Machine Learning

Quick introduction

Digital Vidya

What is Machine Learning?

- **The capability of Artificial Intelligence systems to learn by extracting patterns from data is known as Machine Learning**

- Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given.

What is Machine Learning?

- Data science, machine learning and artificial intelligence are some of the top trending topics in the tech world today

- Machine Learning
  - Study of algorithms that improve their performance at some task with experience

What is Machine Learning?

- Machine learning is a discipline that deals with programming the systems so as to make them automatically learn and improve with experience

- Here, learning implies recognizing and understanding the input data and taking informed decisions based on the supplied data

- It is very difficult to consider all the decisions based on all possible inputs

- To solve this problem, algorithms are developed that build knowledge from a specific data and past experience by applying the principles of statistical science, probability, logic, mathematical optimization, reinforcement learning

ML

- **Machine Learning (ML)** is an automated learning with little or no human intervention
- It involves programming computers so that they learn from the available inputs
- The main purpose of machine learning is to explore and construct algorithms that can learn from the previous data and make predictions on new input data

Growth of Machine Learning

- **Machine learning is preferred approach to**
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology

- **This trend is accelerating**
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment
  - It turns out to be difficult to extract knowledge from human experts→*failure of expert systems in the 1980's.*

Applications of Machine Learning Algorithms

- The developed machine learning algorithms are used in various applications such as:

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Data mining
- Expert systems
- Robotics

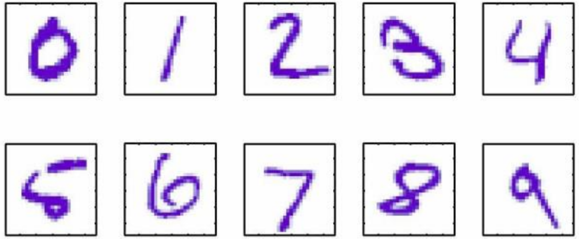
- Vision processing
- Language processing
- Forecasting things like stock market trends, weather
- Pattern recognition
- Games
- [Your favorite area]

Benefits of Machine Learning

- Powerful Processing
- Better Decision Making & Prediction
- Quicker Processing
- Accurate
- Affordable Data Management
- Inexpensive
- Analyzing Complex Big Data

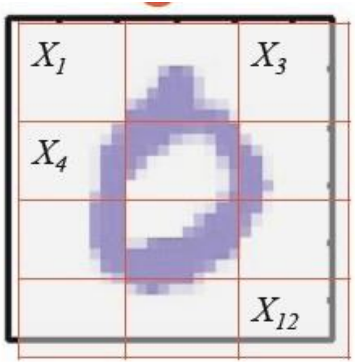
Classification: Digit Recognition

Example:



Input(x_i):ImageFeatures

Output(Y):Clas Labels{$y^0, y^1, .. y^9$}



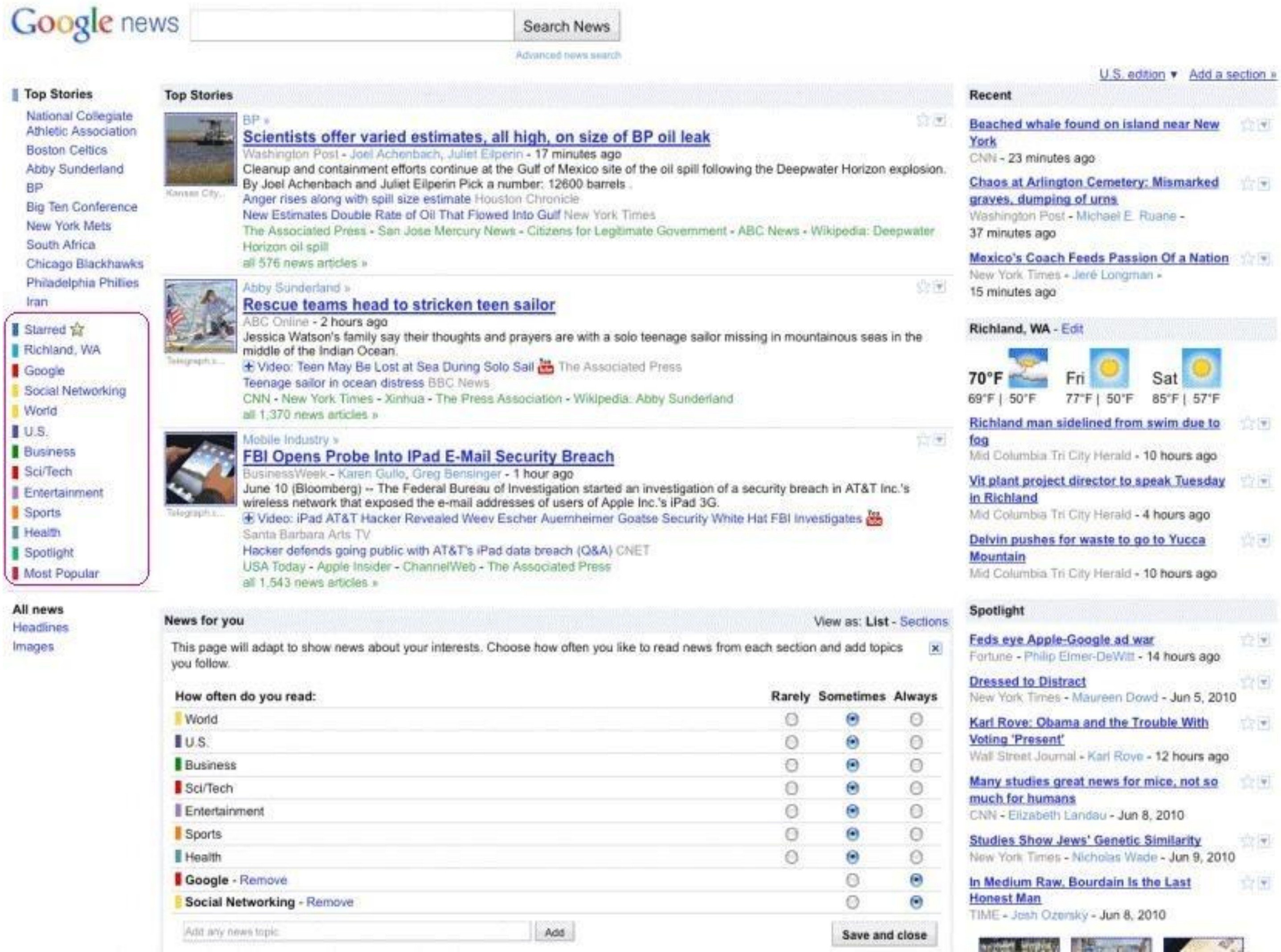
Features(X_i):
Proportion of pixels in each of the 12 cells
X_i *where i*=1,2,. .. , 12

$$\left.\begin{array}{l} x_i^0=0\text{-}10\% \\ x_i^1=10\text{-}20\% \\ \dots \end{array}\right\} \quad Val(Xi)=10$$

No of parameters$=10^{12}-1$

Where ML is used.

Inbox
Outbox
Spam (3015)
Trash



Top Stories
National Collegiate Athletic Association
Boston Celtics
Abby Sunderland
BP
Big Ten Conference
New York Mets
South Africa
Chicago Blackhawks
Philadelphia Phillies
Iran

Starred ☆
Richland, WA
Google
Social Networking
World
U.S.
Business
Sci/Tech
Entertainment
Sports
Health
Spotlight
Most Popular

All news
Headlines
Images

Top Stories

BP »
Scientists offer varied estimates, all high, on size of BP oil leak
Washington Post - Joel Achenbach, Juliet Eilperin - 17 minutes ago
Cleanup and containment efforts continue at the Gulf of Mexico site of the oil spill following the Deepwater Horizon explosion. By Joel Achenbach and Juliet Eilperin Pick a number: 12600 barrels .
Anger rises along with spill size estimate Houston Chronicle
New Estimates Double Rate of Oil That Flowed Into Gulf New York Times
The Associated Press - San Jose Mercury News - Citizens for Legitimate Government - ABC News - Wikipedia: Deepwater Horizon oil spill
all 576 news articles »

Kansas City...

Abby Sunderland »
Rescue teams head to stricken teen sailor
ABC Online - 2 hours ago
Jessica Watson's family say their thoughts and prayers are with a solo teenage sailor missing in mountainous seas in the middle of the Indian Ocean.
+ Video: Teen May Be Lost at Sea During Solo Sail The Associated Press
Teenage sailor in ocean distress BBC News
CNN - New York Times - Xinhua - The Press Association - Wikipedia: Abby Sunderland
all 1,370 news articles »

Telegraph.c...

Mobile Industry »
FBI Opens Probe Into IPad E-Mail Security Breach
BusinessWeek - Karen Gullo, Greg Bensinger - 1 hour ago
June 10 (Bloomberg) -- The Federal Bureau of Investigation started an investigation of a security breach in AT&T Inc.'s wireless network that exposed the e-mail addresses of users of Apple Inc.'s iPad 3G.
+ Video: iPad AT&T Hacker Revealed Weev Escher Auernheimer Goatse Security White Hat FBI Investigates
Santa Barbara Arts TV
Hacker defends going public with AT&T's iPad data breach (Q&A) CNET
USA Today - Apple Insider - ChannelWeb - The Associated Press
all 1,543 news articles »

Telegraph.c...

News for you View as: **List** - Sections

This page will adapt to show news about your interests. Choose how often you like to read news from each section and add topics you follow. ☒

How often do you read:	Rarely	Sometimes	Always
World	○	●	○
U.S.	○	●	○
Business	○	●	○
Sci/Tech	○	●	○
Entertainment	○	●	○
Sports	○	●	○
Health	○	●	○
Google - Remove		○	●
Social Networking - Remove		○	●

Add any news topic [Add] [Save and close]

Recent

Beached whale found on island near New York
CNN - 23 minutes ago

Chaos at Arlington Cemetery: Mismarked graves, dumping of urns
Washington Post - Michael E. Ruane - 37 minutes ago

Mexico's Coach Feeds Passion Of a Nation
New York Times - Jeré Longman » 15 minutes ago

Richland, WA - Edit

70°F	Fri	Sat
69°F | 50°F	77°F | 50°F	85°F | 57°F

Richland man sidelined from swim due to fog
Mid Columbia Tri City Herald - 10 hours ago

Vit plant project director to speak Tuesday in Richland
Mid Columbia Tri City Herald - 4 hours ago

Delvin pushes for waste to go to Yucca Mountain
Mid Columbia Tri City Herald - 10 hours ago

Spotlight

Feds eye Apple-Google ad war
Fortune - Philip Elmer-DeWitt - 14 hours ago

Dressed to Distract
New York Times - Maureen Dowd - Jun 5, 2010

Karl Rove: Obama and the Trouble With Voting 'Present'
Wall Street Journal - Karl Rove - 12 hours ago

Many studies great news for mice, not so much for humans
CNN - Elizabeth Landau - Jun 8, 2010

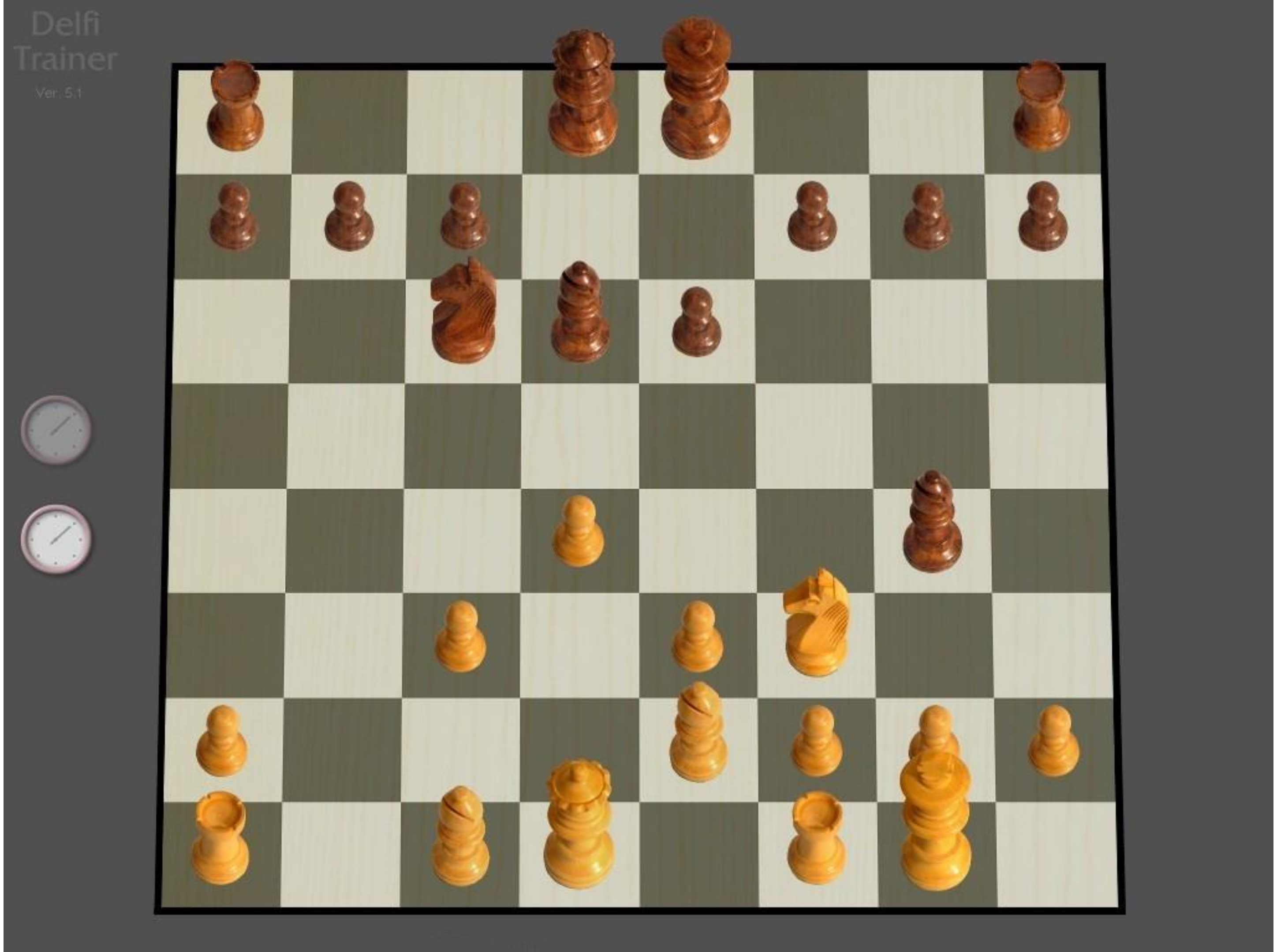
Studies Show Jews' Genetic Similarity
New York Times - Nicholas Wade - Jun 9, 2010

In Medium Raw, Bourdain Is the Last Honest Man
TIME - Josh Ozersky - Jun 8, 2010

GRAND DARPA CHALLENGE

Autonomous Ground Vehicles

IBM
WATSON
POWER7
POWER7
POWER7
POWER7
POWER7

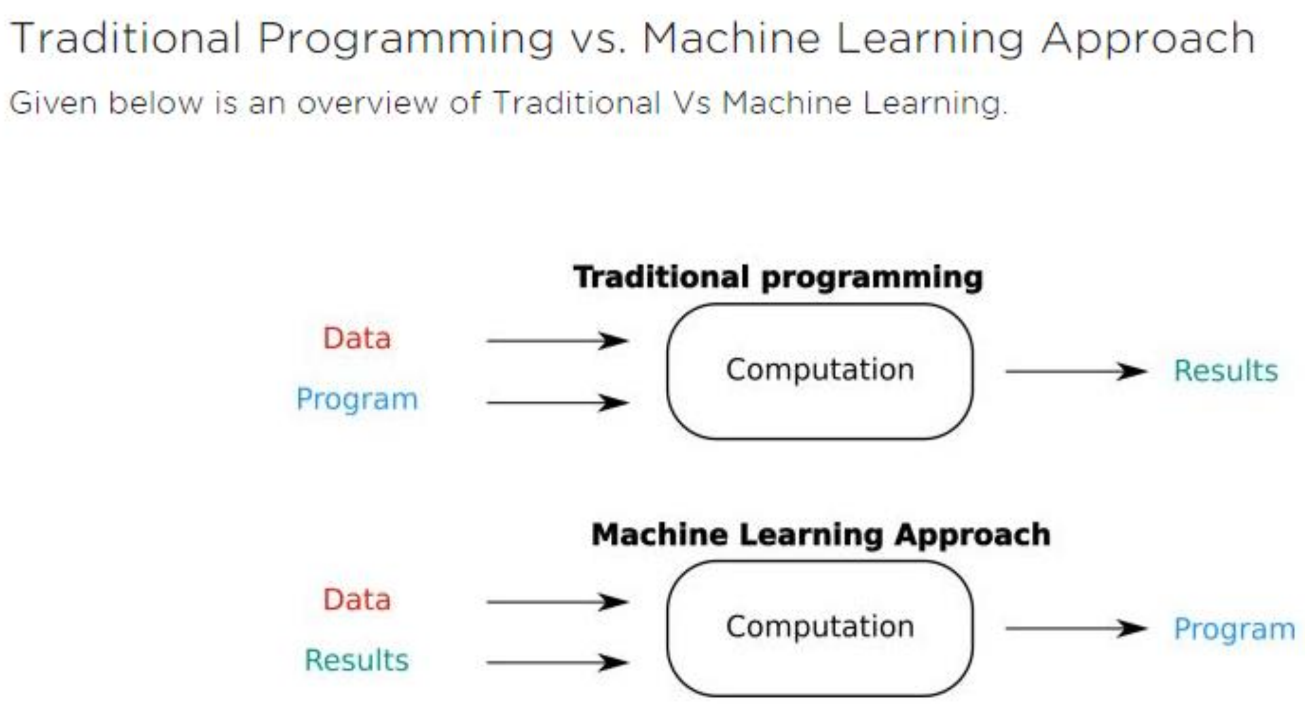


Delfi
Trainer
Ver. 5.1

Steps Involved in Machine Learning

- A machine learning project involves the following steps:
  - Defining a Problem
  - Preparing Data
  - Evaluating Algorithms
  - Improving Results
  - Presenting Results

Comparison with Traditional Programming



Traditional Programming vs. Machine Learning Approach

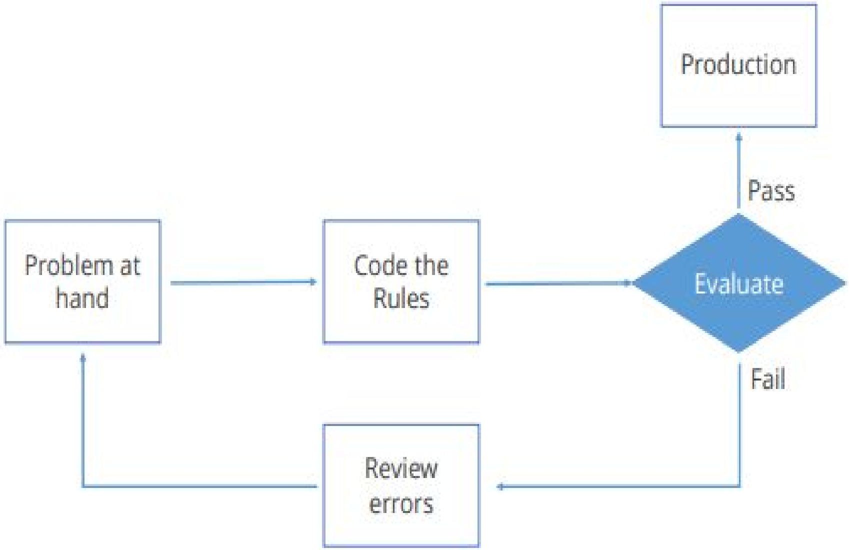
Given below is an overview of Traditional Vs Machine Learning.

Traditional programming

Machine Learning Approach

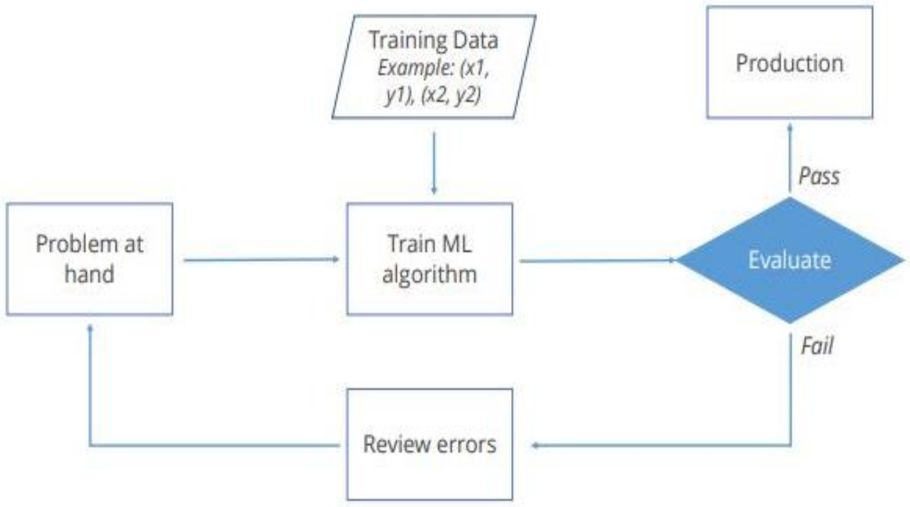
Comparison with Traditional Programming

Traditional programming relies on hard-coded rules.



Comparison with Traditional Programming

Machine Learning relies on learning patterns based on sample data.



Magic?

No, more like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

So what the machine learning is...

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Machine Learning Techniques

Given below are some techniques in Machine Learning world:

- Classification
- Categorization
- Clustering
- Trend analysis
- Anomaly detection
- Visualization
- Decision making

Features of Machine Learning

Let us look at some of the features of Machine Learning

- Machine Learning is computing-intensive and generally requires a large amount of training data
- It involves repetitive training to improve the learning and decision making of algorithms
- As more data gets added, Machine Learning training can be automated for learning new data patterns and adapting its algorithm

Machine Learning Algorithms

- Machine Learning can learn from labeled data (known as supervised learning) or unlabelled data (known as unsupervised learning)

- Machine Learning algorithms involving unlabelled data, or unsupervised learning, are more complicated than those with the labeled data or supervised learning

- Machine Learning algorithms can be used to make decisions in subjective areas as well

Examples

- Logistic Regression can be used to predict which party will win at the ballots.
- Naïve Bayes algorithm can separate valid emails from spam.
- **Face detection**: Identify faces in images (or indicate if a face is present).
- **Email filtering**: Classify emails into spam and not-spam.
- **Medical diagnosis**: Diagnose a patient as a sufferer or non-sufferer of some disease.
- **Weather prediction**: Predict, for instance, whether or not it will rain tomorrow.

Concepts of Learning

- Learning is the process of converting experience into expertise or knowledge

- Learning can be broadly classified into three categories, as mentioned below, based on the nature of the learning data and interaction between the learner and the environment

    - Supervised Learning
    - Unsupervised Learning
    - Semi-supervised learning

Types of Learning

- **Supervised (inductive) learning**
  - Training data includes desired outputs
- **Unsupervised learning**
  - Training data does not include desired outputs
- **Semi-supervised learning**
  - Training data includes a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Supervised Learning

- A majority of practical machine learning uses supervised learning

- In supervised learning, the system tries to learn from the previous examples that are given. (On the other hand, in unsupervised learning, the system attempts to find the patterns directly from the example given.)

- Speaking mathematically, supervised learning is where you have both input variables (x) and output variables(Y) and can use an algorithm to derive the mapping function from the input to the output

- **The mapping function is expressed as Y = f(X).**

Supervised Learning

- When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of Supervised learning

- This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples

Categories of Supervised learningj

- Supervised learning problems can be further divided into two parts, namely **classification**, and **regression**
- **Classification**: A classification problem is when the output variable is a category or a group, such as “black” or “white” or “spam” and “no spam”
- **Regression**: A regression problem is when the output variable is a real value, such as “Rupees” or “height”

Unsupervised Learning

- In unsupervised learning, the algorithms are left to themselves to discover interesting structures in the data
- Mathematically, unsupervised learning is when you only have input data (X) and no corresponding output variables
- This is called unsupervised learning because unlike supervised learning above, there are no given correct answers and the machine itself finds the answers

Unsupervised Learning

- Unsupervised learning is used to detect anomalies, outliers, such as fraud or defective equipment, or to group customers with similar behaviours for a sales campaign

- It is the opposite of supervised learning. There is no labelled data here

- When learning data contains only some indications without any description or labels, it is up to the coder or to the algorithm to find the structure of the underlying data, to discover hidden patterns, or to determine how to describe the data

- This kind of learning data is called **unlabelled data**

Categories of Unsupervised learning

- Unsupervised learning problems can be further divided into **association** and **clustering** problems

- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as “people that buy X also tend to buy Y”

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour

Reinforcement Learning

- A computer program will interact with a dynamic environment in which it must perform a particular goal (such as playing a game with an opponent or driving a car)

- The program is provided feedback in terms of rewards and punishments as it navigates its problem space

- Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it continuously trains itself using trial and error method

- Here learning data gives feedback so that the system adjusts to dynamic conditions in order to achieve a certain objective. The system evaluates its performance based on the feedback responses and reacts accordingly. The best known instances include self-driving cars and chess master algorithm AlphaGo.

Semi-supervised learning

- If some learning samples are labelled, but some other are not labelled, then it is semi- supervised learning

- It makes use of a large amount of **unlabeled data for training** and a small amount of **labelled data for testing**. Semi-supervised learning is applied in cases where it is expensive to acquire a fully labelled dataset while more practical to label a small subset

- ***For example***, it often requires skilled experts to label certain remote sensing images, and lots of field experiments to locate oil at a particular location, while acquiring unlabeled data is relatively easy

- **Here** an incomplete training signal is given: a training set with some (often many) of the target outputs missing. There is a special case of this principle known as Transduction where the entire set of problem instances is known at learning time, except that part of the targets are missing.

Categorizing on the basis of required Output

Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system:

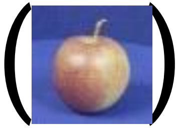
- **Classification :** When inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.

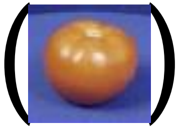
- **Regression :** Which is also a supervised problem, A case when the outputs are continuous rather than discrete

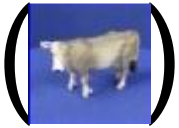
- **Clustering :** When a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

The machine learning framework

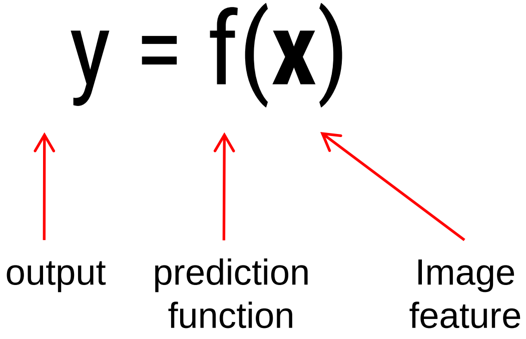
- Apply a prediction function to a feature representation of the image to get the desired output:

f() = “apple”

f() = “tomato”

f() = “cow”

The machine learning framework



$$y = f(\mathbf{x})$$

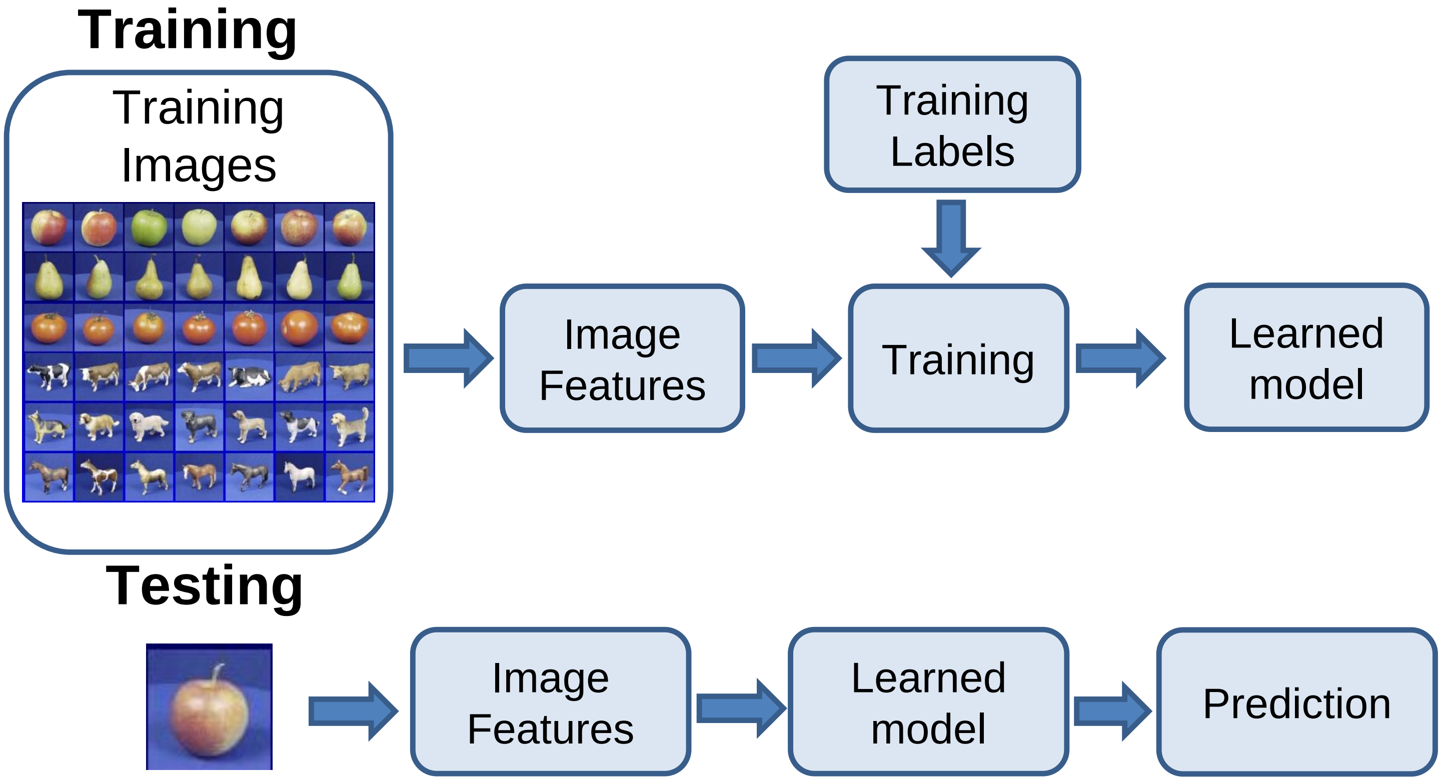
output

prediction function

Image feature

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to a never before seen *test example* **x** and output the predicted value $y = f(\mathbf{x})$

Steps



Training

Training Images

Image Features

Training Labels

Training

Learned model

Testing

Image Features

Learned model

Prediction

Test Image

Generalization

Training set (labels known)

Test set (labels unknown)

- How well does a learned model generalize from the data it was trained on to a new test set?

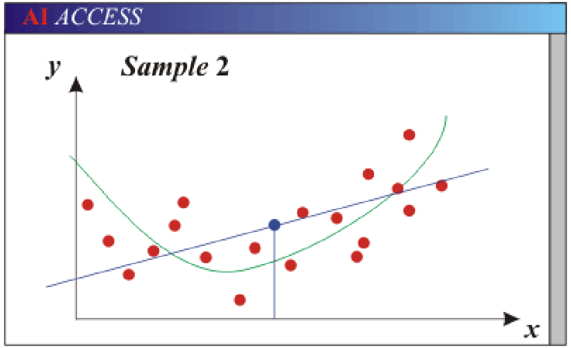
Generalization

- Components of generalization error
  - Bias: how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - Variance: how much models estimated from different training sets differ from each other

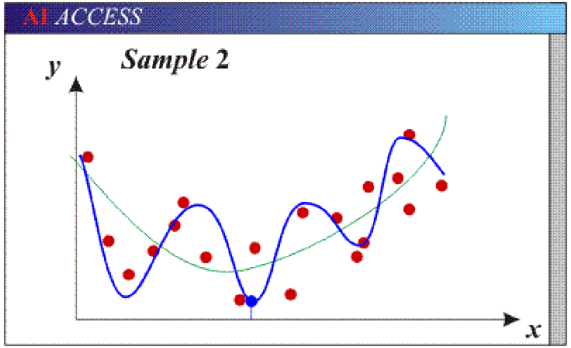
- Underfitting: model is too “simple” to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error

- Overfitting: model is too “complex” and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

Bias-Variance Trade-off



- Models with too few parameters are inaccurate because of a large bias (not enough flexibility).



- Models with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).

Bias-Variance Trade-off

$$E(MSE) = \underbrace{noise^2}_{\text{Unavoidable error}} + \underbrace{bias^2}_{\text{Error due to incorrect assumptions}} + \underbrace{variance}_{\text{Error due to variance of training samples}}$$

Python & Machine Learning

- Python is a popular platform used for research and development of production systems
- It is a vast language with number of modules, packages and libraries that provides multiple ways of achieving a task
- Python and its libraries like NumPy, Pandas, SciPy, Scikit-Learn, Matplotlib are used in data science and data analysis
- They are also extensively used for creating scalable machine learning algorithms

Python & Machine Learning

- Python implements popular machine learning techniques such as Classification, Regression, Recommendation, and Clustering

- Python offers ready-made framework for performing data mining tasks on large volumes of data effectively in lesser time

Libraries and Packages

- To understand machine learning, you need to have basic knowledge of Python programming.

- In addition, there are a number of libraries and packages generally used in performing various machine learning tasks as listed below:

- **numpy** - is used for its N-dimensional array objects
- **pandas** – is a data analysis library that includes dataframes
- **matplotlib** – is 2D plotting library for creating graphs and plots
- **scikit-learn** - the algorithms used for data analysis and data mining tasks
- **seaborn** – a data visualization library based on matplotlib

Thank You

Digital Vidya