Learning Theory

Narayana Santhanam

EE 645
Apr 10, 2024

University of Hawai'i Mānoa

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

Loss (wrt correct labeling f):

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

We cannot observe this in general

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 $\hat{L}(h)$?

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 $X_1, \ldots, X_n$ 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 $\epsilon > 0$,

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

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,

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

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,

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

In a bad set of training samples $B(h)$, $\hat{L}(h)$ 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

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

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

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

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

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

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

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.

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 $|\mathcal{H}(S)| = 2^{|S|}$

Examples

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,

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

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

University of Hawaiʻi Mānoa

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

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

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

log VCDim instead of log #hypothesis

Caveat: not strong enough to explain neural networks

More refinements to come

VC dimension and PAC learnability

Let S be a sample of size n, generated iid D
For each sample, what is the worst deviation between sample error $\hat{L}(h)$ made by some $h \in \mathcal{H}$ and its true generalization error $L(h)$?
Namely

$$\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)|$$

Today' class: bound on this

VC dimension and PAC learnability

Let S be a sample of size n, generated iid D

For each sample, what is the worst deviation between sample error $\hat{L}(h)$ made by some $h \in \mathcal{H}$ and its true generalization error $L(h)$?

Namely

$$\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)|$$

Today' class: bound on this

Full disclosure, we will bound $\mathbb{E}_S \sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)|$, from which we bound $\mathbf{P}(\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)| > \epsilon)$ via a Markov inequality

Ghost sample

Let S' be a "ghost sample" (an imaginary sample also generated iid D)
Let $\hat{L}_{S'}(h)$ be the empirical error of hypothesis h on S'

Ghost sample

Let S' be a "ghost sample" (an imaginary sample also generated iid D)
Let $\hat{L}_{S'}(h)$ be the empirical error of hypothesis h on S'
Still have $L(h) = \mathbb{E}_{S' \sim D} \hat{L}_{S'}(h)$ (expectation over all ghost samples)

Using above, we write for each $h \in \mathcal{H}$,

$$|\hat{L}(h) - L(h)| = |\mathbb{E}_{S'}(\hat{L}(h) - \hat{L}_{S'}(h))| \leq \mathbb{E}_{S'}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

Ghost sample

Let S' be a "ghost sample" (an imaginary sample also generated iid D)

We observed for each h,

$$|\hat{L}(h) - L(h)| \leq \mathbb{E}_{S'}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

Ghost sample

Let S' be a "ghost sample" (an imaginary sample also generated iid D)
We observed for each h,

$$|\hat{L}(h) - L(h)| \leq \mathbb{E}_{S'}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

So for each training sample,

$$\sup_{h\in\mathcal{H}}|\hat{L}(h) - L(h)| \leq \sup_{h\in\mathcal{H}}\mathbb{E}_{S'}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

Ghost sample

Let S' be a "ghost sample" (an imaginary sample also generated iid D)

We observed for each h,

$$|\hat{L}(h) - L(h)| \leq \mathbb{E}_{S'}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

So for each training sample,

$$\sup_{h\in\mathcal{H}}|\hat{L}(h) - L(h)| \leq \sup_{h\in\mathcal{H}}\mathbb{E}_{S'}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

But

$$\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}|\hat{L}(h) - \hat{L}_{S'}(h)| \leq \mathbb{E}_{S'}\sup_{h\in\mathcal{H}}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

and hence

$$|\hat{L}(h) - L(h)| \leq \mathbb{E}_{S'}\sup_{h\in\mathcal{H}}|\hat{L}(h) - \hat{L}_{S'}(h)|$$

Symmetrization with ghost sample

Training sample $S = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and ghost sample $S' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_n)$. Both are drawn *i.i.d.* D.

Symmetrization with ghost sample

Training sample $S = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and ghost sample $S' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_n)$. Both are drawn *i.i.d.* D.

Swapping out the first element of ghost with that of train,

Symmetrization with ghost sample

Training sample $S = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and ghost sample $S' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_n)$. Both are drawn *i.i.d.* D.

Swapping out the first element of ghost with that of train, (so we now consider the function

$$\sup_h \left| \ell(h, \mathbf{z}'_1) - \ell(h, \mathbf{z}_1) + \sum_{i=2}^{n} (\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i) \right|$$

in place of $\sup_h |\hat{L}(h) - \hat{L}_{S'}(h)|$), expectations don't change:

$$\mathbb{E}_S \mathbb{E}_{S'} \sup_h \left| \ell(h, \mathbf{z}'_1) - \ell(h, \mathbf{z}_1) + \sum_{i=2}^{n} (\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i) \right|$$

$$= \mathbb{E}_S \mathbb{E}_{S'} \sup_h \left| \ell(h, \mathbf{z}_1) - \ell(h, \mathbf{z}'_1) + \sum_{i=2}^{n} (\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i) \right|$$

$$= \mathbb{E}_S \mathbb{E}_{S'} \sup_h |\hat{L}(h) - \hat{L}_{S'}(h)|$$

Symmetrization with ghost sample

In fact, can swap as many examples between train/ghos as we want with no change in expectations

We could even pick $\sigma_1, \ldots, \sigma_n$ to be independent random variables taking values in $\{-1, 1\}^n$, with equal probabilities and

$$
\begin{aligned}
E_\sigma \mathbb{E}_S \mathbb{E}_{S'} \sup_h &\Big| \sum_{i=1}^n \sigma_i (\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i)) \Big| \\
&= \mathbb{E}_S \mathbb{E}_{S'} \sup_h \Big| \sum_{i=1}^n (\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i)) \Big| \\
&= \mathbb{E}_S \mathbb{E}_{S'} \sup_h |\hat{L}(h) - \hat{L}_{S'}(h)|
\end{aligned}
$$

University of Hawai'i Mānoa

Reviewing so far

$$\mathbb{E}_S \mathbb{E}_{S'} \sup_h |\hat{L}(h) - \hat{L}_{S'}(h)|$$

$$= E_\sigma \mathbb{E}_S \mathbb{E}_{S'} \sup_h |\sum_{i=1}^n \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))|$$

$$= \mathbb{E}_S \mathbb{E}_{S'} E_\sigma \sup_h |\sum_{i=1}^n \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))|$$

We now fix a train and ghost sample combination and examine

$$\sup_{h \in \mathcal{H}} |\sum_{i=1}^n \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))|$$

Now given a fixed train and ghost sample, and a fixed h, let $W_i = \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))$.

$\mathbb{E} W_i = 0$, $-1 \leq W_i \leq 1$, and W_i independent (not identical!)

Better Heoffding

In fact, our earlier version of Hoeffding's inequality wasn't the full version. We don't need the random variables to be iid, only independent suffices.

Let $W_1, \ldots, W_n$ be independent variables, the variables bounded in range $X_i \in [a, b]$, with $\mathbb{E} W_i = \mu$. Then for any $\epsilon > 0$,

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

Completing the proof

We were examining for a fixed train/ghost sample:

$$\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i)) \right|$$

Now given a fixed train and ghost sample, and a fixed h, let $W_i = \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))$.
$\mathbb{E}W_i = 0$, $-1 \leq W_i \leq 1$, and W_i independent (not identical!)

Therefore for each $h \in \mathcal{H}$,

$$\mathbf{P}\left(\left| \sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i)) \right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

But doesn't $\mathcal{H}$ have infinitely many hypotheses?

Recall: Sauer's lemma

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

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

Proof (simple, and by induction)

We prove a stronger result that

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

Now for Sauer's lemma

Therefore for each $h \in \mathcal{H}$,

$$\mathbf{P}(|\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))| \geq \epsilon) \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Now since we have fixed train/ghost, there are only $L(2n, d)$ labelings from $\mathcal{H}$ if it has VC dimension d. So effectively only $L(2n, d)$ hypotheses! Therefore

$$\mathbf{P}(\sup_{h \in \mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))| \geq \epsilon) \leq 2L(2n, d) \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Remember: this is probability over choice of σ_i (the training and ghost samples are being held fixed)

Now for Sauer's lemma

Therefore for each $h \in \mathcal{H}$,

$$\mathbf{P}(|\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}_i'))| \geq \epsilon) \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Now since we have fixed train/ghost, there are only $L(2n, d)$ labelings from $\mathcal{H}$ if it has VC dimension d. So effectively only $L(2n, d)$ hypotheses! Therefore

$$\mathbf{P}(\sup_{h\in\mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}_i'))| \geq \epsilon) \leq 2L(2n, d) \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Remember: this is probability over choice of σ_i (the training and ghost samples are being held fixed) from which we get a straightforward bound on $\mathbb{E}_\sigma \sup_{h\in\mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}_i'))|$,

Now for Sauer's lemma

Therefore for each $h \in \mathcal{H}$,

$$\mathbf{P}(|\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))| \geq \epsilon) \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Now since we have fixed train/ghost, there are only $L(2n, d)$ labelings from $\mathcal{H}$ if it has VC dimension d. So effectively only $L(2n, d)$ hypotheses! Therefore

$$\mathbf{P}(\sup_{h \in \mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))| \geq \epsilon) \leq 2L(2n, d) \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Remember: this is probability over choice of σ_i (the training and ghost samples are being held fixed)from which we get a straightforward bound on $\mathbb{E}_\sigma \sup_{h \in \mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}'_i))|$, which upper bounds $\mathbb{E}_S \sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)|$,

Now for Sauer's lemma

Therefore for each $h \in \mathcal{H}$,

$$\mathbf{P}(|\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}_i'))| \geq \epsilon) \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Now since we have fixed train/ghost, there are only $L(2n, d)$ labelings from $\mathcal{H}$ if it has VC dimension d. So effectively only $L(2n, d)$ hypotheses! Therefore

$$\mathbf{P}(\sup_{h \in \mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}_i'))| \geq \epsilon) \leq 2L(2n, d) \exp\left(-\frac{n\epsilon^2}{2}\right)$$

Remember: this is probability over choice of σ_i (the training and ghost samples are being held fixed)from which we get a straightforward bound on

$\mathbb{E}_\sigma \sup_{h \in \mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, \mathbf{z}_i) - \ell(h, \mathbf{z}_i'))|$, which upper bounds $\mathbb{E}_S \sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)|$, which upper bounds

$\mathbf{P}(\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)| > \epsilon)$ via a Markov inequality