BITS F464: Machine Learning

23 PAC Learning

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

March 17, 2021 ONLINE (Campus @ BITS-Pilani Jan-May 2021)

`http://ktiwari.in/ml`

PAC learning

Concept class $\mathcal{C}$ is said to be PAC-learnable if $\exists$ and algorithm $\mathcal{A}$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target function $c \in \mathcal{C}$

$$P_{S \sim \mathcal{D}^m}[\mathcal{R}(h_S) \leq \epsilon] \geq 1 - \delta$$

holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$

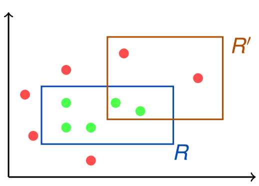
where, $\mathcal{S} = (x_1, x_2, ..., x_m)$ is training samples, and

- $poly(.,.,.,.)$ is a polynomial function
- n is such that the computation cost to represent $x \in \mathcal{X}$ is $O(n)$
- $size(c)$ is maximum computational requirement of $c \in \mathcal{C}$

We have high confidence ($\geq 1 - \delta$) that error would be small ($\leq \epsilon$)

If $\mathcal{A}$ runs in $poly(1/\epsilon, 1/\delta, n, size(c))$ then $\mathcal{C}$ is efficiently PAC-learnable

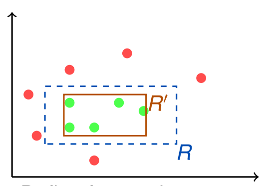
Example: learning axis aligned rectangles



- Instances are points in 2D space, and $R \in \mathcal{C}$
- Hypothesis R' could have some false positive and false negative
- What is the hypothesis space? All possible axis aligned rectangles
- How many parameters? 4, rectangle is (l,b) to (r,t)
- Let the learning algorithm $\mathcal{A}$ returns tightest axis aligned rectangle

Example: contd..

$\mathcal{A}$ returns tightest axis aligned rectangle. $R' = R_S$



- By definition R_S does not have false positive
- Fix some ϵ and let $P(R) \geq \epsilon$
- Define four region r_1, r_2, r_3 and r_4 such that $P(r_i) \geq \epsilon/4$
- [By geometry] If R_S meets all these regions, it will have one side in each of these regions.
- Error area of R_S is the part of R that is not covered by R_S. Note error area would be included in the union of r_i

Example: contd..

$$\begin{aligned}
P_{S \sim \mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] &\leq P_{S \sim \mathcal{D}^m}[\cup_{i=1}^4 \{R_S \bigcap r_i = \phi\}] \\
&\leq \textstyle\sum_{i=1}^4 P_{S \sim \mathcal{D}^m}[\{R_S \bigcap r_i = \phi\}] && \text{by union bound} \\
&\leq 4(1 - \epsilon/4)^m && \text{as } P[r_i] \geq \epsilon/4 \\
&\leq 4.e^{-m\epsilon/4} && \text{as } 1 - x \leq e^{-x}
\end{aligned}$$

To ensure $P_{S \sim \mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] \leq \delta$ we have

$$4.e^{-m\epsilon/4} \leq \delta$$

that gives

$$m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}$$

For any $\epsilon > 0$ and $\delta > 0$, if the sample size m is greater then $\frac{4}{\epsilon} \log \frac{4}{\delta}$, then $P_{S \sim \mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] \leq \delta$ so this concept class is PAC-learnable.

Generalization Bound

Generalization Bound
With probability at least $1 - \delta$ error $\mathcal{R}(R_S)$ is upper bounded by ϵ

We have seen

$$P_{S \sim \mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] \leq \delta$$

With

$$\delta = 4.e^{-m\epsilon/4}$$

we have

$$\epsilon = \frac{4}{m} \log \frac{4}{\delta}$$

So,

$$\mathcal{R}(R_S) \leq \frac{4}{m} \log \frac{4}{\delta}$$

Guarantees for finite hypothesis set

Theorem: Let $|\mathcal{H}|$ be a finite set of functions mapping $\mathcal{X} \to \mathcal{Y}$. Let for any target concept $c \in \mathcal{H}$ and i.i.d samples S, an algorithm $\mathcal{A}$ returns consistent hypothesis $h_S : \hat{\mathcal{R}}(R_S) = 0$. Then for any $\epsilon, \delta > 0$, the inequality $P_{S\sim\mathcal{D}^m}[\mathcal{R}(h_S) \leq \epsilon] \geq 1-\delta$ holds if

$$m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)$$

This result equivalently admits: for any $\epsilon, \delta > 0$, with probability at least $1-\delta$

$$\mathcal{R}(h_S) \leq \frac{1}{m}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)$$

Proof: Fix $\epsilon > 0$ and let $\mathcal{H}_\epsilon = \{h \in \mathcal{H} : R(h) > \epsilon\}$. Probability that a hypothesis $h \in \mathcal{H}_\epsilon$ is consistent on a training sample S drawn i.i.d is bounded as

$$P[\hat{\mathcal{R}}_S(h) = 0] \leq (1-\epsilon)^m$$

Guarantees for finite hypothesis set (Contd...)

$$P[\hat{\mathcal{R}}_S(h) = 0] \leq (1-\epsilon)^m$$

Thus, by union bound

$$\begin{aligned}
P[\exists h \in \mathcal{H}_\epsilon : \hat{\mathcal{R}}_S(h) = 0] &= P[\hat{\mathcal{R}}_S(h_1) = 0 \vee \hat{\mathcal{R}}_S(h_2) = 0 \vee \ldots \\
&\qquad \vee \hat{\mathcal{R}}_S(h_{|\mathcal{H}_\epsilon|}) = 0] \\
&\leq \textstyle\sum_{h\in\mathcal{H}_\epsilon} P[\hat{\mathcal{R}}_S(h) = 0] && \text{union bound} \\
&\leq \textstyle\sum_{h\in\mathcal{H}_\epsilon} (1-\epsilon)^m \\
&= |\mathcal{H}_\epsilon|(1-\epsilon)^m \\
&\leq |\mathcal{H}|(1-\epsilon)^m && 1-\epsilon \leq e^{-\epsilon} \\
&\leq |\mathcal{H}|e^{-m\epsilon}
\end{aligned}$$

Set $|\mathcal{H}|e^{-m\epsilon} \leq \delta$, it derives $m \geq \frac{1}{\epsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta})$

More than m samples lead high chance to get consistent hypothesis

PAC Learning

Probably Approximately Correct (PAC) is a learning framework

Fundamental questions: what can be learned efficiently? What is hard to learn? How many examples are sufficient?

- $\mathcal{X}$: *input space*, all possible examples or instances
- $\mathcal{Y} = \{0, 1\}$: labels or *target values*
- Any subset of $\mathcal{X}$ could be a concept (*concept class* $\mathcal{C} \subseteq 2^{\mathcal{X}}$)
- *Target concept* $c : \mathcal{X} \to \mathcal{Y}$ belongs to $\mathcal{C}$
- Examples are independently and identically distributed (i.i.d) according to fixed and unknown distribution $\mathcal{D}$
- The Learner
  - receives samples $\mathcal{S} \subseteq \mathcal{X}$ drawn according to $\mathcal{D}$
  - Along with $\mathcal{S} = (x_1, x_2, ..., x_m)$ it also gets $(c(x_1), c(x_2), ..., c(x_m))$
  - considers *hypothesis set* a fixed set of possible concepts $\mathcal{H}$ that not necessary coincide with $\mathcal{C}$. The learner returns $h \in \mathcal{H}$
  - so that $\mathcal{R}(h)$ *generalization error* (also risk or true error) is small

Error: Generalization and Empirical

Given hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and an underlying distribution $\mathcal{D}$, *generalization error*, or risk of h is defined as

$$\mathcal{R}(h) = P_{x\sim\mathcal{D}}[h(x) \neq c(x)] = E_{x\sim\mathcal{D}}[1_{h(x)\neq c(x)}]$$

For $h \in \mathcal{H}$, target concept $c \in \mathcal{C}$, samples $\mathcal{S} = (x_1, x_2, ..., x_m)$ the *empirical error*, or empirical risk of h is defined as

$$\hat{\mathcal{R}}_S(h) = \frac{1}{m}\sum_{i=1}^{m} 1_{h(x_i)\neq c(x_i)}$$

What is empirical risk minimization?

Average and expected error!

How generalization and empirical error are related?

For fixed hypothesis $h \in \mathcal{H}$, the expectation of empirical error based on an i.i.d. samples $\mathcal{S}$ is equal to the generalization error

$$\begin{aligned}
E_{S\sim\mathcal{D}^m}[\hat{\mathcal{R}}_S(h)] &= E_{S\sim\mathcal{D}^m}\tfrac{1}{m}\textstyle\sum_{i=1}^m 1_{h(x_i)\neq c(x_i)} && \text{by definition} \\
&= \tfrac{1}{m}\textstyle\sum_{i=1}^m E_{S\sim\mathcal{D}^m}[1_{h(x_i)\neq c(x_i)}] && \text{by linearity of expectation} \\
&= \tfrac{1}{m}\textstyle\sum_{i=1}^m E_{S\sim\mathcal{D}^m}[1_{h(x)\neq c(x)}] && \text{by i.i.d.} \\
&= E_{S\sim\mathcal{D}^m}[1_{h(x)\neq c(x)}] && \text{same values} \\
&= E_{x\sim\mathcal{D}}[1_{h(x)\neq c(x)}] \\
&= \mathcal{R}(h)
\end{aligned}$$

What we want? high accuracy (error less than ϵ) and high confidence (confidence slip δ less than 1)

PAC learning

Concept class $\mathcal{C}$ is said to be PAC-learnable if $\exists$ and algorithm $\mathcal{A}$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target function $c \in \mathcal{C}$

$$P_{S\sim\mathcal{D}^m}[\mathcal{R}(h_S) \leq \epsilon] \geq 1-\delta$$

holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$

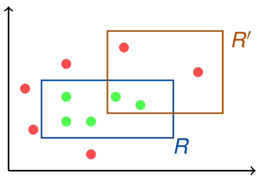
where, $\mathcal{S} = (x_1, x_2, ..., x_m)$ is training samples, and

- $poly(.,.,.,.)$ is a polynomial function
- n is such that the computation cost to represent $x \in \mathcal{X}$ is $O(n)$
- $size(c)$ is maximum computational requirement of $c \in \mathcal{C}$

We have high confidence ($\geq 1-\delta$) that error would be small ($\leq \epsilon$)

If $\mathcal{A}$ runs in $poly(1/\epsilon, 1/\delta, n, size(c))$ then $\mathcal{C}$ is efficiently PAC-learnable

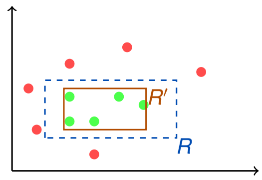
Example: learning axis aligned rectangles



- Instances are points in 2D space, and $R \in \mathcal{C}$
- Hypothesis R' could have some false positive and false negative
- What is the hypothesis space? All possible axis aligned rectangles
- How many parameters? 4, rectangle is (l,b) to (r,t)
- Let the learning algorithm $\mathcal{A}$ returns tightest axis aligned rectangle

Example: contd..

$\mathcal{A}$ returns tightest axis aligned rectangle. $R' = R_S$



- By definition R_S does not have false positive
- Fix some ϵ and let $P(R) \geq \epsilon$
- Define four region r_1, r_2, r_3 and r_4 such that $P(r_i) \geq \epsilon/4$
- If R_S meets all these regions, it will have one side in each of these regions. [Argument by geometry]
- Error area of R_S is the part of R that is not covered by R_S. Note error area would be included in the union of r_i

Example: contd..

$$\begin{aligned} P_{S\sim\mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] &\leq P_{S\sim\mathcal{D}^m}[\cup_{i=1}^4\{R_S \bigcap r_i = \phi\}] \\ &\leq \textstyle\sum_{i=1}^4 P_{S\sim\mathcal{D}^m}[\{R_S \bigcap r_i = \phi\}] && \text{by union bound} \\ &\leq 4(1-\epsilon/4)^m && \text{as } P[r_i] \geq \epsilon/4 \\ &\leq 4.e^{-m\epsilon/4} && \text{as } 1-x \leq e^{-x} \end{aligned}$$

To ensure $P_{S\sim\mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] \leq \delta$ we have

$$4.e^{-m\epsilon/4} \leq \delta$$

that gives

$$m \geq \frac{4}{\epsilon}\log\frac{4}{\delta}$$

For any $\epsilon > 0$ and $\delta > 0$, if the sample size m is greater then $\frac{4}{\epsilon}\log\frac{4}{\delta}$, then $P_{S\sim\mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] \leq \delta$ so this concept class is PAC-learnable.

Generalization Bound

Generalization Bound

With probability at least $1-\delta$ error $\mathcal{R}(R_S)$ is upper bounded by ϵ

We have seen

$$P_{S\sim\mathcal{D}^m}[\mathcal{R}(R_S) > \epsilon] \leq \delta$$

With

$$\delta = 4.e^{-m\epsilon/4}$$

we have

$$\epsilon = \frac{4}{m}\log\frac{4}{\delta}$$

So,

$$\mathcal{R}(R_S) \leq \frac{4}{m}\log\frac{4}{\delta}$$

Guarantees for finite hypothesis set

Theorem: Let $|\mathcal{H}|$ be a finite set of functions mapping $\mathcal{X} \to \mathcal{Y}$. Let for any target concept $c \in \mathcal{H}$ and i.i.d samples S, an algorithm $\mathcal{A}$ returns consistent hypothesis $h_S : \hat{\mathcal{R}}(R_S) = 0$. Then for any $\epsilon, \delta > 0$, the inequality $P_{S\sim\mathcal{D}^m}[\mathcal{R}(h_S) \leq \epsilon] \geq 1-\delta$ holds if

$$m \geq \frac{1}{\epsilon}\Big(\log|\mathcal{H}| + \log\frac{1}{\delta}\Big)$$

This result equivalently admits: for any $\epsilon, \delta > 0$, with probability at least $1-\delta$

$$\mathcal{R}(h_S) \leq \frac{1}{m}\Big(\log|\mathcal{H}| + \log\frac{1}{\delta}\Big)$$

Proof: Fix $\epsilon > 0$ and let $\mathcal{H}_\epsilon = \{h \in \mathcal{H} : R(h) > \epsilon\}$. Probability that a hypothesis $h \in \mathcal{H}_\epsilon$ is consistent on a training sample S drawn i.i.d is bounded as

$$P[\hat{\mathcal{R}}_S(h) = 0] \leq (1-\epsilon)^m$$

Guarantees for finite hypothesis set (Contd...)

$$P[\hat{\mathcal{R}}_S(h) = 0] \leq (1-\epsilon)^m$$

Thus, by union bound

$$\begin{aligned} P[\exists h \in \mathcal{H}_\epsilon : \hat{\mathcal{R}}_S(h) = 0] &= P[\hat{\mathcal{R}}_S(h_1) = 0 \vee \hat{\mathcal{R}}_S(h_2) = 0 \vee \ldots \\ &\qquad \vee \hat{\mathcal{R}}_S(h_{|\mathcal{H}_\epsilon|}) = 0] \\ &\leq \textstyle\sum_{h\in\mathcal{H}_\epsilon} P[\hat{\mathcal{R}}_S(h) = 0] && \text{union bound} \\ &\leq \textstyle\sum_{h\in\mathcal{H}_\epsilon} (1-\epsilon)^m \\ &= |\mathcal{H}_\epsilon|(1-\epsilon)^m \\ &\leq |\mathcal{H}|(1-\epsilon)^m && 1-\epsilon \leq e^{-\epsilon} \\ &\leq |\mathcal{H}|e^{-m\epsilon} \end{aligned}$$

Set $|\mathcal{H}|e^{-m\epsilon} \leq \delta$, it derives $m \geq \frac{1}{\epsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta})$

More than m samples lead high chance to get consistent hypothesis

Thank You!

Thank you very much for your attention!

Queries ?

(Reference[1])

[1] [1] *Bayesian Reasoning and Machine Learning*: ch:1/2/3, by *David Barber*, [1] Foundations of ML, ch:1/2, by *Mehryar Mohri*