Lecture 12

* HW2 posted. Due on Oct 20.

RECAP

Learning with Random Classification Noise (RCN).

ORACLE : $EX^{\eta}(c, D)$

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

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

Statistical Query Learning

Statistical Query oracle : $STAT(c, D)$

Query : (ϕ, τ)

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

$\tau \in (0,1)$

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

$STAT(c, D)$ returns $\hat{P}_\phi$ 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 $STAT(c, D)$ oracle to error ε in time $poly(d, \frac{1}{\varepsilon})$, where every query can be evaluated in time $poly(d, \frac{1}{\varepsilon})$ and has tolerance $\frac{1}{2}$ atleast $poly(d, \frac{1}{\varepsilon})$.

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

TODAY

Thm: If $\mathcal{C}$ is efficiently SQ learnable, then $\mathcal{C}$ 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 => learning with RCN)

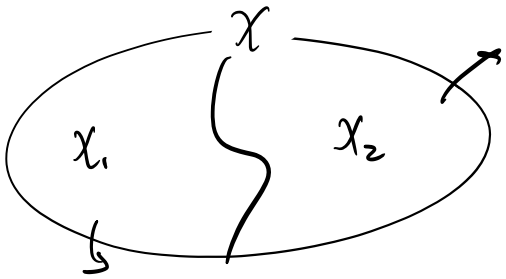
Access to $EX^{\eta}(c, D)$ is simulate $STAT(c, D)$ (with some low failure probability).

Key idea

divide domain X into Two disjoint parts

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

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



D_2: conditional dist. of x on X_2

P_2: $Pr_{x \sim D}[x \in X_2]$

D_1: conditional dist. of x restricted to X_1

P_1: $Pr_{x \sim D}[x \in X_1]$

$$P_\phi = \mathop{\mathbb{E}}_{Ex^\eta(c,D)}[\phi(x, c(x))] = \Pr_{Ex^\eta(c,D)}[(\phi=1) \wedge (x \in X_1)] + \Pr_{Ex^\eta(c,D)}[(\phi=1) \wedge (x \in X_2)]$$

$$= \Pr_{Ex^\eta(c,D)}[x \in X_1] \Pr_{Ex^\eta(c,D)}[\phi=1 \mid x \in X_1] + \Pr_{Ex^\eta(c,D)}[(\phi=1) \wedge (x \in X_2)]$$

$$= \underset{①}{P_1} \cdot \Pr_{Ex^\eta(c,D_1)}[\phi=1] + \Pr_{Ex^\eta(c,D)}[(\phi=1) \wedge (x \in X_2)]$$

$$= P_1 \cdot \Pr_{Ex^\eta(c,D_1)}[\phi=1] + \Pr_{Ex^\eta(c,D)}[(\phi=1) \wedge (x \in X_2)]$$

$$= P_1 \cdot \Pr_{Ex^\eta(c,D_1)}[\phi=1] + \Pr_{Ex^\eta(c,D)}[(\phi=1) \wedge (x \in X_2)]$$

③ In X_2, ϕ does not depend on label.

easy to approximate with access to $Ex^\eta(c,D)$ using concentration bounds.

Assumption: we know noise level η.

① P_1: $\quad P_1 = \Pr_{Ex^\eta(c,D)}[x \in X_1]$

Draw sample from $Ex^\eta(c,D)$, ignore label

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

→ if yes, $x \in X$.

→ otw, $x \notin X$.

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 $\Pr_{Ex(c,D_1)}[\phi=1]$

$$\Pr_{Ex^\eta(c,D_1)}[\phi=1] = (1-\eta)\Pr_{Ex(c,D_1)}[\phi=1] + \eta \Pr_{Ex(c,D_1)}[\phi=0]$$

$$= (1-\eta)\Pr_{Ex(c,D_1)}[\phi=1] + \eta\left(1-\Pr_{Ex(c,D_1)}[\phi=1]\right)$$

$$= \eta + (1-2\eta)\Pr_{Ex(c,D_1)}[\phi=1]$$

$$\therefore \Pr_{Ex(c,D_1)}[\phi=1] = \frac{\Pr_{Ex^\eta(c,D_1)}[\phi=1] - \eta}{1-2\eta}$$

→ we don't have access to D_1.

→ Use rejection sampling. draw (x,b) from $Ex^\eta(c,D)$ Accept sample if $x \in X_1$.

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

→ Gives estimate of $Pr_{Exn(c,D)}[\phi = 1]$ accurate to error τ with $O\left(\frac{1}{\tau^2}\log\left(\frac{1}{\delta'}\right)\right)$ samples.

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

Repeat this for every query ϕ that SQ algorithm makes.

∴ If we know η, we can use SQ algorithm to find $h \in \mathcal{H}$ with $error(h; c, D) \leq \epsilon/100$.

(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 η.

We only $\eta_0 \geq \eta$

for some Δ,

Try all values of $\hat{\eta}$ in set $\left\{i\Delta : 0 \leq i \leq \left\lceil \frac{\eta_0}{\Delta} \right\rceil\right\}$

Atleast one of these values gets error $\leq \epsilon/100$.

∴ we only need to estimate true error.

Recall $error(h; c, D) = Pr_{e \leftarrow (c,D)}[h(x) \neq c(x)]$

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

Define $$r_i = \Pr_{EX^\eta(c,D)}[h(x) \neq b]$$

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

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

$\therefore$ 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 $\mathcal{C}' \subseteq \mathcal{C}$ s.t. 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 & any target $c \in \mathcal{C}$, given access to $EX(c, D)$, w.p. $(1-\delta)$ A produces a hypothesis with $\text{error}(h, c, D) \leq \frac{1}{2} - \gamma$.

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

Thm 1: If $SQ\text{-}DIM_D(\mathcal{C}) = \text{poly}(d)$, 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 $SQ\text{-}DIM_D(\mathcal{C}) > \text{poly}(d)$ 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 $s = SQ\text{-}DIM_D(C)$

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

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

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

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

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

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

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

$\therefore$ at least one h_i satisfy,

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

$\therefore$ either h_i or Th_i is a weak predictor. ∎

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

Parity function:

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

$Y = \{0,1\}$

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

Thm: PARITIES are efficiently PAC learnable.

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

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

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

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 poly(d) using Gaussian elimination.

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

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

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