

Lecture 12

* HW2 posted. Due on Oct 20.

RECAP

Learning with Random Classification Noise (RCN).

ORACLE : $E_{X \sim D}^{\pi}(c, \delta)$

- Draws $x \sim D$ from X
- With prob. $(1-\eta)$, return $(x, c(x))$
 " " π , return $(x, 1 - c(x))$

A concept class C is efficiently PAC learnable with RCN if it can be learnt in time $\text{poly}(d, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{(1-2\eta_0)})$.

Statistical Query Learning

Statistical Query Oracle: $\text{STAT}(C, D)$

Query: (ϕ, τ)

$$\phi: X \times \{0,1\} \rightarrow \{0,1\}$$

$$\tau \in (0,1)$$

$$\text{Let } P_\phi = \Pr_{x \sim D} [\phi(x, c(x)) = 1]$$

$\text{STAT}(C, D)$ returns \hat{P}_ϕ satisfying $\hat{P}_\phi \in [P_\phi - \tau, P_\phi + \tau]$

C is learnable in the SQ model if it can be learnt with access to $\text{STAT}(C, D)$ oracle to error ε in time $\text{poly}(d, \frac{1}{\varepsilon})$, where every query can be evaluated in time $\text{poly}(d, \frac{1}{\varepsilon})$ and has tolerance $\frac{1}{2}$ at least $\text{poly}(d, \frac{1}{\varepsilon})$.

Thm If C is efficiently SQ learnable using H , then C is efficiently PAC learnable using H .

TODAY

Thm: If ℓ is efficiently SQ learnable, then ℓ is PAC learnable in the presence of RCN.

Thm: The class of conjunctions is efficiently learnable in the SQ model. Therefore, it is efficiently learnable with RCN.

Proof (SQ learnability \Rightarrow learning with RCN)

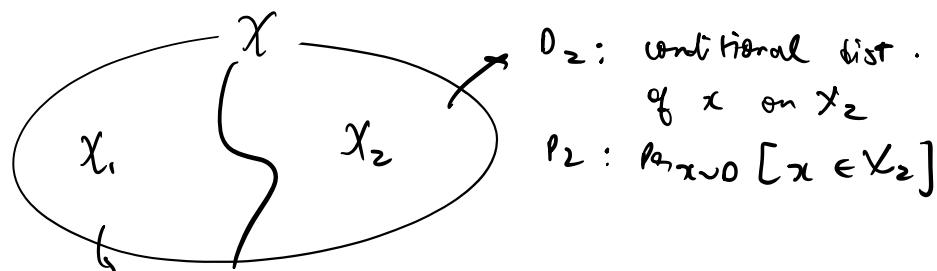
Access to $\text{Ex}^n(c, \delta)$ is simulate $\text{STAT}(c, \delta)$ (with some low failure probability).

Key idea

divide domain X into two disjoint parts

X_1 : all $x \in X$ s.t. $\phi(x, 0) \neq \phi(x, 1) \rightarrow$ need label

X_2 : all $x \in X$ s.t. $\phi(x, 0) = \phi(x, 1) \rightarrow$ no label needed



D₁: conditional dist. of x restricted to X_1 .

P₁: $\Pr_{x \sim D} [x \in X_1]$

$$\begin{aligned}
P_\phi &= \mathbb{E}_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi(x, \zeta(x))] = \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi=1] \wedge (x \in \mathcal{X}_1] \\
&\quad + \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi=1] \wedge (x \in \mathcal{X}_2)] \\
&= \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [x \in \mathcal{X}_1] \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi=1 \mid x \in \mathcal{X}_1] \\
&\quad + \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [x \in \mathcal{X}_2] \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi=1 \mid x \in \mathcal{X}_2] \\
&= \underbrace{\Pr_1 \cdot \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi=1]}_{\textcircled{1}} \underbrace{+ \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi=1 \wedge (x \in \mathcal{X}_2)]}_{\textcircled{2}} \\
&\quad \underbrace{\qquad\qquad\qquad}_{\textcircled{3}} \\
&= \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [\phi=1 \wedge (x \in \mathcal{X}_2)]
\end{aligned}$$

(3)
 { In \mathcal{X}_2 , ϕ does not depend on label.
 Easy to approximate with access to $\mathcal{E} \sim \mathcal{E}(\epsilon, 0)$ using concentration bounds.

Assumption: we know noise level η .

(1) P_1 : $P_1 = \Pr_{\mathcal{E} \sim \mathcal{E}(\epsilon, 0)} [x \in \mathcal{X}_1]$

Draw sample from $\mathcal{E} \sim \mathcal{E}(\epsilon, 0)$, ignore label

Check if $\phi(x, 0) \neq \phi(x, 1)$

→ if yes, $x \in X_1$.

→ o.w., $x \notin X_1$.

Can get enough samples, approximate fraction of data lying in X_1 ,

→ estimate to accuracy $O(\tau)$ with $O\left(\frac{1}{\tau^2} \log\left(\frac{1}{\delta}\right)\right)$ samples (w.p. $1 - \delta$).

② $P_{E \times^n(C, D)}[\phi = 1]$

$$\begin{aligned} P_{E \times^n(C, D)}[\phi = 1] &= (1-\eta) P_{E \times^n(C, D)}[\phi = 1] \\ &\quad + \eta P_{E \times^n(C, D)}[\phi = 0] \\ &= (1-\eta) P_{E \times^n(C, D)}[\phi = 1] \\ &\quad + \eta (1 - P_{E \times^n(C, D)}[\phi = 1]) \\ &= \eta + (1-2\eta) P_{E \times^n(C, D)}[\phi = 1] \end{aligned}$$

$$\therefore P_{E \times^n(C, D)}[\phi = 1] = \frac{P_{E \times^n(C, D)}[\phi = 1] - \eta}{1-2\eta}$$

→ we don't have access to D_1 .

→ Use rejection sampling. draw (x, b) from $E \times^n(C, D)$
Accept sample if $x \in X_1$.

→ Compute fraction of times $\phi(x, b) = 1$

→ Gives estimate of $\Pr_{\mathcal{E} \leftarrow \mathcal{M}(c, \delta)} [d=1]$ accurate to error τ with $O\left(\frac{L}{\tau^2} \log\left(\frac{1}{\delta}\right)\right)$ samples.

∴ We can estimate $\hat{P}_\phi = [\hat{P}_{\phi-\tau}, \hat{P}_{\phi+\tau}]$
with $O\left(\frac{L}{\tau^2} \log\left(\frac{1}{\delta}\right)\right)$ samples (w.p. $\geq 1 - \delta$).

Repeat this for every query ϕ that SQ algorithm makes.

∴ If we know n , we can use SQ algorithm to find $h \in \mathcal{H}$ with $\text{error}(h; c, \delta) \leq \epsilon_{100}$.

(If we set $\delta' = \delta/m$ where m is # queries made by algorithm, get failure probability $\leq \delta$)

Getting rid of assumption that we know n .

We only $n_0 \geq n$

for some Δ ,

Try all values of \hat{n} in set $\left\{ i\Delta : 0 \leq i \leq \left\lceil \frac{n_0}{\Delta} \right\rceil \right\}$

Atleast one of these values gets error $\leq \epsilon_{100}$.

∴ we only need to estimate true error.

Recall $\text{error}(h; c, \delta) = \Pr_{\mathcal{E} \leftarrow \mathcal{M}(c, \delta)} [h(+)] + c(\pi)]$

Let h_i be the hypothesis produced in i th iteration (i.e. noise level η_i)

Define $r_i = \Pr_{x \sim \mathcal{D}} [h_i(x) \neq b]$

$$r_i = (-n) \text{ error}(h_i) + n(1 - \text{error}(h_i))$$

$$\therefore \text{error}(h_i) = \frac{r_i}{1-2n} - \frac{n}{1-2n}$$

\therefore if we can estimate r_i accurately $\forall i$, & choose best h_i , we are done!

•

Characterizing what's learnable using SQ algo.

- Say that two functions f, g are uncorrelated if

$$\Pr_{x \sim D} [f(x) = g(x)] = \frac{1}{2}.$$

Definition (SQ dimension) The SQ-dimension of a class \mathcal{C} wrt a distribution D over X is the size of the largest subset $\mathcal{C}' \subseteq \mathcal{C}$ st. for all $f, g \in \mathcal{C}'$,

$$\left| \Pr_{x \sim D} [f(x) = g(x)] - \frac{1}{2} \right| < \frac{1}{|\mathcal{C}'|}.$$

(size of largest set. of nearly uncorrelated functions in \mathcal{C}).

SQ-dim characterizes learnability in SQ model.

Weak-learning: An algorithm A is a weak learner with edge/advantage γ for class \mathcal{C} if: for any dist. D over examples labelled by any target $c \in \mathcal{C}$, w.h.p A produces a hypothesis with $\text{error}(h; c, D) \leq \frac{1}{2} - \gamma$.

Well (hopefully) see: "Weak learning \Rightarrow Strong (PAC) learning".

Thm 1: If $\text{SQ-DIM}_D(\mathcal{C}) = \text{poly}(d)$, then you can efficiently "weakly-learn" \mathcal{C} over D (get error rate $\leq \frac{1}{2} - \frac{1}{\text{poly}(d)}$) using SQ-algorithm.

Thm 2: If $\text{SQ-DIM}_D(\mathcal{C}) > \text{poly}(d)$ then you cannot efficiently learn \mathcal{C} over D by SQ-algorithms (even "weak-learning" to error $\leq \frac{1}{2} - \frac{1}{\text{poly}(d)}$ is impossible)

Proof of Thm 1

Let $s > \text{SQ-DIM}_D(C)$

Let $\mathcal{H} \subseteq \mathcal{C}$ be maximal subset s.t. $\forall h_i, h_j \in \mathcal{H}$ we have

$$|\Pr_D[h_i(x) = h_j(x)] - \frac{1}{2}| < \frac{1}{s+1}$$

$$\therefore |\mathcal{H}| \leq s.$$

We try every $h_i \in \mathcal{H}$ & use SQ-oracle to estimate its error.

Claim: At least one h_i (or Th_i) must be weak predictor.

Proof If target c satisfied

$$|\Pr_D[c(x) = h_i(x)] - \frac{1}{2}| < \frac{1}{s+1} \quad \forall h_i \in \mathcal{H}$$

then we can include c in the set \mathcal{H} !

But \mathcal{H} is a maximal set \Rightarrow

\therefore at least one h_i satisfy,

$$|\Pr_D[c(x) = h_i(x)] - \frac{1}{2}| > \frac{1}{s+1},$$

\therefore either h_i or Th_i is a weak predictor.

Implication of Thm 2 Cannot learn PARITIES
in the SQ model (efficiently).

Parity function:

$$X^d = \{0, 1\}^d$$

$$Y = \{0, 1\}$$

$$\mathcal{C} = \{w(x) = \langle w, x \rangle \bmod 2 : w \in \{0, 1\}^d\}$$

Thm: PARITIES are efficiently PAC learnable.

Proof first, note that since $|\mathcal{C}| = 2^d$, an ERM algorithm will get error ϵ with $O\left(\frac{d}{\epsilon} \log(1/\delta)\right)$ samples.

We claim we can implement an ERM algorithm in time $\text{poly}(d)$. Let $n = O\left(\frac{d \log(1/\delta)}{\epsilon}\right)$

Get examples $\{(a_1, b_1), \dots, (a_n, b_n)\}$

linear system over \mathbb{F}_2

$$\begin{pmatrix} & a_1 & \\ & a_2 & \\ & \vdots & \\ & a_n & \end{pmatrix} \begin{pmatrix} | \\ w \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Can solve in time $\text{poly}(d)$ using Gaussian elimination.

Let D be the uniform distribution over $\{0,1\}^d$.

Claim: Any two parity functions $(w_1(x))$ & $(w_2(x))$ (where $w_1 \neq w_2$) are uncorrelated:

$$\Pr_D [(w_1(x) = w_2(x))] = \frac{1}{2}.$$