CS 4774 Machine Learning

The Bias-Complexity Tradeoff

Yangfeng Ji

Information and Language Processing Lab
Department of Computer Science
University of Virginia

UVA ILP
UNIVERSITY OF VIRGINIA
INFORMATION AND LANGUAGE PROCESSING LAB

Overview

1. The Bias-Complexity Tradeoff
2. The Bias-Variance Tradeoff
3. The VC Dimension

Readings: [Shalev-Shwartz and Ben-David, 2014, Chapter 5 & 6]

Question

For a real-world machine learning problem, which of the following items are usually available to us?

Question

For a real-world machine learning problem, which of the following items are usually available to us?

- Training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$
- Domain set $\mathcal{X}$
- Label set $\mathcal{Y}$

Question

For a real-world machine learning problem, which of the following items are usually available to us?

- Training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$
- Domain set $\mathcal{X}$
- Label set $\mathcal{Y}$
- Labeling function (the oracle) f
- Distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- The Bayes predictor $f_{\mathcal{D}}(\boldsymbol{x})$

Question

For a real-world machine learning problem, which of the following items are usually available to us?

- Training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$
- Domain set $\mathcal{X}$
- Label set $\mathcal{Y}$
- Labeling function (the oracle) f
- Distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- The Bayes predictor $f_{\mathcal{D}}(\boldsymbol{x})$
- The size of the hypothesis space $\mathcal{H}$
- The empirical risk of a hypothesis $h(\boldsymbol{x}) \in \mathcal{H}$, $L_S(h(\boldsymbol{x}))$
- The true risk of a hypothesis $h(\boldsymbol{x}) \in \mathcal{H}$, $L_{\mathcal{D}}(h(\boldsymbol{x}))$

A 2-dimensional Classification Problem

Consider the following four situations

	Given Data Distribution	Given Training Examples
Nonlinear classifier		
Linear classifier		

Agnostic PAC Learnability

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:

- for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$ and
- for every $\epsilon, \delta \in (0, 1)$,

when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h_S[1] such that, with probability of at least $1 - \delta$,

$$L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon \tag{1}$$

[1]Sometimes, as $h_S(\boldsymbol{x})$ or $h(\boldsymbol{x}, S)$

Agnostic PAC Learnability

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:

- for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$ and
- for every $\epsilon, \delta \in (0, 1)$,

when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h_S[1] such that, with probability of at least $1 - \delta$,

$$L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon \tag{1}$$

This explains the relation between *the hypothesis learned with limited data* (h_S) and *the best hypothesis in the space* ($\operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$).

[1] Sometimes, as $h_S(\boldsymbol{x})$ or $h(\boldsymbol{x}, S)$

The Bayes Optimal Predictor

- The Bayes optimal predictor: given a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$, the predictor is defined as

$$f_{\mathcal{D}}(x) = \begin{cases} +1 & \text{if } \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

- No other predictor can do better: for any predictor h

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(h) \tag{3}$$

The Bayes Optimal Predictor

- The Bayes optimal predictor: given a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$, the predictor is defined as

$$f_{\mathcal{D}}(x) = \begin{cases} +1 & \text{if } \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

- No other predictor can do better: for any predictor h

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(h) \tag{3}$$

- Question: for a given hypothesis space $\mathcal{H}$, does the following relation hold?

$$f_{\mathcal{D}} \in \underset{h' \in \mathcal{H}}{\operatorname{argmin}} \, L_{\mathcal{D}}(h')$$

The Bayes Optimal Predictor

- The Bayes optimal predictor: given a probability distribution $\mathscr{D}$ over $\mathcal{X} \times \{-1, +1\}$, the predictor is defined as

$$f_{\mathscr{D}}(x) = \begin{cases} +1 & \text{if } \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

- No other predictor can do better: for any predictor h

$$L_{\mathscr{D}}(f_{\mathscr{D}}) \leq L_{\mathscr{D}}(h) \tag{3}$$

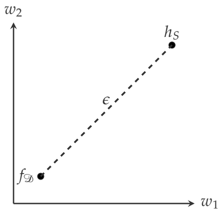
- Question: for a given hypothesis space $\mathcal{H}$, does the following relation hold?

$$f_{\mathscr{D}} \in \operatorname*{argmin}_{h' \in \mathcal{H}} L_{\mathscr{D}}(h')$$

  - Answer: it depends the selection of the hypothesis space $\mathcal{H}$, usually not.
  - Example: if $f_{\mathscr{D}}$ is a nonlinear classifier, while we choose to use logistic regression.

The Gap between h_S and $f_{\mathcal{D}}$

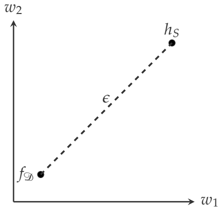
For illustration purpose, let us assume the gap between h_S and $f_{\mathcal{D}}$ can be visualized in the following plot



- $h_S = \text{argmin}_{h' \in \mathcal{H}} L_S(h')$: learned by minimizing the empirical risk
  - Constrained by the selection of $\mathcal{H}$
- $f_{\mathcal{D}}$: the optimal predictor if we know the data distribution $\mathcal{D}$
  - Not constrained by the selection of $\mathcal{H}$

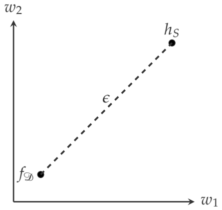
Outline

The previous example implies the error gap between h_S and $f_{\mathscr{D}}$ can be decomposed into two components



Outline

The previous example implies the error gap between h_S and $f_{\mathcal{D}}$ can be decomposed into two components



Two different perspectives of the decomposition

- The bias-complexity tradeoff: from the perspective of learning theory
- The bias-variance tradeoff: from the perspective of statistical estimation

The Bias-Complexity Tradeoff

Basic Learning Procedure

The basic component of formulating a learning process

- Input/output space $\mathcal{X} \times \mathcal{Y}$
- A collection of training examples $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$
- Hypothesis space $\mathcal{H}$
- Learning via empirical risk minimization

$$h_S \in \operatorname*{argmin}_{h' \in \mathcal{H}} L_S(h') = \frac{1}{m} |\{h'(\boldsymbol{x}_i) \neq y_i\}| \tag{4}$$

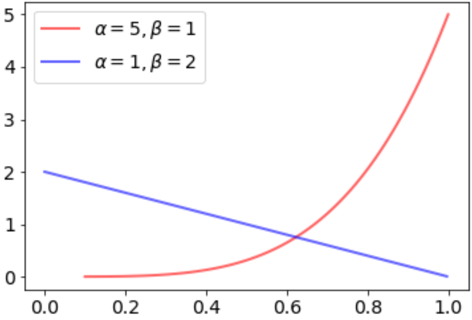
- Analyzing the true error of h_S

$$L_{\mathcal{D}}(h_S) = \mathbb{E}[h_S(x) \neq f(x)] \tag{5}$$

Example

Consider the binary classification problem with the data sampled from the following distribution

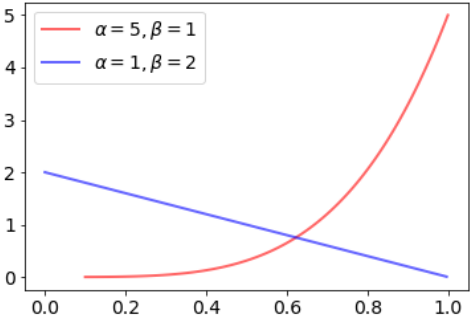
$$\mathscr{D} = \frac{1}{2}\mathscr{B}(x;5,1) + \frac{1}{2}\mathscr{B}(x;1,2) \quad (6)$$



Example (Cont.)

Given the distribution, we can compute the true risk/error of the Bayes predictor $f_{\mathscr{D}}$ as

$$\begin{aligned} L_{\mathscr{D}}(f_{\mathscr{D}}) &= \frac{1}{2}\mathscr{B}(x < b_{\text{Bayes}}; 5, 1) + \frac{1}{2}(1 - \mathscr{B}(x < b_{\text{Bayes}}; 1, 2)) \\ &= 0.11799 \end{aligned} \tag{7}$$



Example (Cont.)

The hypothesis space $\mathcal{H}$ is defined as

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{8}$$

where $N \in \mathbb{N}$ is a predefined integer

Example (Cont.)

The hypothesis space $\mathcal{H}$ is defined as

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{8}$$

where $N \in \mathbb{N}$ is a predefined integer

- The value of N is the size of the hypothesis space

Example (Cont.)

The hypothesis space $\mathcal{H}$ is defined as

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{8}$$

where $N \in \mathbb{N}$ is a predefined integer

- The value of N is the size of the hypothesis space
- The best hypothesis in $\mathcal{H}$

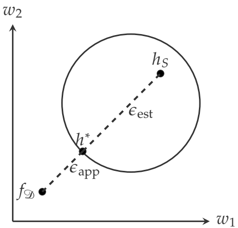
$$h^* \in \operatorname*{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') \tag{9}$$

- Very likely the best predictor in $\mathcal{H}$ is not the Bayes predictor, unless $b_{\text{Bayes}} \in \{\frac{i}{N} : i \in [N]\}$

Error Decomposition

The error gap between h_S and $f_{\mathscr{D}}$ can be decomposed as two parts

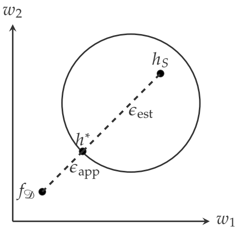
$$L_{\mathscr{D}}(h_S) - L_{\mathscr{D}}(f_{\mathscr{D}}) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \tag{10}$$



Error Decomposition

The error gap between h_S and $f_{\mathcal{D}}$ can be decomposed as two parts

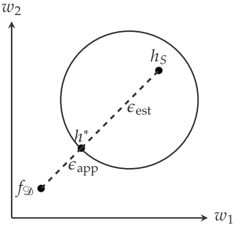
$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f_{\mathcal{D}}) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \tag{10}$$



- Approximation error ϵ_{app} caused by selecting a specific hypothesis space $\mathcal{H}$ (model bias)
- Estimation error ϵ_{est} caused by selecting h_S with a specific training set (model complexity)

Approximation Error ϵ_{app}

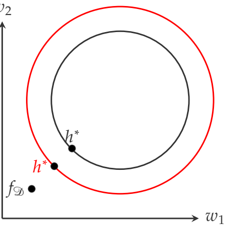
To reduce the approximation error ϵ_{app}, we could increase the size of the hypothesis space



The cost is that we also increase the size of training set, in order to maintain the overall error in the same level (recall the sample complexity of finite hypothesis spaces).

Approximation Error ϵ_{app}

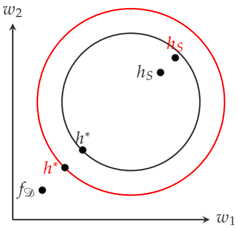
To reduce the approximation error ϵ_{app}, we could increase the size of the hypothesis space



The cost is that we also increase the size of training set, in order to maintain the overall error in the same level (recall the sample complexity of finite hypothesis spaces).

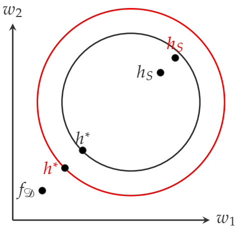
Estimation Error ϵ_{est}

On the other hand, if we use the same training set S, then we *may* have a larger estimation error



Estimation Error ϵ_{est}

On the other hand, if we use the same training set S, then we *may* have a larger estimation error

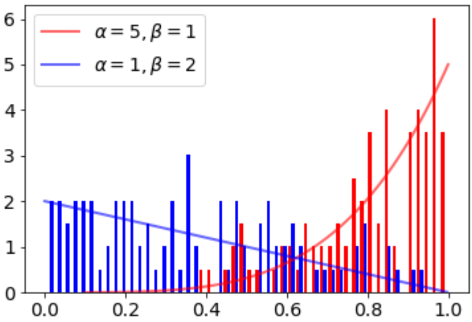


The bias-complexity tradeoff: find the right balance to reduce both approximation error and estimation error.

Example: 200 training examples

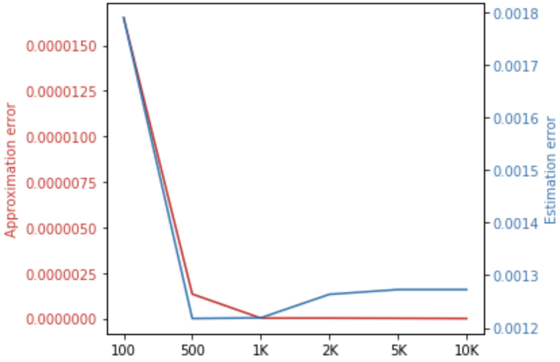
We randomly sampled 100 examples from each class

$$\mathcal{D} = \frac{1}{2}\mathcal{B}(x; 5, 1) + \frac{1}{2}\mathcal{B}(x; 1, 2) \tag{11}$$



Example: 200 training examples

Given 200 training examples, the errors with respect to different hypothesis space is the following (x axis is the size of $\mathscr{H}$)

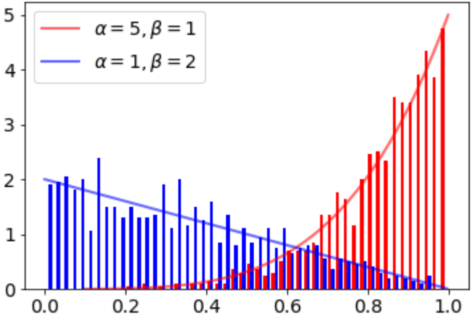


There is a tradeoff with respect to the size of $\mathscr{H}$

Example: 2000 training examples

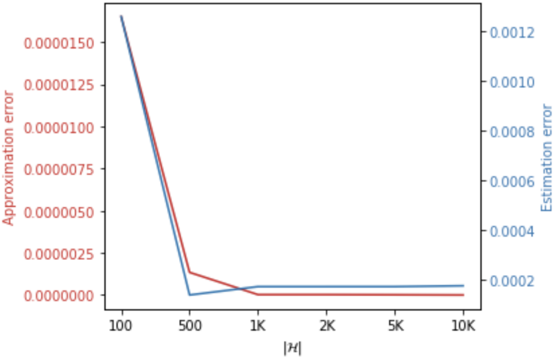
We randomly sampled 1000 examples from each class

$$\mathcal{D} = \frac{1}{2}\mathcal{B}(x; 5, 1) + \frac{1}{2}\mathcal{B}(x; 1, 2) \tag{12}$$



Example: 2000 training examples

With these 2000 training examples, the errors with respect to different hypothesis space is the following



Both errors are smaller, but the tradeoff still exists

Summary

Three components in this decomposition

- $h_S \in \operatorname{argmin}_{h' \in \mathcal{H}} L_S(h')$: the ERM predictor given the training set S
- $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$: the optimal predictor from $\mathcal{H}$
- $f_{\mathcal{D}}$: the Bayes predictor given $\mathcal{D}$

Summary

Three components in this decomposition

- $h_S \in \text{argmin}_{h' \in \mathcal{H}} L_S(h')$: the ERM predictor given the training set S
- $h^* \in \text{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$: the optimal predictor from $\mathcal{H}$
- $f_{\mathcal{D}}$: the Bayes predictor given $\mathcal{D}$

Balancing strategy:

- we can incrase the complexity of hypothesis space to reduce the bias, e.g.,
  - enlarge the hypothesis space (as in the running example)
  - replacing linear predictors with nonlinear predictors

Summary

Three components in this decomposition

- $h_S \in \operatorname{argmin}_{h' \in \mathcal{H}} L_S(h')$: the ERM predictor given the training set S
- $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$: the optimal predictor from $\mathcal{H}$
- $f_{\mathcal{D}}$: the Bayes predictor given $\mathcal{D}$

Balancing strategy:

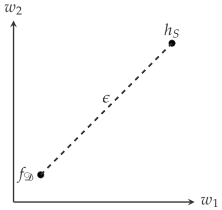
- we can incrase the complexity of hypothesis space to reduce the bias, e.g.,
  - enlarge the hypothesis space (as in the running example)
  - replacing linear predictors with nonlinear predictors
- in the meantime, we have to increase the training size to reduce the approximation error.

The Bias-Variance Tradeoff

A New Perspective

Let us analyze the error ϵ without the assumption of

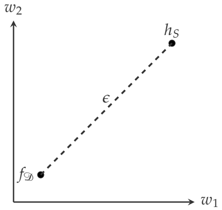
- knowing the best predictor from $\mathcal{H}$, $h^* \in \text{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$
- changing the size of S



A New Perspective

Let us analyze the error ϵ without the assumption of

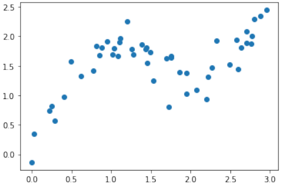
- knowing the best predictor from $\mathcal{H}$, $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$
- changing the size of S



We still need (1) the ERM predictor h_S and (2) the Bayes predictor $f_{\mathcal{D}}$

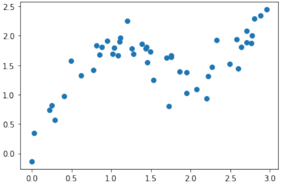
A New Way of Decomposition

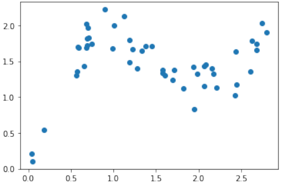
- Consider the randomness in S with m training examples



A New Way of Decomposition

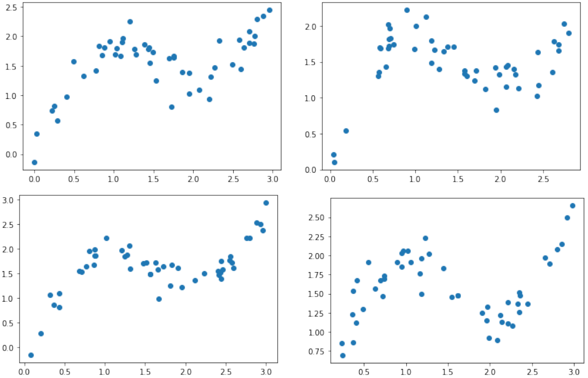
- Consider the randomness in S with m training examples





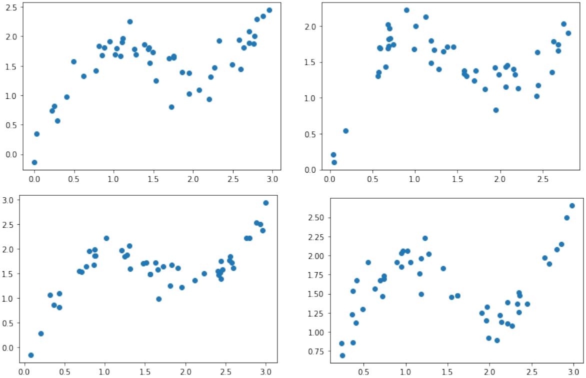
A New Way of Decomposition

- Consider the randomness in S with m training examples



A New Way of Decomposition

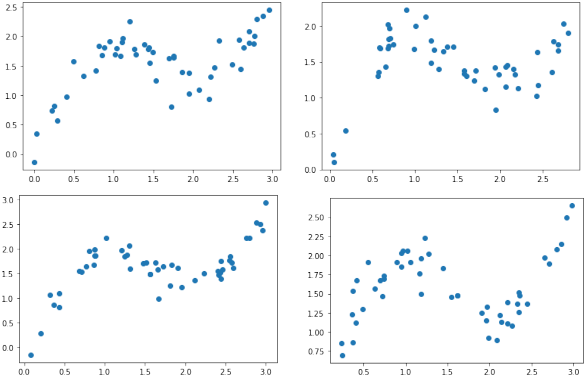
- Consider the randomness in S with m training examples



- In this case, S is a random variable, $h(\boldsymbol{x}, S)$ is a function of S and $\boldsymbol{x}$

A New Way of Decomposition

- Consider the randomness in S with m training examples



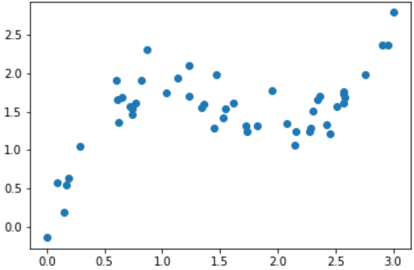
- In this case, S is a random variable, $h(\boldsymbol{x}, S)$ is a function of S and $\boldsymbol{x}$
- The average prediction function given by $E\left[h(\boldsymbol{x}, S)\right]$ where $S \sim \mathscr{D}^m$
  - Overall, $E\left[h(\boldsymbol{x}, S)\right]$ will give good performance on any possible dataset with size m

Data Generation Model

Consider the following *data generation model*

- $X \sim U[0,1]$ uniform distribution
- $Y = \mathcal{N}(X + \sin(2X), \sigma^2)$ with $\sigma^2 = 0.1$

An example of S is

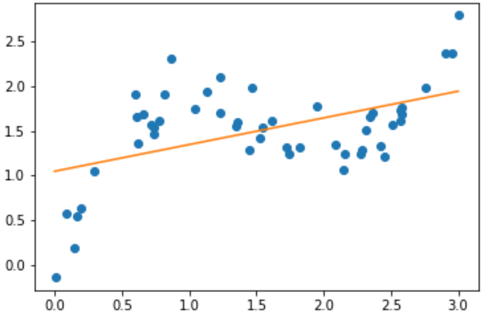


Hypothesis Spaces

Given S and the following hypothesis space $\mathcal{H}_1$

$$\mathcal{H}_1 = \{w_0 + w_1 x : w_0, w_1 \in \mathbb{R}\} \tag{13}$$

the regression result

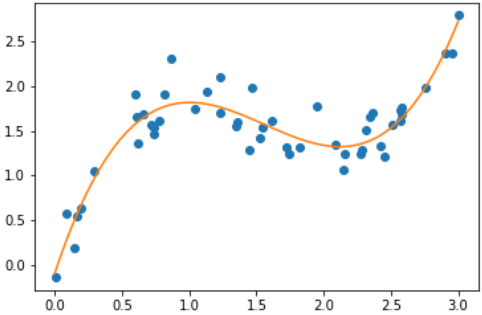


Hypothesis Spaces (Cont.)

Given S and the following hypothesis space $\mathcal{H}_3$

$$\mathcal{H}_3 = \{w_0 + w_1 x + w_2 x^2 + w_3 x^3 : w_0, w_1, w_2, w_3 \in \mathbb{R}\} \tag{14}$$

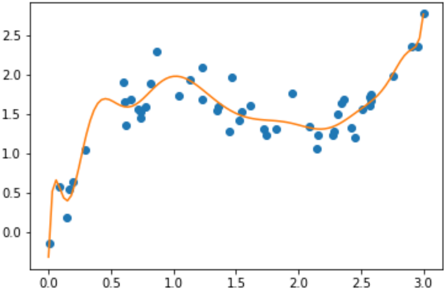
the regression result



Hypothesis Spaces (Cont.)

Given S and the following hypothesis space $\mathscr{H}_{15}$

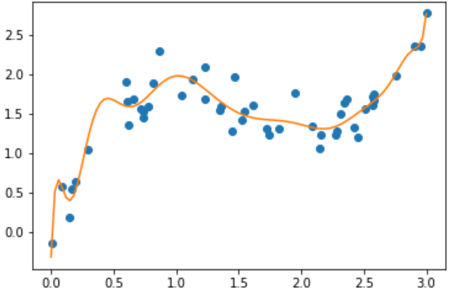
$$\mathscr{H}_{15} = \{w_0 + w_1 x + \cdots + w_{15} x^{15} : w_0, w_1, \cdots, w_{15} \in \mathbb{R}\} \qquad (15)$$



Hypothesis Spaces (Cont.)

Given S and the following hypothesis space $\mathscr{H}_{15}$

$$\mathscr{H}_{15} = \{w_0 + w_1 x + \cdots + w_{15} x^{15} : w_0, w_1, \cdots, w_{15} \in \mathbb{R}\} \qquad (15)$$



- Intuitively, the degree of the polynomials indicates the potential/complexity of the hypothesis space
- Refer to the VC dimension section for more discussion

Error Decomposition

The difference between the best hypothesis $h(\boldsymbol{x}, S)$ and the Bayes predictor $f_{\mathscr{D}}(\boldsymbol{x})$ is measured as

$$\epsilon^2 = \{h(\boldsymbol{x}, S) - f_{\mathscr{D}}(\boldsymbol{x})\}^2 \tag{16}$$

Introduce $E[h(\boldsymbol{x}, S)]$ into the calculation, we have

$$\epsilon^2 = \{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)] + E[h(\boldsymbol{x}, S)] - f_{\mathscr{D}}(\boldsymbol{x})\}^2$$

Error Decomposition

The difference between the best hypothesis $h(\boldsymbol{x}, S)$ and the Bayes predictor $f_{\mathscr{D}}(\boldsymbol{x})$ is measured as

$$\epsilon^2 = \{h(\boldsymbol{x}, S) - f_{\mathscr{D}}(\boldsymbol{x})\}^2 \tag{16}$$

Introduce $E\left[h(\boldsymbol{x}, S)\right]$ into the calculation, we have

$$\begin{aligned}
\epsilon^2 &= \{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right] + E\left[h(\boldsymbol{x}, S)\right] - f_{\mathscr{D}}(\boldsymbol{x})\}^2 \\
&= \{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2 + \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathscr{D}}(\boldsymbol{x})\}^2 \\
&\quad +2\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\} \cdot \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathscr{D}}(\boldsymbol{x})\}
\end{aligned}$$

Review: Mean

Given a random variable X and its probability density function $p(x)$

- Mean: $E[X] = \int xp(x)dx$
- Example: the mean of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$E[X] = \mu \tag{17}$$

Review: Mean

Given a random variable X and its probability density function $p(x)$

- Mean: $E[X] = \int xp(x)dx$
- Example: the mean of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$E[X] = \mu \tag{17}$$

- Approximation to the mean with samples $\{x_1, \ldots, x_m\}$

$$E[X] \approx \frac{1}{m} \sum_{i=1}^{m} x_i \tag{18}$$

Review: Mean

Given a random variable X and its probability density function $p(x)$

- Mean: $E[X] = \int xp(x)dx$
- Example: the mean of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$E[X] = \mu \tag{17}$$

- Approximation to the mean with samples $\{x_1, \ldots, x_m\}$

$$E[X] \approx \frac{1}{m} \sum_{i=1}^{m} x_i \tag{18}$$

- Property: $E[\alpha X] = \alpha E[X]$ for α is determinstic

Review: Variance

Given a random variable X, its probability density function $p(x)$, and its mean $E[X]$

- Variance: $\text{Var}(X) = E\left[(X - E[X])^2\right]$
- Example: the variance of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$\text{Var}(X) = \sigma^2 \tag{19}$$

Review: Variance

Given a random variable X, its probability density function $p(x)$, and its mean $E[X]$

- Variance: $\text{Var}(X) = E\left[(X - E[X])^2\right]$
- Example: the variance of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$\text{Var}(X) = \sigma^2 \tag{19}$$

- Relation between $\text{Var}(X)$ and $E[X]$

$$\begin{aligned} \text{Var}(X) &= E\left[(X - E[X])^2\right] \\ &= E\left[X^2 - 2XE[X] + E[X]^2\right] \\ &= E\left[X^2\right] - 2E[X]E[X] + E[X]^2 \\ &= E\left[X^2\right] - E[X]^2 \end{aligned}$$

Error Decomposition (Cont.)

Recall

$$\begin{aligned}\epsilon^2 &= \{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right] + E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\ &= \{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2 + \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\ &\quad +2\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\} \cdot \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}\end{aligned}$$

Error Decomposition (Cont.)

Recall

$$\begin{aligned}\epsilon^2 &= \{h(\boldsymbol{x},S) - E[h(\boldsymbol{x},S)] + E[h(\boldsymbol{x},S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\ &= \{h(\boldsymbol{x},S) - E[h(\boldsymbol{x},S)]\}^2 + \{E[h(\boldsymbol{x},S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\ &\quad +2\{h(\boldsymbol{x},S) - E[h(\boldsymbol{x},S)]\} \cdot \{E[h(\boldsymbol{x},S)] - f_{\mathcal{D}}(\boldsymbol{x})\}\end{aligned}$$

Taking the expectation of ϵ^2

$$\begin{aligned}E\left[\epsilon^2\right] &= E\left[\{h(\boldsymbol{x},S) - E[h(\boldsymbol{x},S)]\}^2\right] + \{E[h(\boldsymbol{x},S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\ &\quad +2E[\{h(\boldsymbol{x},S) - E[h(\boldsymbol{x},S)]\}] \cdot \{E[h(\boldsymbol{x},S)] - f_{\mathcal{D}}(\boldsymbol{x})\}\end{aligned}$$

Error Decomposition (Cont.)

Recall

$$\begin{aligned}
\epsilon^2 &= \{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)] + E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&= \{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2 + \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&\quad +2\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\} \cdot \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}
\end{aligned}$$

Taking the expectation of ϵ^2

$$\begin{aligned}
E\left[\epsilon^2\right] &= E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right] + \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&\quad +2E[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}] \cdot \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\} \\
&= E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right] + \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&\quad +2\{E[h(\boldsymbol{x}, S)] - E[h(\boldsymbol{x}, S)]\} \cdot \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}
\end{aligned}$$

Error Decomposition (Cont.)

Recall

$$\begin{aligned}
\epsilon^2 &= \{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right] + E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&= \{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2 + \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&\quad +2\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\} \cdot \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}
\end{aligned}$$

Taking the expectation of ϵ^2

$$\begin{aligned}
E\left[\epsilon^2\right] &= E\left[\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2\right] + \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&\quad +2E\left[\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}\right] \cdot \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\} \\
&= E\left[\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2\right] + \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&\quad +2\{E\left[h(\boldsymbol{x}, S)\right] - E\left[h(\boldsymbol{x}, S)\right]\} \cdot \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\} \\
&= E\left[\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2\right] + \{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathcal{D}}(\boldsymbol{x})\}^2
\end{aligned}$$

The Bias-Variance Decomposition

The expected error is decomposed as

$$E\left[\epsilon^2\right] = \underbrace{E\left[\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2\right]}_{\text{variance}} + \underbrace{\{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathscr{D}}(\boldsymbol{x})\}^2}_{\text{bias}^2}$$

The Bias-Variance Decomposition

The expected error is decomposed as

$$E\left[\epsilon^2\right] = \underbrace{E\left[\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2\right]}_{\text{variance}} + \underbrace{\{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathscr{D}}(\boldsymbol{x})\}^2}_{\text{bias}^2}$$

- **bias**: how far the expected prediction $E\left[h(\boldsymbol{x}, S)\right]$ diverges from the optimal predictor $f_{\mathscr{D}}(\boldsymbol{x})$

The Bias-Variance Decomposition

The expected error is decomposed as

$$E\left[\epsilon^2\right] = \underbrace{E\left[\{h(\boldsymbol{x}, S) - E\left[h(\boldsymbol{x}, S)\right]\}^2\right]}_{\text{variance}} + \underbrace{\{E\left[h(\boldsymbol{x}, S)\right] - f_{\mathscr{D}}(\boldsymbol{x})\}^2}_{\text{bias}^2}$$

- **bias**: how far the expected prediction $E\left[h(\boldsymbol{x}, S)\right]$ diverges from the optimal predictor $f_{\mathscr{D}}(\boldsymbol{x})$
- **variance**: how a hypothesis learned from a specific S diverges from the average prediction $E\left[h(\boldsymbol{x}, S)\right]$

Computing $E\left[h(\boldsymbol{x}, S)\right]$

The key of computing $E\left[h(\boldsymbol{x}, S)\right]$ is to eliminate the randomness introduced by S

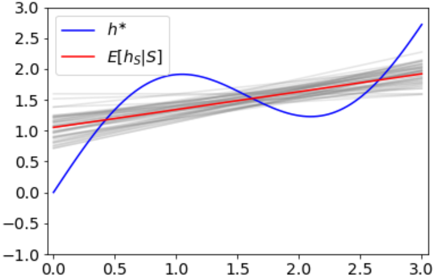
for $k = 1, \cdots, K$ **do**
    Sample a traing set S_k with size m from the data generation model
    Find the best hypothesis via $h(\boldsymbol{x}, S_k) \in \operatorname{argmin}_{h'} L(h', S_k)$
end for
Output:

$$E\left[h(\boldsymbol{x}, S)\right] \approx \frac{1}{K}\sum_{k=1}^{K} h(\boldsymbol{x}, S_k)$$

The larger K, the better approximation

Example: Bias and Variance

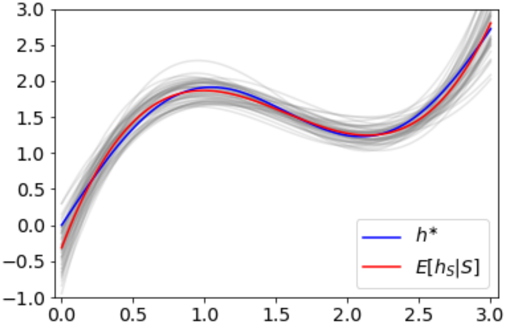
With $K = 50$, $m = 100$, and $\mathscr{H}_1$, we can visualize the bias and variance of a linear regression example as following



High bias and low variance (Underfitting)

Example: Bias and Variance (Cont.)

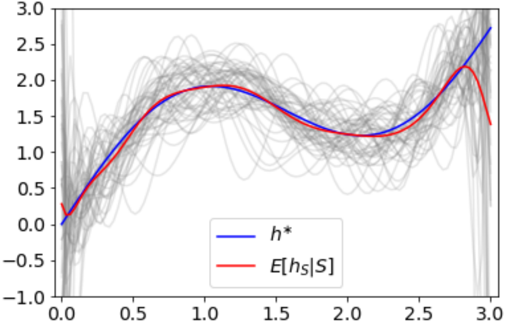
Same training set with $\mathscr{H}_3$



Both bias and variance are fine

Example: Bias and Variance (Cont.)

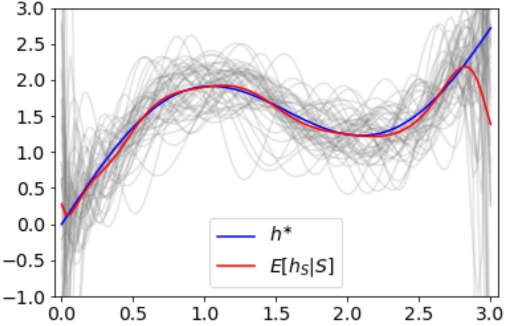
Same training set with $\mathscr{H}_{15}$



Low bias and high variance (Overfitting)

Example: Bias and Variance (Cont.)

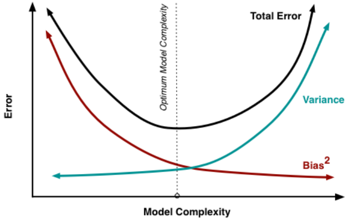
Same training set with $\mathscr{H}_{15}$



Low bias and high variance (Overfitting)

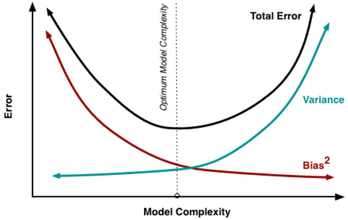
Exercise: The bias-variance tradeoff on linear regression with ℓ_2 regularization

The Bias-Variance Tradeoff



- **bias**: how far the expected prediction $E[h(\boldsymbol{x}, S)]$ diverges from the optimal predictor $f_{\mathcal{D}}(\boldsymbol{x})$
  - Error of this part is caused by *the selection of a hypothesis space*

The Bias-Variance Tradeoff



- **bias**: how far the expected prediction $E[h(\boldsymbol{x}, S)]$ diverges from the optimal predictor $f_{\mathcal{D}}(\boldsymbol{x})$
  - Error of this part is caused by *the selection of a hypothesis space*
- **variance**: how a hypothesis learned from a specific S diverges from the average prediction $E[h(\boldsymbol{x}, S)]$
  - Error of this part is caused by *using a particular data set S*

The VC Dimension

Learnability with Infinite Hypotheses

Infinite-size hypothesis space is learnable

Examples

- Half-space predictor
- Logistic regression predictor
- *Many others*

Shattering

For a given set S and a hypothesis space $\mathcal{H}$,

- A dichotomy of the set S is one of the possible ways of labeling the points in S using a hypothesis $h \in \mathcal{H}$

[Mohri et al., 2018, Page 36]

Shattering

For a given set S and a hypothesis space $\mathcal{H}$,

- A dichotomy of the set S is one of the possible ways of labeling the points in S using a hypothesis $h \in \mathcal{H}$
- A set S of $m \geq 1$ points is said to be shattered by a hypothesis space $\mathcal{H}$, if *all* possible dichotomies of S can be realized by $\mathcal{H}$

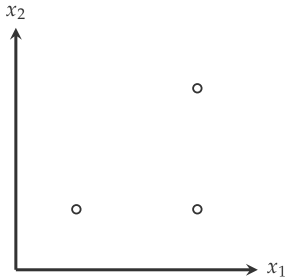
[Mohri et al., 2018, Page 36]

Shattering: Example

Consider the following set S and the half-space hypothesis space

$$\mathscr{H}_{\text{half}} = \{w_0 + w_1 x_1 + w_2 x_2 = 0 : w_0, w_1, w_2 \in \mathbb{R}\} \tag{20}$$

and the following specific set S



There are $2^3 = 8$ different ways to label the points and $\mathscr{H}_{\text{half}}$ can realized all of them.

VC Dimension

The **VC-dimension** of a hypothesis space $\mathcal{H}$, denoted $\text{VCdim}(\mathcal{H})$, is the maximal size of a set $S \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

[Shalev-Shwartz and Ben-David, 2014, Page 70]

VC Dimension

The **VC-dimension** of a hypothesis space $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, is the maximal size of a set $S \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

A: How to find the VC-dimension of a given hypothesis space?
Q: The proof consists of two parts:

- There exists a set S of size d that is shattered by $\mathcal{H}$
- Every set S of size $d+1$ is not shattered by $\mathcal{H}$

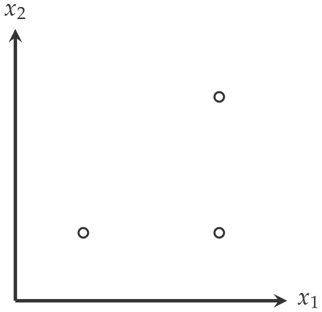
[Shalev-Shwartz and Ben-David, 2014, Page 70]

Half Spaces

Consider a special case as following, where VC-dim($\mathscr{H}_{\text{half}}$) = 3

$$\mathscr{H}_{\text{half}} = \{w_0 + w_1 x_1 + w_2 x_2 = 0 : w_0, w_1, w_2 \in \mathbb{R}\} \tag{21}$$

(1) Exist a case

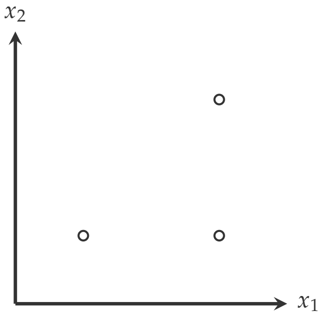


Half Spaces

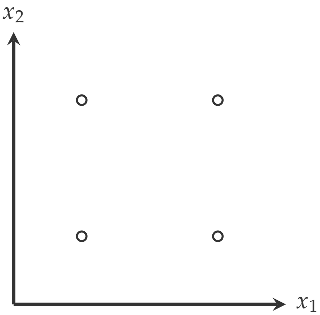
Consider a special case as following, where VC-dim($\mathscr{H}_{\text{half}}$) = 3

$$\mathscr{H}_{\text{half}} = \{w_0 + w_1 x_1 + w_2 x_2 = 0 : w_0, w_1, w_2 \in \mathbb{R}\} \tag{21}$$

(1) Exist a case



(2) For any case



Axis-aligned Rectangles

Let $\mathscr{H}$ be the class of axis-aligned rectangle, formally

$$\mathscr{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\} \tag{22}$$

where

$$h_{(a_1,a_2,b_1,b_2)}(x_1,x_2) = \begin{cases} +1 & x_1 \in [a_1,a_2] \text{and } x_2 \in [b_1,b_2] \\ -1 & \text{otherwise} \end{cases}$$

Axis-aligned Rectangles

Let $\mathscr{H}$ be the class of axis-aligned rectangle, formally

$$\mathscr{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\} \tag{22}$$

where

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} +1 & x_1 \in [a_1, a_2] \text{and } x_2 \in [b_1, b_2] \\ -1 & \text{otherwise} \end{cases}$$

Exist a case

Axis-aligned Rectangles

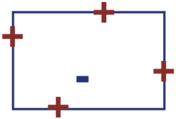
Let $\mathscr{H}$ be the class of axis-aligned rectangle, formally

$$\mathscr{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\} \tag{22}$$

where

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} +1 & x_1 \in [a_1, a_2] \text{and } x_2 \in [b_1, b_2] \\ -1 & \text{otherwise} \end{cases}$$

For any case



Axis-aligned Rectangles

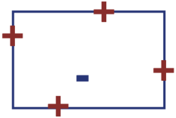
Let $\mathscr{H}$ be the class of axis-aligned rectangle, formally

$$\mathscr{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\} \tag{22}$$

where

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} +1 & x_1 \in [a_1, a_2] \text{and } x_2 \in [b_1, b_2] \\ -1 & \text{otherwise} \end{cases}$$

For any case

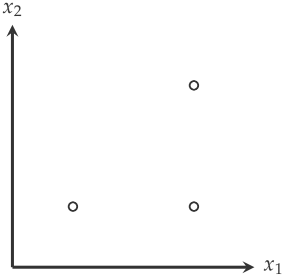


$$\text{VC-dim}(\mathscr{H}_{\text{rect}}) = 4$$

VC Dimension and the Number of Parameters

- For linear predictors, the VC dimensions are equal to the numbers of parameters

$$\mathcal{H}_{\text{half}} = \{w_0 + w_1 x_1 + w_2 x_2 = 0 : w_0, w_1, w_2 \in \mathbb{R}\} \tag{23}$$

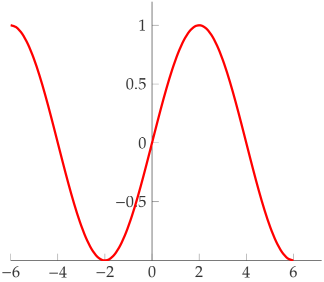


- However, the number of parameters is not always a good indictor for the VC dimension. Considering the following hypothesis space

Sine Functions

The hypothesis space of sine functions is defined as

$$\mathcal{H}_{\sin} = \{\sin(\alpha \cdot x) : \alpha \in \mathbb{R}\} \tag{24}$$

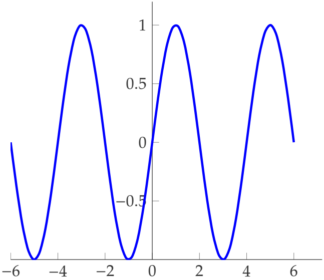


- $\alpha = \frac{\pi}{4}$
- $\alpha = \frac{\pi}{2}$
- $\alpha = \pi$

Sine Functions

The hypothesis space of sine functions is defined as

$$\mathcal{H}_{\sin} = \{\sin(\alpha \cdot x) : \alpha \in \mathbb{R}\} \qquad (24)$$

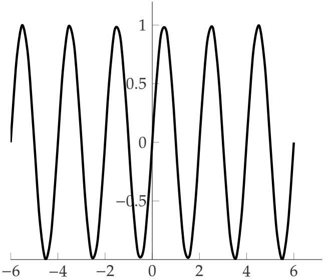


- $\alpha = \frac{\pi}{4}$
- $\alpha = \frac{\pi}{2}$
- $\alpha = \pi$

Sine Functions

The hypothesis space of sine functions is defined as

$$\mathscr{H}_{\sin} = \{\sin(\alpha \cdot x) : \alpha \in \mathbb{R}\} \tag{24}$$

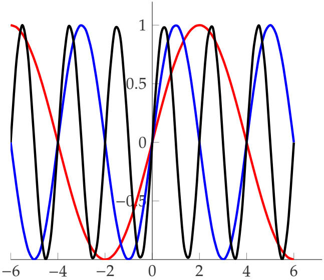


- $\alpha = \frac{\pi}{4}$
- $\alpha = \frac{\pi}{2}$
- $\alpha = \pi$

Sine Functions

The hypothesis space of sine functions is defined as

$$\mathcal{H}_{\sin} = \{\sin(\alpha \cdot x) : \alpha \in \mathbb{R}\} \tag{24}$$



- $\alpha = \frac{\pi}{4}$
- $\alpha = \frac{\pi}{2}$
- $\alpha = \pi$

$$\text{VC-dim}(\mathcal{H}_{\sin}) = \infty$$

Reference

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018).
Foundations of machine learning.
MIT press.

Shalev-Shwartz, S. and Ben-David, S. (2014).
Understanding machine learning: From theory to algorithms.
Cambridge university press.