EECS 598: Statistical Learning Theory, Winter 2014 **Topic 20**

Weakly Supervised Learning

Lecturer: Clayton Scott *Scribe: Yanzhen Deng, Nhat Ho, and Hossein Keshavarz*

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

Weakly supervised learning problems are between supervised and unsupervised learning problems. You can think of them as supervised learning problems where some label information is missing or has been contaminated in some way. We will focus on a specific weakly supervised learning problem, namely, binary classification with one-sided label noise, although the ideas here can be brought to bear on more general WSL problems, such as two-sided label noise [2], label noise in the multiclass case [3], and anomaly detection [4]. The material described here appeared originally in [1].

2 One-sided Label Noise

Let P_0, P_1 be class-conditional distributions. Suppose:

$$\begin{aligned} &X_1^0, ..., X_{n_0}^0 \overset{i.i.d.}{\sim} P_0 \\ &X_1^1, ..., X_{n_1}^1 \overset{i.i.d.}{\sim} \widetilde{P}_1 = (1-\kappa)P_1 + \kappa P_0. \end{aligned} \tag{1}$$

Both P_0, P_1 and the contamination proportion $\kappa \in [0,1]$ are unknown. We think of the second sample as being labeled as belonging to class 1, but some of those labels are wrong.

We measure the performance of a classifier h using the class-conditional error probabilities

$$R_i(h) = P_i(h(X) \neq i).$$

If (X, Y) are jointly distributed and $Pr(Y = 1) = \pi$, then the usual risk is $R(h) = (1-\pi)R_0(h) + \pi R_1(h)$. Alternatively, we can use $R^{max}(h) = R_0(h) \vee R_1(h)$ if we want to be non-Bayesian, or several other performance measures defined in terms of R_0 and R_1.

Suppose we wish to establish a consistent sieve estimator over $(\mathcal{H}_k)_{k\geq 1}$. Then we need estimators $\widehat{R}_0$, $\widehat{R}_1$ s.t.

$$\sup_{h\in\mathcal{H}_k} |\widehat{R}_i(h) - R_i(h)| \to 0 \quad i.p., \quad i = 0, 1$$

as $n \wedge m = \min\{n, m\} \to \infty$ with $k = k(m, n)$ chosen appropriately.

The obvious estimate for R_0 is

$$\widehat{R}_0(h) = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbf{1}_{\{h(X_i^0)=1\}}.$$

There is no obvious estimate of R_1, but we can estimate

$$\widetilde{R}_1(h) := \widetilde{P}_1(h(X) = 0) = (1-\kappa)R_1(h) + \kappa(1 - R_0(h))$$

via

$$\widehat{\widetilde{R}}_1(h) = \frac{1}{n}\sum_{i=1}^{n_1} \mathbf{1}_{\{h(X_i^1)=0\}}.$$

If $\kappa < 1$ and there exists a consistent estimator $\widehat{\kappa} \overset{i.p.}{\to} \kappa$, then it makes sense to estimate R_1 via

$$\widehat{R}_1(h) = \frac{\widehat{\widetilde{R}}_1(h) - \widehat{\kappa}(1 - \widehat{R}_0(h))}{1 - \widehat{\kappa}}.$$

Indeed, we have the following proposition:

Proposition 1. *Let $(\mathcal{H}_k)_{k \geq 1}$ be a sequence of VC classes, where $V_k < \infty$. Let $k(n,m)$ satisfy*

$$\frac{V_{k(n,m)} \log(n \wedge m)}{n \wedge m} \to 0$$

as $n \wedge m \to \infty$. If $\widehat{\kappa} \overset{i.p.}{\to} \kappa < 1$, then

$$\sup_{h \in \mathcal{H}_k} |\widehat{R}_i(h) - R_i(h)| \to 0 \quad i.p., \quad i = 0, 1.$$

The proof is left as an exercise. If we additionally require $k(m,n) \to \infty$ and $(\mathcal{H}_k)$ to possess the universal approximation property, it is straightforward to establish a consistent discrimination rule based on the sieve estimator construction.

3 Irreducibility and the Maximum Mixture Proportion

From the above, it suffices to determine a consistent estimator $\widehat{\kappa}$ of κ. Note, however, that κ in (1) is not identifiable (given $\widetilde{P}_1$ and P_0) without additional assumptions. Indeed, if $\widetilde{P}_1 = (1-\kappa)Q + \kappa P_0$, then it also true that $\widetilde{P}_1 = (1-\kappa')Q' + \kappa' P_0$ for any $\kappa' \in [0, \kappa]$, where

$$Q' = \frac{(1-\kappa)Q + (\kappa - \kappa')P_0}{1 - \kappa'}.$$

Because we have no knowledge of P_1, we cannot decide which representation is the correct one. Therefore, to make κ identifiable (uniquely determined), we adopt the following assumption.

Definition 1. *We say P_1 is* irreducible *with respect to P_0 if there exists no distribution Q and $\gamma > 0$ such that $P_1 = (1-\gamma)Q + \gamma P_0$.*

Examples.

1. If $P_1 = \frac{1}{3}P_0 + \frac{2}{3}P^{'}$, where $P^{'}$ is an arbitrary distribution, then P_1 is clearly not irreducible wrt P_0. See Fig. 1.
2. Suppose $P_0 = N(\mu_0, \sigma_0^2)$ and $P_1 = N(\mu_1, \sigma_1^2)$. If $\mu_0 \neq \mu_1$ and $\sigma_0 \geq \sigma_1$, then P_1 is irreducible with respect to P_0. This can be checked by a property we will state below relating irreducibility and the infimum of the likelihood ratio (see exercises below). See Fig. 2.
3. If $\text{supp}(P_0) \not\subset \text{supp}(P_1)$ then P_1 is irreducible with respect to P_0. See Fig. 3.

Definition 2. *Given $\widetilde{P}_1$, P_0, define $\kappa^*(\widetilde{P}_1 | P_0) = \sup\Big\{\alpha \in [0,1] \,\Big|\, \exists Q \text{ s.t } \widetilde{P}_1 = (1-\alpha)Q + \alpha P_0\Big\}$.*

Remark. P_1 is irreducible with respect to P_0 if and only if $\kappa^*(P_1 | P_0) = 0$.

The following result is the main result concerning identifiability of κ in (1).

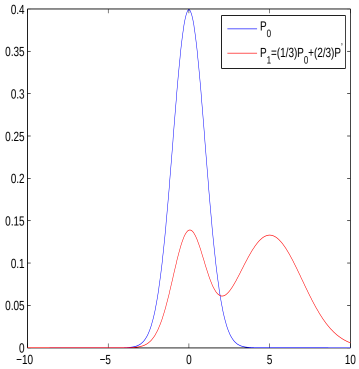

Figure 1: P_1 is not irreducible wrt P_0.

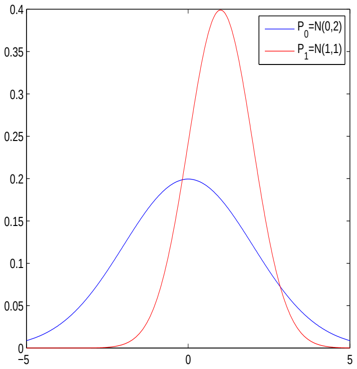

Figure 2: P_1 is irreducible with respect to P_0.

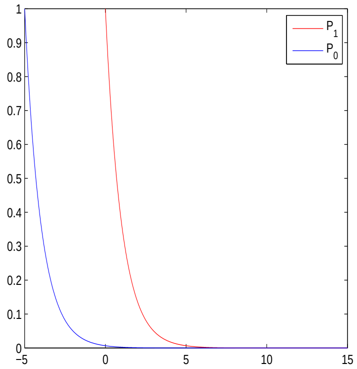

Figure 3: P_1 is irreducible with respect to P_0.

Theorem 1. *Consider a measurable space* $(\mathcal{X}, \mathcal{A})$ *and two distributions* $\widetilde{P}_1 \neq P_0$. *Then there exists a unique* $\kappa < 1$ *and a distribution* P_1 *such that* $\widetilde{P}_1 = (1-\kappa)P_1 + \kappa P_0$ *and* P_1 *is irreducible with respect to* P_0. *If we set* $\kappa = 1$ *when* $\widetilde{P}_1 = P_0$ *then in all cases, we have*

$$\kappa = \kappa^*(\widetilde{P}_1|P_0) = \inf_{A\in\mathcal{A}:P_0(A)>0} \frac{\widetilde{P}_1(A)}{P_0(A)}.$$

Remarks.

1. From the theorem, if $\widetilde{P}_1 = (1-\kappa)P_1 + \kappa P_0$ and P_1 is irreducible with respect to P_0, then $\kappa = \kappa^*(\widetilde{P}_1|P_0)$ if $\kappa < 1$. Thus, irreducibility of P_1 wrt P_0 is sufficient for κ in (1) to be identifiable.

2. If $\widetilde{P}_1$ and P_0 have Lebesgue densities $\widetilde{f}_1$ and f_0, then

$$\kappa^*(\widetilde{P}_1|P_0) = \operatorname{ess\,inf} \frac{\widetilde{f}_1(x)}{f_0(x)} := \inf\left\{\alpha : \exists A \in \mathcal{A} \text{ s.t } P_0(A) > 0 \text{ and } \frac{\widetilde{f}_1}{f_0} \leq \alpha \text{ on } A\right\}.$$

   This is analogous to the last part of the theorem, and the proof is left as as exercise. This property can be used to check irreducibility for specific densities.

3. By this result, $d(\widetilde{P}_1|P_0) := 1 - \kappa^*(\widetilde{P}_1|P_0)$ is a *statistical distance*, meaning it is nonnegative and equal to zero iff the two distributions are the same. This particular statistical distance has been known as the *separation distance* in the study of Markov chains.

Proof. If $\widetilde{P}_1 = P_0$ then it is clear that $\kappa^*(\widetilde{P}_1|P_0) = \inf_{A\in\mathcal{A}:P_0(A)>0} \dfrac{\widetilde{P}_1(A)}{P_0(A)} = 1$. Henceforth, we will assume $\widetilde{P}_1 \neq P_0$. Denote

$$a = \inf_{A\in\mathcal{A}:P_0(A)>0} \frac{\widetilde{P}_1(A)}{P_0(A)}.$$

Claim: $a < 1$. This is seen as follows. Since $\widetilde{P}_1 \neq P_0$, there exists a measurable set $A \in \mathcal{A}$ such that $\widetilde{P}_1(A) \neq P_0(A)$. Consider three cases: If $P_0(A) = 1$, then $\widetilde{P}_1(A) < 1 = P_0(A)$, so $a < 1$. If $0 < P_0(A) < 1$, then either $\widetilde{P}_1(A)/P_0(A) < 1$ or $\widetilde{P}_1(A^c)/P_0(A^c) < 1$. If $P_0(A) = 0$, then $P_1(A) > 0$. Therefore, $P_0(A^c) = 1 > P_1(A^c)$. In all cases we see $a < 1$.

Now define the probability measure

$$P_1 := \frac{\widetilde{P}_1 - aP_0}{1-a},$$

which clearly satisfies $\widetilde{P}_1 = (1-a)P_1 + aP_0$, from which we deduce that $\kappa^*(\widetilde{P}_1|P_0) \geq a$. Now from the definition of inf we have

$$\forall \epsilon > 0,\ \exists A \in \mathcal{A} \text{ such that } P_0(A) > 0 \text{ and } \frac{\widetilde{P}_1(A)}{P_0(A)} < a + \epsilon. \tag{2}$$

If $\widetilde{P}_1 = (1-\alpha)Q + \alpha P_0$ then for all $A \in \mathcal{A}$, $\dfrac{\widetilde{P}_1(A)}{P_0(A)} \geq \alpha$. Therefore, $\alpha \leq a$ from (2), so $\kappa^*(\widetilde{P}_1|P_0) = a$. Since $\dfrac{\widetilde{P}_1}{P_0} = (1-a)\dfrac{P_1}{P_0} + a$, (2) also implies

$$\forall \epsilon > 0,\ \exists A \in \mathcal{A} \ \text{ such that } P_0(A) > 0 \text{ and } \frac{P_1(A)}{P_0(A)} < \frac{\epsilon}{1-a}. \tag{3}$$

This implies that P_1 is irreducible with respect to P_0, for if $P_1 = (1-\gamma)Q + \gamma P_0$ for some $\gamma > 0$, then $\dfrac{P_1(A)}{P_0(A)} \geq \gamma$ for all $A \in \mathcal{A}$, contradicting (3).

Finally we establish uniqueness. Consider a distribution Q such that $\widetilde{P}_1 = (1-\gamma)Q + \gamma P_0$, where $\gamma \leq a$. Then $(1-\gamma)Q + \gamma P_0 = (1-a)P_1 + aP_0$ and so

$$Q = \frac{1-a}{1-\gamma}P_1 + \frac{a-\gamma}{1-\gamma}P_0.$$

Thus, if $\gamma < a$, Q is not irreducible with respect to P_0, while if $\gamma = a$, we obviously have $Q = P_1$. This establishes uniqueness. □

4 Connection to ROC Curves

Given $\alpha \in [0,1]$, define

$$\widetilde{\beta}(\alpha) = \sup\left\{1 - \widetilde{R}_1(h) \,\middle|\, R_0(h) \leq \alpha\right\}.$$

We can think of $\widetilde{\beta}(\alpha)$ as the the receiver operating characteristics (ROC) curve of the most powerful test of size α for

$$\begin{aligned} H_0 &: X \sim P_0 \\ H_1 &: X \sim \widetilde{P}_1 \end{aligned}$$

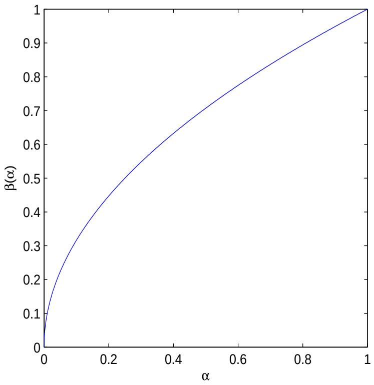

Figure 4: $\widetilde{\beta}(\alpha)$ is an ROC, and when it is concave, κ^* is the slope of this ROC at its right endpoint (note the theory does not assume concavity).

Theorem 2.

$$\kappa^*(\widetilde{P}_1|P_0) = \inf_{\alpha\in[0,1)} \left\{ \frac{1-\widetilde{\beta}(\alpha)}{1-\alpha} \right\}.$$

Proof. Denote $\kappa^* = \kappa^*(\widetilde{P}_1|P_0)$. From Theorem 1, $\forall\epsilon > 0$, we can find $A_\epsilon \in \mathcal{A}$ such that $P_0(A_\epsilon) > 0$ and $\frac{\widetilde{P}_1(A_\epsilon)}{P_0(A_\epsilon)} < \kappa^* + \epsilon$. Define the classifier $h_\epsilon(x) = 1_{\{x \notin A_\epsilon\}}$, and set $\alpha_\epsilon = R_0(h_\epsilon) = 1 - P_0(A_\epsilon) < 1$. Additionally, $\widetilde{R}_1(h_\epsilon) = \widetilde{P}_1(A_\epsilon) = 1 - (1 - \widetilde{P_1}(A_\epsilon)) \geq 1 - \widetilde{\beta}(\alpha_\epsilon)$. Therefore,

$$\inf_{\alpha\in[0,1)} \frac{1-\widetilde{\beta}(\alpha)}{1-\alpha} \leq \frac{1-\widetilde{\beta}(\alpha_\epsilon)}{1-\alpha_\epsilon} \leq \frac{\widetilde{P_1}(A_\epsilon)}{P_0(A_\epsilon)} < \kappa^* + \epsilon$$

ϵ was arbitrary, so

$$\inf_{\alpha\in[0,1)} \frac{1-\beta(\alpha)}{1-\alpha} \leq \kappa^*.$$

On the other hand, from Theorem 1,

$$\begin{aligned} \kappa^* = \inf_{A\in\mathcal{A}:P_0(A)>0} \frac{\widetilde{P_1}(A)}{P_0(A)} = \inf_{\alpha\in[0,1)} \left\{ \inf_{A:P_0(A)\geq 1-\alpha} \frac{\widetilde{P_1}(A)}{P_0(A)} \right\} &\leq \inf_{\alpha\in[0,1)} \left\{ \inf_{A:P_0(A)\geq 1-\alpha} \frac{\widetilde{P_1}(A)}{1-\alpha} \right\} \\ &= \inf_{\alpha\in[0,1)} \frac{1-\widetilde{\beta}(\alpha)}{1-\alpha}. \end{aligned}$$

This completes the proof. □

This result suggests estimating $\kappa^*(\widetilde{P}_1|P_0)$ by estimating the slope of the ROC at its right endpoint, based on an empirical ROC. The next section develops this idea.

5 Consistent Mixture Proportion Estimation

We would like to establish a consistent estimator $\widehat{\kappa}$ of $\kappa^*(\widetilde{P}_1|P_0)$, which will be a consistent estimator of κ in (1) when P_1 is irreducible with respect to P_0.

Let $\mathcal{H}_j$, $j \geq 1$, be VC classes with VC dimensions $V_j < \infty$. For $i = 0, 1$ let $\delta_i > 0$ and denote

$$\epsilon_i\left(j, \delta_i\right) = 2\sqrt{2\frac{V_j \log\left(n_i + 1\right) + \log\left(\frac{2}{\delta_i}\right)}{n_i}}. \tag{4}$$

also define

$$\widehat{\kappa}\left(j, \delta_0, \delta_1\right) = \inf_{h \in \mathcal{H}_j} \frac{\widehat{\widetilde{R}}_1\left(h\right) + \epsilon_1\left(j, \delta_1\right)}{\left(1 - \widehat{R}_0\left(h\right) - \epsilon_0\left(j, \delta_0\right)\right)_+}$$

where $\left(x\right)_+ = \max\left(x, 0\right)$ for any $x \in \mathbb{R}$ and the above ratio is defined to be $+\infty$ if the denominator is 0. Finally, define

$$\widehat{\kappa} = \widehat{\kappa}\left(\delta_0, \delta_1\right) = \inf_{j \geq 1} \widehat{\kappa}\left(j, \delta_0 j^{-2}, \delta_1 j^{-2}\right).$$

The following is assumed about $(\mathcal{H}_j)_{j \geq 1}$.

Assumption 1. *$(\mathcal{H}_j)_{j\geq 1}$ satisfies the following universal approximation property (UAP). For any distribution Q, any positive ϵ and any classifier $h^* : \mathcal{X} \mapsto \{0, 1\}$, there exist $j \geq 1$ and $h \in \mathcal{H}_j$ such that $Q\left(h\left(X\right) \neq h^*\left(X\right)\right) < \epsilon$.*

Theorem 3. *(Universal consistency) Let $\widetilde{P}_1$ and P_0 be arbitrary distributions and consider the estimate $\widehat{\kappa}$ based on iid data drawn according to the one-sided label noise model. Denote $\kappa^* = \kappa^*(\widetilde{P}_1 \,|\, P_0)$.*

1. *(One-sided confidence interval) For any $\delta_0, \delta_1 > 0$,*
$$Pr\left(\widehat{\kappa}\left(\delta_0, \delta_1\right) < \kappa^*\right) \leq 2\left(\delta_0 + \delta_1\right). \tag{5}$$
2. *(Weak consistency). If $\delta_0 = \delta_0\left(n_0\right) = \frac{1}{n_0}$ and $\delta_1 = \delta_1\left(n_1\right) = \frac{1}{n_1}$, then $\widehat{\kappa} \rightarrow \kappa^*$ in probability as $\min\left(n_0, n_1\right) \rightarrow \infty$.*
3. *(Strong consistency). If $\delta_0\left(n_0, n_1\right) = \delta_1\left(n_0, n_1\right) = \left(n_0 n_1\right)^{-2}$, $\log n_0 = o\left(n_1\right)$ and $\log n_1 = o\left(n_0\right)$ as $\min\left(n_0, n_1\right) \rightarrow \infty$, then $\widehat{\kappa} \overset{a.s.}{\rightarrow} \kappa^*$.*

Proof. (**Part** 1). Since $\sum\limits_{j=1}^{\infty} j^{-2} = \frac{\pi^2}{6} < 2$, the VC inequality and union bound imply that with probability at least $1 - 2\left(\delta_0 + \delta_1\right)$, both

$$\forall\, j \geq 1,\ \forall\, h \in \mathcal{H}_j \quad \left|\widehat{R}_0\left(h\right) - R_0\left(h\right)\right| < \epsilon_0\left(j, \delta_0 j^{-2}\right) \tag{6}$$

and

$$\forall\, j \geq 1,\ \forall\, h \in \mathcal{H}_j \quad \left|\widehat{\widetilde{R}}_1\left(h\right) - \widetilde{R}_1\left(h\right)\right| < \epsilon_1\left(j, \delta_1 j^{-2}\right) \tag{7}$$

Thus, with probability at least $1 - 2\left(\delta_0 + \delta_1\right)$,

$$\begin{aligned} \kappa^* &= \inf_{\alpha \in [0,1]} \left(\frac{1 - \widetilde{\beta}\left(\alpha\right)}{1 - \alpha}\right) \overset{(a_0)}{\leq} \inf_{j \geq 1,\, h \in \mathcal{H}_j} \left(\frac{\widetilde{R}_1\left(h\right)}{\left(1 - R_0\left(h\right)\right)_+}\right) \overset{(a_1)}{\leq} \inf_{j \geq 1,\, h \in \mathcal{H}_j} \frac{\hat{\tilde{R}}_1\left(h\right) + \epsilon_1\left(j, \delta_1 j^{-2}\right)}{\left(1 - \hat{R}_0\left(h\right) - \epsilon_0\left(j, \delta_0 j^{-2}\right)\right)_+} \\ &= \widehat{\kappa}\left(\delta_0, \delta_1\right) \end{aligned}$$

where (a_0) holds the infimum is taken over a restricted set and (a_1) is an immediate consequence of deviation inequalities (6) and (7).

(**Part** 2). Let $0 < \epsilon \leq 1$. We previously saw in the proof of Theorem 2 that there exists a classifier h_ϵ such that $R_0\left(h_\epsilon\right) < 1$ and $\left(\frac{\widetilde{R}_1(h)}{1-R_0(h)}\right) \leq \kappa^* + \epsilon$.

By Assumption 1, for any $\gamma \in \left(0, \frac{1}{2}\right)$, there exist $\underline{j} \geq 1$ and $\underline{h} \in \mathcal{H}_{\underline{j}}$ such that

$$\bar{P}\left(h_\epsilon\left(X\right) \neq \underline{h}\left(X\right)\right) < \gamma$$

where $\bar{P} = \frac{1}{2}(P_0 + P_1)$. Since $\bar{P} \geq \frac{1}{2}P_0$ and $\bar{P} \geq \frac{1}{2}P_1$, we have

$$P_0\left(h_\epsilon\left(X\right) \neq \underline{h}\left(X\right)\right) < 2\gamma, \quad P_1\left(h_\epsilon\left(X\right) \neq \underline{h}\left(X\right)\right) < 2\gamma. \tag{8}$$

We claim that $\left|R_i\left(h\right) - R_i\left(h'\right)\right| \leq P_i\left(h_\epsilon\left(X\right) \neq \underline{h}\left(X\right)\right)$ for $i = 1, 2$. To see this, observe

$$\begin{aligned} R_i\left(h\right) - R_i\left(h'\right) &= P_i\left(h\left(X\right) = 1\right) - P_i\left(h'\left(X\right) = 1\right) \\ &= P_i\left(h\left(X\right) = 1,\ h'\left(X\right) = 0\right) - P_i\left(h\left(X\right) = 0,\ h'\left(X\right) = 1\right) \\ &\leq P_i\left(h\left(X\right) = 1,\ h'\left(X\right) = 0\right) \\ &\leq P_i\left(h\left(X\right) \neq h'\left(X\right)\right) \end{aligned}$$

To see the inequality in the reverse direction, just interchange the roles of h and h'.

For brevity, we use ϵ_0 and ϵ_1 instead of $\epsilon_0\left(\underline{j}, \delta_0\underline{j}^{-2}\right)$ and $\epsilon_1\left(\underline{j}, \delta_1\underline{j}^{-2}\right)$. Combining the above, we have that with probability at least $1 - 2\left(\delta_0 + \delta_1\right)$,

$$\widehat{\kappa} \leq \frac{\widehat{\widetilde{R}}_1\left(\underline{h}\right) + \epsilon_1}{\left(1 - \widehat{R}_0\left(\underline{h}\right) - \epsilon_0\right)_+} \leq \frac{\widetilde{R}_1\left(\underline{h}\right) + 2\epsilon_1}{\left(1 - R_0\left(\underline{h}\right) - 2\epsilon_0\right)_+} \leq \frac{\widetilde{R}_1\left(h_\epsilon\right) + 2\left(\epsilon_1 + \gamma\right)}{\left(1 - R_0\left(h_\epsilon\right) - 2\left(\epsilon_0 + \gamma\right)\right)_+}. \tag{9}$$

Note that the first inequality in (9) is a direct consequence of the first part of the theorem, and the last inequality is obtained by the combination of our previous claim and inequality (8). Now choose $\gamma = \frac{\epsilon}{4(1+\epsilon)}\left(1 - R_0\left(h_\epsilon\right)\right)$ and let N be such that $n_0, n_1 \geq N$ implies $\max\left(\epsilon_0, \epsilon_1\right) \leq \gamma$ (possible since ϵ_0 and ϵ_1 tend to zero as n_0 and n_1 go to infinity). Therefore, by using the inequality (9), the following string of inequalities fails with probability at most $2\left(\delta_0 + \delta_1\right)$ for $n_0, n_1 \geq N$:

$$\begin{aligned} \widehat{\kappa} &\leq \frac{\widetilde{R}_1\left(h_\epsilon\right) + 2\left(\epsilon_1 + \gamma\right)}{\left(1 - R_0\left(h_\epsilon\right) - 2\left(\epsilon_0 + \gamma\right)\right)_+} \leq \frac{\widetilde{R}_1\left(h_\epsilon\right) + 4\gamma}{\left(1 - R_0\left(h_\epsilon\right) - 4\gamma\right)_+} = \left(1 + \epsilon\right)\frac{\widetilde{R}_1\left(h_\epsilon\right) + 4\gamma}{1 - R_0\left(h_\epsilon\right)} \\ &= \left(1 + \epsilon\right)\frac{\widetilde{R}_1\left(h_\epsilon\right)}{1 - R_0\left(h_\epsilon\right)} + \frac{4\gamma\left(1 + \epsilon\right)}{1 - R_0\left(h_\epsilon\right)} = \epsilon + \left(1 + \epsilon\right)\frac{\widetilde{R}_1\left(h_\epsilon\right)}{1 - R_0\left(h_\epsilon\right)} \\ &\leq \epsilon + \left(1 + \epsilon\right)\left(\kappa^* + \epsilon\right) \leq \kappa^* + 4\epsilon. \end{aligned}$$

Combining the above bound and that of Part 1, which together hold with probability at least $1 - 2(\delta_0 + \delta_1)$ (because they are based on the same uniform deviation bound), we have that for $n_0, n_1 \geq N$,

$$Pr\left(\left|\kappa^* - \widehat{\kappa}\right| \geq 4\epsilon\right) \leq 2\left(\frac{1}{n_0} + \frac{1}{n_1}\right) \to 0.$$

Thus, $\widehat{\kappa}$ is a weakly consistent estimator of κ^*.

(**Part 3**). In order to show strong consistency, we need the following generalization of the Borel-Cantelli lemma, whose proof is left as an exercise.

Lemma 1. *If Z_{n_0,n_1} is a random process indexed by n_0 and n_1, and for all $\epsilon > 0$ there exists N such that*

$$\sum_{n_0,n_1 \geq N} Pr\left(\left|Z_{n_0,n_1} - Z\right| \geq \epsilon\right) < \infty,$$

then $Z_{n_0,n_1} \to Z$ almost surely as $\min\{n_0, n_1\} \to \infty$.

In the last inequality of Part 2, we have seen that for any $\Delta > 0$, there exists $N > 0$ such that $Pr\left(|\kappa^* - \widehat{\kappa}| \geq \Delta\right) \leq 2\left(\delta_0 + \delta_1\right) = 4\left(n_0 n_1\right)^{-2}$ for any $n_0, n_1 \geq N$. Hence,

$$\sum_{n_0,n_1 \geq N} Pr\left(|\kappa^* - \widehat{\kappa}| \geq \Delta\right) \leq 4 \sum_{n_0,n_1 \geq N} \left(n_0 n_1\right)^{-2} \leq 4\left(\frac{\pi^2}{6}\right)^2 < \infty$$

The result now follows from Lemma 1.

Note the following subtle point: In the argument from Part 2, we need both ϵ_0 and ϵ_1 to tend to zero as $\min\{n_0, n_1\} \to \infty$. In the strong consistency case, where we choose $\delta_0 = \delta_1 = (n_0 n_1)^{-2}$, we must have $\log n_0 = o\left(n_1\right)$ and $\log n_1 = o\left(n_0\right)$ to guarantee this property. For instance,

$$\epsilon_0\left(j, \delta_0 j^{-2}\right) = 3\sqrt{\frac{V_j \log\left(n_0 + 1\right) + 2\log n_0 + 2\log n_1 + \log 2}{n_1}} \to 0$$

provided, $\log n_0 = o\left(n_1\right)$. □

Exercises

1. Prove Proposition 1 and use it to construct a consistent discriminiation rule with respect to the probability of error performance measure.
2. Verify the second remark after Theorem 1
3. Verify the second example (about Gaussian densities) using the likelihood ratio characterization of κ^*.
4. Verify the generalization of the Borel-Cantelli Lemma used in the proof of Theorem 3.
5. Instead of defining $\widehat{\kappa}(\delta_0, \delta_1)$ as an inf over j of $\widehat{\kappa}(j, \delta_0 j^{-2}, \delta_1 j^{-2})$, does the consistency result still hold if we define
$$\widehat{\kappa}(\delta_0, \delta_1) = \widehat{\kappa}(j(n_0, n_1), \delta_0, \delta_1)$$
where $j(n_0, n_1) \to \infty$ at an appropriate rate? If so, does one estimator have any advantage over the other?

References

[1] G. Blanchard, G. Lee, and C. Scott, "Semi-supervised novelty detection," *Journal of Machine Learning Research*, vol. 11, pp. 2973-3009, Nov. 2010.

[2] C. Scott, G. Blanchard, G. Handy, "Classification with Asymmetric Label Noise: Consistency and Maximal Denoising," *Proc. Conf. on Learning Theory* (COLT), JMLR W&CP 30:489-511, 2013.

[3] G. Blanchard and C. Scott, "Decontamination of Mutually Contaminated Models," *Proc. Artifical Intelligence and Statistics* (AISTATS), JMLR W&CP 33 :1-9, 2014

[4] T. Sanderson and C. Scott, "Class Proportion Estimation with Application to Multiclass Anomaly Rejection," *Proc. Artifical Intelligence and Statistics* (AISTATS), JMLR W&CP 33 :850-858, 2014.