Overview of quantum learning theory

Srinivasan Arunachalam (IBM Quantum)

Machine learning

Goals of classical ML

- Grand goal: enable AI systems to improve themselves
- Interacting with environment, providing useful data to "train" the machine
- Underpinning these improvements is better algorithms, more data, computational power

In the last decade:

1. Image processing: Deep neural networks used for image recognition
2. Natural language processing: used for speech recognition
3. Reinforcement learning

DeepMind has algorithms for chess, Go, and protein folding!

Quantum machine learning

What can quantum computing do for machine learning?

- Close to quantum advantage candidate for a practical problem?
- Polynomial speed-ups for many tasks as training Boltzmann machines, clustering, perceptron learning, support vector machines, ...
- Exponential speed-ups for some tasks such as PCA, recommendation systems, linear system solvers, ...

The era of de-quantization

- Tang'18 gave a classical polynomial-time algorithm for recommendation systems
- A flurry of de-quantized algorithms for principal component analysis [T'18], low-rank linear system solvers [GLT'18, CLW'18], SDP solvers [CLLW'18]

A need to prove formal separations in quantum machine learning

Quantum learning theory

In classical ML, the field of computational learning theory deals with understanding ML from a theoretical perspective.

In these lectures:

1. Learning Boolean functions encoded as quantum examples
   - Hardness of PAC learning
   - Some positive and negative under the uniform distribution
2. Learning quantum states
   - General tomography and learning specific class of states
   - Learning in weaker settings: PAC learning and shadows
3. Statistical learning and open questions

A Theory of the Learnable

Valiant gave a complexity-theoretic definition of what it means to learn

Goal: learn a class of functions $\mathcal{C} = \{c_1, c_2 \ldots, \}$ where $c_i : X \rightarrow \{0, 1\}$

Example: $c_i s$ are halfspaces, i.e., each c_i is associated with a separating hyperplane

What does it mean to learn? Let $c^\star \in \mathcal{C}$ (unknown). Given points in X, what is $c^\star$?

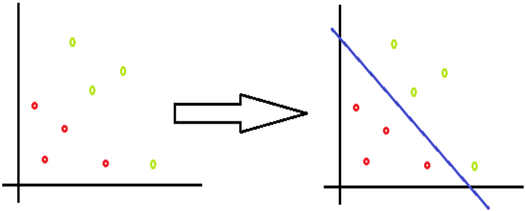
A Theory of the Learnable

Valiant gave a complexity-theoretic definition of what it means to learn

Goal: learn a class of functions $\mathcal{C} = \{c_1, c_2 \ldots, \}$ where $c_i : X \to \{0, 1\}$

Example: $c_i s$ are halfspaces, i.e., each c_i is associated with a separating hyperplane

What does it mean to learn? Let $c^\star \in \mathcal{C}$ (unknown). Given points in X, what is $c^\star$?



Classical learner using classical examples

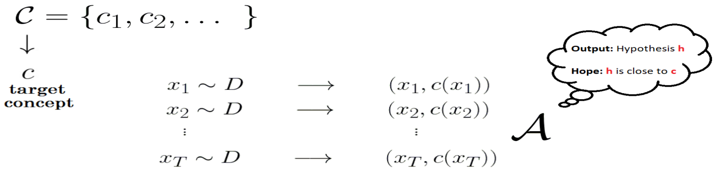
Basic definitions

- Concept class $\mathcal{C}$: collection of Boolean functions on n bits (Known)
- Target concept c: some function $c \in \mathcal{C}$. (Unknown)
- Distribution $D : \{0,1\}^n \to [0,1]$
- Labeled example for $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$

Classical learner using classical examples

Basic definitions

- Concept class $\mathcal{C}$: collection of Boolean functions on n bits (Known)
- Target concept c: some function $c \in \mathcal{C}$. (Unknown)
- Distribution $D : \{0,1\}^n \to [0,1]$
- Labeled example for $c \in \mathcal{C}$: $(x, c(x))$ where $x \sim D$

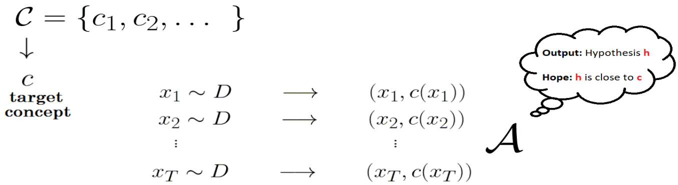


Learner is trying to learn c

Learning model: classical PAC learning

Basic definitions

- Concept class $\mathcal{C}$: collection of Boolean functions on n bits (Known)
- Target concept c: some function $c \in \mathcal{C}$ (Unknown)
- Distribution $D : \{0,1\}^n \to [0,1]$ (Unknown)



Goal of $\mathcal{A}$. For every $c \in \mathcal{C}$ and D, with probability $\geq 1 - \delta$ output a hypothesis h s.t.

$$\Pr_{x \sim D}[h(x) \neq c(x)] \leq \varepsilon$$

Sample complexity of $\mathcal{C}$: Number of examples used on the hardest $c \in \mathcal{C}$ and D

Time complexity of $\mathcal{C}$: Number of time steps used on the hardest $c \in \mathcal{C}$ and D

Quantum PAC learning

Learner is quantum and the data is quantum

Bshouty-Jackson'95 introduced a quantum example as a superposition

$$\sum_{x\in\{0,1\}^n} \sqrt{D(x)}\,|x, c(x)\rangle$$

Measuring this state gives a $(x, c(x))$ with probability $D(x)$,
so quantum examples are at least as powerful as classical

$$\mathcal{C} = \{c_1, c_2, \ldots\ \}$$

$\downarrow$

c
target concept

$\mathcal{QA}$

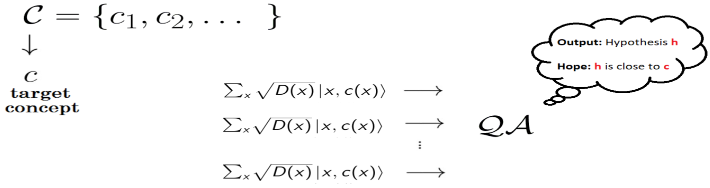
Quantum PAC learning

Learner is quantum and the data is quantum

Bshouty-Jackson'95 introduced a quantum example as a superposition

$$\sum_{x\in\{0,1\}^n} \sqrt{D(x)}\,|x, c(x)\rangle$$

Measuring this state gives a $(x, c(x))$ with probability $D(x)$, so quantum examples are at least as powerful as classical



Goal of $\mathcal{QA}$. For every $c \in \mathcal{C}$ and D, with prob. $\geq 1-\delta$ output a hypothesis h s.t.

$$\Pr_{x\sim D}[h(x) \neq c(x)] \leq \varepsilon$$

Motivating question: Do quantum examples give an advantage for PAC learning?

Vapnik and Chervonenkis (VC) dimension

VC dimension of $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$

Let M be the $|\mathcal{C}| \times 2^n$ Boolean matrix whose c-th row is the truth table of concept $c : \{0,1\}^n \to \{0,1\}$

VC-dim($\mathcal{C}$): largest d s.t. the $|\mathcal{C}| \times d$ rectangle in M contains $\{0,1\}^d$ These d column indices are shattered by $\mathcal{C}$

Vapnik and Chervonenkis (VC) dimension

VC dimension of $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$

M is the $|\mathcal{C}| \times 2^n$ Boolean matrix whose c-th row is the truth table of c

VC-dim($\mathcal{C}$): largest d s.t. the $|\mathcal{C}| \times d$ rectangle in M contains $\{0,1\}^d$ These d column indices are shattered by $\mathcal{C}$

Table: VC-dim($\mathcal{C}$) $= 2$

Concepts	Truth table			
c_1	0	1	0	1
c_2	0	1	1	0
c_3	1	0	0	1
c_4	1	0	1	0
c_5	1	1	0	1
c_6	0	1	1	1
c_7	0	0	1	1
c_8	0	1	0	0
c_9	1	1	1	1

Vapnik and Chervonenkis (VC) dimension

VC dimension of $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$

M is the $|\mathcal{C}| \times 2^n$ Boolean matrix whose c-th row is the truth table of c

VC-dim($\mathcal{C}$): largest d s.t. the $|\mathcal{C}| \times d$ rectangle in M contains $\{0,1\}^d$ These d column indices are shattered by $\mathcal{C}$

Table: VC-dim($\mathcal{C}$) = 2

Concepts	Truth table			
c_1	0	1	0	1
c_2	0	1	1	0
c_3	1	0	0	1
c_4	1	0	1	0
c_5	1	1	0	1
c_6	0	1	1	1
c_7	0	0	1	1
c_8	0	1	0	0
c_9	1	1	1	1

Table: VC-dim($\mathcal{C}$) = 3

Concepts	Truth table			
c_1	0	1	1	0
c_2	1	0	0	1
c_3	0	0	0	0
c_4	1	1	0	1
c_5	1	0	1	0
c_6	0	1	1	1
c_7	0	0	1	1
c_8	0	1	0	1
c_9	0	1	0	0

VC dimension characterizes PAC sample complexity

VC dimension of $\mathcal{C}$

M is the $|\mathcal{C}| \times 2^n$ Boolean matrix whose c-th row is the truth table of c

VC-dim($\mathcal{C}$): largest d s.t. the $|\mathcal{C}| \times d$ rectangle in M contains $\{0,1\}^d$ These d column indices are shattered by $\mathcal{C}$

Fundamental theorem of PAC learning

Suppose VC-dim($\mathcal{C}$) $= d$

- Blumer-Ehrenfeucht-Haussler-Warmuth'86:
  every (ε, δ)-PAC learner for $\mathcal{C}$ needs $\Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ examples
- Hanneke'16: exists an (ε, δ)-PAC learner for $\mathcal{C}$ using $O\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ examples

Quantum sample complexity

Quantum upper bound

Classical upper bound $O\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ carries over to quantum

Best known quantum lower bounds

Atici & Servedio'04: lower bound $\Omega\left(\frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon}\right)$

Zhang'10 improved first term to $\frac{d^{1-\eta}}{\varepsilon}$ for all $\eta > 0$

Quantum sample complexity = Classical sample complexity

Quantum upper bound

Classical upper bound $O\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ carries over to quantum

Best known quantum lower bounds

Atici & Servedio'04: lower bound $\Omega\left(\frac{\sqrt{d}}{\varepsilon} + d + \frac{\log(1/\delta)}{\varepsilon}\right)$

Zhang'10 improved first term to $\frac{d^{1-\eta}}{\varepsilon}$ for all $\eta > 0$

Our result: Tight lower bound

[AW'18]: $\Omega\left(\frac{d}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$ quantum examples are necessary

Two proof approaches

- Information theory: conceptually simple, nearly-tight bounds
- Optimal measurement: tight bounds, some messy calculations

Proof approach: Pretty Good Measurement

State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z \in [m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob p_z. Goal: identify z
- Optimal measurement could be quite complicated,
  but we can always use the Pretty Good Measurement. This has POVM operators
  $M_z = p_z \rho^{-1/2} |\psi_z\rangle\langle\psi_z| \rho^{-1/2}$, where $\rho = \sum_z p_z |\psi_z\rangle\langle\psi_z|$
- Success probability of PGM: $P_{PGM} = \sum_i p_z \mathrm{Tr}(M_z |\psi_z\rangle\langle\psi_z|)$
- Crucial property: if P_{opt} is the success probability of the optimal measurement,
  then $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill'02)

How does learning relate to identification?

- Quantum PAC: Given $|\psi_c\rangle = \left|E_{c,D}\right\rangle^{\otimes T}$, learn c *approximately*
- Goal: show $T \geq d/\varepsilon$, where $d = \text{VC-dim}(\mathcal{C})$
- Suppose $\{s_0, \ldots, s_d\}$ is shattered by $\mathcal{C}$. Fix a nasty distribution D:
  $D(s_0) = 1 - 16\varepsilon$, $D(s_i) = 16\varepsilon/d$ on $\{s_1, \ldots, s_d\}$
- Let $E : \{0,1\}^k \to \{0,1\}^d$ be a good error-correcting code
  s.t. $k \geq d/4$ and $d_H(E(y), E(z)) \geq d/8$
- Pick concepts $\{c^z\}_{z \in \{0,1\}^k} \subseteq \mathcal{C}$: $c^z(s_0) = 0$, $c^z(s_i) = E(z)_i\ \forall\ i$

Pick concepts $\{c^z\} \subseteq \mathcal{C}$: $c^z(s_0) = 0$, $c^z(s_i) = E(z)_i \; \forall \; i$

Suppose $VC(\mathcal{C}) = d + 1$ and $\{s_0, \ldots, s_d\}$ is shattered by $\mathcal{C}$, i.e., $|\mathcal{C}| \times (d+1)$ rectangle of $\{s_0, \ldots, s_d\}$ contains $\{0,1\}^{d+1}$

Concepts $c \in \mathcal{C}$	Truth table						
	s_0	s_1	$\cdots$	s_{d-1}	s_d	$\cdots$	$\cdots$
c_1	0	0	$\cdots$	0	0	$\cdots$	$\cdots$
c_2	0	0	$\cdots$	1	0	$\cdots$	$\cdots$
c_3	0	0	$\cdots$	1	1	$\cdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$
c_{2^d-1}	0	1	$\cdots$	1	0	$\cdots$	$\cdots$
c_{2^d}	0	1	$\cdots$	1	1	$\cdots$	$\cdots$
c_{2^d+1}	1	0	$\cdots$	0	1	$\cdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$
$c_{2^{d+1}}$	1	1	$\cdots$	1	1	$\cdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$

} $c(s_0) = 0$

Among $\{c_1, \ldots, c_{2^d}\}$, pick 2^k concepts that correspond to codewords of $E : \{0,1\}^k \to \{0,1\}^d$ on $\{s_1, \ldots, s_d\}$

Proof approach: Pretty Good Measurement

State identification: Ensemble $\mathcal{E} = \{(p_z, |\psi_z\rangle)\}_{z\in[m]}$

- Given state $|\psi_z\rangle \in \mathcal{E}$ with prob p_z Goal: identify z
- Optimal measurement could be quite complicated, but we can always use the Pretty Good Measurement
- Crucial property: $P_{opt} \geq P_{pgm} \geq P_{opt}^2$ (Barnum-Knill'02)

How does learning relate to identification?

- Given $|\psi_{c^z}\rangle = \left|E_{c^z,D}\right\rangle^{\otimes T}$, learn c^z *approximately*. Show $T \geq d/\varepsilon$
- Suppose $\{s_0, \ldots, s_d\}$ is shattered by $\mathcal{C}$. Fix a nasty distribution D: $D(s_0) = 1 - 16\varepsilon$, $D(s_i) = 16\varepsilon/d$ on $\{s_1, \ldots, s_d\}$
- Let $E : \{0,1\}^k \to \{0,1\}^d$ be a good error-correcting code s.t. $k \geq d/4$ and $d_H(E(y), E(z)) \geq d/8$
- Pick concepts $\{c^z\}_{z\in\{0,1\}^k} \subseteq \mathcal{C}$: $c^z(s_0) = 0$, $c^z(s_i) = E(z)_i$ $\forall$ i
- *Learning* c^z approximately (wrt D) is equivalent to *identifying* z!

Sample complexity lower bound via PGM

Recap

- Learning c^z approximately (wrt D) is equivalent to identifying z!
- If sample complexity is T, then there is a good learner that *identifies* z from $|\psi_{c^z}\rangle = |E_{c^z,D}\rangle^{\otimes T}$ with probability $\geq 1 - \delta$

Analysis of PGM

- For the ensemble $\{|\psi_{c^z}\rangle : z \in \{0,1\}^k\}$ with uniform probabilities $p_z = 1/2^k$, we have $P_{pgm} \geq P_{opt}^2 \geq (1-\delta)^2$
- $P_{pgm} \leq \cdots$ 4-page calculation $\cdots \leq \exp(T^2\varepsilon^2/d + \sqrt{Td\varepsilon} - d - T\varepsilon)$
- This implies $T = \Omega(d/\varepsilon)$

Random classification noise

Classical model

- There is a fixed noise parameter $\eta \in [0,1]$
- A learning algorithm for $c \in \mathcal{C}$ obtains an (x,b) where $b = c(x)$ with probability $1-\eta$ and $b = 1 + c(x)$ with probability η
- Given such noisy examples, learn c

Quantum model

- There is a fixed noise parameter $\eta \in [0,1]$
- A quantum learner obtains

$$\sum_x \sqrt{D(x)}\,|x\rangle \left(\sqrt{1-\eta}\,|c(x)\rangle + \sqrt{\eta}\,|1+c(x)\rangle\right).$$

Given copies of this state, learn c

Strengths and weaknesses of noisy examples

- [AW'18] quantum noisy examples do not provide an advantage
- When D is the uniform distribution, even learning parities is open classically but quantum learning parities is possible in quantum polynomial time.

Agnostic learning

Lets get real!

- So far, examples were generated according to a target concept $c \in \mathcal{C}$
- In realistic situations we could have *"noisy" examples* for the target concept, or maybe *no fixed target concept* even exists

How do we model this? Agnostic learning

- Unknown distribution D on (x, ℓ) generates examples
- Suppose "best" concept in $\mathcal{C}$ has error $\text{OPT} = \min_{c \in \mathcal{C}} \Pr_{(x,\ell)\sim D}[c(x) \neq \ell]$
- Goal of the agnostic learner: output $h \in \mathcal{C}$ with error $\leq \text{OPT} + \varepsilon$

What about sample complexity?

- Classical sample complexity: $\Theta\left(\frac{d}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$ [VC74,Tal94]
- No quantum bounds known before (unlike PAC model)
- We show the quantum examples do not reduce sample complexity