Machine Learning Theory (CSC 431/531) - Lecture 14

Nishant Mehta

1 Computational hardness of agnostically learning halfspaces

In the problem of *efficiently* agnostically learning halfspaces over $\mathbb{R}^d$, the goal is to learn a hypothesis (not necessarily a linear separator) from a training sample which, with probability at least $1-\delta$, obtains risk at most ε in excess of the best linear separator using runtime polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, and the dimension d. If the learning algorithm is restricted to be a proper learner (which may only output a halfspace), the problem is known to be NP-hard; moreover, the problem is NP-hard even to approximate: obtaining risk $\varepsilon + \alpha \cdot R(f^*)$ for some constant α is NP-hard.[1] One wonders if the situation might change if we allow improper learners, but, at least under prevailing complexity assumptions on the hardness of various problems, the problem continues to be computationally hard.

On the other hand, in the realizable case (where there is a linear separator that perfectly classifies the data), one can use linear programming to efficiently identify an empirical risk minimizer. Using our risk bounds based on VC dimension (which is $d+1$ in this case) and sufficiently many samples (polynomial in the same 3 quantities as above), we can be assured that any such minimizer has risk at most ε with high probability. While this may seem like progress, it turns out that we can do much better from the statistical perspective when the data is separable by some margin γ.

2 Margin Bounds

When learning linear classifiers in the mistake bound model, we saw that data that is linearly separable with a large margin γ can be learned with a mistake bound whose scaling with γ is $\frac{1}{\gamma^2}$. In short, a larger margin guarantees that Perceptron makes fewer mistakes. An algorithm obtaining a small mistake bound could in turn be converted into an algorithm for the statistical learning setting which obtains a hypothesis with correspondingly low risk. Let us turn now to the statistical learning setting. Suppose that we have a training set which is linearly separable with some margin γ. It is not hard to see that there are infinitely many linear separators that obtain zero training error, and thus there are infinitely many empirical risk minimizers. However, as we will see shortly, not all empirical risk minimizers are created equal: those linear separators that achieve large margin admit much smaller risk bounds as a result.

We will begin by deriving a generalization error bound that is small when an algorithm has learned a linear separator that achieves large margin on most of the examples. Concretely, considering γ as a tuning parameter, the bound will be best when γ is large and when, for all but a small number of examples, the margin achieved by the linear separator is at least γ. Conversely, the bound will degrade either when there are few examples for which we achieve margin γ or when γ gets smaller.

[1] See (Daniely, 2015) for more details, including hardness results for improper learning.

Before continuing, it is important to formally define the margin. The attentive reader may remember that we already saw a definition of the margin near the beginning of the course (when we covered the Perceptron algorithm); however, that definition assumed that the bias term b is equal to zero. The definition below will include the bias term. That said, in deriving our generalization bound, we will again ignore the bias term to keep the derivation simple and focused on the core ideas; for more information, see the remark at the end of this section.

The geometric margin. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$, and take $\mathcal{H}$ to be the set of nonhomogeneous linear separators

$$\left\{x \mapsto \operatorname{sgn}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

Each classifier in $\mathcal{H}$ can be identified with a separating hyperplane

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

Let us use (w, b) to refer to the corresponding hyperplane. A useful observation that we will often use is that the hyperplane (w, b) is invariant to scaling w and b by the same positive constant. That is, for any $\alpha > 0$, replacing w and b by αw and αb gives the same set in (1).

The (geometric) margin of a hyperplane (w, b) is defined as the minimum distance of the hyperplane to a correctly classified point in the training sample. For a given example x_j, the distance from x_j to the hyperplane (w, b) is

$$\frac{|\langle w, x_j\rangle + b|}{\|w\|}. \tag{2}$$

Assuming the example is correctly classified, this distance is the same as the margin the hyperplane obtains on the example, which we write as

$$\frac{y(\langle w, x_j\rangle + b)}{\|w\|}.$$

From these formulas, we can easily see that this distance (and margin) is invariant to scaling w and b by the same positive constant; this is not surprising, given that the hyperplane itself is invariant to such scaling.

We are almost ready to see how to derive risk bounds that benefit from the learned hypothesis obtaining large margin on most of the data. But first, recall the story of our Rademacher complexity-based risk bounds in the case of classification with VC classes.

We began with a risk bound based on empirical Rademacher complexity:

For any (learned) classifier $\hat{f}$, with probability at least $1 - \delta$ over the training sample,

$$\mathsf{E}\left[\mathbf{1}\left[Y \neq \hat{f}(X)\right]\right] \leq \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[Y_j \neq \hat{f}(X_j)\right] + 2\widehat{\mathcal{R}}_n(\ell_{0-1} \circ \mathcal{H}) + O\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}}\right). \tag{3}$$

We then used the very special relationship

$$\widehat{\mathcal{R}}_n(\ell_{0-1} \circ \mathcal{H}) = \frac{1}{2}\widehat{\mathcal{R}}_n(\mathcal{H}), \tag{4}$$

which was useful because we could then bound $\widehat{\mathcal{R}}_n(\mathcal{H})$ via the growth function (which in turn is bounded in terms of the VC dimension), yielding the final bound

$$\mathsf{E}\left[\mathbf{1}\left[Y \neq \hat{f}(X)\right]\right] \leq \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}\left[Y_j \neq \hat{f}(X_j)\right] + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{n}}\right) + O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

Unfortunately, when the dimension d is large, this bound scales like $O(\sqrt{d/n})$. But this is really the best that we can hope for in a worst-case scenario. To see one reason why the above approach cannot be used to get margin-dependent bounds, let us look at (3). In this expression, we see the appearance of the function class $\mathcal{H}$. Because each hypothesis $f \in \mathcal{H}$ computes the sign of $\langle w, x\rangle + b$, we lose all information about the magnitude of $\frac{\langle w,x\rangle+b}{\|w\|}$; this magnitude information is critical to keep around if we care about the margin. We therefore will proceed differently.

A first important step is to work with an analogue of $\mathcal{H}$ that avoids taking the sign, hence retaining magnitude information. To this end, for any hyperplane (w, b), let $f_{w,b}$ be the real-valued predictor defined as

$$f_{w,b}(x) = \frac{\langle w, x\rangle + b}{\|w\|}.$$

Next, we define the normalized class $\mathcal{F}_1$ as

$$\mathcal{F}_1 := \{f_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

The normalization by $\|w\|$ present in the definition of $f_{w,b}$ can equivalently be viewed as scaling w and b together so that w has unit norm; recall that such scaling does not change the hyperplane (and hence does not change the margin achieved on any example). Moreover, the normalization is convenient because the margin achieved by (w, b) on a correctly classified example (x, y) may now be expressed as $yf_{w,b}(x)$.

Formally, to work with $\mathcal{F}_1$ rather than $\mathcal{H}$, we need to extend the zero-one loss to accommodate real-valued predictions. To this end, from this point onwards, we define the zero-one loss as a mapping from $\{-1, +1\} \times \mathbb{R}$ to $\{0, 1\}$, defined as $\ell_{0-1}(y, \hat{y}) = \mathbf{1}[y\hat{y} \leq 0]$. Written this way, the zero-one loss is a member of a general family of losses known as *margin losses*.

Definition 1 (margin loss)**.** A loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is a margin loss if it can be expressed in the form $\ell(y, \hat{y}) = \Phi(y\hat{y})$ for some function $\Phi : \mathbb{R} \to \mathbb{R}$.

We can express the zero-one loss as a margin loss using the function

$$\Phi_{0-1}(t) = \mathbf{1}[t \leq 0].$$

We now proceed to derive a risk bound that improves with the margin using two key ingredients infused into a Rademacher complexity-based approach:

1. We will try to upper bound the risk under the zero-one loss by the risk under a carefully-selected Lipschitz loss whose Lipschitz constant is inversely proportional to the margin. If we are successful, we will avoid the VC dimension-based upper bound on the Rademacher complexity of a set of classifiers. Instead, we will deal with the Rademacher complexity of loss-composed real-valued predictors (where the loss is the aforementioned Lipschitz loss that we will design).

2. We will upper bound the Rademacher complexity of the set of these loss-composed real-valued predictors by using results from the last lecture. Namely, we will use the result that lets us peel off the loss function and replace it with a Lipschitz constant (which gives the dependence on the margin), after which it remains to bound the Rademacher complexity of a class of linear predictors (which we already did last lecture).

Now, consider some example (x, y) that is correctly classified by some $f_{w,b} \in \mathcal{F}_1$ with margin at least $\gamma > 0$. Then

$$y f_{w,b}(x) = \frac{y\left(\langle w, x\rangle + b\right)}{\|w\|} \geq \gamma. \tag{5}$$

The zero-one loss of $f_{w,b}$ on this example is clearly zero since $\mathbf{1}\left[y f_{w,b} \leq 0\right] = 0$. Moreover, even if we were to increase the threshold for correct classifications to just under γ, i.e., $\mathbf{1}\left[y f_{w,b} < \gamma\right]$, the loss is still zero. By making this change, we now are free to "charge" for errors by linearly interpolating between the threshold γ and the threshold 0. This linear interpolation gives rise to a particularly useful subclass of margin losses known as "ramp losses."

The γ-*ramp loss* is the margin loss defined via the function

$$\Phi_\gamma(t) = \begin{cases} 0 & \text{if } t \geq \gamma \\ 1 - \frac{t}{\gamma} & \text{if } 0 < t < \gamma \\ 1 & \text{if } t \leq 0. \end{cases}$$

Why is the γ-ramp loss useful in developing a margin bound? From (5), if $f_{w,b}$ classifies (x, y) with margin at least γ, then $\Phi_\gamma(y f_{w,b}(x)) = 0$, so there is a clear link between the γ-ramp loss and correctly classifying an example with margin γ. Moreover, as $t = y f_{w,b}(x)$ decreases from γ to zero, $\Phi_\gamma(t)$ increases at the rate of $\frac{1}{\gamma}$. Thus, Φ_γ is $\frac{1}{\gamma}$-Lipschitz, and, consequently, the γ-ramp loss is $\frac{1}{\gamma}$-Lipschitz in its second argument.

A useful observation is that the zero-one loss is upper bounded by the γ-ramp loss for any $\gamma > 0$. Consequently, we have for any (real-valued) hypothesis f that

$$\mathsf{E}\left[\Phi_{0-1}(Y f(X))\right] \leq \mathsf{E}\left[\Phi_\gamma(Y f(X))\right]. \tag{6}$$

This is incredibly useful, as we now can upper bound the risk under γ-ramp loss using our Rademacher complexity-style analysis, and whatever bound we obtain will also be an upper bound on the risk under zero-one loss

Everything is now in place to obtain a risk bound that depends on the margin. From the uniform convergence bound based on empirical Rademacher complexity (and using (6)), it holds with probability at least $1 - \delta$ that for all $f \in \mathcal{F}_1$,

$$\begin{aligned} \mathsf{E}\left[\Phi_{0-1}(Y f(X))\right] &\leq \mathsf{E}\left[\Phi_\gamma(Y f(X))\right] \\ &\leq \frac{1}{n}\sum_{j=1}^{n} \Phi_\gamma(Y_j f(X_j)) + 2\widehat{\mathcal{R}}_n(\Phi_\gamma \circ \mathcal{F}_1) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}} \\ &\leq \frac{1}{n}\sum_{j=1}^{n} \Phi_\gamma(Y_j f(X_j)) + \frac{2}{\gamma}\widehat{\mathcal{R}}_n(\mathcal{F}_1) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \end{aligned}$$

Next, just like I mentioned we would do, we adopt the simplifying assumption that $b = 0$. Since we are in the homogeneous case, from our upper bound on the Rademacher complexity of linear

prediction classes (stated in the last lecture), we have

$$\widehat{\mathcal{R}}_n(\mathcal{F}_1) \leq \frac{\max_{j\in[n]} \|X_j\|}{\sqrt{n}},$$

where we used the fact that $\|w\|_2 = 1$ for all $f_{w,b} \in \mathcal{F}_1$.

Thus, we have the following risk bound: with probability at least $1-\delta$, for all $f \in \mathcal{F}_1$,

$$\mathsf{E}\left[\mathbf{1}\left[Yf(x) \leq 0\right]\right] \leq \frac{1}{n}\sum_{j=1}^{n} \Phi_\gamma(Y_j f(X_j)) + \frac{2\max_{j\in[n]} \|X_j\|}{\gamma\sqrt{n}} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \tag{7}$$

Lastly, if we wish to make the bound more interpretable, we can use the fact that $\Phi_\gamma(t) \leq \mathbf{1}\left[t < \gamma\right]$, where we call the margin loss defined by threshold γ the *γ-margin error*.

Then we have, for any $\gamma > 0$, with probability at least $1-\delta$, for any $f \in \mathcal{F}_1$,

$$\mathsf{E}\left[\mathbf{1}\left[Yf(x) \leq 0\right]\right] \leq \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[Y_j f(X_j) < \gamma\right] + \frac{2\max_{j\in[n]} \|X_j\|}{\gamma\sqrt{n}} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}.$$

Consider either of the above bounds. The bound is valid for any choice of γ, as long as the choice is made before seeing the data. In practice, we would like to have a valid bound which holds for the particular, data-dependent choice of γ that minimizes the bound. We leave obtaining this bound as a simple exercise.

An analysis that does not ignore the bias term. Handling the case of general b (i.e., the nonhomogeneous case) is rarely discussed. A common, textbook approach to handle the nonhomogeneous case is to add an extra dummy dimension to each input x which always takes the value 1 (so that we also increase w by one dimension, and the last component of w now plays the role of b), but this transformation can have a drastic effect on the norm of w and hence on the margin. There is a proper treatment of the nonhomogeneous case, i.e., a result that does not introduce a dummy dimension nor do anything else that would change the margin. This satisfying result is Theorem 15 of Hanneke and Kontorovich (2019), whose proof does not involve all that much extra effort.

3 An algorithmic approach to minimize the margin bound

We have now worked out risk bounds that depend on the margin. Focusing on the first bound, (7), let us think about designing an algorithm that minimizes the bound; that is, using the data we have observed, we would like to select a pair (w, b) and a value for γ that minimizes the bound in (7). There are two challenges here:

- The loss function, the γ-ramp loss, has a parameter γ. So, different choices of γ lead to different losses. It would be convenient if we could somehow separate γ from the loss function. It turns out that we can achieve this separation; in the process, we will introduce a different regularization parameter, but this new parameter is more standard in machine learning and optimization.
- Ramp losses are non-convex. Therefore, even for a fixed value of γ, in general we will not be able to efficiently minimize the bound. Our solution to this problem is a simple application of an idea called *convex relaxation*. In a convex relaxation, we shift from a non-convex optimization problem to a convex optimization problem; in this case, we will do so by upper bounding the ramp loss by something called the *hinge loss*.

Regarding the first challenge, suppose that on some training example (x, y), we have $\Phi_\gamma(yf_{w,b}(x)) = 0$. By definition of Φ_γ, this is the same as

$$y \cdot \frac{\langle w, x\rangle + b}{\|w\|} \geq \gamma,$$

which may be rewritten as

$$y(\langle w, x\rangle + b) \geq \gamma\|w\|.$$

Now, recalling that we are free to set γ however we wish in the bound (7), suppose we set γ as $\frac{1}{\|w\|}$. Then the above becomes

$$y(\langle w, x\rangle + b) \geq 1.$$

Now, observe that the above inequality is equivalent to $\Phi_1(y(\langle w, x\rangle + b)) = 0$. Therefore, by making the choice $\gamma = \frac{1}{\|w\|}$, minimizing the bound (7) is the following problem:

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{n}\sum_{j=1}^{n} \Phi_1(Y_j(\langle w, X_j\rangle + b)) + \frac{2\max_{j\in[n]}\|X_j\|}{\sqrt{n}} \cdot \|w\|,$$

where we dropped the last term in (7) since it only depends on δ and n.

Now, since γ has been replaced by $\frac{1}{\|w\|}$, it would be convenient if we had control over the norm of w. This can be achieved by introducing a regularization parameter $\lambda > 0$ as follows

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{n}\sum_{j=1}^{n} \Phi_1(Y_j(\langle w, X_j\rangle + b)) + \lambda\|w\|;$$

note that we dropped the coefficient $\frac{2\max_{j\in[n]}\|X_j\|}{\sqrt{n}}$ since this can be handled by λ.

We make two last adjustments. First to control the size of $\|w\|$, we can instead penalize by the squared norm $\|w\|^2$; this is more convenient in terms of optimization. Next, and this is a major change: because the ramp loss is non-convex, in general the empirical risk under the ramp loss also is non-convex. In order to obtain a convex loss function, we will instead use the *hinge loss*, which upper bounds the 1-ramp loss. The hinge loss is defined

$$\ell_{\text{hinge}}(y, f(x)) = \max\{0, 1 - yf(x)\}.$$

The hinge loss can be expressed as a margin loss via the choice $\Phi_{\text{hinge}}(t) = \max\{0, 1 - t\}$. By drawing a picture, it is easy to see that $\Phi_1(t) \leq \Phi_{\text{hinge}}(t)$ for all $t \in \mathbb{R}$. Taking into account these two changes, we arrive at a problem known as the *soft-margin support vector machine (SVM) problem*:

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{n}\sum_{j=1}^{n} \ell_{\text{hinge}}(Y_j, \langle w, X_j\rangle + b) + \frac{\lambda}{2}\|w\|^2.$$

This problem is an adaptation of a problem known as the hard-margin SVM problem. For historical reasons, let us take a brief detour to introduce the hard-margin SVM problem.

The hard-margin version involves minimizing the norm of w subject to the constraint that the hinge loss on *every* example is zero. A direct translation of this statement (while still using the squared norm of w) gives

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & \ell_{\text{hinge}}(y_j, \langle w, x_j \rangle + b) = 0, \ j \in [n]. \end{aligned}$$

However, this problem is usually written as

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_j(\langle w, x_j \rangle + b) \geq 1, \ j \in [n]. \end{aligned}$$

Coming back to the soft-margin SVM problem, we can see that as λ increases, the optimization objective encourages $\|w\|$ to be smaller. Keeping in mind the setting $\gamma = \frac{1}{\|w\|}$ (which is inherent to our derivation of the soft-margin SVM problem), larger λ means smaller $\|w\|$ and hence larger margin γ. Finally, remember that the soft-margin SVM problem is simply the technique that we use to *try* to get a good generalization error bound of the form (7). Here is a practical theory approach one can use:

1. For some choice of $\lambda > 0$, solve the soft-margin SVM problem, yielding solution (w, b).
2. Set $\gamma = \frac{1}{\|w\|}$, and apply the bound (7). Note that this step requires having established a version of (7) that holds simutaneously for all values of γ.

Now, what value of λ should we use? One strategy is to try many values of λ, each one giving a different (w, b), and then select the λ for which our bound (7).

References

Amit Daniely. A PTAS for agnostically learning halfspaces. In *Conference on Learning Theory*, pages 484–502, 2015.

Steve Hanneke and Aryeh Kontorovich. Optimality of svm: novel proofs and tighter bounds. *Theoretical Computer Science*, 796:99–113, 2019.