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

Loss (wrt correct labeling f):

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

We cannot observe this in general

Empirical loss on a sample of size n,

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

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 $\epsilon > 0$,

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

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 $\epsilon > 0$,

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

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

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

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 $\epsilon > 0$,

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

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

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

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

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

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

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

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

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

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.

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

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

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

and

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

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

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

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

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

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 $\mathrm{P}(\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)| > \epsilon)$ 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

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

Symmetrization with ghost sample

Training sample $S = (z_1, \ldots, z_n)$ and ghost sample $S' = (z'_1, \ldots, z'_n)$. Both are drawn *i.i.d.* D.

Symmetrization with ghost sample

Training sample $S = (z_1, \ldots, z_n)$ and ghost sample $S' = (z'_1, \ldots, z'_n)$. Both are drawn *i.i.d.* D.

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

Symmetrization with ghost sample

Training sample $S = (z_1, \ldots, z_n)$ and ghost sample $S' = (z'_1, \ldots, z'_n)$. Both are drawn *i.i.d.* D.

Swap out the first element of ghost with that of train, (so we now consider the function in place of $\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)|$

$$\sup_h |\ell(h, z'_1) - \ell(h, z_1) + \sum_{i=2}^{n} (\ell(h, z_i) - \ell(h, z'_i)|$$

Above function different, its expectation (over S, S') is not, that is:

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

$$= \mathbb{E}_S \mathbb{E}_{S'} \sup_h |\ell(h, z_1) - \ell(h, z'_1) + \sum_{i=2}^{n} (\ell(h, z_i) - \ell(h, z'_i)|$$

$$= \mathbb{E}_S \mathbb{E}_{S'} \sup_h |\ell(h, z'_1) - \ell(h, z_1) + \sum_{i=2}^{n} (\ell(h, z_i) - \ell(h, z'_i)|$$

Symmetrization with ghost sample

In fact, can swap as many examples between train/ghost as we want to get new functions with the same expectation

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

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

$$= \mathbb{E}_S \mathbb{E}_{S'} \sup_h |\sum_{i=1}^{n} (\ell(h, z_i) - \ell(h, z_i'))|$$

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

Reviewing so far

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

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

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

We now fix a train and ghost sample combination and examine

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

The quantity above is very close to the Rademacher complexity, which will give us another way to deal with generalization error.

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 $\epsilon > 0$,

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

Completing the proof for classes with small VC dimension

We were examining for a fixed train/ghost sample:

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

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

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

$$P\left(\left|\sum_{i=1}^{n} \sigma_i(\ell(h, z_i) - \ell(h, z_i'))\right| \ge \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

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 $h \in \mathcal{H}$,

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

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

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

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 $h \in \mathcal{H}$,

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

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

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

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

$\mathbb{E}_S \mathbb{E}_{S'} \mathbb{E}_\sigma \sup_{h\in\mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, z_i) - \ell(h, z_i'))|$.

Now for Sauer's lemma

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

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

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

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

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

Now for Sauer's lemma

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

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

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

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

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 $\mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_\sigma \sup_{h\in\mathcal{H}} |\sum_{i=1}^{n} \sigma_i(\ell(h, z_i) - \ell(h, z_i'))|$, which upper bounds $\mathbb{E}_S \sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)|$, which upper bounds $P(\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| > \epsilon)$ via a Markov inequality

Putting everything together

For a class $\mathcal{H}$ with VC dimension d,

$$\mathrm{P}\left(\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| \geq \frac{\sqrt{\log L(2n,d)}}{\eta\sqrt{2n}}\right) \leq \eta.$$

Putting everything together

For a class $\mathcal{H}$ with VC dimension d,

$$P\left(\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| \geq \frac{\sqrt{\log L(2n, d)}}{\eta\sqrt{2n}}\right) \leq \eta.$$

For a given accuracy ϵ and confidence η, we need a training sample of size

$$n \geq 4\frac{2d}{(\eta\epsilon)^2}\log\left(\frac{2d}{(\eta\epsilon)^2}\right) + \frac{4d\log(2e/d)}{(\eta\epsilon)^2}.$$

But even this isn't enough

We need to do beter. For kernel methods, we lift up the points to very high-d space. Even linear classifiers in this high-d space have VC dimension equal to high-d $+1$, which is too large.

Similar line of argument works, but we focus on Rademacher complexity

But even this isn't enough

We need to do beter. For kernel methods, we lift up the points to very high-d space. Even linear classifiers in this high-d space have VC dimension equal to high-d $+1$, which is too large.

Similar line of argument works, but we focus on Rademacher complexity

Let $\mathcal{F}$ be a set of functions on $S = (\mathrm{z}_1, \ldots, \mathrm{z}_n)$. Then

$$\mathcal{R}(\mathcal{F}(S)) = \frac{1}{m}\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{F}}\sum_i \sigma_i f(\mathrm{z}_i)\Big].$$

Note that we don't think of the worst case training sample or an average. The Rademacher complexity is defined per sample.

But even this isn't enough

We need to do beter. For kernel methods, we lift up the points to very high-d space. Even linear classifiers in this high-d space have VC dimension equal to high-d $+1$, which is too large.

Similar line of argument works, but we focus on Rademacher complexity

Let $\mathcal{F}$ be a set of functions on $S = (\mathrm{z}_1, \ldots, \mathrm{z}_n)$. Then

$$\mathcal{R}(\mathcal{F}(S)) = \frac{1}{m}\mathbb{E}_\sigma\big[\sup_{f\in\mathcal{F}}\sum_i \sigma_i f(\mathrm{z}_i)\big].$$

Note that we don't think of the worst case training sample or an average. The Rademacher complexity is defined per sample.

If we have a hypothesis class $\mathcal{H}$, and let $f(\mathrm{z}_i) = \ell(h, \mathrm{z}_i)$, we pretty much can get a bound in terms of Rademacher complexity in place of VC dimension. But Rademacher complexity doesn't blow up automatically like with VC dimension of classifiers in high-d space

Rademacher complexity

For a training sample $S = (z_1, \ldots, z_n)$, using the ghost sample idea

$$
\begin{aligned}
&\sup_h L(h) - \hat{L}(h) \\
&\quad = \sup_h \mathbb{E}_{S'} \big[\hat{L}_{S'}(h) - \hat{L}(h) \big] \\
&\quad \le \mathbb{E}_{S'} \sup_h \big[\hat{L}_{S'}(h) - \hat{L}(h) \big] \\
&\quad = \mathbb{E}_{S'} \sup_h \big[\frac{1}{n} \sum_i (\ell(h, z') - \ell(h, z)) \big]
\end{aligned}
$$

We can now swap between the training/ghost samples just like before to ge new functions with the same expectation (under S and S')

Rademacher central argument

By arguments very similar to before, where $\sigma_i \sim$iid B(1/2)

$$
\begin{aligned}
&\mathbb{E}_S \sup_h L(h) - \hat{L}(h) \\
&\qquad \leq \mathbb{E}_S \mathbb{E}_{S'} \sup_h \big[\hat{L}_{S'}(h) - \hat{L}(h)\big] \\
&\qquad = \mathbb{E}_S \mathbb{E}'_S \mathbb{E}_\sigma \sup_h \Big[\frac{1}{n} \sum_i \sigma_i (\ell(h, z') - \ell(h, z))\Big] \\
&\qquad \leq 2 \mathbb{E}_S \mathbb{E}_\sigma \sup_h \frac{1}{n} \Big[\sum_i \sigma_i \ell(h, z)\Big] \\
&\qquad = 2 \mathbb{E}_S \mathcal{R}(\ell(\mathcal{H}(S)))
\end{aligned}
$$

Rademacher central argument

By arguments very similar to before, where $\sigma_i \sim$iid B(1/2)

$$
\begin{aligned}
&\mathbb{E}_S \sup_h L(h) - \hat{L}(h) \\
&\quad \leq \mathbb{E}_S \mathbb{E}_{S'} \sup_h \big[\hat{L}_{S'}(h) - \hat{L}(h)\big] \\
&\quad = \mathbb{E}_S \mathbb{E}'_S \mathbb{E}_\sigma \sup_h \big[\frac{1}{n} \sum_i \sigma_i(\ell(h, z') - \ell(h, z))\big] \\
&\quad \leq 2\mathbb{E}_S \mathbb{E}_\sigma \sup_h \frac{1}{n} \big[\sum_i \sigma_i \ell(h, z)\big] \\
&\quad = 2\mathbb{E}_S \mathcal{R}(\ell(\mathcal{H}(S)))
\end{aligned}
$$

where $\ell(\mathcal{H}(S))$ is the set of loss sequences obtained on S from all the labelings in $\mathcal{H}$.

Rademacher central argument

By arguments very similar to before, where $\sigma_i \sim$iid B(1/2)

$$
\begin{aligned}
&\mathbb{E}_S \sup_h L(h) - \hat{L}(h) \\
&\qquad \le \mathbb{E}_S \mathbb{E}_{S'} \sup_h \left[\hat{L}_{S'}(h) - \hat{L}(h)\right] \\
&\qquad = \mathbb{E}_S \mathbb{E}'_S \mathbb{E}_\sigma \sup_h \left[\frac{1}{n}\sum_i \sigma_i(\ell(h, z') - \ell(h, z))\right] \\
&\qquad \le 2\mathbb{E}_S \mathbb{E}_\sigma \sup_h \frac{1}{n}\left[\sum_i \sigma_i \ell(h, z)\right] \\
&\qquad = 2\mathbb{E}_S \mathcal{R}(\ell(\mathcal{H}(S)))
\end{aligned}
$$

where $\ell(\mathcal{H}(S))$ is the set of loss sequences obtained on S from all the labelings in $\mathcal{H}$.

In the VC bound, we just used Sauer's lemma to bound second line above

Generalization in terms of Rademacher complexity

Given a sample S, we want to bound

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

Generalization in terms of Rademacher complexity

Given a sample S, we want to bound

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

From Rademacher central argument,

$$\begin{aligned}
&\mathbb{E}_S \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h) \\
&\quad \leq \mathbb{E}_S \mathbb{E}_{S'} \sup_{h} \left[\hat{L}_{S'}(h) - \hat{L}(h) \right] \\
&\quad \leq 2 \mathbb{E}_S \mathcal{R}(\ell(\mathcal{H}(S)))
\end{aligned}$$

Generalization in terms of Rademacher complexity

Given a sample S, we want to bound

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

From Rademacher central argument,

$$\begin{aligned} &\mathbb{E}_S \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h) \\ &\leq \mathbb{E}_S \mathbb{E}_{S'} \sup_h \left[\hat{L}_{S'}(h) - \hat{L}(h) \right] \\ &\leq 2\mathbb{E}_S \mathcal{R}(\ell(\mathcal{H}(S))) \end{aligned}$$

So we now compare the expected sample representativeness with the expected Rademacher complexity. But how to compare directly, ie $\sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$ with $\mathcal{R}(\ell(\mathcal{H}(S)))$?

McDiarmid's inequality

When $X_1, \ldots, X_n$ are independent random variables, McDiarmid's inequality bounds how far $f(X_1, \ldots, X_n)$ can be from $\mathbb{E}f(X_1, \ldots, X_n)$ for certain functions f

McDiarmid's inequality

When $X_1, \ldots, X_n$ are independent random variables, McDiarmid's inequality bounds how far $f(X_1, \ldots, X_n)$ can be from $\mathbb{E}f(X_1, \ldots, X_n)$ for certain functions f

Specifically, f must satisfy for all x_i and x_i'

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c$$

McDiarmid's inequality

When $X_1, \ldots, X_n$ are independent random variables, McDiarmid's inequality bounds how far $f(X_1, \ldots, X_n)$ can be from $\mathbb{E} f(X_1, \ldots, X_n)$ for certain functions f

Specifically, f must satisfy for all x_i and x_i'

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c$$

Then

$$\mathrm{P}(|f(X_1, \ldots, X_n) - \mathbb{E} f(X_1, \ldots, X_n)| > \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{nc^2}\right)$$

Representativeness of a sample

$S = z_1 z_2 \cdots \quad z_n$

$z_1' z_2 \cdots \quad z_n$

Representativeness of a sample S

If L is hinge loss
$$f(S) = \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$$

$w \cdot z$
$w' \cdot z'$

If we change any one example in S, the above changes by $\frac{2}{n}$ $\left(\frac{2BR}{n}\right)$

Applying McDiarmid's inequality,

$$P(|f(S) - \mathbb{E}f(S)| > \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{n(2/n)^2}\right) = 2\exp\left(-\underbrace{\frac{n\epsilon^2}{2}}_{\delta}\right)$$

or equivalently

$$P\left(|f(S) - \mathbb{E}f(S)| > \sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}\right) \leq \delta$$

Rademacher-McDiarmid bound on representativeness of a sa

Representativeness of a sample S

$$f(S) = \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$$

We just saw with probability $\geq 1 - \delta$,

$$f(S) \leq \mathbb{E} f(S) + \sqrt{\frac{2}{n} \ln\left(\frac{2}{\delta}\right)}$$

But Rademacher's bound: $\mathbb{E} f(S) \leq 2\mathbb{E}\mathcal{R}(\ell(\mathcal{H}(S)))$, so with probability $\geq 1 - \delta$,

$$f(S) \leq 2\mathbb{E}\mathcal{R}(\ell(\mathcal{H}(S))) + \sqrt{\frac{2}{n} \ln\left(\frac{2}{\delta}\right)}$$

Tightening the Rademacher-McDiarmid bound

In fact, the Rademacher complexity of S, $\mathcal{R}(\ell(\mathcal{H}(S)))$ only changes by $2/n$ when we change any sample.

Tightening the Rademacher-McDiarmid bound

In fact, the Rademacher complexity of S, $\mathcal{R}(\ell(\mathcal{H}(S)))$ only changes by $2/n$ when we change any sample.

So $\mathcal{R}(\ell(\mathcal{H}(S)))$ concentrates around $\mathbb{E}\mathcal{R}(\ell(\mathcal{H}(S)))$. Allows us to rewrite: w.p. $\geq 1-\delta$

$$f(S) \leq 2\mathbb{E}\mathcal{R}(\ell(\mathcal{H}(S))) + \sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}$$

as w.p. $\geq 1-\delta$

$$f(S) \leq 2\mathcal{R}(\ell(\mathcal{H}(S))) + 3\sqrt{\frac{2}{n}\ln\left(\frac{4}{\delta}\right)}$$

Rademacher complexity with kernel machines

$z_1 \quad z_n$

$x_1 \ldots x_n$

In kernel machines, SVM in particular, we do linear classification in a very high-d space.

$S = (\mathrm{z}_1, \ldots, \mathrm{z}_n)$ in a high-d space and our hypothesis class $\mathcal{H}$ is linear classifiers.

(Soft) predictions on S for some w are $(\mathrm{w} \cdot \mathrm{z}_1, \ldots, \mathrm{w} \cdot \mathrm{z}_n)$.

Want to explore

$$\mathcal{R}(\mathcal{H}(S)) = \frac{1}{m}\mathbb{E}_\sigma \left[\sup_{\mathrm{w}} \sum_i \sigma_i (\mathrm{w} \cdot \mathrm{z}_i) \right]$$

But in SVMs, while w is in a very high-d space, its length in that space is optimized (in fact, the SVM picks the smallest possible length satisfying constraints).

Rademacher complexity of linear classes

$S = (z_1 \ldots z_n)$

$L(h) - \hat{L}(h)$

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

So bound $||\mathrm{w}||^2$— we will bound the length by B for analysis here

$$\mathcal{R}(\mathcal{H}(S)) = \frac{1}{n}\mathbb{E}_\sigma\Big[\sup_{\substack{\mathrm{w} \\ ||\mathrm{w}|| \le B}} \sum_i \sigma_i(\mathrm{w}\cdot \mathrm{z}_i)\Big] \le \frac{B \max_i ||\mathrm{z}_i||}{\sqrt{n}}$$

Work out on the board

SVM analysis

In support vector machines, we consider the hinge loss

$$\ell(\mathrm{w}, \mathrm{z}) = \max(0, 1 - y\mathrm{w}^T\mathrm{z})$$

rather than $\mathrm{w}^T\mathrm{z}$ itself. Easy to show that

$$\mathcal{R}(\ell(\mathcal{H}(S))) \leq \mathcal{R}(\mathcal{H}(S))$$

(often called a contraction lemma, and happens because the hinge loss $\ell(\mathrm{w}, \mathrm{z})$ is a 1-Lipschitz function of $\mathrm{w}^T\mathrm{z}$)

SVM bound

Suppose all training examples satisfy $||z|| \leq R$

SVM bound

Suppose all training examples satisfy $||\mathrm{z}|| \leq R$
Suppose the hypothesis $||\mathrm{w}|| \leq B$

SVM bound

Suppose all training examples satisfy $||z|| \leq R$
Suppose the hypothesis $||w|| \leq B$

Only new thing is that if we consider the hinge loss, change of any one example could change the sample hinge loss by $R||w||/n \leq RB/n$ instead of $2/n$.

SVM bound

Suppose all training examples satisfy $||\mathrm{z}|| \leq R$
Suppose the hypothesis $||\mathrm{w}|| \leq B$

Only new thing is that if we consider the hinge loss, change of any one example could change the sample hinge loss by $R||\mathrm{w}||/n \leq RB/n$ instead of $2/n$. With that one change, wp $\geq 1 - \delta$

$$\leq 2\frac{BR}{\sqrt{n}}$$

$$L(h) \leq \hat{L}(h) + \underbrace{2\mathcal{R}(\ell(\mathcal{H}(S)))} + 3RB\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}$$

where L is the generalization hinge loss and $\hat{L}$ is the sample hinge loss.

$$w = \sum \alpha_i x_i$$

$$\|w\|^2_{RKHS} = \sum_{i,j} \alpha_i \alpha_j k(x_i x_j).$$

SVM bound

$\sup_{h} L(h) - \hat{L}(h)$

Suppose all training examples satisfy $||\mathbf{z}|| \leq R$
Suppose the hypothesis $||\mathbf{w}|| \leq B$

Only new thing is that if we consider the hinge loss, change of any one example could change the sample hinge loss by $R||\mathbf{w}||/n \leq RB/n$ instead of $2/n$. With that one change, wp $\geq 1 - \delta$

$$L(h) \leq \hat{L}(h) + 2\underbrace{\mathcal{R}(\ell(\mathcal{H}(S)))}_{\leq 2\frac{BR}{\sqrt{n}}} + 3RB\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}$$

where L is the generalization hinge loss and $\hat{L}$ is the sample hinge loss.

In hard margin SVM, the sample hinge loss is 0, while in the soft margin SVM it isn't so.

Putting it all together

Suppose all training examples satisfy $||\mathrm{z}|| \leq R$
Suppose the hypothesis $||\mathrm{w}|| \leq B$. Then w.p. $\geq 1 - \delta$

$$L(h) \leq \hat{L}(h) + 2\mathcal{R}(\ell(\mathcal{H}(S))) + 3RB\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}$$

But we saw already $\mathcal{R}(\ell(\mathcal{H}(S))) \leq \mathcal{R}(\mathcal{H}(S)) \leq RB/\sqrt{n}$. So wp $\geq 1 - \delta$

Putting it all together

Suppose all training examples satisfy $||\mathrm{z}|| \leq R$
Suppose the hypothesis $||\mathrm{w}|| \leq B$. Then w.p. $\geq 1 - \delta$

$$L(h) \leq \hat{L}(h) + 2\mathcal{R}(\ell(\mathcal{H}(S))) + 3RB\sqrt{\frac{2}{n} \ln \left(\frac{2}{\delta}\right)}$$

But we saw already $\mathcal{R}(\ell(\mathcal{H}(S))) \leq \mathcal{R}(\mathcal{H}(S)) \leq RB/\sqrt{n}$. So wp $\geq 1 - \delta$

$$L(h) \leq \hat{L}(h) + 2RB/\sqrt{n} + 3RB\sqrt{\frac{2}{n} \ln \left(\frac{2}{\delta}\right)}$$

Putting it all together

Suppose all training examples satisfy $||z|| \leq R$
Suppose the hypothesis $||w|| \leq B$. Then w.p. $\geq 1 - \delta$

$$L(h) \leq \hat{L}(h) + 2\mathcal{R}(\ell(\mathcal{H}(S))) + 3RB\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}$$

But we saw already $\mathcal{R}(\ell(\mathcal{H}(S))) \leq \mathcal{R}(\mathcal{H}(S)) \leq RB/\sqrt{n}$. So wp $\geq 1 - \delta$

$$L(h) \leq \hat{L}(h) + 2RB/\sqrt{n} + 3RB\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}$$

If discrepancy between true/empirical is to be ϵ, we need $\mathcal{O}(R^2B^2/\epsilon^2)$ samples

Rademacher complexity and neural networks

single (hidden) layer NN, real valued output

Rademacher complexity and neural networks

single (hidden) layer NN, real valued output
input x $\rightarrow$ First layer U $\rightarrow$ relu ϕ $\rightarrow$ Second w $\rightarrow$ output
Network computes $\mathrm{w}^T\phi(U\mathrm{x})$ (Denote: u_j: j'th row of U)

Rademacher complexity $\leq \frac{2B_1R}{\sqrt{n}}$, where

$$B_1 = \sup_{\mathrm{w},U} \sum |w_j| ||\mathrm{u}_j||_2$$

We can prove w.p $\geq 1-\delta$

$$L(h) \leq \hat{L}(h) + \frac{2B_1R}{\gamma\sqrt{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{n}}$$

where γ is margin of classification

Number of hidden layer neurons

Interesting phenomenon with hidden layer
Increase hidden layer size (with ℓ_2 regularization)

Number of hidden layer neurons

Interesting phenomenon with hidden layer
Increase hidden layer size (with ℓ_2 regularization)
Generalization should improve, even after 0 training loss

Number of hidden layer neurons

Interesting phenomenon with hidden layer
Increase hidden layer size (with ℓ_2 regularization)
Generalization should improve, even after 0 training loss

The above bound confirms its validity since

$$B_1 = \sup_{\substack{\mathrm{w}, U \\ ||\mathrm{w}||^2 + ||U||_F^2 \leq 1}} \sum |w_j| ||\mathrm{u}_j||_2 \leq \frac{1}{2}$$

therefore, w.p $\geq 1 - \delta$

$$L(h) \leq \hat{L}(h) + \frac{R}{\gamma\sqrt{n}} + \sqrt{\frac{\log \frac{2}{\delta}}{n}}$$

and γ improves with increasing hidden layer size!

More in this direction

- Can be extended to deep neural networks
  - Still somewhat loose,
  - emphasizes regularization in training DNNs
- Can be extended to adverserial settings as well

Stronger? PAC Bayes

All starts with Donsker-Varadhan variational formula from 1976

Let Θ be our parameter space

Stronger? PAC Bayes

All starts with Donsker-Varadhan variational formula from 1976

Let Θ be our parameter space
For all $h : \Theta \to \mathbb{R}$, measure μ over Θ, Jensen's inequality

$$\ln \mathbb{E}_{\theta\sim\mu} \exp(h(\theta)) \geq \mathbb{E}_{\theta\sim\mu} h(\theta)$$

However, very interestingly for all ρ over Θ,

$$\log \mathbb{E}_{\theta\sim\mu} \exp(h(\theta)) \geq \mathbb{E}_{\nu\sim\rho} h(\nu) - D(\rho||\mu)$$

with equality under the "Gibbs" measure

$$\rho(\theta) \propto \mu(\theta) \exp(h(\theta))$$