Bias-Variance Tradeoff
Overview of the second half the semester

1. A little bit Learning Theory

2. Make our linear models nonlinear (Kernel)

3. How to combine multiple classifiers into a stronger one (Bagging & Boosting)?

4. Intro of Neural Networks (old and new)
Outline of Today

1. Intro on Underfitting/Overfitting and Bias/Variance

2. Derivation of the Bias-Variance Decomposition

3. Example on Ridge Linear Regression
Bayes optimal predictor

Consider regression problem with dataset $\mathcal{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$
Consider regression problem with dataset $\mathcal{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$

The Bayes optimal regressor:

$$\bar{y}(x) := \mathbb{E}[y | x]$$

(e.g., size of the house)
Bayes optimal predictor

Consider regression problem w/ dataset $\mathcal{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$

The Bayes optimal regressor:

$$\bar{y}(x) := \mathbb{E}[y \mid x]$$

The best we could do, cannot beat this one
Consider regression problem with dataset $\mathcal{D} = \{x, y\}, (x, y) \sim P, x \in \mathbb{R}, y \in \mathbb{R}$

The Bayes optimal regressor:

$$\bar{y}(x) = w_0 + w_1 x + w_2 x^2$$

The best we could do, cannot beat this one

$$\bar{y}(x) := \mathbb{E}[y \mid x]$$
Underfitting

$\tilde{y}(x) = w_0 + w_1 x + w_2 x^2$

(Just right)
Underfitting

\[ \bar{y}(x) = w_0 + w_1 x + w_2 x^2 \]

\[ h_D(x) = w_0 + w_1 x \]

\( X \) (e.g., size of the house)
Underfitting

\[ \bar{y}(x) = w_0 + w_1 x + w_2 x^2 \]

\[ h_{\mathcal{D}}(x) = w_0 + w_1 x \]
Underfitting

Just right versus Underfitting

\( x \) (e.g., size of the house)

\( y \)

Bias:

Bias towards to linear models
Underfitting

Now let’s redo linear regression on a different dataset $D'$, but from the same distribution.
Underfitting

Now let’s redo linear regression on a different dataset \( \mathcal{D}' \), but from the same distribution.
Underfitting

Now let’s redo linear regression on a different dataset $D'$, but from the same distribution.

The new linear function does not differ too much from the old one.
Underfitting

Now let’s redo linear regression on a different dataset $D'$, but from the same distribution.

The new linear function does not differ too much from the old one.

This is called low variance.
Now let’s redo linear regression on a different dataset $D'$, but from the same distribution. The new linear function does not differ too much from the old one. This is called low variance.

Q: what happens when our linear predictor is $h(x) = w_0$?
Now let’s redo linear regression on a different dataset \( \mathcal{D}' \), but from the same distribution.

The new linear function does not differ too much from the old one.

This is called low variance.

Q: what happens when our linear predictor is \( h(x) = w_0 \)?

A: in this case, \( w_0 \) models the mean of the \( y \) in data.
Summary on underfitting

1. Often our model is too simple, i.e., we bias towards too simple models
Summary on underfitting

1. Often our model is too simple, i.e., we bias towards too simple models

2. This causes underfitting, i.e., we cannot capture the trend in the data
Summary on underfitting

1. Often our model is too simple, i.e., we bias towards too simple models.

2. This causes underfitting, i.e., we cannot capture the trend in the data.

3. In this case, we have large bias, but low variance (think about the $h(x) = w_0$ case).
Overfitting

\[ \ddot{y}(x) = w_0 + w_1 x + w_2 x^2 \]

\( y \quad x \) (e.g., size of the house)
Overfitting

\[ \bar{y}(x) = w_0 + w_1 x + w_2 x^2 \]

\[ h_{\mathcal{D}}(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_6 x^5 \]
Overfitting

\[ \bar{y}(x) = w_0 + w_1 x + w_2 x^2 \]

\[ h_{\mathcal{D}}(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_6 x^5 \]

\( \mathcal{X} \) (e.g., size of the house)

(Just right)

Overfitting
Overfitting

Just right versus Overfitting

\[ y \]

\[ X \text{ (e.g., size of the house)} \]
Overfitting

Just right versus Overfitting

No strong bias:
Our hypothesis class is all polynomials up to 5-th order

\[ y = w_0 + w_1 x + w_2 x^2 + \ldots + w_5 x^5 \]
Overfitting

Just right versus Overfitting

No strong bias:

Our hypothesis class is all polynomials up to 5-th order

i.e., in a priori, no strong bias towards linear or quadratic, or cubic, etc
Overfitting

Redo the higher-order polynomial fitting on different dataset $\mathcal{D}'$
Overfitting

Redo the higher-order polynomial fitting on different dataset $\mathcal{D}'$

$y$

$\mathcal{X}$ (e.g., size of the house)
Overfitting

Redo the higher-order polynomial fitting on different dataset $\mathcal{D}'$

The new linear function does differ a lot from the old one.
Overfitting

Redo the higher-order polynomial fitting on different dataset $\mathcal{D}'$

The new linear function does differ a lot from the old one

This is called high variance

$y$

$x$ (e.g., size of the house)
Summary on Overfitting

1. Often our model is too complex (e.g., can fit noise perfectly to achieve zero training error)
Summary on Overfitting

1. Often our model is too complex (e.g., can fit noise perfectly to achieve zero training error)

2. This causes overfitting, i.e., cannot generalize well on unseen test example
Summary on Overfitting

1. Often our model is too complex (e.g., can fit noise perfectly to achieve zero training error)

2. This causes overfitting, i.e., cannot generalize well on unseen test example

3. In this case, we have small bias, but large variance (tiny change on the dataset cause large change in the fitted functions)
Outline of Today

1. Intro on Underfitting/Overfitting and Bias/Variance

2. Derivation of the Bias-Variance Decomposition

3. Example on Ridge Linear Regression
Generalization error

Given dataset $\mathcal{D}$, a hypothesis class $\mathcal{H}$, squared loss $\ell(h, x, y) = (h(x) - y)^2$, denote $h_\mathcal{D}$ as the ERM solution
Generalization error

Given dataset $\mathcal{D}$, a hypothesis class $\mathcal{H}$, squared loss $\ell(h, x, y) = (h(x) - y)^2$, denote $h_\mathcal{D}$ as the ERM solution.

We are interested in the generalization bound of $h_\mathcal{D}$:

$$\mathbb{E}_\mathcal{D} \mathbb{E}_{x, y \sim p}(h_\mathcal{D}(x) - y)^2$$
Generalization error

Given dataset $\mathcal{D}$, a hypothesis class $\mathcal{H}$, squared loss $\ell(h, x, y) = (h(x) - y)^2$, denote $h_{\mathcal{D}}$ as the ERM solution.

We are interested in the generalization bound of $h_{\mathcal{D}}$:

$$\mathbb{E}_\mathcal{D} \mathbb{E}_{x, y \sim p}(h_{\mathcal{D}}(x) - y)^2$$

Q: how to estimate this in practice?
The expectation of our model $h_{\mathcal{D}}$

Since $h_{\mathcal{D}}$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_{\mathcal{D}} [h_{\mathcal{D}}]$$
The expectation of our model $h_{\mathcal{D}}$

Since $h_{\mathcal{D}}$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_{\mathcal{D}} [h_{\mathcal{D}}]$$

In other words, we have:

$$\bar{h}(x) = \mathbb{E}_{\mathcal{D}} [h_{\mathcal{D}}(x)], \forall x$$
The expectation of our model $h_D$

Since $h_D$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_D \left[ h_D \right]$$

In other words, we have:

$$\bar{h}(x) = \mathbb{E}_D \left[ h_D(x) \right], \forall x$$

Q: what is $\bar{h}$ is the case where hypothesis is $h(x) = w_0$?
The expectation of our model $h_{\mathcal{D}}$

Since $h_{\mathcal{D}}$ is random, we consider its expected behavior:

$$\bar{h} := \mathbb{E}_{\mathcal{D}} \left[ h_{\mathcal{D}} \right]$$

In other words, we have:

$$\bar{h}(x) = \mathbb{E}_{\mathcal{D}} \left[ h_{\mathcal{D}}(x) \right], \forall x$$

Q: what is $\bar{h}$ is the case where hypothesis is $h(x) = w_0$?

A: $\bar{h}(x) = \mathbb{E}_{y}[y]$
Formal definition of Bias and Variance

\[ \bar{h} := \mathbb{E}_\mathcal{D} [h_\mathcal{D}] \quad \bar{y}(x) := \mathbb{E}[y|x] \]

Bias: difference between \( \bar{h} \) and the best \( \bar{y}(x) \), i.e., \( \mathbb{E}_x (\bar{y}(x) - \bar{h}(x))^2 \)
Formal definition of Bias and Variance

\[
\bar{h} := \mathbb{E}_\mathcal{D} \left[ h_\mathcal{D} \right] \quad \bar{y}(x) := \mathbb{E}[y | x]
\]

Bias: difference between \( \bar{h} \) and the best \( \bar{y}(x) \), i.e.,

\[
\mathbb{E}_x \left( \bar{y}(x) - \bar{h}(x) \right)^2
\]

Difference between our mean and the best
Formal definition of Bias and Variance

\[ \bar{h} := \mathbb{E}_D [h_D] \quad \bar{y}(x) := \mathbb{E}[y \mid x] \]

Bias: difference between \( \bar{h} \) and the best \( \bar{y}(x) \), i.e., \( \mathbb{E}_x \left( \bar{y}(x) - \bar{h}(x) \right)^2 \)

Variance: difference from \( \bar{h} \) and \( h_D \), i.e., \( \mathbb{E}_D \mathbb{E}_x \left( h_D(x) - \bar{h}(x) \right)^2 \)
Formal definition of Bias and Variance

$$\bar{h} := \mathbb{E}_\mathcal{D} [h_\mathcal{D}] \quad \bar{y}(x) := \mathbb{E}[y \mid x]$$

Bias: difference between $\bar{h}$ and the best $\bar{y}(x)$, i.e., $\mathbb{E}_x (\bar{y}(x) - \bar{h}(x))^2$

Difference between our mean and the best

Variance: difference from $\bar{h}$ and $h_\mathcal{D}$, i.e., $\mathbb{E}_\mathcal{D} \mathbb{E}_x (h_\mathcal{D}(x) - \bar{h}(x))^2$

Fluctuation of our random model around its mean
Generalization error decomposition

\[ \bar{h} := \mathbb{E}_\mathcal{D} [h_\mathcal{D}] \quad \bar{y}(x) := \mathbb{E}[y | x] \]

What we gonna show now:
Generalization error decomposition

\[ \bar{h} := \mathbb{E}_D [h_D] \quad \bar{y}(x) := \mathbb{E}[y | x] \]

What we gonna show now:

\[ \mathbb{E}_D \mathbb{E}_{x,y \sim P}(h_D(x) - y)^2 \]

= **Bias** + **Variance** + Noise (unavoidable, independent of Algs/models)
Generalization error decomposition

\[ \bar{h} := \mathbb{E}_D \left[ h_D \right] \quad \bar{y}(x) := \mathbb{E}[y | x] \]

What we gonna show now:

\[ \mathbb{E}_D \mathbb{E}_{x,y \sim p}(h_D(x) - y)^2 \]

= **Bias** + **Variance** + Noise (unavoidable, independent of Algs/models)

We will use the following trick twice: \((x - y)^2 = (x - z)^2 + (z - y)^2 + 2(x - z)(z - y)\)
\[ E(h_{\mathcal{D}}(x) - y)^2 = E(h_{\mathcal{D}}(x) - \bar{h}(x) + \bar{h}(x) - y)^2 \]
\[ \mathbb{E}(h_\mathcal{D}(x) - y)^2 \]

\[ = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x) + \bar{h}(x) - y)^2 \]

\[ = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 + 2\mathbb{E}_{x,y,\mathcal{D}} [(h_\mathcal{D}(x) - \bar{h}(x))(\bar{h}(x) - y)] \]
\[ \mathbb{E}(h_{\mathcal{D}}(x) - y)^2 \]

\[ = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x) + \bar{h}(x) - y)^2 \]

\[ = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 - 2\mathbb{E}_{x,y,\mathcal{D}}[(h_{\mathcal{D}}(x) - \bar{h}(x))(\bar{h}(x) - y)] \]

This term is zero since:
\[ \mathbb{E}(h_\mathcal{D}(x) - y)^2 \]
\[ = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x) + \bar{h}(x) - y)^2 \]
\[ = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 - 2\mathbb{E}_{x,y}\mathcal{D}[(h_\mathcal{D}(x) - \bar{h}(x))(\bar{h}(x) - y)] \]

This term is zero since:
\[ \mathbb{E}\left[(h_\mathcal{D}(x) - \bar{h}(x))(\bar{h}(x) - y)\right] \]
\[ = \mathbb{E}_{x,y}\left[\mathbb{E}_\mathcal{D}(h_\mathcal{D}(x) - \bar{h}(x)) \cdot (\bar{h}(x) - y)\right] \]
\[ = \mathbb{E}_\mathcal{D}\left[h_\mathcal{D}(x)\right] \cdot \mathbb{E}(\bar{h}(x) - y) \]
\[
\mathbb{E}(h_\mathcal{D}(x) - y)^2 \\
= \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x) + \bar{h}(x) - y)^2 \\
= \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 - 2\mathbb{E}_{x,y,\mathcal{D}} \left[ (h_\mathcal{D}(x) - \bar{h}(x))(\bar{h}(x) - y) \right] \\
\]

This term is zero since:

\[
\mathbb{E} \left[ (h_\mathcal{D}(x) - \bar{h}(x))(\bar{h}(x) - y) \right] \\
= \mathbb{E}_{x,y} \left[ \mathbb{E}_\mathcal{D}(h_\mathcal{D}(x) - \bar{h}(x)) \cdot (\bar{h}(x) - y) \right] \\
= \mathbb{E}_{x,y} \left[ (\bar{h}(x) - \bar{h}(x)) \cdot (\bar{h}(x) - y) \right] = 0
\]
\[ \mathbb{E}(h(x) - y)^2 \]

\[ = \mathbb{E}(h(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 \]
\[ \mathbb{E}(h_{\mathcal{D}}(x) - y)^2 \]

\[ = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 \]

Bias: \( \bar{h}(x) - \bar{y}(x) \) where \( \bar{y}(x) = \mathbb{E}_{y | x}[y] \)

Variance
\[ \mathbb{E}(h_{\mathcal{D}}(x) - y)^2 \]

\[ = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 \]

Variance
\[ \mathbb{E}(h_\mathcal{D}(x) - y)^2 = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 \]

\[ = \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2 \]
\[ \mathbb{E}(h_\mathcal{D}(x) - y)^2 = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 \]

\[ = \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2 \]

\[ = \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2 \]

\[ + 2\mathbb{E}(\bar{h}(x) - \bar{y}(x))(\bar{y}(x) - y) \]
\[
\mathbb{E}(h(x) - y)^2 = \mathbb{E}(h(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2
\]

Variance

\[
= \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2
\]

\[
= \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2
\]

\[
+ 2\mathbb{E}(\bar{h}(x) - \bar{y}(x)))(\bar{y}(x) - y)
\]

This term is zero since:

\[
\mathbb{E}_{x, y} \left[ (h(x) - \bar{y}(x)) (\bar{y}(x) - y) \right]
\]

\[
= \mathbb{E}_x \mathbb{E}_{y|x} (\bar{h}(x) - \bar{y}(x))(\bar{y}(x) - y)
\]

\[
= 0
\]

\[
\mathbb{E}(x, y) = p(x) p(y|x)
\]

\[
\Rightarrow x \sim p, \quad y \sim p(y|x)
\]
\[ \mathbb{E}(h(x) - y)^2 = \mathbb{E}(h(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2 \]

This term is zero since:

\[ = \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2 \]

\[ = \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2 + 2\mathbb{E}(\bar{h}(x) - \bar{y}(x))(\bar{y}(x) - y) \]

\[ = \mathbb{E}_\mathcal{D}\mathbb{E}_x(\bar{h}(x) - \bar{y}(x)) \cdot \mathbb{E}_{y|x}(\bar{y}(x) - y) \]
\[
\mathbb{E}(h_{\mathcal{D}}(x) - y)^2 = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - y)^2
\]

Variance

\[
= \mathbb{E}(\bar{h}(x) - \bar{y}(x) + \bar{y}(x) - y)^2
\]

\[
+ 2\mathbb{E}(\bar{h}(x) - \bar{y}(x)) (\bar{y}(x) - y)
\]

This term is zero since:

\[
= \mathbb{E}_{\mathcal{D}} \mathbb{E}_x(\bar{h}(x) - \bar{y}(x)) \cdot \mathbb{E}_{y|x}(\bar{y}(x) - y)
\]

\[
= \mathbb{E}_{\mathcal{D}} \mathbb{E}_x(\bar{h}(x) - \bar{y}(x)) \cdot (\bar{y}(x) - \mathbb{E}_{y|x}[y])
\]

\[
\bar{y}(x) = \mathbb{E}_{y|x}[y]
\]
Putting the derivations together, we arrive at:

\[ \mathbb{E}(h_\mathcal{D}(x) - y)^2 = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2 \]

Variance Bias

\[ \mathbb{E}(h_\mathcal{D}(x) - y)^2 = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2 \]

Variance Bias

\[ \mathbb{E}(h_\mathcal{D}(x) - y)^2 = \mathbb{E}(h_\mathcal{D}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2 \]
Putting the derivations together, we arrive at:

$$\mathbb{E}(h_{\mathcal{D}}(x) - y)^2 = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2$$

- Variance
- Bias
- Noise
Putting the derivations together, we arrive at:

\[
\mathbb{E}(h_{\mathcal{D}}(x) - y)^2 = \mathbb{E}(h_{\mathcal{D}}(x) - \bar{h}(x))^2 + \mathbb{E}(\bar{h}(x) - \bar{y}(x))^2 + \mathbb{E}(\bar{y}(x) - y)^2
\]

Variance
Bias
Noise

Note that the noise term is independent of training algorithms / models.
Outline of Today

1. Intro on Underfitting/Overfitting and Bias/Variance
2. Derivation of the Bias-Variance Decomposition
3. Example on Ridge Linear Regression
Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., $x_1, \ldots, x_n$ fixed (no randomness)
Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., \( x_1, \ldots, x_n \) fixed (no randomness)

But \( y_i \sim (w^*)^T x_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0,1) \)
Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., $x_1, \ldots, x_n$ fixed (no randomness)

$$y_i \sim (w^*)^T x_i + \epsilon_i, \; \epsilon_i \sim \mathcal{N}(0,1)$$

(This is called LR w/ fixed design)
Ex: Ridge Linear regression

Let us consider the case where features are fixed, i.e., $x_1, \ldots, x_n$ fixed (no randomness)

But $y_i \sim (w^*)^T x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,1)$

(This is called LR w/ fixed design)

(So the only randomness of our dataset $\mathcal{D} = \{x_i, y_i\}$ is coming from the noises $\epsilon_i$)
Ex: Ridge Linear regression

Ridge Linear Regression formulation

\[ \hat{w} = \arg \min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2 \]
Ex: Ridge Linear regression

Ridge Linear Regression formulation

\[ \hat{w} = \arg \min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2 \]

What we will show now:

Larger \( \lambda \) (model becomes “simpler”) \( \Rightarrow \) larger bias, but smaller variance
Ex: Ridge Linear regression

Ridge Linear Regression formulation

$$\hat{w} = \arg\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$$

What we will show now:

Larger $\lambda$ (model becomes “simpler”) => larger bias, but smaller variance

(Q: think about the case where $\lambda \to \infty$, what happens to $\hat{w}$?)