Machine Learning

S. Sridhar and M. Vijayalakshmi
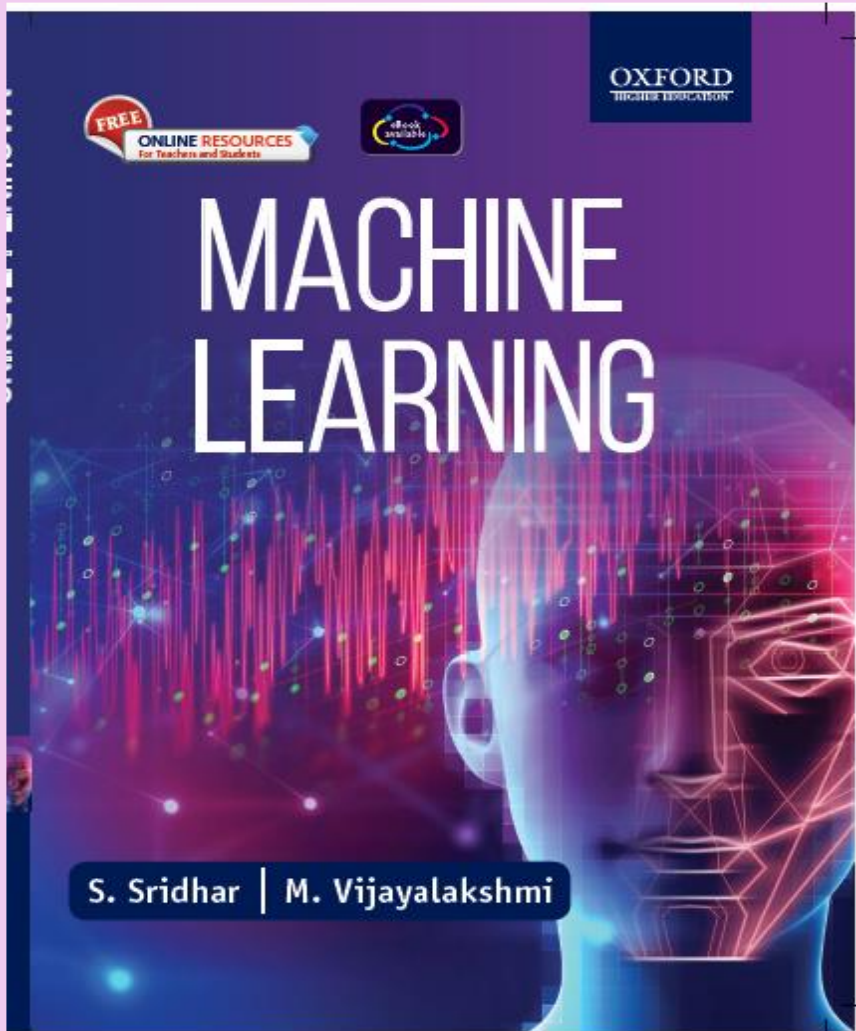
Chapter 1

Introduction to Machine Learning

Need for Machine Learning?

- BUSINESS ORGANIZATIONS HAVE NUMEROUS DATA

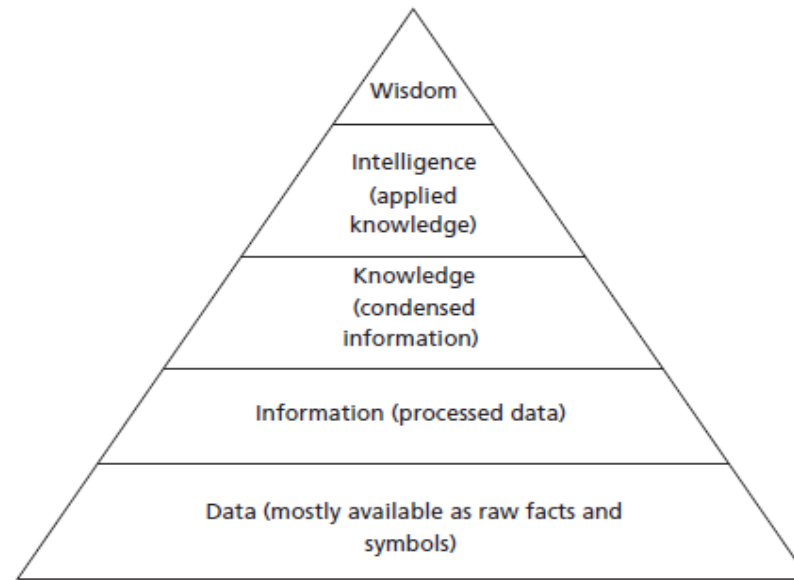- NEED TO ANALYZE DATA FOR TAKING DECISIONS

Figure 1.1: The Knowledge Pyramid

Popularity for Machine Learning

- HIGH VOLUME PF DATA

- REDUCED COST OF STORAGE

- AVAILABILITY OF COMPLEX ALGORITHMS AND TOOLS

What is Machine Learning?

- "MACHINE LEARNING IS A FIELD OF STUDY THAT GIVES THE COMPUTERS TO LEARN WITHOUT BEING EXPLICITLY PROGRAMMED"

Figure 1.2: (a) A Learning System for Humans (b) A Learning System for Machine Learning
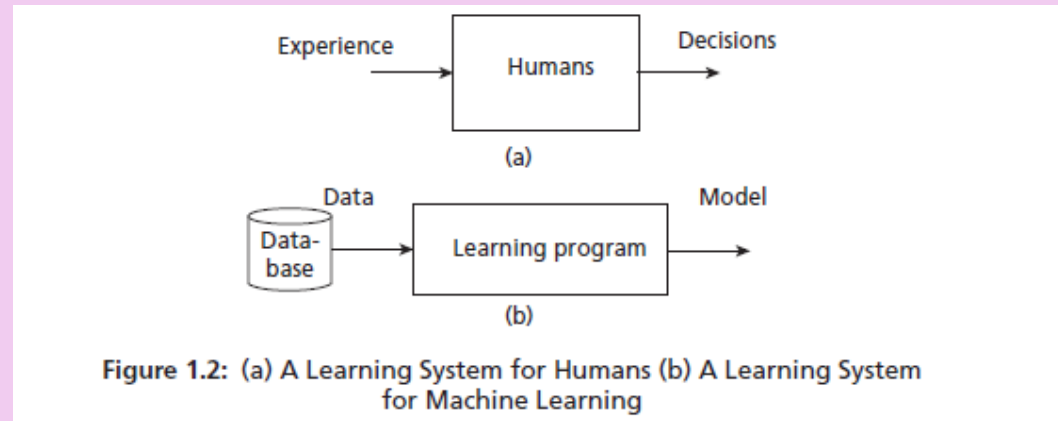
What is a Model?

A MODEL CAN BE ANY ONE OF THE FOLLOWING –

- MATHEMATICAL EQUATION

- RELATIONAL DIAGRAMS LIKE GRAPHS/TREES

- LOGICAL IF/ELSE RULES

- GROUPINGS CALLED CLUSTERS

Another View of Machine Learning

- TOM MITCHELL DEFINITION OF MACHINE LEARNING

Another pioneer of AI, Tom Mitchell's definition of machine learning states that, "*A computer program is said to learn from experience E, with respect to task T and some performance measure P, if its performance on T measured by P improves with experience E.*" The important components of this definition are experience E, task T, and performance measure P.

Machine Learning and AI

- RELATION BETWEEN MACHINE LEARNING AND AI

Figure 1.3: Relationship of AI with Machine Learning
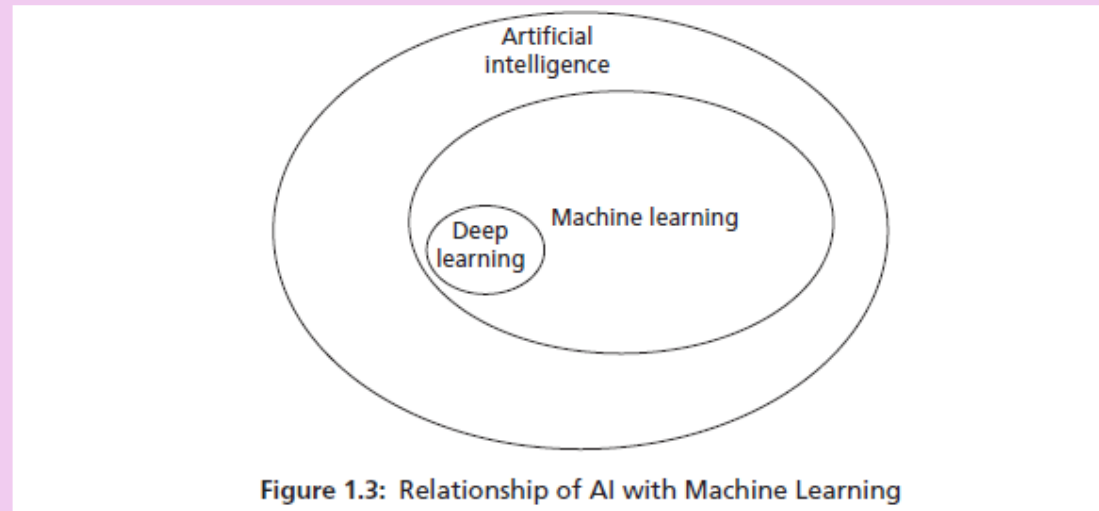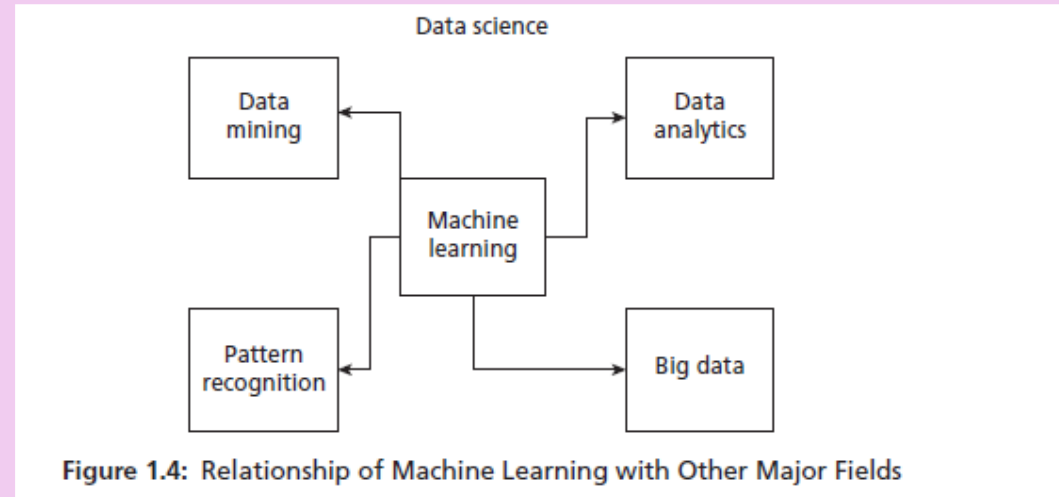
Machine Learning and Data Science

- DATA SCIENCE IS AN "UMBRELLA TERM" COVERING FROM DATA COLLECTION TO DATA ANALYSIS.

Figure 1.4: Relationship of Machine Learning with Other Major Fields

Machine Learning and Data Science

- CHARACTERISTICS OF BIG DATA

Big Data Data science concerns about collection of data. Big data is a field of data science that deals with data's following characteristics:

1. Volume: Huge amount of data is generated by big companies like Facebook, Twitter, YouTube.
2. Variety: Data is available in variety of forms like images, videos, and in different formats.
3. Velocity: It refers to the speed at which the data is generated and processed.

Machine Learning and Data Science

- DATA SCIENCE AND DATA MINING

Data Mining Data mining's original genesis is in the business. Like while mining the earth one gets into precious resources, it is often believed that unearthing of the data produces hidden information that otherwise would have eluded the attention of the management. Nowadays, many consider that data mining and machine learning are same. There is no difference between these fields except that data mining aims to extract the hidden patterns that are present in the data, whereas, machine learning aims to use it for prediction.

Machine Learning and Data Science

- DATA SCIENCE AND DATA ANALYTICS / PATTERN RECOGNITION

Data Analytics Another branch of data science is data analytics. It aims to extract useful knowledge from crude data. There are different types of analytics. Predictive data analytics is used for making predictions. Machine learning is closely related to this branch of analytics and shares almost all algorithms.

Pattern Recognition It is an engineering field. It uses machine learning algorithms to extract the features for pattern analysis and pattern classification. One can view pattern recognition as a specific application of machine learning.

Machine Learning and Statistics

ROLE OF STATISTICS

Statistics is a branch of mathematics that has a solid theoretical foundation regarding statistical learning. Like machine learning (ML), it can learn from data. But the difference between statistics and ML is that statistical methods look for regularity in data called patterns. Initially, statistics sets a hypothesis and performs experiments to verify and validate the hypothesis in order to find relationships among data.
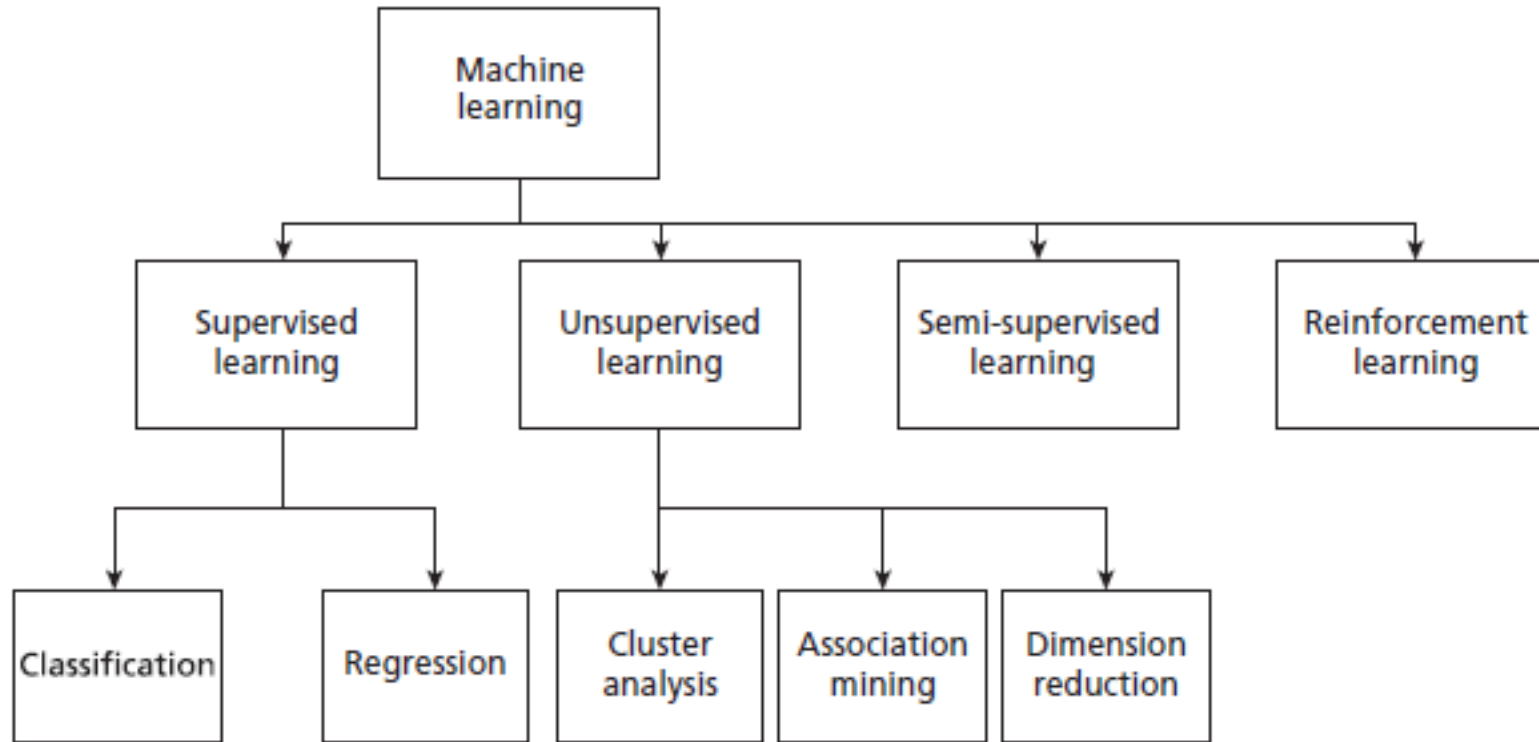
Machine Learning Types

Figure 1.5: Types of Machine Learning

Labelled Data

Table 1.1: Iris Flower Dataset

S.No.	Length of Petal	Width of Petal	Length of Sepal	Width of Sepal	Class
1.	5.5	4.2	1.4	0.2	Setosa
2.	7	3.2	4.7	1.4	Versicolor
3.	7.3	2.9	6.3	1.8	Virginica

A dataset need not be always numbers. It can be images or video frames. Deep neural networks can handle images with labels. In the following Figure 1.6, the deep neural network takes images of dogs and cats with labels for classification.

Unlabelled Data

DATA THAT IS NOT ASSOCIATED WITH LABELS IS CALLED UNLABELLED DATA.

Supervised Learning

Figure 1.7: An Example Classification System

Supervised Learning

Some of the key algorithms of classification are:

- Decision Tree
- Random Forest
- Support Vector Machines
- Naïve Bayes
- Artificial Neural Network and Deep Learning networks like CNN

Supervised Learning

REGRESSION ALGORITHMS

Figure 1.8: A Regression Model of the Form $y = ax + b$
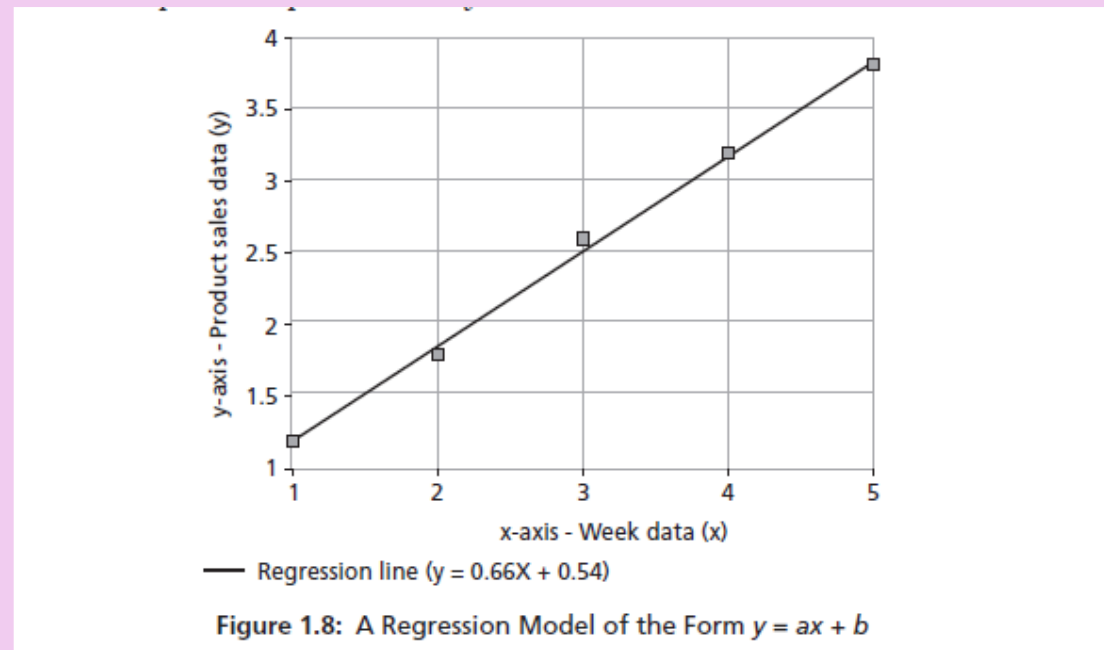
product sales = 0.66 × Week + 0.54.

Unsupervised Learning

CLUSTERING IS A GROUPING PROCESS.

Figure 1.9: An Example Clustering Scheme
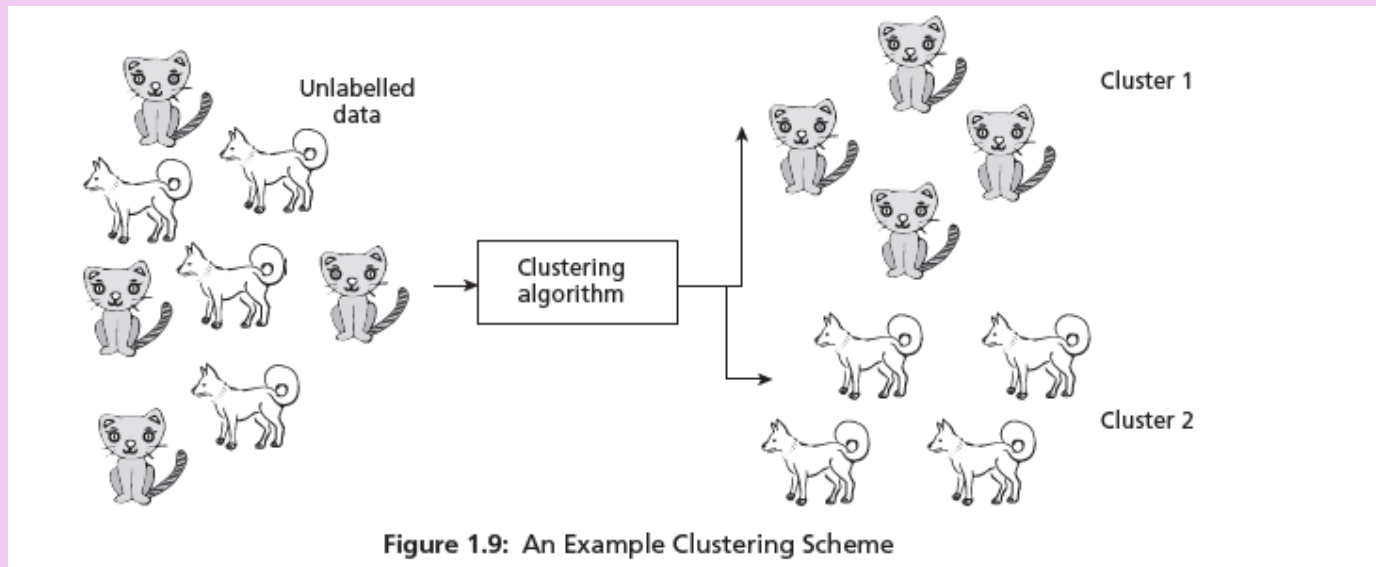
Unsupervised Learning

KEY ALGORITHMS OF UNSUPERVISED LEARNING

- k-means algorithm
- Hierarchical algorithms

Key Differences

S.No.	Supervised Learning	Unsupervised Learning
1.	There is a supervisor component	No supervisor component
2.	Uses Labelled data	Uses Unlabelled data
3.	Assigns categories or labels	Performs grouping process such that similar objects will be in one cluster

Semi-supervised Learning

There are circumstances where the dataset has a huge collection of unlabelled data and some labelled data. Labelling is a costly process and difficult to perform by the humans. Semi-supervised algorithms use unlabelled data by assigning a pseudo-label. Then, the labelled and pseudo-labelled dataset can be combined.

Reinforcement Learning

Reinforcement learning mimics human beings. Like human beings use ears and eyes to perceive the world and take actions, reinforcement learning allows the agent to interact with the environment to get rewards. The agent can be human, animal, robot, or any independent program. The rewards enable the agent to gain experience. The agent aims to maximize the reward.

The reward can be positive or negative (Punishment). When the rewards are more, the behavior gets reinforced and learning becomes possible.

Figure 1.10: A Grid game
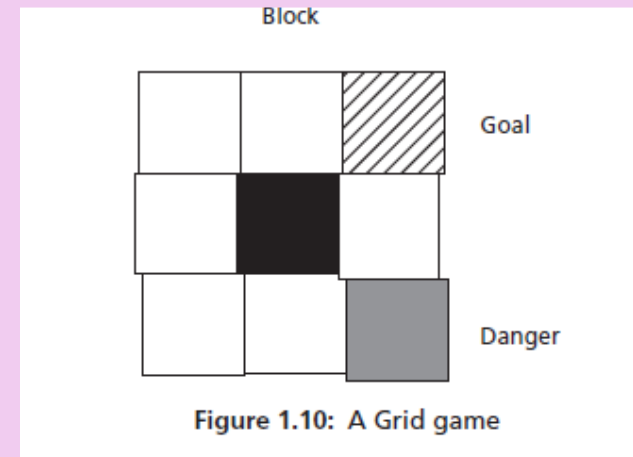
Challenges of Machine Learning

1. ILL-POSED PROBLEMS – PROBLEMS WHOSE SPECIFICATIONS ARE NOT CLEAR

2. HUGE DATA

3. HUGE COMPUTATION POWER

4. COMPLEXITY OF ALGORITHMS

5. BIAS-VARIANCE

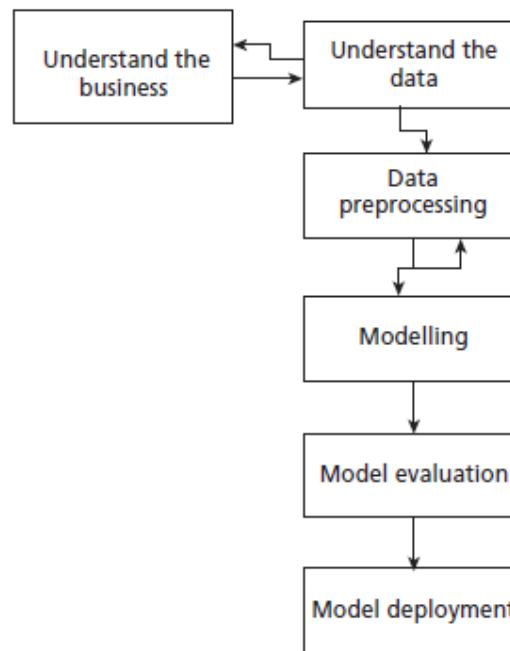# Machine Learning Process

MACHINE LEARNING/DATA MINING PROCESS

Figure 1.11: A Machine Learning/Data Mining Process

Machine Learning Applications

MACHINE LEARNING MAJOR APPLICATIONS

1. Sentiment analysis – This is an application of natural language processing (NLP) where the words of documents are converted to sentiments like happy, sad, and angry which are captured by emoticons effectively. For movie reviews or product reviews, five stars or one star are automatically attached using sentiment analysis programs.

2. Recommendation systems – These are systems that make personalized purchases possible. For example, Amazon recommends users to find related books or books bought by people who have the same taste like you, and Netflix suggests shows or related movies of your taste. The recommendation systems are based on machine learning.

3. Voice assistants – Products like Amazon Alexa, Microsoft Cortana, Apple Siri, and Google Assistant are all examples of voice assistants. They take speech commands and perform tasks. These chatbots are the result of machine learning technologies.

4. Technologies like Google Maps and those used by Uber are all examples of machine learning which offer to locate and navigate shortest paths to reduce time.

Machine Learning Applications

S.No.	Problem Domain	Applications
1.	Business	Predicting the bankruptcy of a business firm
2.	Banking	Prediction of bank loan defaulters and detecting credit card frauds
3.	Image Processing	Image search engines, object identification, image classification, and generating synthetic images
4.	Audio/Voice	Chatbots like Alexa, Microsoft Cortana. Developing chatbots for customer support, speech to text, and text to voice
5.	Telecommuni-cation	Trend analysis and identification of bogus calls, fraudulent calls and its callers, churn analysis
6.	Marketing	Retail sales analysis, market basket analysis, product performance analysis, market segmentation analysis, and study of travel patterns of customers for marketing tours
7.	Games	Game programs for Chess, GO, and Atari video games
8.	Natural Language Translation	Google Translate, Text summarization, and sentiment analysis
9.	Web Analysis and Services	Identification of access patterns, detection of e-mail spams, viruses, personalized web services, search engines like Google, detection of promotion of user websites, and finding loyalty of users after web page layout modification
10.	Medicine	Prediction of diseases, given disease symptoms as cancer or diabetes. Prediction of effectiveness of the treatment using patient history and Chatbots to interact with patients like IBM Watson uses machine learning technologies.
11.	Multimedia and Security	Face recognition/identification, biometric projects like identification of a person from a large image or video database, and applications involving multimedia retrieval
12.	Scientific Domain	Discovery of new galaxies, identification of groups of houses based on house type/geographical location, identification of earthquake epicenters, and identification of similar land use

Summary

1. Machine learning can enable top management of an organization to extract the knowledge from the data stored in various archives to facilitate decision making.

2. Machine learning is an important subbranch of Artificial Intelligence (AI).

3. A model is an explicit description of patterns within the data.

4. A model can be a formula, procedure or representation that can generate data decisions.

5. Humans predict by remembering the past, then formulate the strategy and make a prediction. In the same manner, the computers can predict by following the process.

6. Machine learning is an important branch of AI. AI is a much broader subject. The aim of AI is to develop intelligent agents. An agent can be a robot, humans, or other autonomous systems.

Summary

7. Deep learning is a branch of machine learning. The difference between machine learning and deep learning is that models are constructed using neural network technology in deep learning. Neural networks are models constructed based on the human neuron models.

8. Data science deals with gathering of data for analysis. It is a broad field that includes other fields.

9. Data analytics aims to extract useful knowledge from crude data. There are many types of analytics. Predictive data analytics is an area that is dedicated for making predictions. Machine learning is closely related to this branch of analytics and shares almost all algorithms.

10. One can say thus there are two types of data – labelled data and unlabelled data. The data with a label is called labelled data and those without a label are called unlabelled data.

11. Supervised algorithms use labelled dataset. As the name suggests, there is a supervisor or teacher component in supervised learning. A supervisor provides the labelled data so that the model is constructed and gives test data for checking the model.

12. Classification is a supervised learning method. The input attributes of the classification algorithms are called independent variables. The target attribute is called label or dependent variable. The relationship between the input and target variables is represented in the form of a structure which is called a classification model.

13. Cluster analysis is an example of unsupervised learning. It aims to assemble objects into disjoint clusters or groups.

14. Semi-supervised algorithms assign a pseudo-label for unlabelled data.

15. Reinforcement learning allows the agent to interact with the environment to get rewards. The agent can be human, animal, robot, or any independent program. The rewards enable the agent to gain experience.

16. The emerging process model for the data mining solutions for business organizations is CRISP-DM. This model stands for Cross Industry Standard Process – Data Mining.

17. Machine Learning technologies are used widely now in different domains.