AIFE: UNIT-2
MS. SHWETA TIWARI
Published: NOVEMBER, 2021

"Artificial Intelligence for Engineering/Engineers (KMC-201)"_

UNIT-2: DATA And ALGORITHMS

FALL SEMESTER, YEAR (2ND, 1ST)

FALL SESSION (2021-22)
(AIFE)
MS. SHWETA TIWARI
Published: NOVEMBER, 2021

Rajkiya Engineering College | Ambedkar Nagar, UP, India

Faculty Name: Ms. Shweta Tiwari, Subject: Artificial Intelligence for Engineering (KMC201) Year- 1st Year, Semester- 2nd Sem, Branch- CE, Session- Even Semester (2021-22)

PREPARED FOR
Engineering Students
All Engineering College

PREPARED BY
SHWETA TIWARI
Guest Faculty

November, 2021

Notes Part-1
MS. SHWETA TIWARI

UNIT-2

“Artificial Intelligence for Engineering/Engineers (KMC-201)”

By SHWETA TIWARI
Artificial Intelligence for Engineering/Engineers (KMC-201)”

UNIT-2: DATA And ALGORITHMS

November, 2021

Notes Part-1
MS. SHWETA TIWARI

UNIT-2: DATA And ALGORITHMS

November, 2021

Notes Part-1

MS. SHWETA TIWARI

Data & Algorithms Data

In general, data is any set of characters that is gathered and translated for some purpose, usually analysis. If data is not put into context, it doesn't do anything to a human or computer.

There are multiple types of data. Some of the more common types of data include the following:

Single character

Boolean (true or false)

Text (string)

Number (integer or floating-point)

Picture

Sound

Video

In a computer's storage, data is a series of bits (binary digits) that have the value one or zero. Data is processed by the CPU, which uses logical operations to produce new data (output) from source data (input).

Algorithms

Big data is data so large that it does not fit in the main memory of a single machine, and the need to process big data by efficient algorithms arises in Internet search, network traffic monitoring, machine learning, scientific computing, signal processing, and several other areas. This course will cover mathematically rigorous models for developing such algorithms, as well as some provable limitations of algorithms operating in those models. Some topics we will cover include:

Sketching and Streaming. Extremely small-space data structures that can be updated on the fly in a fast-moving stream of input.

Dimensionality reduction. General techniques and impossibility results for reducing data dimension while still preserving geometric structure.

Numerical linear algebra. Algorithms for big matrices (e.g. a user/product rating matrix for Netflix or Amazon). Regression, low rank approximation, matrix completion, ...

Compressed sensing. Recovery of (approximately) sparse signals based on few linear measurements.

November, 2021

Notes Part-1

MS. SHWETA TIWARI

- **External memory and cache-obliviousness.** Algorithms and data structures minimizing I/Os for data not fitting on memory but fitting on disk. B-trees, buffer trees, multiway mergesort, ...

2.1. History of Data

The history of big data starts many years before the present buzz around Big Data. Seventy years ago the first attempt to quantify the growth rate of data in the terms of volume of data was encountered. That has popularly been known as "information explosion". We will be covering some major milestones in the evolution of "big data".

1944: Fremont Rider, based upon his observation, speculated that Yale Library in 2040 will have "approximately 200,000,000 volumes, which will occupy over 6,000 miles of shelve. From 1944 to 1980, many articles and presentations were presented that observed the 'information explosion' and the arising needs for storage capacity.

1980: In 1980, the sociologist Charles Tilly uses the term big data in one sentence "none of the big questions has actually yielded to the bludgeoning of the big-data people." in his article "The old- new social history and the new old social history". But the term used in this sentence is not in the context of the present meaning of Big Data today.

1997: In 1977, Michael Cox and David Ellsworth published the article "Application-controlled demand paging for out-of-core visualization" in the Proceedings of the IEEE 8th conference on Visualization. The article uses the big data term in the sentence"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources.".

1998: In 1998, John Mashey, who was Chief Scientist at SGI presented a paper titled “Big Data… and the Next Wave of Infrastress.” at a USENIX meeting. John Mashey used this term in his various speeches and that’s why he got the credit for coining the term Big Data.

November, 2021

Notes Part-1

MS. SHWETA TIWARI

2000: In 2000, Francis Diebold presented a paper titled "' Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting" to the Eighth World Congress of the Econometric Society.

2001: In 2001, Doug Laney, who was an analyst with the Meta Group (Gartner), presented a research paper titled "3D Data Management: Controlling Data Volume, Velocity, and Variety." The 3V's have become the most accepted dimensions for defining big data.

2005: In 2005, Tim O'Reilly published his groundbreaking article "What is Web 2.0?". In this article, Tim O'Reilly states that the "data is the next Intel inside". O'Reilly Media explicitly used the term 'Big Data' to refer to the large sets of data which is almost impossible to handle and process using the traditional business intelligence tools.

In 2005 Yahoo used Hadoop to process petabytes of data which is now made open-source by Apache Software Foundation. Many companies are now using Hadoop to crunch Big Data.

So we can say that 2005 is the year that the Big data revolution has truly begun and the rest they say is history.

2.2. Data Storage And Importance of Data and its Acquisition

The systems, used for data acquisition are known as **data acquisition systems**. These data acquisition systems will perform the tasks such as conversion of data, storage of data, transmission of data and processing of data.

Data acquisition systems consider the following **analog signals**.

> Analog signals, which are obtained from the direct measurement of electrical quantities such as DC & AC voltages, DC & AC currents, resistance and etc.

Analog signals, which are obtained from transducers such as LVDT, Thermocouple & etc.

Types of Data Acquisition Systems

Data acquisition systems can be classified into the following **two types**.

Analog Data Acquisition Systems

Digital Data Acquisition Systems

Now, let us discuss about these two types of data acquisition systems one by one.

Analog Data Acquisition Systems

The data acquisition systems, which can be operated with analog signals are known as **analog data acquisition systems**. Following are the blocks of analog data acquisition systems.

- **Transducer** − It converts physical quantities into electrical signals.
- **Signal conditioner** − It performs the functions like amplification and selection of desired portion of the signal.
- **Display device** − It displays the input signals for monitoring purpose.
- **Graphic recording instruments** − These can be used to make the record of input data permanently.
- **Magnetic tape instrumentation** − It is used for acquiring, storing & reproducing of input data.

Digital Data Acquisition Systems

The data acquisition systems, which can be operated with digital signals are known as **digital data acquisition systems**. So, they use digital components for storing or displaying the information.

Mainly, the following **operations** take place in digital data acquisition.

- Acquisition of analog signals
- Conversion of analog signals into digital signals or digital data
- Processing of digital signals or digital data

Following are the blocks of **Digital data acquisition systems**.

- **Transducer** − It converts physical quantities into electrical signals.
- **Signal conditioner** − It performs the functions like amplification and selection of desired portion of the signal.
- **Multiplexer** − connects one of the multiple inputs to output. So, it acts as parallel to serial converter.
- **Analog to Digital Converter** − It converts the analog input into its equivalent digital output.
- **Display device** − It displays the data in digital format.

Digital Recorder – It is used to record the data in digital format.

Data acquisition systems are being used in various applications such as biomedical and aerospace. So, we can choose either analog data acquisition systems or digital data acquisition systems based on the requirement.

2.3. The Stages of data Processing

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

Six stages of data processing

- **Data collection**

Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

- **Data preparation**

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as “pre-processing” is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create

high-quality data for the best business intelligence.

- **Data input**

The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

- **Processing**

During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks,

connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

- **Data output/interpretation**

The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

- **Data storage**

The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Plus, properly stored data is a necessity for compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

The future of data processing

The future of data processing lies in the cloud. Cloud technology builds on the convenience of current electronic data processing methods and accelerates its speed and effectiveness. Faster, higher-quality data means more data for each organization to utilize and more valuable insights to extract.

2.4. Data Visualization

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to

drill down into charts and graphs for more detail, interactively changing what data you see and how it’s processed.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

The different types of visualizations

When you think of data visualization, your first thought probably immediately goes to simple bar graphs or pie charts. While these may be an integral part of visualizing data and a common baseline for many data graphics, the right visualization must be paired with the right set of information. Simple graphs are only the tip of the iceberg. There's a whole selection of visualization methods to present data in effective and interesting ways.

Common general types of data visualization:

Charts

Tables

Graphs

Maps

Infographics

Dashboards

More specific examples of methods to visualize data:

Area Chart

Bar Chart

Box-and-whisker Plots

Bubble Cloud

Bullet Graph

Cartogram

Circle View

Dot Distribution Map

Gantt Chart

November, 2021

Notes Part-1

MS. SHWETA TIWARI

Heat Map

Highlight Table

Histogram

Matrix

Network

Polar Area

Radial Tree

Scatter Plot (2D or 3D)

Streamgraph

November, 2021

Notes Part-1

MS. SHWETA TIWARI

- Text Tables
- Timeline
- Treemap
- Wedge Stack Graph
- Word Cloud
- And any mix-and-match combination in a dashboard!

2.5 Regression, Prediction & Classification

Regression Analysis

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Table 2.1: Sales data

The company wants to do the advertisement of $200 in the year 2019 and wants to know the prediction about the sales for this year. So to solve such type of prediction problems in machine learning, we need regression analysis.

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, **"Regression shows a line or curve that passes through all the datapoints on target- predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum."** The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving

Linear Regression:

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and

the dependent variable (Y-axis), hence called linear regression.

- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.

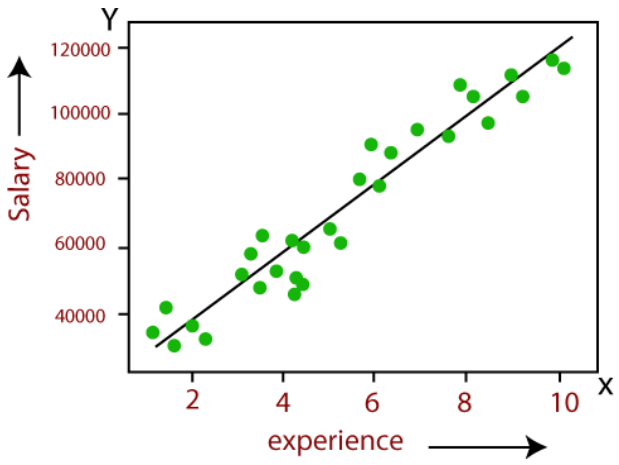

Figure 2.1: Linear regression

- Below is the mathematical equation for Linear regression:

$$Y= aX+b$$

Here, Y = dependent variables (target variables),

X= Independent variables (predictor variables), a

and b are the linear coefficients

Some popular applications of linear regression are:

- Analyzing trends and sales estimates

- Salary forecasting
- Real estate prediction
- Arriving at ETAs in traffic.

Classification

A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

For example, when filtering emails “spam” or “not spam”, when looking at transaction data, “fraudulent”, or “authorized”. In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and the values (class labels) in

classifying attributes and uses it in classifying new data. There are a number of classification models. Classification models include logistic regression, decision tree, random forest, gradient- boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes.

Types of Classification Algorithms

Linear Models

- Logistic Regression
- Support Vector Machines

Nonlinear models

- K-nearest Neighbors (KNN)
- Kernel Support Vector Machines (SVM)
- Naïve Bayes
- Decision Tree Classification
- Random Forest Classification

2.6. Clustering & Recommender Systems

Introduction to Clustering

It is basically a type of unsupervised learning method . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples. **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between

them.

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.

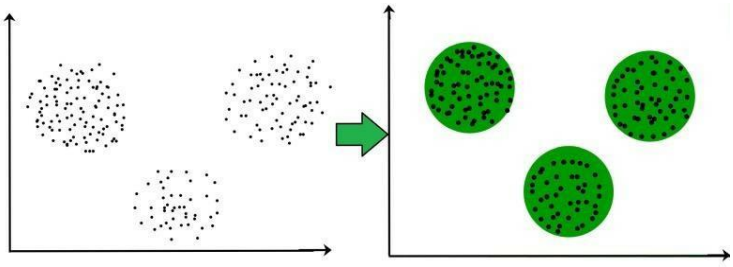

Figure 2.2: Clustering

Clustering Methods :

- **Density-Based Methods :** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters.Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise) , OPTICS (Ordering Points to Identify Clustering Structure) etc.

- **Hierarchical Based Methods :** The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category

  - **Agglomerative** (bottom up approach)
  - **Divisive** (top down approach)

  examples CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies) etc.

- **Partitioning Methods :** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon Randomized Search) etc.

- **Grid-based Methods :** In this method the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operation done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest) etc.

Recommender systems

Recommender systems are the systems that are designed to recommend things to the user based on many different factors. These systems predict the most likely product that the users are most likely to purchase and are of interest to. Companies like Netflix, Amazon, etc. use recommender systems to help their users to identify the correct product or movies for them.

The recommender system deals with a large volume of information present by filtering the most important information based on the data provided by a user and other factors that take care of the user's preference and interest. It finds out the match between user and item and imputes the similarities between users and items for recommendation.

November, 2021

Notes Part-1

MS. SHWETA TIWARI

Both the users and the services provided have benefited from these kinds of systems. The quality and decision-making process has also improved through these kinds of systems.

Why the Recommendation system?

- Benefits users in finding items of their interest.
- Help item providers in delivering their items to the right user.
- Identity products that are most relevant to users.
- Personalized content.
- Help websites to improve user engagement.

What can be Recommended?

There are many different things that can be recommended by the system like movies, books, news, articles, jobs, advertisements, etc. Netflix uses a recommender system to recommend movies & web-series to its users. Similarly, YouTube recommends different videos. There are many examples of recommender systems that are widely used today.

Types of Recommendation System

Popularity-Based Recommendation System

It is a type of recommendation system which works on the principle of popularity and or anything which is in trend. These systems check about the product or movie which are in trend or are most popular among the users and directly recommend those.

For example, if a product is often purchased by most people then the system will get to know that that product is most popular so for every new user who just signed it, the system will recommend that product to that user also and chances becomes high that the new user will also purchase that.

Merits of popularity based recommendation system

- It does not suffer from cold start problems which means on day 1 of the business also it can recommend products on various different filters.
- There is no need for the user's historical data.
    - Demerits of popularity based recommendation system
- Not personalized
- The system would recommend the same sort of products/movies which are solely based upon popularity to every other user.

Example

- Google News: News filtered by trending and most popular news.
- YouTube: Trending videos.

Questions

1. What is Data and Big Data?
2. What is algorithm and is properties?
3. Explain data and its acquisition.
4. What are the stages involve in data processing?
5. Define data visualization.
6. How many types of data visualization.
7. What is data classification and Regression?
8. What is data clustering? Explain any one method in details.
9. What are recommender systems? How is working in OTT.
10. How many types of data acquisition systems.