When can unlabeled data improve the learning rate?

COLT 2019

Notation

$$R_P(f) := P(f(X) \neq Y)$$

$$\eta(x) = P(Y = 1 | X = x)$$

$$R_{P,\mathcal{H}} := \inf_{h \in \mathcal{H}} R_P(h)$$

$$P^{(\ell,u)} := P^{\ell} \times P_X^u \qquad A : (\mathcal{X} \times \mathcal{Y})^{\ell} \times \mathcal{X}^u \to \mathcal{Y}^{\mathcal{X}}$$

Learning Rates

Definition 1 (Minimax expected risk) *The* minimax expected excess risk *of learning a hypothesis class $\mathcal{H}$ on a sample of size (ℓ, u) over the set of admissible distributions $\mathcal{P}$ is the expected excess risk of the best algorithm A under the worst distribution P:*

$$L(\ell, u, \mathcal{H}, \mathcal{P}) := \inf_{A} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^{(\ell,u)}} \left[\left(R(A(S)) - R_{P,\mathcal{H}} \right)_+ \right] . \tag{1}$$

Definition 2 (SL and SSL learnability) *We say that a problem $(\mathcal{H}, \mathcal{P})$ is* SL learnable *if $L(\ell, 0, \mathcal{H}, \mathcal{P})$ converges to zero as ℓ goes to infinity. We say that a problem $(\mathcal{H}, \mathcal{P})$ is* SSL learnable *if, for some function $u : \mathbb{N} \to \mathbb{N}$, $L(\ell, u(\ell), \mathcal{H}, \mathcal{P})$ converges to zero as ℓ goes to infinity.*

$$\textit{if there exists } c_1, c_2 \in \boldsymbol{R}^{>0}, c_1 g(l) \leq f(l) \leq c_2 g(l), \textit{we say } f \textit{ has rate } g$$

Concepts of unlabeled data helping

Definition 3 (Unlabeled data helps) *We say that* unlabeled data helps *to learn* $\mathcal{H}$ *over the set of admissible distributions* $\mathcal{P}$ *if there exists some* $u : \mathbb{N} \to \mathbb{N}$ *such that*

$$\lim_{\ell \to \infty} \frac{\inf_{A_{SSL}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^{(\ell, u(\ell))}} \left[\left(R(A_{SSL}(S)) - R_{\mathcal{H}} \right)_+ \right]}{\inf_{A_{SL}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^{\ell}} \left[\left(R(A_{SL}(S)) - R_{\mathcal{H}} \right)_+ \right]} = 0 \,. \tag{2}$$

Definition 4 (Unlabeled data helps non-uniformly) *We say that* unlabeled data helps non-uniformly *to learn* $\mathcal{H}$ *over the sequence of distributions* $(\mathcal{P}_\ell)_{\ell \in \mathbb{N}}$ *if there exists some* $u : \mathbb{N} \to \mathbb{N}$ *such that*

$$\lim_{\ell \to \infty} \frac{\inf_{A_{SSL}} \sup_{P \in \mathcal{P}_\ell} \mathbb{E}_{S \sim P^{(\ell, u(\ell))}} \left[\left(R(A_{SSL}(S)) - R_{\mathcal{H}} \right)_+ \right]}{\inf_{A_{SL}} \sup_{P \in \mathcal{P}_\ell} \mathbb{E}_{S \sim P^{\ell}} \left[\left(R(A_{SL}(S)) - R_{\mathcal{H}} \right)_+ \right]} = 0 \,. \tag{3}$$

Concepts of unlabeled data helping

Definition 5 (Unlabeled data helps weakly non-uniformly) *We say that* unlabeled data helps weakly non-uniformly *to learn* $\mathcal{H}$ *over the sequence of distributions* $(\mathcal{P}_\ell)_{\ell\in\mathbb{N}}$ *with* $\mathcal{P}_\ell \subseteq \mathcal{P}_{\ell+1}$ *if there exists some* $u : \mathbb{N} \to \mathbb{N}$ *such that*

$$\lim_{\ell\to\infty} \frac{\inf_{A_{SSL}} \sup_{P\in\mathcal{P}_\ell} \mathbb{E}_{S\sim P^{(\ell,u(\ell))}} \left[(R(A_{SSL}(S)) - R_{\mathcal{H}})_+\right]}{\inf_{A_{SL}} \sup_{P\in\bigcup_{i\in\mathbb{N}}\mathcal{P}_i} \mathbb{E}_{S\sim P^\ell} \left[(R(A_{SL}(S)) - R_{\mathcal{H}})_+\right]} = 0\,. \tag{4}$$

Definition 6 (Knowing the marginal helps) *We say that* knowing the marginal helps *to learn* $\mathcal{H}$ *over the set of admissible distributions* $\mathcal{P}$ *if*

$$\lim_{\ell\to\infty} \frac{\inf_{A_{SSL}} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^\ell} \left[(R(A_{SSL}(S, P_X)) - R_{\mathcal{H}})_+\right]}{\inf_{A_{SL}} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^\ell} \left[(R(A_{SL}(S)) - R_{\mathcal{H}})_+\right]} = 0\,. \tag{5}$$

It has been shown e.g. by Seeger (2000) that, if the parameters determining marginal and labeling are independent, unlabeled data is not useful in estimating the parameters that determine the labeling, it may still be helpful in finding a low-risk classifier.

No free learnability

Theorem 7 *If a problem $(\mathcal{H}, \mathcal{P})$ is unlearnable in the SL setting, i.e. $L(\ell, 0, \mathcal{H}, \mathcal{P})$ does not converge to 0, then it is also unlearnable in the SSL setting, i.e. for any $u : \mathbb{N} \to \mathbb{N}$, $L(\ell, u(\ell), \mathcal{H}, \mathcal{P})$ does not converge to 0.*

Proof To avoid cluttered notation, we here omit $\mathcal{H}$ and $\mathcal{P}$ from the error rates. We prove that if there is some u such that $\lim L(\ell, u(\ell)) = 0$ then $\lim L(\ell, 0) = 0$. Indeed, $L(\ell, 0) \geq 0$ by definition and for any ℓ and u, $L(\ell+u(\ell), 0) \leq L(\ell, u(\ell))$, since an SL algorithm can simply opt to forget the labels of $u(\ell)$ labeled samples and treat them as unlabeled. Furthermore, $L(\ell, 0)$ is non-increasing since an algorithm receiving an $\ell + 1$ sample can always ignore an example, hence $L(\ell, 0) \leq L(\ell + u(\ell), 0) \leq L(\ell, u(\ell))$. ■

Why we relate the labeling and the marginal

Definition 8 *We say that a family of probability distributions $\mathcal{P}$ is* rich for a class $\mathcal{H}$ *if there exist hypotheses $h, h' \in \mathcal{H}$ and marginal P_X such that $P_X(\{x : h(x) \neq h'(x), h(x) = 0) \neq P_X(\{x : h(x) \neq h'(x), h(x) = 1\})$, and for every $\alpha \in (0, \frac{1}{2})$, $\mathcal{P}$ contains P_α and $P_{-\alpha}$ which consist of P_X paired with labeling functions η_α and $\eta_{-\alpha}$, that agree with h where $h = h'$ and take values $\frac{1}{2} + \alpha$ and $\frac{1}{2} - \alpha$ respectively where $h \neq h'$.*

Theorem 9 *Let $\mathcal{H}$ be a class of finite VC dimension. Then for every set of probability distributions $\mathcal{P}$ that is rich for $\mathcal{H}$, knowing the marginal does not help to learn $\mathcal{H}$ over the set of admissible distributions $\mathcal{P}$.*

Since $\mathcal{H}$ has finite VC dimension, the SL rate of $(\mathcal{H}, \mathcal{P})$ is upper-bounded by $\frac{1}{\sqrt{\ell}}$. Showing that the rate of any algorithm with access to the marginal distribution is lower-bounded by $\frac{1}{\sqrt{\ell}}$ proves that both the SL and SSL rates are of order $\frac{1}{\sqrt{\ell}}$, since the SSL rate cannot be slower than the SL rate. Let $C := \{x : h(x) \neq h'(x)\}$, let $c := P_X(C)$ and $c' := P_X(\{x : x \in C \wedge h(x) = 1\})$. By requirement, $c' \neq c/2$. Without loss of generality, assume $c' > c/2$. A simple calculation shows that

$$\mathrm{KL}(P_\alpha^\ell, P_{-\alpha}^\ell) = 2c\ell\alpha \log\left(\frac{1+2\alpha}{1-2\alpha}\right) . \tag{11}$$

Using

$$\frac{1+x}{1-x} = 1 + \frac{2x}{1-x} \qquad \text{and} \qquad \log(1+x) < x \;\forall x > 0\,, \tag{12}$$

we find that

$$\mathrm{KL}(P_\alpha^\ell, P_{-\alpha}^\ell) \leq 2c\ell\alpha \frac{4\alpha}{1-2\alpha} . \tag{13}$$

For $\alpha < \frac{1}{4}$, this can be bounded by

$$\mathrm{KL}(P_\alpha^\ell, P_{-\alpha}^\ell) \leq 16c\ell\alpha^2 \,. \tag{14}$$

(Tsybakov, 2009, Theorem 2.2, (i)) shows that for any hypothesis test that decides between P_α and $P_{-\alpha}$, the probability of choosing incorrectly is lower-bounded by $\frac{1-a}{2}$, where $a \geq \mathrm{TV}(P_\alpha, P_{-\alpha})$ and TV denotes the total variation distance. Since

$$\mathrm{TV}(P_\alpha^\ell, P_{-\alpha}^\ell) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(P_\alpha^\ell, P_{-\alpha}^\ell)}\,, \tag{15}$$

the probability of a test choosing incorrectly between P_α and $P_{-\alpha}$ can be lower-bounded by

$$\frac{1 - \sqrt{8c\ell\alpha^2}}{2}\,. \tag{16}$$

Since P_α and $P_{-\alpha}$ have the same marginal distribution, this probability is independent of whether or not the marginal is known to the learner. If P_α is the true underlying distribution, any hypothesis that does not majorly label C with 1 incurs an excess risk of at least $2\alpha(c' - c/2)$. Likewise, if $P_{-\alpha}$ is the true underlying distribution, any hypothesis that does not majorly label C with 0 incurs an excess risk of at least $2\alpha(c' - c/2)$. Let $\alpha = \frac{1}{\sqrt{32\ell c}}$. Then Equation (16) shows that the probability of choosing incorrectly between P_α and $P_{-\alpha}$ can be lower-bounded by $\frac{1}{4}$, and the expected excess risk of any algorithm can be lower-bounded by

$$\frac{2c' - c}{16\sqrt{2c}} \frac{1}{\sqrt{\ell}} \asymp \frac{1}{\sqrt{\ell}} \,. \tag{17}$$

a helpful and a non-helpful example

Example 1 *Let $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{H} = \{0,1\}^{\mathcal{X}}$. Then every marginal distribution P_X on $\mathcal{X}$ can be parameterized by $\beta \in (-\frac{1}{2}, \frac{1}{2})$ with $P_X^\beta(x_1) = \frac{1}{2} + \beta$, $P_X^\beta(x_2) = \frac{1}{2} - \beta$. Now, for each P_X^β, restrict $\mathcal{P}$ to contain only those P, denoted by $P^{\alpha\beta}$, such that $P(Y = 1|x_1) = \frac{1}{2} + \alpha = P(Y = 0|x_2)$ with $\alpha\beta > 0$, i.e. restrict the possible labelings such that the Bayes classifier assigns opposite labels to the two points, and labels with 1 the point that is seen more often.*

In Example 1, the Bayes classifier is completely determined by the marginal distribution. As such, this is an example where we can observe improvements via idealistic SSL, i.e. knowing the marginal helps (Definition 6): an SSL algorithm that knows the marginal has expected excess risk zero, while the SL rate of learning $(\mathcal{H}, \mathcal{P})$ is $\frac{1}{\sqrt{\ell}}$.

Example 3 *Let $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{H} = \{0,1\}^{\mathcal{X}}$. Now, restrict $\mathcal{P}$ to those distributions such that $P(Y = 1|x_i) = P_X(x_i)$. That is, the Bayes classifier labels x_1 and x_2 with opposite labels, the noise of each labeling is equal to the noise in choosing x_i, and the point that is less likely to be seen is the one which the Bayes classifier labels with 0.*

Let A be the algorithm that disregards all labels, assigns 1 to the x_i that appears more often in the sample and 0 to the x_i that appears less often. The expected excess loss of this algorithm is bounded by $\frac{1}{\sqrt{\ell+u(\ell)}}$, so the SSL minimax rate for $u(\ell) = \ell^2$ is at least $f(\ell) = \frac{1}{\ell}$, the SSL minimax rate for $u(\ell) = \ell^4$ is at least $f(\ell) = \frac{1}{\ell^2}$ and the SSL minimax rate for $u(\ell) = \exp(\ell)$ is exponential. The SL rate of $(\mathcal{H}, \mathcal{P})$ is $\frac{1}{\sqrt{\ell}}$.[5] Proofs of lower and upper bounds can be found in Appendix B.

How much unlabeled data is enough?

Proposition 11 *If the SL rate of* $(\mathcal{H},|\mathcal{P})$ *is* $\ell^{-\alpha}$ *with* $\alpha > 0$, *and unlabeled data helps for* $u : \mathbb{N} \to \mathbb{N}$, *then* $\frac{u(\ell)}{\ell} \to \infty$, *i.e.* u *must grow superlinearly.*

Proof Since $L(\ell+u(\ell),0) < L(\ell,u(\ell))$, $\frac{L(\ell,u(\ell))}{L(\ell,0)} \to 0$ implies that $\frac{L(\ell+u(\ell),0)}{L(\ell,0)} \to 0$. Applying these conditions to $L(\ell,0) = \ell^{-\alpha}$ yields the result. ■