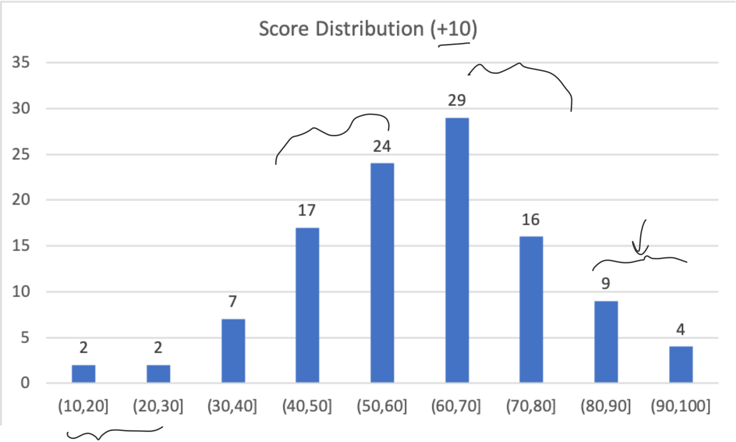
Learning From Data
Lecture 7: Learning Theory

Yang Li **yangli@sz.tsinghua.edu.cn**

TBSI

November 18, 2022

Midterm Results



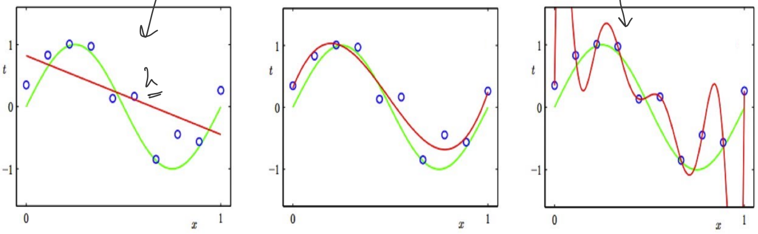
	max	mean	median
curved score	100	60	61

Review

Overfit & Underfit

Underfit Both training error and testing error are large

Overfit Training error is small, testing error is large

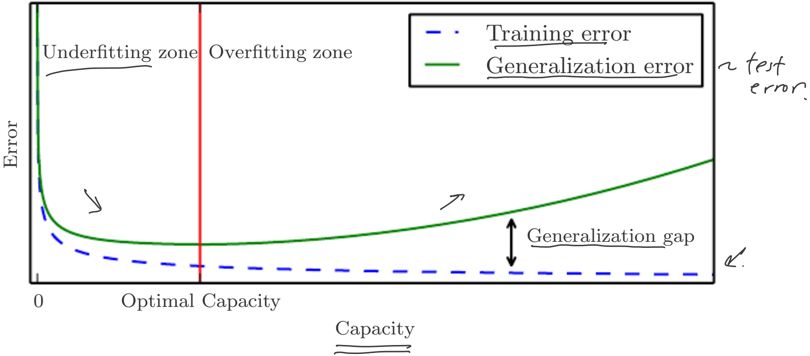


underfit overfit

Model capacity: the ability to fit a wide variety of functions

Model Capacity

Changing a model's **capacity** controls whether it is more likely to overfit or underfit



How to formalize this idea?

Bias and Variance

Suppose data is generated by the following model:

$$y = h(x) + \epsilon \qquad \sim P_{XY}$$

with $\mathbb{E}[\epsilon] = 0$, $Var(\epsilon) = \sigma^2$

- $h(x)$: true hypothesis function, unknown
- $\hat{h}_D(x)$: estimated hypothesis function based on training data $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ sampled from P_{XY}
- **Model bias:** $Bias(\hat{h}_D(x)) = \mathbb{E}_D[\hat{h}_D(x) - h(x)]$ *Expected estimation error of the model over all choices of training data D*
- **Model variance:** $Var(\hat{h}_D(x)) = \mathbb{E}_D[\hat{h}_D(x)^2] - \mathbb{E}_D[\hat{h}_D(x)]^2$ *Variance of the model over all choices of D*

Bias - Variance Tradeoff

If we measure generalization error by MSE $y = h(x) + \varepsilon$ $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

test error

$$\underline{MSE} = \mathbb{E}[(\hat{h}_D(x) - y)^2] = \boxed{\underline{Bias}(\hat{h}_D(x))^2 + \underline{Var}(\hat{h}_D(x))} + \underline{\sigma}^2,$$

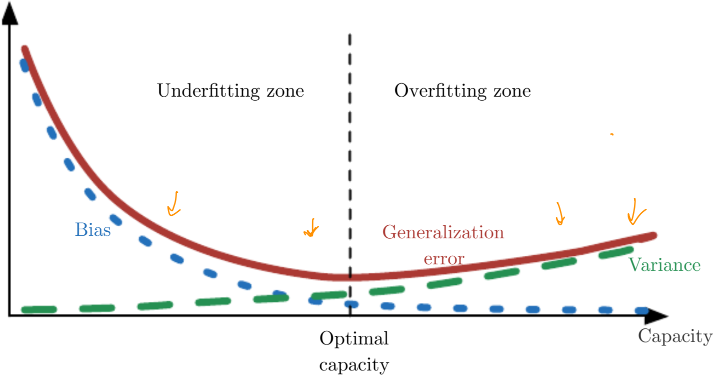
- σ^2 represents irreducible error *(caused by noisy data)*
- in practice, increasing capacity tends to increase variance and decrease bias.

Bias - Variance Tradeoff

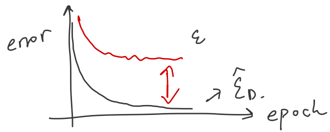
If we measure generalization error by MSE

$$MSE = \mathbb{E}[(\hat{h}_D(x) - y)^2] = Bias(\hat{h}_D(x))^2 + Var(\hat{h}_D(x)) + \sigma^2,$$

- σ^2 represents irreducible error *(caused by noisy data)*
- in practice, increasing capacity tends to increase variance and decrease bias.



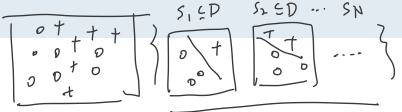
Exercise:



overfitting

When the training error is much smaller than the testing error in a regression problem, what should be done? Select all that apply.

- A) Add more training data. ✓
- B) Reduce model complexity. ✓
- C) Add more features. → reduce Bias ×
- D) Apply random transformation to the training data *(data augmentation)*. ✓
  → more training data
  → equivalent to adding implicit regularization
- E) Train multiple models on random subsets of the training data; Make prediction by averaging of the output of each model. *(bagging a.k.a. boostrap aggregation)* ✓



$Var(h_i) = \sigma^2$.

$Var(\sum_{i=1}^{N} h_i) = \frac{\sigma^2}{N}$. reduces variance

Today's Lecture

- How to measure model capacity?
- Can we find a theoretical guarantee for model generalization?

A brief introduction to learning theory

- Empirical risk ~~estimation~~ minimization
- Generalization bound for finite and infinite hypothesis space

Final project information.

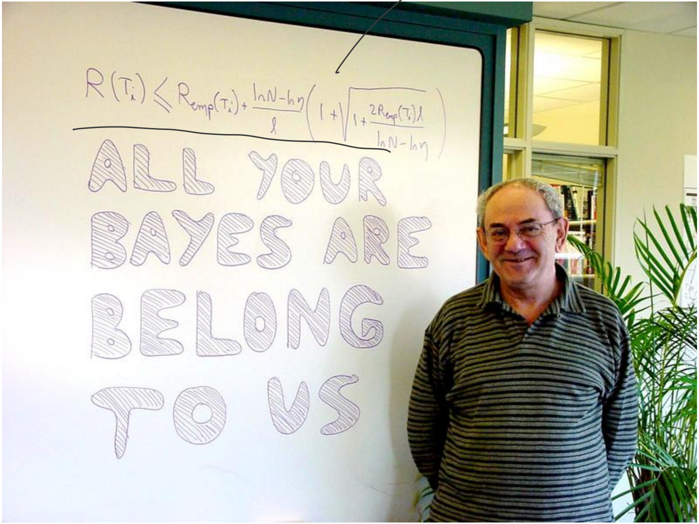
Introduction to Learning Theory

- Empirical risk estimation
- Learning bounds
  - Finite Hypothesis Class
  - Infinite Hypothesis Class

Learning theory

How to quantify generalization error?

uniform convergence



Prof. Vladimir Vapnik in front of his famous theorem

Empirical risk

Simplified assumption: $y \in (0, 1)$

- Training set: $S = (x^{(i)}, y^{(i)}); i = 1, \ldots, m$ with $(x^{(i)}, y^{(i)}) \sim \mathcal{D}$
- For hypothesis h, the **training error** or **empirical risk/error** in learning theory is defined as

0-1 error

$$\checkmark \quad \hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\}$$

- The **generalization error** is

P_{xy}

$$\epsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$$

- **PAC assumption**: assume that training data and test data (for evaluating generalization error) were drawn from the same distribution $\mathcal{D}$

Hypothesis Class and ERM

$\mathcal{H} = \{ h_\theta(x) = 1\{\theta^T x \geq 0\} \mid \theta \in \mathbb{R}^n \}$

Hypothesis class

The **hypothesis class** $\boxed{\mathcal{H}}$ **used by a learning algorithm** is the set of all classsifiers considered by it.

e.g. Linear classification considers $h_{\boxed{\theta}}(x) = 1\{\theta^T x \geq 0\} = \begin{cases} 1 & \theta^T x \geq 0. \\ 0. & \text{o.w.} \end{cases}$

Empirical Risk Minimization (ERM): the "simplest" learning algorithm: pick the best hypothesis h from hypothesis class $\mathcal{H}$

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \underbrace{\hat{\epsilon}(h)}_{\text{empirical risk.}}$$

$\hat{\varepsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x^i) \neq y^i\}$

How to measure the generalization error of empirical risk minimization over $\mathcal{H}$*?*

- Case of finite $\mathcal{H}$
- Case of infinite $\mathcal{H}$

Case of Finite $\mathcal{H}$

Goal: give guarantee on generalization error $\epsilon(h)$ = $\mathbb{E}_{x,y\sim D} 1\{h(x) \neq y\}$

- Show $\hat{\epsilon}(h)$ (training error) is a good estimate of $\epsilon(h)$
- Derive an upper bound on $\epsilon(h)$

For any $h_i \in \mathcal{H}$, the event of h_i miss-classification given sample $(x, y) \sim \mathcal{D}$:

$$Z = 1\{h_i(x) \neq y\}$$

$Z \in \{0, 1\}$.

$Z_j = 1\{h_i(x^{(j)}) \neq y^{(j)}\}$: event of h_i miss-classifying sample $x^{(j)}$

Case of Finite $\mathcal{H}$

Goal: give guarantee on generalization error $\epsilon(h)$

- Show $\hat{\epsilon}(h)$ (training error) is a good estimate of $\epsilon(h)$
- Derive an upper bound on $\epsilon(h)$

For any $h_i \in \mathcal{H}$, the event of h_i miss-classification given sample $(x, y) \sim \mathcal{D}$:

$$Z = 1\{h_i(x) \neq y\}$$

$Z_j = 1\{h_i(x^{(j)}) \neq y^{(j)}\}$: event of h_i miss-classifying sample $x^{(j)}$

Training error of $h_i \in \mathcal{H}$ is:

$$\hat{\epsilon}(h_i) = \underbrace{\frac{1}{m} \sum_{j=1}^{m} 1\{\underbrace{h_i(x^{(j)}) \neq y^{(j)}}\}}$$

$$\hat{\epsilon}(h_i) = \frac{1}{m} \sum_{j=1}^{m} Z_j$$

Preliminaries

Here we make use of two famous inequalities:

Lemma 1 (Union Bound)

Let $A_1, A_2, \ldots, A_k$ be k different events, then

$$P(A_1 \cup \ldots \cup A_k) \leq P(A_1) + \ldots + P(A_k)$$

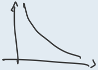
Probability of any one of k events happening is less the sums of their probabilities.

Preliminaries

Lemma 2 (Hoeffding Inequality, Chernoff bound)

Let $Z_1, \ldots, Z_m$ be m i.i.d. random variables drawn from a Bernoulli(ϕ) distribution. i.e. $P(Z_i = 1) = \phi$, $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ be the sample mean of RVs.
For any $\gamma > 0$,

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2\exp(-2\gamma^2 m)$$



The probability of $\hat{\phi}$ having large estimation error is small when m is large!

sample mean

Case of Finite $\mathcal{H}$

Training error of $h_i \in \mathcal{H}$ is:

ϕ

$$\hat{\epsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j$$

ϕ

where $Z_j \sim Bernoulli(\epsilon(h_i))$

Case of Finite $\mathcal{H}$

Training error of $h_i \in \mathcal{H}$ is:

Hoeffding.

Given γ,

$P(|\phi - \hat{\phi}| > \gamma) \le 2e^{-2\gamma^2 m}$

$$\hat{\epsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j$$

where $Z_j \sim Bernoulli(\epsilon(h_i))$ Given γ.

By Hoeffding inequality, P for every $h_i \in \mathcal{H}$

$$P(|\epsilon(h_i) - \hat{\epsilon}(h_i)| > \gamma) \le \underline{2e^{-2\gamma^2 m}} \quad (1)$$

Let A_i be the event (R.V) that $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$, $i = 1, \dots, k$

Then $P_r(\exists h \in \mathcal{H} \,|\, |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) = P(A_1 \cup A_2 \cdots \cup A_k)$

there exists

By the Union bound, $\le \sum_{i=1}^{k} P(A_i) = \sum_{i=1}^{k} \underbrace{P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma)}$.

By (1), $\le \sum_{i=1}^{k} 2e^{-2\gamma^2 m} = 2ke^{-2\gamma^2 m}$

Then, by negation

$P_r(\forall h \in \mathcal{H} \,|\, |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \le \gamma) \ge 1 - 2ke^{-2\gamma^2 m}$

Case of Finite $\mathcal{H}$

Training error of $h_i \in \mathcal{H}$ is:

$$\hat{\epsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j$$

where $Z_j \sim Bernoulli(\epsilon(h_i))$

By Hoeffding inequality,

$$P(\underbrace{|\epsilon(h_i) - \hat{\epsilon}(h_i)| > \gamma}_{A_i}) \leq 2e^{-2\gamma^2 m}$$

By Union bound,

$$P(\forall h \in \mathcal{H}.|\epsilon(h) - \hat{\epsilon}(h)| \leq \gamma) \geq 1 - \underbrace{2ke^{-2\gamma^2 m}}_{\delta}$$

Uniform Convergence Results

Proposition.

Given γ, m

$P(\forall h \in \mathcal{H} \mid |\varepsilon(h) - \hat{\varepsilon}(h)| \le \gamma) \ge 1 - 2k e^{-2\gamma^2 m}$

Corollary 3

Given γ and $\delta > 0$, If

$$m \ge \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

where $k = |\mathcal{H}|$

Then with probability at least $1 - \delta$, we have $|\epsilon(h) - \hat{\epsilon}(h)| \le \gamma$ for all $\mathcal{H}$. m is called the algorithm's **sample complexity**.

let. $\delta = 2ke^{-2\gamma^2 m}$

$\log \delta = \log(2k) + (-2\gamma^2 m)$.

$m = \frac{\log \delta - \log 2k}{-2\gamma^2} = \frac{\log 2k - \log \delta}{2\gamma^2} = \frac{1}{2\gamma^2} \log\left(\frac{2k}{\delta}\right)$ ← minimum sample size for

$P(\forall h \in \mathcal{H} \mid |\varepsilon(h) - \hat{\varepsilon}(h)| \le \gamma) \ge 1 - \delta$.

Uniform Convergence Results

Corollary 3

Given γ *and* $\delta > 0$, *If*

$$m \geq \frac{1}{2\gamma^2} \log \frac{2\boxed{k}}{\delta}$$

Then with probability at least $1 - \delta$, *we have* $|\epsilon(h) - \hat{\epsilon}(h)| \leq \gamma$ *for all* $\mathcal{H}$.
m is called the algorithm's **sample complexity**.

Remarks

- Lower bound on m tell us how many training examples we need to make generalization guarantee.
- # of training examples needed is logarithm in k

Uniform Convergence Results

$\left.\begin{array}{l}\gamma \\ \delta \\ m\end{array}\right\}$

Proposition.

Given γ, m

$$P(\forall h \in \mathcal{H} \mid |\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma) \geq 1 - 2ke^{-2\gamma^2 m}$$

Corollary 3. For $|\hat{\varepsilon}(h) - \varepsilon(h)| \leq \gamma$ to hold for all $h \in \mathcal{H}$, $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$.

Corollary 4

With probability $1 - \delta$, *for all* $h \in \mathcal{H}$, given m samples

$$|\hat{\epsilon}(h) - \epsilon(h)| \leq \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

By corollary 3, solve for γ: $2\gamma^2 = \frac{1}{m} \log \frac{2k}{\delta}$.

$$\gamma = \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}.$$

Then $|\hat{\varepsilon}(h) - \varepsilon(h)| \leq \gamma = \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$. for all $h \in \mathcal{H}$.

Uniform Convergence Results

Corollary 4

With probability $1 - \delta$, *for all* $h \in \mathcal{H}$,

$$|\hat{\epsilon}(h) - \epsilon(h)| \leq \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

What is the convergence result when we pick $\hat{h} = \text{argmin}_{h \in \mathcal{H}} \hat{\epsilon}(h)$

Uniform Convergence Theorem for Finite $\mathcal{H}$

Using previous corollaries, we can bound $\epsilon(\hat{h})$:

Theorem 5 (Uniform convergence)

Let $|\mathcal{H}| = k$, and m,δ be fixed. With probability at least $1-\delta$, we have

$$\boxed{\epsilon(\hat{h})} \leq \Big(\underbrace{\min_{h\in\mathcal{H}'} \epsilon(h)}_{\varepsilon(h^*)}\Big) + \boxed{2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}} \sim \text{variance}$$

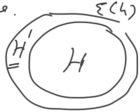
1), Choose a larger $H' \supseteq H$, $\min_{h\in H} \varepsilon(h)$. will decrease.

i.e. $\min_{h\in H'} \varepsilon(h) \le \min_{h\in H} \varepsilon(h)$ → smaller bias.

2) When H is larger, $K = |H|$ increases.

then $2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$. increases

bias. variance tradeoff.



training risk/error $\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x) \neq y\}$,

testing/generalization risk. $\varepsilon(h) = \mathbb{E}_{x,y\sim D} 1\{h(x) \neq y\}$.

$\hat{h} = \arg\min_{h\in H} \hat{\varepsilon}(h)$ ← empirical estimator (using ERM)

$h^* = \arg\min_{h\in H} \varepsilon(h)$ ← true hypothesis

With probability at least $1-\delta$, sample size $m \geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta}$. we have

$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ for all $h \in H$. (by corollary 3)

Then $|\varepsilon(\hat{h}) - \hat{\varepsilon}(\hat{h})| \leq \gamma$ since $\hat{h} \in H$

Fact. $|y| \leq a$

$\boxed{-a \leq y \leq a}$

$\varepsilon(\hat{h}) - \hat{\varepsilon}(\hat{h}) \leq \gamma$.

$\varepsilon(\hat{h}) \leq \gamma + \hat{\varepsilon}(\hat{h}) \leq \gamma + \hat{\varepsilon}(h^*)$. (1).

since $\hat{h} = \arg\min_{h\in H} \hat{\varepsilon}(h)$, $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h^*)$

Similarly, $|\varepsilon(h^*) - \hat{\varepsilon}(h^*)| \leq \gamma$

$\varepsilon(h^*) - \hat{\varepsilon}(h^*) \geq -\gamma$.

$\hat{\varepsilon}(h^*) \leq \varepsilon(h^*) + \gamma$,

By (1), $\varepsilon(\hat{h}) \leq \gamma + \hat{\varepsilon}(h^*) \leq \gamma + \varepsilon(h^*) + \gamma$.

$\leq \varepsilon(h^*) + 2\gamma \leq \underbrace{\varepsilon(h^*)}_{\min_{h\in H}\varepsilon(h)} + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$.

$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$.

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d-1$ unknowns.

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

$\theta_d \in \mathbb{R}.$

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d-1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$ ← # of parameter.

one real #

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d - 1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$
- How many samples do we need to guarantee $\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\gamma$ to hold with probability at least $1 - \delta$?

$$\underline{m} \geq O\left(\underline{\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}}\right) = O\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right) = \underline{O_{\gamma,\delta}(d)}$$

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \dots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d-1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$
- How many samples do we need to guarantee $\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\gamma$ to hold with probability at least $1-\delta$?

$$m \geq O\left(\frac{1}{\gamma^2}\log\frac{2^{64d}}{\delta}\right) = O\left(\frac{d}{\gamma^2}\log\frac{1}{\delta}\right) = O_{\gamma,\delta}(d)$$

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d-1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$
- How many samples do we need to guarantee $\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\gamma$ to hold with probability at least $1-\delta$?

$$m \geq O\left(\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}\right) = O\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right) = O_{\gamma,\delta}(d)$$

To learn ***well****, the number of samples has to be linear in d*

Infinite hypothesis class: Challenges

Size of $\mathcal{H}$ depends on the choice of parameterization

Example

$2n+2$ parameters:

$$h_{u,v} = \mathbf{1}\{\overset{\theta_0}{(u_0^2 - v_0^2)} + \overset{\theta_1}{(u_1^2 - v_1^2)}x_1 + \ldots + \overset{\theta_n.}{(u_n^2 - v_n^2)}x_n \geq 0\}$$

is equivalent the hypothesis with $n+1$ parameters:

$$h_\theta(x) = \mathbf{1}\{\theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n \geq 0\}$$

Infinite hypothesis class: Challenges

Size of $\mathcal{H}$ depends on the choice of parameterization

Example

$2n+2$ parameters:

$$h_{u,v} = \mathbf{1}\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + \ldots + (u_n^2 - v_n^2)x_n \geq 0\}$$

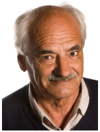
is equivalent the hypothesis with $n+1$ parameters:

$$h_\theta(x) = \mathbf{1}\{\theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n \geq 0\}$$

We need a complexity measure of a hypothesis class invariant to parameterization choice

Infinite hypothesis class: Vapnik-Chervonenkis theory

A computational learning theory developed during 1960-1990 explaining the learning process from a statistical point of view.



Alexey Chervonenkis (1938-2014), Russian mathematician

Vladimir Vapnik (Facebook AI Research, Vencore Labs) Most known for his contribution in statistical learning theory

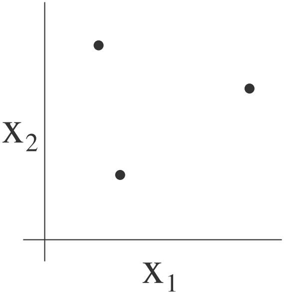
Shattering a point set

- Given d points $x^{(i)} \in \mathcal{X}$, $i = 1, \ldots, d$, $\mathcal{H}$ **shatters** S if $\mathcal{H}$ can realize any labeling on S. y.

① $\mathcal{H}$ realize a labeling on S: $\exists\, h \in \mathcal{H}$, $h(x)$ makes no mistakes in predicting the labeling

② $\mathcal{H}$ can realize any labeling on S.

Example: $S = \{x^{(1)}, x^{(2)}, x^{(3)}\}$ where $x^{(i)} \in \mathbb{R}^2$.



Suppose $y^{(i)} \in \{0, 1\}$, how many possible labelings does S have? 2^3

Shattering a point set

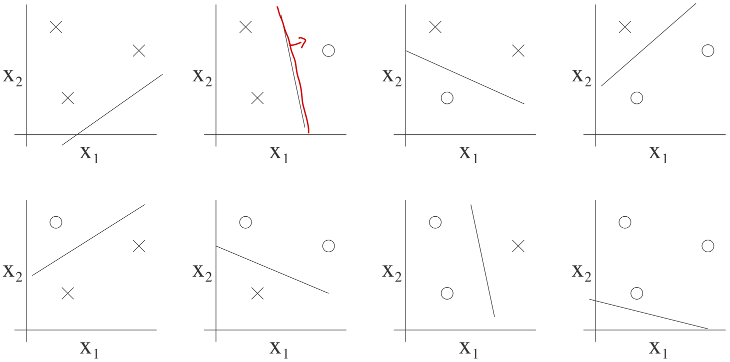
linear threshold function in $\mathbb{R}^2$.

- Example: Let $\mathcal{H}_{LTF,2}$ be the linear threshold function in $\mathbb{R}^2$ (e.g. in the perceptron algorithm)

$\mathcal{H}_{LTF,2}$ can shatter S

$|S| = 3$.

$$h(x) = \begin{cases} 1 & w_1x_1 + w_2x_2 \geq b \\ 0 & \text{otherwise} \end{cases}$$

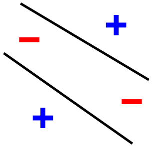


$\mathcal{H}_{LTF,2}$ shatters $S = \{x^{(1)}, x^{(2)}, x^{(3)}\}$

VC Dimension

The **Vapnik-Chervonenkis** dimension of $\mathcal{H}$, or $VC(\mathcal{H})$, is the cardinality of the largest set shattered by $\mathcal{H}$. $VC(H_{LTF,2}) \geq 3.$

- Example: $\boxed{VC(H_{LTF,2}) = 3}$



$\Rightarrow VC(H_{LTF,2}) < 4.$

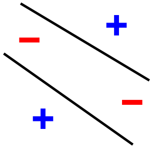
$\mathcal{H}_{LTF}$ can not shatter 4 points: for any 4 points, label points on the diagonal as '+'. (See Radon's theorem)

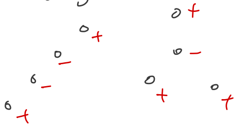
VC Dimension

The **Vapnik-Chervonenkis** dimension of $\mathcal{H}$, or $VC(\mathcal{H})$, is the cardinality of the largest set shattered by $\mathcal{H}$.

- Example: $VC(H_{LTF,2}) = 3$

To show $VC(\mathcal{H}_{LTF,2}) < 4$,
Given any set S, with $|S| = 4$.
Construct a labeling by assigning opposite labels on the diagonal.





$\mathcal{H}_{LTF}$ can not shatter 4 points: for any 4 points, label points on the diagonal as '+'. (See Radon's theorem)

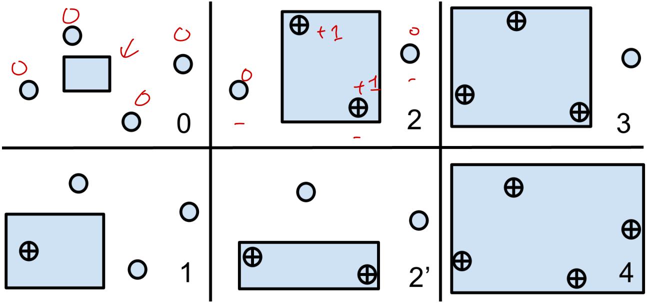
- To show $VC(\mathcal{H}) \geq d$, it's sufficient to find **one** set of d points shattered by $\mathcal{H}$ ✓
- To show $VC(\mathcal{H}) < d$, need to prove $\mathcal{H}$ doesn't shatter any set of d points ✓

VC Dimension

$h_{AAR.} = \begin{cases} 1. & x \in \text{Rectangle.} \\ 0. & \text{otherwise} \end{cases}$

- Example: $VC(AxisAlignedRectangles) = 4$

$|S| = 4.$



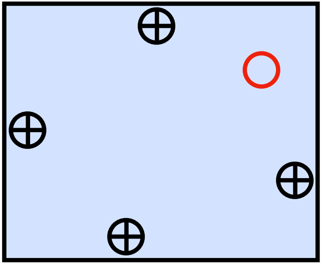
2^4

Axis-aligned rectangles can shatter 4 points. $VC(AxisAlignedRectangles) \geq 4$

VC Dimension

- Example: $VC(AxisAlignedRectangles) = 4$

$h \in H_{AAR}$



For all $|S| = 5$, we can find such labeling that H_{AAR} can not realize!

For any 5 points, label topmost, bottommost, leftmost and rightmost points as "+".
$VC(AxisAlignedRectangles) < 5$

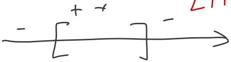
Discussion on VC Dimension

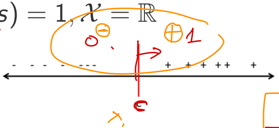
① $VC(PHL) \le 1$.

More VC results of common $\mathcal{H}$:

- $VC(PositiveHalf\text{-}Lines) = 1, \mathcal{X} = \mathbb{R}$

LTF in $\mathbb{R}$.

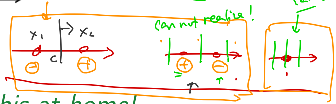




② $VC(PHL) < 2$.

label by diff labels

can not realize!



- $VC(Intervals) = 2, \mathcal{X} = \mathbb{R}$
- $VC(LTF \text{ in } \mathbb{R}^n) = n + 1, \mathcal{X} = \mathbb{R}^n$ ← *prove this at home!*

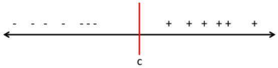
Let $c = \frac{x_2 - x_1}{2}$ assuming $x_2 > x_1$

label $\min(x_1, x_2)$ as $+1$, and the other as 0.

Discussion on VC Dimension

More VC results of common $\mathcal{H}$:

- $VC(PositiveHalf\text{-}Lines) = 1, \mathcal{X} = \mathbb{R}$



- $VC(Intervals) = 2, \mathcal{X} = \mathbb{R}$
- $VC(LTF \text{ in } \mathbb{R}^n) = n + 1, \mathcal{X} = \mathbb{R}^n \leftarrow$ *prove this at home!*

Proposition 1

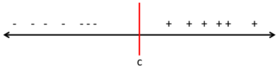
If $\mathcal{H}$ is finite, VC dimension is related to the cardinality of $\mathcal{H}$:

$$VC(\mathcal{H}) \leq log|\mathcal{H}|$$

Discussion on VC Dimension

More VC results of common $\mathcal{H}$:

- $VC(PositiveHalf\text{-}Lines) = 1, \mathcal{X} = \mathbb{R}$



- $VC(Intervals) = 2, \mathcal{X} = \mathbb{R}$
- $VC(LTF \ in \ \mathbb{R}^n) = n + 1, \mathcal{X} = \mathbb{R}^n$ ← *prove this at home!*

Proposition 1

If $\mathcal{H}$ is finite, VC dimension is related to the cardinality of $\mathcal{H}$:

$$VC(\mathcal{H}) \leq \log|\mathcal{H}|$$

Let $VC(H) = d$.
H has to shatter d points
2^d labelings. $|H| \geq 2^d$.

Proof. Let $d = VC|\mathcal{H}|$. There must exists a shattered set of size d on which H realizes all possible labelings. Every labeling must have a corresponding hypothesis, then $|\mathcal{H}| \geq 2^d$ □

Learning bound for infinite $\mathcal{H}$

Theorem 6

Given $\mathcal{H}$, let $d = VC(\mathcal{H})$.

replaced $k = |H|$

by $VC(H) \le \log|H|$

- *With probability at least $1 - \delta$, we have that for all h*

generation risk

VC(H)

$$|\epsilon(h) - \hat{\epsilon}(h)| \le O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

training error

sample size

Learning bound for infinite $\mathcal{H}$

Theorem 6

Given $\mathcal{H}$, let $d = VC(\mathcal{H})$.

- *With probability at least $1-\delta$, we have that for all h*

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

- *Thus, with probability at least $1-\delta$, we also have*

$h = \arg\min_{h \in \mathcal{H}} \mathcal{E}(h)$

$$\epsilon(\hat{h}) \leq \epsilon(h^*) + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

testing error ~ learned from ERM

variance

Learning bound for infinite $\mathcal{H}$

Corollary 7

For $|\epsilon(h) - \hat{\epsilon}(h)| \leq \gamma$ to hold for all $h \in \mathcal{H}$ with probability at least $1 - \delta$, it suffices that $m = O_{\gamma,\delta}(d)$.

Learning bound for infinite $\mathcal{H}$

Corollary 7

For $|\epsilon(h) - \hat{\epsilon}(h)| \leq \gamma$ to hold for all $h \in \mathcal{H}$ with probability at least $1-\delta$, it suffices that $m = O_{\gamma,\delta}(d)$.

Remarks

- Sample complexity using $\mathcal{H}$ is linear in $VC(\mathcal{H})$
- For "most"[a] hypothesis classes, the VC dimension is linear in terms of parameters
- For algorithms minimizing training error, # training examples needed is roughly linear in number of parameters in $\mathcal{H}$.

[a] Not always true for deep neural networks

VC Dimension of Deep Neural Networks

Theorem 8 (Cover, 1968; Baum and Haussler, 1989)

Let $\mathcal{N}$ be an arbitrary feedforward neural net with w weights that consists of linear threshold activations, then $VC(\mathcal{N}) = O(w \log w)$.

VC Dimension of Deep Neural Networks

Theorem 8 (Cover, 1968; Baum and Haussler, 1989)

Let $\mathcal{N}$ be an arbitrary feedforward neural net with w weights that consists of linear threshold activations, then $VC(\mathcal{N}) = O(w \log w)$.

Recent progress

- For feed-forward neural networks with piecewise-linear activation functions (e.g. ReLU), let w be the number of parameters and l be the number of layers, $VC(\mathcal{N}) = O(wl \log(w))$ [Bartlett et. al., 2017]

Bartlett and W. Maass (2003) Vapnik-Chervonenkis Dimension of Neural Nets
Bartlett et. al., (2017) Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks.

VC Dimension of Deep Neural Networks

Theorem 8 (Cover, 1968; Baum and Haussler, 1989)

Let $\mathcal{N}$ be an arbitrary feedforward neural net with w weights that consists of linear threshold activations, then $VC(\mathcal{N}) = O(w \log w)$.

Recent progress

- For feed-forward neural networks with piecewise-linear activation functions (e.g. ReLU), let w be the number of parameters and l be the number of layers, $VC(\mathcal{N}) = O(wl \log(w))$ [Bartlett et. al., 2017]
- *Among all networks with the same size (number of weights), more layers have larger VC dimension* , thus more training samples are needed to learn a deeper network

Bartlett and W. Maass (2003) Vapnik-Chervonenkis Dimension of Neural Nets
Bartlett et. al., (2017) Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks.

Final Project Information

See http://yangli-feasibility.com/home/classes/lfd2022fall/project.html