Simultaneous Private Learning of Multiple Concepts

January 16, 2016

Mark Bun  
Harvard U.

Kobbi Nissim  
Harvard & Ben-Gurion

Uri Stemmer  
Ben-Gurion U.

No obfuscation!
Privacy-Preserving Data Analysis

Want curators that are:  
- Private  
- Accurate  
- Efficient
Privacy-Preserving Data Analysis

Want curators that are:
- Differentially Private
- Accurate for Learning Tasks
- Sample Efficient
What can be Done with Differential Privacy?

Histograms [DMNS06]
Contingency tables [BCDKMT07, GHRU11, TUV12, DNT14]
PAC learning [BDMN05, KLNRS08]
Clustering [BDMN05, NRS07]
Streaming algorithms [DNRY10, DNPR10, MMNW11]
SVD [HR12, HR13, KT13, DTTZ14]
Mechanism Design [MT07, NST10, X11, NOS12, CCKMV12, HK12, KPRU12]

Question: Can these tasks be performed as efficiently as their non-private counterparts?

This work: Sample complexity of privately PAC learning multiple concepts over the same example set

Refs. thanks to Salil Vadhan
PAC Learning [Valiant84]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>4Chan?</th>
<th>MLP:FiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>38</td>
<td>Y</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>27</td>
<td>N</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ h = ((\text{Age} < 10) \text{ AND } (\text{Gender} = \text{F})) \text{ OR } ((17 < \text{Age} < 40) \text{ AND } (\text{Gender} = \text{M}) \text{ AND } (4\text{Chan}\? = \text{Y})) \]
PAC Learning [Valiant84]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>4Chan?</th>
<th>MLP:FiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>38</td>
<td>Y</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>27</td>
<td>N</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
h = ((\text{Age} < 10) \text{ AND } (\text{Gender} = F)) \text{ OR } ((17 < \text{Age} < 40) \text{ AND } (\text{Gender} = M) \text{ AND } (\text{4Chan?} = Y))
\]
PAC Learning [Valiant84]

\( \mathcal{P} = \) unknown distribution over domain \( X \)

\( \mathcal{C} = \) concept class \( \{c : X \rightarrow \{0, 1\}\} \) e.g. DNF of intervals

Fact: \( n = \Theta(\text{VC}(\mathcal{C})) \) samples suffice to generalize

\( \text{VC}(\mathcal{C}) \leq \log |\mathcal{C}|, \) but can be much smaller
PAC Multi-Learning [Valiant06]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>4Chan?</th>
<th>MLP:FiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>38</td>
<td>Y</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>27</td>
<td>N</td>
<td>0</td>
</tr>
</tbody>
</table>

F 65 N  h  0
## PAC Multi-Learning [Valiant06]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>4Chan?</th>
<th>MLP:FiM</th>
<th>AdvTime</th>
<th>A:tLA</th>
<th>Dora</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>38</td>
<td>Y</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>N</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td>N</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>27</td>
<td>N</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Multi-Learner**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>4Chan?</th>
<th>MLP:FiM</th>
<th>AdvTime</th>
<th>A:tLA</th>
<th>Dora</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>65</td>
<td>N</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
PAC Multi-Learning [Valiant06]

\[ \mathcal{P} = \text{unknown distribution over domain } X \]

\[ \mathcal{C} = \text{concept class } \{c: X \to \{0, 1\}\} \]

Goal: For all \( \mathcal{P} \) and \( c_1, \ldots, c_k \in \mathcal{C} \), output \( h \) s.t.

\[ h_i \approx c_i \text{ on } \mathcal{P} \text{ for every } i = 1, \ldots, k \]
**PAC Multi-Learning** [Valiant06]

\[ \mathcal{P} = \text{unknown distribution over domain } X \]
\[ \mathcal{C} = \text{concept class } \{c: X \rightarrow \{0, 1\}\} \]

<table>
<thead>
<tr>
<th>(x_1)</th>
<th>(c_1(x_1))</th>
<th>(c_k(x_1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_2)</td>
<td>(c_1(x_2))</td>
<td>(c_k(x_2))</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(c_i(x_i))</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(x_n)</td>
<td>(c_1(x_n))</td>
<td>(c_k(x_n))</td>
</tr>
</tbody>
</table>

**Fact:** \( n = \Theta(\text{VC}(\mathcal{C})) \) samples suffice to generalize

Uniform convergence: Over a random sample \( S \) of size \( O_{\alpha, \beta}(\text{VC}(\mathcal{C})) \),

\[
\Pr\left[ \exists f, g \in \mathcal{C} : (f \mid_S = g \mid_S) \land \text{err}_P(f, g) > \alpha \right] \leq \beta
\]
What about Privacy?

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>4Chan?</th>
<th>MLP:FiM</th>
<th>AdvTime</th>
<th>A:tLA</th>
<th>Dora</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>38</td>
<td>Y</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>N</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td>N</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>27</td>
<td>N</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The data is anonymized, so it’s safe to release, right?

| F | 65 | N | h | 0 | 0 | 1 | 0 |

What about Privacy?

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>4Chan?</th>
<th>MLP:FiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>38</td>
<td>Y</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>27</td>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>65</td>
<td>N</td>
<td>0</td>
</tr>
</tbody>
</table>

The data is anonymized, so it’s safe to release, right?

Wrong! [Narayanan-Shmatikov08]

Pinkie Pie is best pony!

Author: Kobbi Nissim
20 December 2014

151 out of 205 people found the following review useful:

Jake The Dog is my spirit animal

Author: Uri Stemmer
20 December 2014

120 out of 164 people found the following review useful:

Motivates need for rigorous privacy guarantees
Private PAC Multi-Learning

Extending Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith ‘08
Private PAC Multi-Learning

Extending Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith ‘08

<table>
<thead>
<tr>
<th>Attributes</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th></th>
<th>$c_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$b_{11}$</td>
<td>$b_{21}$</td>
<td></td>
<td>$b_{k1}$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$b_{12}$</td>
<td>$b_{22}$</td>
<td></td>
<td>$b_{k2}$</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$b_{13}$</td>
<td>$b_{23}$</td>
<td></td>
<td>$b_{k3}$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>$b_{ji}$</td>
<td></td>
</tr>
<tr>
<td>$x_n$</td>
<td>$b_{1n}$</td>
<td>$b_{2n}$</td>
<td></td>
<td>$b_{kn}$</td>
</tr>
</tbody>
</table>

Differentially Private Multi-Learner

$h_1$ | $h_2$ |  | $h_k$ |
Private PAC Multi-Learning

<table>
<thead>
<tr>
<th>Attributes</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>...</th>
<th>$c_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$b_{11}$</td>
<td>$b_{21}$</td>
<td>$b_{k1}$</td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td>$b_{12}$</td>
<td>$b_{22}$</td>
<td>$b_{k2}$</td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>$b_{13}$</td>
<td>$b_{23}$</td>
<td>$b_{k3}$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>$b_{ji}$</td>
<td></td>
</tr>
<tr>
<td>$x_n$</td>
<td>$b_{1n}$</td>
<td>$b_{2n}$</td>
<td>$b_{kn}$</td>
<td></td>
</tr>
</tbody>
</table>

$D$ and $D'$ are **neighbors** if they differ on one row

$M$ is **differentially private** if for all neighbors $D, D'$:

$$M(D) \approx M(D')$$

Extending Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith ‘08

DN03+Dwork, DN04, BDMN05, DMNS06, DKMMN06
**Private PAC Multi-Learning**

Extending Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith ‘08

<table>
<thead>
<tr>
<th>Attributes</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>...</th>
<th>$c_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$b_{11}$</td>
<td>$b_{21}$</td>
<td>$b_{k1}$</td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td>$b_{12}$</td>
<td>$b_{22}$</td>
<td>$b_{k2}$</td>
<td></td>
</tr>
<tr>
<td>$x'_3$</td>
<td>$b'_{13}$</td>
<td>$b'_{23}$</td>
<td>$b'_{k3}$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>$b_{ji}$</td>
<td></td>
</tr>
<tr>
<td>$x_n$</td>
<td>$b_{1n}$</td>
<td>$b_{2n}$</td>
<td>$b_{kn}$</td>
<td></td>
</tr>
</tbody>
</table>

$D$ and $D'$ are **neighbors** if they differ on one row

$M$ is **differentially private** if for all neighbors $D$, $D'$:

$$M(D) \approx M(D')$$

DN03+Dwork, DN04, BDMN05, DMNS06, DKMMN06
Private PAC Multi-Learning

Extending Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith ‘08

<table>
<thead>
<tr>
<th>Attributes</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$\ldots$</th>
<th>$c_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$b_{11}$</td>
<td>$b_{21}$</td>
<td>$\ldots$</td>
<td>$b_{k1}$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$b_{12}$</td>
<td>$b_{22}$</td>
<td>$\ldots$</td>
<td>$b_{k2}$</td>
</tr>
<tr>
<td>$x'_3$</td>
<td>$b'_{13}$</td>
<td>$b'_{23}$</td>
<td>$\ldots$</td>
<td>$b'_{k3}$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_n$</td>
<td>$b_{1n}$</td>
<td>$b_{2n}$</td>
<td>$\ldots$</td>
<td>$b_{kn}$</td>
</tr>
</tbody>
</table>

$D$ and $D'$ are neighbors if they differ on one row

$M$ is $(\varepsilon, \delta)$-differentially private if for all neighbors $D$, $D'$ and $T \subseteq \text{Range}(M)$:

$$\Pr[M(D') \in T] \leq (1 + \varepsilon) \Pr[M(D) \in T] + \delta$$

Questions:

- How does the sample complexity depend on $k$?

DN03+Dwork, DN04, BDMN05, DMNS06, DKMMMN06
Samp. Cx. of Private Multi-Learning

- For $k = 1$, can privately learn $\mathcal{C}$ with sample complexity $n = \text{SCDP}_1(\mathcal{C})$ where:
  \[
  \text{VC}(\mathcal{C}) \leq \text{SCDP}_1(\mathcal{C}) \leq \log |\mathcal{C}| \quad [\text{KLNRS08}]
  \]

- For arbitrary $k$, can learn each concept independently: $\text{SCDP}_k(\mathcal{C}) \leq k^{1/2} \text{SCDP}_1(\mathcal{C})$ [DRV10]

- Can we do better? Is the dependence on $k$ necessary?
Our Results

Upper bounds:

<table>
<thead>
<tr>
<th>$C$</th>
<th>PAC learning (proper and improper)</th>
<th>Agnostic learning (proper and improper)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{POINT}_X$</td>
<td>1</td>
<td>$\sqrt{k}$</td>
</tr>
<tr>
<td>$\text{THRESH}_X$</td>
<td>$2^{\log^*</td>
<td>X</td>
</tr>
<tr>
<td>General $C$</td>
<td>$\min{\sqrt{k} \log</td>
<td>C</td>
</tr>
<tr>
<td>$\text{PAR}_d$ (uniform)</td>
<td>$\log</td>
<td>C</td>
</tr>
</tbody>
</table>

Lower bounds:

<table>
<thead>
<tr>
<th>$C$</th>
<th>PAC learning proper</th>
<th>PAC learning improper</th>
<th>Agnostic learning proper</th>
<th>Agnostic learning improper</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{POINT}_X$</td>
<td>1</td>
<td></td>
<td>$\sqrt{k}$</td>
<td></td>
</tr>
<tr>
<td>$\text{THRESH}_X$</td>
<td>$\log^*</td>
<td>X</td>
<td>+ k^{1/3}$</td>
<td>$k^{1/3}$</td>
</tr>
<tr>
<td>$\text{PAR}_d$ (uniform)</td>
<td>$\log</td>
<td>C</td>
<td>$</td>
<td></td>
</tr>
</tbody>
</table>
Our Results (Human Readable Version)

• Upper bounds
  - Generic multi-learner achieving
  \[ \text{SCDP}_k(C) \leq \text{VC}(C) \log |X| + k^{1/2} \text{VC}(C) \]
  - Improved multi-learners for specific classes

• Lower bounds via fingerprinting codes
  - \( k^{1/3} \) lower bound for multi-learning thresholds
  - \( k^{1/2} \) lower bound for agnostically learning...
    ...anything
Threshold Functions

$X$ a totally ordered domain

$C = \{f_t : f_t(x) = 1 \text{ iff } x \leq t\}$
Fingerprinting Codes [Boneh-Shaw95]

I want to distribute my new movie...

...but Equestria is full of pirates!
Fingerprinting Codes [Boneh-Shaw95]

I want to distribute my new movie

...but Equestria is full of pirates!

Who collude against me!
Fingerprinting Codes [Boneh-Shaw95]
Fingerprinting Codes [Boneh-Shaw95]

Gen(1^n) outputs \( W \in (\{0,1\}^k)^n \)

For all coalitions \( T \) and all pirate alg. for producing \( w \),

\[ \Pr[\text{Trace}(w) \in T] \approx 1 \]
FP Codes vs. Diff. Privacy [B.-Ullman-Vadhan14]

Coalition of n pirates

Feasible codeword $w$

Pr[Trace($w$) = $\text{Princess}$] $\geq 1/n$
FP Codes vs. Diff. Privacy [B.-Ullman-Vadhan14]

Coalition of n pirates

Pr[Trace(w) = 🦄] << 1/n

Feasible codeword w
FP Codes vs. Diff. Privacy [B.-Ullman-Vadhan14]

Trace behaves very differently depending on whether 🪄 is in the coalition

Fingerprinting codes are the “opposite” of differential privacy!
Lower Bound for Thresholds

Suppose (for contradiction) we have
• A FP code of length k for (n+1) users
• A diff. private M that learns k threshold functions

Reduction: Use M to break security of the FP code

Labeled sample of n users = coalition of n users
C1 C2 C3 C4 C5
X1 1 1 0 1 1
X2 0 1 0 1 1
Xn 0 0 0 0 1

M accurate ⇒ w feasible
Lower Bound for Thresholds

Labeled sample of n users = coalition of n users

How do we ensure M is accurate?

Each column of the codebook needs to be consistent with a threshold concept

Magic observation: The FP code of [BS95] has this structure

M accurate ⇒ w feasible
Lower Bound for Thresholds

Labeled sample of n users = coalition of n users

Suppose (for contradiction) we have

- A nice FP code of length k for (n+1) users
- A diff. private M that learns k threshold functions

Reduction: Use M to break security of the FP code

\[ w_j = \text{round} \left( \frac{1}{n} \sum_{i=1}^{n} h_j(x_i) \right) \]
Lower Bound for Thresholds

Labeled sample of n users = coalition of n users

\[
\begin{align*}
\mathbf{x}_1 & \quad \begin{array}{ccccc}
1 & 1 & 0 & 1 & 1 \\
\end{array} \\
\mathbf{x}_2 & \quad \begin{array}{ccccc}
0 & 1 & 0 & 1 & 1 \\
\end{array} \\
\vdots & \\
\mathbf{x}_n & \quad \begin{array}{ccccc}
0 & 0 & 0 & 0 & 1 \\
\end{array}
\end{align*}
\]

\[
\text{Pr}[\text{Trace}(w) = \begin{array}{ccccc}
1 & 1 & 1 & 1 & 1 \\
\end{array}] \geq \frac{1}{n}
\]

M accurate ⇒ \(w\) feasible

M Pirate algorithm
Lower Bound for Thresholds

Labeled sample of n users = coalition of n users

Contradicts security of FP code!

Pr[Trace(w) = \text{success}] \geq \frac{(1/n) - \delta}{1 + \epsilon}

\geq \frac{1}{3n}

M accurate ⇒ w feasible

M private ⇒ Trace fails
Lower Bound for Thresholds

• \( \exists \) “nice” FP code for \( n \) users with length \( k \)
  \( \Rightarrow \) learning \( k \) thresholds requires \( n \) samples

• [BS95] \( \exists \) “nice” FP code for \( \Omega(k^{1/3}) \) users of length \( k \)
  \( \therefore \) learning \( k \) thresholds requires \( n \geq \Omega(k^{1/3}) \)
Conclusions

• Introduce study of private multi-learning
• Paint a complex picture of how sample complexity depends on $k$

• Open questions
  – Is dependence on $\text{poly}(k)\text{VC}(C)$ necessary?
  – Other examples of “direct-sum” tasks?

Thank you!
Generic Multi-Learner

• Apply technique from [Beimel-Nissim-Stemmer15] for reducing labeled sample complexity

• Idea: 1. Identify set $H$ of $2^{VC(C)}$ “important” concepts via sanitization
   2. Run [KLNRS08] generic learner $k$ times using $H$ as hypothesis class

• Total sample complexity = fixed cost of sanitization + $k^{1/2} VC(C)$