Theorem[lecnum] [theorem]Lemma [theorem]Proposition [theorem]Corollary [theorem]Definition

CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 8: September 28

Lecturer: Csaba Szepesvári *Scribes: Tian Tian*

8.1 Recap

At the end of the last class, we talked about a variance condition, which we recall as follows. Let Z be a set and P a distribution over Z. Let $\mathcal{G}$ be a set of measurable functions that maps Z to $\mathbb{R}$ (denote as $\mathcal{M}(Z, R)$). For $c_0, c_1 > 0$, we define

$$Var_Z(c_0, c_1, P) = \{g \in \mathcal{M}(Z, R) : Var_P(g) \leq c_0^2 + c_1 Pg\}, \tag{8.1}$$

where $Pg = \int g dP$ and $Var_P(g) = \int (g - Pg)^2 dP$.

The reason why we introduce such variance condition is that we can apply Bernstein inequality and obtain a tight uniform convergence rate. Recall the Bernstein inequality has a variance term in it, and if the function class satisfies the above variance condition, then the variance can be upper-bounded in terms of the integral of the expected value of the function. To see more examples of problem where such variance condition is satisfied, we begin today's lecture with binary classification.

In PAC learning, we got fast rate for binary classification under no misspecification and no noise. What happens when noise is present? Would noise make sample complexity that much worse? The suspicion is that we should still see fast rate of $O(1/\varepsilon)$. In binary classification, one only needs to decide whether if the label is 1 or 0 based on whether the mean of the Bernoulli random variable is above or below 0.5. The further the mean is from 0.5, one would be more definitive in the decision of 1 or 0. We expect the misclassification error decreases the further the mean is from 0.5. We shall see in this lecture, why our suspicion could be true.

We begin this lecture by introducing another class of noise called the Tsybakov noise condition. Then we define what it is, how the binary classification problem satisfies this Tsybakov noise condition, and how it all relates to the variance condition **??**. Finally in the last section, we show how the variance condition can be used with Bernstein inequality to obtaining a uniform convergence. From the bound we get, we can see how the binary classification could also achieve a fast rate of $O(1/\varepsilon)$.

8.2 Tsybakov noise condition

Consider the binary classification problem on $\mathcal{X} \times \{0, 1\}$. Let P be a distribution on $\mathcal{X} \times \{0, 1\}$ and $(X, Y) \sim P$.

Definition 8.1. We say P satisfies the Tsybakov's noise condition (i.e., $P \in Tsyb_{\mathcal{X}}(c, \varepsilon_0, \beta)$) if there exists $\beta \in (0, 1], c > 0, \varepsilon_0 \in (0, 0.5]$ such that

$$\mathbb{P}(|\mathbb{P}(Y = 1|X) - 0.5| \leq \varepsilon) \leq c\varepsilon^{\frac{\beta}{1-\beta}}, \quad \text{for all } \varepsilon \in [0, \varepsilon_0]. \tag{8.2}$$

Note that $\beta \in (0, 1]$ and $\beta/1 - \beta$ goes from zero is infinity in a monotonous fashion. If $\beta = 1$, $\mathbb{P}(|\mathbb{P}(Y = 1|X) - 0.5| \leq \varepsilon)$ is interpreted as 0. This is the case where there is no probability mass in the

region where $\mathbb{P}(Y=1|X)$ is ε_0 close to 0.5. In other words, $|\mathbb{P}(Y=1|X)-0.5| \geq \varepsilon_0$ for all X. Since we want to learn P, Tsybakov noise condition characterizes how hard is going to be. The hardness measure is β. If β is closer to 1 means that P will be easier to learn as oppose to if β is closer to 0.

How does all this relate to the variance condition **??**? For ease of writing, let $\eta_P(x) = \mathbb{P}(Y=1|X=x)$. The optimal decision at P, $f_P(x) = \mathbb{I}\{\eta_P(x) \geq 0.5\}$. The binary function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$. For a $f \in \mathcal{F}$, we define the loss $g_f : \mathcal{X} \times \{0,1\} \to \{0,1\}$ by $g_f(x,y) = \mathbb{I}\{f(x) \neq y\}$. Then the loss set

$$\mathcal{G} = \ell_{01} \circ \mathcal{F} = \{g_f : f \in \mathcal{F}\}, \tag{8.3}$$

and the shifted loss set

$$\tilde{\mathcal{G}} = \mathcal{G} - \{g_{f_P}\}. \tag{8.4}$$

Claim 8.2. If $P \in Tsyb_{\mathcal{X}}(c, \varepsilon_0, \beta)$, then there exists $c' > 0$ s.t. for all $\tilde{g} \in \tilde{\mathcal{G}}$:

$$P\tilde{g}^2 \leq (c')^{2-\beta}(P\tilde{g})^{\beta}. \tag{8.5}$$

Following from the claim,

1. $\beta = 1, \tilde{\mathcal{G}} \subseteq Var_{\mathcal{X}\times\{0,1\}}(0, c', P)$,

2. $\beta < 1$, for all $\gamma > 0$, $\tilde{\mathcal{G}} \subseteq Var_{\mathcal{X}\times\{0,1\}}((1-\beta)^{0.5}\gamma^{\frac{0.5}{1-\beta}}(c'), \beta c'\gamma^{-\frac{1}{\beta}}, P)$,

3. $\mathcal{X} = [0,1], Y = \{0,1\}, (X,Y) \sim P \in M_1(\mathcal{X} \times \{0,1\})$

$$\mathbb{P}(Y=1|X=x) = \begin{cases} p & x \in X_0, \\ 1-p & x \in X_0^c, \end{cases} \tag{8.6}$$

where $p \in [0, 0.5)$ is the uniform noise and $X_0 \subseteq \mathcal{X}$. Intuitively, for some subset of the input, we have 1 with probability p and for all other input, we have 0 with probability $1-p$. Pick an ε and consider the lower-bracket cover $\tilde{\mathcal{G}}(\varepsilon)$ of $\tilde{\mathcal{G}}$, then there exist $c_0(p), c_1(p) > 0$ s.t. $\tilde{\mathcal{G}}(\varepsilon) \in Var_{\mathcal{X}\times[0,1]}(\sqrt{\varepsilon c_0(p)}, c_1(p), P)$.

How do we use all this with the Bernstein inequality to get a tighter bound on the losses? Before we state the main theorem, recall Bernstein inequality.

8.3 Bernstein inequality

We use $Ber(b, V)$ to denote the Bernstein condition with parameter b and variance V. For any random variable X that satisfies the Bernstein condition (i.e., $X \in Ber(b, V)$), then w.p. $1-\delta$,

$$\bar{X}_n < \mu + \sqrt{\frac{2V\ln(1/\delta)}{n}} + \frac{b\ln(1/\delta)}{3n}, \tag{8.7}$$

$$\bar{X}_n \geq \mu - \sqrt{\frac{2V\ln(1/\delta)}{n}} - \frac{b\ln(1/\delta)}{3n} \tag{8.8}$$

8.4 Uniform Bernstein

Theorem 8.3. *Fix $\varepsilon_0, b, c_0, c_1 > 0$, $G \subseteq \mathbb{R}^Z$, $0 \leq \varepsilon \leq \varepsilon_0$, $P \in M_1(Z)$. Let Z be a random variable whose distribution is P, and assume*

1. *for all* $g \in \mathcal{G}(\varepsilon), g(z) \in Ber(b, Var_P g)$
2. *fix* ε*, take a lower bracketting cover* $\mathcal{G}(\varepsilon)$ *for* $\mathcal{G}$*.* $\mathcal{G}(\varepsilon) \subseteq Var_Z(c_0, c_1, P)$*,*

then w.p. $1-\delta$*, for all* $g \in G$

$$Pg - P_n g \leq \sqrt{\frac{2c_0^2 \ln(N_\varepsilon/\delta)}{n}} + \varepsilon + \sqrt{\frac{2c_1(Pg)_+ \ln(N_\varepsilon/\delta)}{n}} + \frac{c_0 \ln(N_\varepsilon/\delta)}{3n}, \tag{8.9}$$

where $N_\varepsilon = |G(\varepsilon)|$*.*

Proof. Pick $g \in \mathcal{G}$, there exit $j \in [N_\varepsilon]$ such that

$$g_j \leq g, \quad Pg \leq Pg_j + \varepsilon. \tag{8.10}$$

Then, it follows that

$$Pg - P_n g \leq Pg_j - P_n g_j + \varepsilon. \tag{8.11}$$

We want to apply Bernstein to every element of the cover. That is, with probability $1-\delta$, for all $i \in [N_\varepsilon]$,

$$Pg_i - P_n g_i \leq \sqrt{\frac{2V_i \ln(N_\varepsilon/\delta)}{n}} + \frac{b \ln(N_\varepsilon/\delta)}{3n}. \tag{8.12}$$

Recall $V_i = Var_p(g_i)$, since $\mathcal{G}(\varepsilon) \subseteq Var_Z(c_0, c_1, P)$, this means $V_i = Var_p(g_i) \leq c_0^2 + c_1(Pg_i)_+$. Then it follows that

$$Pg_i - P_n g_i \leq \sqrt{\frac{2c_0^2 \ln(N_\varepsilon/\delta)}{n}} + \sqrt{\frac{2c_1(Pg_i)_+ \ln(N_\varepsilon/\delta)}{n}} + \frac{b \ln(N_\varepsilon/\delta)}{3n}. \tag{8.13}$$

The inequality in **??** uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. The **??** holds for all elements of the cover, and thus it holds for $Pg_j - P_n g_j$. Since $(Pg_j)_+ \leq (Pg)_+$, then we get the result. □

Note if we could make c_0 in the order of $1/n$, then we get multiplicative Chernoff. Then it all boils down to what cases is c_0 in the order of $1/n$. If so, then we get $O(1/\varepsilon)$ sample complexity, and this will be the case with squared loss that we mentioned in the last lecture as another example of problem that satisfies the variance condition **??**. Similarly, we get a $O(1/\varepsilon)$ sample complexity for the binary classification case with Tsybokov noise condition when $\beta = 1$. In general, we get a worse rate between $1/\varepsilon$ and $1/\varepsilon^2$ depending on β.