EECS 598: Statistical Learning Theory, Winter 2014 **Topic 8**

Oracle Inequalities and Adaptive Rates

Lecturer: Clayton Scott *Scribe: Yue Wang*

1 Introduction

We have previously seen how sieve estimators give rise to rates of convergence to the Bayes risk by performing empirical risk minimization over $\mathcal{H}_{k(n)}$, where $(\mathcal{H}_k)k \geq 1$ is an increasing sequence of sets of classifiers, and $k(n) \to \infty$. However, the rate of convergence depends on $k(n)$. Usually this rate is chosen to minimize the worst-case rate over all distributions of interest. However, it would be nice if we could automatically get a faster rate of convergence when the distribution is more favorable. Since we don't know whether our distribution is worst-case or not a priori, we don't know how to choose $k(n)$, and adaptive rates of convergence are not possible with sieve estimators.

Adaptive rates are possible, however, using another learning strategy called penalized empirical risk minimization. With this approach we can prove a so-called "oracle inequality" that expresses the ability of penalized ERM to automatically select a classifier of the appropriate complexity so as to achieve improved rates of convergence, even when the best model class $\mathcal{H}_k$ depends on some unknown property of the distribution.

In these notes we will consider oracle inequalities in the context of dyadic decision trees and of general VC classes. In the latter case, penalized empirical risk minimization is known as structural risk minimization.

2 Dyadic Decision Trees

We have actually already seen an example of penalized empirical risk minimization and an associated oracle inequality in the context of dyadic decision trees. However, the results we proved did not actually require penalized ERM and the oracle inequality (see exercise at the end of the notes on dyadic decision trees). We review those results here, and then use the oracle inequality to obtain an adaptive rate of convergence.

Let $\mathcal{X} \in [0,1]^d$. Recall $\mathcal{T}_m = \{\text{all DDTs whose cells all have sidelength} \geq 1/m\}$. Denote by $\Pi(h) = \{A_i\}$ the recursive dyadic partition associated with h. Recall the penalized empirical risk minimizer

$$\widehat{h}_n = \arg\min_{h \in \mathcal{T}_m} \widehat{R}_n(h) + \Phi_n(h), \tag{1}$$

where $\Phi_n(h)$ is the complexity penalty

$$\Phi_n(h) = \sum_{A \in \Pi(h)} \sqrt{2B2^{-j(A)} \frac{[\kappa j(A) + \log(n)]}{n}}$$

where $j(A)$ is the depth of cell A (the number of splits needed to form A), B is a bound on the assumed density of P_X, and κ is a constant. We have shown that with probability at least $1 - 1/n$,

$$R(\widehat{h}_n) - R^* \leq \inf_{h \in \mathcal{T}_m} \left\{ \underbrace{R(h) - R^*}_{\text{approximation error}} + \underbrace{2\Phi_n(h)}_{\text{bound on estimation error}} \right\} \tag{2}$$

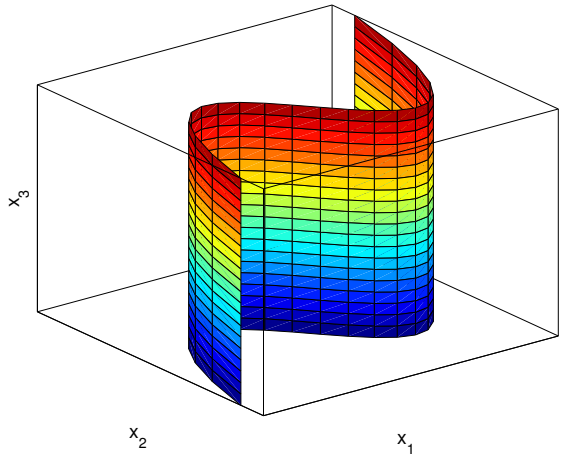

Figure 1: An example of a Bayes decision boundary when there exists feature dimensions irrelevant to the label. In this case, x_3 is the irrelevant feature dimension.

By setting $m \sim \left(\frac{n}{\log n}\right)^{\frac{1}{d}}$, both error terms converge to zero as $O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d}}\right)$ whenever P_{XY} belongs to the box-counting class (defined in the notes on the sieve estimators). Inequality (2) is an example of an *oracle inequality*: the discrimination rule $\widehat{h}_n$ in (1) achieves an optimal trade-off between approximation error and (a bound on) estimation error.

In many real applications, only some dimensions of the feature space are relevant for classification. For example, in the case of Figure 1, there is no need to split along dimension x_3. It can be shown that for a distribution in the box counting class with only $d_r < d$ relevant features, the optimal rate of convergence (that holds uniformly for all such distributions) is $O(n^{-1/d_r})$ [1]. We will see that the discrimination rule in (1) can attain this optimal rate *without knowledge of* d_r. In this sense, the discrimination rule is adaptive.

We first extends the definition of box-counting class to incorporate the number of relevant features.

Definition 1. *Let $d_r \leq d$. Define $\mathcal{B}(d_r)$ to be the set of all P_{XY} such that*

(A) *P_X has a bounded probability density function f, $\|f\|_\infty \leq B$;*

(B) *$\exists C$ s.t. $\forall m \geq 1$, the Bayes decision boundary intersects at most Cm^{d-1} of the m^d cells in a regular partition of $\mathcal{X}$;*

(C) *d_r of the features are statistically dependent on Y*

Theorem 1. *Let $m \sim \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$. If $P_{XY} \in \mathcal{B}(d_r)$, then*

$$\mathbb{E}\left[R(\widehat{h}_n) - R^*\right] = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d_r}}\right).$$

The convergence rate is optimal up to the logarithmic factor.

Proof. (Sketch – proof of the lemmas is left as an exercise.) Without loss of generality, assume the first d_r features are relevant. Let $m_r = 2^J \leq m$ for some integer $J > 0$; Let $h_{m_r} \in \mathcal{T}_m$ be the DDT obtained by cycling through the first d_r dimensions J times each. Let $h^*_{m_r} \in \mathcal{T}_m$ be obtained by pruning all cells of h_{m_r} except those intersecting the Bayes decision boundary (and their siblings).

Lemma 1.

$$R(h^*_{m_r}) - R^* = O\left(\frac{1}{m_r}\right)$$

The proof of Lemma 1 is left as an exercise.

Lemma 2.

$$\mathbb{E}\left[R(h_n) - R(h^*_{m_r})\right] = O\left(m_r^{\frac{d_r}{2}-1}\sqrt{\frac{\log n}{n}}\right)$$

The proof of Lemma 2 is left as an exercise. To prove Theorem 1, now choose

$$m_r \sim \left(\frac{n}{\log n}\right)^{\frac{1}{d_r}}$$

□

A key point to realize is that because of the oracle inequality, the user does not actually need to know d_r.

3 Structural Risk Minimization

Let $\{\mathcal{H}_k\}_{k\geq 1}$ be a collection of VC classes with VC dimensions V_k. Recall that for each k, we have the uniform deviation bound: with probability at least $1-\delta_k$, $\forall h \in \mathcal{H}_k$,

$$\left|\widehat{R}_n(h) - R(h)\right| \leq \sqrt{\frac{32\left[V_k \log(n+1) + \log(8/\delta_k)\right]}{n}}.$$

Further, we know that with probability at least $1-\delta_k$,

$$R(\widehat{h}_{n,k}) \leq R^*_{\mathcal{H}_k} + 2\sqrt{\frac{32\left[V_k \log(n+1) + \log(8/\delta_k)\right]}{n}}, \tag{3}$$

where $\widehat{h}_{n,k}$ is ERM over $\mathcal{H}_k$. We can unify these bounds into one, as stated below.

Corollary 1. *With probability at least* $1-1/n$, $\forall k,\ \forall h \in \mathcal{H}_k$,

$$\left|\widehat{R}_n(h) - R(h)\right| \leq \sqrt{\frac{32\left[V_k \log(n+1) + k\log 2 + \log(8n)\right]}{n}} =: \Phi_n(k) \tag{4}$$

Proof. Think of the complementary event on which the inequality (4) fails to hold. That means at least one $h \in \mathcal{H}_k$ fails the bound (3) for some k, which occurs with probability at most δ_k. By the union bound, the inequality (4) fails with probability at most $\sum_k \delta_k$. Set $\delta_k = \delta 2^{-k}$ and $\delta = 1/n$ to obtain the result. □

Note that $\Phi_n(k)$ is an increasing function of the VC dimension V_k. We think of $\Phi_n(k)$ as a complexity penalty. The following discrimination rule formulates penalized ERM in this setting.

Definition 2. *Structural risk minimization (SRM) is defined via* $\widehat{h}_n = \widehat{h}_{n,\widehat{k}}$, *where*

$$\widehat{k} = \arg\min_{k\geq 1} \underbrace{\widehat{R}_n(\widehat{h}_{n,k})}_{\textit{empirical risk}} + \underbrace{\Phi_n(k)}_{\textit{complexity}}$$

The chosen $\widehat{k}$ *effects the trade-off between data fidelity and model complexity.*

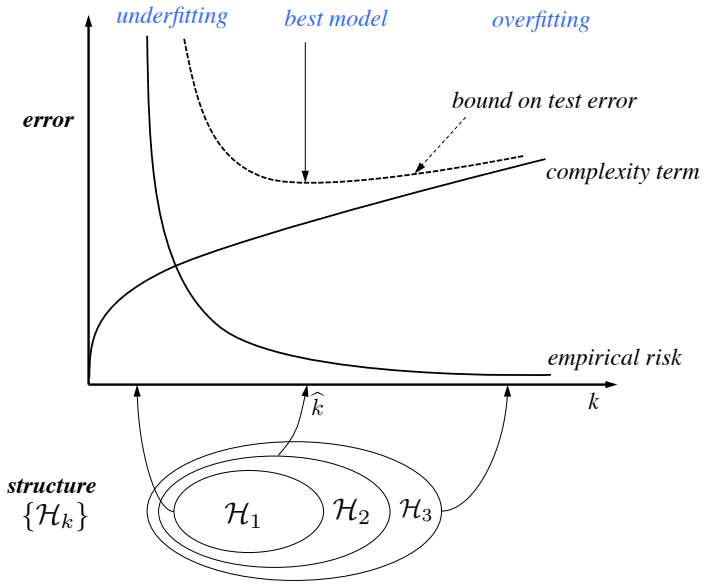

Figure 2: The bound on the test error is the sum of the empirical risk and a complexity term. The empirical risk $\widehat{R}_n(\widehat{h}_{n,k})$ decreases with the index k of the VC class $\mathcal{H}_k$ (structure), while the model complexity $\Phi_n(k)$ increases with k. The optimal bound is achieved on some appropriate $\widehat{k}$.

The SRM principle is illustrated in Figure 2. On the left side of x-axis, the model complexity is low, which means the classifier does not have enough capacity to the fit training data well, leading to high empirical risk or *underfitting*. On the right side of the x-axis, the model is complex enough to fit the training data well (hence low empirical error), but an over-complex model tends to perfectly fit the idiosyncrasies (noise) in training data, leading to poor generalization on test data, a phenomenon called *overfitting*. SRM selects the best model that achieves a near-optimal trade-off between empirical error and model complexity. This is reflected in the following result.

Theorem 2. *With probability at least* $1 - 1/n$,

$$R(\widehat{h}_n) - R^* \leq \inf_{k \geq 1} \left\{ R^*_{\mathcal{H}_k} - R^* + 4\Phi_n(k) \right\}.$$

Proof. Consider the event $\Omega = \{\text{bound in Corollary 1 holds}\}$. By assumption, $\Pr(\Omega) \geq 1 - 1/n$. On event Ω, for any $k \geq 1$,

$$\begin{aligned}
R(\widehat{h}_n) &= R(\widehat{h}_{n,\widehat{k}}) \\
&\leq \widehat{R}(\widehat{h}_{n,\widehat{k}}) + \Phi_n(\widehat{k}) && \text{Inequality (4)} \\
&\leq \widehat{R}(\widehat{h}_{n,k}) + \Phi_n(k) && \text{Definition of SRM } \widehat{h}_{n,\widehat{k}} \\
&\leq R(\widehat{h}_{n,k}) + 2\Phi_n(k) && \text{Inequality (4)} \\
&\leq R^*_{\mathcal{H}_k} + 4\Phi_n(k) && \text{Inequality (3)}
\end{aligned}$$

Now subtract R^* from both sides. Since k was arbitrary, the proof is complete. □

Exercises

1. Complete the proof of Theorem 1 as follows.

   (a) Prove Lemma 1. *Hint:* show that the projection of the Bayes decision boundary onto the relevant dimensions intersects no more than Cm^{d_r-1} cells in $[0,1]^{d_r}$. You may also want to look at the analysis of approximation error in our study of the histogram rule.

   (b) Prove Lemma 2. *Hint:* This should be a relatively straightforward modification of our bound on the estimation area from the notes on dyadic decision trees.

2. Suppose we have access to some *holdout data* $X_{n+1}, \ldots, X_{n+m}$ that is not used for training. Define the holdout error of a classifier h to be

$$\tilde{R}_n(h) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{h(X_{n+i}) \neq Y_{n+i}\}}.$$

   Now suppose we have a sequence of sets of classifiers $(\mathcal{H}_k)$ of increasing complexity. Consider the discrimination rule $\tilde{h}_n = \widehat{h}_{h,\widehat{k}}$, where $\widehat{h}_{n,k}$ is empirical risk minimization over $\mathcal{H}_k$, and

$$\widehat{k} = \arg\min_{k \geq 1} \tilde{R}_n(h) + \tilde{\Phi}_n(k).$$

   Your task in this problem is to define $\tilde{\Phi}_n(k)$ and prove an oracle inequality for $\tilde{h}_n$. Your result should basically say that SRM does about as well as the best empirical risk minimizer, which of course is not known. *Note:* We are not assuming that the $\mathcal{H}_k$ are VC classes. Also, "with high probability" refers to the draw of both the training and holdout data.

3. DDTs also adapt to the intrinsic dimension of the data distribution. Modify conditions **(A)** and **(B)** in the definition of the box-counting class to reflect the idea that the data lie on a manifold of dimension $d_m < d$. Then show that DDTs can adaptively come within a log factor of the optimal rate of $O(n^{-1/d_m})$. *Hint:* The modification of **(B)** is fairly straightforward. To modify **(A)**, don't mess with trying to define a density on a manifold; instead, use box-counting ideas and formulate this condition in terms of the probability of the data occurring in a hypercube with side length $1/m$ (i.e., bound this probability in terms of some power of m).

4. Give conditions under which SRM is strongly universally consistent. Provide a theorem statement with proof. Compare to sieve estimation.

References

[1] C. Scott and R. Nowak, "Minimax-Optimal Classification with Dyadic Decision Trees," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1335-1353, 2006.