Introduction to Machine Learning

PSC 8185: Machine Learning for Social Science

Iris Malone

January 10, 2022

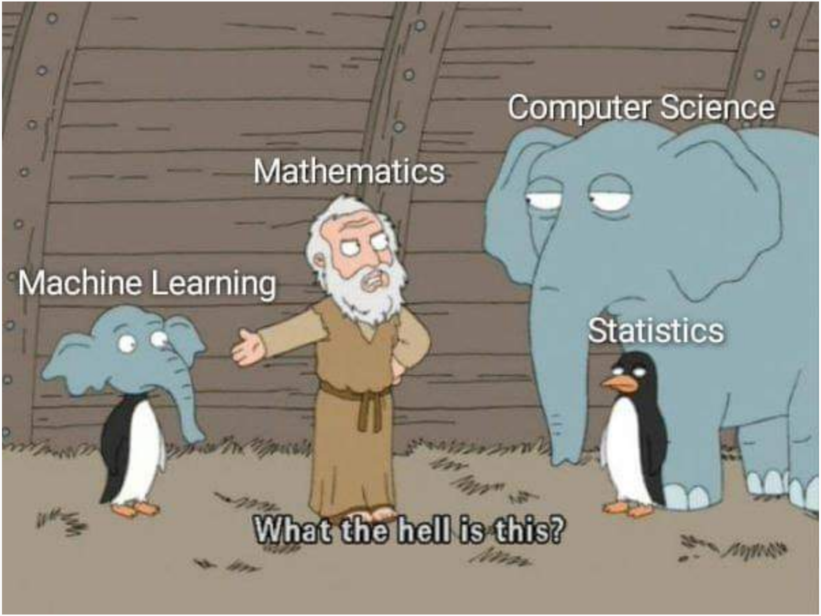
Materials adapted from Sergio Ballacado and Rochelle Terman

Agenda

1. Motivation
2. Course Overview
3. What is Machine Learning?
4. Preview: Model Assessment and Selection

Motivation

Machine Learning (ML) Seems Intimidating...



But it isn't.

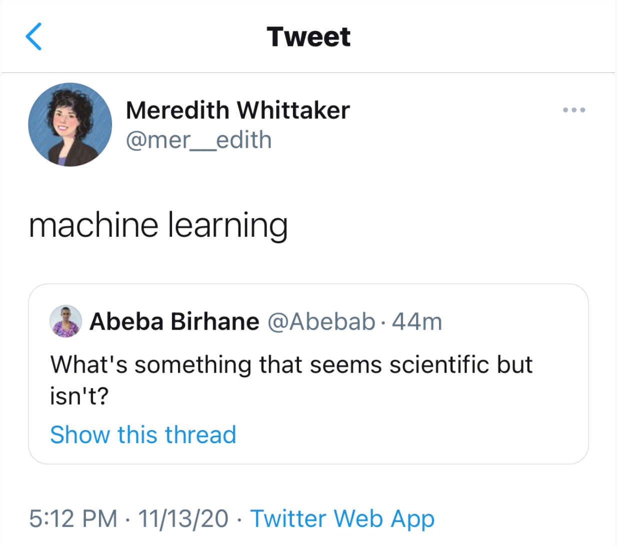


Figure 1: Faculty Director AI Now Institute, Research Prof NYU. Ex-Google.

Main Intuition Behind ML

Main Idea: ML is a form of artificial intelligence increasingly used in social science to solve complex prediction problems.

Main Intuition Behind ML

Main Idea: ML is a form of artificial intelligence increasingly used in social science to solve complex prediction problems. It involves a set of computer algorithms which 'learn' patterns in existing data to assist in prediction and description.

Main Intuition Behind ML

Main Idea: ML is a form of artificial intelligence increasingly used in social science to solve complex **prediction problems**. It involves a set of computer algorithms which ‘learn’ patterns in existing data to assist in prediction and description.

Examples of Prediction Problems

- Predict outcome of interest based on existing information

Examples of Prediction Problems

- Predict outcome of interest based on existing information
- Predict who will win the 2022 Midterm Elections based on public opinion polls and economic data

Examples of Prediction Problems

- Predict outcome of interest based on existing information
- Predict who will win the 2022 Midterm Elections based on public opinion polls and economic data
- Estimate Apple's stock price in March 2022 based on historical values

Examples of Prediction Problems

- Predict outcome of interest based on existing information
- Predict who will win the 2022 Midterm Elections based on public opinion polls and economic data
- Estimate Apple's stock price in March 2022 based on historical values
- Classify Facebook posts as 'fake' or 'real' news based on words in the title

Examples of Prediction Problems

- Predict outcome of interest based on existing information
- Predict who will win the 2022 Midterm Elections based on public opinion polls and economic data
- Estimate Apple's stock price in March 2022 based on historical values
- Classify Facebook posts as 'fake' or 'real' news based on words in the title
- Identify future Covid spikes based on Google Trends

Brief History of ML

- 1800s-1940s: Linear Models + Principal Component Analysis

Brief History of ML

- 1800s-1940s: Linear Models + Principal Component Analysis
- 1950s-1980s: Computing Power $\rightarrow$ "Machine Learning" + Neural Nets (Perceptron)

Brief History of ML

- 1800s-1940s: Linear Models + Principal Component Analysis
- 1950s-1980s: Computing Power $\rightarrow$ "Machine Learning" + Neural Nets (Perceptron)
- 1990s-2010s: Random Forests, Boosting, Support Vector Machines, Bayesian

Brief History of ML

- 1800s-1940s: Linear Models + Principal Component Analysis
- 1950s-1980s: Computing Power $\rightarrow$ "Machine Learning" + Neural Nets (Perceptron)
- 1990s-2010s: Random Forests, Boosting, Support Vector Machines, Bayesian
- Today:
  - More Computational Power
  - More Data
  - New Algorithms
  - Broader applications, more demand, bigger audience

Image Classification (Tensorflow)

ML Applications

- Industry
  - Measure consumer opinion
  - Deliver engaging content to users

ML Applications

- Industry
  - Measure consumer opinion
  - Deliver engaging content to users
- Public Sector
  - Predict disease onset
  - Assist criminal sentencing

ML Applications

- Industry
  - Measure consumer opinion
  - Deliver engaging content to users
- Public Sector
  - Predict disease onset
  - Assist criminal sentencing
- Campaigns
  - Target likely voters and donors
  - Identify ideology based on social media behavior

ML Applications

- Industry
  - Measure consumer opinion
  - Deliver engaging content to users
- Public Sector
  - Predict disease onset
  - Assist criminal sentencing
- Campaigns
  - Target likely voters and donors
  - Identify ideology based on social media behavior
- Social Science
  - Measure polarization in political institutions (Clinton, Jackman, and Rivers 2004)
  - Infer extent and strategy of Chinese censorship (King, Pan, and Roberts 2014)
  - Assess risk of conflict onset and escalation (Malone 2022)

Course Logistics

Course Presumptions

1. ML is relevant and useful in a wide range of academic and non-academic fields

Course Presumptions

1. ML is relevant and useful in a wide range of academic and non-academic fields
2. Growing and diverse audience should be able to understand the models, intuitions, and applications of various approaches

Course Presumptions

1. ML is relevant and useful in a wide range of academic and non-academic fields
2. Growing and diverse audience should be able to understand the models, intuitions, and applications of various approaches
3. Applying ML methods to real-world problems requires quantitative skills + social science reasoning

Course Prerequisites

1. Familiarity with R
2. Basic understanding of statistical regression

Course Outline

1. Supervised Learning

2. Unsupervised Learning

Course Outline

1. Supervised Learning
    1.1 Regression and Classification
    1.2 Cross-Validation
    1.3 Regularization and Feature Selection
    1.4 Random Forests, Boosting, Bagging
2. Unsupervised Learning

Course Outline

1. Supervised Learning
    1.1 Regression and Classification
    1.2 Cross-Validation
    1.3 Regularization and Feature Selection
    1.4 Random Forests, Boosting, Bagging
2. Unsupervised Learning
    2.1 Principal Component Analysis
    2.2 Clustering
    2.3 Topic Models (Text Analysis)

This Course Does Not

- Go into the technical details behind optimizing different ML algorithms

This Course Does Not

- Go into the technical details behind optimizing different ML algorithms
- Cover all ML tools or even most of them

This Course Does Not

- Go into the technical details behind optimizing different ML algorithms
- Cover all ML tools or even most of them
- Teach you to be a professional programmer

Format and Materials

Lecture

- Semi-Flipped Classroom (1/2 Lecture, 1/2 R Coding)
- Recommend R, RStudio, and RMarkdown

Materials

- Lecture Notes, Code, and Data (Blackboard)
- Discussion Board (Blackboard)
- Text: Introduction to Statistical Learning (Free Online)

Evaluation

- Problem Sets (70%):
  - 7 problem sets, approx. every 2 weeks
  - Programming in R should be submitted via R markdown (.Rnw or .Rmd)
  - Collaboration is encouraged, but write up your own
  - First problem set released **Jan 24** $\rightarrow$ due **Feb 7**

Evaluation

- Problem Sets (70%):
  - 7 problem sets, approx. every 2 weeks
  - Programming in R should be submitted via R markdown (.Rnw or .Rmd)
  - Collaboration is encouraged, but write up your own
  - First problem set released **Jan 24** $\rightarrow$ due **Feb 7**
- Final Project (30%)
  - Option 1: Replication Study
  - Option 2: Original Research Design and Prelim Results
  - Let professor know which option by Spring Break

What is Machine Learning?

What is ML?

- **Non-Technical Take:** ML involves a set of computer algorithms which ‘learn’ patterns in existing data to assist in prediction and inference.

What is ML?

- **Non-Technical Take:** ML involves a set of computer algorithms which ‘learn’ patterns in existing data to assist in prediction and inference.
- **Technical Take:** We want to build a model f that optimizes a given loss function in order to maximize model performance

2 Types of Machine Learning

1. Unsupervised Learning
2. Supervised Learning

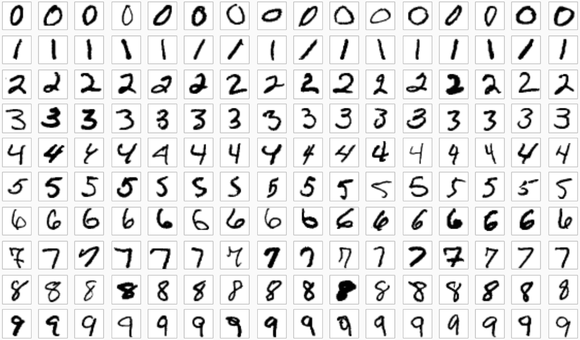
Unsupervised Learning

- Main Idea: Descriptive Data Analysis
- Common Objectives:
    - Identify meaningful groupings of the data $\rightarrow$ **clustering**
    - Simplify high-dimensional data to explain variation in as few dimensions as possible $\rightarrow$ **principal component analysis**

Unsupervised Learning

Real-World Applications:

- Stock Market Anomaly Detection (Insider Trading)
- Hand-Writing Analysis
- Measure Consumer Opinion
- Defining ‘Nationalism’ or ‘State Capacity’



Supervised Learning

- Main Idea: Learn patterns in existing data and extrapolate good predictions based on this information.

Supervised Learning

- Main Idea: Learn patterns in existing data and extrapolate good predictions based on this information.
- Real-World Applications: Predict terrorist attacks, predict covid trends, predict election results.

The Economist is analysing polling, economic and demographic data to predict America's elections in 2020

→ **Read more of our election coverage**

harts, maps and analysis of the presidential and congressional races in one place

nate House *Last updated on November 3rd*

Our final pre-election forecast is that **Joe Biden is very likely to beat Donald Trump** in the electoral college.

House	Chance of winning the electoral college	Chance of winning the most votes	Predicted range of electoral college votes (270 to win)
 **Joe Biden** Democrat	**better than 19 in 20** or 97%	**better than 19 in 20** or >99%	**259-415**
 **Donald Trump** Republican	**less than 1 in 20** or 3%	**less than 1 in 20** or <1%	**123-279**

The probability of an electoral-college tie is <1%

Supervised Learning

- Common Objective: Learn relationship between outcome variable (Y) and input variables ($X = (X_1, X_2, \ldots, X_i)$) by estimating f
- Assume relationship between Y and X_i such that...

$$y = f(X) + \epsilon \tag{1}$$

  - f is fixed, but unknown function.
  - f captures information (systematic patterns) about how X affects Y
  - ϵ is “noise” in the model (error term)

Why Learn the Relationship Between X and Y?

1. Inference

2. Prediction

Why Learn the Relationship Between X and Y?

1. Inference
    1.1 Inputs and outputs readily available
    1.2 Want to understand how Y changes as $X = (X_1, X_2, \ldots, X_i)$ changes
    1.3 Better model $\rightarrow$ more interpretable
    1.4 e.g. Which factors *explain* covid cases?
2. Prediction

Why Learn the Relationship Between X and Y?

1. Inference
   1.1 Inputs and outputs readily available
   1.2 Want to understand how Y changes as $X = (X_1, X_2, \ldots, X_i)$ changes
   1.3 Better model $\rightarrow$ more interpretable
   1.4 e.g. Which factors *explain* covid cases?
2. Prediction
   2.1 Inputs are readily available, but Y is not
   2.2 Want to predict $\hat{Y} = \hat{f}(X)$
   2.3 Better model $\rightarrow$ more accurate predictions ($\hat{Y} \approx Y$)
   2.4 e.g. What factors *predict* covid cases?

An Analogy

- Inference: Why is the car running?
- Prediction: Where is the car going?

Supervised learning aims to estimate f

How do we estimate f?

1. Data Collection and Processing
    1.1 Collect a set of n data points with p predictors

Supervised learning aims to estimate f

How do we estimate f?

1. Data Collection and Processing
   1.1 Collect a set of n data points with p predictors
   1.2 Partition the data into a training (in-sample) and test (out-of-sample) set of observations

Supervised learning aims to estimate f

How do we estimate f?

1. Data Collection and Processing
   1.1 Collect a set of n data points with p predictors
   1.2 Partition the data into a training (in-sample) and test (out-of-sample) set of observations
   1.3 Select a learning algorithm to estimate f

Supervised learning aims to estimate f

How do we estimate f?

1. Data Collection and Processing
   1.1 Collect a set of n data points with p predictors
   1.2 Partition the data into a training (in-sample) and test (out-of-sample) set of observations
   1.3 Select a learning algorithm to estimate f
2. Model Training and Assessment
   2.1 Use training data to fit prediction function estimate $\hat{f}$

Supervised learning aims to estimate f

How do we estimate f?

1. Data Collection and Processing
   1.1 Collect a set of n data points with p predictors
   1.2 Partition the data into a training (in-sample) and test (out-of-sample) set of observations
   1.3 Select a learning algorithm to estimate f
2. Model Training and Assessment
   2.1 Use training data to fit prediction function estimate $\hat{f}$
   2.2 Use $\hat{f}$ to predict outcomes $\hat{f}(x)$ using test set inputs

Supervised learning aims to estimate f

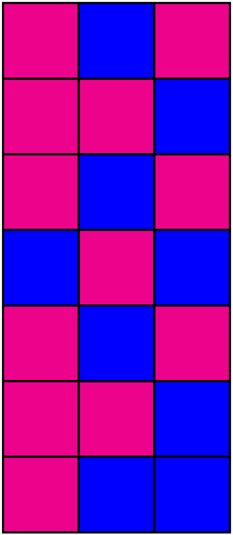
How do we estimate f?

1. Data Collection and Processing
   1.1 Collect a set of n data points with p predictors
   1.2 Partition the data into a training (in-sample) and test (out-of-sample) set of observations
   1.3 Select a learning algorithm to estimate f
2. Model Training and Assessment
   2.1 Use training data to fit prediction function estimate $\hat{f}$
   2.2 Use $\hat{f}$ to predict outcomes $\hat{f}(x)$ using test set inputs
   2.3 Evaluate whether $\hat{f}$ good model by comparing predicted response $\hat{f}(x)$ (aka $\hat{Y}$) with true response Y

Data Set

Data: $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

Test and Training Set



Partition Data into Test and Training Set

Figure 2: Training Set

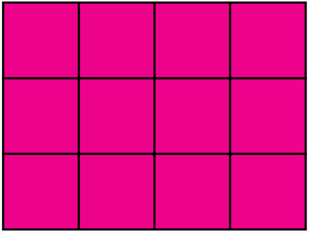
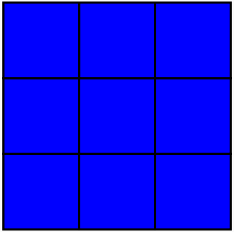
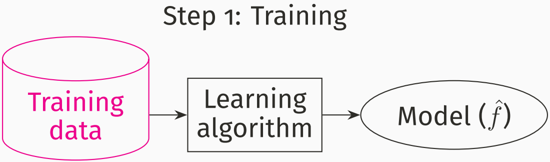


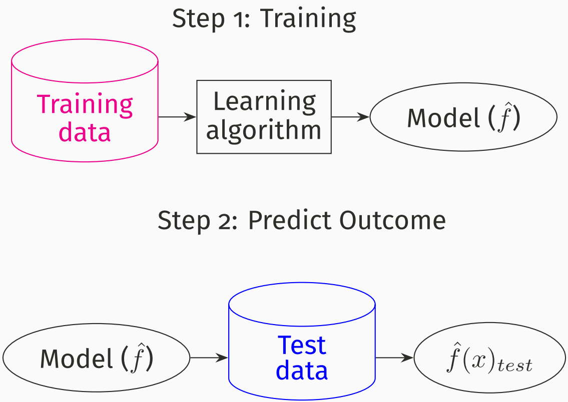
Figure 3: Test Set



Train Model $\hat{f}$ then Test Accuracy of Predictions

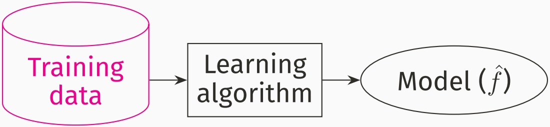


Train Model $\hat{f}$ then Test Accuracy of Predictions

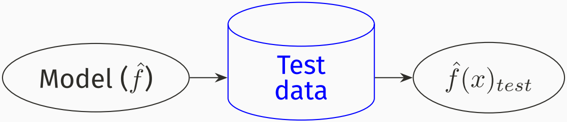


Train Model $\hat{f}$ then Test Accuracy of Predictions

Step 1: Training



Step 2: Predict Outcome



Step 3: Evaluate Test Predictions

$$\boxed{\hat{f}(x)_{test} \approx Y_{test}?}$$

Preview: Model Assessment and Selection

New Terminology So Far

Training Data: $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$
Test Data: $(x'_1, y'_1), (x'_2, y'_2), \ldots (x_m, y_m)$
Loss Function: Optimization Function to Maximize Model Performance
Prediction Function Estimate: $\hat{f}$

Lingering Questions

- How do I evaluate test predictions?
- How do I choose learning algorithm?

Lingering Questions

- How do I evaluate test predictions? Model Assessment
- How do I choose learning algorithm?

Lingering Questions

- How do I evaluate test predictions? Model Assessment
- How do I choose learning algorithm? Model Selection

Preview: Model Assessment

Question: How do we know if $\hat{f}$ is a good estimate?

Preview: Model Assessment

Question: How do we know if $\hat{f}$ is a good estimate?
Main Idea: $\hat{f}$ is good if it predicts well

Preview: Model Assessment

Question: How do we know if $\hat{f}$ is a good estimate?
Main Idea: $\hat{f}$ is good if it predicts well

- If (x_m, y_m) is an out-of-sample (not used in training) datapoint, then $\hat{f}(x_m)$ and y_m should be close
- Popular measure of closeness is mean squared error (MSE) $(y_m - \hat{f}(x_m))^2$

Assess Model Using Test Mean Squared Error

Given many test set datapoints $\{(x_i', y_i'); i = 1, \ldots, m\}$, estimate model performance using loss function known as test mean squared error:

$$\frac{1}{m}\sum_{i=1}^{m}(y_i' - f(\hat{x}_i'))^2 \tag{2}$$

Why Not Use Training MSE?

If you don't have extra test data, why not assess model performance using training mean squared error?

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\hat{x}_i))^2 \tag{3}$$

Why Not Use Training MSE?

If you don't have extra test data, why not assess model performance using training mean squared error?

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\hat{x}_i))^2 \tag{3}$$

Answer: Model will always fit training data well, but tells us nothing about if it fits test data well. Small training error does not imply small test error.

How to Maximize Model Performance?

- Better model $\rightarrow$ smaller test MSE

How to Maximize Model Performance?

- Better model $\rightarrow$ smaller test MSE
- Select modeling method (learning algorithm) to minimize average test error:

$$E(y_m - f(x_m))^2$$

How to Maximize Model Performance?

- Better model $\rightarrow$ smaller test MSE
- Select modeling method (learning algorithm) to minimize average test error:

$$E(y_m - f(x_m))^2 \tag{4}$$

- Best Model will achieve:
    - Low Variance
    - Low Bias

How to Maximize Model Performance?

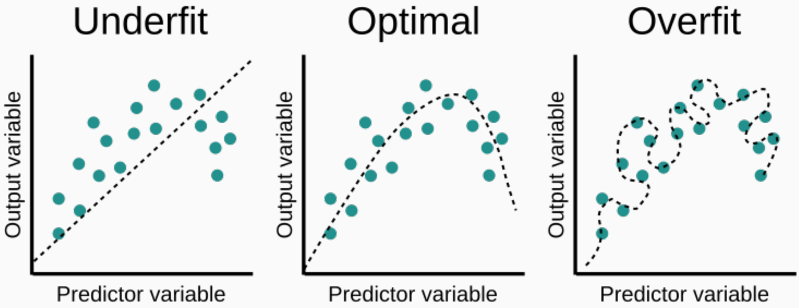
- Better model $\rightarrow$ smaller test MSE
- Select modeling method (learning algorithm) to minimize average test error:

$$E(y_m - f(x_m))^2 \tag{4}$$

- Best Model will achieve:
    - Low Variance
    - Low Bias
- **Problem:** Easier said than done.

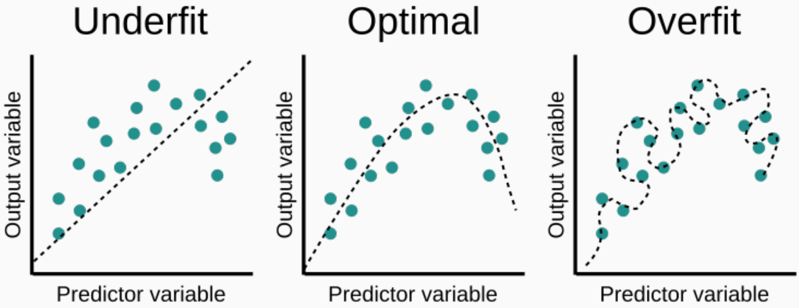
Bias-Variance Trade-Off

- Bias-Variance Trade-Off: Models tend to result in either (1) low variance and high bias (under-fitting) or (2) high variance and low bias (over-fitting).



Bias-Variance Trade-Off

- Bias-Variance Trade-Off: Models tend to result in either (1) low variance and high bias (under-fitting) or (2) high variance and low bias (over-fitting).



- A central ML challenge is finding a method that minimizes *both* variance and bias.
- Rule of Thumb: More flexible methods will result in higher variance, but lower bias.

Preview: Model Selection

Supervised learning algorithms fall into 2 classes:

1. Parametric

2. Non-Parametric

Preview: Model Selection

Supervised learning algorithms fall into 2 classes:

1. Parametric
    1.1 More rigid $\rightarrow$ low variance
    1.2 Assumes f has fixed form with fixed number of parameters $(\beta_1, \ldots \beta_p)$
    1.3 Estimating f $\rightarrow$ estimating parameters
    1.4 Ex. Linear Regression Model

$$\hat{f}(X) = X_1\beta_1 + \cdots + X_p\beta_p \tag{5}$$

2. Non-Parametric

Preview: Model Selection

Supervised learning algorithms fall into 2 classes:

1. Parametric
   1.1 More rigid $\rightarrow$ low variance
   1.2 Assumes f has fixed form with fixed number of parameters $(\beta_1, \ldots \beta_p)$
   1.3 Estimating f $\rightarrow$ estimating parameters
   1.4 Ex. Linear Regression Model

$$\hat{f}(X) = X_1\beta_1 + \cdots + X_p\beta_p \tag{5}$$

2. Non-Parametric
   2.1 More flexible $\rightarrow$ low bias
   2.2 No fixed f to describe data
   2.3 $\hat{f}$ is “black box”

Conclusion

- ML aims to learn patterns and make good predictions about out-of-sample (test) data
- Best ML model minimizes test MSE
- Picking best model means optimizing bias-variance trade-off
- Non-parametric methods reduce bias, but increase variance