E0 370 Statistical Learning Theory **Lecture 6 (Aug 30, 2011)**

Margin Analysis

Lecturer: Shivani Agarwal *Scribe: Narasimhan R S*

1 Introduction

In the last few lectures we have seen how to obtain high confidence bounds on the generalization error of functions learned from function classes of limited capacity, measured in terms of the growth function and VC-dimension for binary-valued function classes in the case of binary classification, and in terms of the covering numbers, pseudo-dimension, and fat-shattering dimension for real-valued function classes in the case of regression. Table 1 summarizes the nature of results we have obtained.

In this lecture, we return to binary classification, but focus on function classes $\mathcal{H}$ of the form $\mathcal{H} = \text{sign}(\mathcal{F})$ for some real-valued function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. If an algorithm learns a function $h_S : \mathcal{X} \to \{-1, 1\}$ of the form $h_S(x) = \text{sign}(f_S(x))$ for some $f_S \in \mathcal{F}$, can we say more about its generalization error than for a general binary-valued function $h_S : \mathcal{X} \to \{-1, 1\}$? Note that the SVM algorithm falls in this category, while the basic formulations of decision tree and nearest neighbour classification algorithms do not.[1]

Let us first see what we can say about the generalization error of $h_S = \text{sign}(f_S)$ using the techniques we have learned so far.

2 Bounding Generalization Error of $\text{sign}(f_S)$ via $\text{VCdim}(\text{sign}(\mathcal{F}))$

One approach to bounding the generalization error of $h_S \in \text{sign}(\mathcal{F})$ is to consider the VC-dimension of the binary-valued function class $\text{sign}(\mathcal{F})$. If $\text{VCdim}(\text{sign}(\mathcal{F}))$ is finite, then this gives that for any $0 < \delta \leq 1$, with probability at least $1 - \delta$ (over the draw of $S \sim D^m$),

$$\text{er}_D^{0\text{-}1}[h_S] \;\leq\; \text{er}_S^{0\text{-}1}[h_S] + c\sqrt{\frac{\text{VCdim}(\text{sign}(\mathcal{F}))\ln m + \ln\left(\frac{1}{\delta}\right)}{m}} \tag{1}$$

where $c > 0$ is some fixed constant.

However this approach is limited. Consider for example the two classifiers shown in Figure 1. Both are selected from the class of linear classifiers in $\mathbb{R}^2$: $\mathcal{H} = \text{sign}(\mathcal{F})$, where $\mathcal{F} = \{f : \mathbb{R}^2 \to \mathbb{R} \mid f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \text{ for some } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$. We saw earlier that the VC-dimension of this class is 3. The empirical error of h_1 on the training sample is $\text{er}_S[h_1] = 1/m$; for h_2, this is $\text{er}_S[h_2] = 0$. So the above result would give a better bound on the generalization error of h_2 than that of h_1, even though many of us would intuitively expect h_1 to perform better on new examples than h_2.

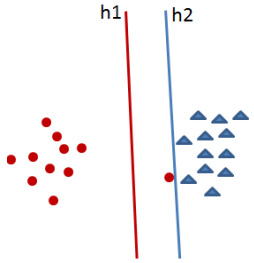

Figure 1: Two linear classifiers.

3 Bounding Generalization Error of $\text{sign}(f_S)$ via Bounds for f_S

As noted above, bounding the generalization error of $h_S = \text{sign}(f_S)$ using only the binary values output by $\text{sign}(f_S)$ has its limitations. Therefore we would like to take into account the real-valued predictions made

[1] There are variants of decision tree and nearest neighbour algorithms that take into account class probabilities or distance-based weights and could be described in the above form, but the basic formulations are not naturally described in this way.

Table 1: Summary of generalization error bounds we have seen so far.

Learning problem	**Function class**	**Loss function**	**Capacity measures**	**Bound on generalization error**
Binary classification	$\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$	$\ell_{0\text{-}1}$	$\Pi_{\mathcal{H}}(m)$ $\mathrm{VCdim}(\mathcal{H})$	w.p. $\geq 1-\delta$ (over $S \sim D^m$): $\mathrm{er}_D^{0\text{-}1}[h_S] \leq \mathrm{er}_S^{0\text{-}1}[h_S] + c\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) \ln m + \ln\left(\frac{1}{\delta}\right)}{m}}$
Regression	$\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$	ℓ : $0 \leq \ell \leq B$ L-Lipschitz	$\mathcal{N}_1(\epsilon, \mathcal{F}, m)$ $\mathrm{Pdim}(\mathcal{F})$ $\mathrm{fat}_{\mathcal{F}}(\epsilon)$	w.p. $\geq 1-\delta$ (over $S \sim D^m$): $\mathrm{er}_D^{\ell}[f_S] \leq \mathrm{er}_S^{\ell}[f_S] + \epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, B, L)$ (Here $\epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, B, L)$ is the smallest value of ϵ for which the bound obtained in the last lecture is smaller than δ; no closed-form expression in general.)
Binary classification using real-valued functions	$\mathcal{H} = \mathrm{sign}(\mathcal{F})$ $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$	$\ell_{0\text{-}1}$	??	??

by f_S. To this end, and to simplify later statements, let us first extend the $\ell_{0\text{-}1}$ loss to real-valued predictions as follows: define $\ell_{0\text{-}1} : \{-1,1\} \times \mathbb{R} \to \{0,1\}$ as

$$\ell_{0\text{-}1}(y, \hat{y}) \;=\; \mathbf{1}(\mathrm{sign}(\hat{y}) \neq y)\,. \tag{2}$$

This gives for example $\mathrm{er}_D^{0\text{-}1}[f_S] = \mathbf{P}_{(x,y)\sim D}(\mathrm{sign}(f_S(x)) \neq y)$ $(= \mathrm{er}_D^{0\text{-}1}[h_S]$ for $h_S = \mathrm{sign}(f_S))$.

We can then consider obtaining bounds on the generalization error $\mathrm{er}_D^{0\text{-}1}[f_S]$ using results derived in the last lecture in terms of the covering numbers or fat-shattering dimension of the real-valued function class $\mathcal{F}$. Unfortunately, we cannot use these results to bound $\mathrm{er}_D^{0\text{-}1}[f_S]$ directly in terms of $\mathrm{er}_S^{0\text{-}1}[f_S]$ using covering numbers of $\mathcal{F}$, since $\ell_{0\text{-}1}$ is not Lipschitz.[2] However, for any bounded, Lipschitz loss function $\ell \geq \ell_{0\text{-}1}$, we have $\mathrm{er}_D^{0\text{-}1}[f_S] \leq \mathrm{er}_D^{\ell}[f_S]$, which in turn can be bounded in terms of $\mathrm{er}_S^{\ell}[f_S]$ using covering numbers of $\mathcal{F}$.

Specifically, let $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$ for some $\widehat{\mathcal{Y}} \subseteq \mathbb{R}$, and consider any loss function $\ell : \{-1,1\} \times \widehat{\mathcal{Y}} \to [0,\infty)$ for which there are constants $B, L > 0$ such that for all $y \in \{-1,1\}$ and all $\hat{y}, \hat{y}_1, \hat{y}_2 \in \widehat{\mathcal{Y}}$:

(i) $0 \leq \ell(y, \hat{y}) \leq B$;

(ii) $|\ell(y, \hat{y}_1) - \ell(y, \hat{y}_2)| \leq L\,|\hat{y}_1 - \hat{y}_2|$; and

(iii) $\ell_{0\text{-}1}(y, \hat{y}) \leq \ell(y, \hat{y})$.

Then with probability at least $1-\delta$ (over $S \sim D^m$),

$$\mathrm{er}_D^{0\text{-}1}[f_S] \;\leq\; \mathrm{er}_D^{\ell}[f_S] \;\leq\; \mathrm{er}_S^{\ell}[f_S] + \epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, B, L)\,, \tag{3}$$

where $\epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, B, L)$ is as described in Table 1.

Below we shall give several concrete examples of loss functions satisfying the above conditions. Before we do so, it will be useful to introduce an alternative description of loss functions $\ell : \{-1,1\} \times \widehat{\mathcal{Y}} \to [0,\infty)$ that satisfy the following 'symmetry' condition (assuming $\widehat{\mathcal{Y}} \subseteq \mathbb{R}$ is closed under negation):

$$\ell(-1, \hat{y}) = \ell(1, -\hat{y}) \quad \forall \hat{y} \in \widehat{\mathcal{Y}}\,. \tag{4}$$

[2] We could bound $\mathrm{er}_D^{0\text{-}1}[f_S]$ in terms of $\mathrm{er}_S^{0\text{-}1}[f_S]$ using covering numbers of the *loss function* class $(\ell_{0\text{-}1})_{\mathcal{F}}$, but this simply takes us back to the growth function/VC-dimension of $\mathrm{sign}(\mathcal{F})$.

For any such loss function ℓ, we can write

$$\ell(y, \hat{y}) = \ell(1, y\hat{y}) \quad \forall y \in \{-1, 1\}, \hat{y} \in \widehat{\mathcal{Y}}\,, \tag{5}$$

and therefore the loss function can equivalently be described using a single-variable function $\phi : \widehat{\mathcal{Y}} \to [0, \infty)$ given by $\phi(u) = \ell(1, u)$:

$$\ell(y, \hat{y}) = \phi(y\hat{y}) \quad \forall y \in \{-1, 1\}, \hat{y} \in \widehat{\mathcal{Y}}\,. \tag{6}$$

Such loss functions are often called *margin-based* loss functions, for reasons that will be clear later. As an example, the 0-1 loss above can be viewed as a margin-based loss with $\phi(u) = \mathbf{1}(u \leq 0)$. On the other hand, an 'asymmetric' misclassification loss $\ell : \{-1, 1\} \times \{-1, 1\} \to \{0, a, b\}$ where $a \neq b$, $\ell(y, y) = 0$, $\ell(1, -1) = a$, and $\ell(-1, 1) = b$, cannot be written as a margin-based loss. Margin-based losses are easily visualized through a plot of the function $\phi(\cdot)$; they also satisfy the following properties, which can be easily verified:

(M1) $0 \leq \ell \leq B \iff 0 \leq \phi \leq B$

(M2) ℓ L-Lipschitz in 2nd argument $\iff$ ϕ L-Lipschitz

(M3) $\ell \geq \ell_{0\text{-}1} \iff \phi(u) \geq \mathbf{1}(u \leq 0)\ \forall u \in \widehat{\mathcal{Y}}$

(M4) ℓ convex in 2nd argument $\iff$ ϕ convex.

We now give several examples of (margin-based) losses satisfying conditions (i)-(iii) above, which can be used to derive bounds on the 0-1 generalization error $\mathrm{er}_D^{0\text{-}1}[f_S]$ in terms of the covering numbers/fat-shattering dimension of $\mathcal{F}$.

Example 1 (Squared loss)**.** Let $\widehat{\mathcal{Y}} = [-1, 1]$, and

$$\ell_{\mathrm{sq}}(y, \hat{y}) = (y - \hat{y})^2 = (1 - y\hat{y})^2 \quad \forall y \in \{-1, 1\}, \hat{y} \in \widehat{\mathcal{Y}}\,.$$

Then clearly, ℓ_{sq} satisfies conditions (i)-(iii) above with $B = 4$ and $L = 4$ (over $\widehat{\mathcal{Y}}$ as above; see Figure 2). Therefore we can bound the generalization error $\mathrm{er}_D^{0\text{-}1}[f_S]$ as follows: with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\mathrm{er}_D^{0\text{-}1}[f_S] \;\leq\; \mathrm{er}_S^{\mathrm{sq}}[f_S] + \epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, 4, 4)\,. \tag{7}$$

Example 2 (Absolute loss)**.** Let $\widehat{\mathcal{Y}} = [-1, 1]$, and

$$\ell_{\mathrm{abs}}(y, \hat{y}) = |y - \hat{y}| = |1 - y\hat{y}| \quad \forall y \in \{-1, 1\}, \hat{y} \in \widehat{\mathcal{Y}}\,.$$

Then clearly, ℓ_{abs} satisfies properties (i)-(iii) above with $B = 2$ and $L = 1$ (see Figure 2). Therefore we have with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\mathrm{er}_D^{0\text{-}1}[f_S] \;\leq\; \mathrm{er}_S^{\mathrm{abs}}[f_S] + \epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, 2, 1)\,. \tag{8}$$

Example 3 (Hinge loss)**.** Let $\widehat{\mathcal{Y}} = [-1, 1]$, $0 < \gamma \leq 1$, and

$$\ell_{\mathrm{hinge}(\gamma)}(y, \hat{y}) = \frac{1}{\gamma}(\gamma - y\hat{y})_+ \quad \forall y \in \{-1, 1\}, \hat{y} \in \widehat{\mathcal{Y}}\,.$$

(Recall $z_+ = \max(0, z)$.) Clearly, $\ell_{\mathrm{hinge}(\gamma)}$ satisfies properties (i)-(iii) above with $B = (\gamma + 1)/\gamma$ and $L = 1/\gamma$ (see Figure 2). Therefore we have with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\mathrm{er}_D^{0\text{-}1}[f_S] \;\leq\; \mathrm{er}_S^{\mathrm{hinge}(\gamma)}[f_S] + \epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, (\gamma + 1)/\gamma, 1/\gamma)\,. \tag{9}$$

Example 4 (Ramp loss)**.** Let $\widehat{\mathcal{Y}} = [-1, 1]$, $0 < \gamma \leq 1$, and

$$\ell_{\mathrm{ramp}(\gamma)}(y, \hat{y}) = \begin{cases} 1 & \text{if } y\hat{y} \leq 0 \\ \ell_{\mathrm{hinge}(\gamma)}(y, \hat{y}) & \text{if } y\hat{y} > 0 \end{cases} \quad \forall y \in \{-1, 1\}, \hat{y} \in \widehat{\mathcal{Y}}\,.$$

Clearly, $\ell_{\mathrm{ramp}(\gamma)}$ satisfies properties (i)-(iii) above with $B = 1$ and $L = 1/\gamma$ (see Figure 2). Therefore we have with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\mathrm{er}_D^{0\text{-}1}[f_S] \;\leq\; \mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f_S] + \epsilon^*(\mathrm{fat}_{\mathcal{F}}, m, \delta, 1, 1/\gamma)\,. \tag{10}$$

Note that this bound is uniformly better than that using the hinge loss in Example 3, since $\mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f_S] \leq \mathrm{er}_S^{\mathrm{hinge}(\gamma)}[f_S]$ and the bound B here is also smaller.

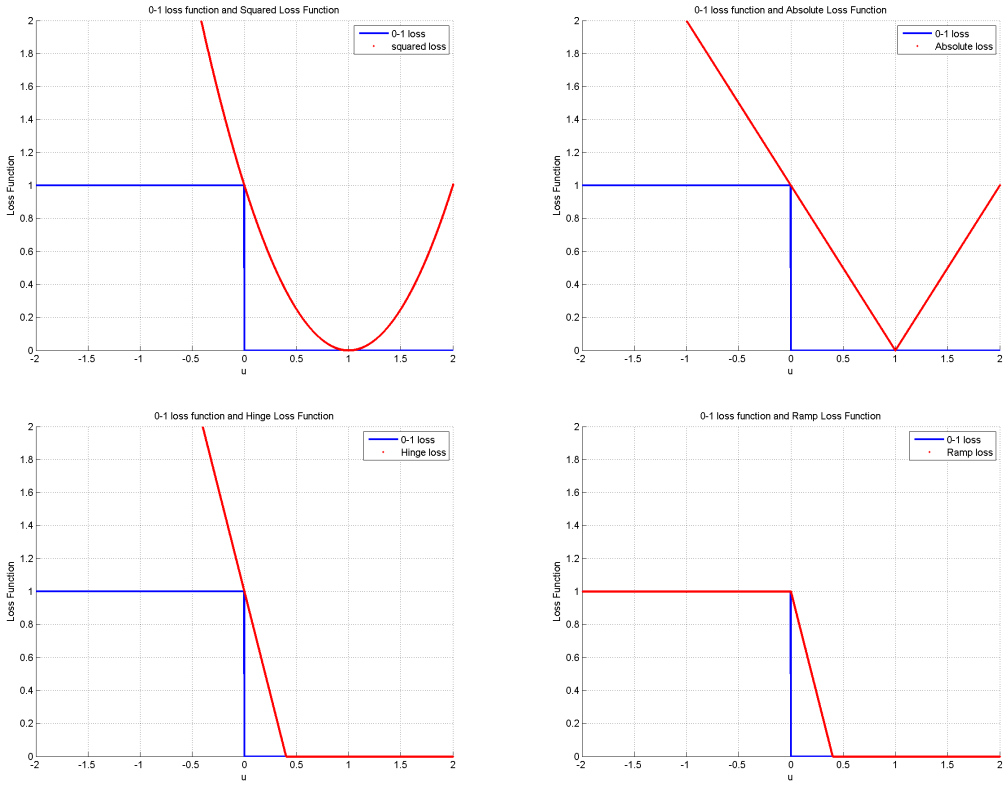

Figure 2: Loss functions used in Examples 1–4. **Top left:** Squared loss ℓ_{sq}. **Top right:** Absolute loss ℓ_{abs}. **Bottom left:** Hinge loss $\ell_{\text{hinge}(\gamma)}$ for $\gamma = 0.4$. **Bottom left:** Ramp loss $\ell_{\text{ramp}(\gamma)}$ for $\gamma = 0.4$. The zero-one loss $\ell_{\text{0-1}}$ is shown in each case for comparison. All are margin-based loss functions and are plotted in terms of the corresponding functions ϕ (see Section 3).

The above examples illustrate how the generalization error $\text{er}_D^{\text{0-1}}[f_S]$ can be bounded in terms of the covering numbers/fat-shattering dimension of $\mathcal{F}$ via generalization error bounds for f_S using a variety of different loss functions typically applied to real-valued learning problems. However there are several limitations with this approach as well: (1) ϵ^* is often not available in a closed form; (2) bounds in terms of $\text{fat}_{\mathcal{F}}$ can often be quite loose; and (3) the bounds do not provide intuition about the problem, in particular they do not guide algorithm design. Below we discuss a different approach for bounding the 0-1 generalization error $\text{er}_D^{\text{0-1}}[f_S]$ of a real-valued classifier $\text{sign}(f_S)$ which avoids these difficulties.

4 Bounding Generalization Error of $\text{sign}(f_S)$ via Margin Analysis

Consider again the example in Figure 1. Can we explain our intuition that h_1 should perform better than h_2? One possible explanation for this is that while h_2 classifies all training examples correctly, h_1 has a better 'margin' on most of the examples.

Formally, define the *margin* of a real-valued classifier $h = \text{sign}(f)$ (or simply of f) on an example (x, y) as

$$\text{margin}(f, (x, y)) = yf(x)\,.$$

If the margin $yf(x)$ is positive, then $\text{sign}(f(x)) = y$, i.e. f makes the correct prediction on (x, y); if the margin $yf(x)$ is negative, then $\text{sign}(f(x)) \neq y$, i.e. f makes the wrong prediction on (x, y). Moreover, a larger positive value of the margin $yf(x)$ can be viewed as a more 'confident' prediction. In estimating the generalization error of $h_S = \text{sign}(f_S)$, we may then want to take into account not only the number of

examples in S that are classified correctly by f_S, but the number of examples that are classified correctly with a large margin.

To formalize this idea, let $\gamma > 0$, and define the *γ-margin loss* $\ell_{\text{margin}(\gamma)} : \{-1, 1\} \times \mathbb{R} \to \{0, 1\}$ as

$$\ell_{\text{margin}(\gamma)}(y, \hat{y}) \;=\; \mathbf{1}(y\hat{y} \leq \gamma)\,. \tag{11}$$

(In the terminology of the previous section, this is clearly a margin-based loss.) The empirical γ-margin error of f_S then counts the fraction of training examples that have margin less than γ:

$$\text{er}_S^{\text{margin}(\gamma)}[f_S] \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(y_i f_S(x_i) \leq \gamma)\,.$$

We cannot bound $\text{er}_D^{0\text{-}1}[f_S]$ in terms of $\text{er}_S^{\text{margin}(\gamma)}[f_S]$ using the technique of the previous section since $\ell_{\text{margin}(\gamma)}$ is not Lipschitz. Instead, one can show directly the following one-sided 'uniform convergence' result (strictly speaking, this is not uniform 'convergence', but rather a uniform 'bound'):

Theorem 4.1 (Bartlett, 1998)**.** Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. Let $\gamma > 0$ and let D be any distribution on $\mathcal{X} \times \{-1, 1\}$. Then for any $\epsilon > 0$,

$$\mathbf{P}_{S\sim D^m}\left(\exists f \in \mathcal{F} : \text{er}_D^{0\text{-}1}[f] \geq \text{er}_S^{\text{margin}(\gamma)}[f] + \epsilon\right) \;\leq\; 2\mathcal{N}_\infty(\gamma/2, \mathcal{F}, 2m)\, e^{-m\epsilon^2/8}\,.$$

The proof is similar in structure to the previous uniform convergence proofs we have seen, with the same four main steps. Again, the main difference is in Step 3 (reduction to a finite class). Details can be found in [2, 1]; you are also encouraged to try the proof yourself!

It is worth noting that, as in the case of previous uniform convergence results we have seen, the proof actually yields a more refined bound in terms of the *expected covering numbers* $\mathbf{E}_{\mathbf{x}_1^{2m}\sim\mu^{2m}}\left[\mathcal{N}(\gamma/2, \mathcal{F}_{|\mathbf{x}_1^{2m}}, d_\infty)\right]$ (where μ is the marginal of D on $\mathcal{X}$) rather than the uniform covering numbers $\mathcal{N}_\infty(\gamma/2, \mathcal{F}, 2m) = \max_{\mathbf{x}_1^{2m}} \mathcal{N}(\gamma/2, \mathcal{F}_{|\mathbf{x}_1^{2m}}, d_\infty)$. However the expected covering numbers are typically difficult to estimate and so one generally falls back on the upper bound in terms of uniform covering numbers.[3]

The above bound makes use of the d_∞ covering numbers of $\mathcal{F}$ rather than the d_1 covering numbers that were used in the results of the last lecture. A more important difference is that the covering number is measured at a scale of $\gamma/2$, and is independent of ϵ; this means that when converting to a confidence bound, one gets a closed form expression. In particular, we have that with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\text{er}_D^{0\text{-}1}[f_S] \;\leq\; \text{er}_S^{\text{margin}(\gamma)}[f_S] + \sqrt{\frac{8\ln\left(\mathcal{N}_\infty(\gamma/2, \mathcal{F}, 2m)\right) + \ln\left(\frac{2}{\delta}\right)}{m}}\,. \tag{12}$$

As before, the covering number $\mathcal{N}_\infty(\gamma/2, \mathcal{F}, 2m)$ needs to be sub-exponential in m for the above result to be meaningful. For function classes $\mathcal{F}$ with finite pseudo-dimension or finite fat-shattering dimension, it is possible to obtain polynomial bounds on the d_∞ covering numbers of $\mathcal{F}$ in terms of these quantities, similar to the bounds we discussed for the d_1 covering numbers in the previous lecture; see [1] for details. For certain function classes of interest, the d_∞ covering numbers can also be bounded directly. For example, we state below a bound on the covering numbers of classes of linear functions over bounded subsets of $\mathbb{R}^n$ with bounded-norm weight vectors; details can be found in [3]:[4]

Theorem 4.2 (Zhang, 2002)**.** Let $R, W > 0$. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq R\}$ and $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} \mid f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} \text{ for some } \mathbf{w} \in \mathbb{R}^n, \|\mathbf{w}\|_2 \leq W\}$. Then $\forall \gamma > 0$,

$$\mathcal{N}_\infty(\gamma, \mathcal{F}, m) \;\leq\; \frac{36R^2W^2}{\gamma^2} \cdot \ln\left(2\left\lceil\frac{4RW}{\gamma} + 2\right\rceil m + 1\right)\,.$$

[3] Note that the expected covering numbers are sometimes referred to as 'random' covering numbers in the literature; this is a misnomer as there is no randomness left once you take the expectation.

[4] Note that since $d_1 \leq d_\infty$, any upper bounds on the d_∞ covering numbers of a function class also imply upper bounds on its d_1 covering numbers.

Thus for linear classifiers $h_S = \text{sign}(f_S)$ where f_S is learned from the class $\mathcal{F}$ in the above theorem, we have the following generalization error bound: for any $\gamma > 0$ and $0 < \delta \leq 1$, with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\text{er}_D^{\text{0-1}}[f_S] \;\leq\; \text{er}_S^{\text{margin}(\gamma)}[f_S] + c\sqrt{\frac{\frac{R^2W^2}{\gamma^2}\ln m + \ln\left(\frac{1}{\delta}\right)}{m}}\,. \tag{13}$$

Note that if we can make the first term in the bound small for a larger value of γ (i.e. achieve a good separation on the data with a larger margin γ), then the complexity term on the right becomes smaller, implying a better generalization error bound. This suggests that algorithms achieving a large margin on the training sample may lead to good generalization performance. However it is important to note a caveat here: the above bound holds only when the margin parameter $\gamma > 0$ is fixed in advance, *before* seeing the data; it cannot be chosen after seeing the sample S. In a later lecture, we will see how to extend this to a uniform margin bound that holds uniformly over all γ in some bounded range, and will therefore apply even when γ is chosen based on the data.

In closing, we note that the margin idea can be extended to learning problems other than binary classification; see for example [3].

Exercise. Let $\mathcal{X} \subseteq \mathbb{R}^n$, and let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathcal{X} \times \{-1, 1\})^m$. Say you learn from S a linear classifier $h_S = \text{sign}(\mathbf{w}_S \cdot \mathbf{x})$ using the SVM algorithm with some value of C:

$$\mathbf{w}_S \;=\; \arg\min_{\mathbf{w}\in\mathbb{R}^n}\left(\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\ell_{hinge}(y_i, \mathbf{w}\cdot\mathbf{x}_i)\right).$$

Can you bound $\|\mathbf{w}_S\|_2$? (Hint: consider what you get with $\mathbf{w} = \mathbf{0}$.)

5 Next Lecture

In the next lecture, we will derive generalization error bounds for both classification and regression using Rademacher averages, which provide a distribution-dependent measure of the complexity of a function class whose empirical (data-dependent) analogues can often be estimated reliably from data, and can lead to tighter bounds than what can be obtained with distribution-independent complexity measures such as the growth function/VC-dimension or covering numbers/fat-shattering dimension.

References

[1] Martin Anthony and Peter L. Bartlett. *Learning in Neural Networks: Theoretical Foundations.* Cambridge University Press, 1999.

[2] Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[3] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.