Learning from positive and unlabeled examples - Finite size sample bounds

Farnam Mansouri **Shai Ben-David**

University of Waterloo and Vector Institute

PU Learning

Learning from positive and unlabeled data (PU learning) is a variant of the classical machine learning where the training data consists of positive and unlabeled examples.

Application 1: Personalized advertising [BD20]

Positive examples: visited pages and clicks.
Unlabeled examples: all other pages (shouldn't be labeled negative).
Note that positive and unlabeled examples are not independently drawn.
We call such scenarios *training-set scenario*.

Application 2: Predicting users of mobile application

Positive examples: from individuals who are already users of an application.
Unlabeled examples: from a random selection of users.
Note that positive and unlabeled examples are independently drawn.
We call such scenarios *control-set scenario*. This poster entirely focuses on **case-control scenarios**.

Let $\mathcal{D}$ be an unknown distribution over the domain $\mathcal{X}$. Denote $\mathcal{D}_\mathcal{X}$ as the marginal distribution and $\mathcal{D}^P_\mathcal{X}(x) = \mathcal{D}_\mathcal{X}(x \mid y = 1)$ as the positive distribution. A **PU learner** has access to

Positive Sample S^P i.i.d. drawn over $\tilde{\mathcal{D}}^P_\mathcal{X}$
Unlabeled sample S^U i.i.d drawn over $\mathcal{D}_\mathcal{X}$

Target of PU learner is to find a concept $c : \mathcal{X} \to \{0, 1\}$ that minimizes $\mathrm{err}_\mathcal{D}(c) := Pr_{(x,y)\sim\mathcal{D}}[c(x) \neq y]$.

Class of $\tilde{\mathcal{D}}^P_\mathcal{X}$	Description
SCAR	$\tilde{\mathcal{D}}^P_\mathcal{X} = \mathcal{D}^P_\mathcal{X}$
SAR	$\tilde{\mathcal{D}}^P_\mathcal{X}(x) \sim \mathcal{D}^P_\mathcal{X}(x)e(x)$, where e is the *individual propensity score*
PCS	$\tilde{\mathcal{D}}^P_\mathcal{X}$ has zero weight in an area with all negative labels
APDS	any arbitrary $\tilde{\mathcal{D}}^P_\mathcal{X}$

Overview of results

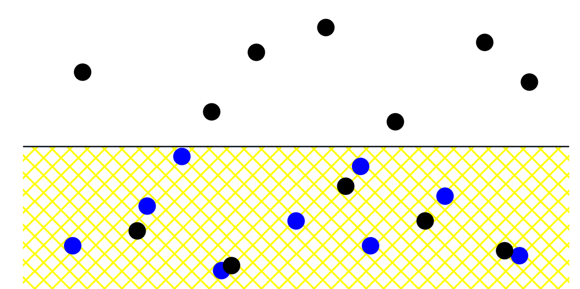
Setting	Upper Bound	Lower Bound	Source
Realizable (SCAR)	$\lvert S^P\rvert = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon}\right)$ $\lvert S^U\rvert = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon}\right)$	$\lvert S^P\rvert = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon}\right)$ $\lvert S^U\rvert = \tilde{\Omega}\left(\frac{\mathrm{CLAW}(\mathcal{C})}{\varepsilon}\right)$	[Liu+02], New
Realizable (SAR)	$\lvert S^P\rvert = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{r\varepsilon}\right)$ $\lvert S^U\rvert = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon}\right)$	$\lvert S^P\rvert = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{r\varepsilon}\right)$ $\lvert S^U\rvert = \tilde{\Omega}\left(\frac{\mathrm{CLAW}(\mathcal{C})}{\varepsilon}\right)$	New
Realizable (PCS)	$\lvert S^P\rvert = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{r^2\varepsilon}\right)$ $\lvert S^U\rvert = \tilde{O}\left(\frac{\left(\frac{\sqrt{k}}{\gamma}\right)^k + \alpha\,\mathrm{VCD}(\mathcal{C})}{\alpha\,\varepsilon}\right)$	$\lvert S^P\rvert + \lvert S^U\rvert = \tilde{\Omega}\left((1 + \frac{1}{2\gamma})^{k/2}\right)$	New
Agnostic (SCAR, known α)	$\lvert S^P\rvert = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$ $\lvert S^U\rvert = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$	$\lvert S^P\rvert = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$ $\lvert S^U\rvert = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$	[DPNS15], New

$\alpha := \Pr[y = 1]$; k is dimensionality of the space; γ is the margin parameter; and r is a weight ratio between $\tilde{\mathcal{D}}^P_\mathcal{X}$ and $\mathcal{D}^P_\mathcal{X}$.

Realizable (SCAR)

Definition 1. *We define* claw number *of concept class $\mathcal{C}$ denoted by* $\mathrm{CLAW}(\mathcal{C})$ *to be the largest $\mathfrak{h} \in \mathbb{N}$ such that for every $m \geq \mathfrak{h}$, there exists a $B \subseteq \mathcal{X}$ with $|B| = m$ such that $\{O \subseteq B \mid |O| = m - \mathfrak{h}\} \subseteq \mathcal{C} \mid B$. If no such $\mathfrak{h}$ exists, we say the claw number of $\mathcal{C}$ is 0.*

Postive Empirical Risk Minimizer (PERM): $\mathrm{argmin}_{c\in\mathcal{C}, S^P\subseteq c} |c \cap S^U|$



Demonstration of PERM. Blue points represent positive examples and Black ones represent unlabeled ones. The cross hatch part represents $c(x) = 1$ and black points in crosshatched parts represent $c \cap S^U$.

Theorem 1. *(informal) Consider concept class $\mathcal{C}$ over domain $\mathcal{X}$. In the Realizable SCAR case (i) When $|S^P|, |S^U| = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon}\right)$, with probability $1 - \delta$ error of PERM algorithm is at most ε. (ii) No algorithm can achieve error less than ε, with a probability more than $1-\delta$ if $|S^P| = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon}\right)$ or $|S^U| = \tilde{O}\left(\frac{\mathrm{CLAW}(\mathcal{C})}{\varepsilon}\right)$.*

Realizable (PCS)

Algorithm $\mathcal{A}$ we introduce next, is the same algorithm introduced by [BDU12] for learning with distribution shift adapted to PU learning.

Algorithm 1: Algorithm $\mathcal{A}$

Input: S^P i.i.d. sampled from $\tilde{\mathcal{D}}^P_\mathcal{X}$ with label 1 and an unlabeled i.i.d. sample S^U from $\mathcal{D}_\mathcal{X}$ and a margin parameter γ.

- Partition the domain $[0, 1]^k$ into a collection $\mathcal{B}$ of boxes (axis-aligned rectangles) with sidelength $(\gamma/\sqrt{k})$. ;
- Obtain sample S' by removing every point in S^P, which is sitting in a box that is not hit by S^U ;
- **return** $\mathrm{argmin}_{c\in\mathcal{C}, X(S')\subseteq c} |c \mid S^U|$

Definition 2. *(informal) For distributions $\mathcal{Q}_1$, $\mathcal{Q}_2$ over $\mathcal{X}$ and $\mathcal{B} \subseteq 2^X$, we define the weight ratio of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ with respect to $\mathcal{B}$ as*

$$R_\mathcal{B}(\mathcal{Q}_1, \mathcal{Q}_2) = \inf_{\substack{A\in\mathcal{B} \\ \mathcal{Q}_2(A)\neq 0}} \frac{\mathcal{Q}_1(A)}{\mathcal{Q}_2(A)}.$$

Theorem 2. *(informal) Consider concept class $\mathcal{C}$ over domain $\mathcal{X} = [0, 1]^k, \gamma > 0$ a margin parameter and labeling be deterministic. Suppose (i) $\mathcal{D}_\mathcal{X}$ is realizable by $\mathcal{C}$ with margin γ (ii) there is a deterministic labeling function l that is γ-margin classifier with respect to $\mathcal{D}_\mathcal{X}$ (not formally defined). In the PCS case, when $|S^P| = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon R_\mathcal{I}(\tilde{\mathcal{D}}^P_\mathcal{X}, \mathcal{D}^P_\mathcal{X})^2}\right)$ and $|S^U| = \tilde{\Omega}\left(\frac{\left(\sqrt{k}/\gamma\right)^k + \alpha\,\mathrm{VCD}(\mathcal{C})}{\varepsilon\alpha}\right)$ where $\mathcal{I} = (\mathcal{C}\Delta\mathcal{C}) \sqcap \mathcal{B}$, then algorithm $\mathcal{A}$ outputs a classifier c with $\mathrm{err}_\mathcal{D}(c) \leq \varepsilon$ with probability at least $1 - \delta$.*

Theorem 3. *(informal) Consider any finite domain $\mathcal{X}$. There exists a concept class $\mathcal{C}_{0,1}$ with $\mathrm{VCD}(\mathcal{C}_{0,1}) = 1$ such that for the class of realizable distributions $\mathcal{D}$ and $\tilde{\mathcal{D}}^P_\mathcal{X}$ with positive covariate shift with bounded weight ratio no algorithm can achieve error less than ε with probability $1 - \delta$ unless $|S^P| + |S^U| = \Omega(\sqrt{|\mathcal{X}|})$.*

Agnostic (SCAR)

Theorem 4. *(informal) Let $\mathcal{C}$ be a concept class over $\mathcal{X}$. In the Agnostic SCAR case with known α, (i) When $|S^P|, |S^U| = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$ there is an algorithm achieving error ε with probability $1 - \delta$. (ii) If $|S^P| = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$ or $|S^U| = \tilde{O}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$ no algorithm achieves error less than ε with probability $1 - \delta$.*

Theorem 5. *(informal) Let $\mathcal{C}$ be a concept class over $\mathcal{X}$ containing at least two distinct concepts. Then, for every $\eta \in (0, 1)$ and any PU learner $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ with $\alpha \in \{\eta, 1 - \eta\}$, where $\alpha := \Pr(y = 1)$, such that for every positive sample S^P and unlabeled sample S^U satisfies*

$$\mathrm{err}_\mathcal{D}\left(\mathcal{A}(S^P, S^U)\right) \geq \frac{\max(\alpha, 1 - \alpha)}{\min(\alpha, 1 - \alpha)} \min_{c\in\mathcal{C}} \mathrm{err}_\mathcal{D}(c)$$

Theorem 6. *(informal) Consider concept class $\mathcal{C}$ over domain $\mathcal{X}$. Let $c^\gamma = \mathrm{argmin}_{c\in C} \frac{|c|S^U|}{|S^U|} + \gamma\frac{|S^P|-|c|S^P|}{|S^P|}$. In the Agnostic SCAR case, when $|S^P|, |S^U| = \tilde{\Omega}\left(\frac{\mathrm{VCD}(\mathcal{C})}{\varepsilon^2}\right)$ and $\gamma \geq \alpha$ then with probability $1 - \delta$ we have*

$$\mathrm{err}_\mathcal{D}(c^\gamma) \leq \max\left(\frac{\gamma - \alpha}{\alpha}, \frac{\alpha}{\gamma - \alpha}\right)\left(\min_{c\in\mathcal{C}} \mathrm{err}_\mathcal{D}(c) + 2(1 + \gamma)\varepsilon\right)$$

References

[BD20] Jessa Bekker and Jesse Davis. "Learning from positive and unlabeled data: A survey". In: *Machine Learning* 109.4 (2020), pp. 719–760.

[Liu+02] Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. "Partially supervised classification of text documents". In: *ICML*. Vol. 2. 485. Sydney, NSW. 2002, pp. 387–394.

[DPNS15] Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. "Convex formulation for learning from positive and unlabeled data". In: *International conference on machine learning*. PMLR. 2015, pp. 1386–1394.

[BDU12] Shai Ben-David and Ruth Urner. "On the hardness of domain adaptation and the utility of unlabeled target samples". In: *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*. Springer. 2012, pp. 139–153.