Machine Learning and Intelligent Systems

Bias-Variance Decomposition

Maria A. Zuluaga

Dec 1, 2023

EURECOM - Data Science Department

Table of contents

Recap

So far in this course

- We have covered several machine learning methods:
    - Linear regression
    - Linear discriminant analysis
    - Logistic regression
    - The perceptron

So far in this course

- We have covered several machine learning methods
- We have stated that it is important that our trained models **generalize**
- We started to discuss some problems, such as the variance
- But, we **have not** really looked into it nor asked too many questions about our models:
    - Will they perform well on testing data?
    - Is the training data available enough?
    - Among many models, which one should be chosen?

So far in this course

- We have covered several machine learning methods
- We have stated that it is important that our trained models **generalize**
- We started to discuss some problems, such as the variance
- But, we **have not** really looked into it nor asked too many questions about our models:
  - Will they perform well on testing data?
  - Is the training data available enough?
  - Among many models, which one should be chosen?

This lecture:

We will answer these questions from a theoretical and practical perspective

Definitions

- **Generalization:** Ability of a model to perform well on unseen data

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[l(y, h(\mathbf{x}))]$$

Generalization loss (from Lecture 1)

Definitions

- **Generalization:** Ability of a model to perform well on unseen data

$$\epsilon = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{P}}[l(y, h(\mathbf{x}))]$$

Generalization loss (from Lecture 1)

- **Model Selection:** Task of selecting a model from a set of candidate models given the data
  - Intuitively, we arrived to the conclusion that this is a necessary step but, we have not addressed it properly.
  - Examples: order of the polynomial features, features to use for a specific problem (lab 1), neural networks, the λ term in regularization

Bias-Variance Trade-off

Generalization Error

- We have focused on the minimization of the training error (loss)
- So far, we expect that this is good enough for our trained model to generalize.

Generalization Error

- We have focused on the minimization of the training error (loss)
- So far, we expect that this is good enough for our trained model to generalize.

- **Generalize:** Perform well on unseen data
- Lets have a detailed look into the generalization error:

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[l(y, h(\mathbf{x}))] \tag{1}$$

Definitions: Expected Label

- The training data comes in input pairs $(\mathbf{x}, y)$, with $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathcal{C}$.
- We will focus on the case where $\mathcal{C} = \mathbb{R}^O$, $O = 1$
- The **training set** points $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ i.i.d. drawn from an unknown probability distribution $\mathcal{P}(X, Y)$.

Definitions: Expected Label

- The training data comes in input pairs $(\mathbf{x}, y)$, with $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathcal{C}$.
- We will focus on the case where $\mathcal{C} = \mathbb{R}^O,\ O = 1$
- The **training set** points $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ i.i.d. drawn from an unknown probability distribution $\mathcal{P}(X, Y)$.
- **Important:** For a given $\mathbf{x}$ there is a distribution of y

$$\mathcal{P}(X, Y) = \mathcal{P}(Y|X)\mathcal{P}(X)$$

Definitions: Expected Label

- The training data comes in input pairs $(\mathbf{x}, y)$, with $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathcal{C}$.
- We will focus on the case where $\mathcal{C} = \mathbb{R}^O,\ O = 1$
- The **training set** points $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ i.i.d. drawn from an unknown probability distribution $\mathcal{P}(X, Y)$.
- **Important:** For a given $\mathbf{x}$ there is a distribution of y

$$\mathcal{P}(X, Y) = \mathcal{P}(Y|X)\mathcal{P}(X)$$

Definition 1: Expected Label

$$\bar{y}(\mathbf{x}) = \mathbb{E}_{y|\mathbf{x}}[y] = \int y\mathcal{P}(y|\mathbf{x})dy \tag{2}$$

Machine Learning Algorithms, Learning Process & Hypothesis

- Let us now denote a machine learning algorithm as $\mathcal{A}$
- **Learning Process:** We use an algorithm $\mathcal{A}$ in conjunction with the training set $\mathcal{D}$ to learn a hypothesis $h \in \mathcal{H}$

Machine Learning Algorithms, Learning Process & Hypothesis

- Let us now denote a machine learning algorithm as $\mathcal{A}$
- **Learning Process:** We use an algorithm $\mathcal{A}$ in conjunction with the training set $\mathcal{D}$ to learn a hypothesis $h \in \mathcal{H}$
- Formally,

$$h_{\mathcal{D}} = \mathcal{A}(\mathcal{D}) \tag{3}$$

Definitions: Expected Generalization Error Given $h_{\mathcal{D}}$

- Since we are working with regression, we can use the quadratic loss in Eq. 1

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[(h(\mathbf{x}) - y)^2]$$

Definitions: Expected Generalization Error Given $h_{\mathcal{D}}$

- Since we are working with regression, we can use the quadratic loss in Eq. 1

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[(h(\mathbf{x}) - y)^2]$$

- We also redefined h as $h_{\mathcal{D}}$ to denote its dependency to the dataset

Definitions: Expected Generalization Error Given $h_{\mathcal{D}}$

- Since we are working with regression, we can use the quadratic loss in Eq. 1

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[(h(\mathbf{x}) - y)^2]$$

- We also redefined h as $h_{\mathcal{D}}$ to denote its dependency to the dataset

Definition 2: Expected Generalization Error (given $h_{\mathcal{D}}$)

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] \tag{4}$$

Definitions: Expected Generalization Error Given $h_{\mathcal{D}}$

- Since we are working with regression, we can use the quadratic loss in Eq. 1

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[(h(\mathbf{x}) - y)^2]$$

- We also redefined h as $h_{\mathcal{D}}$ to denote its dependency to the dataset

Definition 2: Expected Generalization Error (given $h_{\mathcal{D}}$)

$$\begin{aligned} \epsilon &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] \\ &= \int_{\mathbf{x}}\int_{y}(h_{\mathcal{D}}(\mathbf{x}) - y)^2\mathcal{P}(\mathbf{x}, y)\partial y\partial\mathbf{x} \end{aligned} \tag{4}$$

Definitions: Expected Predictor

- The term in Eq. 4 is valid for a given $\mathcal{D}$
- $\mathcal{D}$ is drawn from $\mathcal{P}^N$. It is a random variable.

Definitions: Expected Predictor

- The term in Eq. 4 is valid for a given $\mathcal{D}$
- $\mathcal{D}$ is drawn from $\mathcal{P}^N$. It is a random variable.
- Since $h_{\mathcal{D}}$ is a function of $\mathcal{D}$, it is a random variable too.
- We can compute its expectation

Definitions: Expected Predictor

- The term in Eq. 4 is valid for a given $\mathcal{D}$
- $\mathcal{D}$ is drawn from $\mathcal{P}^N$. It is a random variable.
- Since $h_{\mathcal{D}}$ is a function of $\mathcal{D}$, it is a random variable too.
- We can compute its expectation

Definition 3: Expected Predictor (given $\mathcal{A}$)

$$\bar{h} = \mathbb{E}_{\mathcal{D} \sim \mathcal{P}^N}[h_{\mathcal{D}}] = \int_{\mathcal{D}} h_{\mathcal{D}} \mathcal{P}(\mathcal{D}) \tag{5}$$

Definitions: Expected Generalization Error of $\mathcal{A}$

- We are interested in knowing the generalization error of a type of method in general. Not just $h_{\mathcal{D}}$
- We need to integrate over all $h_{\mathcal{D}}$

Definitions: Expected Generalization Error of $\mathcal{A}$

- We are interested in knowing the generalization error of a type of method in general. Not just $h_{\mathcal{D}}$
- We need to integrate over all $h_{\mathcal{D}}$

Definition 4: Expected Generalization Error of $\mathcal{A}$

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P},\mathcal{D}\sim\mathcal{P}^N}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] \tag{6}$$

Definitions: Expected Generalization Error of $\mathcal{A}$

- We are interested in knowing the generalization error of a type of method in general. Not just $h_{\mathcal{D}}$
- We need to integrate over all $h_{\mathcal{D}}$

Definition 4: Expected Generalization Error of $\mathcal{A}$

$$\begin{aligned}\epsilon &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P},\mathcal{D}\sim\mathcal{P}^N}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] \\ &= \int_{\mathcal{D}}\int_{\mathbf{x}}\int_{y}(h_{\mathcal{D}}(\mathbf{x}) - y)^2\mathcal{P}(\mathbf{x}, y)\mathcal{P}(\mathcal{D})\partial y\partial\mathbf{x}\partial\mathcal{D}\end{aligned} \tag{6}$$

Definitions: Expected Generalization Error of $\mathcal{A}$

- We are interested in knowing the generalization error of a type of method in general. Not just $h_{\mathcal{D}}$
- We need to integrate over all $h_{\mathcal{D}}$

Definition 4: Expected Generalization Error of $\mathcal{A}$

$$\begin{aligned}\epsilon &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P},\mathcal{D}\sim\mathcal{P}^N}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] \\ &= \int_{\mathcal{D}}\int_{\mathbf{x}}\int_{y}(h_{\mathcal{D}}(\mathbf{x}) - y)^2\mathcal{P}(\mathbf{x},y)\mathcal{P}(\mathcal{D})\partial y\partial\mathbf{x}\partial\mathcal{D}\end{aligned} \tag{6}$$

Important: In Eq. 6, $(\mathbf{x}, y) \sim \mathcal{P}$ denotes points at testing and $\mathcal{D} \sim \mathcal{P}^N$ is training data

Definitions: Expected Generalization Error of $\mathcal{A}$

- We are interested in knowing the generalization error of a type of method in general. Not just $h_{\mathcal{D}}$
- We need to integrate over all $h_{\mathcal{D}}$

Definition 4: Expected Generalization Error of $\mathcal{A}$

$$\begin{aligned}\epsilon &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P},\mathcal{D}\sim\mathcal{P}^N}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] \\ &= \int_{\mathcal{D}}\int_{\mathbf{x}}\int_{y}(h_{\mathcal{D}}(\mathbf{x}) - y)^2\mathcal{P}(\mathbf{x}, y)\mathcal{P}(\mathcal{D})\partial y\partial\mathbf{x}\partial\mathcal{D}\end{aligned} \tag{6}$$

Important: In Eq. 6, $(\mathbf{x}, y) \sim \mathcal{P}$ denotes points at testing and $\mathcal{D} \sim \mathcal{P}^N$ is training data

In words: We measure how well an algorithm $\mathcal{A}$ generalizes with respect to a data distribution $\mathcal{P}(X, Y)$

Decomposition of the Generalization Error: Step 1

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P},\mathcal{D}\sim\mathcal{P}^N}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] \rightarrow \mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2]$$

Let's have a deeper look into it:

Decomposition of the Generalization Error: Step 1

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P},\mathcal{D}\sim\mathcal{P}^N}[(h_\mathcal{D}(\mathbf{x})-y)^2] \to \mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_\mathcal{D}(\mathbf{x})-y)^2]$$

Let's have a deeper look into it:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_\mathcal{D}(\mathbf{x})-y)^2] &= \mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_\mathcal{D}(\mathbf{x})-\bar{h}(\mathbf{x})+\bar{h}(\mathbf{x})-y)^2] \\
&= \mathbb{E}_{\mathbf{x},y,\mathcal{D}}[((h_\mathcal{D}(\mathbf{x})-\bar{h}(\mathbf{x}))+(\bar{h}(\mathbf{x})-y))^2] \\
&= \mathbb{E}_{\mathbf{x},\mathcal{D}}[(h_\mathcal{D}(\mathbf{x})-\bar{h}(\mathbf{x}))^2]+\mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x})-y)^2] \\
&\quad + 2\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_\mathcal{D}(\mathbf{x})-\bar{h}(\mathbf{x}))(\bar{h}(\mathbf{x})-y)]
\end{aligned}$$

The last term is equal to zero:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_\mathcal{D}(\mathbf{x})-\bar{h}(\mathbf{x}))(\bar{h}(\mathbf{x})-y)^2] &= \mathbb{E}_{\mathbf{x},y}[\mathbb{E}_\mathcal{D}[(h_\mathcal{D}(\mathbf{x})-\bar{h}(\mathbf{x}))](\bar{h}(\mathbf{x})-y)] \\
&= \mathbb{E}_{\mathbf{x},y}[(\mathbb{E}_\mathcal{D}[h_\mathcal{D}(\mathbf{x})])-\mathbb{E}_\mathcal{D}[\bar{h}(\mathbf{x})](\bar{h}(\mathbf{x})-y)] \\
&= \mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x})-\bar{h}(\mathbf{x}))(\bar{h}(\mathbf{x})-y)] = 0
\end{aligned}$$

Decomposition of the Generalization Error: Step 2

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - \bar{h}(\mathbf{x}))^2]}_{\text{variance}} + \mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x}) - y)^2] \tag{7}$$

Let's do a similar manipulation to the second term in Eq. 7:

Decomposition of the Generalization Error: Step 2

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - \bar{h}(\mathbf{x}))^2]}_{\text{variance}} + \mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x}) - y)^2] \tag{7}$$

Let's do a similar manipulation to the second term in Eq. 7:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x}) - y)^2] &= \mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}) + \bar{y}(\mathbf{x}) - y)^2] \\
&= \mathbb{E}_{\mathbf{x},y}[((\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})) + (\bar{y}(\mathbf{x}) - y))^2] \\
&= \mathbb{E}_{\mathbf{x}}[(\bar{h}(\mathbf{x}) - \bar{y})^2] + \mathbb{E}_{\mathbf{x},y}[(\bar{y}(\mathbf{x}) - y)^2] + 2\mathbb{E}_{\mathbf{x},y}[(\bar{y}(\mathbf{x}) - y)(\bar{h}(\mathbf{x}) - \bar{y})] = 0
\end{aligned}$$

The last term is equal to zero:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y}[(\bar{y}(\mathbf{x}) - y)(\bar{h}(\mathbf{x}) - \bar{y})] &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[(\bar{y}(\mathbf{x}) - y)](\bar{h}(\mathbf{x}) - \bar{y})] \\
&= \mathbb{E}_{\mathbf{x}}[(\mathbb{E}_{y|\mathbf{x}}[\bar{y}(\mathbf{x})] - \mathbb{E}_{y|\mathbf{x}}[y])(\bar{h}(\mathbf{x}) - \bar{y})] \\
&= \mathbb{E}_{\mathbf{x}}[(\bar{y}(\mathbf{x}) - \bar{y}(\mathbf{x}))(\bar{h}(\mathbf{x}) - \bar{y})] = 0
\end{aligned}$$

Bias Variance Decomposition

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - \bar{h}(\mathbf{x}))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathbf{x},y}[(\bar{y}(\mathbf{x}) - y)^2]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}))^2]}_{\text{Bias}^2}$$

Bias Variance Decomposition

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - \bar{h}(\mathbf{x}))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathbf{x},y}[(\bar{y}(\mathbf{x}) - y)^2]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}))^2]}_{\text{Bias}^2}$$

Variance: Error caused from sensitivity to fluctuations in the training set. How much does the model change if it is trained in a different dataset. High variance can cause an algorithm to model noise from the training data rather than the intended targets (overfitting)

Bias Variance Decomposition

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - \bar{h}(\mathbf{x}))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathbf{x},y}[(\bar{y}(\mathbf{x}) - y)^2]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}))^2]}_{\text{Bias}^2}$$

Variance: Error caused from sensitivity to fluctuations in the training set. How much does the model change if it is trained in a different dataset. High variance can cause an algorithm to model noise from the training data rather than the intended targets (overfitting)

Bias: The inherent error that you obtain from the model even with infinite training data. This is due to the classifier being biased to a particular solution (e.g. linear classifier)

Bias Variance Decomposition

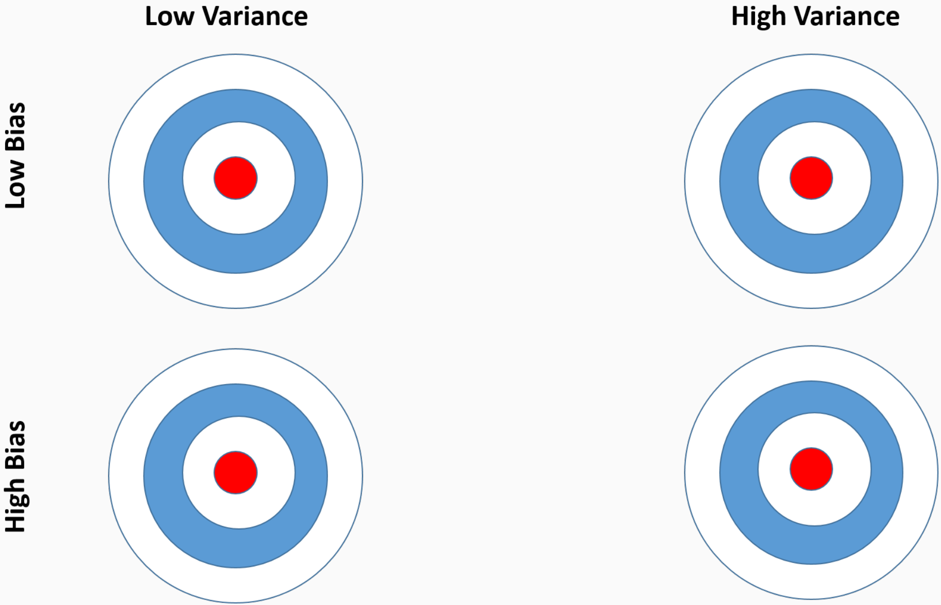
$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - y)^2] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(h_{\mathcal{D}}(\mathbf{x}) - \bar{h}(\mathbf{x}))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathbf{x},y}[(\bar{y}(\mathbf{x}) - y)^2]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}))^2]}_{\text{Bias}^2}$$

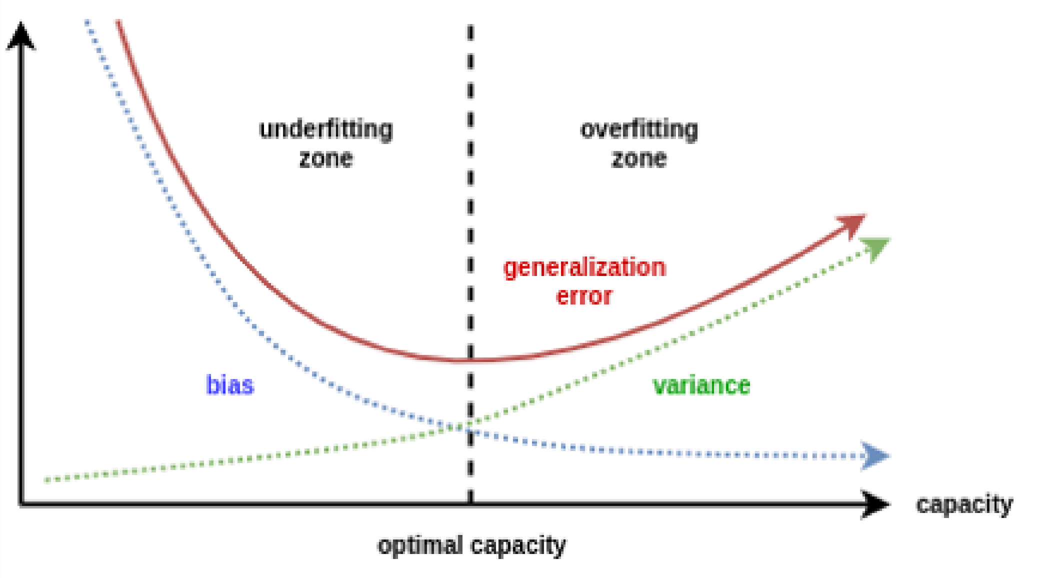
Variance: Error caused from sensitivity to fluctuations in the training set. How much does the model change if it is trained in a different dataset. High variance can cause an algorithm to model noise from the training data rather than the intended targets (overfitting)

Bias: The inherent error that you obtain from the model even with infinite training data. This is due to the classifier being biased to a particular solution (e.g. linear classifier)

Noise: The error associated to the data. It measures ambiguity due to your data distribution and feature representation. You can never beat this.

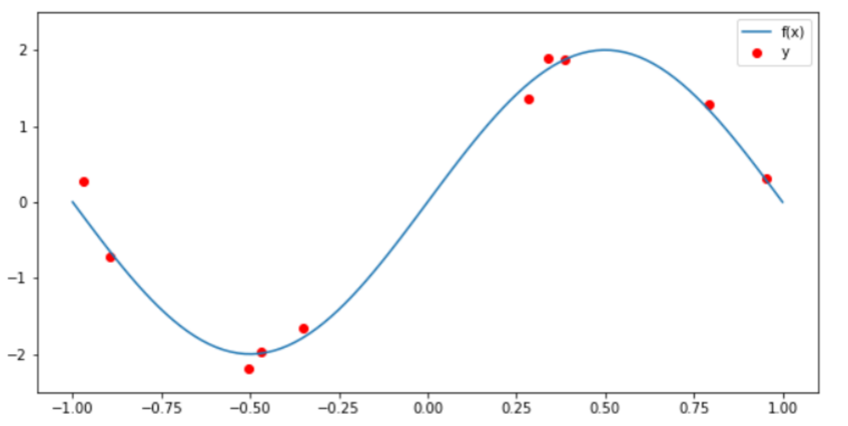
Analysis


Low Variance
High Variance
Low Bias
High Bias



Source: `https://djsaunde.wordpress.com/2017/07/17/the-bias-variance-tradeoff/`

A Simulated Example

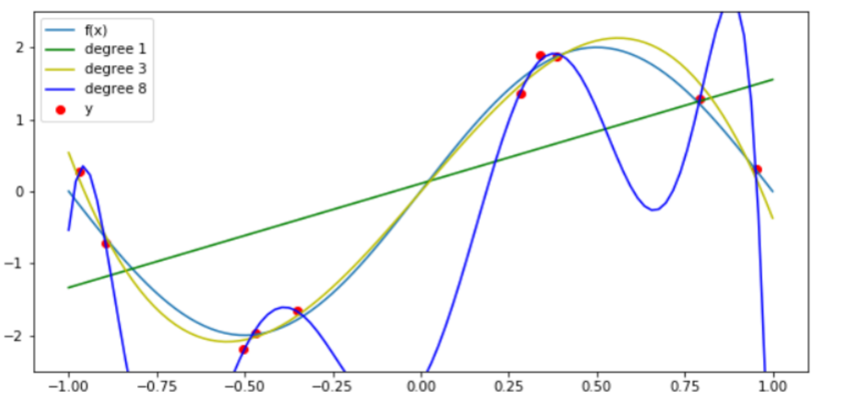


$$f(\mathbf{x}) = \sin(\pi \mathbf{x})$$

$$y = f(\mathbf{x}) + \mathcal{N}(0, \sigma^2)$$

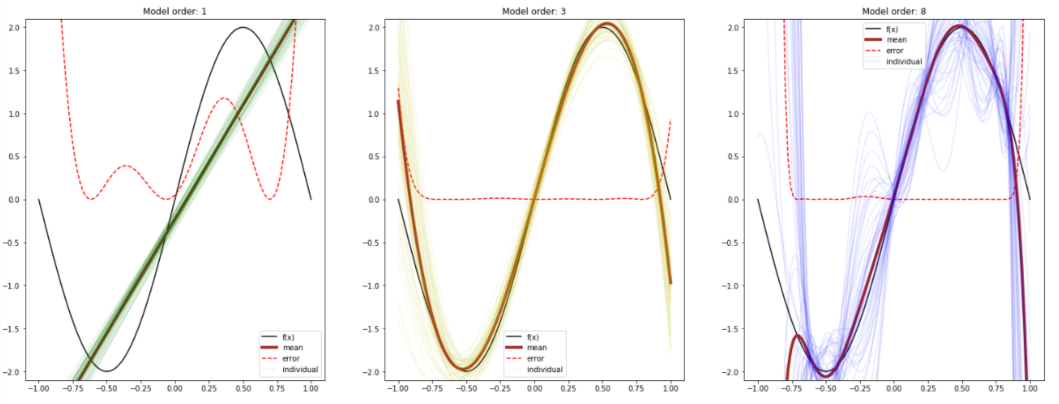
Notebook: See `04_bias_variance.ipynb`

A Simulated Example: Model Fitting

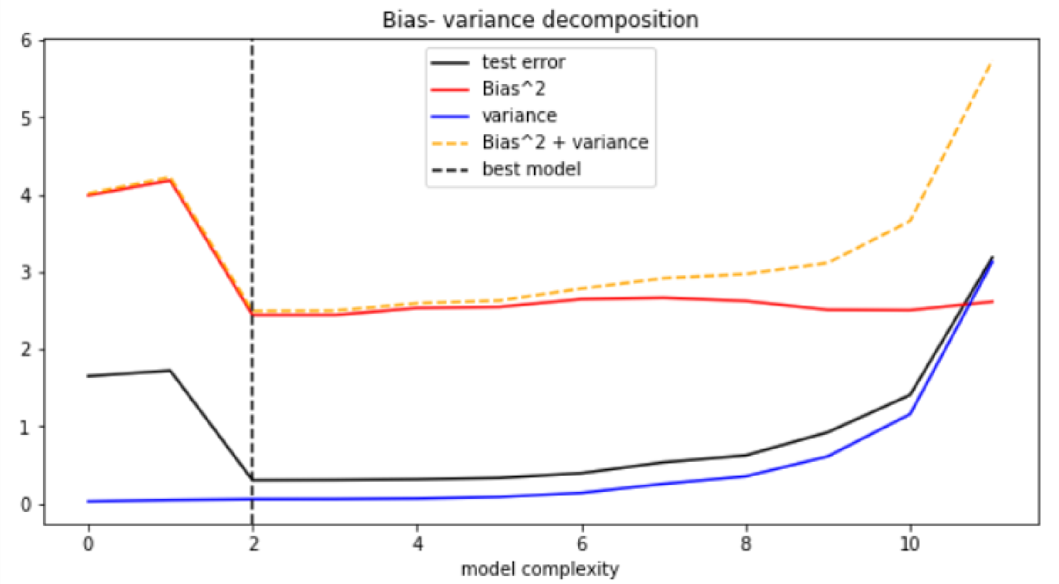


We fit models with different polynomial orders

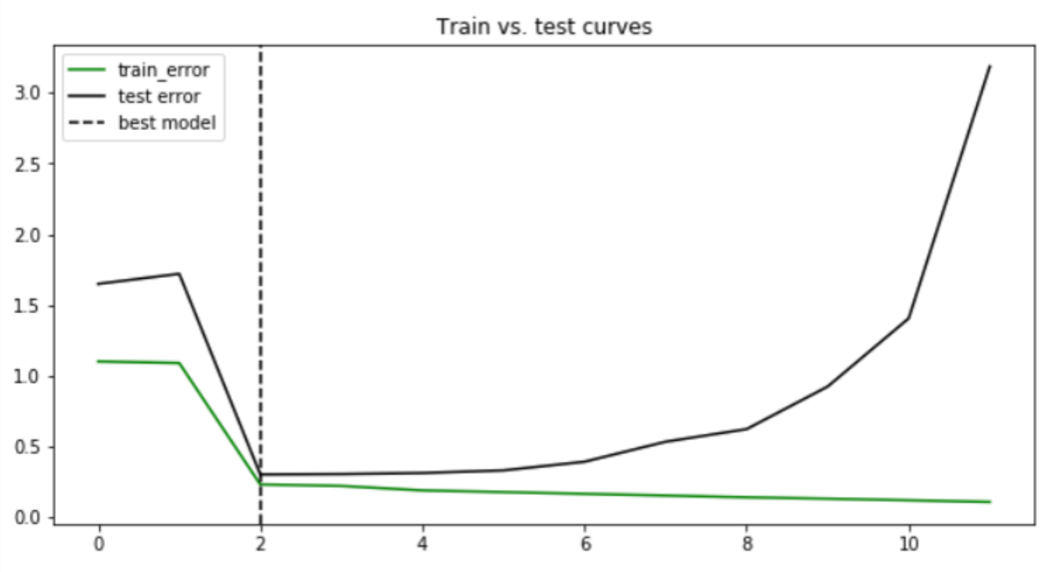
Simulation with 50 sampled datasets



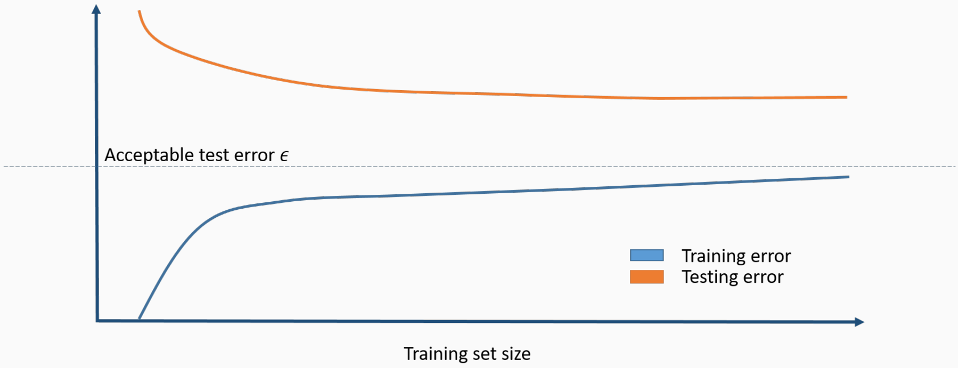
Bias-variance Decomposition (100 datasets)



Training vs. Testing Curve (100 datasets)



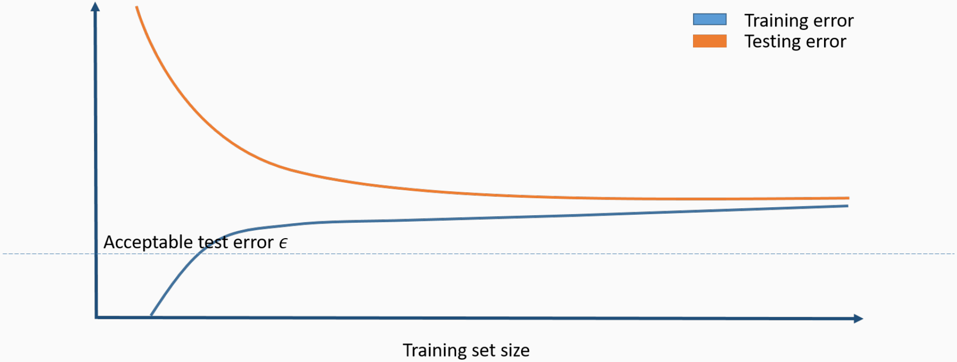
Detection of High Bias / Variance: Scenario 1



High Variance Symptoms: Test error $> \epsilon$, training error $\ll \epsilon$, training error $\ll$ test error

Potential solutions: More training data, reduce model complexity, bagging

Detection of High Bias / Variance: Scenario 2



High Bias Symptoms: Training error $> \epsilon$

Potential solutions: More complex model,add features, boosting

Training vs. Test Error: Summary

	Small training error	**Large training error**
Small testing error	Generalizes and performs	Plausible but weird (*)
Large testing error	Fails to generalize	Generalizes but performs poorly

Wrap-up

Wrap-up

- We studied in detail the generalization error
- We introduced the concept of bias and formalized that one of variance
- We had some insights on how to detect these problems in practice

Key Concepts

- Bias
- Variance
- Overfitting
- Noise
- Generalization Error