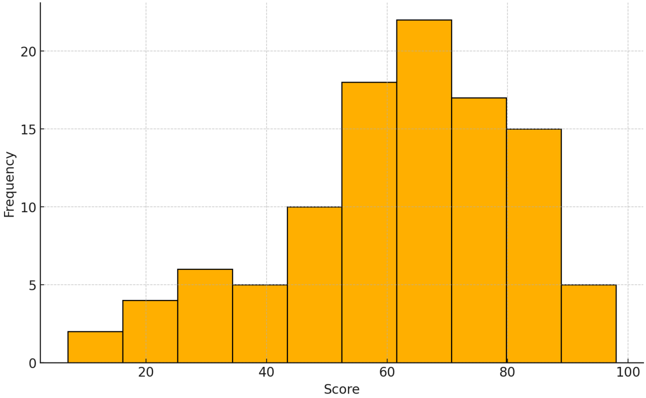
Learning From Data
Lecture 8: Learning Theory

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

TBSI

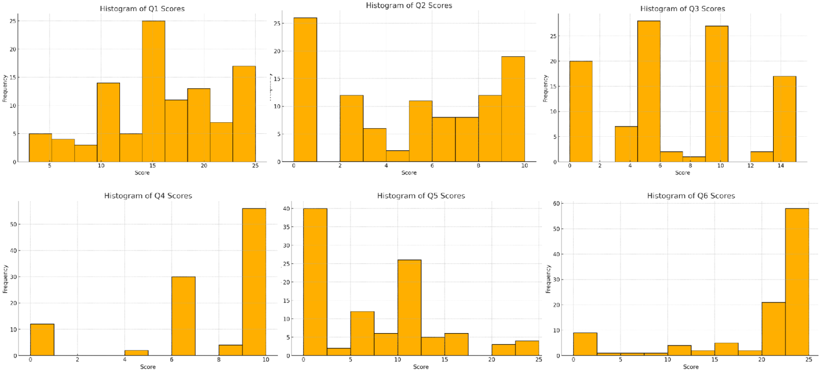
November 8, 2024

Midterm Results



	max	mean	median
curved score	98	62.13	66.5

Midterm Breakdown

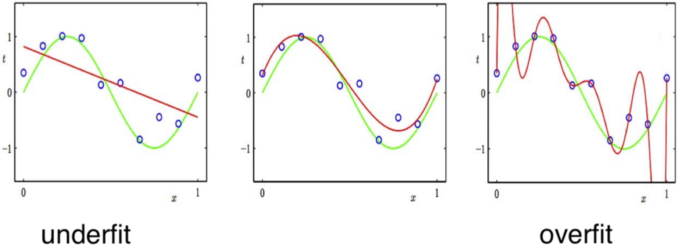


Review

Overfit & Underfit

Underfit Both training error and testing error are large

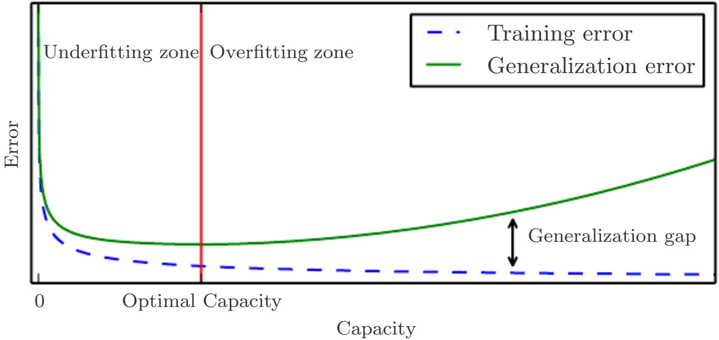
Overfit Training error is small, testing error is large



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:

$$y = h(x) + \epsilon$$

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 for test sample (x, y)

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

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

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 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?

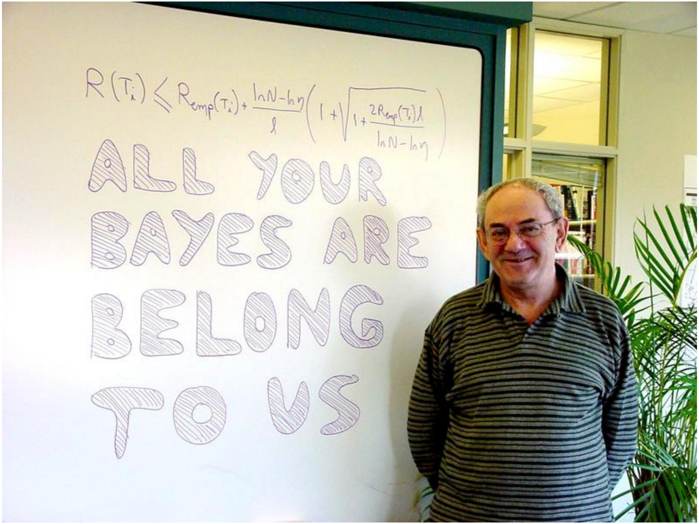


Figure: Prof. Vladimir Vapnik in front of his famous theorem

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

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

- The **generalization error** is

$$\epsilon(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\, 1\{h(x) \neq y\}$$

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

Hypothesis class

The **hypothesis class $\mathcal{H}$ used by a learning algorithm** is the set of all classsifiers considered by it.
e.g. Linear classification considers $h_\theta(x) = 1\{\theta^T x \geq 0\}$

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

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\epsilon}(h)$$

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 $\epsilon(h)$

- Show $\hat{\epsilon}(h)$ (training error) is a good estimate of $\epsilon(h)$ for all 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:

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

$$\hat{\epsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j = \hat{\mathbb{E}}[Z]$$

Testing error of $h_i \in \mathcal{H}$ is: $\epsilon(h_i) = \mathbb{E}[Z]$

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

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

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!

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

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,

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

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,

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

By Union bound,

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

Uniform Convergence Results

Corollary 3

Given γ and $\delta > 0$, If

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

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**.

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

- 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: $|\mathcal{H}| = 2^{64d}$

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 $1-\delta$?

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

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 $\epsilon(\hat{h}) \le \epsilon(h^*) + 2\gamma$ to hold with probability at least $1-\delta$?

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

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:

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

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

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

Infinite hypothesis class: Challenges

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

Example

$2n+2$ parameters:

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

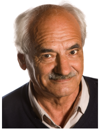
is equivalent the hypothesis with $n+1$ parameters:

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

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

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

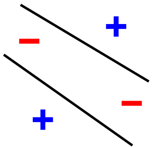


Figure: $\mathcal{H}_{LTF}$ can not shatter 4 points: for any 4 points, label points on the diagonal as '+'. (See Radon's theorem)

- 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

- Example: $VC(AxisAlignedRectangles) = 4$

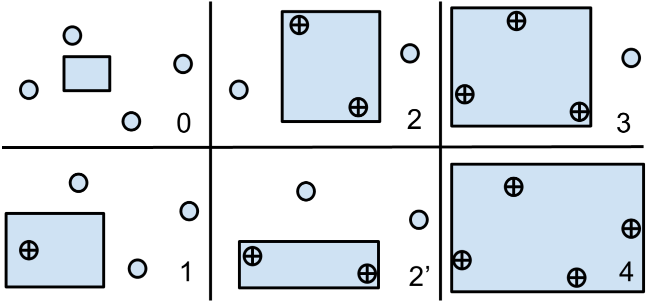


Figure: Axis-aligned rectangles can shatter 4 points.
$VC(AxisAlignedRectangles) \geq 4$

VC Dimension

- Example: $VC(AxisAlignedRectangles) = 4$

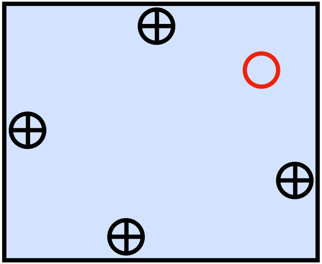
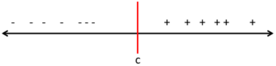


Figure: For any 5 points, label topmost, bottommost, leftmost and rightmost points as "+". $VC(AxisAlignedRectangles) < 5$

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

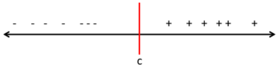
If $\mathcal{H}$ is finite, VC dimension is related to the cardinality of $\mathcal{H}$:

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

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 $\mathcal{H}$:

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

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 $\mathcal{H}$

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.

Summary

- We can control generalization by adjusting the complexity of hypothesis $\mathcal{H}$
- VC dimension as a useful measure of complexity.

We could bound the performance of a learning algorithm in terms of $VC(\mathcal{H})$ and the amount of data we have.

Limitation of VC Dimension

- The bound is not very tight as VC is distribution independent
- Only defined for binary classification

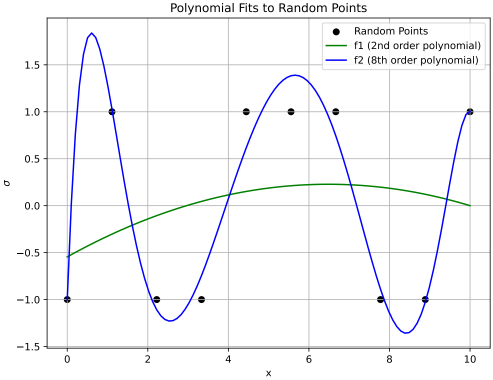
Rademacher Complexity

Rademacher Complexity

- Named after German-American Mathematician Hans Rademacher
- A more modern notion of complexity that is **distribution dependent** and defined for **any class of real-valued functions**.

Rademacher Complexity

- Named after German-American Mathematician Hans Rademacher
- A more modern notion of complexity that is **distribution dependent** and defined for **any class of real-valued functions**.
- **Rademacher Complexity (Informal)**: The ability of a hypothesis (function) class to fit random noise $\sigma_i \in \{+1, -1\}$. *Higher Rademacher complexity, greater capacity to overfit.*



Mathematical Definition

Empirical Rademacher Complexity

- Let $\mathcal{F}$ be a class of real-value functions $f : \mathcal{Z} \to \mathbb{R}$.
- Given a set of examples $S = \{z^{(1)}, z^{(2)}, \ldots, z^{(m)}\}$, each drawn from a fixed distribution D, the empirical Rademacher complexity of $\mathcal{F}$ is:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right],$$

where σ_i are i.i.d. Rademacher variables (taking values ± 1 with equal probability).

Mathematical Definition

Empirical Rademacher Complexity

- Let $\mathcal{F}$ be a class of real-value functions $f : \mathcal{Z} \to \mathbb{R}$.
- Given a set of examples $S = \{z^{(1)}, z^{(2)}, \ldots, z^{(m)}\}$, each drawn from a fixed distribution D, the empirical Rademacher complexity of $\mathcal{F}$ is:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right],$$

where σ_i are i.i.d. Rademacher variables (taking values ± 1 with equal probability).

Rademacher Complexity

- The Rademacher complexity of $\mathcal{F}$ over a distribution $\mathcal{D}$ is:

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{S \sim \mathcal{D}^m} \left[\hat{\mathfrak{R}}_S(\mathcal{F}) \right].$$

measures the expected noise-fitting-ability of $\mathcal{F}$ over all data sets S drawn according to D

Rademacher-based uniform convergence

For a function $f \in \mathcal{F}$ and a sample $S = \{z_1, \ldots, z_m\}$, the empirical expectation (sample mean) of f is:

$$\hat{\mathbb{E}}_S[f] = \frac{1}{m} \sum_{i=1}^{m} f(z_i).$$

Theorem 9

Let $\mathcal{F} \subseteq \{f : \mathcal{Z} \to [a, a+1]\}$ be any class of bounded real-value functions.
With probability at least $1 - \delta$ (for a confidence level $\delta \in (0,1)$), for any function $f \in \mathcal{F}$:

$$\mathbb{E}_{z \sim \mathcal{D}}[f(z)] \leq \hat{\mathbb{E}}_S[f] + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{m}}$$

We bound the expectation of each function in terms of its sample mean, the Rademacher complexity of the class, and an error term.

Connection to Loss Functions

Take binary classification as an example:

- Let $X = \mathbb{R}^d$, $Y = \{1, 1\}$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- Given a hypothesis class $\mathcal{H}$, we can define a class of loss functions $L(\mathcal{H}) = \{l_h : \mathcal{Z} \to \mathbb{R} | h \in \mathcal{H}\}$
- Example: The 0-1 loss can be written as $l_h(z) = l_h(x, y) = \mathbb{1}_{h(x) \neq y}$.
- The (empirical) expectation of the loss function are the (empirical) error of classifier h

$$\mathbb{E}_D[l_h(z)] = E_D[\mathbb{1}_{h(x)y}] = \epsilon(h)$$

$$\hat{\mathbb{E}}_S[l_h(z)] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h(x)y} = \hat{\epsilon}(h)$$

- By Theorem 9, we can show

$$\begin{aligned}\epsilon(h) \leq& \hat{\epsilon}(h) + 2R_m(L(\mathcal{H})) + \sqrt{\frac{\ln(1/\delta)}{m}}\\ =& \hat{\epsilon}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{m}}\end{aligned}$$

Summary

- Rademacher complexity $\mathcal{R}_m(\mathcal{H})$ depends on the underlying distribution D from which sample points are drawn.
- Uniform convergence of the generalization error can be derived using Rademacher complexity for any bounded loss function

Final Project Information

See http://yangli-feasibility.com/home/classes/lfd2024fall/project.html

- Project Timeline

Deadline	Task
11-Nov	Submit group assignment
22-Nov	Submit project proposal
6-Dec	Team meeting with course staff
25-Dec	Submit poster PDF file (Submission will be closed at 11:59am)
27-Dec	Poster session
3-Jan	Submit final report