Updates: HW #1 now due 3/24

Make sure to download latest version with corrections

In advance of project, if looking for partners, post an introduction about yourself, your research interests on #408partners

Generalization Bounds

So far we've studied the fitting problem:

> "Is there a choice of parameters s.t. my deep net perfectly fits the training data?"

And can we find it efficiently?

This isn't necessarily meaningful by itself (ref: Q3 on HW1)

In the next few lectures we will study the generalization problem:

"If my deep net fits the training data, does that mean it will work on new data?"

Some key definitions:

def: For a hypothesis $h: \mathbb{R}^d \to \{0,1\}$ the empirical error on a training sample

$$S = \{(x_1, y_1) \dots (x_N, y_N)\}$$

is defined as

$$\text{err}_S(h) \triangleq \frac{1}{N} \sum_i \mathbb{1}_{h(x_i) \neq y_i}$$

i.e. the fraction (mislabelled) of the training data

And for a distribution D on $\mathbb{R}^d \times \{0,1\}$ the population error is

$$\text{err}_D(h) \triangleq \mathbb{P}_{(x,y) \sim D}[h(x) \neq y]$$

Finally we define the generalization error as

$$|err_D(h) - err_S(h)|$$

Main Question: when is the generalization error is small?

Standard approach in learning theory is to constrain h to come from some not too complex set $\mathcal{H}$

we'll study how the sample size and the "complexity" of $\mathcal{H}$ give rise to bounds on the generalization error

Case #1: $\mathcal{H}$ is finite

In this case, we'll get bounds from:

tail bounds + union bound

Throughout, assume samples are i.i.d from D

Lemma 1: Fix any $\varepsilon, \delta > 0$. Then if we draw a sample of size

$$N \geq \frac{1}{2\varepsilon^2}\left(\ln 2|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

with probability $\geq 1-\delta$ we have $\forall h \in \mathcal{H}$

$$|err_D(h) - err_S(h)| \leq \varepsilon$$

Here, we do not need to assume $y = h^*(x)$ for some $h^* \in \mathcal{H}$

i.e. there might be no hypothesis that fits the data perfectly

approx. fit the training data $\Rightarrow$ fits the true distribution

The proof is a straightforward application of:

Theorem [Hoeffding] Suppose $X_1, \ldots X_N$ are iid Bernoulli r.v.s. with $\mathbb{E}[X_i] = p$. Then for any $0 < \varepsilon < 1$

$$\mathbb{P}\left[\left|\frac{\sum X_i}{N} - p\right| \geq \varepsilon\right] \leq 2e^{-2N\varepsilon^2}$$

Proof of Lemma 1: For any $h \in \mathcal{H}$ we can view $err_S(h)$ as the average of N i.i.d r.v.s. with expectation $err_D(h)$

Thus by Hoeffding, we have

$$\mathbb{P}\left[|err_D(h) - err_S(h)| \geq \varepsilon\right] \leq 2e^{-2N\varepsilon^2}$$

Now if we set $N = \frac{1}{2\varepsilon^2}\left(\ln(2|\mathcal{H}|) + \ln\frac{1}{\delta}\right)$ we get

$$\mathbb{P}\left[|err_D(h) - err_S(h)| \geq \varepsilon\right] \leq \frac{\delta}{|\mathcal{H}|}$$

and now we can union bound ▨

Remark: Under the assumption $y = h^*(x)$, we can improve the dependence on ε from $\frac{1}{\varepsilon^2}$ to $\frac{1}{\varepsilon}$

Much of generalization theory is about being able to handle richer hypotheses classes

Are there reasonable bounds on generalization error when $\mathcal{H}$ is infinite?

def: We say that a set $S = \{x_1, \ldots x_N\}$ is shattered by $\mathcal{H}$ if

$$y_1, \ldots, y_N \in \{0,1\}$$

$\exists\, h \in \mathcal{H}$, with $h(x_i) = y_i \;\; \forall i$

def: The Vapnik-Chervonenkis (VC) dimension for $\mathcal{H}$ is the largest k s.t. $\exists$ a set S of k samples that is shattered

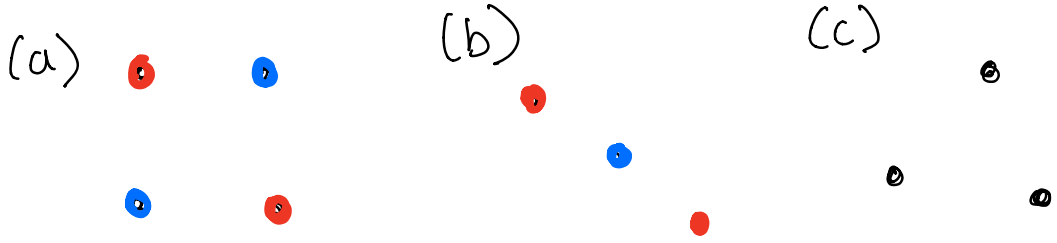
Some Examples

Recall the set of halfspaces:

$$\mathcal{H} = \{h \mid h(x) = \text{sgn}(a^T x + b)\}$$

Which of the following point sets does it shatter?

(a) (b) (c)



Fact: The VC-dimension of a halfspace in $\mathbb{R}^d$ is $d+1$

Warning Many people think of

VC-dimension $\sim$ # parameters in the model

But what about

$$\mathcal{H} = \{h \mid h(x) = \operatorname{sgn}(\sin ax)\}$$

It has only one parameter. Nevertheless:

Fact: The VC-dimension is infinite

It turns out whenever VC-dimension is finite (even though it could be $|\mathcal{H}|=\infty$) we can get bounds on the generalization error:

Theorem 2: Fix any $\varepsilon, \delta > 0$. Suppose $\mathcal{H}$ has VC-dimension k and we draw a sample of size

$$N \geq \frac{c}{\varepsilon^2}\left(k \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}\right)$$

then with probability $\geq 1-\delta$, we have that $\forall h \in \mathcal{H}$

$$|err_D(h) - err_S(h)| \leq \varepsilon$$

we can restate this in a useful form

Corollary:

$$err_D(h) \leq err_S(h) + \sqrt{\frac{c\left(k \ln \frac{N}{k}\right) + \ln\left(\frac{1}{\delta}\right)}{N}}$$

In particular, we need $N \geq k$ before we can get a non-trivial bound

The key idea is to show:

$$\#\text{labellings of } N \text{ samples} \sim N^k$$

def: For a set $S = \{x_1, \ldots, x_N\}$ let $\mathcal{H}(S) \subseteq \{0,1\}^N$ be defined as

$$\mathcal{H}(S) = \{(h(x_1), \ldots, h(x_N)) \mid h \in \mathcal{H}\}$$

In this notation, S is shattered iff

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

The following classic lemma will be the key to our generalization bound

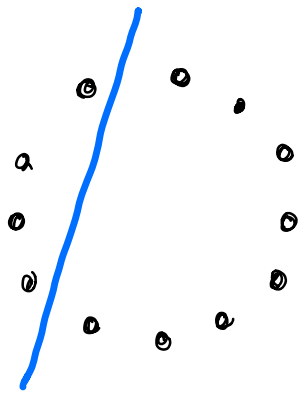
Lemma [Sauer-Shelah] If $\mathcal{H}$ has VC-dim. k then

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

In particular, when $|S| > k$ we have

$$|\mathcal{H}(S)| \leq \left(\frac{e|S|}{k}\right)^k$$

We will omit the proof, but to get some intuition consider



with N points on the boundary. Then

$$|H(S)| \approx 2 \times \binom{N}{2}$$

↑ choice of two arcs to pass thru

To show just the main idea, let's prove a weaker version of theorem 2 that bounds the <u>expected generalization error</u>

<u>Proof:</u> we will be interested in

$$\mathbb{E}_S \left[\sup_{h \in H} |err_D(h) - err_S(h)| \right]$$

which we can rewrite as

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \left| \mathbb{E}_{S'}[err_{S'}(h)] - err_S(h)\right|\right]$$

Now because sup is convex, we can use Jensen's inequality to push it inside the expectation

$$\leq \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \left| err_{S'}(h) - err_S(h)\right|\right] \quad (*)$$

The key idea is instead of sampling $S = \{x_1, \ldots, x_N\}$ and $S' = \{x'_1, \ldots x'_N\}$ we can sample

$$\{x_1, x'_1\}, \ldots \quad \{x_N, x'_N\}$$

and then flip a coin to decide which is x_i and which is x'_i

Thus we can rewrite $(*)$ as $=$

$$\mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}} \left| \frac{1}{N}\sum_i \sigma_i \left(\mathbb{1}_{h(x_i)\neq y_i} - \mathbb{1}_{h(x'_i)\neq y_i}\right)\right|\right]$$

σ: sequence of N ± 1 r.v.s.

But once we fix S and S', any two hypotheses h and h' that are the same restricted to $S \cup S'$ are interchangeable

Thus we can union bound over at most

$$\left(\frac{2eN}{k}\right)^k$$

In particular for any fixed $h \in \mathcal{H}$ we have

$$\mathbb{P}_{S,S',\sigma}\left[\left|\frac{\sum_i \sigma_i (\mathbb{1}_{h(x_i)\neq y_i} - \mathbb{1}_{h(x_i')\neq y_i'})}{N}\right|\right] \leq 2e^{-\frac{\varepsilon^2 N}{2}}$$

and if we plug in $N = \frac{C}{\varepsilon^2} k \ln\frac{1}{\varepsilon}$ we get

$$(*) \leq \frac{\varepsilon}{2} + \left(\frac{eN}{k}\right)^k \times 2e^{-\frac{\varepsilon^2 N}{2}} \leq \varepsilon$$

which completes the proof. ☒

Back to Deep Learning

Main Question: Can classic results give non-trivial generalization bounds?

Let's restrict to sgn activations. Then our basic model becomes

$$f = f_L \circ f_{L-1} \circ \dots \circ f_1$$

where each $f_i = \operatorname{sgn}(A_i x + b_i)$

A_i: $d_i \times d_{i-1}$, b_i: d_i

Theorem 3 The VC dimension of $\mathcal{H} = \{f \mid \text{given architecture}\}$ is $O(p \log p)$ where

$$p = \sum_{i=1}^{L} d_i (d_{i-1} + 1) = \#\text{ of parameters}$$

We'll prove this by working with more general labelling functions

$$X \longrightarrow Y$$

Y: finite set

def: The growth function of $\mathcal{H}$ is

$$G_{\mathcal{H}}(N) \triangleq \max_{x_1,\dots x_N \in \mathcal{X}} |\mathcal{H}(\{x_1,\dots x_N\})|$$

i.e. it is the largest # of distinct hypotheses, when restricting to N examples

Next we will prove that the growth function behaves nicely under some elementary operations

"functions from $\mathcal{X}$ to $\mathcal{Y}_1$"

Lemma 2 Let $\mathcal{H}_1 \subseteq \mathcal{Y}_1^{\mathcal{X}}$ and $\mathcal{H}_2 \subseteq \mathcal{Y}_2^{\mathcal{X}}$

Now set $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$

Then we have $G_{\mathcal{H}}(N) \leq G_{\mathcal{H}_1}(N)\, G_{\mathcal{H}_2}(N)$

Proof: Each $h \in \mathcal{H}$ is a function from $\mathcal{X}$ to $\mathcal{Y}_1 \times \mathcal{Y}_2$. Thus for any $x_1,\dots x_N \in \mathcal{X}$ we have

$$|\mathcal{H}(\{x_1,\dots x_N\})| = |\mathcal{H}_1(\{x_1,\dots x_N\})|\,|\mathcal{H}_2(\{x_1,\dots x_N\})|$$

$$\leq G_{\mathcal{H}_1}(N)\, G_{\mathcal{H}_2}(N)$$

□

Lemma 3: Let $\mathcal{H}_1 \leq Y_1^X$ and $\mathcal{H}_2 \leq Y_2^{Y_1}$ and set $\mathcal{H} = \mathcal{H}_2 \circ \mathcal{H}_1$

Then we have $G_{\mathcal{H}}(N) \leq G_{\mathcal{H}_1}(N)\, G_{\mathcal{H}_2}(N)$

Proof: Consider any $x_1, \ldots, x_N \in X$. Then by definition we have

$$\mathcal{H}(\{x_1 \ldots x_N\}) = \left\{ (h_2(h_1(x_1)), \ldots, h_2(h_1(x_N))) \,\middle|\, \begin{matrix} h_1 \in \mathcal{H}_1 \\ h_2 \in \mathcal{H}_2 \end{matrix} \right\}$$

$$= \bigcup_{\vec{u} \in \mathcal{H}_1(\{x_1 \ldots x_N\})} \{ (h_2(u_1), \ldots h_2(u_N)) \,|\, h_2 \in \mathcal{H}_2 \}$$

Thus we have

$$|\mathcal{H}(\{x_1 \ldots x_N\})| \leq \sum_{\vec{u} \in \mathcal{H}_1(\{x_1 \ldots x_N\})} |\{ (h_2(u_1), \ldots h_2(u_N)) \,|\, h_2 \in \mathcal{H}_2 \}|$$

$$\leq \sum_{\vec{u} \in \mathcal{H}_1(\{x_1 \ldots x_N\})} G_{\mathcal{H}_2}(N)$$

$$\leq G_{\mathcal{H}_1}(N)\, G_{\mathcal{H}_2}(N)$$

Now let's return to Theorem 3

Proof of Theorem 3: We can write

$$\mathcal{H} = \mathcal{H}_L \circ \mathcal{H}_{L-1} \cdots \circ \mathcal{H}_1$$

where each $\mathcal{H}_i = \mathcal{H}_{(i,1)} \times \cdots \times \mathcal{H}_{(i,d_i)}$

where $\mathcal{H}_{(i,j)}$ is a set of halfspaces $\operatorname{sgn}(a^T x + b)$, j^{th} row of A_i

The VC-dimension of $\mathcal{H}_{(i,j)}$ is $d_{i-1}+1$ and by the Sauer-Shelah Lemma we get

$$G_{\mathcal{H}_{(i,j)}}(N) \leq \left(\frac{eN}{d_{i-1}+1}\right)^{d_{i-1}+1}$$

Now using Lemma 2 and Lemma 3 we have

$$G_{\mathcal{H}}(N) \overset{\text{lemma 3}}{\leq} \prod_{i=1}^{L} G_{\mathcal{H}_i}(N)$$

$$\overset{\text{lemma 2}}{\leq} \prod_{i=1}^{L} \prod_{j=1}^{d_i} G_{\mathcal{H}(i,j)}(N)$$

$$\leq \prod_{i=1}^{L} \prod_{j=1}^{d_i} \left(\frac{Ne}{d_{i-1}+1}\right)^{d_{i-1}+1}$$

$$\leq \prod_{i=1}^{L} \prod_{j=1}^{d_i} (Ne)^{d_{i-1}+1} \leq (Ne)^{p}$$

Now we can go the other direction and use a bound on the growth function to get a bound on the VC-dimension:

Suppose $\mathcal{H}$ shatters S. Then

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

However from our bound on the growth function

$$|\mathcal{H}(S)| \leq (e|S|)^{p}$$

$\Rightarrow 2^{|S|} \leq (e|S|)^p$

$\Rightarrow |S| \leq p \log e|S| \Rightarrow |S| = O(p \log p)$ 

So what's the problem with applying VC-bounds in the context of DL?

NTKs need width $\geq$ # examples

VC theory needs #examples $\geq$ #parameters

Actually this disconnect is inevitable because:

NTKs fit any collection of data, w/o assumptions on the structure

Generalization is only possible when the data has structure

Returning to VC-bounds, it is actually a tight characterization of learnability

Let's unpack what this means

def: [Valiant] A concept class is PAC learnable if for any distribution D on X there is an algorithm that (not necessarily efficient)

(1) takes $N \geq N(\varepsilon, \delta, \mathcal{H})$

$x \sim D$ and $y = h^*(x)$ for some $h^* \in \mathcal{H}$

and outputs a hypothesis $h \in \mathcal{H}$ that satisfies

(2) $$\mathbb{P}_{\substack{x \sim D \\ y = h^*(x)}} [h(x) \neq h^*(x)] \leq \varepsilon$$

with probability $\geq 1 - \delta$.

Here PAC stands for

Probably Approximately Correct

A key result in the area is:

Theorem [Blumer, Ehrenfeucht, Haussler, Warmuth]: A concept class $\mathcal{H}$ is PAC learnable iff it has finite VC-dimension

The generalization bound in theorem 3 proves one side:

finite VC-dim $\Rightarrow$ PAC learnable

But why is it impossible to PAC learn a concept class with ∞ VC-dimension?

Main Idea: Consider a sequence of distr. $D_1, D_2 \ldots$ that are uniform over larger and larger sets that are shattered

You can ensure that every new example has its label set randomly, which means you can't learn until you've seen most examples

Note: This characterization works even when there is noise. In particular a concept class $\mathcal{H}$ is

agnostically PAC learnable $\iff$ it has finite VC-dim

(want h with $\text{err}_D(h) \leq \text{opt} + \varepsilon$)

While this characterization is clean and elegant, it also points a way forward:

what if the distr. D is not worst-case?

Are there better "sample dependent" generalization bounds?

def: [Koltchinski, Panchenko] The Rademacher complexity of a concept class $\mathcal{H}$ is

$$R_N(\mathcal{H}) \triangleq \mathop{\mathbb{E}}_{S} \left[\mathop{\mathbb{E}}_{\sigma} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{N} \sum \sigma_i \mathbb{1}_{h(x_i) \neq y_i} \right| \right] \right]$$

Literally, we are labelling $x_1, \ldots, x_N$ with random labels $\sigma_1, \ldots, \sigma_N$ and asking for the best correlation with an $h \in \mathcal{H}$

Intuitively

worse correlation $\Rightarrow$ less expressive concept class $\Rightarrow$ better generalization bounds

In particular:

Theorem 4: Fix any $\varepsilon, \delta > 0$. Then with probability $\geq 1-\delta$ we have

$$\sup_{h \in H} |err_D(h) - err_S(h)| \leq 2R_N(H) + \sqrt{\frac{\log 1/\delta}{N}}$$

we will not prove this, but note that in our bound on the expected generalization error what we actually showed was

$$\mathbb{E}_S\left[\sup_{h \in H} |err_D(h) - err_S(h)|\right] \leq 2\,\mathbb{E}_S\left[\sup_{h \in H} \left|\frac{1}{N}\sum \sigma_i \mathbb{1}_{h(x_i) \neq y_i}\right|\right]$$

This bound can be, and often is, much sharper

e.g. if the VC-dimension is large but typical set cannot be shattered

This leads us to a natural question:

"Even if the VC-dim is large, could it be that Rademacher complexity is small?"

Maybe we never find some settings of parameters that are needed to shatter real world data?

e.g. because we get stuck in a local minima, or maybe we're different problem

Unfortunately in a landmark experiment, Zhang et al. dashed these hopes

Experimental Findings: Deep nets learned via SGD can fit perfectly random labels on real data

So deep nets truly are that expressive

Epilogue

It's not all about the # of parameters

First, there are many standard ways to generalize notions like VC-dimension to real-valued functions:

def: We say that $x_1, \ldots, x_N$ are γ-shattered if $\exists\, r_1, \ldots r_k \in \mathbb{R}$ such that for any $b_1, \ldots, b_k \in \{\pm 1\}$

$$(h(x_i) - r_i) b_i \geq \gamma$$

$\forall h \in \mathcal{H}$

And the fat shattering dimension at scale γ, denoted by $\text{fat}_{\mathcal{H}}(\gamma)$, is largest k s.t.

$\exists\, x_1 \ldots x_k$ that are γ-shattered

In a seminal work, Bartlett showed

$$\text{fat-shattering dimension} << \#\text{parameters}$$

Theorem [Bartlett]: Suppose

$$X = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq B\}$$

Further suppose for any hidden unit its weight vector w satisfies $\|w\|_1 \leq A$. Then

$$fat_H(\gamma) = O\left(\frac{B^2 (cA)^{L(L+1)}}{\gamma^{2L}} \log n\right)$$

Recently, other generalization bounds in this vein have been established, e.g. see Bartlett, Foster, Telgarsky

Also you can get better bounds on the generalization error if you can bound how much your hypothesis depends on the samples

To do so, we will work with distributions p and q over hypotheses

$$err_D(p) = \mathop{\mathbb{E}}_{h \sim p}[err_D(p)]$$

and similarly for $\text{err}_S(p)$. Then $KL = \int \log \frac{q(w)}{p(w)} dq$

Theorem [McAllester] For any $\gamma, \delta > 0$ and sample of size N, we have

$$\text{err}_D(q) \leq \frac{1}{\gamma}\text{err}_S(q) + \frac{KL(q\|p) + \ln\frac{1}{\delta}}{2\gamma(1-\gamma)N}$$

(q: distribution on hypotheses learned from data)

with probability $\geq 1-\delta$, and p is a reference distribution on hypotheses

Led to non-vacuous generalization bounds for MNIST, see Dziugaite, Roy

General Mantra, a la Occam's Razor

If you can describe your hypothesis using $o(N)$ bits, and it fits your data $\Rightarrow$ must be learning

We can think of many of our bounds in terms of increasingly sophisticated ways to describe our hypothesis

Case #1: $\mathcal{H}$ is finite

Then describe h using $\log|\mathcal{H}|$ bits

Case #2: $\mathcal{H}$ has finite VC-dimension

On a large enough sample of size N, it only takes $k \log N$ bits to describe the labelling

Case #3: PAC Bayes

If you take a reference p as given, you only need to describe how p changes into q ($KL(q||p)$)

Next time: the role of regularization (explicit vs. implicit) in generalization, and double-descent