Learning with Statistical Queries

Sunday, November 7, 2021 7:43 PM

Yunfan

PAC model ✓

What about errors in the labels ?

One model for this :

Random Classification Noise Model

each label is flipped with prob. η.

$$(x, \ell(x)) \quad \ell(x) = \begin{cases} f(x) & \text{w.p. } 1-\eta \\ 1-f(x) & \text{w.p. } \eta \end{cases}$$

Q. Can we still learn the underlying concept?

E.g. suppose we are learning an <u>OR</u>.

of Boolean variables.

Let $p_i = \Pr(f(x)=0 \text{ and } x_i = 1)$

We can let output hypothesis h be

all variables for which $p_i = 0$

(any variable with $p_i = 0$ is in the true OR) (any variable with $p_i = 0$ ...

and no variables for which $p_i > \frac{\varepsilon}{n}$.

So can we estimate each $p_i \pm \frac{\varepsilon}{2n}$ (say)?

$$p_i = \underbrace{\Pr\left(f(x)=0 / x_i = 1\right)}_{q_i} \cdot \Pr(x_i = 1)$$

independent of label noise

$$\Pr_{\text{noise } \eta}\left(l(x)=0 / x_i=1\right) = (1-\eta)\, q_i + \eta\,(1-q_i)$$
$$= \eta + q_i(1-2\eta)$$

So if we have LHS, we can subtract η, divide by $(1-2\eta)$.

suffices to approximate to within $\pm \frac{\varepsilon}{2n}(1-2\eta)$.
which we can do from samples!

Statistical Query Model.
Can ask for expectations of bounded functions $g(x, l(x))$ up to additive

[illegible]
functions of $(x, \ell(x))$ up to additive error.

E.g. $\mathbb{E}(x_i \mid \ell(x) = 1)$

$\mathbb{E}(x_i(1 - x_j))$

$$\boxed{\begin{array}{l} \chi : \underset{\text{example}}{X} \times \underset{\text{label}}{\{0,1\}} \to [0,1], \quad \tau > 0 \\[1em] \text{SQ oracle responds with } \mathbb{E}_D(\chi) \pm \tau. \end{array}}$$

χ: poly time computable $\quad \tau \geq \frac{1}{\text{poly}}$

Many (almost all) known learning algorithms can be implemented with SQ.

e.g. gradient descent. $\quad L(w) = \mathbb{E}_{x,y}(\ell(x,y,w))$

$$\nabla_w L(w) = \nabla_w \mathbb{E}_{x,y}(\ell(x,y,w))$$

$$= \mathbb{E}_{x,y}(\nabla_w(\ell(x,y,w))$$

$$= \mathbb{E}_{x,y}\left(\nabla_w(\ell(x,y,w))\right)$$

SQ. $\longrightarrow$

Thm. SQ-learnable $\Rightarrow$ PAC learnable with Random classification noise.

Estimate $\Pr(\chi(x, f(x)) = 1)$

Let

CLEAN $= \{x : \chi(x,0) = \chi(x,1)\}$

NOISY $= \{x : \chi(x,0) \neq \chi(x,1)\}$

example $\chi(x,\ell) = 1$ if $x_i = 1$ and $\ell = 0$.

$$\Pr(\chi(x,f(x))=1) = \Pr(\chi(x,f(x))=1 \text{ and } x \in \text{CLEAN}) + \Pr(\chi(x,f(x))=1 \text{ and } x \in \text{NOISY})$$

Can. estimate $\Pr(x \in \text{CLEAN})$ — not affected by noise.

and hence first term.

Also $\Pr(x \in \text{NOISY})$

Also $\Pr(x \in \text{NOISY})$

and $\Pr_\eta(\chi(x, f(x)) = 1 \mid x \in \text{NOISY}) \} q$

$= (1-\eta)p + \eta(1-p) = \eta + p(1-2\eta)$

where $p = \Pr(\chi(x, f(x)) = 1 \mid x \in \text{NOISY})$

Hence $p = \dfrac{q - \eta}{1 - 2\eta}$.

Need to estimate p to within $\pm \tau$

So need to estimate q to within $\pm \tau(1-2\eta)$

Some functions are hard to learn with SQ.

e.g. parity!

Suppose the hypothesis class is the set of parities. How many $SQ(\tau)$ do we need?

Thm Weak Learning PARITY needs $2^{\Omega(n)}$ SQs.

Thm. Weak Learning PARITY needs 2 SQs.

Pf. Recall that $\langle \chi_S, \chi_T \rangle_D = \begin{cases} 1 & S = T \\ 0 & \text{o.w.} \end{cases}$

Take an SQ $\chi(x, \ell(x))$

it is function from $\{-1,1\}^n \times \{-1,1\} \longrightarrow [-1,1]$

Let us extend $\{\chi_S\}$ to a basis for $\{-1,1\}^{n+1}$

$$\chi_S(x,\ell) = \chi_S(x)$$

$$h_S(x,\ell) = \ell(x) \cdot \chi_S(x)$$

These are 2^{n+1} functions.

Check: $\langle h_S, \chi_T \rangle = 0$

$$\langle h_S, h_T \rangle = \begin{cases} 1 & S = T \\ 0 & \text{o.w.} \end{cases}$$

So any query g can be written as

So any query g can be written as

$$g(x,\ell(x)) = \sum_S \alpha_S h_S(x,\ell) + \underbrace{\sum \hat{f}_S \chi_S(x)}_{\text{independent of } \ell(x)}$$

$\therefore$

$$\mathbb{E}_D\big(g(x),\ell(x)\big) = \sum_S \alpha_S \mathbb{E}_D\big(\ell(x)\chi_S(x)\big) + g_0$$

Now $\ell(x) = \chi_T$ for some T. So the first term is 0 except when $S = T$, when we get

$$= \alpha_T$$

$\# T$ for which $\alpha_T \geq \tau$ is $\leq \frac{1}{\tau^2}$.

(since $\sum_S \alpha_S^2 + \sum_S \hat{f}_S^2 \leq 1$)

So if the SQ oracle responds "0", at most $\frac{1}{\tau^2}$ are eliminated.

at least $2^n . \tau^2$ queries.

$\therefore$ need at least $2^{n} \cdot \tau^{2}$ queries.

More generally,

$$\text{SQ-dim}(\mathcal{H}, \gamma) = \max \{ d : \exists\ d \text{ concepts } h_1, h_2, \dots h_d \in \mathcal{H} \text{ s.t. } \|h_i\|^2 \leq 1,\ \langle h_i, h_j \rangle_D \leq \gamma \}.$$

Thm. To learn $\mathcal{H}$ in the SQ model, we need $\Omega(d\gamma)$ queries to $SQ(\sqrt{\gamma})$.