MACHINE LEARNING

Dr. Ir. Kurnianingsih, S.T., M.T.

kurnianingsih@polines.ac.id

What is Machine Learning?

…learning from data

Machine Learning is…

Machine Learning is…

… learning from data

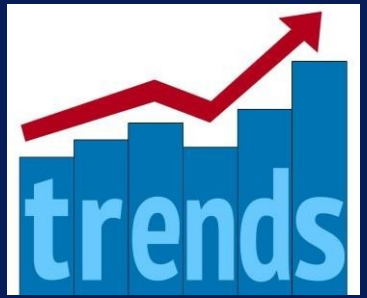
… no explicit programming

Machine Learning is…

… learning from data

… no explicit programming

… discovering hidden patterns



Machine Learning is…

… learning from data

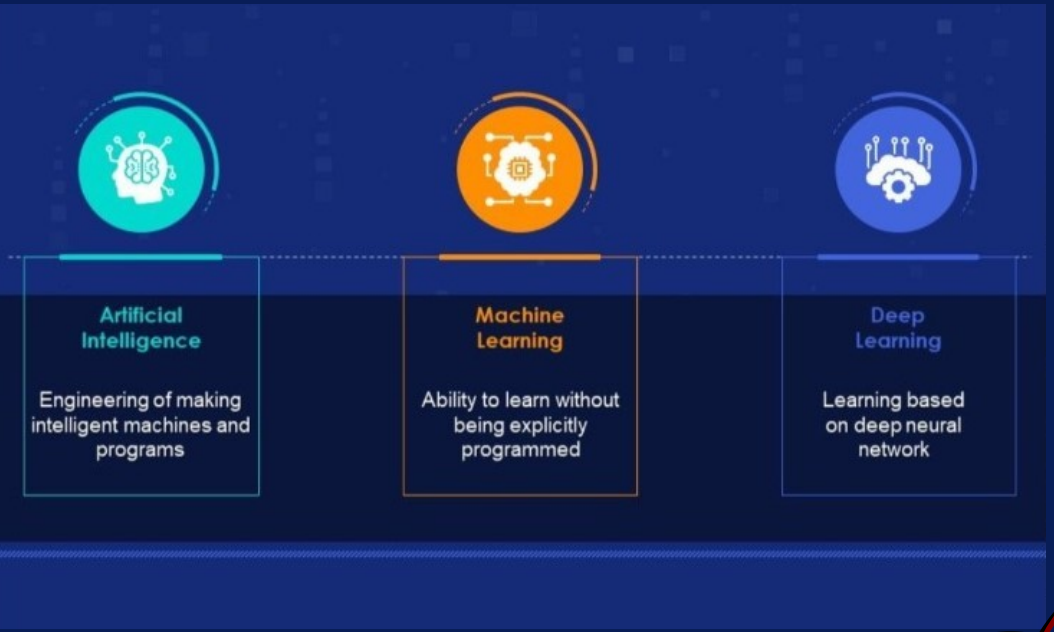
… no explicit programming

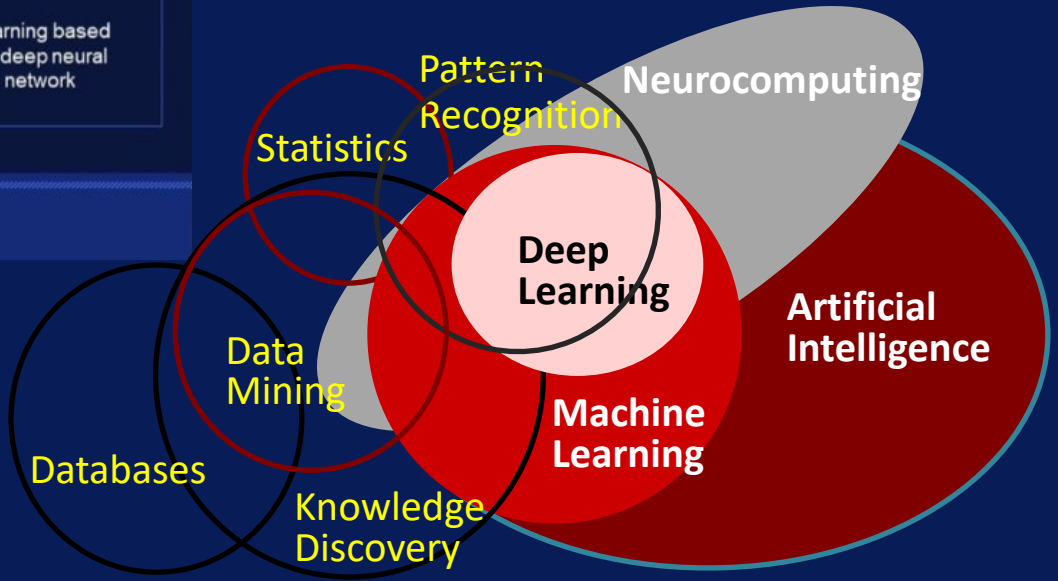
… discovering hidden patterns

… data-driven decisions

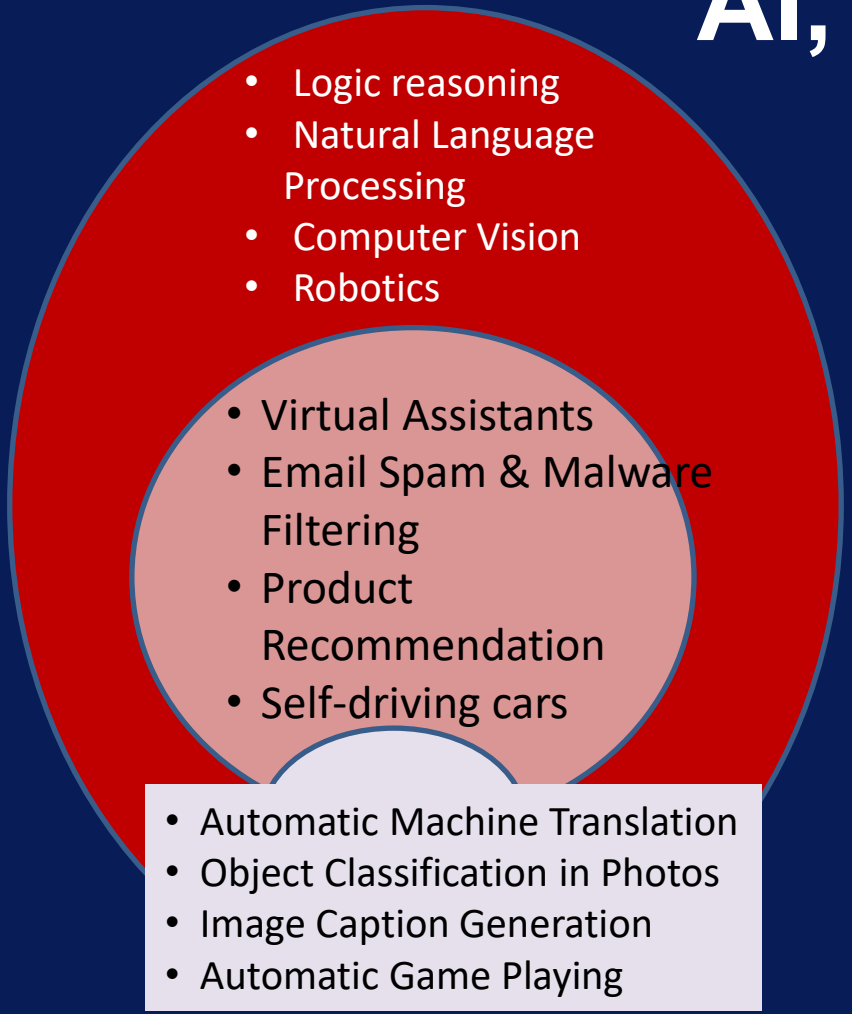
Machine learning is concerned with building systems that improve their performance on a task when given examples of ideal performance on the task or improve their performance with repeated experience on the task.

AI, ML, DL


Artificial Intelligence
Engineering of making intelligent machines and programs
Machine Learning
Ability to learn without being explicitly programmed
Deep Learning
Learning based on deep neural network


Pattern Recognition
Neurocomputing
Statistics
Deep Learning
Artificial Intelligence
Data Mining
Machine Learning
Databases
Knowledge Discovery

AI, ML, DL



ARTIFICIAL INTELLIGENCE

- Computer systems simulate human intelligence processes, including: learning, reasoning, self-correction.
- "Cognitive computing" find solutions in complex situations where answers may be ambiguous and uncertain.
- Strong vs. Weak AI

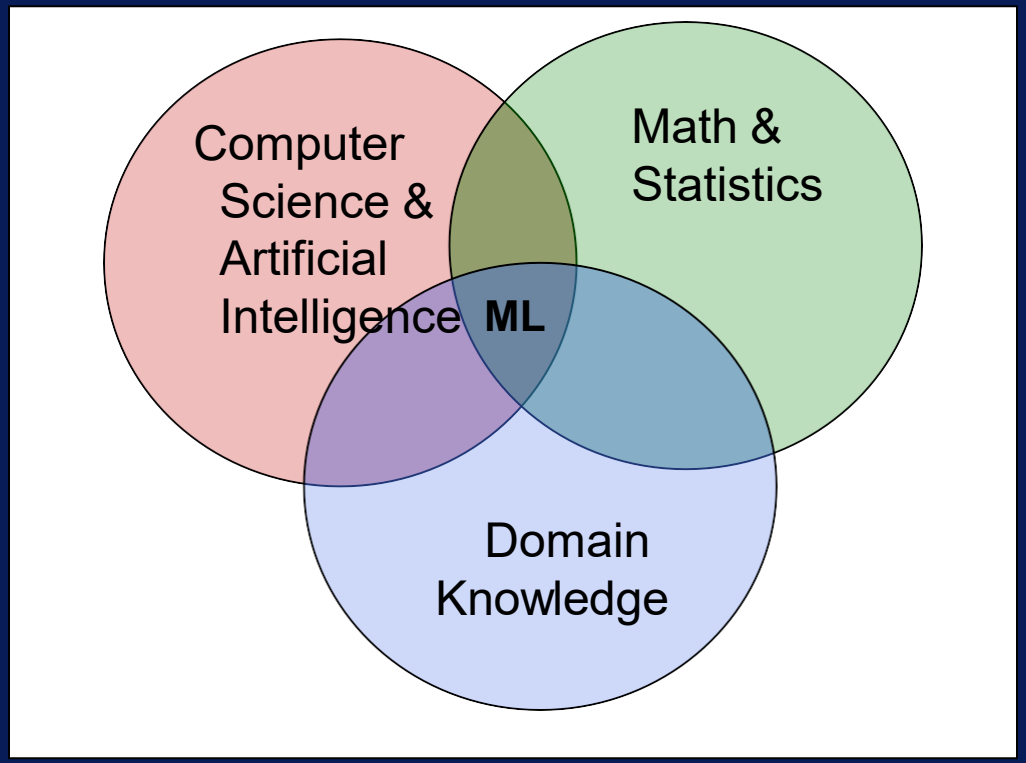
MACHINE LEARNING

- Computer systems that learn by generalizing data examples without relying on rules-based programming.
- Supervised, Unsupervised, and Reinforcement Learning

DEEP LEARNING

- Large neural networks and huge amounts of data create hierarchy of models which allows computer to learn complicated concepts by building them out of simpler ones.

Machine Learning (ML) is an Interdisciplinary Field

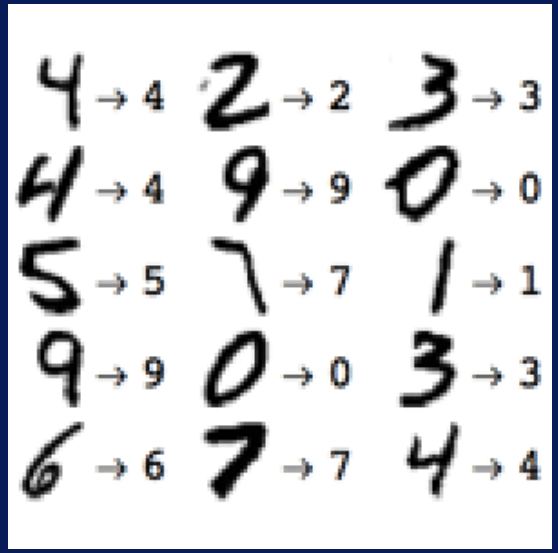


Example Application of Machine Learning

- Credit card fraud detection

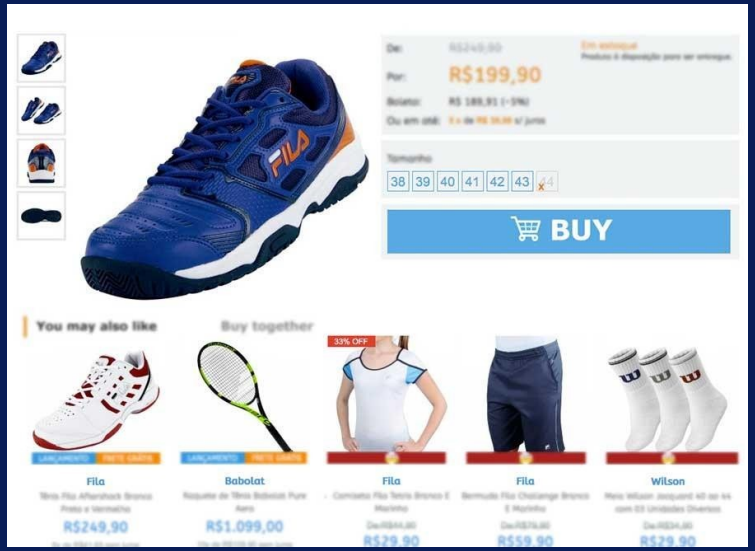
Example Application of Machine Learning

- Handwritten digit recognition



Example Application of Machine Learning

- Recommendations on websites



More Applications of Machine Learning

- Targeted ads on mobile apps
- Sentiment analysis
- Climate monitoring
- Crime pattern detection
- Drug effectiveness analysis

What's in a Name?

Machine learning

Data mining

Predictive analytics

Data science

Machine Learning Models

- Learn from data
- Discover patterns and trends
- Allow for data-driven decisions
- Used in many different applications

Machine Learning Process

After this lecture, students will be able to..

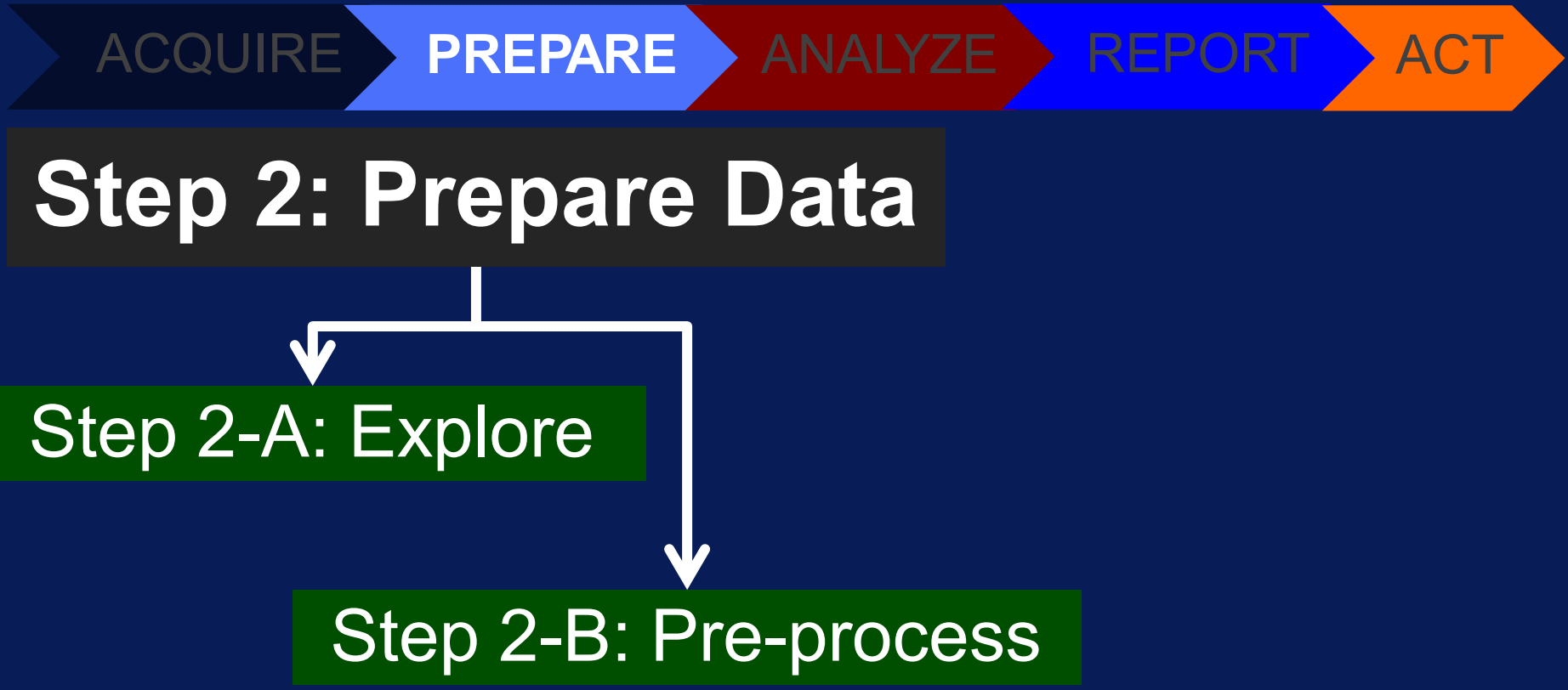
- Identify the steps in the machine learning process
- Discuss why the machine learning process is iterative

ACQUIRE PREPARE ANALYZE REPORT ACT

ACQUIRE PREPARE ANALYZE REPORT ACT

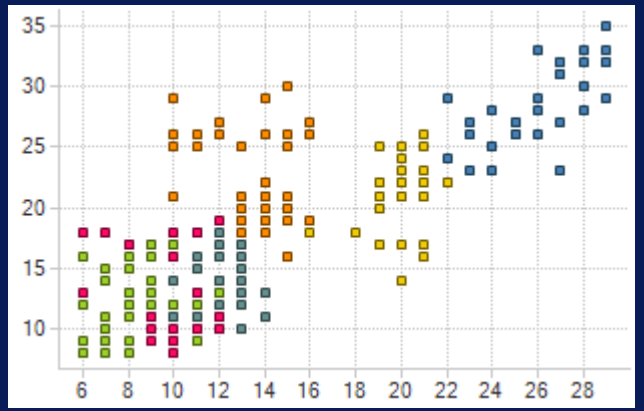
Step 1: Acquire Data

- Identify data sources
- Collect data
- Integrate data

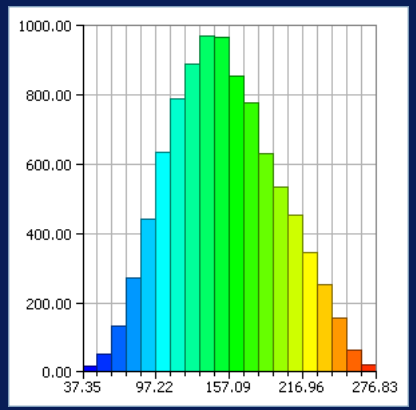

ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Step 2: Prepare Data
Step 2-A: Explore
Step 2-B: Pre-process

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

Step 2-A: Explore Data



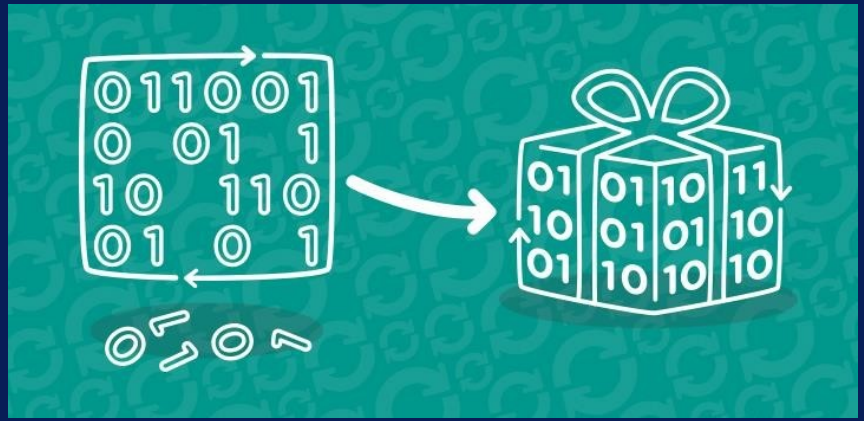
Understand nature of data



Preliminary analysis

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

Step 2-B: Pre-process Data



Clean | Select | Transform

ACQUIRE | PREPARE | **ANALYZE** | REPORT | ACT

Step 3: Analyze Data

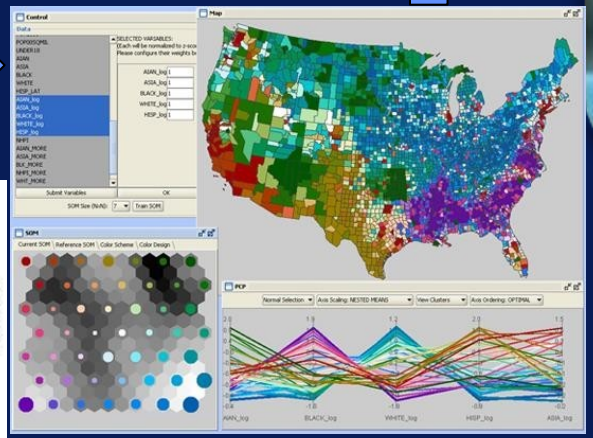
Select analytical techniques

Build models

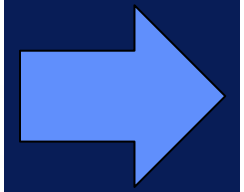
Assess results

ACQUIRE PREPARE ANALYZE **REPORT** ACT

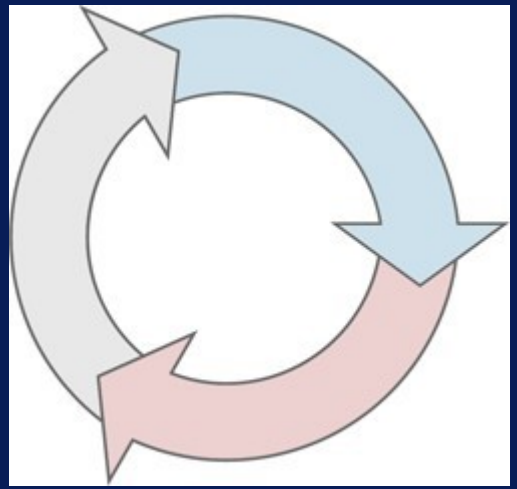
Step 4: Communicate Results



Step 5: Apply Results

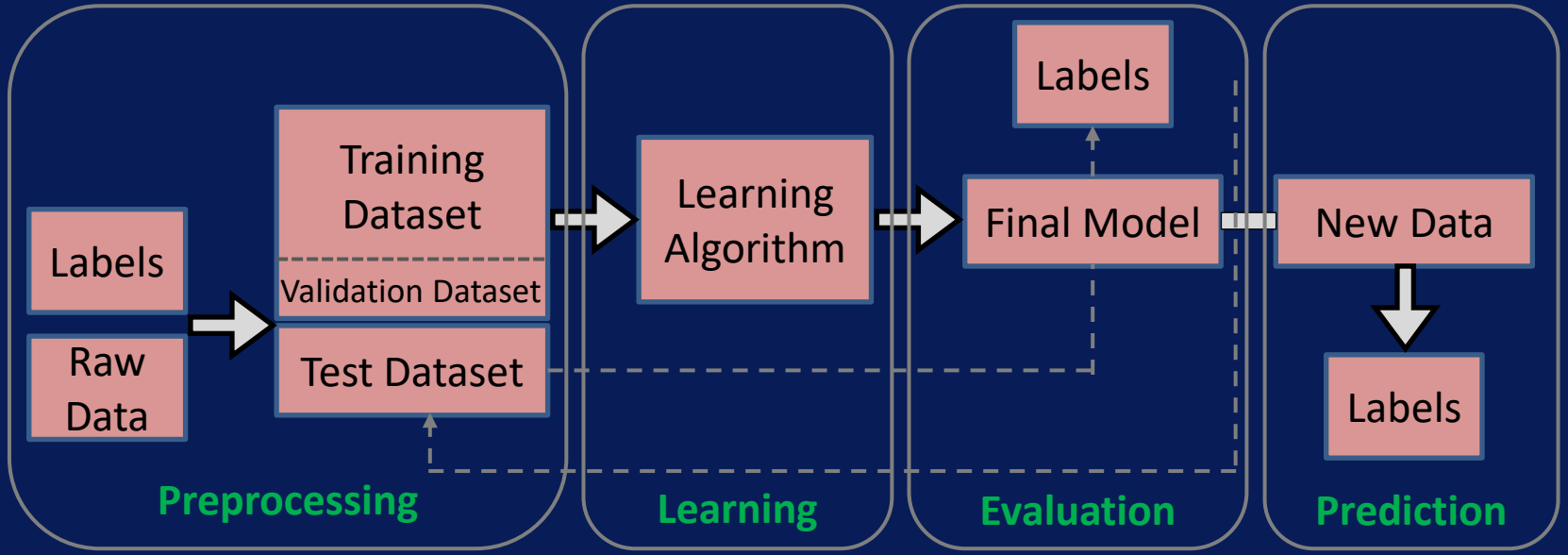


ACQUIRE PREPARE ANALYZE REPORT ACT



Iterative process

Roadmap for Building Machine Learning System



- Feature Extraction & Scaling
- Feature Selection
- Dimensionality Reduction
- Sampling

- Model Selection
- Cross-Validation
- Performance Metrics
- Hyperparameter Optimization

Categories of Machine Learning Techniques

After this lecture, students will be able to..

- Describe the main categories of machine learning techniques
- Summarize how supervised learning differs from unsupervised learning

Categories of Machine Learning Techniques

- Classification
- Regression
- Cluster Analysis
- Association Analysis

Classification

Goal: Predict category

Sunny

Windy

Rainy

Cloudy

Image source: http://www.davidson.k12.nc.us/parents___students/inclement_weather

Classification Examples

- Classify tumor as benign or malignant
- Predict if it will rain tomorrow
- Determine if loan application is high-, medium-, or low-risk
- Identify sentiment as positive, negative, or neutral

Regression

Goal: Predict numeric value

Regression Examples

- Estimate demand for a product based on time of year
- Predict score on a test
- Determine likelihood of drug effectiveness for patient
- Predict amount of rain

Cluster Analysis

Goal: Organize similar items into groups.

Image source: http://www.monetate.com/blog/the-intrinsic-value-of-customer-segmentation

Cluster Analysis Examples

- Identify areas of similar topography (desert, grass, etc.)
- Categorize different types of tissues from medical images
- Determine different groups of weather patterns
- Discover crime hot spots

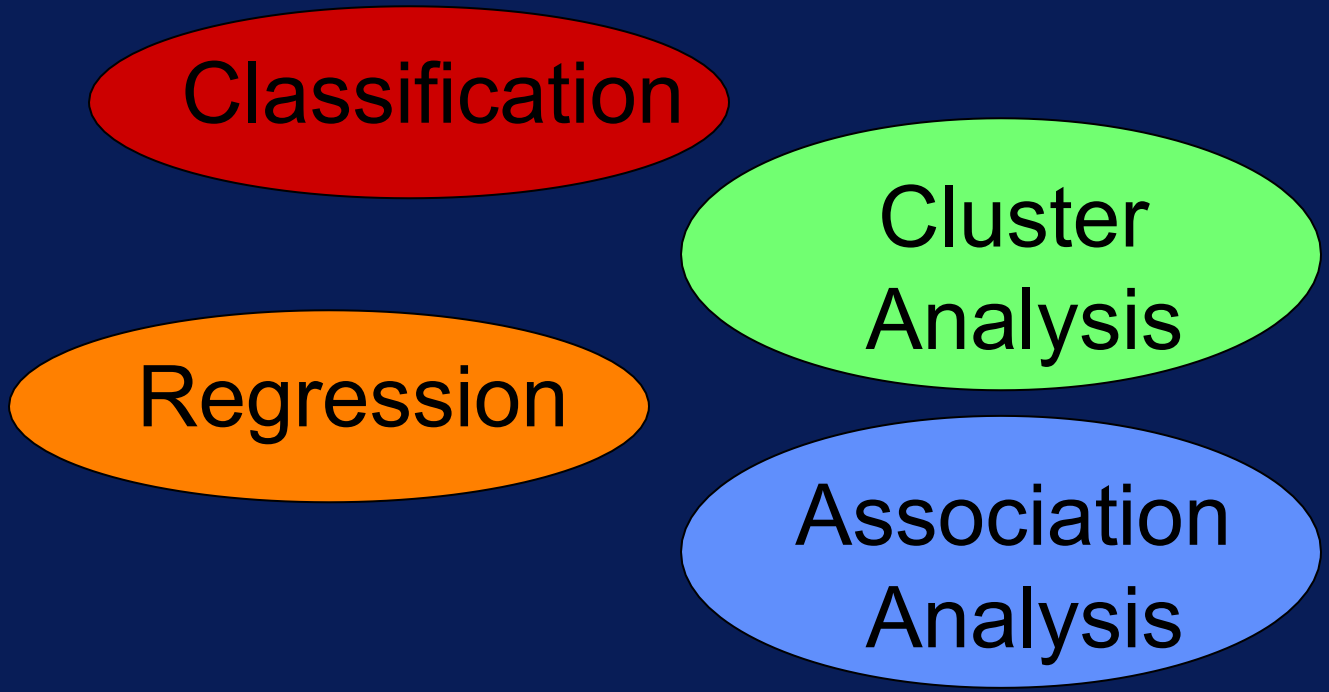
Association Analysis

Goal: Find rules to capture associations between items.

Association Analysis Examples

- Recommend items based on purchase/browsing history
- Have sales on related items often purchased together
- Identify web pages accessed together

Categories of Machine Learning Techniques



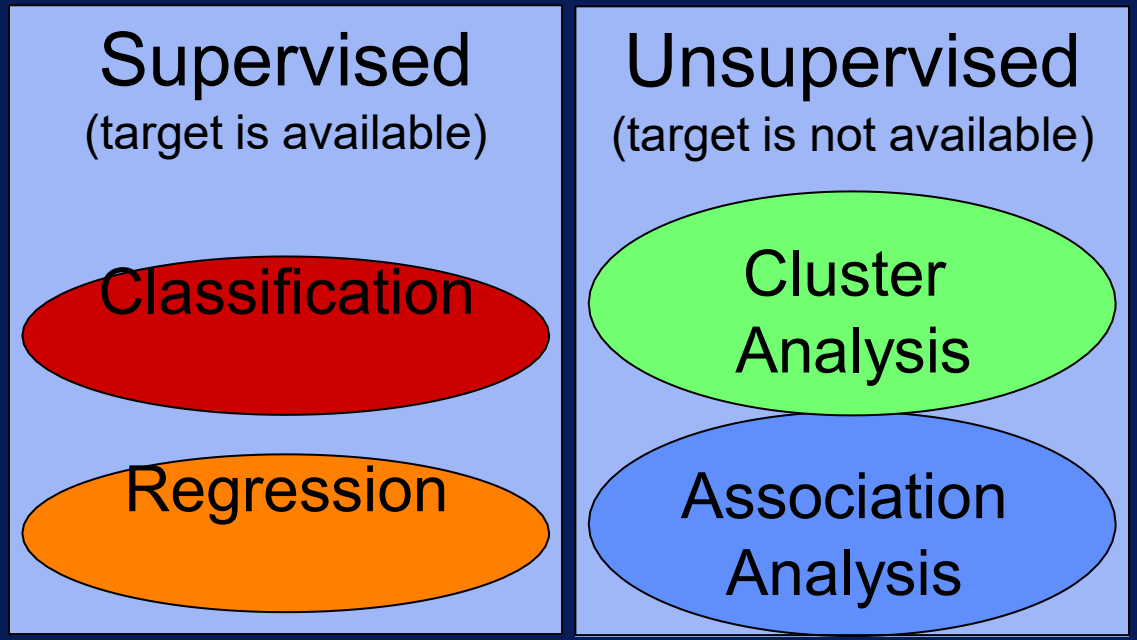
Supervised vs. Unsupervised

- **Supervised Approaches**
  - Target (what model is predicting) is provided
  - ‘Labeled’ data
  - Classification & regression are supervised.

Supervised vs. Unsupervised

- **Unsupervised Approaches**
  - Target is unknown or unavailable
  - ‘unlabeled’ data
  - Cluster analysis & association analysis are unsupervised.

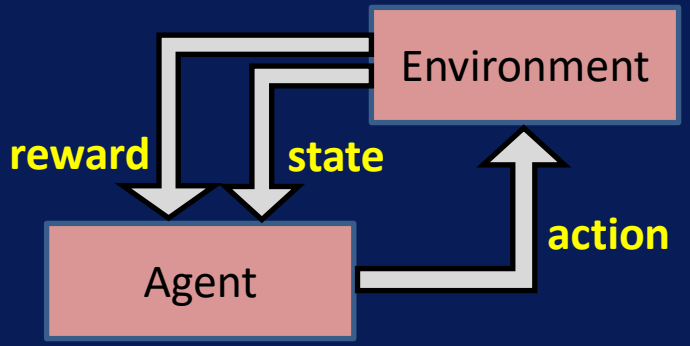
Categories of Machine Learning Techniques



Reinforcement Learning

- Employ a system (*agent*) that improves its performance based on interactions with the *environment*.
- Agent uses reinforcement learning, through interaction with the environment, to learn a series of actions (inputs) that maximizes the reward signal, measured by a reward function.

Example: chess engine

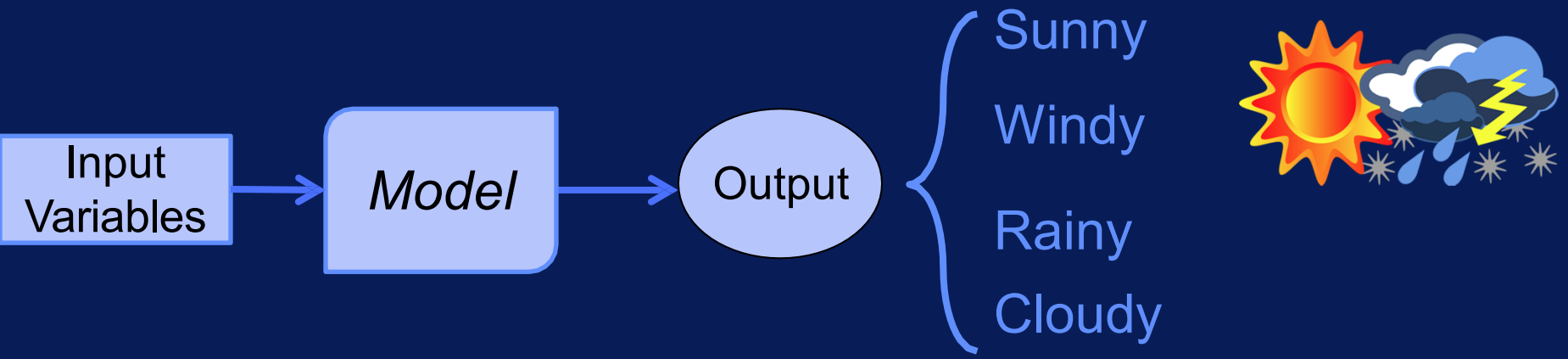


Regression

After this lecture, students will be able to..

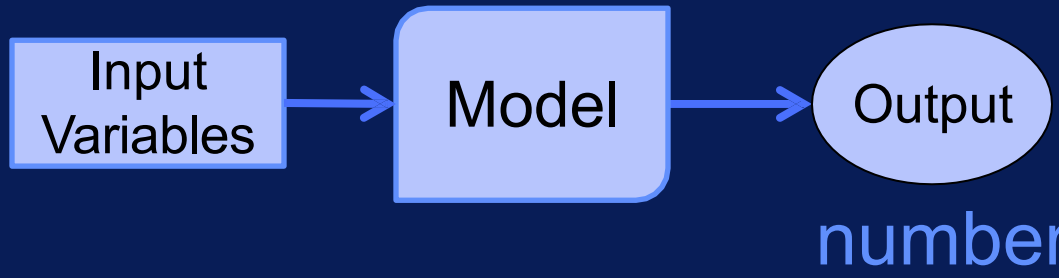
- Define what regression is
- Explain the difference between regression and classification
- Name some applications of regression

Classification Review



Classification:
Given input variables,
predict category

Regression

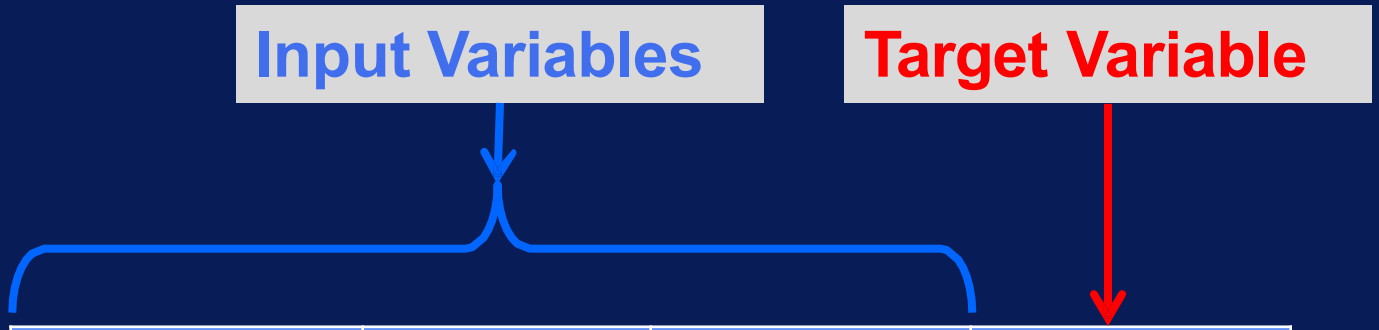


Regression:
Given input variables,
predict numeric value

Regression Examples

- **Forecast** high temperature for next day
- **Estimate** average house price for a region
- **Determine** demand for a new product
- **Predict** power usage

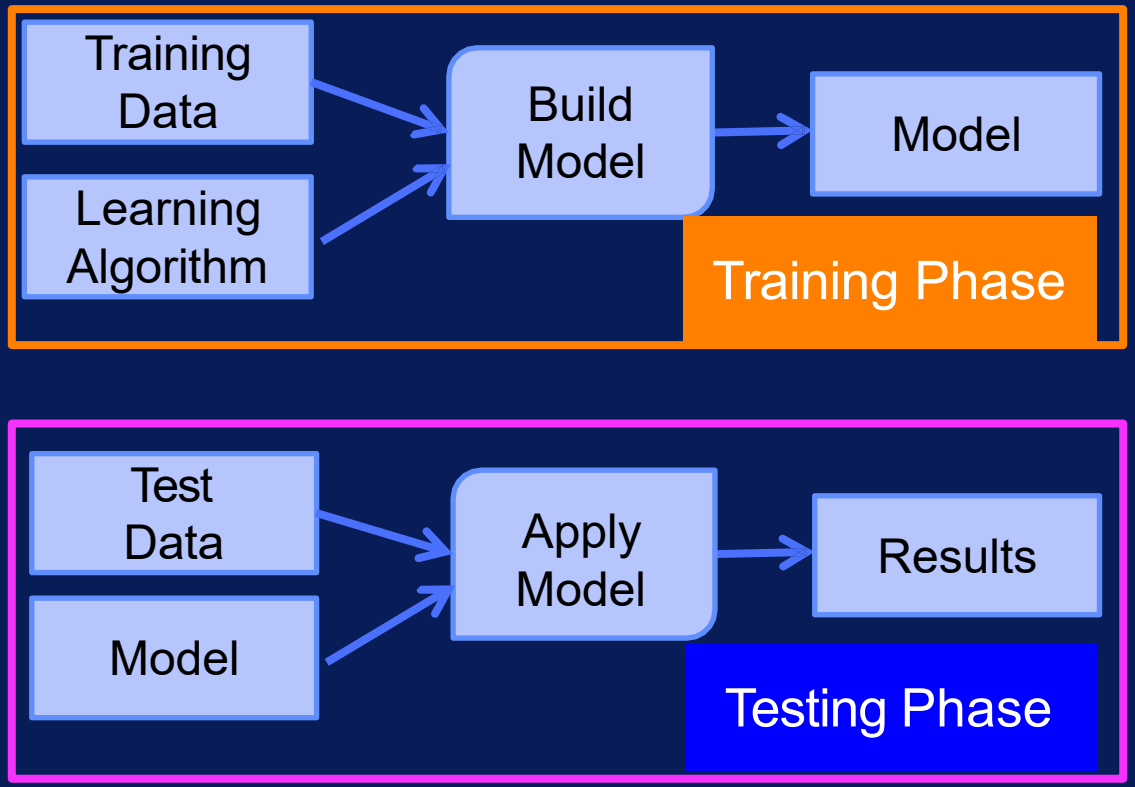
Regression is Supervised



Today's High	Today's Low	Month	Tomorrow's High
79	64	July	81
60	45	October	58
68	49	May	65
57	47	January	54

Target is provided

Building vs. Applying Model

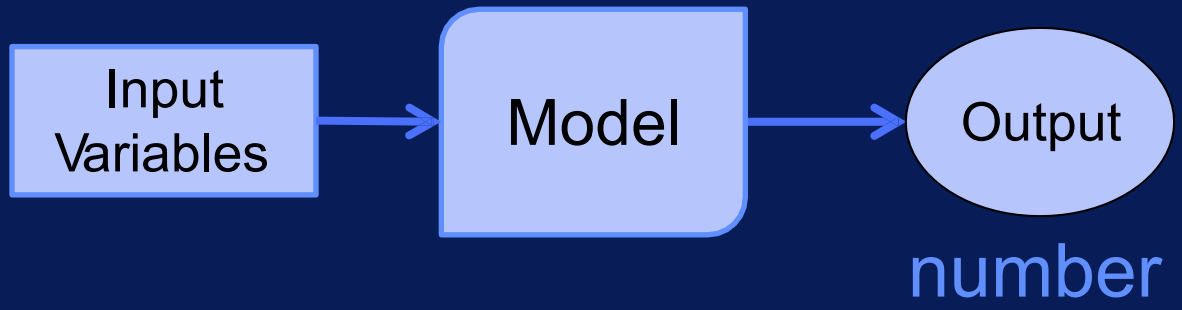


Datasets

Training Data	Validation Data	Test Data
Adjust model parameters	Determine when to stop training (avoid overfitting) Estimate generalization performance	Evaluate performance on new data

Regression Main Points

- Predict number from input variables
- Regression is a supervised task
- Target variable is numerical

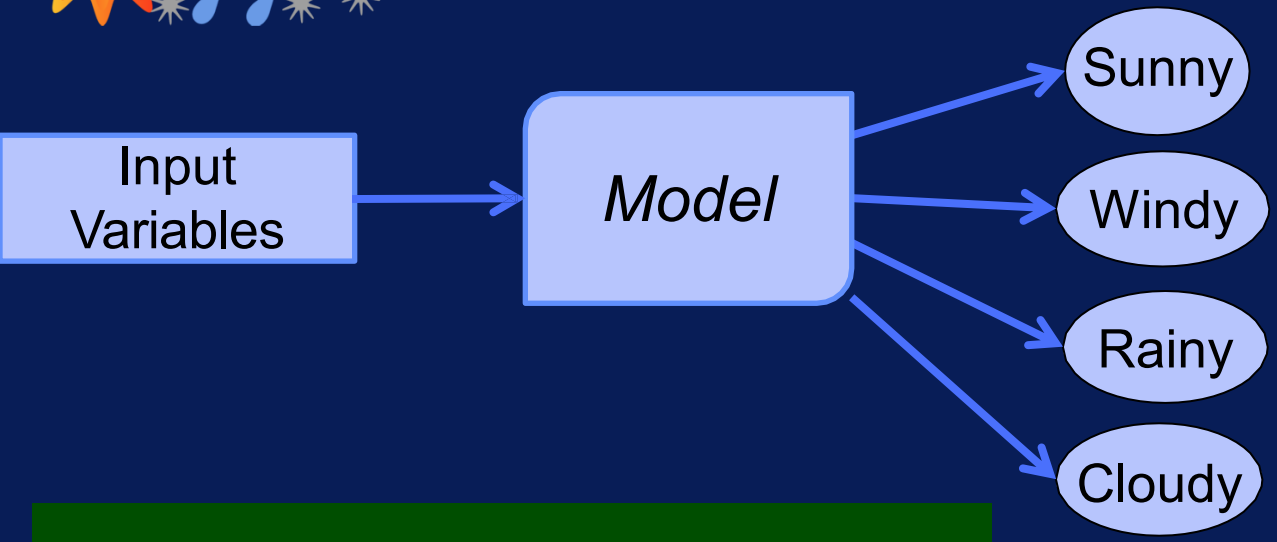


Classification Overview

After this lecture, students will be able to..

- Define what classification is
- Discuss whether classification is supervised or unsupervised
- Describe how binomial classification differs from multinomial classification

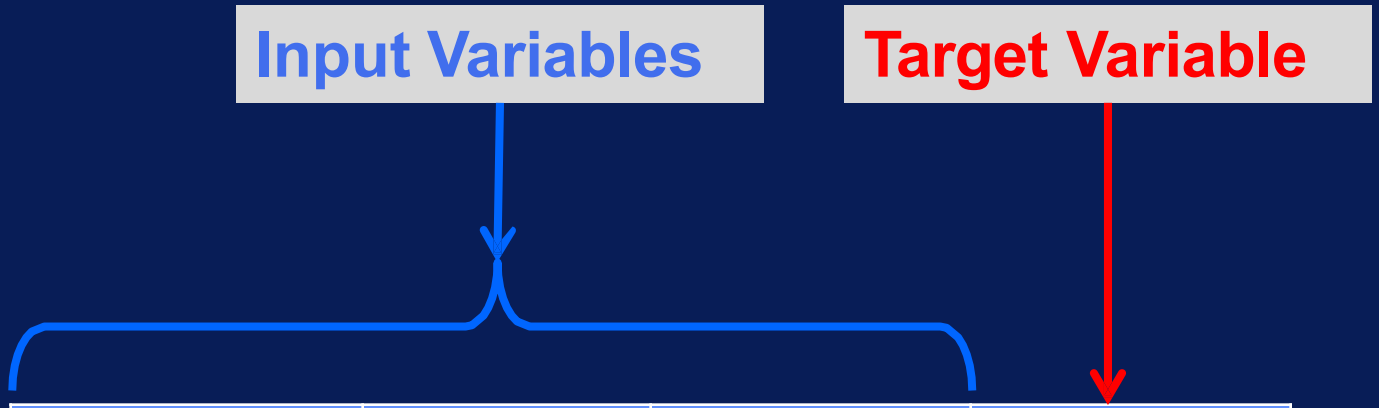
Classification



Target variable is categorical

Goal:
Given input variables, predict category

Data for Classification



Temperature	Humidity	Wind Speed	Weather
79	48	2.7	Sunny
60	80	3.8	Rainy
68	45	17.9	Windy
57	77	4.2	Cloudy

Classification is Supervised

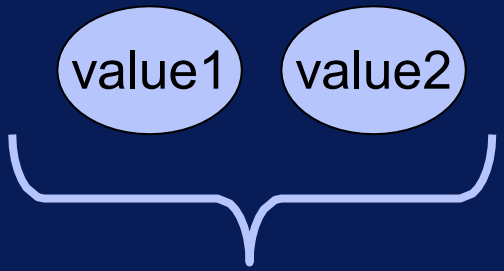
Target | Label | Output

Class Variable | Class | Category

Temperature	Humidity	Wind Speed	Weather
79	48	2.7	Sunny
60	80	3.8	Rainy
68	45	17.9	Windy
57	77	4.2	Cloudy

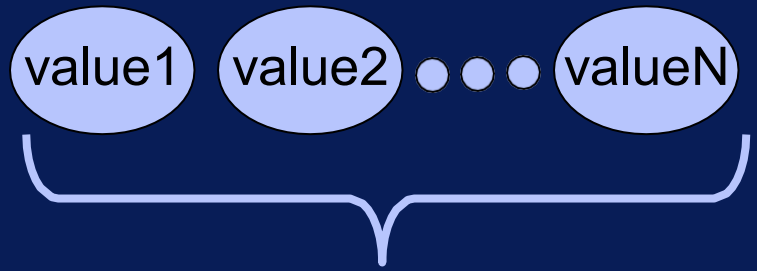
Types of Classification

Binary Classification



Target has two values

Multi-class Classification



Target has > 2 values

Classification Examples

Binary Classification

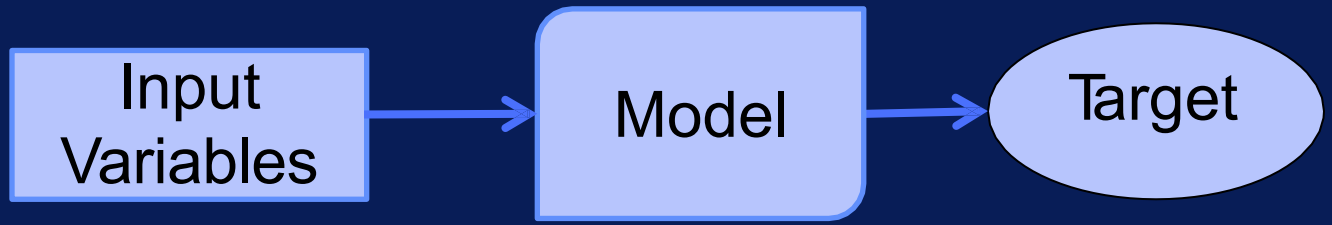
- Will it rain tomorrow or not?
- Is this transaction legitimate or fraudulent

Multi-Class Classification

- What type of product will this customer buy?
- Is this tweet positive, negative, or neutral

Classification Main Points

- Predict category from input variables
- Classification is a supervised task
- Target variable is categorical



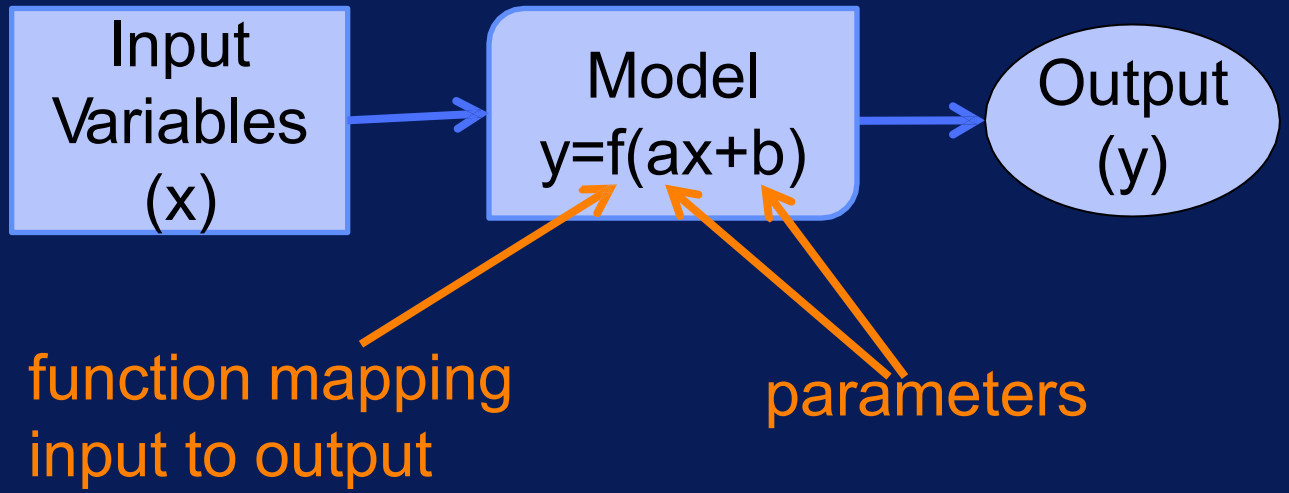
Building and Applying a Classification Model

After this lecture, students will be able to..

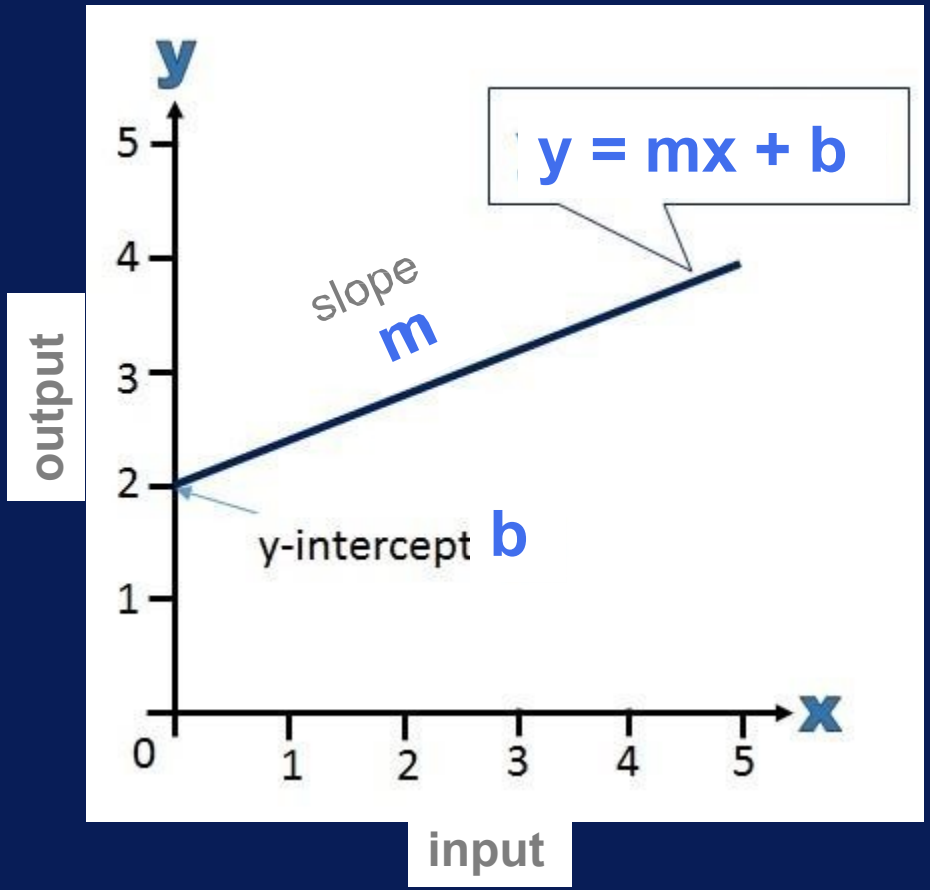
- Discuss what building a classification model means
- Explain the difference between building and applying a model
- Summarize why the parameters of a model need to be adjusted

What is a Machine Learning Model?

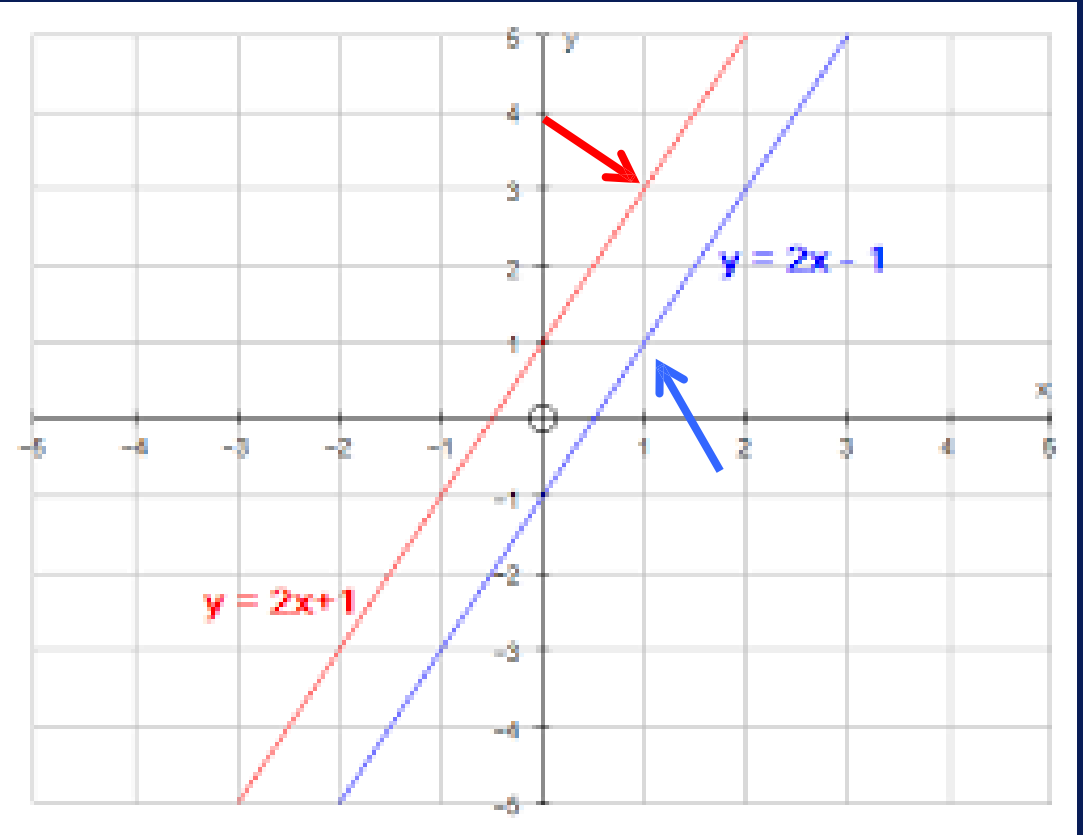
- A mathematical model with parameters that map input to output



Example of



Adjusting Model

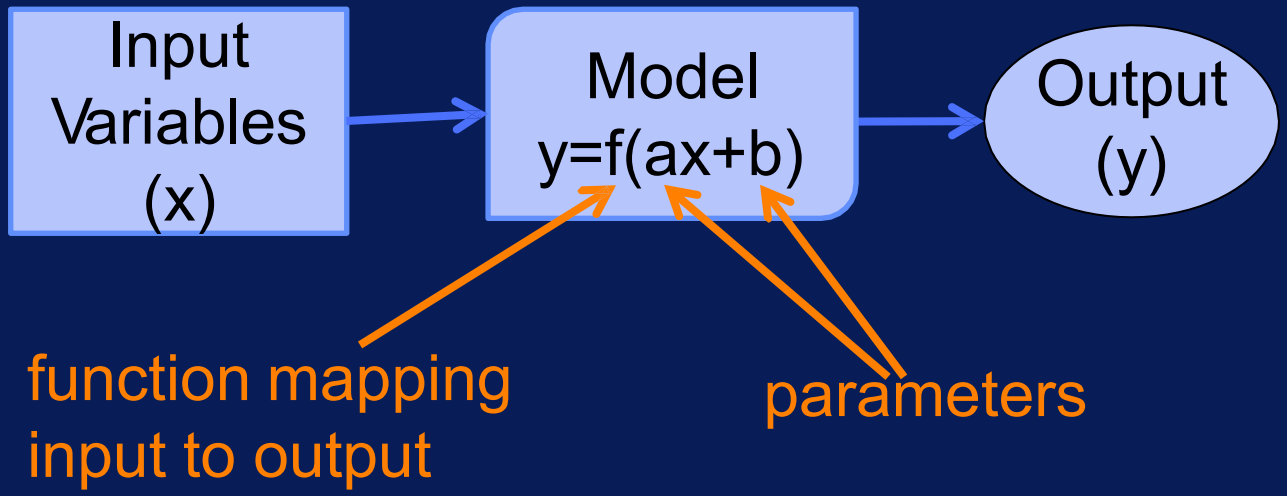


slope m = 2
y-intercept b = -1
x=1 => y=2*1-1=1

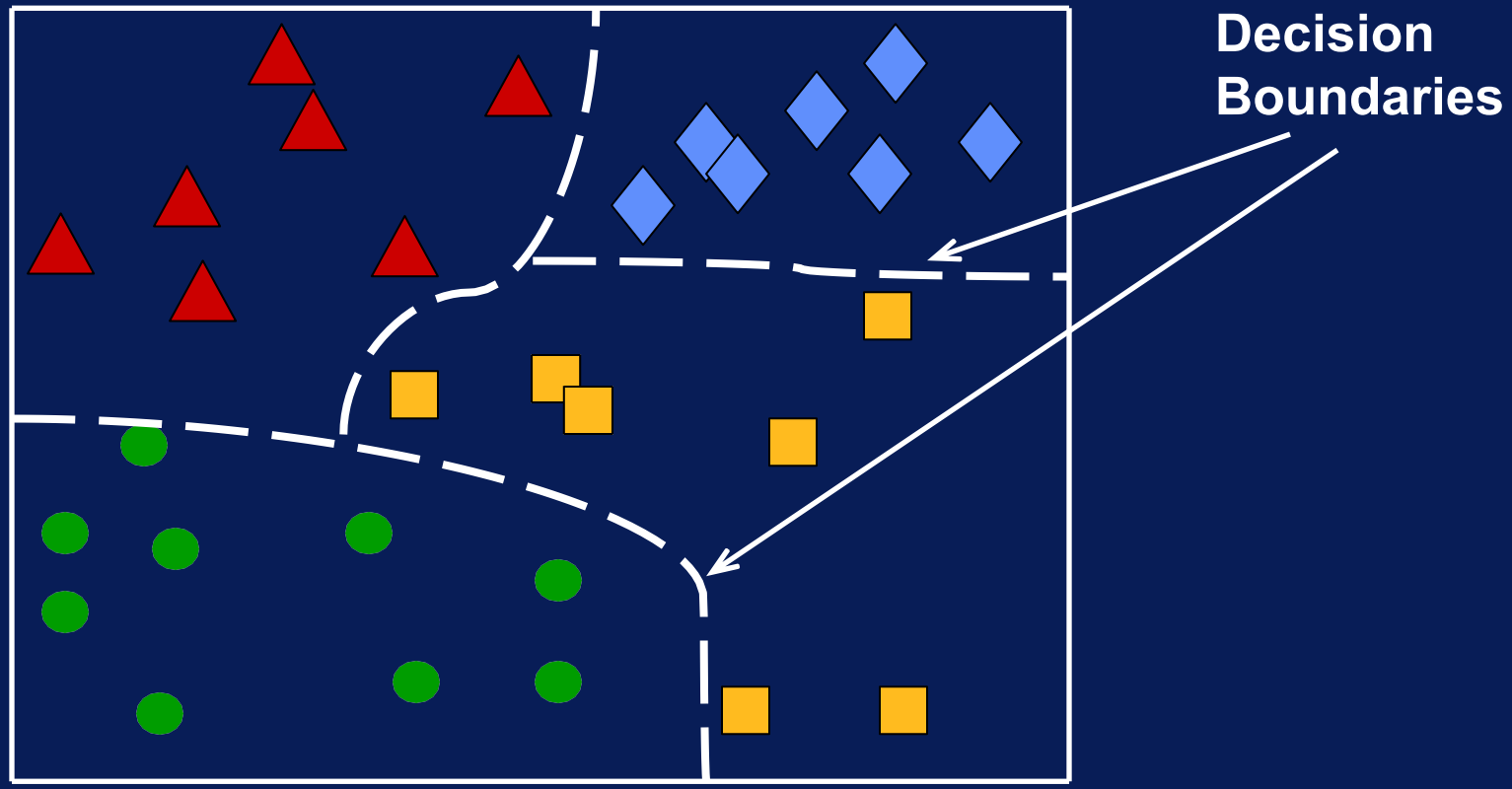
slope m = 2
y-intercept b = +1
x=1 => y=2*1+1= 3

Building Machine Learning Model

Model parameters are adjusted during model training to change input-output mapping.



Building Classification Model

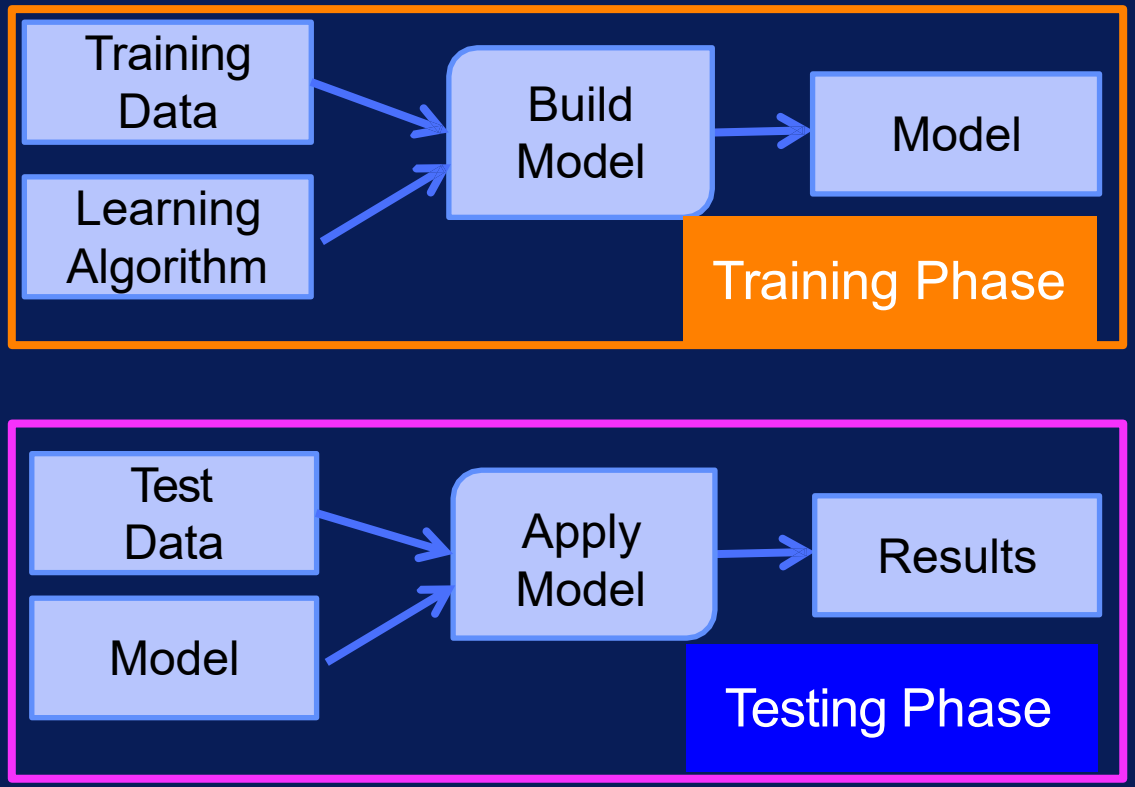


Building vs. Applying Model

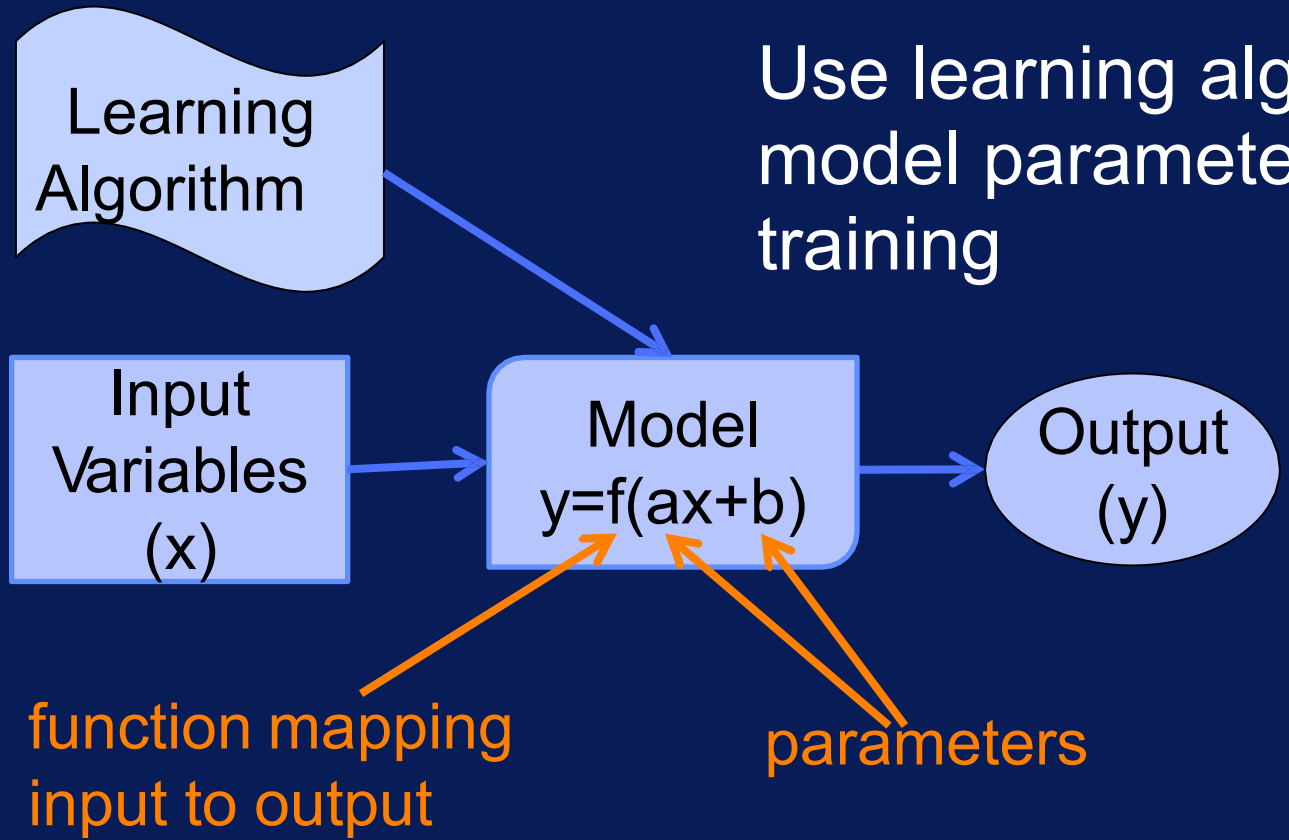
- Training Phase
  - Adjust model parameters
  - Use training data

- Testing Phase
  - Apply learned model
  - Use new data

Building vs. Applying Model



Building a Classification Model



Overfitting in Machine Learning

After this lecture, students will be able to..

- Discuss overfitting in the machine learning
- Discuss underfitting in the machine learning
- Discuss a good fit in the machine learning

Bias & Variance

Bias – error caused because the model can not represent the concept

Variance – error caused because the learning algorithm overreacts to small changes (noise) in the training data

TotalLoss = Bias + Variance (+ noise)

Overfitting in Machine Learning

- Overfitting refers to a model that models the training data too well.

- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

- Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

Underfitting in Machine Learning

- Underfitting refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
- Underfitting is often not discussed as it is easy to detect given a good performance metric.
- The remedy is to move on and try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of overfitting.

How to Limit Overfitting

2 techniques to limit overfitting:

1. Use a resampling technique to estimate model accuracy.
2. Hold back a validation dataset.

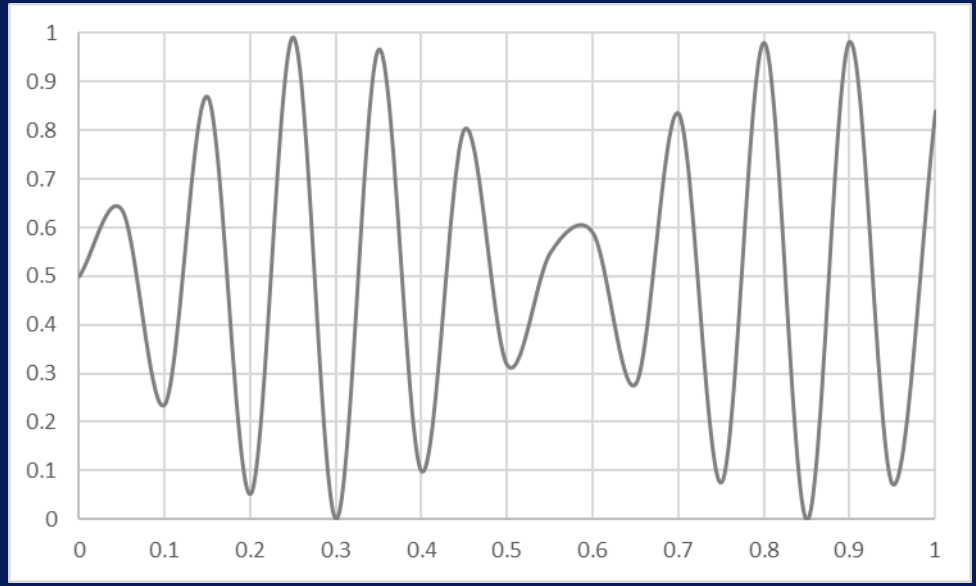
- The most popular resampling technique is k-fold cross validation. It allows you to train and test your model k-times on different subsets of training data and build up an estimate of the performance of a machine learning model on unseen data.
- A validation dataset is simply a subset of your training data that you hold back from your machine learning algorithms until the very end of your project. After you have selected and tuned your machine learning algorithms on your training dataset you can evaluate the learned models on the validation dataset to get a final objective idea of how the models might perform on unseen data.
- Using cross validation is a gold standard in applied machine learning for estimating model accuracy on unseen data. If you have the data, using a validation dataset is also an excellent practice.

Overfitting & Underfitting

OVERFITTING

Fitting the data too well

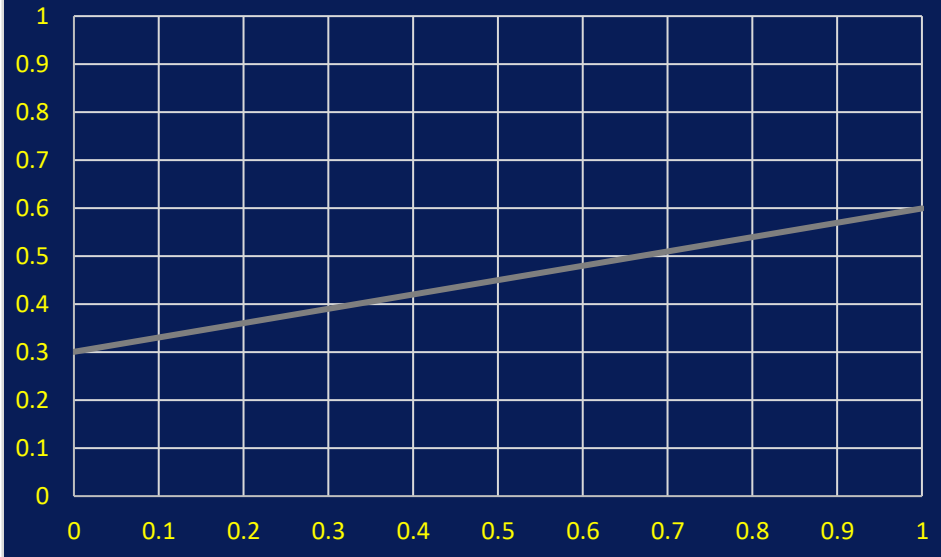
- Features are noisy / uncorrelated to concept
- Modeling process very sensitive (powerful)
- Too much search



UNDERFITTING

Learning too little of the true concept

- Features don't capture concept
- Too much bias in model
- Too little search to fit model



Modeling to Balance Under/Over Fitting

- Data
- Learning Algorithms
- Feature Sets
- Complexity of Concept
- Search and Computation

Parameter sweeps!