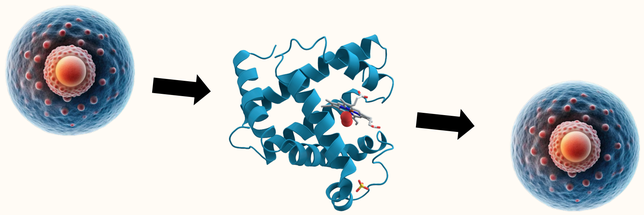
Learning from Imperfect and Missing Information

Jane H. Lee, Anay Mehrotra, Manolis Zampetakis

Yale

Proteins Involved in Signaling

Goal: Classify proteins that are involved in signaling across cellular membranes [Sair Tran Barabote '06, Elkan Noto '08]



Classical Learning Theory Approach
Collect positive and negative samples and train a classifier

Issue [Elkan Noto'08]

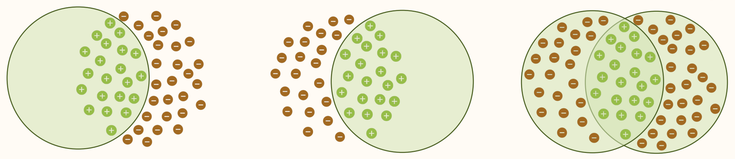
- "…TCDB is a database containing information about over 4000 proteins that are involved in signaling…"
- "If we ask a biologist for proteins that are not involved in this process, the only answer is *"all other proteins"*

We have a good quality data set with samples of the form $(x_i, 1)$
But, we do not have collected data of the form $(x_i, -1)$!

Q. *Can we solve classification problems using only positive data?*

Theorem [Natarajan 1987] [Lee **M.** Zampetakis]
Binary classification with only positive samples can be solved **only for very restricted** function classes

Scenario A | Scenario B | *Impossible to distinguish*



Idea [Elkan Noto'08]. Use public protein dataset as a proxy

Issues. Some proteins in this data may impact signaling

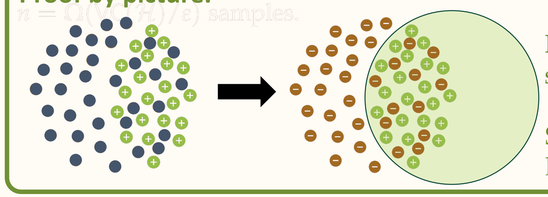
PU Learning – Prior Work

INPUT: n samples $x_i \sim \mathcal{P}$, and n samples $\tilde{x}_j \sim \mathcal{U}$.
OUTPUT: a concept $\widehat{h} \in \mathcal{H}$ such that with high probability

$$\mathbb{P}_{x\sim\mathcal{U}}(\widehat{h}(x) \neq h^\star(x)) \leq \varepsilon.$$

Theorem [Denis 1998] PU Learning reduces to agnostic.

Proof by picture.



Idea: Label unlabeled samples as negative.

Solve agnostic learning via ERM.

PU learning has many applications:

1. Proteins involved in signaling [Elkan Noto '08]
   Public dataset has poor quality
2. Classify x-rays [Hassanzadeh Kholghi Nguyen Chu '08]
   X-rays dataset is biased
3. Fraud detection [M.S. Yuan Wu '22]
   Unlabeled activity dataset is massive and difficult to clean

Positive and Imperfect Unlabeled Learning

PIU Learning
INPUT: n samples $x_i \sim \mathcal{P}$, and n samples $\tilde{x}_j \sim \widetilde{\mathcal{U}}$.
OUTPUT: a concept $\widehat{h} \in \mathcal{H}$ such that with high probability

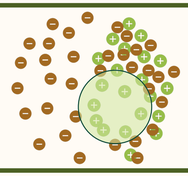
$$\mathbb{P}_{x\sim\mathcal{U}}(\widehat{h}(x) \neq h^\star(x)) \leq \varepsilon.$$

Issue: Denis's reduction does not apply!

Assumption 1 (SMOOTHNESS) It holds that $\chi^2(\mathcal{U}\|\widetilde{\mathcal{U}}) \leq C$.

Technical Vignette.

ERM does not work!



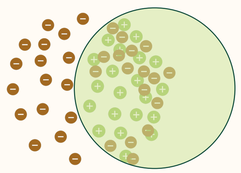
Our Results on PIU Learning

Theorem [Lee **M.** Zampetakis] If Assumption 1 holds, then we can solve PIU learning using $n = \Omega(\mathsf{VC}(\mathcal{H})/\varepsilon^2)$ samples.

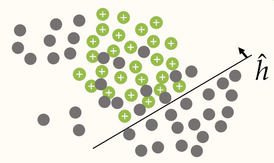
Simplifying assumption

$$\min_{h\in\mathcal{H}} \sum_j \mathbf{1}\{h(\tilde{x}_j) = 1\}$$
$$\text{s.t.}, \quad \sum_i \mathbf{1}\{h(x_i) = 1\} \geq (1-\varepsilon)n$$

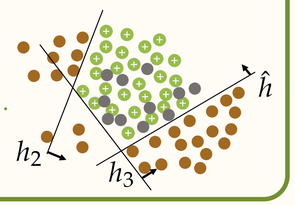
Decouples false-positives and false-negatives.



Pessimistic ERM does not work without $\mathrm{supp}(\mathcal{U}) = \mathrm{supp}(\widetilde{\mathcal{U}})$!



Idea: Find $h_1 = \text{Pessimistic ERM}(P, U)$,
$h_2 = \text{Pessimistic ERM}(P, U \cap h_1)$,
$h_3 = \text{Pessimistic ERM}(P, U \cap h_1 \cap h_2), \ldots$
return $\bigcap_{i=1}^{1/\varepsilon} h_i$



Assumption 2 (NON-TRIVIAL FRACTION OF POSITIVES) For some known $\alpha > 0$, $\mathbb{P}_{x\sim\mathcal{U}}(h^\star(x) = 1) \geq \alpha$.

Assumption 3 (APPROXIMABLE BY POLYNOMIALS)
We assume that $\mathcal{H}$ is approximable by polynomials.

Concept Class	$k(\varepsilon)$
PTFs of degree k	$O(k^2/\varepsilon^2)$ [Kane '11]
Intersections of k halfspaces	$O(\log k/\varepsilon^2)$ [KOS '08]
General convex sets	$O(d^{1/2}/\varepsilon^2)$ [Ball 1993]

Theorem [Lee **M.** Zampetakis]
If Assumptions 1, 2, and 3 hold, then we can efficiently solve PIU learning in using $n = \mathrm{poly}(d^{k(\varepsilon)})$ samples.

Method for efficient agnostic learning: **ℓ_1-regression**.
[Kalai Klivans Mansour Servedio '08]

$$\min_{\deg(p)\leq k} \sum_j \left|p(\tilde{x}_j)\right|, \quad \text{s.t.}, \quad \sum_i \min\{p(x_i), 1\} \geq (1-\varepsilon)n.$$

(Constrained ℓ_1-regression)

Can be solved using linear program.
Challenging to show that p is: (1) feasible, and (2) approximate optimizer of Pessimistic ERM.

Applications to Learning Theory + Statistics

1. **Learning from despite *corruptions* in data**
2. **Learning with smooth positive examples** - bypasses impossibility!
3. **Truncated statistics:** (a) detecting truncation and (b) parameter estimation with truncation - leads to fastest algorithms