathcal{G}_0) = \frac{1}{m}\mathbb{E}_{\sigma\sim\mathrm{Unif}\{-1,1\}^m}\left[\sup_{g\in\mathcal{G}_0}\sum_{i=1}^{m}\sigma_i g(x_i,y_i)\right] = \frac{1}{m}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i y_i f(x_i)\right] = \frac{1}{m}\mathbb{E}_{\sigma}\left[\sup_{f}\sum_{i=1}^{m}\sigma_i f(x_i)\right]$$

$= \hat{\mathcal{R}}_S(\mathcal{F})$

$(\sigma_1,\dots,\sigma_m)$ and $(y_1\sigma_1,\dots,y_m\sigma_m)$ have the same distribution

By taking the expectation over S, we can also get

$$R_m(G_0) = R_m(F).$$

Thus, we get the following generalization bound:

Thm 5.8. [Margin Bound for Binary Classification]

$\forall D \in Pr(X \times Y)$ $\forall m \in \mathbb{N}$ $\forall \delta > 0$.

$$\mathbb{P}_{S \sim D^m}\left[\forall f \in F \;\; R(\operatorname{sgn} \circ f) \le \hat{R}_{S,\rho}(f) + \frac{2}{\rho} R_m(F) + \sqrt{\frac{\log 1/\delta}{2m}}\right] \ge 1-\delta.$$

and

$$\mathbb{P}_{S \sim D^m}\left[\forall f \in F \;\; R(\operatorname{sgn} \circ f) \le \hat{R}_{S,\rho}(f) + \frac{2}{\rho} \hat{R}_S(F) + 3\sqrt{\frac{\log 2/\delta}{2m}}\right] \ge 1-\delta.$$

Note that using large ρ reduces the model-complexity part of the bound, but increases the margin error. So, there is a trade-off here.

To complete the proof of the thm, we need to show the next lemma

Lemma 5.7 [Talagrand's lemma]

$\Phi_1, \dots, \Phi_m$ ---- l-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$.

G_1 ---- any set of measurable functions from Z to $\mathbb{R}$ (bounded)

$S = (z_1, \dots, z_m) \in Z^m$.

$\Rightarrow$

$$\frac{1}{m} \mathbb{E}_{\vec{\sigma} \in \text{Unif}\{-1,1\}^m} \left[\sup_{g \in G_1} \sum_{i=1}^m \sigma_i \times (\Phi_i(g(x_i))) \right]$$

$$\le \frac{l}{m} \mathbb{E}_{\vec{\sigma}} \left[\sup_{g \in G_1} \sum_{i=1}^m \sigma_i g(x_i) \right] = l \hat{R}_S(G_1).$$

In particular, if $\Phi_i = \Phi$ for all $i \in [m]$,

$$\hat{R}_S(\Phi \circ G_1) \le l \hat{R}_S(G_1).$$

means $\{\Phi \circ g \mid g \in G_1\}$.

Proof We will show that

$$\mathop{\mathbb{E}}_{\vec{\sigma}}\left[\sup_{g\in\mathcal{G}_1}\sum_{i=1}^{m}\sigma_i\times\Phi_i(g(z_i))\right]\le l\mathop{\mathbb{E}}_{\vec{\sigma}}\left[\sup_{g\in\mathcal{G}_1}\sum_{i=1}^{m}\sigma_i g(z_i)\right].$$

The proof is by induction on m.

The base case is $m=0$. Then, both sides of the inequality are $-\infty$. So, the inequality holds.

For the inductive case $m>0$, we will show that

$$\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{g} u_{m-1}(g)+\sigma_m\Phi_m(g(z_m))\right] \le \mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{g} u_{m-1}(g)+\sigma_m l g(z_m)\right]$$

where $u_{m-1}(g)=\sum_{i=1}^{m-1}\sigma_i\Phi_i(g(z_i))$

Pick any $\varepsilon>0$. Then, for all $(\sigma_1,\dots,\sigma_{m-1})$, $\exists g_1,g_2\in\mathcal{G}_1$ s.t.

$$u_{m-1}(g_1)+\Phi(g_1(z_m)) \ge (1-\varepsilon)\sup_{g\in\mathcal{G}_1} u_{m-1}(g)+\Phi(g(z_m))$$
$$u_{m-1}(g_2)-\Phi(g_2(z_m)) \ge (1-\varepsilon)\sup_{g\in\mathcal{G}_1} u_{m-1}(g)-\Phi(g(z_m)),$$

because the supremums in both are bounded. Note that $u_{m-1}(g)$ depends on $(\sigma_1,\dots,\sigma_{m-1})$, and so do the choices of g_1 and g_2.

Thus, $(1-\varepsilon)\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{g\in\mathcal{G}_1} u_{m-1}(g)+\sigma_m\Phi_m(g(z_m))\right]$

$$=(1-\varepsilon)\left(\frac{1}{2}\left[\sup_g u_{m-1}(g)+\Phi_m(g(z_m))\right]+\frac{1}{2}\left[\sup_g u_{m-1}(g)-\Phi_m(g(z_m))\right]\right)$$
$$\le \frac{1}{2}\left(u_{m-1}(g_1)+\Phi_m(g_1(z_m))\right)+\frac{1}{2}\left(u_{m-1}(g_2)-\Phi_m(g_2(z_m))\right)$$
$$\le \frac{1}{2}\left(u_{m-1}(g_1)+u_{m-1}(g_2)\right)+\frac{1}{2}\left(\Phi_m(g_1(z_m))-\Phi_m(g_2(z_m))\right).$$

Let $s=\operatorname{sgn}(g_1(z_m)-g_2(z_m))$.

$$\le \frac{1}{2}\left[\left(u_{m-1}(g_1)+u_{m-1}(g_2)\right)+s l\left(g_1(z_m)-g_2(z_m)\right)\right] \quad (\because l\text{-Lipschitz})$$
$$=\frac{1}{2}\left(u_{m-1}(g_1)+s l g_1(z_m)\right)+\frac{1}{2}\left(u_{m-1}(g_2)-s l g_2(z_m)\right)$$
$$\le \frac{1}{2}\left(\sup_{g\in\mathcal{G}_1}\left(u_{m-1}(g)+s l g(z_m)\right)\right)+\frac{1}{2}\left(\sup_{g\in\mathcal{G}_1}\left(u_{m-1}(g)-s l g(z_m)\right)\right)$$

$= \mathbb{E}\left[\sup_{g \in G} \sigma_m \, l_{g(z_m)} \right]$.

↑ Symmetrization by the Rademacher variable σ_m.

D.

(4). To get a meaningful generalization bound for SVM, we have to adjust two aspects of Thm 5.8. First, we need to make it uniform with respect to the margin ρ. Doing so would let us say something about an algorithm that computes ρ based on a sample S and so uses a random ρ. Second, we need to replace $R_m(F)$ by something easier to deal with.

The next results do these adjustments.

Thm 5.9 $\quad F \subseteq [X \xrightarrow{\text{measurable}} \mathbb{R}]$.

$P \in Pr(X \times Y) \quad \delta > 0 \quad r > 0 \quad m \in \mathbb{N}$.

$\Rightarrow$

$$P^m\Big[\forall f \in F.\ \forall \rho \in (0, r].\ R(\operatorname{sgn} \circ f) \le \hat{R}_{S,\rho}(f) + \frac{4}{\rho} R_m(F) + \sqrt{\frac{\log \log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \Big] \ge 1 - \delta.$$

and

$$P^m\Big[\forall f \in F\ \forall \rho \in (0, r].\ R(\operatorname{sgn} \circ f) \le \hat{R}_{S,\rho}(f) + \frac{4}{\rho} \hat{R}_S(F) + \sqrt{\frac{\log \log_2 \frac{2r}{\rho}}{m}} + 3\sqrt{\frac{\log \frac{4}{\delta}}{2m}} \Big] \ge 1 - \delta.$$

We prove the first bound only. The proof of the second is similar.

Proof $\quad (\rho_k \in (0, r])_{k \ge 1}$ and $(\varepsilon_k \in (0, \infty))_{k \ge 1}$ ... two sequences.

=

Then, by Thm 5.8.,

$$P^{2m}\Big[\sup_{f \in F} R(\operatorname{sgn} \circ f) - \hat{R}_{S,\rho}(f) > \frac{2}{\rho_k} R_m(F) + \sqrt{\frac{\log \frac{1}{\delta_k}}{2m}} \Big] \le \delta_k$$

$\sqrt{\frac{\log \frac{1}{\delta_k}}{2m}} = \varepsilon_k$. $\qquad \delta_k = \exp(-2m\varepsilon_k^2)$

Choose $\varepsilon_k = \varepsilon + \sqrt{\frac{\log k}{m}}$

$$\mathbb{P}\left[\sup_{\substack{f\in\mathcal{F}\\ k\geq 1}} R(\mathrm{sgn}\circ f) - \widehat{R}_{S,\rho_k}(f) - \frac{2}{\rho_k}\mathcal{R}_m(\mathcal{F}) - \varepsilon_k > 0\right]$$

$$\leq \sum_{k\geq 1} \mathbb{P}\left[\sup_{f} R(\mathrm{sgn}\circ f) - \widehat{R}_{S,\rho_k}(f) - \frac{2}{\rho_k}\mathcal{R}_m(\mathcal{F}) - \varepsilon_k > 0\right]$$

$$\leq \sum_{k\geq 1} \exp(-2m\varepsilon_k^2) \qquad \leq \sum_{k\geq 1} \exp\left(-2m\left(\varepsilon^2 + \frac{\log k}{m}\right)\right)$$

$$= \exp(-2m\varepsilon^2) \sum_{k\geq 1} \frac{1}{k^2} = \exp(-2m\varepsilon^2) \times \frac{\pi^2}{6} = 2\exp(-2m\varepsilon^2).$$

Choose $\rho_k = r/2^k$. Let $\rho_0 = r$.

Then, for any $\rho \in (0, r]$, $\exists k \geq 1$ s.t. $\rho \in (\rho_k, \rho_{k-1}]$.

Note $\frac{1}{\rho_k} = \frac{2}{\rho_{k-1}} \leq \frac{2}{\rho}$ and so

$$\sqrt{\log k} = \sqrt{\log \log_2(r/\rho_k)} \leq \sqrt{\log \log_2 2r/\rho}$$

Also,

$$\widehat{R}_{S,\rho_k}(f) \leq \widehat{R}_{S,\rho}(f) \quad \text{for all } f \in \mathcal{F}.$$

As a result,

$$\sup_{\substack{f\in\mathcal{F}\\ \rho\in(0,r]}} R(\mathrm{sgn}\circ f) - \widehat{R}_{S,\rho}(f) - \frac{4}{\rho}\mathcal{R}_m(\mathcal{F}) - \sqrt{\frac{\log\log_2 2r/\rho}{m}} - \varepsilon$$

$$\leq$$

$$\sup_{\substack{f\in\mathcal{F}\\ k\geq 1}} R(\mathrm{sgn}\circ f) - \widehat{R}_{S,\rho_k}(f) - \frac{2}{\rho_k}\mathcal{R}_m(\mathcal{F}) - \sqrt{\frac{\log k}{m}} - \varepsilon$$

$= -\varepsilon_k$ from above.

Thus,

$$\mathbb{P}\left[\sup^{m} \cdots > 0\right] \leq \mathbb{P}\left[\sup^{m} \cdots > 0\right] \leq 2\exp(-2m\varepsilon^2).$$

$= \delta$ $\qquad = \sqrt{\frac{\log 2/\delta}{2m}}$ □

Thm 5.10 $S_x \in \{x \in \mathcal{X} \mid \|x\| \le r\}^m$

$\mathcal{F} = \{x \mapsto \langle w, x\rangle \mid \|w\| \le \Lambda\}$.

$\Rightarrow$

$$\widehat{R}_{S_x}(\mathcal{F}) \le \sqrt{\frac{r^2\Lambda^2}{m}}$$

.... bound on empirical Rademacher complexity.

Note that the thm implies that if $\mathbb{P}_{(x,y)\sim D}[\|x\| \le r]$, then $\widehat{R}_m(\mathcal{F}) \le \sqrt{\frac{r^2\Lambda^2}{m}}$ ..... bound on Rademacher complexity.

Proof. Let $S_x = (x_1, \dots, x_m)$.

$$\widehat{R}_{S_x}(\mathcal{F}) = \frac{1}{m} \mathbb{E}_{\vec{\sigma} \sim \text{unif}\{-1,1\}^m} \left[\sup_{\substack{w \\ \|w\| \le \Lambda}} \sum_{i=1}^m \sigma_i \langle w, x_i\rangle \right]$$

$$= \frac{1}{m} \mathbb{E}_{\vec{\sigma}} \left[\sup_{\substack{w \\ \|w\| \le \Lambda}} \left\langle w, \sum_{i=1}^m \sigma_i x_i \right\rangle \right]$$

$$\le \frac{1}{m} \mathbb{E}_{\vec{\sigma}} \left[\sup_{\substack{w \\ \|w\| \le \Lambda}} \|w\| \left\| \sum_{i=1}^m \sigma_i x_i \right\| \right] = \frac{\Lambda}{m} \mathbb{E}_{\vec{\sigma}} \left[\left\| \sum_{i=1}^m \sigma_i x_i \right\| \right]$$

Cauchy-Schwartz inequality.

$$\le \frac{\Lambda}{m} \mathbb{E}_{\vec{\sigma}} \left[\left\| \sum_{i=1}^m \sigma_i x_i \right\|^2 \right]^{\frac{1}{2}}$$

Jensen's inequality

$$\le \frac{\Lambda}{m} \mathbb{E}_{\vec{\sigma}} \left[\sum_{i=1}^m \sigma_i^2 \|x_i\|^2 + \sum_{\substack{i,j=1 \\ i \ne j}}^m \sigma_i \sigma_j \langle x_i, x_j\rangle \right]^{\frac{1}{2}}$$

$$= \frac{\Lambda}{m} \left(\sum_{i=1}^m \|x_i\|^2 \right)^{\frac{1}{2}} = \sqrt{\frac{r^2\Lambda^2}{m}}$$

□

We now bring Thm 5.9 and Thm 5.10 together.

Corollary $F = \{ x \mapsto \langle w, x \rangle \mid \|w\| \le \Lambda \}$.
$D \in Pr(X \times Y)$ s.t. $\mathbb{P}_{(x,y)\sim D}[\|x\| \le r] = 1$.

$m \in \mathbb{N}$. $\delta > 0$. $r' > 0$.

$\Rightarrow$.

$$\mathbb{P}_{S\sim D^m}\Big[\forall f \in F \; \forall \rho \in (0, r'] .\; R(\text{sgn} \circ f) \le \hat{R}_{S,\rho}(f) + \frac{4}{\rho}\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\log\log_2 \frac{2r'}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\Big] \ge 1-\delta.$$

※ The upper bound doesn't say anything about the dimension of the input. But the bound r on the size of the input x plays an important role here.

(5) From the above corollary, we can derive the upper bound mentioned when we started our discussion on margin theory. We just need to replace $\hat{R}_{S,\rho}(f)$ by its upper bound:

$$\hat{R}_{S,\rho}(f) \le \frac{1}{m}\sum_{i=1}^{m} \max\left(0, 1 - y_i \left\langle \frac{w}{\rho}, x_i \right\rangle\right)$$

where $f(x) = \langle w, x \rangle$.