Evolvability of Linear Threshold Functions

Vitaly Feldman
IBM Almaden Research Center
Learning

- Predict $f$ on $x \in X$
- Direct programming is infeasible
- Use examples of $f$

Evolvability

- Realize $h$ on $x \in X$
- No programmer
- Try variants of current $h$ and choose the “fittest”

Human brain

Biological evolution
The gene expression example

- The expression of each gene in the DNA is regulated by multiple transcription factors (TF)
  - Proteins that detect various conditions of the environment
    - Internal signals/other proteins
    - Presence/absence of a chemical element
    - Temperature, light etc.
- More than 2000 known TFs
- How do these regulation mechanisms evolve?
- E.g. Optimally the gene should be expressed whenever several fixed (but unknown) TFs are jointly present
  - Can the optimal combination of TFs evolve efficiently
Model [Valiant 06]

- Based on PAC learning model [Valiant 84]

Random examples: $(x, f(x))$
$x \sim D$ over $X, f \in C$

Learning algorithm:
For every $f, D, \epsilon > 0$, output $h$
s.t. $\Pr[f(x) = h(x)] \geq 1 - \epsilon$
Poly in $\frac{1}{\epsilon}, |x|, |f|$ time

Selection model:
Poly in $\frac{1}{\epsilon}, |x|, |f|$ time

Mutation algorithm:
For every $f, D, \epsilon > 0$, reach $h$
s.t. $\Pr[f(x) = h(x)] \geq 1 - \epsilon$
Poly in $\frac{1}{\epsilon}, |x|, |f|$ time
Mutation algorithm

- $R$ - representation class of functions over $X$
  - E.g. all linear thresholds over $\mathbb{R}^n$
- $M$ - randomized algorithm that given $r \in R$ outputs (a random mutation) $r' \in R$
  - Efficient: poly in $\frac{1}{\epsilon}, n$
  - E.g. choose a random $i$ and adjust $w_i$ by 0, +1 or -1 randomly
Selection

- Fitness/performance $P_D(f, r) \in [-1, 1]$
  - Correlation: $E_D[f(x)r(x)]$
  - Quadratic loss: $1 - E_D[(f(x) - r(x))^2]/2$
- For $r \in R$ sample $M(r)$ $p$ times: $r_1, r_2, \ldots, r_p$
- Estimate empirical performance of $r$ and each $r_i$ using $s$ samples: $\tilde{P}_D(f, r_i)$

$p, s$ and $1/t$ are “feasible” (polynomial in $n, 1/\epsilon$)
Evolvability

- Class of functions $C$ is evolvable over $D$ if exists an evolution algorithm $(R, M)$ and a polynomial $\ell(\cdot, \cdot)$ s.t.

For every $f \in C, r \in R, \varepsilon > 0$ and a sequence $r_0 = r, r_1, r_2, \ldots$ where $r_{i+1} \leftarrow \text{Select}(R, M, r_i)$ it holds: $P_D(f, r_{\ell(\frac{1}{\varepsilon})}) \geq 1 - \varepsilon$ w.h.p.

- Evolvable (distribution-independently)
  - Evolvable for all $D$ by the same mutation algo $(R, M)$
  - Evolvable weakly
    - $P_D(f, r_{\ell(\frac{1}{\varepsilon})}) \geq 1/poly(n)$
Prior work

- \( \text{EV} \subseteq \text{PAC} \)
- \( \text{EV} \subseteq \text{SQ} \neq \text{PAC} \) [Valiant 06]
  - Statistical Query learning [Kearns 93]: estimates of \( E_D[\psi(x, f(x))] \) for an efficiently evaulatable \( \psi \)
- Monotone conjunctions are evolvable over the uniform distribution on \( \{0,1\}^n \) [Valiant 06]
  - Improved to general conjunctions [Jacobson 09, KVV 10]
- \( \text{EV} = \text{CSQ} \) [F 08]
  - Learnability by correlational statistical queries
    \( \text{CSQ}: E_D[\phi(x)f(x)] \)
- Fixed \( D \): \( \text{EV} = \text{CSQ} = \text{SQ} \) [Bshouty, F 01]
- All \( D \): \( \text{CSQ} \neq \text{SQ} \)
  - General linear threshold functions are not evolvable, even weakly
- Singletons are evolvable [F 09α]
This work

- Are conjunctions evolvable distribution-independently?
  - [F, Valiant COLT 08 Open problem]
- Evolvable weakly [F 08]
- Singletons are evolvable [F 09a]
- General linear threshold functions are not evolvable, even weakly [F 08]
- NO
  - Not CSQ learnable
  - Monotone conjunctions of a superconstant number of variables are not CSQ learnable to subconstant accuracy $\epsilon$
  - No boosting for evolvability
Overview

- Information-theoretic lower bound on CSQ learnability
  - Weak SQ learning, fixed $D$ [BFJKMR 94; Bshouty, F 01; Yang 05]
  - Weak CSQ learning [F 08]
  - (General) SQ learning, fixed $D$ [BCGKL 07; Simon 07; F 09b; Szoreniy 09]

For every function $\phi: X \rightarrow [-1,1]$ and distribution $D$ over $X$ there exists a poly-sized set of functions $G$ that “distinguishes” between any $f \in C$ over any distribution $D'$ from $\phi$ over $D$

Exists $g \in G$, $|E_{D'}[f \cdot g] - E_D[\phi \cdot g]| \geq \tau$,

Unless $f$ is $\epsilon$ −close over $D'$ to a fixed $h$
Hard to distinguish function-distribution pairs

- Let $T_A$ be the conjunction of variables in a set $A$
- Define $D_A$ and $\theta_A : X \rightarrow [-1,1]$ such that
  - $T_A$ over $D_A$ is indistinguishable from $c + \theta_A$ over $U$ ($c \in [-1,1]$)
  - for size-$k$ $A$ and $B$ that share less than $k/3$ variables $E_U[\theta_A \cdot \theta_B] = 0$
- For $k = \omega(1)$ there are $n^{\omega(1)}$ size-$k$ sets of variables s.t. no two share $\geq \frac{k}{3}$ variables
- Cannot be distinguished using a polynomial number of functions from $c$ over $U$
- Main idea: Correlation is determined by Fourier coefficients. $D_A$ “erases” low order Fourier coefficients of $T_A$ ($\leq k/3$). QED.
Selection

- Fitness/performance $P_D(f, r) \in [-1, 1]$
  - Correlation: $E_D[f(x)r(x)]$
  - Quadratic loss: $1 - E_D[(f(x) - r(x))^2]/2$
- For $r \in R$, sample $M(r)$ $p$ times: $r_1, r_2, \ldots, r_p$
- Estimate empirical performance of $r$ and each $r_i$ using $s$ samples: $\tilde{P}_D(f, r_i)$

Robust to selection mechanism [F 09a]
Robust to drift [KVV 10]
Related work (quadratic loss)

- Decision lists over $U$ on $\{0,1\}^n$ [Michael 07]
- Any SQ-learnable $C$ (non-monotone) [F 09a]
- Singletons [F 09a]
- Any SQ-learnable $C$, fixed $D$ [F 09b]
- Conjunctions [F 09b]
- Linear functions [P. Valiant 11]
This work: evolvability of LTFs

- LTFs are evolvable strictly monotonically with quadratic loss
  - Polynomial in $1/\gamma$ where $\gamma$ is the margin of $f$
  - Implies: conjunctions, polynomial integer weight LTFs over $\{0,1\}^n$
  - Everything that can be efficiently embedded into LTFs with margin

- Simple mutation algorithm
  - Choose random variable $x_i$ (or a constant)
  - Add or subtract $ax_i$ from $r(x)$
  - Clip-off values of $r(x) + ax_i$ outside of $[-1,1]$: $P(r(x) + ax_i)$

- Can be extended to other loss functions and has additional robustness properties
If $E_D[(f(x) - r(x))^2] \geq \epsilon$ then exists $i \in [0,1,\ldots,n]$ and $\alpha$ s.t.

$$|\alpha| \geq \frac{1}{\text{poly}(n,\frac{1}{\epsilon})} \text{ and }$$

$$E_D[(f(x) - P(r(x) + \alpha x_i))^2] \leq E_D[(f(x) - r(x))^2] - \alpha^2$$

Strictly beneficial neighborhood implies strictly monotone evolvability

Proof:

1. Gradient of $E_D[(f(x) - r(x))^2]$ at $r(x)$ is $-2(f(x) - r(x))$
2. Margin $\gamma$ implies that $E_D[(f - r)(\sum w_i x_i - \theta)] \geq \epsilon \gamma / 2$
3. Exists $i \in [0,1,\ldots,n]$ s.t. $|E_D[(f - r) x_i]| \geq \epsilon \gamma / (3 \sqrt{n})$
4. $\alpha = +/ - \epsilon \gamma / (3 \sqrt{n})$ gives

$$E_D[(f(x) - (r(x) + \alpha x_i))^2] \leq E_D[(f(x) - r(x))^2] - \alpha^2$$

$P()$ can only reduce the loss. QED.
Conclusions and open problems

• Cannot learn much from looking only at disagreement/correlation
• Can evolve robustly with quadratic loss most of PAC learnable classes
• Characterize (strong) CSQ learning
  o Strengthen the lower bound to constant accuracy
• Monotone evolvability of other SQ learnable concept classes
  o Decision lists
  o General LTFs
• Proper evolvability: use only LTFs as representations