EECS 598: Statistical Learning Theory, Winter 2014 **Topic 4**

Empirical Risk Minimization

Lecturer: Clayton Scott *Scribe: John Lipor*

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

Let (X_i, Y_i), $i = 1, \ldots, n$ be i.i.d. with distribution P_{XY}. Recall that P_{XY} is a distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{Y} = \{0, 1\}$ and define a set of classifiers $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$. A natural choice for a learning algorithm is *empirical risk minimization* (ERM)

$$\widehat{h}_n = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$$

where $\widehat{R}_n(h) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{h(X_i) \neq Y_i\}}$. An important question is how close is $R(\widehat{h}_n)$ to $R^*_{\mathcal{H}} := \inf_{h \in \mathcal{H}} R(h)$. This will be explored in the following sections.

2 Uniform Deviation Bounds

Previously we saw that for any fixed h (not dependent on data)

$$\Pr\left(|\widehat{R}_n(h) - R(h)| \geq \epsilon\right) \leq \delta$$

where $\delta = 2e^{-2n\epsilon^2}$. Since we don't know $\widehat{h}_n$ a priori, we will look for a *uniform deviation bound* (UDB), which has the form

$$\Pr\left(\sup_{h \in \mathcal{H}} |\widehat{R}_n(h) - R(h)| \geq \epsilon\right) \leq \delta. \tag{1}$$

Note that in this case the random quantity is the training data. Consider as a first example the case where $|\mathcal{H}| < \infty$.

Proposition 1. *Assume $|\mathcal{H}| < \infty$. Then*

$$\Pr\left(\sup_{h \in \mathcal{H}} |\widehat{R}_n(h) - R(h)| \geq \epsilon\right) \leq 2|\mathcal{H}|e^{-2n\epsilon^2}.$$

Proof. Let $\Omega_\epsilon(h) \subseteq (\mathcal{X} \times \mathcal{Y})^n$ be the event that $|\widehat{R}_n(h) - R(h)| \geq \epsilon$. Let $\Omega_\epsilon = \cup_{h \in \mathcal{H}} \Omega_\epsilon(h)$. Then

$$\begin{aligned}
\Pr\left(\sup_{h \in \mathcal{H}} |\widehat{R}_n(h) - R(h)| \geq \epsilon\right) &= \Pr(\Omega_\epsilon) \\
&\leq \sum_{h \in \mathcal{H}} \Pr(\Omega_\epsilon(h)) \\
&\leq \sum_{h \in \mathcal{H}} 2e^{-2n\epsilon^2} \\
&= 2|\mathcal{H}|e^{-2n\epsilon^2}.
\end{aligned}$$

□

A key point is that the result is distribution free, i.e., it requires no assumptions on P_{XY}. A UDB let's us bound the performance of ERM.

Proposition 2. *Suppose* $\mathcal{H}$ *satisfies (1). Then with probability at least* $1-\delta$

$$R(\widehat{h}_n) \leq R^*_{\mathcal{H}} + 2\epsilon.$$

Proof. Let Ω_ϵ be the event that $\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| \geq \epsilon$. By assumption, $\Pr(\Omega_\epsilon) \leq \delta$. Let $h \in \mathcal{H}$ be any classifier. Then on Ω_ϵ^c we have

$$\begin{aligned} R(\widehat{h}_n) &\leq \widehat{R}_n(\widehat{h}_n) + \epsilon \\ &\leq \widehat{R}_n(h) + \epsilon \\ &\leq R(h) + 2\epsilon, \end{aligned}$$

where the second step follows from the definition of ERM. Note that the choice of $h \in \mathcal{H}$ was arbitrary, so $R(\widehat{h}_n) \leq R^*_{\mathcal{H}} + 2\epsilon$.

□

Remark. Note that the above proof assumes the existence of an empirical risk *minimizer*. For finite $\mathcal{H}$, this is guaranteed. For infinite $\mathcal{H}$, however, the existence of an empirical risk minimizer needs to be checked. If a minimizer does not exist, one can modify the above argument by taking $\widehat{h}_n$ to come within $\tau > 0$ of the infimum of the empirical risk, where τ may be arbitrarily small.

Remark. Note that as an intermediate result we established the non-probabilistic statment

$$R(\widehat{h}_n) - R^*_{\mathcal{H}} \leq 2 \sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)|.$$

Corollary 1. *If* $\mathcal{H}$ *is finite, then*

$$\Pr\left(R(\widehat{h}_n) \geq R^*_{\mathcal{H}} + \epsilon\right) \leq \underbrace{2|\mathcal{H}|e^{-n\epsilon^2/2}}_{\delta}.$$

(Note that the term 2ϵ *was replaced by* ϵ*.) Equivalently, with probability at least* $1-\delta$

$$R(\widehat{h}_n) \leq R^*_{\mathcal{H}} + \sqrt{\frac{2\left[\log|\mathcal{H}| + \log(2/\delta)\right]}{n}}.$$

3 Histogram Classifier

Let $\mathcal{X} = [0,1]^d$, $k \geq 1$, $k \in \mathbb{Z}$. Let $\mathcal{H}_k$ be the set of classifiers that are piecewise constant on regular partitions of $\mathcal{X}$ into hypercubes of sidelength $1/k$. Note that $\widehat{h}_n(x)$ is the majority vote in each cell. An example of one such classifier can be seen in Figure 1. With the given parameters, we have

$$|\mathcal{H}_k| = 2^{k^d}.$$

Then with probability at least $1-\delta$

$$R(\widehat{h}_n) \leq R^*_{\mathcal{H}} + \sqrt{\frac{2\left[k^d\log(2) + \log(2/\delta)\right]}{n}}.$$

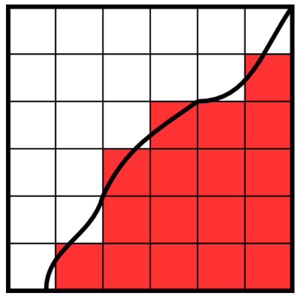

Figure 1: Example histogram classifier where the white squares represent class 0 and the red squares class 1. In this case, $d = 2$ and $k = 6$. ERM assigns labels to cells by a majority vote of data points X_i in each cell.

4 PAC Learning & Sample Complexity

Definition 1. *We say $\widehat{h}_n$ is an (ϵ, δ)-learning algorithm for $\mathcal{H}$ if there exists a function $N(\epsilon, \delta)$ such that $\forall \epsilon, \delta > 0$*

$$n \geq N(\epsilon, \delta) \Rightarrow \Pr\left(R(\widehat{h}_n) - R^*_{\mathcal{H}} \geq \epsilon\right) \leq \delta.$$

Terminology:

- $N(\epsilon, \delta)$ is called the *sample complexity*
- $\mathcal{H}$ is said to be *uniformly learnable*
- $\widehat{h}_n$ is *probably approximately correct* (PAC)

For finite $\mathcal{H}$, we have $\delta = 2|\mathcal{H}|e^{-n\epsilon^2/2}$. Solving for n,

$$N(\epsilon, \delta) = \frac{2 \log \frac{2|\mathcal{H}|}{\delta}}{\epsilon^2}.$$

Therefore $\mathcal{H}$ is uniformly learnable and ERM is PAC.

5 Zero Error Case

If $\widehat{R}_n(\widehat{h}_n) = 0$, we can obtain a tighter bound.

Proposition 3. *Let $|\mathcal{H}| < \infty$. Then*

$$\Pr\left(\exists h \in \mathcal{H} : \widehat{R}_n(h) = 0,\ R(h) \geq \epsilon\right) \leq \underbrace{|\mathcal{H}|e^{-n\epsilon}}_{\delta}$$

i.e., with probability at least $1 - \delta$, if $\widehat{R}_n(h) = 0$, then $R(h) \leq \frac{\log|\mathcal{H}| + \log(1/\delta)}{n}$.

Proof. Let $\Omega_0(h) = \{\widehat{R}_n(h) = 0\}$ and $\Omega_\epsilon = \cup_{h:R(h)\geq\epsilon} \Omega_0(h)$. Then for any h such that $R(h) \geq \epsilon$

$$\begin{aligned}\Pr(\Omega_0(h)) &\leq (1-\epsilon)^n \\ &= e^{n\log(1-\epsilon)} \\ &\leq e^{-n\epsilon}\end{aligned}$$

where we used $\log(1-\epsilon) \leq -\epsilon$. Therefore

$$\begin{aligned}\Pr(\Omega_\epsilon) &\leq \sum_{h: R(h) \geq \epsilon} e^{-n\epsilon} \\ &\leq |\mathcal{H}| e^{-n\epsilon}.\end{aligned}$$

□

Exercises

1. The probability of error is not the only performance measure for binary classification. Indeed, the probability of error depends on the prior probability of the class label Y, and it may be that the frequency of the classes changes from training to testing data. In such cases, it is desirable to have a performance measure that does not require knowledge of the prior class probability. Let P_y be the class conditional distribution of class y, $y = 0, 1$. For $y = 0, 1$ define $R_y(h) := P_y(h(X) \neq y)$. Also let $\alpha \in (0, 1)$. For $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ define

$$\begin{aligned}R^*_{\mathcal{H},1} = \inf_{h \in \mathcal{H}} &\; R_1(h) \\ s.t.\; &R_0(h) \leq \alpha.\end{aligned}$$

   In this problem you will investigate a discrimination rule that is probably approximately correct with respect to the above criterion, which is sometimes called the Neyman-Pearson criterion based on connections to the Neyman-Pearson lemma in hypothesis testing.

   Suppose we observe $X^y_1, \ldots, X^y_{n_y} \stackrel{iid}{\sim} P_y$ for $y = 0, 1$. Define the empirical errors

$$\widehat{R}_y(h) = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathbf{1}_{\{h(X^y_i) \neq y\}}.$$

   Fix $\epsilon > 0$ and consider the discrimination rule

$$\begin{aligned}\widehat{h}_n = \arg\min_{h \in \mathcal{H}} &\; \widehat{R}_1(h) \\ s.t.\; &\widehat{R}_0(h) \leq \alpha + \frac{\epsilon}{2}.\end{aligned}$$

   Suppose $\mathcal{H}$ is finite. Show that with high probability

$$R_0(\widehat{h}_n) \leq \alpha + \epsilon \text{ and } R_1(\widehat{h}_n) \leq R^*_{\mathcal{H},1} + \epsilon.$$