Learning Theory

Narayana Santhanam

EE 645
Apr 10, 2024

This section

PAC Learning

VC dimension and Sauer's lemma

Learnability of
- finite classes
- bounded VC dimension

PAC learning

Example/Instance space $\mathcal{X}$, label set $\mathcal{Y}$
Hypothesis class $\mathcal{H}$ (set of functions from $\mathcal{X} \to \mathcal{Y}$)
Distribution D over $\mathcal{X}$
Training sample S generated by distribution D

Prediction rule $h : \mathcal{X} \to \mathcal{Y}$ that is somehow good

Loss of a prediction rule

Prediction

Loss (wrt correct labeling f):

$L_{D,f}(h) = \mathbf{P}_{X \sim D}[h(x) \neq f(x)]$ $= \mathbb{E}[\mathbb{1}(h(x) \neq f(x))]$

We cannot observe this in general

$(x_i, f(x_i))$

Empirical loss on a sample of size n,

$\hat{L}(h) = \frac{1}{n} \sum \mathbf{1}(h(x_i) \neq f(x_i))$

This we observe in a supervised setting

What can we infer about $L(h)$ from $\hat{L}(h)$?

$$\mathbb{1}(\text{condition}) = \begin{cases} 1 & \text{if True} \\ 0 & \text{if False} \end{cases}$$

UNIVERSITY of HAWAI'I MĀNOA

IID assumption

Generally expect every example of our training sample to be generated independently

In this case we can expect $\hat{L}(h)$ to concentrate around $L(h)$

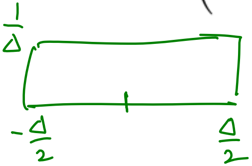
Empirical average $\approx$ real expectation

But by how much? What is the deviation?

Hoeffding's Inequality

Let $X_1, \dots, X_n$ be *i.i.d.* variables, the variables bounded in range $X_i \in [a, b]$, and let $\mu = \mathbb{E}X_i$. Then for any $\epsilon > 0$,

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_i X_i - \mu\right| > \epsilon \right) \le 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2} \right)$$



$$2\int_0^{\Delta/2} x^2 \cdot \frac{1}{\Delta}\, dx = 2\cdot\left(\frac{x^3}{3}\Big|_0^{\Delta/2} \right)\cdot\frac{1}{\Delta}$$

$$= \frac{2}{3}\cdot\frac{\Delta^3}{8}\frac{1}{\Delta} = \frac{\Delta^2}{12}.$$

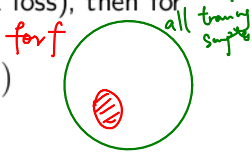
Hoeffding's Inequality

Let $X_1, \ldots, X_n$ be *i.i.d.* variables, the variables bounded in range $X_i \in [a, b]$, and let $\mu = \mathbb{E}X_i$. Then for any $\epsilon > 0$,

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_i X_i - \mu\right| > \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

If we are working with binary classification (with 0-1 loss), then for each h,

$$\mathbf{P}\left(\left|\hat{L}(h) - L(h)\right| > \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right)$$



$$\frac{1}{n}\sum 1(h(x_i) \neq f(x_i))$$

$$\mathbb{E}\left[\frac{1}{n}\sum 1(h(x_i) \neq f(x_i))\right] = \frac{1}{n}\sum \mathbb{E}\left[1(h(x_i) \neq f(x_i))\right]$$

$$= \frac{1}{n}\sum P(h(x) \neq f(x)) = \frac{n L(h)}{n}$$

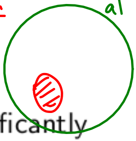
Hoeffding's Inequality

Let $X_1, \ldots, X_n$ be *i.i.d.* variables, the variables bounded in range $X_i \in [a, b]$, and let $\mu = \mathbb{E}X_i$. Then for any $\epsilon > 0$,

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_i X_i - \mu\right| > \epsilon \right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2} \right)$$

If we are working with binary classification (with 0-1 loss), then for each h,

$$\mathbf{P}\left(\left|\hat{L}(h) - L(h)\right| > \epsilon \right) \leq 2\exp\left(-2n\epsilon^2 \right)$$ for f



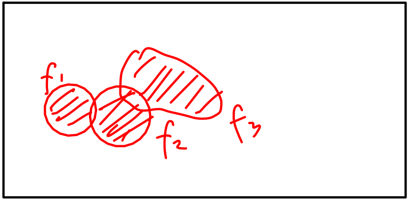
In a bad set of training samples $B(\cancel{h}f)$, $\hat{L}(h)$ deviates significantly from $L(h)$, but the set of misleading training samples have small probability if n is large enough

Union bound

If we have finite number of hypothesis, we can argue that collectively, all the bad sets of all $h \in \mathcal{H}$ don't matter: Union bound

$$\mathbf{P}\left(\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)| > \epsilon\right) \leq 2|\mathcal{H}| \exp\left(-2n\epsilon^2\right)$$

Here $|\cdot|$ denotes the size of a set



Union bound

If we have finite number of hypothesis, we can argue that collectively, all the bad sets of all $h \in \mathcal{H}$ don't matter: Union bound

$$\mathbf{P}\left(\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)| > \epsilon\right) \leq 2|\mathcal{H}| \exp\left(-2n\epsilon^2\right)$$

Here $|\cdot|$ denotes the size of a set

This is not artificial—in fact, given we only use finite precision and a finite number of network weights, most deep networks also form finite classes in practice.

Union bound

$2|H| e^{-2n\epsilon^2} = \eta .$

If we have finite number of hypothesis, we can argue that collectively, all the bad sets of all $h \in \mathcal{H}$ don't matter: Union bound

$n = \frac{1}{2\epsilon^2} \ln\left(\frac{2|H|}{\eta}\right)$

$$\mathbf{P}\left(\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| > \epsilon\right) \leq \underbrace{2|\mathcal{H}| \exp\left(-2n\epsilon^2\right)}_{\eta}$$

$f \quad f \quad f$

Here $|\cdot|$ denotes the size of a set

This is not artificial—in fact, given we only use finite precision and a finite number of network weights, most deep networks also form finite classes in practice.

Catch is, we don't have to wait till we are guaranteed convergence like above: usually our estimators work good well before we need to sample to reduce the right side to within a given confidence

Vapnik Chervonenkis dimension

Again, binary classification, 0-1 loss.

Vapnik Chervonenkis dimension

Again, binary classification, 0-1 loss.

A set of points S is shattered by a hypothesis class $\mathcal{H}$ if all $2^{|S|}$ labelings on S are produced by hypothesis in $\mathcal{H}$, namely $|\mathcal{H}(S)| = 2^{|S|}$

$\mathcal{H}(S) =$ $\quad$ $|\mathcal{H}(S)| = 7.$

Examples

$$\left\{ \begin{array}{ccc} - & - & - \\ - & - & + \\ - & + & - \\ - & + & + \\ + & - & - \\ + & + & - \\ + & + & + \end{array} \right\}$$

Vapnik Chervonenkis dimension

The VC dimension of $\mathcal{H}$ is the size of the largest set S of points it shatters.

If the VC dimension of $\mathcal{H}$ is d, it doesn't mean every set of d points is shattered by $\mathcal{H}$

only that some set of d points is

But it does mean *no* set of $d + 1$ points can be shattered by $\mathcal{H}$

Larger VC dimension, more power

Sauer's lemma

If $\mathcal{H}$ has VC dimension d, how many labelings on a sample S of size n can it generate?

Trivially, if $n > d$, then number of labelings is $< 2^n$
But one would imagine 2^n is a gross overestimate
Proposed by Erdös, solved (1972) and re-proved several times in other contexts

including by Vapnik and Chervonenkis

Sauer's lemma

If $\mathcal{H}$ has VC dimension d and S is a sample of size n,

$$|\mathcal{H}(S)| \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Proof (simple, and by induction)

We prove a stronger result that

$$|\mathcal{H}(S)| \leq |B \subset S : \mathcal{H} \text{ shatters } B|.$$

Induction argument

To prove:

$$|\mathcal{H}(S)| \leq |B \subset S : \mathcal{H} \text{ shatters } B|.$$

Proof: When $n = 1$, either both sides are 1 or both are 2.

Induction hypothesis: Assume true for all sets S with size $< n$, will prove for all S of size n

Hence qed

Proof

To prove:

$$|\mathcal{H}(S)| \leq |B \subset S : \mathcal{H} \text{ shatters } B|.$$

Let S' be the sample with the last example removed (so size $n-1$) and let

$$Y_0 = \{\mathbf{y}(S') : \mathbf{y}(S') \in \mathcal{H}(S')\}$$

and

$$Y_1 = \{\mathbf{y}(S') : (\mathbf{y}(S'), 0) \text{ and } (\mathbf{y}(S'), 1) \in \mathcal{H}(S)\}$$

Clearly

$$|\mathcal{H}(S)| = |Y_0| + |Y_1|$$

Proof

To prove:

$$|\mathcal{H}(S)| \leq |B \subset S : \mathcal{H} \text{ shatters } B|.$$

Recall S' be the sample with the last example removed (so size $n-1$) and that

$$Y_0 = \{\mathbf{y}(S') : \mathbf{y}(S') \in \mathcal{H}(S')\}$$

From induction hypothesis

$$|Y_0| \leq |\{B \in S : \mathcal{H} \text{ shatters } B \text{ and } y_n \notin B\}|$$

University of Hawaiʻi Mānoa

Proof

To prove:

$$|\mathcal{H}(S)| \leq |B \subset S : \mathcal{H} \text{ shatters } B|.$$

Recall S' be the sample with the last example removed (so size $n-1$) and

$$Y_1 = \{\mathbf{y}(S') : (\mathbf{y}(S'), 0) \text{ and } (\mathbf{y}(S'), 1) \in \mathcal{H}(S)\}$$

Let $\mathcal{H}'$ be a subset of $\mathcal{H}$. We put a pair h, h' into $\mathcal{H}'$ if h, h' agree on S' but disagree on the last example.

Now we claim $|Y_1| = |\mathcal{H}'(S')|$ and therefore that

$$|\mathcal{H}'(S')| \leq |\{B \in S : \mathcal{H} \text{ shatters } B \text{ and } y_n \in B\}|$$

Proof

Therefore

$$|\mathcal{H}(S)| = |Y_0| + |Y_1|,$$

but

$$|Y_0| \leq |\{B \in S : \mathcal{H} \text{ shatters } B \text{ and } y_n \notin B\}|$$

and

$$|Y_1| \leq |\{B \in S : \mathcal{H} \text{ shatters } B \text{ and } y_n \in B\}|$$

and so, the result follows!

Next steps

How Sauer's lemma gives us learnability results for infinite classes

Still, not strong enough to explain neural networks

PAC Bayes approaches