Subject: Intro. to Info. Theory

No.: 1

Date: 12 / 31 / 2019

* $$P_*(S_t \mid S_{t-1}) = \frac{P_*(S_t)}{P_*(S_{t-1})} \in (0,1)$$

$$P_*(\lambda \mid S_{t-1}) = 1 - P_*(0 \mid S_{t-1}) - P_*(1 \mid S_{t-1})$$

* if do probabilistic prediction

$$\hat{S}_t = \begin{cases} 1 & \\ 0 & \text{with prob.} \\ \lambda & \end{cases} \quad \begin{matrix} P_*(1 \mid S_{t-1}) \\ P_*(0 \mid S_{t-1}) \\ P_*(\lambda \mid S_{t-1}) \end{matrix}$$

$$\underbrace{[\![\hat{S}_t \neq S_t]\!]}_{\text{error } e_t} \quad \text{w/ prob.} \quad \underbrace{1 - P_*(S_t \mid S_{t-1})}_{q_t}$$

$$\underbrace{\mathbb{E}\left(\sum_{t=1}^{T} e_t\right)}_{\text{total error}} = \sum_{t=1}^{T} q_t$$

$$= \sum_{t=1}^{T} \left(1 - \frac{P_*(S_t)}{P_*(S_{t-1})} \right)$$

$$\leq \sum_{t=1}^{T} -\ln \frac{P_*(S_t)}{P_*(S_{t-1})}$$

$$= -\sum_{t=1}^{T} \left(\ln P_*(S_t) - \ln P_*(S_{t-1}) \right)$$

$$= -\ln P_*(S_T) + \cancel{\ln P_*(S_0)}$$

$$\leq -\ln 2^{-\underbrace{K_*(S_T)}}$$

↓

(prefix) Kolmogorov complexity of generating S_T *

$$= K_*(S_T) \cdot \ln 2$$

$$\leq \underbrace{\left| \text{actual program that generates } S_\infty \right| \cdot \ln 2}_{\text{"constant"}}$$

avg. error $\leq \frac{1}{T} \cdot$ constant $\Rightarrow$ $O(\text{complexity of } S_\infty)$ "examples" to learn (Solomonoff Inference)

* if do deterministic prediction

$$\tilde{s}_t = \underset{s \in \{1,0,2\}}{\operatorname{argmax}} \; P_*(s \mid s_{t-1})$$

if $[\tilde{s}_t \neq s_t] \implies P_*(s_t \mid s_{t-1}) < \frac{1}{2}$

error $\tilde{e}_t$

$$\Updownarrow$$

$$q_t \geq \frac{1}{2}$$

$$\sum_{t=1}^{T} \tilde{e}_t \leq \sum_{t=1}^{T} [q_t \geq \tfrac{1}{2}]$$

$$\leq \sum_{t=1}^{T} 2 \cdot q_t$$

$$\vdots$$

* learning :

consider a hypothesis set $\mathcal{H} = \{h\}$

and examples $D = \{(x_n, y_n = h_*(x_n))\}_{n=1}^{N}$

generated iid from some $P(\underline{x})$ and $h_* \in \mathcal{H}$

called realizable

$s \rightarrow$ examples (x_n, y_n)

programs $\rightarrow$ hypotheses h

program that generates s_∞ $\rightarrow$ generating hypothesis h_*

error on future bits $\rightarrow$ error on future $x \sim P(x)$ and $y = h_*(x)$

$$e(h) = \underset{x \sim P(x)}{\mathbb{E}}\left([h(x) \neq h_*(x)] \right)$$

goal : find $g \in \mathcal{H}$ s.t. $e(g) \leq \epsilon$ w/ high probability

how many examples (N) needed ?

Subject : No. : 3 Date : 12/31/2019

* conclusion

training error ($e_D(g) = 0$)

if always choosing one $g \in \mathcal{H}$ s.t. $g(x_n) = y_n$ for all n

(one must exists because $h_* \in \mathcal{H}$)

and have a preference $p(h)$ on $h \in \mathcal{H}$

w/ $p(h) \geq 0$, $\sum_h p(h) = 1$ "prior"

then for any given $0 < \delta \leq 1$

$$\Pr_{D \sim P^N} \Big(\underbrace{\forall h \text{ s.t. } e_D(h) = 0}_{\text{i.e. "}g\text{"}} , \quad e(h) \leq \frac{1}{N}\Big(\ln\frac{1}{\delta} + \ln\frac{1}{p(h)}\Big)\Big) \gtrsim 1-\delta$$

* proof : consider one h with $e(h) = \lambda$

$$\Pr(e_D(h) = 0) = (1-\lambda)^N = e^{N\ln(1-\lambda)} \leq e^{-N\lambda}$$

want: $\Pr(e_D(h)=0) < \delta \cdot p(h)$

$$\Uparrow$$

$$e^{-N\lambda} < \delta \cdot p(h)$$

$$\Uparrow$$

$$\lambda > \frac{1}{N}\Big(\ln\frac{1}{\delta} + \ln\frac{1}{p(h)}\Big)$$

So w/ prob $\geq 1 - \delta\, p(h)$

$$e(h) \leq \frac{1}{N}\Big(\ln\frac{1}{\delta} + \ln\frac{1}{p(h)}\Big)$$

$\Downarrow$ union bound over $e_D(h) = 0$

w/ prob $\geq 1 - \delta \sum_{e_D(h)=0} p(h) \geq 1-\delta$

$\forall h$ w/ $e_D(h) = 0$, $\quad e(h) \leq \frac{1}{N}\Big(\ln\frac{1}{\delta} + \ln\frac{1}{p(h)}\Big)$

\# examples (N) proportion to self information $\ln\frac{1}{p(h)}$

to reach certain $e(h)$ guarantee