Lecture 11

* HW1 due today
* I'll send a survey about HW1 & lectures, do fill!

RECAP

Last time:

* Cryptographic assumptions => Hardness of improper learning
* Agnostically learning halfspaces is hard over some complexity assumptions (hardness of "refuting random k-XOR")
* Agnostic learning is quite a challenging model.

> **Thm** There is no efficient proper learning algorithm for agnostically learning rectangles in $\mathbb{R}^d$ & conjuctions, unless $RP = P$.

Agnostic learning allows "worst-case noise" which makes learning hard.

Today: Learning under **random** noise.

Learning with Random Classification Noise (RCN).

* Only the labels are noisy
* All labels are flipped with probability η (white-noise model)

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

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

It costs one unit time to make call to $EX^{\eta}(c, D)$

Criterion for success

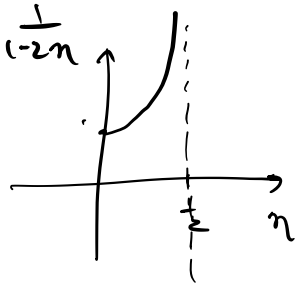
With $1-\delta$, find hypothesis h s.t.

$$\text{error}(h; c, D) = \Pr_{x \sim D}[c(x) \neq h(x)] \leq \varepsilon.$$

(comparing with true labels)

Restrict η to lie in $[0, \frac{1}{2})$

(learning impossible with $\eta = 1/2$)



As $\eta \to \frac{1}{2}$, learning becomes harder.

Allow factors of $\left(\frac{1}{1-2\eta}\right)$ in complexity bounds.

Definition (PAC learning with RCN)

Let $\mathcal{C}$ be a concept class and $\mathcal{H}$ be a hypothesis class over X^d. We say that $\mathcal{C}$ is efficiently PAC learnable with RCN if : $\exists$ a learning algo. A s.t. $\forall\ c \in \mathcal{C}$, $\forall$ dist. D over X^d, $\forall\ \epsilon, \delta \in (0, \frac{1}{2})$, $\forall\ \eta \in (0, \frac{1}{2})$ if A is given access to $EX^{\eta}(c, D)$ and inputs ϵ, δ & η_0 s.t. $\eta_0 \in (\eta, \frac{1}{2})$, with prob $1-\delta$, A outputs h s.t. $\text{error}(h; c, D) \leq \epsilon$ (Prob. includes randomness in noisy labels).

We say that $\mathcal{C}$ is efficiently PAC learnable if A runs in time $\text{poly}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{1-2\eta_0}\right)$.

Alg. for learning conjunctions

Previous alg. fails miserably on random noise!

Consider a robust algo. that relies on aggregate statistics instead of a single example.

- Suppose ℓ is a literal ($\bar{x}_i$ or x_i)

Define $p_0(\ell) = \Pr_{a \sim D}[\ell \text{ is } 0 \text{ in } a]$

$p_{bad}(\ell) = \Pr_{a \sim D}[\ell \text{ is } 0 \text{ in } a \wedge c(a) = 1]$

Note that if ℓ is in $c(x)$, then $p_{bad}(\ell) = 0$.

→ ℓ is <u>significant</u> if $p_0(\ell) \geq \varepsilon/8d$.

→ ℓ is <u>harmful</u> if $p_{bad}(\ell) \geq \varepsilon/8d$.

Claim: If h is conjunction of all <u>significant</u> literals that are not <u>harmful</u>, then $error(h; c, D) \leq \varepsilon$.

Proof

$$error(h; c, D) = \Pr_{a \sim D}[h(a) = 1 \wedge c(a) = 0] + \Pr_{a \sim D}[h(a) = 0 \wedge c(a) = 1]$$

Caused if $h(x)$ does not have some literal which $c(x)$ has (& that literal is set to 0 in example a).

∴ Caused due to insignificant literals

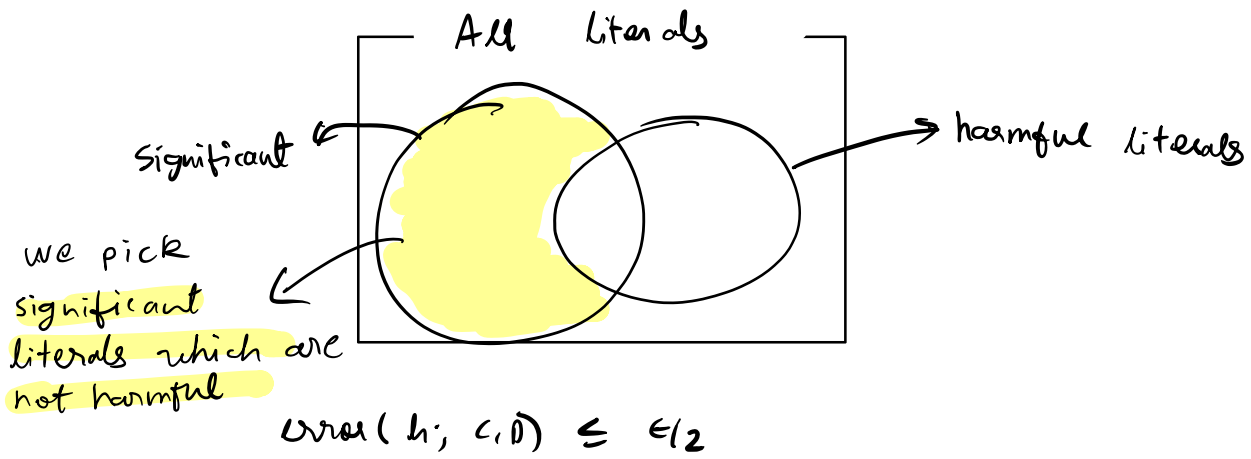
$\leq 2d \cdot \frac{\varepsilon}{8d} \quad \frac{\varepsilon}{4}$

Caused if $h(x)$ has a literal which $c(x)$ does not have (& that literal is set to 0 in a)

$h(x)$ does not have any harmful literals

∴ Caused due to significant literals which are not harmful

$\leq 2d \cdot \frac{\varepsilon}{8d} \quad \frac{\varepsilon}{4}$



$error(h, c, D) \leq \epsilon/2$

Still in noiseless case:

for any literal ℓ, $p_o(\ell)$ & $p_{bad}(\ell)$ can be estimated based on samples from $Ex(c, D)$. Can take a large set to get concentration with Chernoff.

This implies we can PAC learn conjunctions using this algorithm.

This algorithm can also be run with $Ex^{\eta}(c, D)$!

Statistical Query learning (Kearns ~90)

Do not have access to example oracle.

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 $\hat{P}_\phi \in [P_\phi - \tau, P_\phi + \tau]$

Examples

Consider a spam classification task.

"What fraction of emails labelled as spam (label is 1), have the words "Urgent" & "free", but not "USC", and the number of words is ≤ 20?"

for our conjunction algorithm :

$P_0(\ell) = \Pr_{a \sim D}(\ell \text{ is } 0 \text{ in } a)$

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

$\phi_\ell(x, y) = \begin{cases} 1 & \text{if } \ell \text{ is } 0 \text{ in } a \\ 0 & \text{otw} \end{cases}$

$$p_{\phi_\ell} = \mathbb{E}_{x \sim D}\left[\phi_\ell(x, c(x))\right] = p_0(\ell)$$

$$p_{bad}(\ell) = \Pr_{a \sim D}\left(\ell \text{ is } 0 \text{ in } a \;\&\; c(a) = 1\right)$$

$$\phi'_\ell(x, y) = \begin{cases} 1 & \text{if } \ell \text{ is } 0 \text{ in } a \;\&\; y = 1 \\ 0 & \text{otw} \end{cases}$$

$$p_{\phi'_\ell} = \mathbb{E}_{x \sim D}\left[\phi'_\ell(x, c(x))\right] = p_{bad}(\ell).$$

Definition (SQ learning)

Let $\mathcal{C}$ be a concept class and $\mathcal{H}$ be a hypothesis class. We say $\mathcal{C}$ is <u>efficiently</u> learnable from statistical queries (SQ) using $\mathcal{H}$, if $\exists$ alg. A, and polynomials $p(\cdot,\cdot)$, $q(\cdot,\cdot)$ & $r(\cdot,\cdot)$ s.t. $\forall\ c \in \mathcal{C}$, $\forall$ dist. D over X^d, $\forall\ \epsilon \in (0, \frac{1}{2})$, if A is gives access to $STAT(c, D)$ and input ϵ then,

- $\forall$ query (ϕ, τ) made by A, the predicate ϕ can be evaluated in time $q(\frac{1}{\epsilon}, d)$ and $\frac{1}{\tau}$ is bounded by $r(\frac{1}{\epsilon}, d)$.

- A runs in time $p(\frac{1}{\epsilon}, d)$

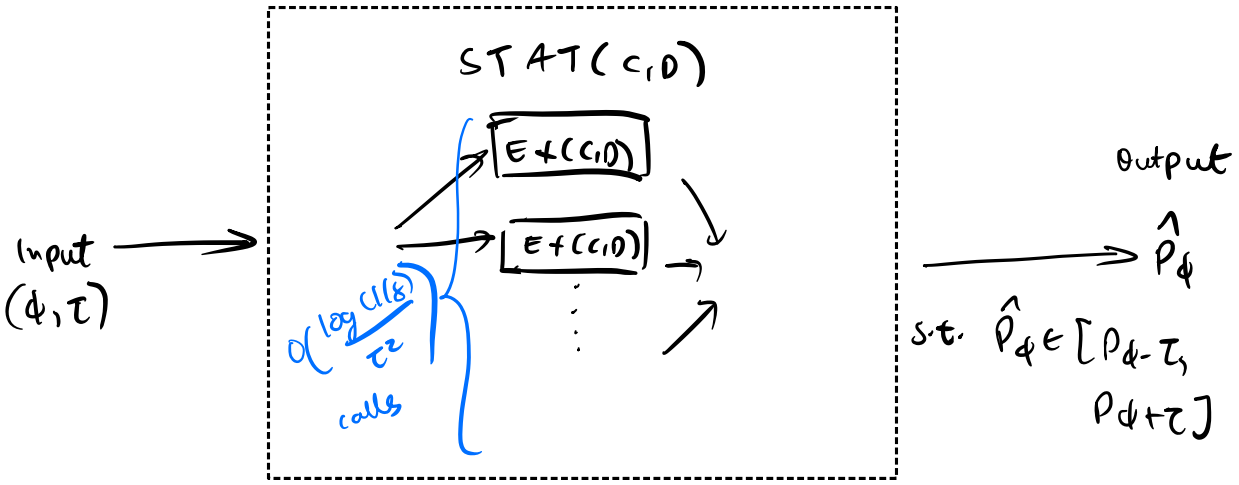
- A outputs $h \in \mathcal{H}$ s.t. $error(h; c, D) \leq \epsilon$.

Note: no confidence parameter δ. Before, δ took care of the probability of seeing a set of "bad examples". But the $EX(c,D)$ has been replaced by $STAT(c,D)$, which is deterministic.

Thm If $\mathcal{C}$ is efficiently SQ learnable using $\mathcal{H}$, then $\mathcal{C}$ is efficiently PAC learnable using $\mathcal{H}$.

Proof Exercise. More details below.

We can implement $STAT(c,D)$ using $Ex(c,D)$



By Chernoff, if we take $O\left(\frac{\log(1/\delta)}{\tau^2}\right)$ samples from $Ex(c,D)$ then $\hat{p}_\phi$ satisfy

$$\hat{p}_\phi \in [p_\phi - \tau, p_\phi + \tau] \quad \text{w.p. } (1-\delta)$$

Aside: Differential privacy. If $\mathcal{C}$ is efficiently SQ learnable, then $\mathcal{C}$ is efficiently PAC learnable with differential privacy.

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 Our conjunction learning algo. can be implemented in the SQ model. ∎