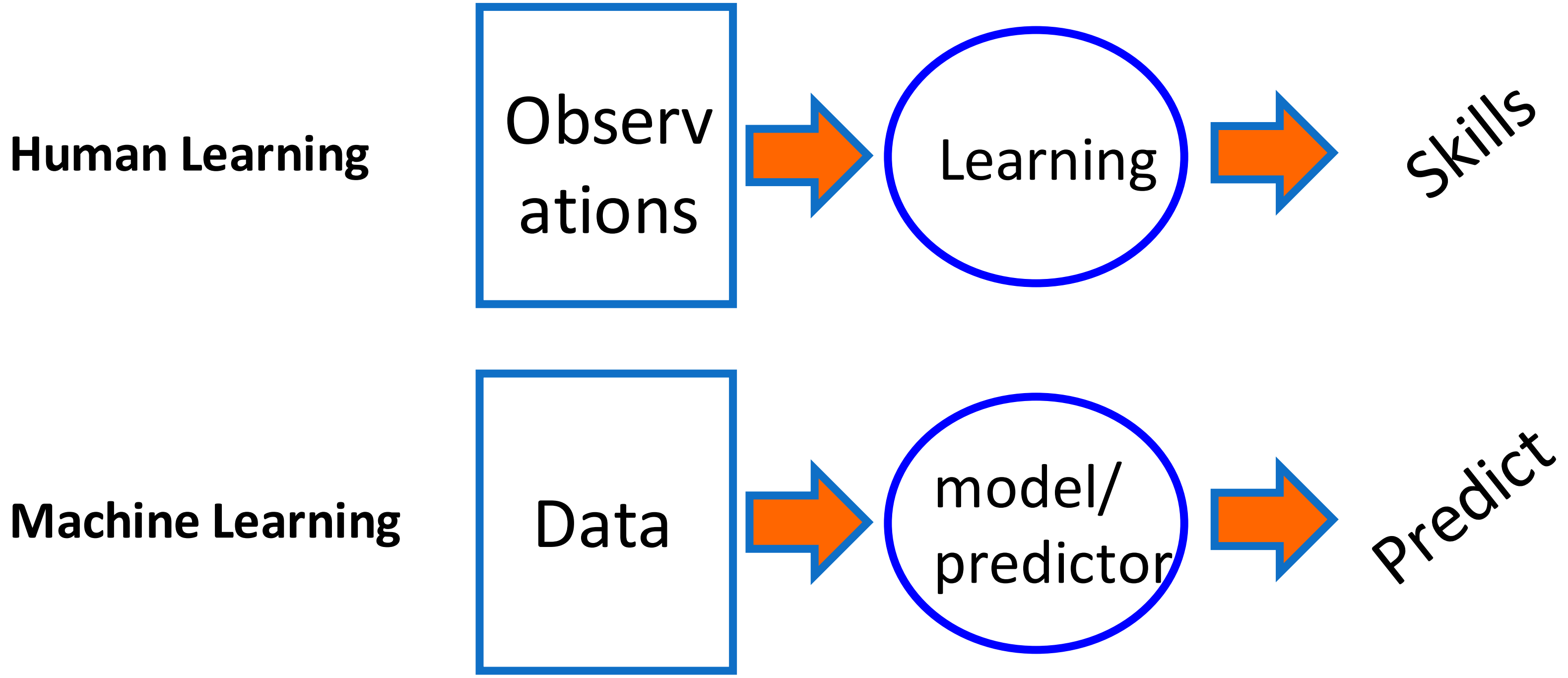
Start-Tech Academy

MACHINE LEARNING

Machine learning is programming computers to optimize a performance criterion using example data or past experience. Machine learning can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.

What



MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Why

Organizations/governments are collecting a lot of data

Information from data is being used to take key business/ political decisions

At lower levels in organization, data is used for MIS reporting

At higher levels data based prescriptive and predictive models are being built

Machine learning is the most popular technique of creating these predictive and prescriptive model

MACHINE LEARNING

Machine learning is closely associated with Statistics, AI and Data mining

ML vs others

Machine Learning Vs. Statistics

- Traditional Statistics focuses on provable results with mathematical assumptions, and care less about computation
- “Statistics: A useful tool for Machine Learning”

Machine Learning Vs. Artificial Intelligence

- “Machine Learning is one possible route to realize AI”

Machine Learning Vs. Data Mining

- Traditional DM focuses on provable results with math assumptions along with efficient computation in large dataset
- “Difficult to distinguish ML and DM in reality”

MACHINE LEARNING

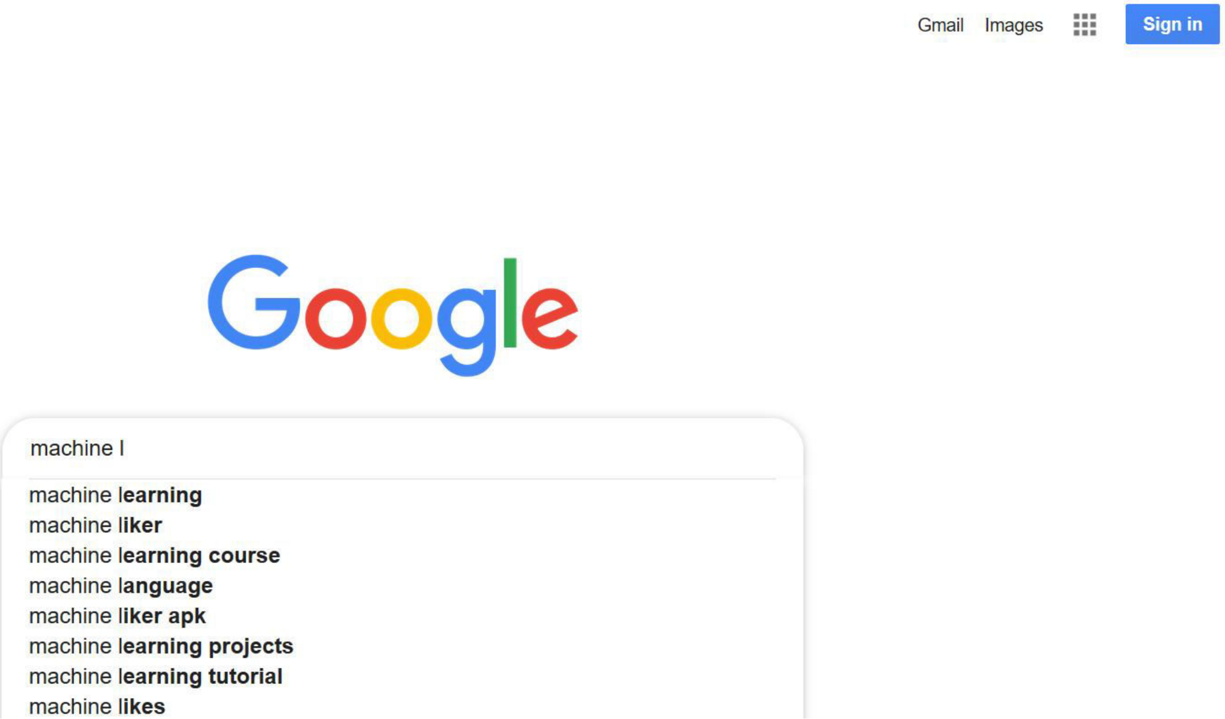
Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Example

MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Example



machine l

machine **learning**
machine **liker**
machine **learning course**
machine **language**
machine **liker apk**
machine **learning projects**
machine **learning tutorial**
machine **likes**

MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Use cases

Banking / Telecom / Retail

- Identify:
  - Prospective customers
  - Dissatisfied customers
  - Good customers
  - Bad payers
- Obtain:
  - More effective advertising
  - Less credit risk
  - Fewer fraud
  - Decreased churn rate

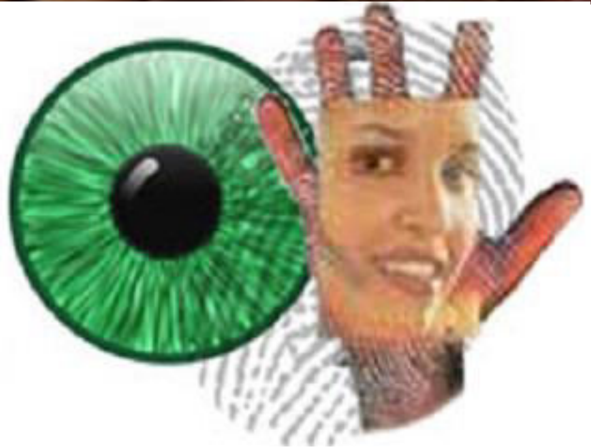
MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Use cases

Biomedical / Biometrics

- Medicine:
  - Screening
  - Diagnosis and prognosis
  - Drug discovery
- Security:
  - Face recognition
  - Signature / fingerprint / iris verification
  - DNA fingerprinting



MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Use cases
Computer / Internet

- Computer interfaces:
  - Troubleshooting wizards
  - Handwriting and speech
  - Chat bots
- Internet
  - Hit ranking
  - Spam filtering
  - Text categorization
  - Text translation
  - Recommendation

Google Translate

MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Example

For example, a real estate agent who wants to price a particular property will have:

Output variable: Price of property (Y)
Input variables: Area covered (X1), Number of bedrooms (X2), proximity to a landmark (X3), proximity to market (X4), recent sale price of a neighborhood property (X5) and so on

The real estate wants to find out

$$Y = f(X1, X2, X3, X4, X5\ldots)$$

So that whenever s/he gives a value of the input variables to this function, s/he can get the price of the property.

WHY ESTIMATE f(x)

f(x) defines the relationship between dependent and independent variables.

Types

There are two major reasons to estimate f(x):

1. **Prediction** – When the values of input variables is available and output variable is to be predicted. We are only interested in the value of Y, not in the relationship of Y with other variables

2. **Inference** – When the relationship between input and output variable is important. We want to establish how output variable varies with change in each predictor variable

WHY ESTIMATE f(x)

f(x) defines the relationship between dependent and independent variables.

Choice of Model

Choice of model for estimating will depend on whether we want to **predict or infer**.

- For Prediction, **accuracy** of predicted function is the most important
- For Inference, **interpretability** of predicted function is most important

For example, **linear regression** is simple to interpret but may not give very accurate predicted values of Y

Whereas highly **non-linear models** may be predicting very accurately but the relationship may be very difficult to interpret

HOW TO ESTIMATE F(x)

Next, we need to specify the type of **learning method**.

Parametric vs Non parametric

In Parametric approach, we assume the functional form of the relationship between predictor and predicted variable

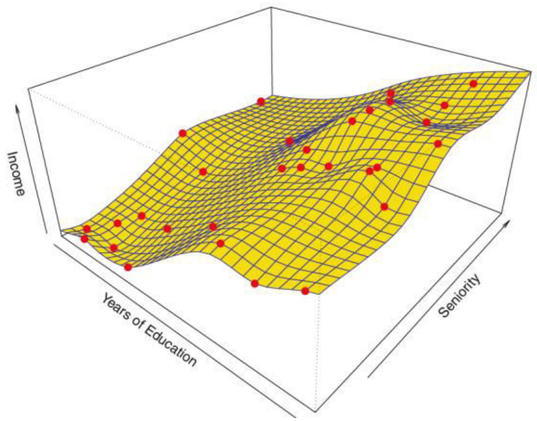
For example, we may assume linear relationship between house price with other variables

Price (Y) = a0 + a1*x1 + a2*x2 + a3*x3 .....an*xn

Then we will use the training data to estimate the values of a0, a1, a2, a3... an

In non-parametric approach, we do not assume any functional form for the relationship. Instead f is estimated by getting as close to the training points

For example, in the image shown, for three variables, a three dimensional spleen is created which is as close to the points and has a smooth surface

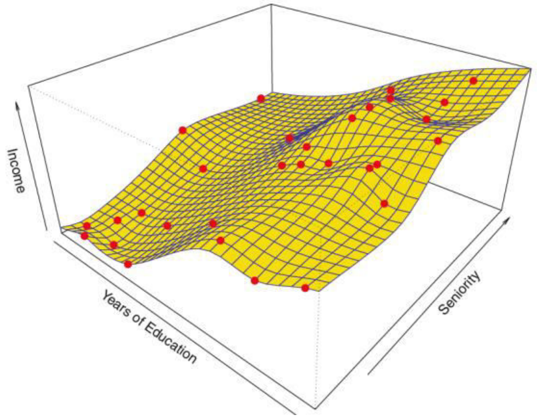


Start-Tech ACADEMY

HOW TO ESTIMATE F(x)

Parametric vs Non parametric

Parametric vs Non parametric



Parametric approach

- Usually more interpretable
- May not be as accurate
- Preferable if inference is the reason estimating f(x)

Non-parametric approach, w

- Less interpretable
- Potentially more accurate
- Needs large amount of data to train
- Preferable if prediction is the priority

TYPES OF LEARNING

Supervised vs Unsupervised learning

Supervised Vs Unsupervised

Supervised Learning:

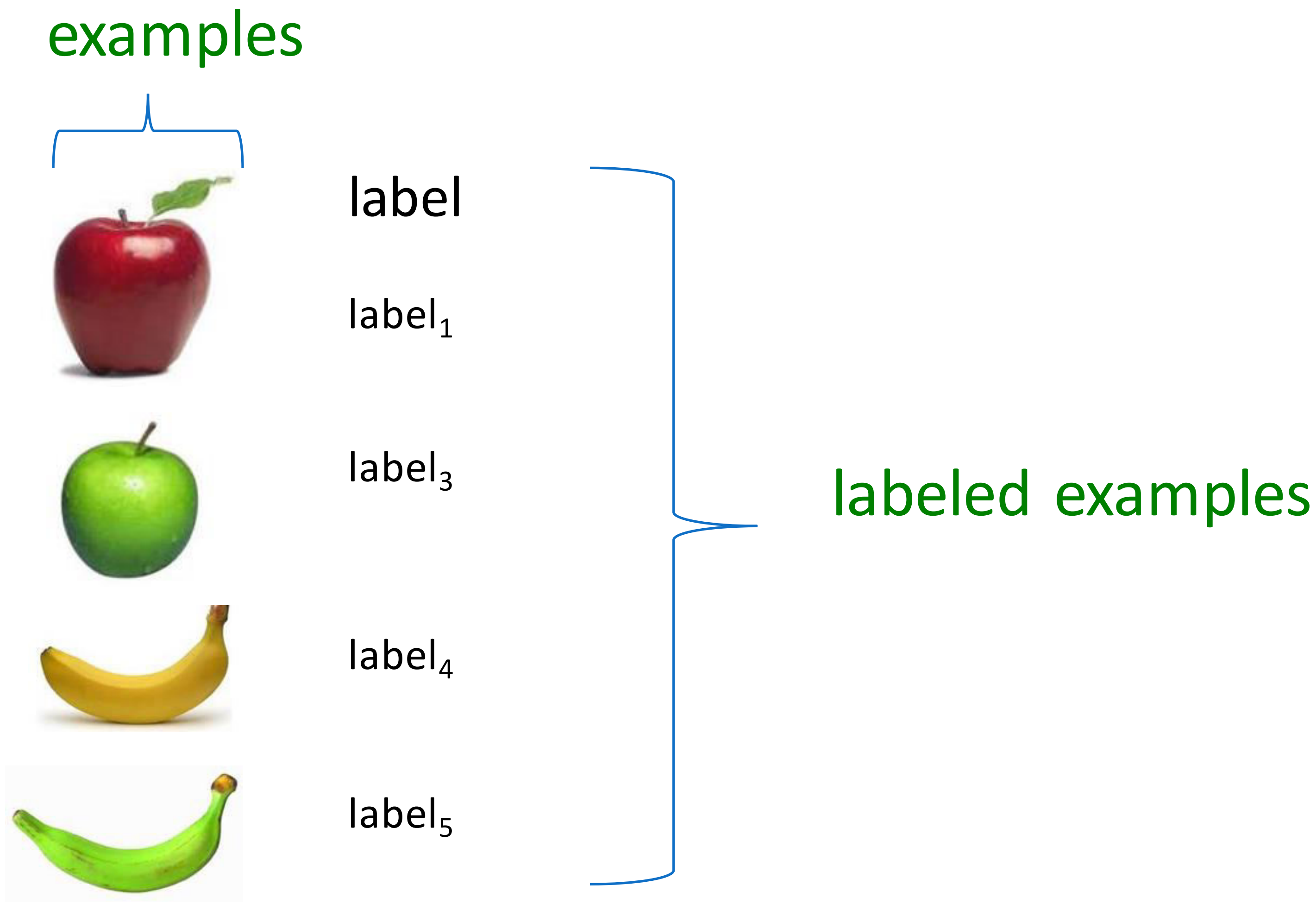
- Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.
- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Unsupervised Learning:

- Unsupervised learning is where you only have input data (X) and no corresponding output variables.
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Supervised Learning: Example

Supervised Learning Example



examples

label

label$_1$

label$_3$

label$_4$

label$_5$

labeled examples

Supervised Learning: Example

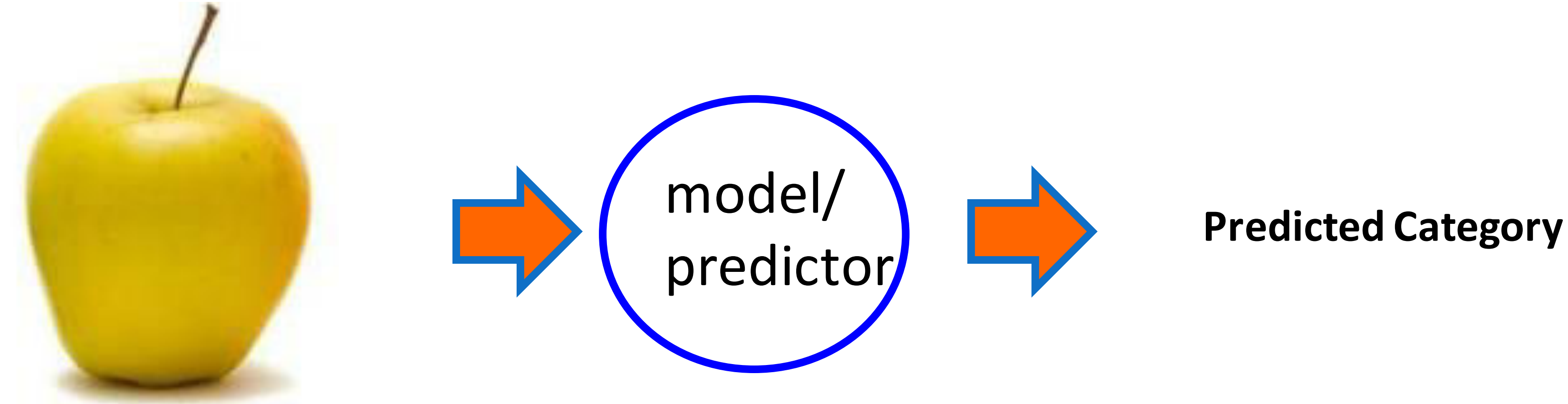
Supervised Learning Example

	Category	Weight
	Apple	100 gm
	Apple	80 gm
	Banana	40 gm
	Banana	60 gm

model/ predictor

Supervised Learning: classification

Supervised Learning Example (classification)



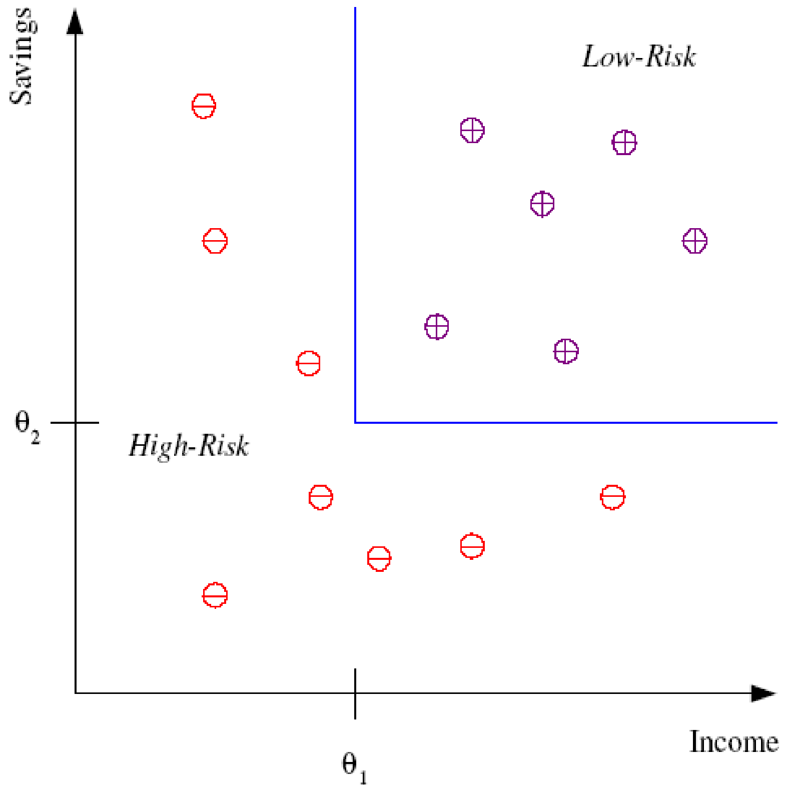
Supervised Learning (classification)

Supervised Learning (classification)

Classification:

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their income and savings
- Model - Discriminant

IF *income* > θ_1 AND *savings* > θ_2
THEN low-risk ELSE high-risk

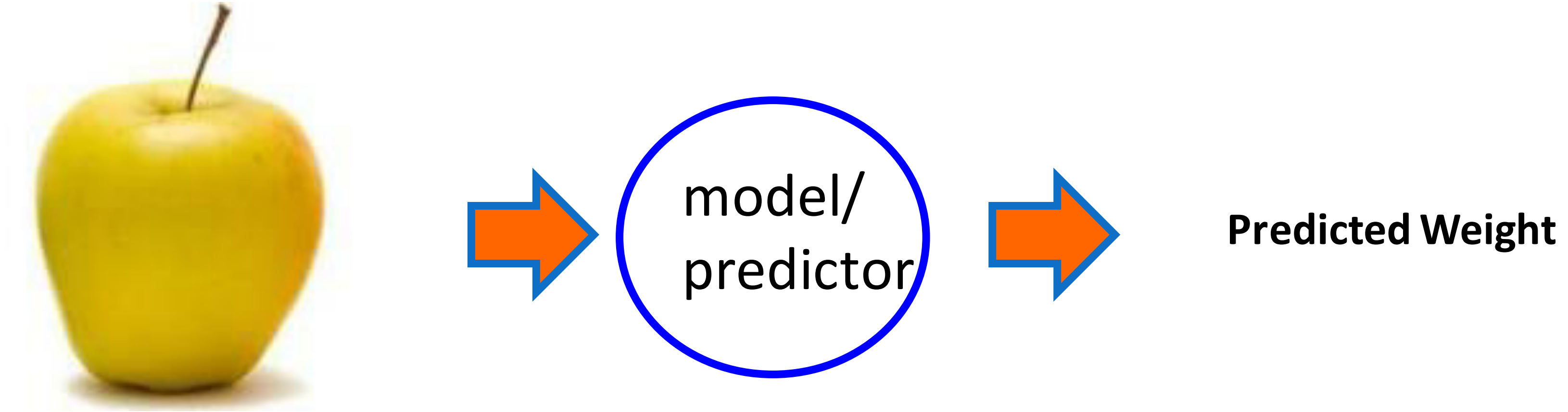


Applications :

- Pattern recognition
- Face recognition
- Character recognition
- Medical diagnosis
- Web Advertising

Supervised Learning: Regression

Supervised Learning Example (Regression)



model/ predictor

Predicted Weight

Supervised Learning (Regression)

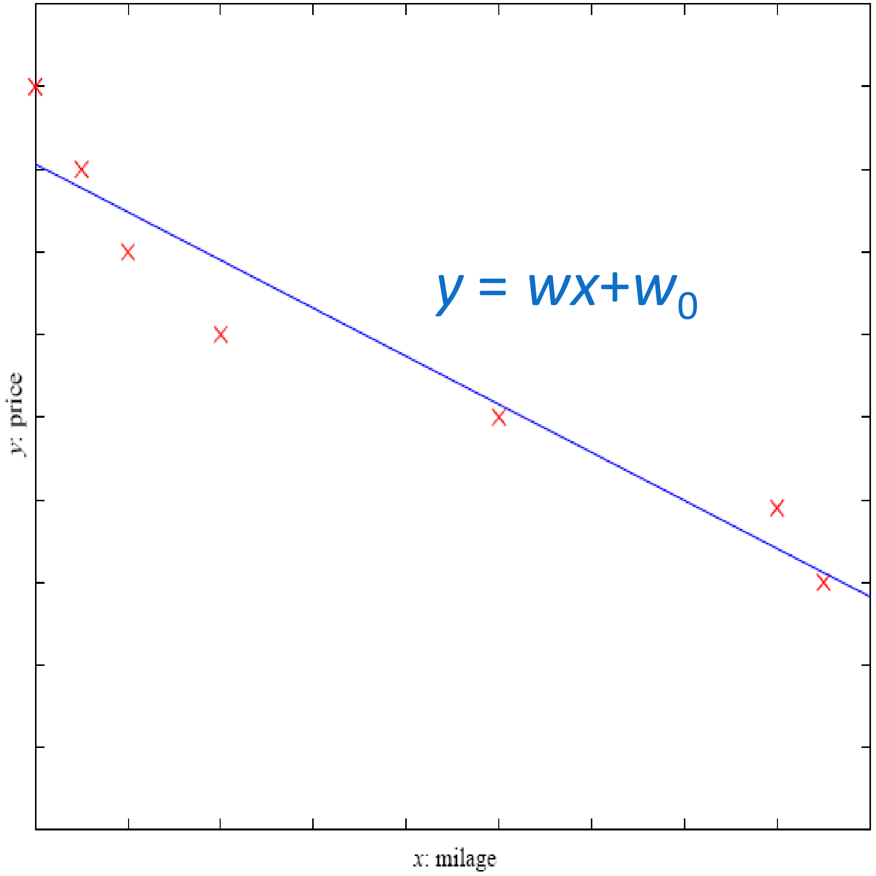
Supervised Learning (Regression)

Regression:

- Example: Price of a used car
- x : car attributes
  (e.g. mileage)
  y : price
- Model – Linear Regression
  $y = wx + w0$

Applications :

- Weather forecast
- Sales forecasting
- Advertising budget allocation
- Product pricing



Start-Tech ACADEMY

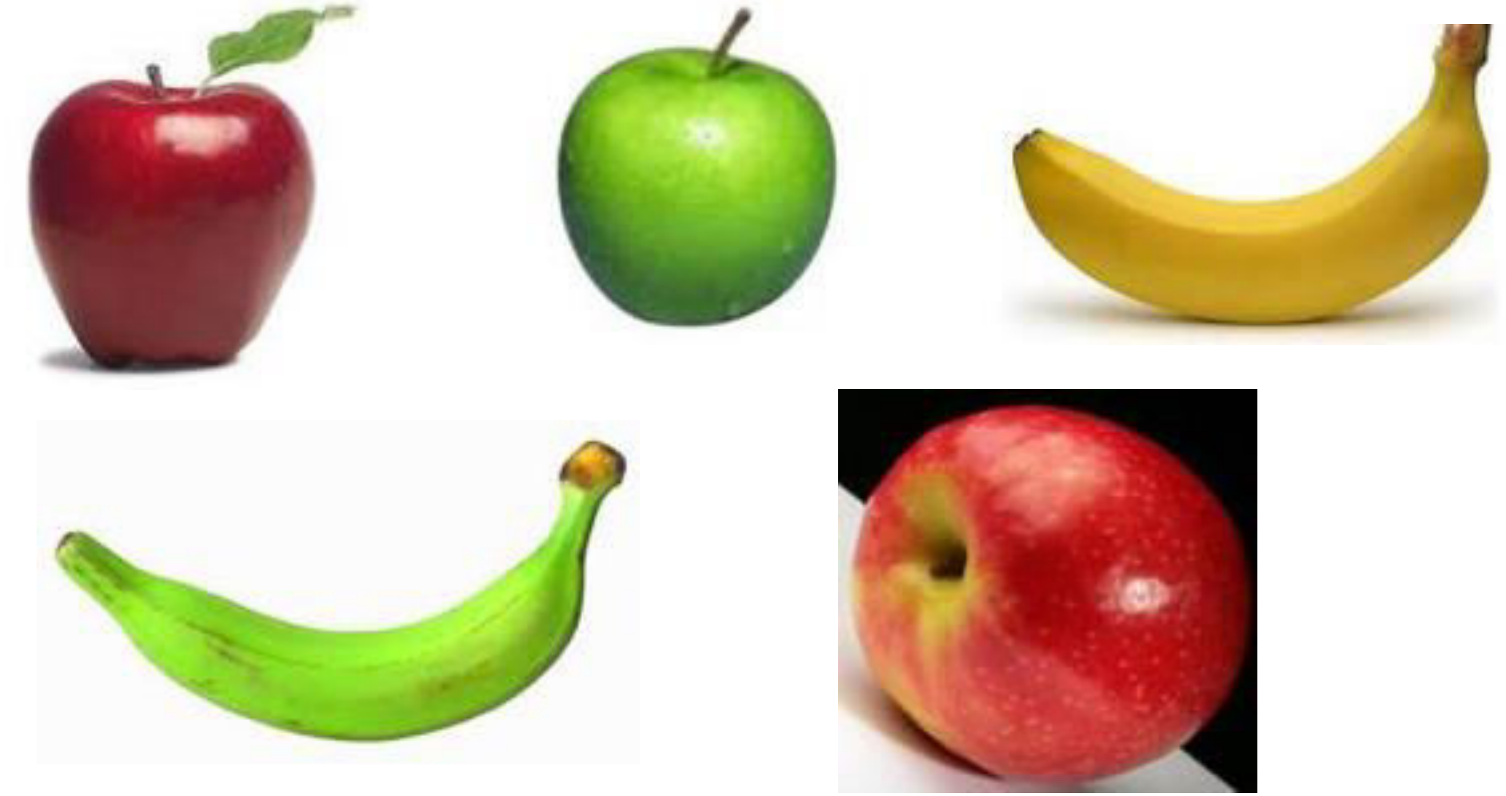
Supervised Learning Algorithms

Supervised Learning Algorithms

ANN
Neural
Naive
CART
Networks
EnsembleBayes
Discriminant
Regression
AnalysisSVM
KNN
Linear
Random
Forest
Logistic
CHAID

Unsupervised Learning: Example

Unsupervised Learning Example



Unupervised learning: given data, i.e. examples, but no labels

Unsupervised Learning Algorithms

Unsupervised Learning Algorithms

Unsupervised Learning - Algorithms:

- Clustering
  - K means
  - Hierarchical clustering
- Hidden Markov Models (HMM)
- Dimension Reduction (Factor Analysis, PCA)
- Feature Extraction methods
- Self-organizing Maps (Neutral Nets)

Machine Learning Model

Steps

Steps in Building ML Model

1. Problem formulation
2. Data Tidying
3. Pre-Processing
4. Train-Test Split
5. Model Building
6. Validation and Model Accuracy
7. Prediction

Machine Learning Model

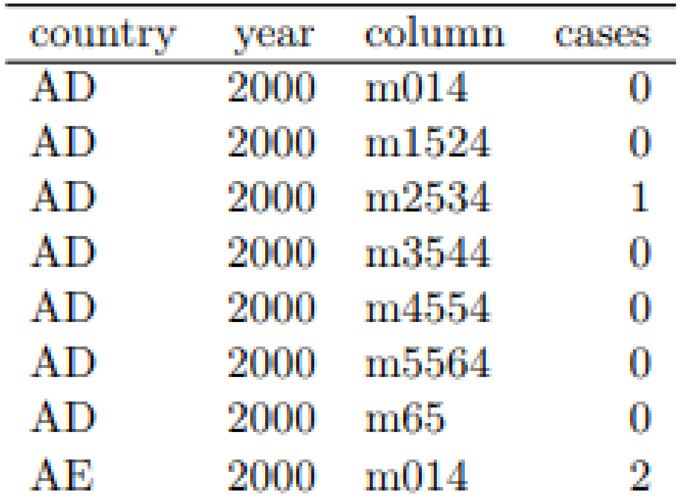
1. Problem formulation

- Convert your business problem into a Statistical problem
- Clearly define the dependent and independent variable
- Identify whether you want to predict or infer

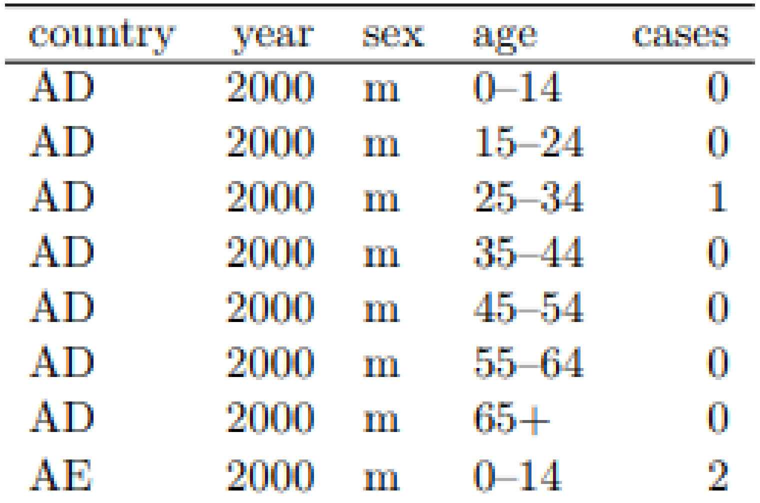
Machine Learning Model

2. Data Tidying

- Transform collected data into a useable data table format
- Example



country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2



country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2

Machine Learning Model

3. Data Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction

Machine Learning Model

4. Test - Train Split

Training data is the information used to train an algorithm.
The training data includes both input data and the corresponding expected output.
Based on this data, the algorithm can learn the relationship between input and output variables.

Testing data includes only input data, not the corresponding expected output.
The testing data is used to assess the accuracy of model created or the predictor function created using the training data.

- There should not be any overlap between the two.
- Usually, 70-80% of the available data is used as training data and 20-30% as testing data

Machine Learning Model

5. Model Training

$$\underbrace{y}_{\text{Output}} = \underbrace{f}_{\text{Function}}(\underbrace{x}_{\text{Input variables}})$$

Machine Learning Model

6. Performance Metrics and Validation

In Sample error

- Error resulted from applying your prediction algorithm to the dataset you built it with

Out of Sample error

- Error resulted from applying your prediction algorithm to a new data set

Machine Learning Model

7. Prediction

- Setup a pipeline to use your model in real life scenario
- Improve by monitoring your model over time
- Try to automate