Machine Learning Theory (CSC 482A/581B) - Lecture 12

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 Support vector machines

In the realm of classification using linear classifiers, we saw in the mistake bound model that data that is linearly separable with a large margin can be learned with far 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's turn now to the statistical learning setting. Suppose that we have a training set which is linearly separable with some margin γ. Clearly, 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 an algorithm, the support vector machine (SVM), whose goal is to find a linear separator which maximizes the margin. We first will consider the case where the data is indeed linearly separable; this is often called the "hard margin" case. After that, we will relax the requirement that the data be perfectly linearly separable, giving rise to the "soft margin" SVM.

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

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

[1] See Daniely (2015) for more details.

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

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

and we use (w, b) to refer to the corresponding hyperplane.

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

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

(I drew a picture explaining this in class)

Observe that from the equation defining any hyperplane (w, b), the hyperplane is invariant to scaling w and b by the same non-zero constant. We can and will pick a particular convenient scaling in order to simplify the computation of the margin. If the data is separable by a hyperplane (as we assume here), the hyperplane does not pass through any point, and we can scale (w, b) such that

$$\min_{j\in[n]} |\langle w, x_j\rangle + b| = 1.$$

In the context of linear separators, any hyperplane with such scaling is referred to as a *canonical hyperplane*. We will refer to the above condition as the canonicality condition.

The margin then takes a particularly simple form:

$$\min_{j\in[n]} \frac{|\langle w, x_j\rangle + b|}{\|w\|} = \frac{1}{\|w\|}.$$

Hard margin SVM. The SVM seeks to maximize the margin, which is equivalent to minimizing $\|w\|$ subject to the canonicality condition. Equivalently, the SVM is the solution to the following optimization problem:

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

Observe that for any candidate solution for which all the constraints are inactive, the objective can be improved by scaling down (w, b) until a constraint becomes active. Therefore, the solution to the SVM problem indeed satisfies the canonicality condition.

The SVM problem is a convex optimization problem. At this point, we can walk over to our friends in optimization and they will happily give us a solution to the SVM problem. Were this a standard machine learning course, we would discuss in more detail efficient methods for solving the SVM problem.

Soft margin SVM. For training sets that are not linearly separable, the hard margin SVM optimization problem does not have a solution. The issue is that the linear inequality constraint cannot be satisfied for all of the training examples simultaneously. The soft margin SVM offers a solution to this issue by relaxing these constraints with slack variables $\xi_1, \ldots, \xi_n$ and introducing a regularization parameter $C \geq 0$:

$$\begin{aligned} \underset{\substack{w,b \\ \xi\geq 0}}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{j=1}^{n} \xi_j \\ \text{subject to} \quad & y_j(\langle w, x_j\rangle + b) \geq 1 - \xi_j,\ j \in [n]. \end{aligned} \tag{Soft-SVM}$$

Similar to before, the above problem is a convex optimization problem and can be solved efficiently.

There is a particularly useful way to interpret the ξ_j variables. These variables can be thought of as indicating "margin errors":

Let $\gamma = \frac{1}{\|w\|}$. For any ξ_j that is equal to zero, it holds that

$$\frac{y_j(\langle w, x_j\rangle + b)}{\|w\|} \geq \frac{1}{\|w\|} = \gamma,$$

and hence (x_j, y_j) is correctly classified with margin at least γ. For any positive $\xi_j \in (0, 1)$, however, the example is correctly classified but with margin only $(1-\xi_j)\gamma$. Lastly, if $\xi_j \geq 1$, this example has been misclassified. Now, we can define a *margin error* (with respect to margin γ) to be any example that is not classified correctly with margin at least γ, which corresponds precisely to the set of indices $\{j \in [n] : \xi_j > 0\}$.

3 Margin bounds

Now that we have an algorithm that attempts to obtain large margin over the training sample, it makes sense to try to develop risk bounds that benefit from hypotheses obtaining large margin.

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 estimator $\hat{f}$, with probability at least $1 - \delta$ over the training sample,

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

We then used the very special relationship

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

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

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

Unfortunately, when the dimension d is large, this bound scales like $O(\sqrt{\frac{d}{n}})$. But this is really the best that we can hope for in a worst-case scenario, and at the point when we have leveraged inequality (1), we have thrown away the labels and hence given up all hope of obtaining a better bound for data that is separable by a large margin.

We therefore will proceed differently, and instead of invoking (1), we will try to find some way to instead obtain a bound that can improve as the margin gets larger, while simulataneously not depending on the dimension.

We begin by considering a useful rewrite of the soft-margin SVM problem. First, we rearrange the terms in the constraints of problem (Soft-SVM), yielding the problem

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

Next, observe that this problem is equivalent to the problem

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 + C\sum_{j=1}^{n} \max\{0, 1 - y_j(\langle w, x_j\rangle + b)\}.$$

This motivates the definition of the *hinge loss*:

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

Unlike the zero-one loss, the second parameter of hinge loss can be real-valued. In fact, for a real-valued predictor $f\colon \mathcal{X} \to \mathbb{R}$, we can also extend the zero-one loss to real-valued predictions via the extended definition

$$\ell_{0-1}(y, f(x)) = \mathbf{1}[yf(x) \leq 0].$$

Problem (Soft-SVM) may now be rewritten as

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

where we have the correspondence $\lambda = \frac{1}{Cn}$.

This rewrite makes evident that the soft-margin SVM problem involves regularized empirical hinge loss minimization; specifically, this problem involves minimizing the empirical hinge risk plus a penalty on the squared ℓ_2-norm of w, with λ modulating the magnitude of the penalty.

The hinge 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 $\Phi : \mathbb{R} \to \mathbb{R}$.

The hinge loss can be expressed as a margin loss using the function

$$\Phi_{\text{hinge}}(t) = \max\{0, 1 - t\}.$$

(I drew a picture in class)

The zero-one loss also can be expressed as a margin loss via 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 swap the zero-one loss for a Lipschitz loss in our analysis, but while still obtaining an upper bound on the risk under zero-one loss. If we can do this, we can avoid the VC-dimension-based upper bound on the Rademacher complexity of a set of classifiers. Instead we need only deal with a set of linear predictors, for which we already have a Rademacher complexity bound (we did this in the last lecture).

2. We will try to relate the Lipschitz loss to the margin. Ideally, if the margin is large, the Lipschitz constant is small, and so our Rademacher complexity-based risk bound will improve with the margin.

For the sake of our analysis, it will be useful to normalize our set of linear predictors such that w always has unit norm. To this end, for any hyperplane (w, b), let $f_{w,b}(x) = \frac{\langle w,x\rangle + b}{\|w\|}$ and define the normalized class $\mathcal{F}_1$ as $\mathcal{F}_1 := \{f_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}$.

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

$$y f_{w,b}(x) = \frac{y\left(\langle w, x\rangle + b\right)}{\|w\|} \geq \gamma. \tag{2}$$

The zero-one loss of $f_{w,b}$ on this example is clearly zero since $\mathbf{1}\left[y f_{w,b} \leq 0\right] = 0$, but, 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. Moreover, 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

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

(I drew a picture in class)

Why is the γ-ramp loss useful in developing a margin bound? From (2), 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, which also held for hinge loss, is that zero-one loss is upper bounded by the γ-ramp loss for any $\gamma > 0$.[2] Consequently, we have for any (real-valued) hypothesis f that

$$\mathsf{E}\left[\mathbf{1}\left[Y f(X) \leq 0\right]\right] \leq \mathsf{E}\left[\Phi_\gamma(Y f(X))\right]. \tag{3}$$

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 (3)), it holds with probability at least $1 - \delta$ that for all $f \in \mathcal{F}_1$,

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

Next, we are going to commit a grave injustice *which has been commited by the authors of most (and possibly all) previous lecture notes and textbooks*: we will assume that we are in the

[2] Note that this observation also held for the hinge loss; in fact, the hinge loss upper bounds the 1-ramp loss.

homogeneous case, so that $b = 0$; this technically is possible by adding 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 note that doing this transformation will have a nontrivial effect on the norm of w and hence on the margin.

Since we are in the homogeneous case, from our upper bound on the Rademacher complexity of linear prediction classes (stated in 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 $f \in \mathcal{F}_1$,

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

Lastly, to make the bound more interpretable, we 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 $f \in \mathcal{F}_1$,

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

Note that the above bound is valid for any choice of γ, as long as the choice is made before seeing the data. It is straightforward to form an essentially equivalent bound that holds for all γ (in some closed interval) simultaneously.

Note that $\Phi_\gamma(y(\langle w, x\rangle + b)\cdot\gamma) = \Phi_1(y(\langle w, x\rangle + b)) \leq \Phi_{\text{hinge}}(y(\langle w, x\rangle + b))$.

Consequently, in the special case that we take $\gamma = \frac{1}{\|w\|}$ (recall that γ is now a parameter of our risk bound), we can view the SVM objective as trying to make the γ-ramp loss small while also trying to make γ (which is now the margin) large.

References

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