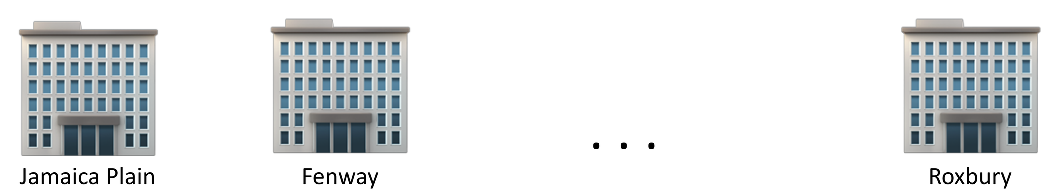
Improved Algorithms for Collaborative PAC Learning

Huy L. Nguyen and Lydia Zakynthinou

1) Collaborative Learning

- Introduced by [Blum, Haghtalab, Procaccia, Qiao '17].

Bank Stores

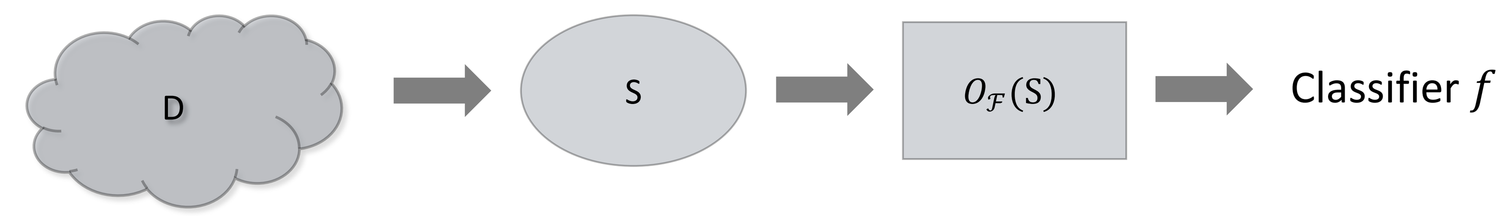


Distributions D_1 D_2 . . . D_k

- Goal: Draw labeled samples from all the distributions and use them to learn classifier(s) s.t. with high probability the error is low on all distributions.
  - Personalized: Can return different classifiers.
  - **Centralized**: Returns a single classifier.

2) Existing Results

- For a single distribution D:



  - VC dimension of concept class $\mathcal{F}$: d
  - If $|S| = m_{\epsilon,\delta} = O\left(\frac{1}{\epsilon}\left(d\ln\left(\frac{1}{\epsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right)\right)$:

    Classifier $f = O_{\mathcal{F}}(S)$ minimizes the error on S
    $\Rightarrow$ has error at most ϵ on D with probability $1-\delta$.
- If each were to learn a classifier independently, they would need $k \cdot m_{\epsilon,\delta}$ samples in total.
- With collaboration [BHPQ'17]:
  - Personalized $\approx \ln(k)\cdot m_{\epsilon,\delta}$.
  - Centralized $\approx \ln^2(k)\cdot m_{\epsilon,\delta}$.
  - Lower bound: $\Omega\left(\frac{k}{\epsilon}\ln\left(\frac{k}{\delta}\right)\right)$ for $d = \Theta(k)$.

3) Our Algorithms

Centralized Problem

Realizable setting
- Algorithm $R1$ matches the sample complexity for the personalized variant.
- Algorithm $R2$ matches the lower bound (better that $R1$ for most parameter regimes).

Non-realizable setting
- Deterministic classifier with error $(2+a)\cdot \mathrm{OPT} + \epsilon$, sample complexity matching the realizable setting, where a is constant.
- Randomized classifier with error $(1+a)\cdot \mathrm{OPT} + \epsilon$, using $\frac{1}{\epsilon}$ times more samples.

Key Idea: **Multiplicative Weight Updates**

4) Realizable Setting

Algorithm $R2$

Initialize weights $w_1^{(0)},\dots,w_k^{(0)} = 1$.

For $r = 1$ to $t = O(\ln(k/\delta))$ rounds:

Draw sample set $S^{(r)}$, $|S^{(r)}| = m_{\frac{\epsilon'}{16},\delta}$ from

$$\tilde{D}^{(r-1)} = \frac{\sum_{i=1}^k w_i^{(r-1)}\cdot D_i}{\sum_{i=1}^k w_i^{(r-1)}}.$$

Find a classifier $f^{(r)} = O_{\mathcal{F}}(S)$.

Draw $|T_i| = O(1/\epsilon')$ samples from each distribution, find $G^{(r)} = \{i: \mathrm{err}_{T_i}(f^{(r)}) \le 3\epsilon'/4\}$.

Update the weights: $w_i^{(r)} = 2w_i^{(r-1)}$, if $i \notin G^{(r)}$.

Return maj$\{f^{(r)}\}_{r=1}^t$.

$f^{(r)}$ has error $\epsilon'/2$ for at most 1/8 of the distributions' weight

Distinguishes between distributions with error $\le \epsilon'/2$ and $\ge \epsilon'$ with probability 99%.

For each D_i at least $0.6t$ classifiers have error $< \epsilon'$.

5) Non-Realizable Setting

- Need a smoother update rule.
- **Deterministic:**
  - $w_i^{(r)} = \left(1+\min\left(\frac{\mathrm{err}_{T_i}(f^{(r)})\cdot a^2}{(1+3a)\cdot \mathrm{err}_{S^{(r)}}(f^{(r)})+3\epsilon'}, a\right)\right)\cdot w_i^{(r-1)}$
  - Return maj$\{f^{(r)}\}_{r=1}^t$
- **Randomized:**
  - $w_i^{(r)} = \left(1+\frac{\mathrm{err}_{T_i}(f^{(r)})\cdot \epsilon' \cdot a}{(1+3a)\cdot \mathrm{err}_{S^{(r)}}(f^{(r)})+3\epsilon'}\right)\cdot w_i^{(r-1)}$
  - Return $f \overset{R}{\leftarrow} \{f^{(r)}\}_{r=1}^t$

Good classifiers are now the ones for which $\mathrm{err}_{T_i}(f^{(r)})$ is low and close to $\mathrm{err}_{D_i}(f^{(r)})$. ➡ For each D_i at least $\approx (1-a)t$ classifiers are good in the deterministic case, $\approx (1-\epsilon' a)t$ in the randomized.

6) Conclusion

	Alg 1	Alg 2
Realizable	$\frac{\ln(k)}{\epsilon}\left(d\ln\left(\frac{1}{\epsilon}\right)+k\ln\left(\frac{k}{\delta}\right)\right)$	$\frac{\ln(k/\delta)}{\epsilon}\left(d\ln\left(\frac{1}{\epsilon}\right)+k+\ln\left(\frac{k}{\delta}\right)\right)$
Non-realizable (determ.)	$\frac{\ln(k)}{\epsilon}\left(d\ln\left(\frac{1}{\epsilon}\right)+k\ln\left(\frac{k}{\delta}\right)\right)$	$\frac{\ln(k/\delta)}{\epsilon}\left(d\ln\left(\frac{1}{\epsilon}\right)+k+\ln\left(\frac{k}{\delta}\right)\right)$
Non-realizable (random.)	$\frac{\ln(k)}{\epsilon^2}\left(d\ln\left(\frac{1}{\epsilon}\right)+k\ln\left(\frac{k}{\delta}\right)\right)$	$\frac{\ln(k/\delta)}{\epsilon^2}\left((d+k)\ln\left(\frac{1}{\epsilon}\right)+\ln\left(\frac{k}{\delta}\right)\right)$

- Can we avoid the multiplicative factor of 2 in the non-realizable setting, without using $\frac{1}{\epsilon}$ times more samples?
- Can this classifier be adapted to perform well on a new related distribution?

Northeastern University
College of Computer and Information Science