Advanced Analytics:
Machine Learning with R and Python

Dr. Ash Pahwa

California Institute of Technology
Center for Technology and Management Education
1200 East California Blvd., Mail Code 121-79
Pasadena, California 91125
Phone: 626.395.4042

Advanced Analytics: Machine Learning with R and Python

Lesson 1.0: Machine Learning and Predicted Analytics

Lesson 1.3

Machine Learning Techniques

Outline

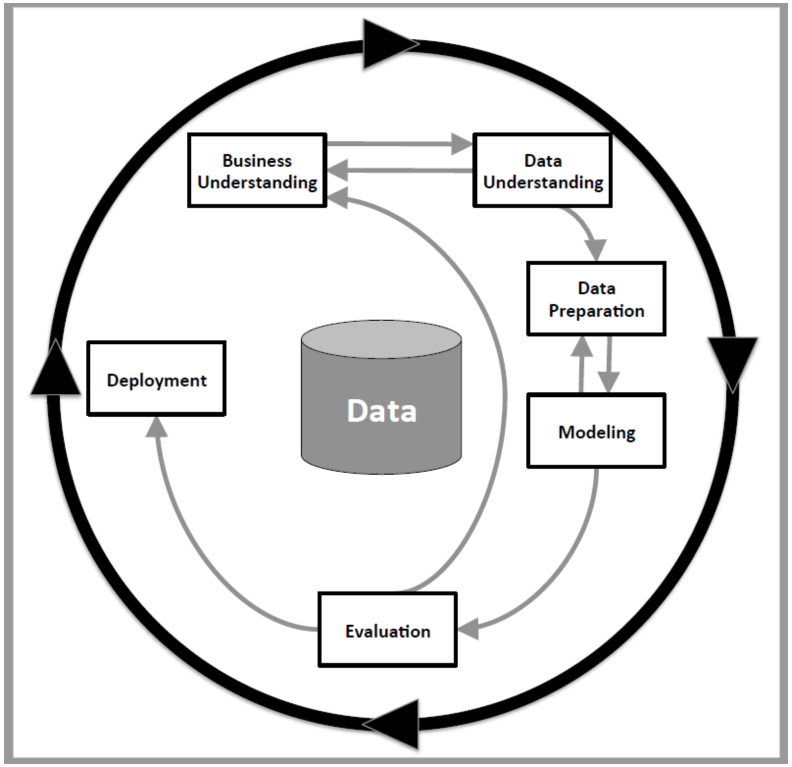
- CRISP/DM Model
- Classifying Modeling Methods
  - Response Variable:
    - Numerical or Categorical
  - Supervised or Unsupervised
  - Strategy:
    - Error Based
    - Information Based
    - Similarity Based
    - Probability Based
    - Mimicking the Human Brain (Neural networks)

CRISP/DM Process for Modeling

CRISP-DM Process for Modeling

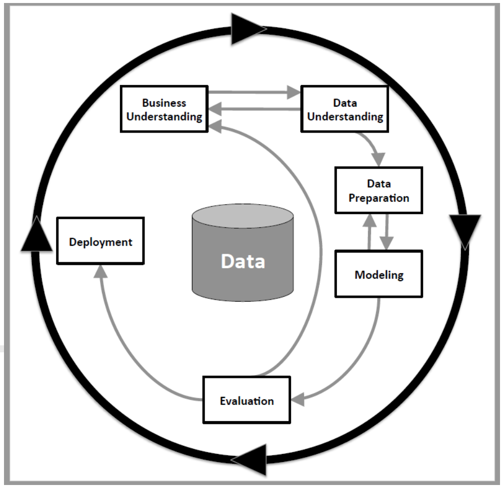
- www.crisp-dm.org
- CRoss Industry Standard Process for Data Mining

The word Data Mining can be interchanged with Predictive Analytics.



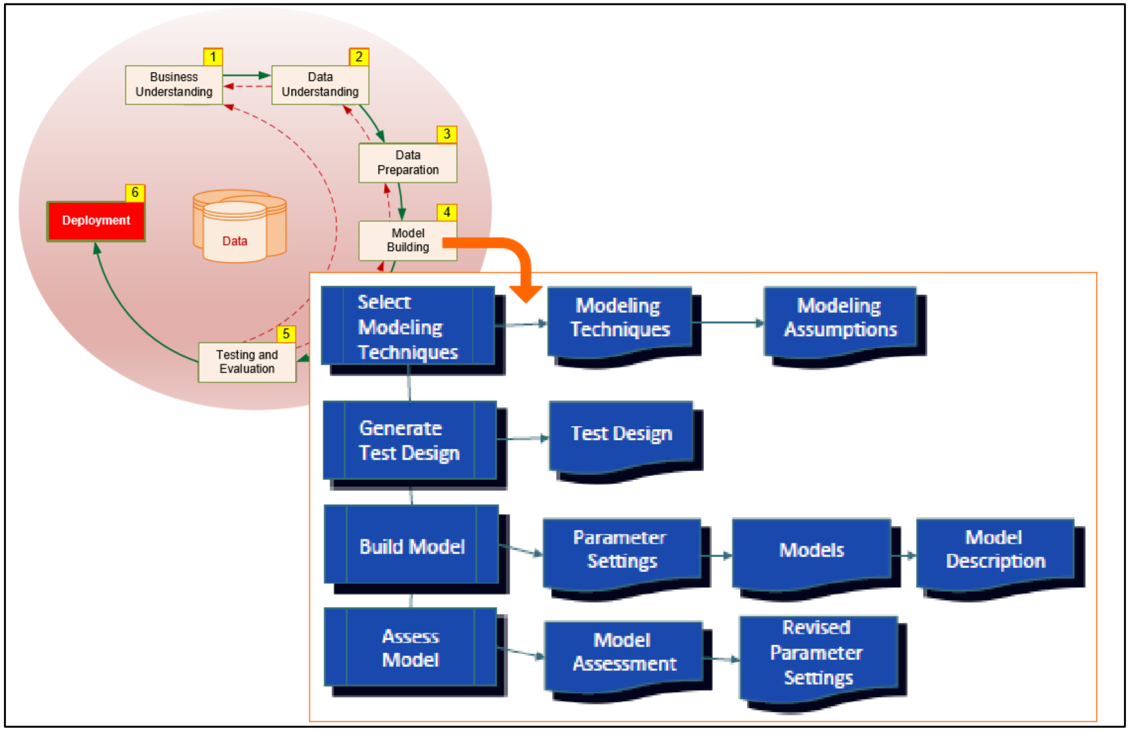
Copyright 2018 Dr. Ash Pahwa

CRISP-DM Process Model



- Step #1
  - Start with business understanding of what you want to do with data mining
- Step #2
  - Data understanding
  - Interactions between business understanding and data understanding
- Step #3
  - Data preparation
  - Interactions between data preparation and modeling
- Step #4
  - Modeling + Assessment (Evaluation)
  - Interactions between model evaluation and business understanding
- Step #5
  - Deployment of Model
- Step #6
  - Results achieved from PA should be compared with the business understanding

Modeling – CRISP-DM Step 4



Common Modeling Methods

Modeling Methods

#	Modeling Methods
1	Linear & Polynomial Regression
2	Logistic Regression
3	Discriminant Analysis
4	K Nearest Neighbor
5	Decision and Regression Trees
6	Naïve Bayes
7	Neural Networks
8	Clustering
9	Principal Component Analysis
10	Support Vector Machines
11	ARIMA : Time Series

Which ML Technique is the Best?

- Why do we consider many different techniques?
- Which one is the best?
- No one technique is the best.
- All depends upon the data.
- Some techniques will work better on certain data.

Goals of Machine Learning Application: Estimation or Classification

- Estimation – Regression modeling technique is used
  - Output is a number
    - House price
    - Product sales for next quarter
    - GNP growth for the next quarter
    - Employment
- Classification – Naïve Bayes, Decision Trees etc. modeling techniques are used
  - Output is a categorical variable
    - Sports team will win or lose
    - Email is junk or not
    - Which grade student will get
    - Tweet is positive or negative

Classification of Modeling Methods

Classification of Modeling Methods

- Response Variable
  - Numerical or Categorical
- Supervised or unsupervised
- Strategy
  - Error based learning
  - Similarity Based Learning
  - Information Based Learning
  - Probability Based Learning
  - Mimicking the Human Brain

Response Variable

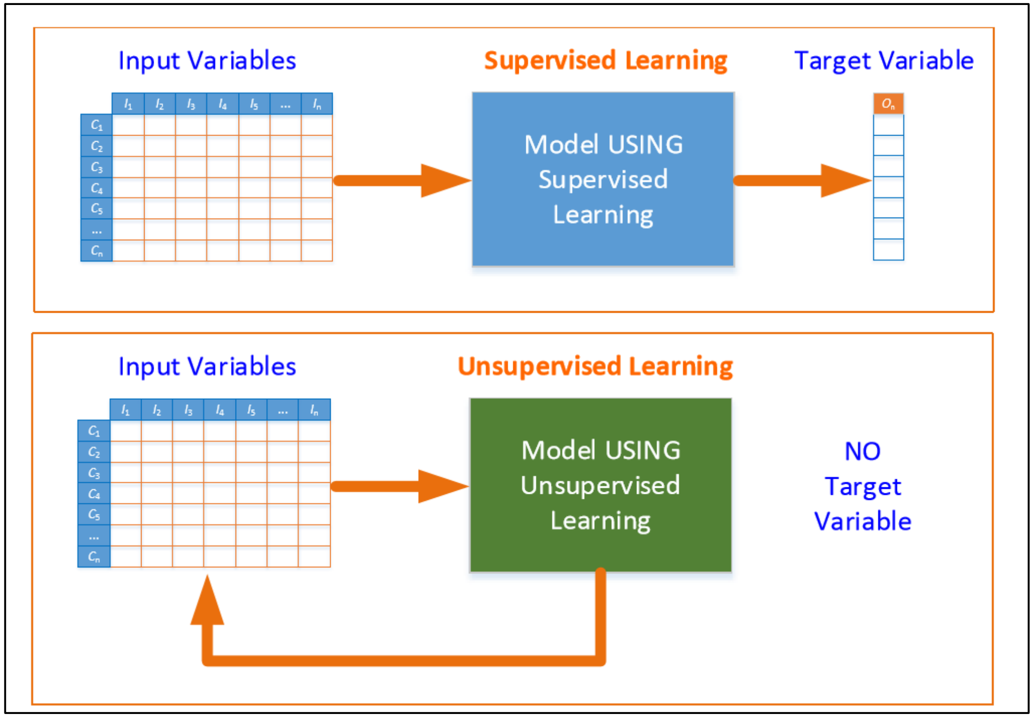
Response Variable

#	Modeling Methods	Response Variable: Numerical / Categorical
1	**Linear & Polynomial Regression**	Numerical
2	**Logistic Regression**	Categorical (Binary)
3	**Discriminant Analysis**	Categorical
4	**K Nearest Neighbor**	Categorical
5	**Decision and Regression Trees**	Categorical + Numerical
6	**Naïve Bayes**	Categorical
7	**Neural Networks**	Numerical + Categorical
8	**Clustering**	
9	**Principal Component Analysis**	
10	**Support Vector Machines**	Categorical
11	**ARIMA : Time Series**	Numerical

Supervised vs Unsupervised Learning

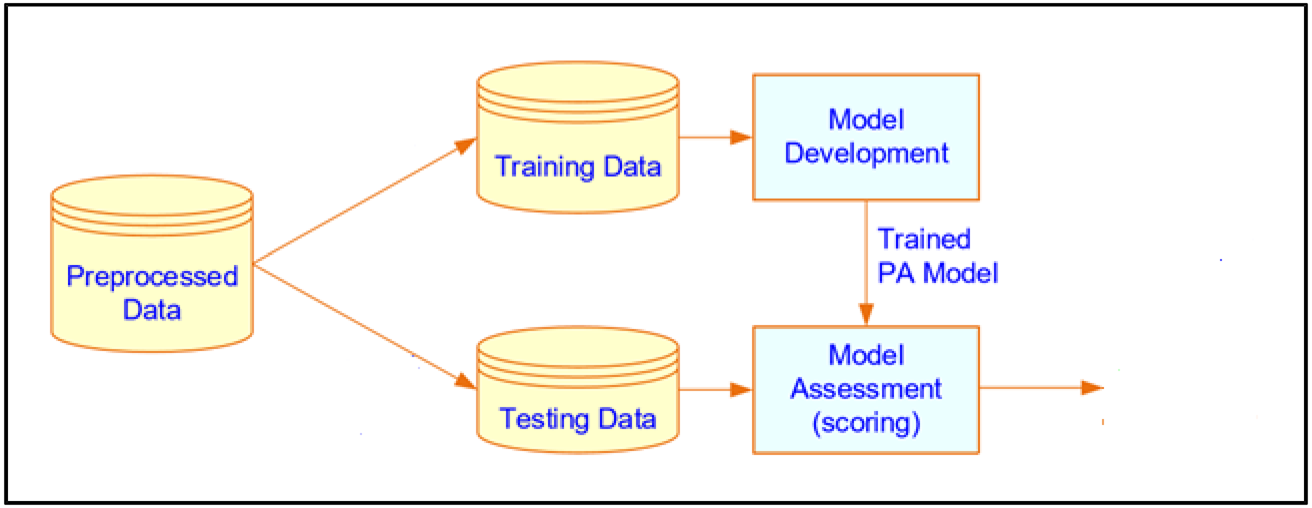
Supervised vs. Unsupervised Learning in PA

- Supervisor learning is the most common learning type where there is a target/output variable (which is also called supervisor)
  - Supervisor (target variable) teaches the algorithm how to build/learn the pattern model
  - In PA, supervised learning ≈ <u>predictive modeling</u>
- Unsupervised learning has NO target variable
  - No supervisor to teach → algorithm has to learn by itself
  - In PA, unsupervised learning ≈ <u>descriptive modeling</u>


Input Variables
Supervised Learning
Target Variable
I_1 I_2 I_3 I_4 I_5 ... I_n
C_1 C_2 C_3 C_4 C_5 ... C_n
Model USING Supervised Learning
O_n
Input Variables
Unsupervised Learning
I_1 I_2 I_3 I_4 I_5 ... I_n
C_1 C_2 C_3 C_4 C_5 ... C_n
Model USING Unsupervised Learning
NO Target Variable

Supervised Learning
Model Development and Deployment

- Single split model assessment methodology
- The model is tested on hold-out sample
  - Only the hold-out sample accuracy is reported



Copyright 2018 Dr. Ash Pahwa

Modeling Methods

#	Modeling Methods	Supervised or Unsupervised
1	**Linear & Polynomial Regression**	Supervised
2	**Logistic Regression**	Supervised
3	**Discriminant Analysis**	Supervised
4	**K Nearest Neighbor**	Supervised
5	**Decision and Regression Trees**	Supervised
6	**Naïve Bayes**	Supervised
7	**Neural Networks**	Supervised
8	**Clustering**	Unsupervised
9	**Principal Component Analysis**	Unsupervised
10	**Support Vector Machines**	Supervised
11	**ARIMA : Time Series**	Supervised

Classifying Based on Strategy to Build a Model

Classifying Based on Strategy to Build a Model

- Error based learning
  - Regression
  - Support Vector Machine
- Similarity Based Learning
  - K Nearest Neighbor
- Information Based Learning
  - Decision Trees
- Probability Based Learning
  - Naïve Bayes
- Mimicking the Human Brain
  - Neural networks

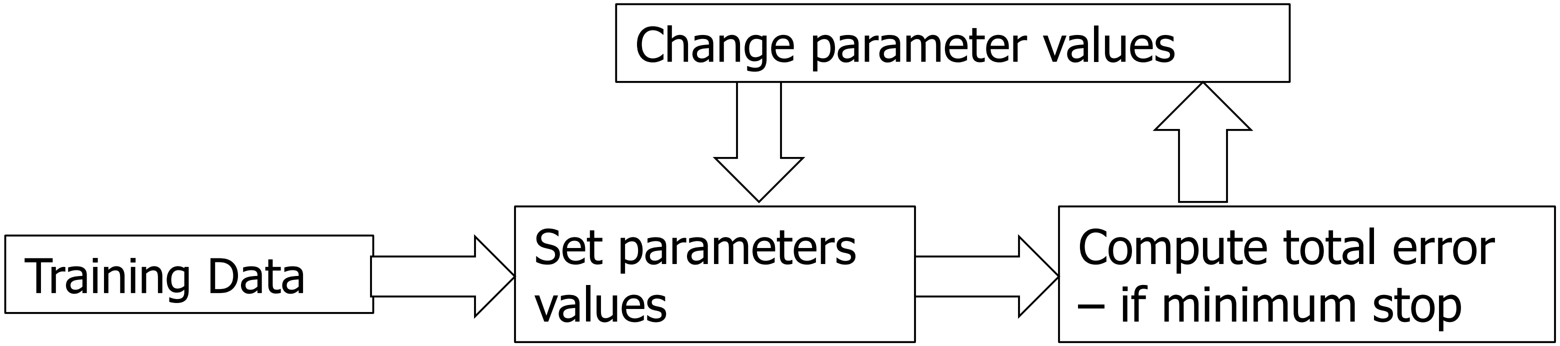
Error Based Learning

Linear Multi Variable Regression

Support Vector Machines

Error Based Learning

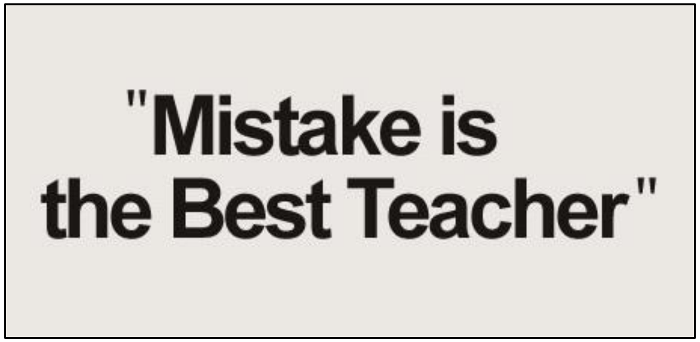
- In error-based machine learning
  - We perform a search for a set of parameters for a parameterized model
  - That minimizes the total error across the predictions made by the model
  - With respect to a set of training instances (training data)



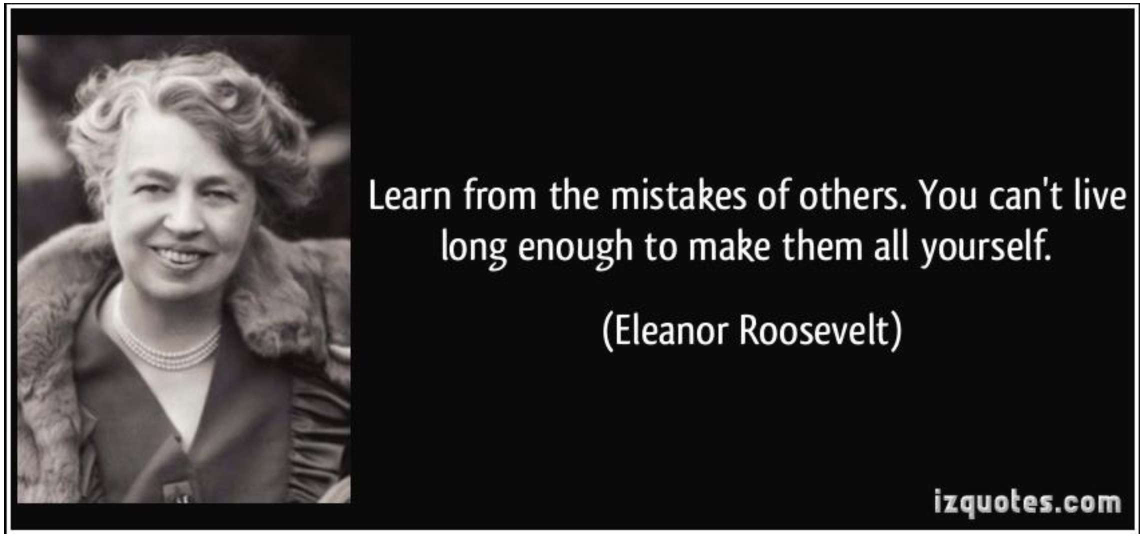
Copyright 2018 Dr. Ash Pahwa

Error Based Learning

- All humans learn using this technique
- Most natural form of learning



"Mistake is the Best Teacher"



Learn from the mistakes of others. You can't live long enough to make them all yourself.

(Eleanor Roosevelt)

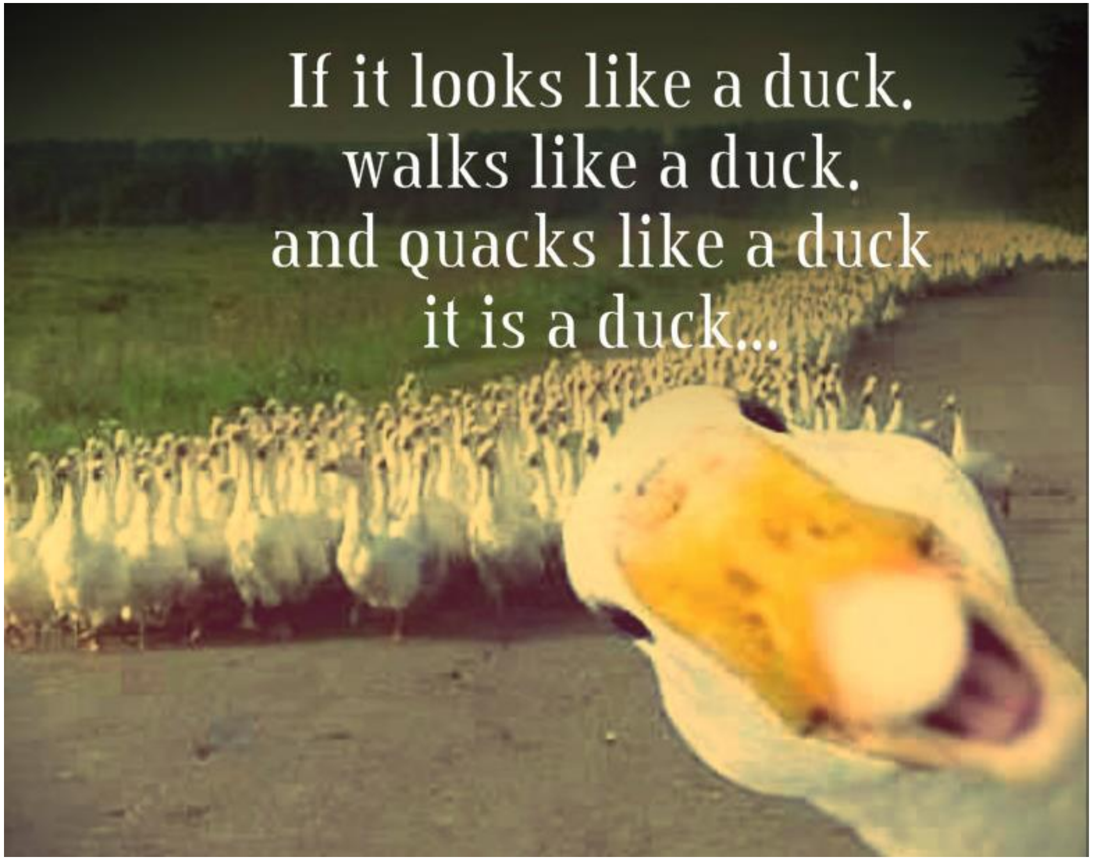
Error Based Machine Learning Techniques

- Linear Multi Variable Regression
- Support Vector Machine

Similarity Based Learning

k Nearest Neighbor

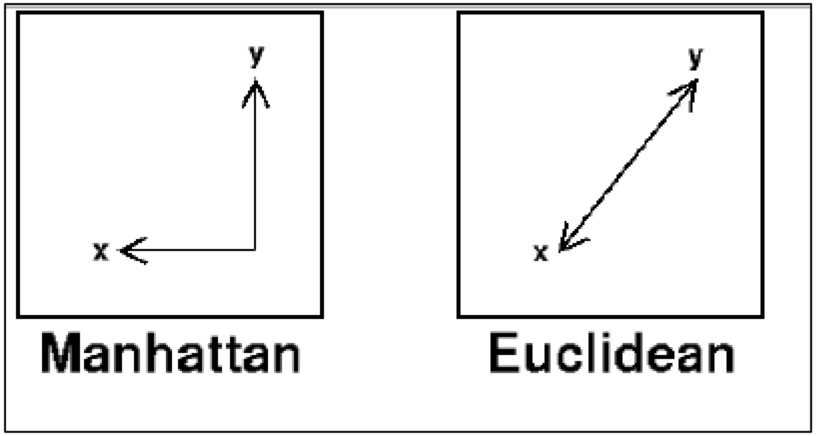
Similarity Based Learning



Similarity Based Learning
k Nearest Neighbor

- Compute the distance matrices between objects

$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N} (Xi - Yi)^2}$$



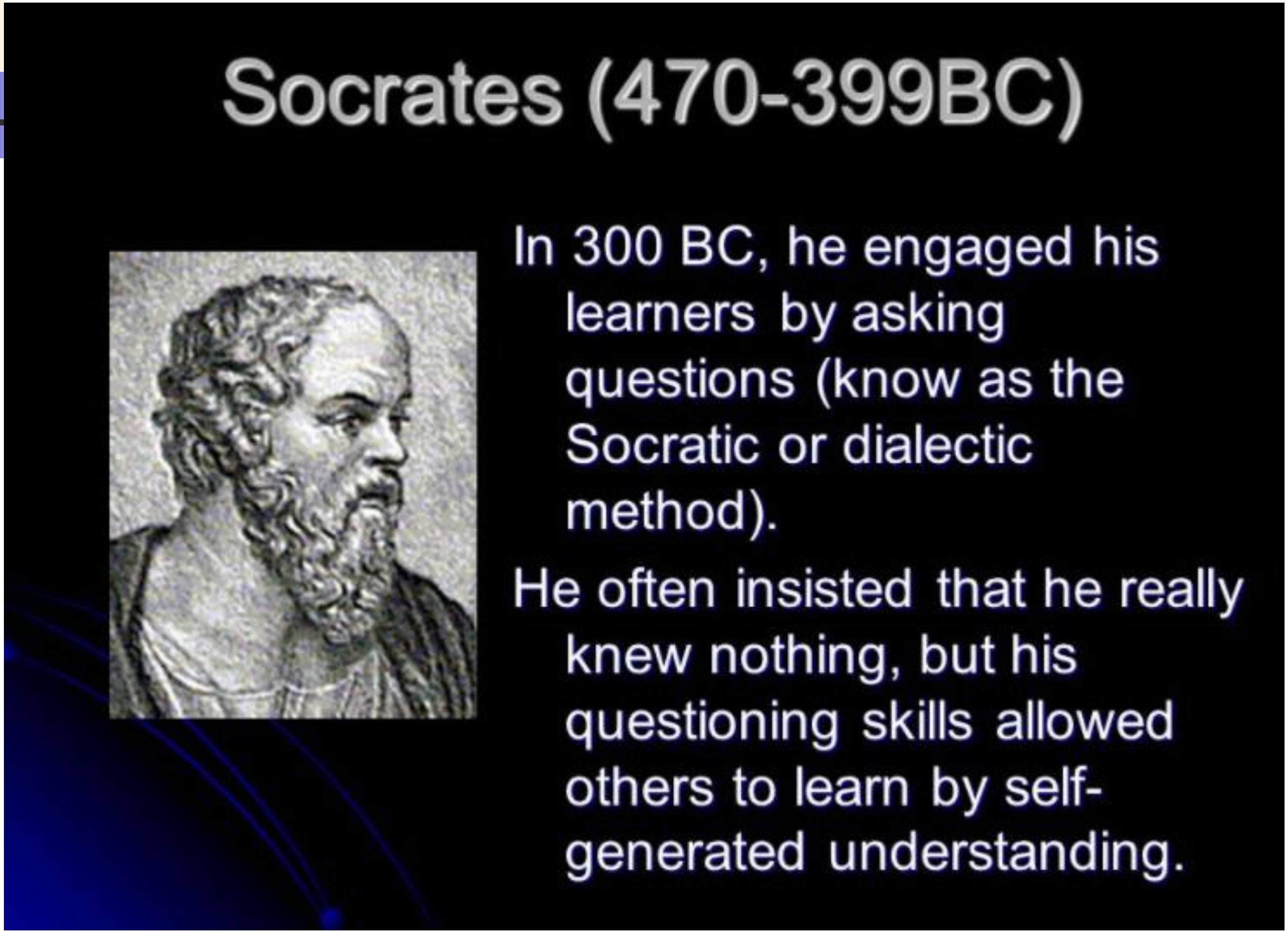
Information Based Learning

Decision Trees

Information Based Learning

- Learn by Asking Questions
- The **Socratic** approach to **questioning** is based on the practice of disciplined, thoughtful dialogue.
- **Socrates**, the early Greek philosopher/teacher, believed that disciplined practice of thoughtful **questioning** enabled the student to examine ideas logically and to determine the validity of those ideas.

Socrates: Greek Philosopher



Socrates (470-399BC)

- In 300 BC, he engaged his learners by asking questions (know as the Socratic or dialectic method).
- He often insisted that he really knew nothing, but his questioning skills allowed others to learn by self-generated understanding.

What is Decision Tree?

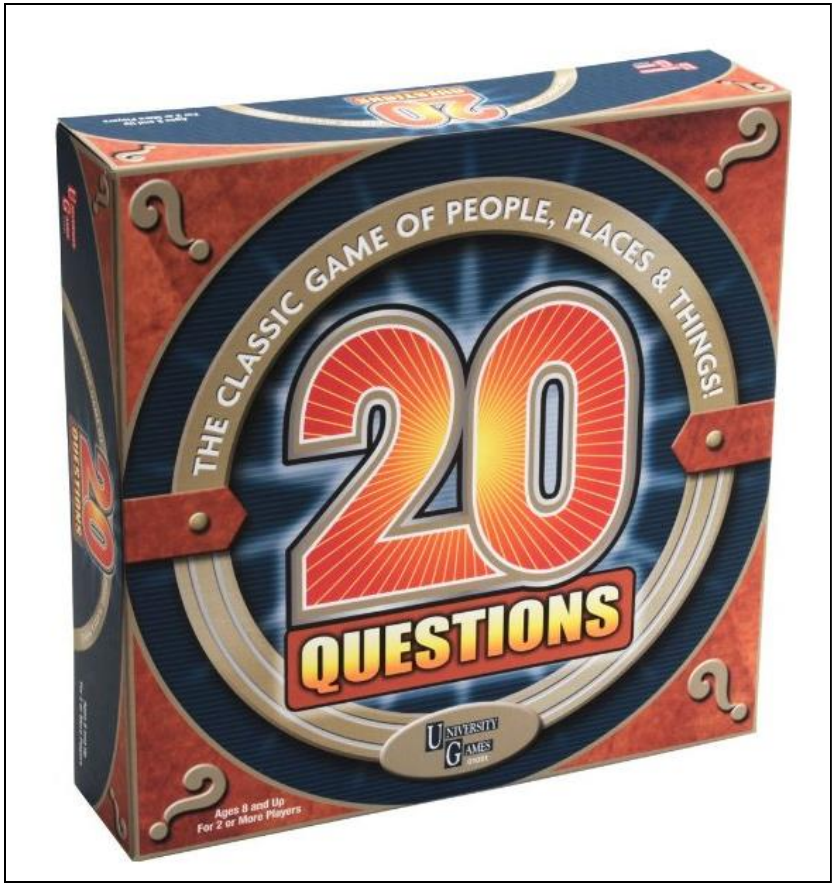
- Identical to 20 questions game for kids

20 Questions Board Game

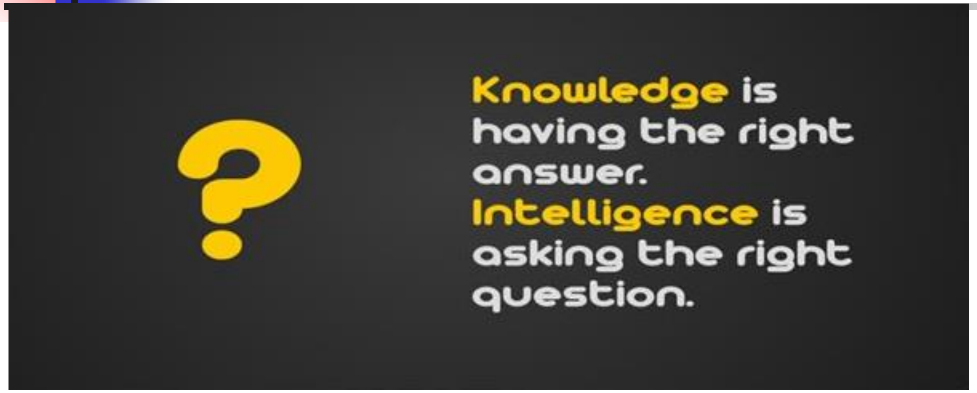
~~$29.99~~ $18.99 ✓Prime

Only 1 left in stock - order soon.

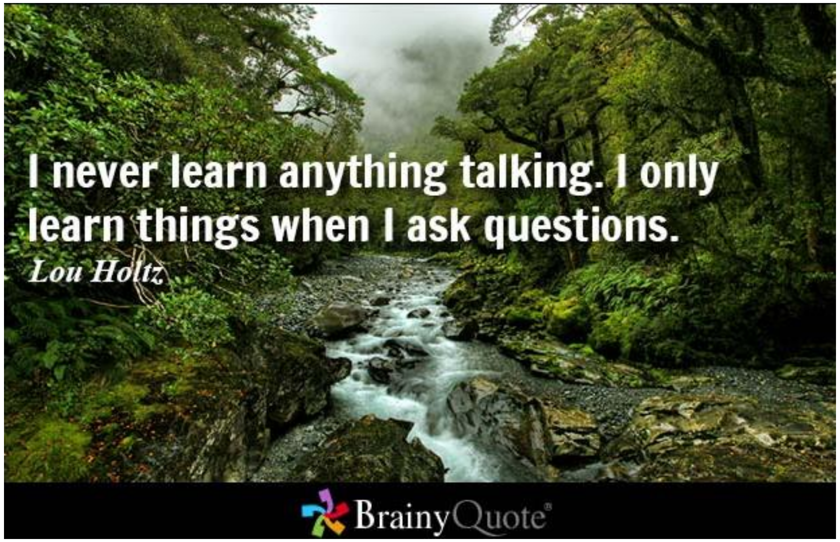
FREE Shipping on orders over $35



Learning by Asking Questions



Knowledge is having the right answer.
Intelligence is asking the right question.



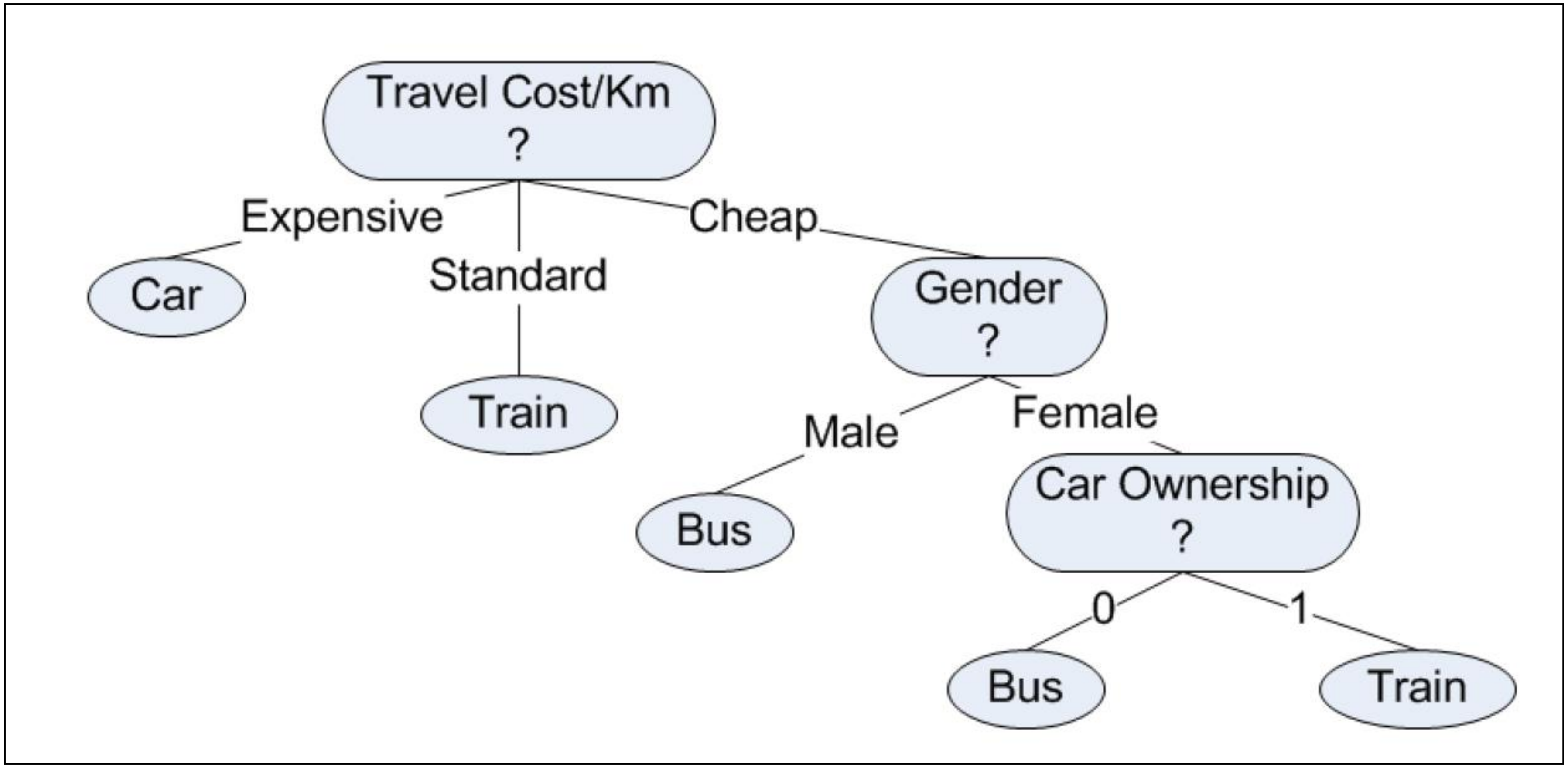
I never learn anything talking. I only learn things when I ask questions.

Lou Holtz

BrainyQuote

Decision Tree

	Gender	Car Ownership	Travel Cost	Income Level	Transportation Mode
1	Male	0	Cheap	Low	Bus
2	Male	1	Cheap	Medium	Bus
3	Female	1	Cheap	Medium	Train
4	Female	0	Cheap	Low	Bus
5	Male	1	Cheap	Medium	Bus
6	Male	0	Standard	Medium	Train
7	Female	1	Standard	Medium	Train
8	Female	1	Expensive	High	Car
9	Male	2	Expensive	Medium	Car
10	Female	2	Expensive	High	Car



Information Based Machine Learning Techniques

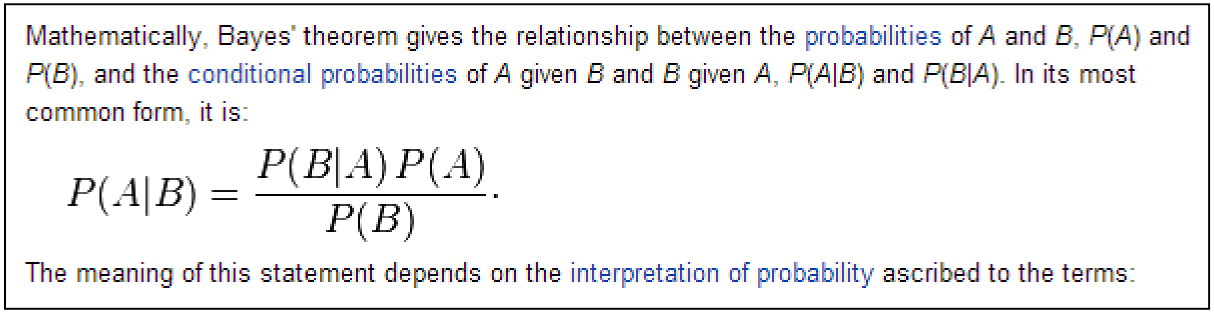
- Decision Trees
- Regression Trees
- Split of decision trees are based on the entropy of the tables

Probability Based Learning

Naïve Bayes

Thomas Bayes

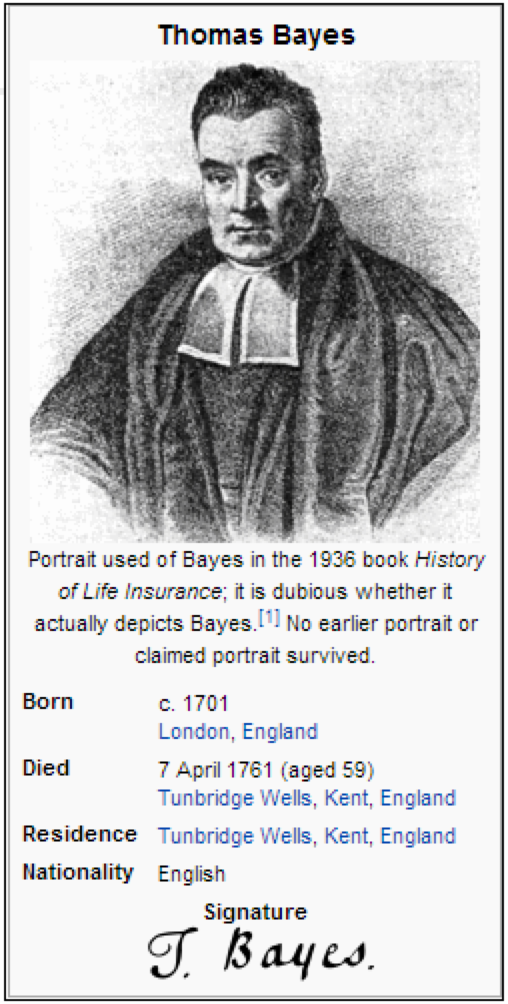
- **Thomas Bayes** (1701 – 1761) was an
  - English statistician,
  - Philosopher and
  - Presbyterian minister
- Known for having formulated a specific case of the theorem that bears his name:
  - Bayes' theorem



Mathematically, Bayes' theorem gives the relationship between the probabilities of *A* and *B*, *P*(*A*) and *P*(*B*), and the conditional probabilities of *A* given *B* and *B* given *A*, *P*(*A*|*B*) and *P*(*B*|*A*). In its most common form, it is:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

The meaning of this statement depends on the interpretation of probability ascribed to the terms:



Thomas Bayes

Portrait used of Bayes in the 1936 book *History of Life Insurance*; it is dubious whether it actually depicts Bayes.[1] No earlier portrait or claimed portrait survived.

Born	c. 1701 London, England
Died	7 April 1761 (aged 59) Tunbridge Wells, Kent, England
Residence	Tunbridge Wells, Kent, England
Nationality	English

Signature

T. Bayes.

Bayes Rule

- Provides a way to compute *reverse* probability
- Given P(B|A)
  - We can compute P(A|B)

$$\boxed{P(A|B)=\frac{P(B|A).P(A)}{P(B)}}$$

Naïve Assumption
Assuming Variable Independence

- What is the probability that a person will respond
  - Given that customer is urban AND a golfer

$$P(\text{response} = Y \mid \text{urban and golfer}) =$$

$$= \frac{P(\text{response} = Y) * P(\text{urban \& golfer}) \mid \text{response} = Y)}{P(\text{urban \& golfer})}$$

$$* \text{ Naive Assumption : If 'urban' and 'golfer' are independent variables}$$

$$P(\text{response} = Y \mid \text{urban and golfer}) =$$

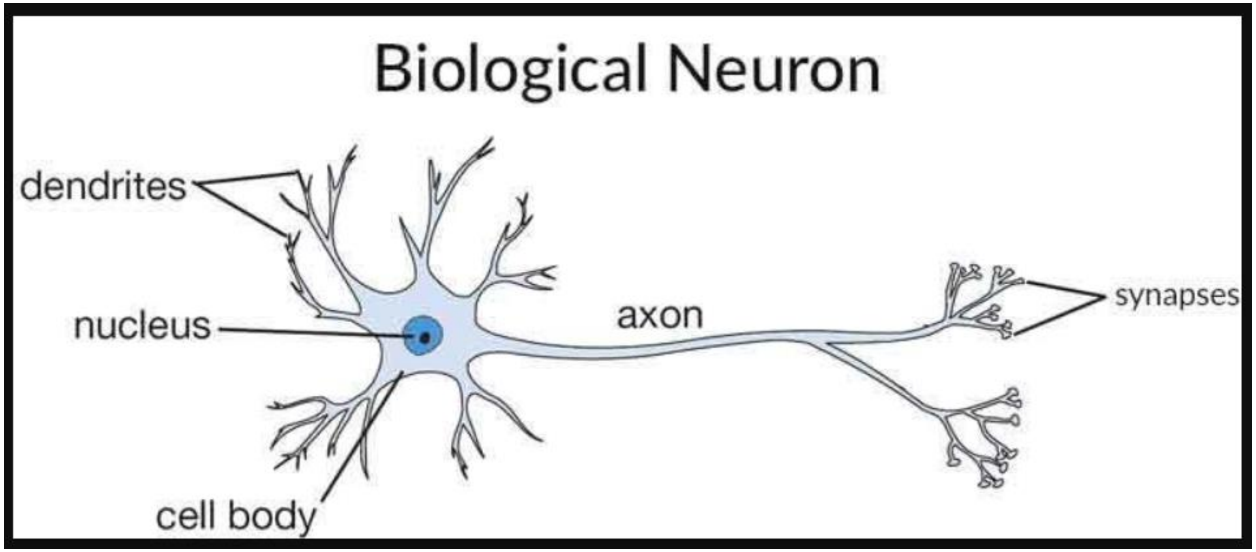
$$= \frac{P(\text{response} = Y) * P(\text{urban} \mid \text{response} = Y) * P(\text{golfer} \mid \text{response} = Y)}{P(\text{urban \& golfer})}$$

Mimicking the Human Brain

Neural Networks

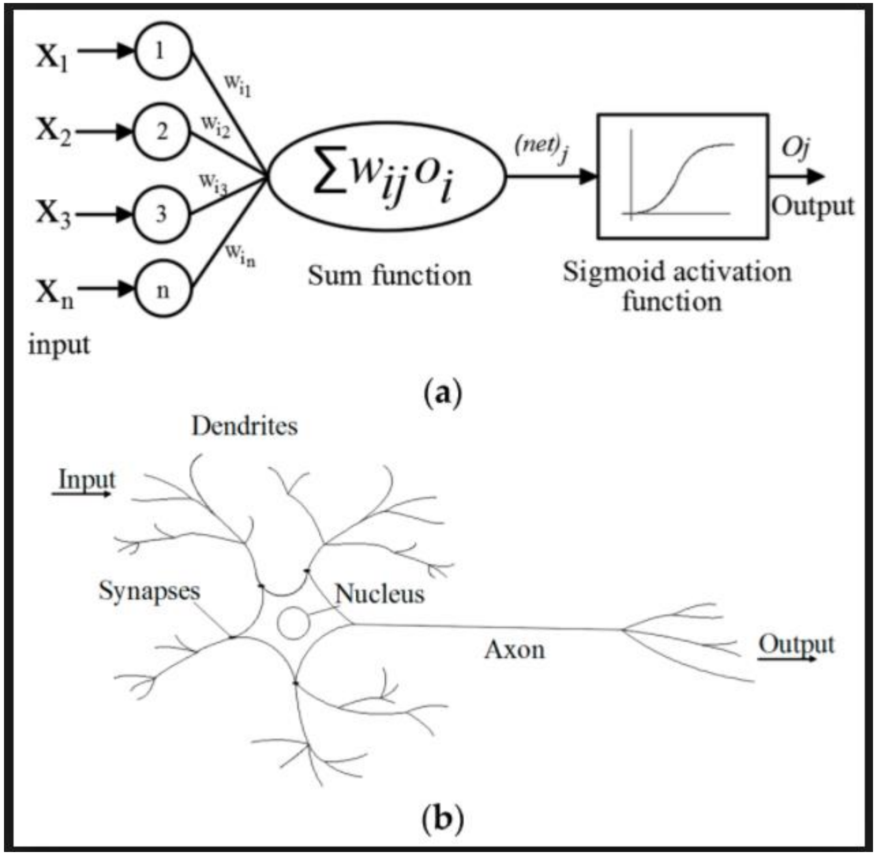
Inspiration for Neural Networks
Biological Neuron

- Human Brains have 86 billion neurons



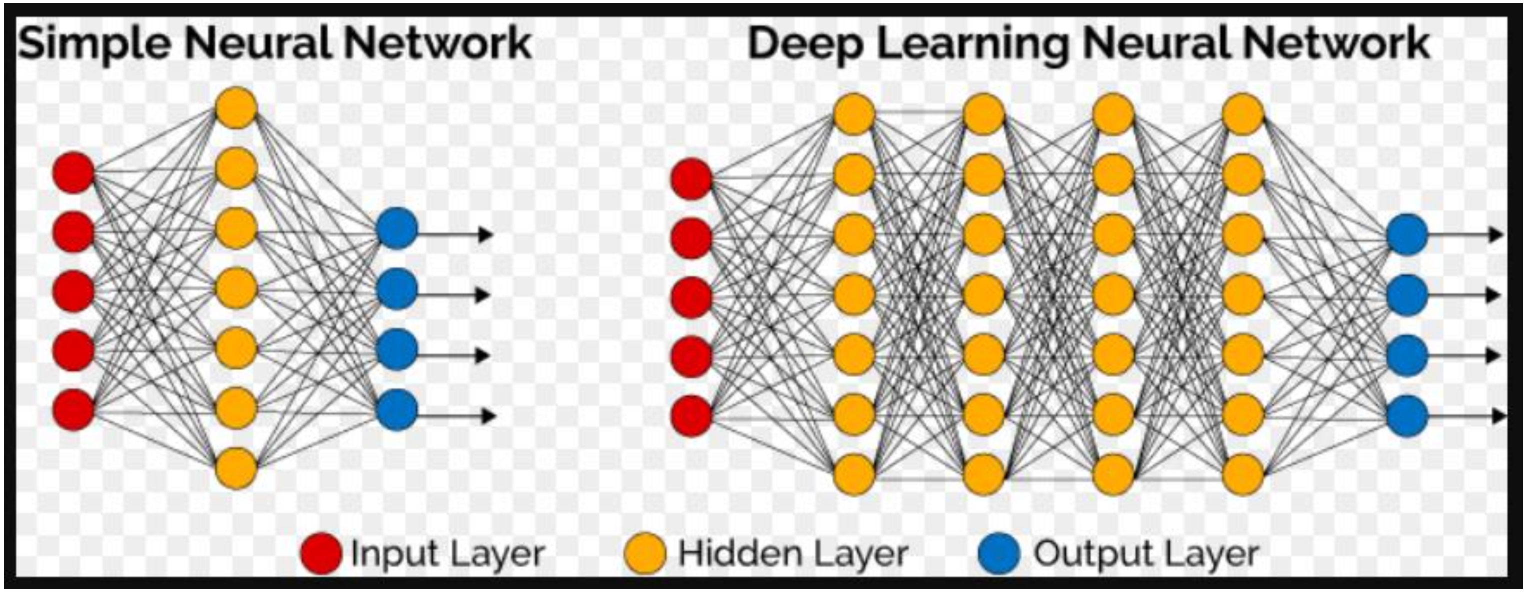
Neural Networks

- Neural Networks behave similar to human brain
- Central Idea
  - Extract linear combinations of the inputs
  - Model the target as the non-linear functions of these features



Deep Learning

- Complex set of Neural Networks with many layers of processing



Main Applications of Deep Learning Neural Networks

- Image Recognition
  - Convolution Neural Networks
- Image Classification
  - Convolution Neural Networks
- Hand Writing Identification
- Speech Recognition
  - Long Short-Term Memory Networks

Modeling Methods

#	Modeling Methods	Strategy
1	**Linear & Polynomial Regression**	Error Based Minimizing Error
2	**Logistic Regression**	Maximizing Likelihood
3	**Discriminant Analysis**	
4	**K Nearest Neighbor**	Similarity Based
5	**Decision and Regression Trees**	Information Based
6	**Naïve Bayes**	Probability Based
7	**Neural Networks**	Mimicking Human Brain
8	**Clustering**	
9	**Principal Component Analysis**	
10	**Support Vector Machines**	Error Based
11	**ARIMA : Time Series**	Auto Regression & Moving Average

Summary

#	Modeling Methods	Response Variable: Numerical /Categorical	Supervised or Unsupervised	Strategy
1	**Linear & Polynomial Regression**	Numerical	Supervised	Error Based Minimizing Error
2	**Logistic Regression**	Categorical (Binary)	Supervised	Maximizing Likelihood
3	**Discriminant Analysis**	Categorical	Supervised	
4	**K Nearest Neighbor**	Categorical	Supervised	Similarity Based
5	**Decision and Regression Trees**	Categorical + Numerical	Supervised	Information Based
6	**Naïve Bayes**	Categorical	Supervised	Probability Based
7	**Neural Networks**	Numerical + Categorical	Supervised	Mimicking Human Brain
8	**Clustering**		Unsupervised	
9	**Principal Component Analysis**		Unsupervised	
10	**Support Vector Machines**	Categorical	Supervised	Error Based
11	**ARIMA : Time Series**	Numerical	Supervised	Auto Regression & Moving Average

Summary

- CRISP/DM Model
- Classifying Modeling Methods
  - Response Variable:
    - Numerical or Categorical
  - Supervised or Unsupervised
  - Strategy:
    - Error Based
    - Information Based
    - Similarity Based
    - Probability Based
    - Mimicking the Human Brain (Neural networks)