EECS 598: Statistical Learning Theory, Winter 2014 **Topic 18**

Rates for Linear SVMs under the Hard Margin Assumption

Lecturer: Clayton Scott *Scribe: Andrew Melfi*

1 Introduction

Note: There are some errors in the bounds below stemming from a missing constant in the second term on the RHS of the Rademacher complexity bound.

In these notes we will establish rates of convergence for linear support vector machines under the hard margin assumption. To set the stage, we first introduce a little notation and a basic result the follows easily from previous discussions.

Denote $\mathcal{X}_M = \{x \in \mathbb{R}^d \,|\, \|x\|_2 \leq M\}$, a ball of radius M. The *linear kernel* is $k(x, x') = \langle x, x' \rangle := x^T x'$. It is called linear because its RKHS contains linear functions. Given $w \in \mathbb{R}^d$, denote by f_w the function $f_w(x) = \langle w, x \rangle$.

Proposition 1. *$\mathcal{F}_M = \{f_w \,|\, w \in \mathcal{X}_M\}$ is the RHKS of k on $\mathcal{X}_M$ and $\|f_w\|_{\mathcal{F}_M} = \|w\|_2 \quad \forall w \in \mathcal{X}_m$*

Proof. Notice that $\mathrm{Id} : \mathcal{X}_m \to \mathcal{X}_m$ is a feature map for the kernel. Futhermore, the map $\mathcal{V} : \mathcal{X}_M \to \mathcal{F}_M$ given by $\mathcal{V}(w) = f_w$ is injective. By Lemma 1 from Topic 12, the result follows.

□

2 Hard Margin Assumption

We know from earlier discussions that rates of convergence are not possible without distributional assumptions. We make the following assumption on the joint distribution P_{XY} of (X, Y).

Definition 1. *We say $\mathcal{L}$ is a* separating hyperplane *if $\exists\, w$ such that*

- $\mathcal{L} = \{x : \langle w, x \rangle = 0\}$
- $\Pr(Y \langle w, X \rangle > 0) = 1$

Definition 2. *We say P_{XY} has a* hard margin *of $\Delta > 0$ if there exists a separating hyperplane $\mathcal{L}$ such that $Pr(X \in \mathcal{L} + \Delta) = 0$ where $\mathcal{L} + \Delta$ means all points whose distance to $\mathcal{L}$ is less than Δ.*

See Fig. 1 for an illustration.

3 Rates for Soft-Margin SVMs

Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{-1, 1\}$ be training data. The soft-margin linear SVM is the discrimination rule $\widehat{f}_n := f_{\widehat{w}_n}$, where $\widehat{w}_n$ is the solution of the optimization problem:

$$\arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \langle w, X_i \rangle) + \lambda \|w\|_2^2,$$

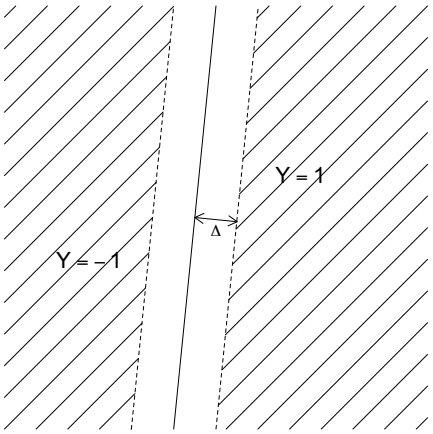

Figure 1: An example of a separating hyperplane for a distribution with a hard margin of Δ.

where L is the hinge loss. This is equivalent to the quadratic program:

$$\min_{w \in \mathbb{R}^d} \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i$$
$$\text{s.t.} \quad Y_i \langle w, \quad X_i \rangle \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

Theorem 1. *Assume P_{XY} has a hard margin of $\Delta > 0$, and also $Pr(\|X\|_2 \leq B) = 1$ for some $B > 0$. Then for any $\delta > 0$, with probability $1 - \delta$,*

$$R(\widehat{f}_n) \leq \frac{2B}{\sqrt{n}} \max\{\lambda^{-1/2}, \Delta^{-1}\} + \frac{\lambda}{\Delta^2} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

In particular, if $\delta \sim n^{-\gamma}$ for some $\gamma \geq \frac{1}{3}$, and $\lambda \sim n^{-3}$, then

$$\mathbb{E}[R(\widehat{f}_n)] = O(n^{-1/3}).$$

The following intermediate result will be used in established the above.

Lemma 1. *If P_{XY} has a hard margin of Δ, then there exists w_Δ such that $\|w_\Delta\|_2 = \frac{1}{\Delta}$ and*

$$\Pr\left(Y \langle w_\Delta, X \rangle \geq 1\right) = 1.$$

Proof. Recall that for any $w, x \in X_M$, the shortest distance from x to $\{x : \langle w, x \rangle = 0\}$ is

$$\frac{|\langle w, x \rangle|}{\|w\|}.$$

Let w_Δ be such that $\mathcal{L} = \{x : \langle w_\Delta, x \rangle = 0\}$ from the hard margin assumption, normalized such that

$\|w_\Delta\| = \frac{1}{\Delta}$ (This is allowed because rescaling the vector w does not alter the hyperplane.) Then

$$\begin{aligned}
\Pr\left(Y\langle w_\Delta, X\rangle \geq 1\right) &= \Pr\left(|\langle w_\Delta, X\rangle| \geq 1\right) \\
&= \Pr\left(|\langle w_\Delta, X\rangle| \geq \frac{\Delta}{\Delta}\right) \\
&= \Pr\left(\frac{|\langle w_\Delta, X\rangle|}{(\frac{1}{\Delta})} \geq \Delta\right) \\
&= \Pr\left(\frac{|\langle w_\Delta, X\rangle|}{\|w_\Delta\|} \geq \Delta\right) \\
&= \Pr\left(\text{distance to hyperplane is} \geq \Delta\right) = 1 \quad \text{(by hard margin assumption).}
\end{aligned}$$

□

Now we can prove Theorem 1:

Proof of Theorem 1. In the proof we refer to w and f_w interchangeably since the two are equivalent and have the same norm (see introduction).

Denote $J(w) = \frac{1}{n}\sum_{i=1}^n L(Y_i, \langle w, x_i\rangle) + \lambda\|w\|_2^2$. Observe also that

$$\lambda\|\widehat{w}_n\|_2^2 \leq J(\widehat{w}_n) \leq J(0) = 1.$$

So $\|\widehat{w}_n\|_2 \leq \frac{1}{\sqrt{\lambda}}$. Let w_Δ be as in the lemma. Then both $\widehat{f}_n = f_{\widehat{w}_n}$ and and f_Δ belong to $\mathcal{F}_M$ where $M = \max\{\lambda^{-1/2}, \Delta^{-1}\}$. We know use the uniform deviation bound (UDB) on balls in a RKHS from a previous lecture. The assumption that $\Pr(\|X\| \leq B) = 1$ means that the kernel is bounded, which is necessary for the UDB.

Let $\delta > 0$. By the one sided uniform deviation bound for balls in a RKHS, with probability $1 - \delta$,

$$\begin{aligned}
R(\widehat{w}_n) &\leq R_L(\widehat{w}_n) \\
&\quad \text{[since hinge loss upper bounds } 0-1 \text{ loss]} && (1)\\
&\leq \widehat{R}_L(\widehat{w}_n) + \frac{2MB}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \\
&\quad \text{[UDB]} && (2)\\
&\leq \widehat{R}_L(f_\Delta) + \lambda\|w_\Delta\|^2 - \lambda\|\widehat{w}_n\|^2 + \frac{2MB}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} && (3)\\
&\quad \text{[since } J(\widehat{w}_n) \leq J(w_\Delta)\text{]} && (4)\\
&\leq \widehat{R}_L(f_\Delta) + \lambda\|w_\Delta\|^2 + \frac{2MB}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} && (5)\\
&\leq \frac{\lambda}{\Delta^2} + \frac{2MB}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. && (6)\\
&\quad [\widehat{R}_L(f_\Delta) = 0] && (7)
\end{aligned}$$

This establishes the first part of the theorem. To prove the second part, let Ω be the event where the bound holds. Then

$$\mathbb{E}[R(\widehat{f}_n)] = \mathbb{E}[R(\widehat{f}_n)|\Omega]\Pr(\Omega) + \mathbb{E}[R(\widehat{f}_n)|\Omega^c]\Pr(\Omega^c)$$

Looking at the above expression, it is clear that $\Pr(\Omega) < 1$, $\Pr(\Omega^c) \leq \delta = n^{-\gamma}$, and $\mathbb{E}[R(\widehat{f}_n)|\Omega^c] < 1$. The remaining term is bounded using the first part of the theorem. So that the first two terms of the bound

grow at an equivalent rate, select λ such that $\lambda = \frac{1}{\sqrt{n\lambda}}$. This yields $\lambda \sim n^{-1/3}$. Plugging this into the above expression, we deduce

$$\mathbb{E}[R(\widehat{f}_n)] = O(n^{-1/3}).$$

□

4 Rates for Hard-Margin Linear SVM

If we know the hard margin assumption holds, we can attain a faster rate using the so-called *hard margin SVM*, where $\widehat{w}_n$ is the solution to

$$\arg\min_{w \in \mathbb{R}^d} \|w\|^2 \quad \text{such that} \quad Y_i \langle w, X_i \rangle \geq 1.$$

So $\widehat{w}_n$ maximizes the distance (among all separating hyperplanes) to the nearest X_i.

Theorem 2. *Assume P_{XY} has hard margin $\Delta > 0$ and $\Pr(\|X\| \geq B) = 1$ for some B. Then for $\delta > 0$, with probability $1 - \delta$,*

$$R(\widehat{f}_n) \leq \frac{2B}{\Delta\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

In particular, if $\delta \sim n^{-\gamma}$ for some $\gamma \geq \frac{1}{2}$, then

$$\mathbb{E}[R(\widehat{f}_n)] = O\left(\sqrt{\frac{\log n}{n}}\right)$$

Proof. Notice that w_Δ is feasible for this optimization problem (that is, it satisfies the constraints but isn't necessarily the minimum). This means that

$$\|\widehat{w}_n\| \leq \|w_\Delta\| = \frac{1}{\Delta}$$

Arguing as in the previous proof, with probability at least $1 - \delta$,

$$R(\widehat{w}_n) \leq R_L(\widehat{w}_n) \leq \widehat{R}_L(\widehat{w}_n) + \frac{2B}{\Delta\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} = \frac{2B}{\Delta\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

The second part of the proof is similar to last time, except with $\delta \sim n^{-\gamma}$, the final term is dominant, and that is why we attain the improved rate of

$$\mathbb{E}[R(\widehat{f}_n)] = O\left(\sqrt{\frac{\log n}{n}}\right).$$

□

The primary reason why we use the soft margin rather than the hard (even though it has more favorable characteristics) is that the hard margin assumption may not be satisfied, which causes the algorithm for the hard margin SVM to fail. The soft margin SVM can still produce a usable hyperplane even without perfect separation.

Exercises

1. Can you generalize the above results to other Lipshitz losses?