EECS 598: Statistical Learning Theory, Winter 2014 **Topic 17**

Universal Consistency of SVMs and Other Kernel Methods

Lecturer: Clayton Scott *Scribe: Kristjan Greenewald*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

As before, the following supervised learning setup is considered. There are available n iid training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. Let k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{F}$, and let $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss. Consider the kernel method

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda \|f\|_{\mathcal{F}}^2.$$

It will be shown that under sufficient conditions on k, L, and $\lambda = \lambda_n$ that $R_L(\widehat{f}_n) \xrightarrow{a.s} R_L^* \; \forall P_{XY}$ and $R(\widehat{f}_n) \xrightarrow{a.s} R^* \; \forall P_{XY}$.

Definition 1 (Lipschitz Loss)**.** *A* Lipschitz loss *is any loss L such that for every $y \in \mathcal{Y}$, $L(y, \cdot)$ is C-Lipschitz where C does not depend on y.*

2 Large RKHSs

Proofs of several results is the section may be found in [1], Chs. 4 and 5, along with additional results and discussion. Recall

$$R_L^* = \inf\{R_L(f) \,|\, f : \mathcal{X} \to \mathbb{R}\}$$

and define

$$R_{L,\mathcal{F}}^* = \inf\{R_L(f) \,|\, f \in \mathcal{F}\}.$$

There exist kernels for which these are equal.

Definition 2 (Universal Kernels)**.** *Let $\mathcal{X}$ be a compact metric space. We say a kernel k on $\mathcal{X}$ is universal if its RKHS $\mathcal{F}$ is dense in $\mathcal{C}(\mathcal{X})$, the set of continuous functions $\mathcal{X} \to \mathbb{R}$, with respect to the supremum norm. That is, $\forall \epsilon > 0$, $\forall g \in \mathcal{C}(\mathcal{X})$, $\exists f \in \mathcal{F}$ such that*

$$\|f - g\|_\infty := \sup_{x \in \mathcal{X}} |f(x) - g(x)| < \epsilon.$$

Facts about universal kernels:

1. If k is universal, then $R_{L,\mathcal{F}}^* = R_L^*$ for any Lipschitz loss L.
2. If $p(t) = \sum_{n \geq 0} a_n t^n$ for $|t| < r$ and $a_n > 0 \, \forall n$, then

$$k(x, x') = p(\langle x, x' \rangle_{\mathbb{R}^d})$$

is universal on $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| < \sqrt{r}\}$. Example: $e^{\beta \langle x, x' \rangle}$ is universal on any compact set in $\mathbb{R}^d$. The proof uses the Stone-Weierstrass Theorem.

3. If k is universal on $\mathcal{X}$, then so is the associated normalized kernel. Hence the Gaussian kernel $e^{-\gamma\|x-x'\|^2}$ is universal on any compact set in $\mathbb{R}^d$. The proof follows from definitions relatively easily.

4. Every nonconstant radial kernel of the form

$$k(x,x') = \int_0^\infty e^{-u\|x-x'\|^2} d\mu(u)$$

   where μ is a nonnegative finite measure, is universal on any compact set in $\mathbb{R}^d$. See [2]. This includes the Gaussian, Laplacian, and multivariate Student kernels.

5. If k is universal, then k is characteristic, which means the map $P \mapsto \int k(\cdot,x)dP(x) \in \mathcal{F}$ is injective.

6. If k is universal on $\mathcal{X}$, and $A, B \subseteq \mathcal{X}$ are disjoint and compact, then $\exists f \in \mathcal{F}$ such that $f(x) > 0 \, \forall x \in A$ and $f(x) < 0 \, \forall x \in B$.

   Proof. Let d be the metric on $\mathcal{X}$. For $C \subseteq \mathcal{X}$ define $d(x,C) = \inf_{x' \in C} d(x,x')$. Consider the function

$$g(x) = \frac{d(x,B)}{d(x,A)+d(x,B)} - \frac{d(x,A)}{d(x,A)+d(x,B)}.$$

   Since $d(x,C)$ is continuous in x (proof left as an exercise), $g \in \mathcal{C}(\mathcal{X})$. Observe that $g(x) = 1$ for $x \in A$ and $g(x) = -1$ for $x \in B$. Let $\epsilon > 0$ and let $f \in \mathcal{F}$ such that $\|f - g\|_\infty < \epsilon$. Then $f \geq 1 - \epsilon$ on A and $f \leq -1 + \epsilon$ on B. $\square$

   This means $\mathcal{F}$ has infinite VC dimension. This can be seen by letting $\{X_1, \ldots X_n\} \in \mathcal{X}$, distinct, $Y_1, \ldots, Y_n \in \{+1, -1\}$, and setting $A = \{X_i | Y_i = +1\}$, $B = \{X_i | Y_i = -1\}$.

   This property has another interesting consequence. Let $\Phi_0 : \mathcal{X} \to \mathcal{F}_0$ be any feature map for k. By an exercise in Topic 12, we know that

$$\mathcal{F} = \{f = \langle w, \Phi_0(\cdot)\rangle_{\mathcal{F}_0}, w \in \mathcal{F}_0\}.$$

   If $f \in \mathcal{F}$, let w be such that $f = \langle w, \Phi_0(\cdot)\rangle_{\mathcal{F}_0}$. Then f is a linear classifier with respect to the transformed data, $(\Phi_0(X_1), Y_1), \ldots, (\Phi_0(X_n), Y_n)$. Let

$$A = \{\Phi_0(X_i) | Y_i = 1\}, \quad B = \{\Phi_0(X_i) | Y_i = -1\}.$$

   Then by Prop. 6, there exists a linear classifier such that the distance from that hyperplane to *every* training point

$$\frac{|\langle w, \Phi_0(x_i)\rangle|}{\|w\|}$$

   is approximately the same. This property is certainly not true for the standard dot product kernel on $\mathbb{R}^d$, and therefore we must be careful when applying our intuition from 2 and 3 dimension to universal kernels.

One drawback of universal kernels is that $\mathcal{X}$ must be compact. While this may not be a limitation in practical applications, it does exclude the theoretically interesting case $\mathcal{X} = \mathbb{R}^d$. Fortunately, the following is true.

Theorem 1. *If k is a Gaussian kernel on $\mathcal{X} = \mathbb{R}^d$ and L is Lipschitz, then $R^*_{L,\mathcal{F}} = R^*_L$.*

3 Universal Consistency

Theorem 2. *Let k be a kernel such that $R^*_{L,\mathcal{F}} = R^*_L$. Let L be a Lipschitz loss for which $L_0 := \sup_{y\in\mathcal{Y}} L(y,0) < \infty$. Assume $\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} = B < \infty$. Let $\lambda = \lambda_n \to 0$, such that $n\lambda_n \to \infty$ as $n \to \infty$. Then $R_L(\widehat{f}_n) - R^*_L \xrightarrow{a.s.} 0 \quad \forall P_{XY}$.*

Corollary 1. *If in addition L is classification calibrated, then $R(\widehat{f}_n) - R^* \xrightarrow{a.s.} 0 \quad \forall P_{XY}$.*

Note. The condition $L_0 < \infty$ always holds for classification problems since $\mathcal{Y}$ is finite.

Proof of Theorem 2. Denote

$$\begin{aligned} J(f) &= \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda_n \|f\|^2 \\ &= \widehat{R}_L(f) + \lambda_n\|f\|^2. \end{aligned}$$

Observe that $J(\widehat{f}_n) \le J(0) \le L_0$. Therefore $\lambda_n\|\widehat{f}_n\|^2 \le L_0 - \widehat{R}_L(\widehat{f}_n) \le L_0$ and so $\|\widehat{f}_n\|^2 \le L_0/\lambda_n$.

Set $M_n = \sqrt{L_0/\lambda_n}$ so that $\widehat{f}_n \in B_k(M_n)$. Let $\epsilon > 0$. By the Borel-Cantelli Lemma it suffices to show

$$\sum_{n\ge 0} \Pr(R_L(\widehat{f}_n) - R^*_L \ge \epsilon) < \infty.$$

Fix $f_\epsilon \in \mathcal{F}$ s.t. $R_L(f_\epsilon) \le R^*_L + \epsilon/2$. Note that $f_\epsilon \in B_k(M_n)$ for n sufficiently large. By the two-sided Rademacher complexity bound for balls in a RKHS (Topic 15), for such large n and with probability $\ge 1-\delta$ w.r.t. the training data,

$$\begin{aligned} R_L(\widehat{f}_n) &\le \widehat{R}_L(\widehat{f}_n) + \frac{2CBM_n}{\sqrt{n}} + (L_0 + CBM_n)\sqrt{\frac{\ln 2/\delta}{2n}} \\ &\le \widehat{R}_L(f_\epsilon) + \lambda_n\|f_\epsilon\|^2 - \lambda_n\|\widehat{f}_n\|^2 + 2CBM_n + (L_0 + CBM_n)\sqrt{\frac{\ln 2/\delta}{2n}} \\ &\qquad \text{(because } J(\widehat{f}_n) \le J(f_\epsilon) \text{ by definition of } \widehat{f}_n\text{)} \\ &\le \widehat{R}_L(f_\epsilon) + \lambda_n\|f_\epsilon\|^2 + 2CBM_n + (L_0 + CBM_n)\sqrt{\frac{\ln 2/\delta}{2n}} \\ &\le R_L(f_\epsilon) + \lambda_n\|f_\epsilon\|^2 + 4CBM_n + 2(L_0 + CBM_n)\sqrt{\frac{\ln 2/\delta}{2n}}. \end{aligned}$$

Note the Rademacher complexity bound is used twice, in the first and last steps. Take $\delta = n^{-2}$, and let N be such that $n \ge N$ implies that both $f_\epsilon \in B_k(M_n)$ and

$$\lambda_n\|f_\epsilon\|^2 + 4CB\sqrt{\frac{L_0}{n\lambda_n}} + 2(L_0 + CB\sqrt{\frac{L_0}{n\lambda_n}})\sqrt{\frac{\ln 2n^2}{2n}} < \epsilon/2.$$

Then for $n \ge N$, w.p. $\ge 1 - n^{-2}$

$$\begin{aligned} R_L(\widehat{f}_n) &< R_L(f_\epsilon) + \epsilon/2 \\ &\le R^*_L + \epsilon. \end{aligned}$$

Therefore

$$\sum_{n\ge 1} \Pr(R_L(\widehat{f}_n) - R^*_L \ge \epsilon) \le N - 1 + \sum_{n\ge N}\frac{1}{n^2} < \infty.$$

$\square$

Remark. Both the hinge and logistic losses are Lipschitz and classification calibrated. Therefore, both the support vector machine and kernel logistic regression, together with a bounded and universal kernel (such as a nonconstant radial kernel, e.g., Gaussian, Laplacian, multivariate Student), and regularization parameter tending to zero slower than $1/n$, are universally consistent on any compact subset of $\mathbb{R}^d$.

Remark. Note that the consistency result does not require $\mathcal{Y} = \{-1, +1\}$. Thus, consider a regression problem with $\mathcal{Y} = [a, b] \subset \mathbb{R}$ and a clipped loss such as

$$L(y, t) = \min\{L_B, |y - t|^p\},$$

$p \geq 1$, $L_B > 0$. Then L satisifies the assumptions of the theorem. However, note that this loss is nonconvex. Other techniques exist for addressing unbounded output spaces and convex regression losses.

Exercises

1. In the definition of universal kernels, why is $\mathcal{X}$ required to be compact?
2. Prove Fact 3 about universal kernels.
3. Rates for linear SVMs under hard margin assumption (there are some errors in the constants below).
   (a) Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$, and L be the hinge loss. Consider the linear classifer $\widehat{f}_n(x) = \widehat{w}_n^T x$ where $\widehat{w}_n$ is the solution of

   $$\min_w \frac{1}{n} \sum_{i=1}^{n} L(Y_i, w^T X_i) + \lambda \|w\|^2$$

   Assume the following about P_{XY}. We say $\mathcal{L}$ is a *separating hyperplane* if there exists w such that $\mathcal{L} = \{x : w^T x = 0\}$ and $\Pr(Y w^T X > 0) = 1$.
   - (Hard margin assumption) There exists a separating hyperplane $\mathcal{L}$ and a $\Delta > 0$ such that $\Pr(X \in \mathcal{L} + \Delta) = 0$, where $\mathcal{L} + \Delta$ is the set of all points within Δ of $\mathcal{L}$.
   - $\Pr(\|X\| \leq B) = 1$ for some $B > 0$.

   Show that with probability at least $1 - \delta$,

   $$R(\widehat{f}_n) \leq \frac{4MB}{\sqrt{n}} + \frac{\lambda}{\Delta} + 2\sqrt{\frac{\log(2/\delta)}{2n}},$$

   for some constant M, and express M in terms of Δ and λ (M should be inversely proportional to both). Show that for appropriate growth of λ, $\mathbb{E}R(\widehat{f}_n) = O(n^{-1/3})$.

   (b) If we know the hard margin condition holds a priori, it makes sense to let $\widehat{w}_n$ be the *hard margin SVM*, obtained by solving

   $$\min_w \|w\|^2 \quad s.t.\ Y_i w^T X_i \geq 1.$$

   This classifier maximizes the distance from the hypeplane $\{x : w^T x = 0\}$ to the nearest training data point, subject to being a separating hyperplane. Show that the same bound as in (a) holds but without the λ/Δ term, and with an M that is no larger than the M from (a). Deduce that $\mathbb{E}R(\widehat{f}_n) = O(n^{-1/2})$.

References

[1] I. Steinwart and A. Christmann, *Support vector machines*, Springer 2008.

[2] C. Micchelli, Y. Xu and H. Zhang, "Universal Kernels," *Journal of Machine Learning Research*, vol. 7, pp. 2651-2667, 2006.