UNIT-I

Introduction-Data Analytics, Data Mining and Knowledge discovery- Data and Relations- Dissimilarity Measures- Similarity Measures- Sequence Relations- Sampling and Quantization- Analysis vs Reporting- Modern Data Analytic Tools- Statistical Concepts- Probability- Sampling and Sampling Distribution- Statistical Inference- Prediction and Prediction Error- Resampling.

Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses.
Data analytics is about uncovering various relationships between different variables and uncovering patterns using automation and other techniques. e.g. if there is a relation between the purchase of soft drinks and an ice box.
Types of data analytics applications

At a high level, data analytics methodologies include **exploratory data analysis (EDA)**, which aims to find patterns and relationships in data, and **confirmatory data analysis (CDA)**, which appliesstatistical techniques to determine whether hypotheses about a data set are true or false. Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

The analytics process starts with data collection, in which data scientists identify the information they need for a particular analytics application and

then work on their own or with data engineers and IT staffers to assemble it for use. Data from different source systems may need to be combined via data integration routines, transformed into a common format and loaded into an analytics system, such as a Hadoop cluster, NoSQL database or data warehouse. In other cases, the collection process may consist of pulling a relevant subset out of a stream of raw data that flows into, say, Hadoop and moving it to a separate partition in the system so it can be analyzed without affecting the overall data set.

Once the data that's needed is in place, the next step is to find and fix data quality problems that could affect the accuracy of analytics applications. That includes running data profilingand data cleansing jobs to make sure that the information in a data set is consistent and that errors and duplicate entries are eliminated. Additional data preparation work is then done to manipulate and organize the data for the planned analytics use, and data governance policies are applied to ensure that the data hews to corporate standards and is being used properly.

At that point, the data analytics work begins in earnest. A data scientist builds an analytical model, using predictive modeling tools or other analytics software and programming languages such as Python, Scala, R and SQL. The model is initially run against a partial data set to test its accuracy; typically, it's then revised and tested again, a process known as "training" the model that continues until it functions as intended. Finally, the model is run in production mode against the full data set, something that can be done once to address a specific information need or on an ongoing basis as the data is updated.

Data Mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications –

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Data Mining Applications

Data mining is highly useful in the following domains –
- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid.

Market Analysis and Management

Listed below are the various fields of market where data mining is used –
- Customer Profiling − Data mining helps determine what kind of people buy what kind of products.

• Identifying Customer Requirements − Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
• Cross Market Analysis − Data mining performs association/correlations between product sales.
• Target Marketing − Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
• Determining Customer purchasing pattern − Data mining helps in determining customer purchasing pattern.
• Providing Summary Information − Data mining provides us various multidimensional summary reports.

Corporate Analysis and Risk Management

Data mining is used in the following fields of the Corporate Sector –

• **Finance Planning and Asset Evaluation** − It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
• **Resource Planning** − It involves summarizing and comparing the resources and spending.
• **Competition** − It involves monitoring competitors and market directions.

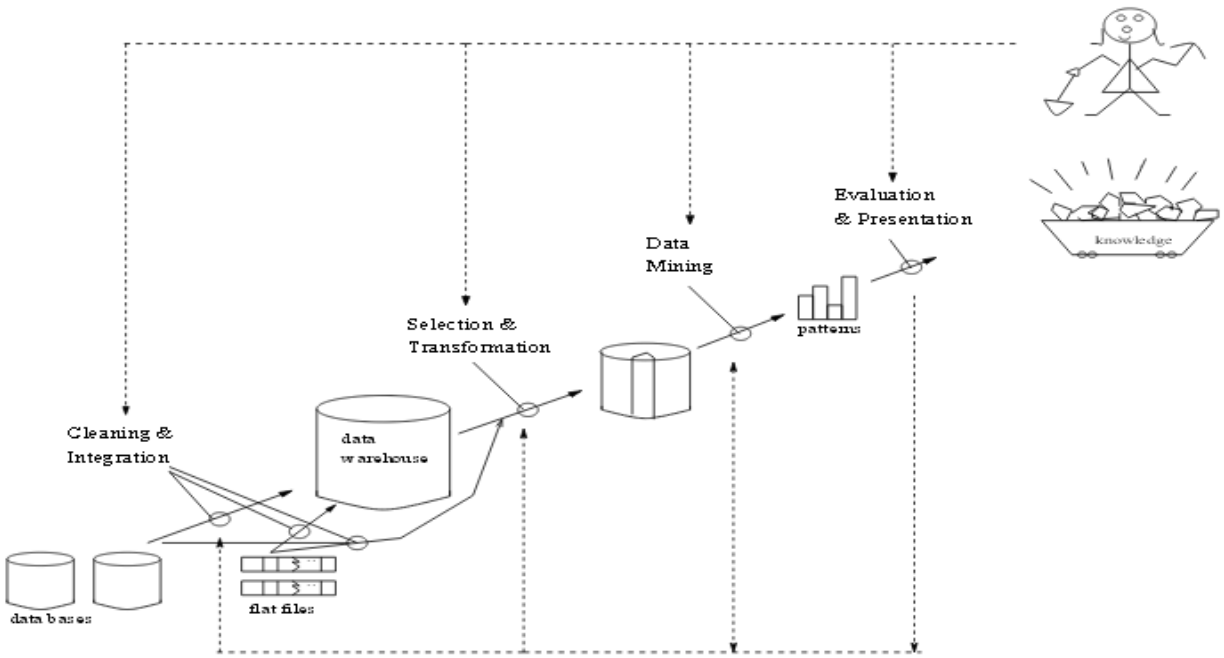
Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

Knowledge discovery in databases (KDD)

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.
Here is the list of steps involved in the knowledge discovery process –

• Data Cleaning − In this step, the noise and inconsistent data is removed.
• Data Integration − In this step, multiple data sources are combined.
• Data Selection − In this step, data relevant to the analysis task are retrieved
from the database.
• Data Transformation − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
• Data Mining − In this step, intelligent methods are applied in order to extract
data patterns.
• Pattern Evaluation − In this step, data patterns are evaluated.
• Knowledge Presentation − In this step, knowledge is represented.



Analysis Vs Reporting

Reporting: The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

Analysis: The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

Reporting translates raw data into **information**. Analysis transforms data and information into **insights**. Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected

ranges. Good reporting should **raise questions** about the business from its end users.

The goal of analysis is to **answer questions** by interpreting the data at a deeper level and providing actionable recommendations. Through the process of performing analysis you may raise additional questions, but the goal is to identify answers, or at least potential answers that can be tested. In summary, reporting shows you ***what is happening*** while analysis focuses on explaining ***why it is happening*** and ***what you can do about it.***

Three main types of reporting: canned reports, dashboards, and alerts.

1. **Canned reports:** These are the out-of-the-box and custom reports that you can access within the analytics tool or which can also be delivered on a recurring basis to a group of end users. Canned reports are fairly static with fixed metrics and dimensions. In general, some canned reports are more valuable than others, and a report's value may depend on how relevant it is to an individual's role (e.g., SEO specialist vs. web producer).
2. **Dashboards:** These custom-made reports combine different KPIs and reports to provide a comprehensive, high-level view of business performance for specific audiences. Dashboards may include data from various data sources and are also usually fairly static.
3. **Alerts:** These conditional reports are triggered when data falls outside of expected ranges or some other pre-defined criteria is met. Once people are notified of what happened, they can take appropriate action as necessary.

In contrast, analysis follows a **pull approach**, where particular data is pulled by an analyst in order to answer specific business questions.

There are two main types: *ad hoc responses* and *analysis presentations*.

1. **Ad hoc responses:** Analysts receive requests to answer a variety of business questions, which may be spurred by questions raised by the reporting. Typically, these urgent requests are time sensitive and demand a quick turnaround. The analytics team may have to juggle multiple requests at the same time. As a result, the analyses cannot go as deep or wide as the analysts may like, and the deliverable is a short and concise report, which may or may not include any specific recommendations.
2. **Analysis presentations:** Some business questions are more complex in nature and require more time to perform a comprehensive, deep-dive analysis. These analysis projects result in a more formal deliverable, which includes two key sections: key findings and recommendations. The key findings section highlights the most meaningful and actionable insights gleaned from the analyses performed. The recommendations section provides guidance on what actions to take based on the analysis findings.

Sampling

- Sampling has always been central to statistics, and has been extensively researched
- In survey research collecting data is expensive (while analyzing it is cheap), so sampling to limit data
- In big data analyzing data is expensive (while collecting it is cheap), and again sampling to limit data
- Most straightforward idea is to just sample uniformly at random

Contrast to classical setting

- Often in statistics we think about sampling values at random from some parametric distribution is order to estimate the parameters

- For example we might sample a normal distribution to estimate the mean, and in this case we know very well how for example sample mean function behaves (Student's t-distribution etc.)
- Now we are doing something different, that is sampling a finite collection which we can, if we so desire, scan several times. We can for example find the minimum and maximum values in this collection.
- This difference in setting will lead to some theory which is not so familiar from elementary statistics

Problems with uniform sampling

- Uniform sampling will sometimes yield abysmal results
- If the data is spread on a large interval, obtaining useful estimates can require large samples
- A very specific method to handle this situation

Sampling sparse data

Sparse here means that the values are spread over a long interval (not to be mixed with the usual definition of sparseness)

Stratified sampling

Idea is to split the dataset into buckets and ensure each is represented in the sample. This way we can have outliers (which are disproportionaly important for estimates) represented

Probability

The most important tool in statistical inference is probability theory. This section provides a short review of the important concepts

Random Experiments

A random experiment is an experiment that satises the following conditions

1. all possible distinct outcomes are known in advance,

2. in any particular trial, the outcome is not known in advance,
3. the experiment can be repeated under identical conditions.

The outcome space of an experiment is the set of all possible outcomes of the experiment.

Example 1. Tossing a coin is a random experiment with outcome space = {H,T}
Example 2. Rolling a die is a random experiment with outcome space = {1,2,3,4,5,6}

Something that might or might not happen, depending on the outcome of the experiment, is called an event. Examples of events are \coin lands heads" or \die shows an odd number". An event A is represented by a subset of the outcome space. For the above examples we have A = {H} and A = {1,3,5} respectively. Elements of the outcome space are called elementary events.

Classical definition of probability

If all outcomes in are equally likely, the probability of A is the number of outcomes in A, which we denote by M(A) divided by the the total number of outcomes M

$$P(A)=M(A)/M$$

Probability axioms

Probability is dened as a function from subsets of $\Omega$to the real line IR, that satisfies the following axioms

1.Non-negativity: P (A)>=0
2. Additivity: If $A \cap B = \varnothing$; then $P (A \cup B) = P (A) + P (B)$

3. $P (\Omega)=1$ The classical, frequency and subjective definitions of probability all satisfy these axioms. Therefore every property that may be deduced from these axioms holds for all three interpretations of probability.

Conditional probability and independence

The probability that event A occurs may be influenced by information concerning the occurrence of event B. The probability of event A, given that B will occur or has occurred, is called the conditional probability of A given B, denoted by P (A / B). It follows from the axioms of probability that

$$P(A / B) = P(A \cap B) / P(B)$$

for P (B) > 0. Intuitively we can appreciate this equality by considering that B effectively becomes the new outcome space. The events A and B are called independent if the occurrence of one event does not influence the probability of occurrence of the other event, i.e.

$$P(A / B) = P(A), \text{ and consequently } P(B / A) = P(B)$$

Since independence of two events is always mutual, it is more concisely expressed by the product rule

$$P(A \cap B) = P(A)\, P(B)$$

Statistical Inference

The relation between sample data and population may be used for reasoning in two directions: from known population to yet to be observed sample data, and from observed data to (partially) unknown population. This last direction of reasoning is of inductive nature and is addressed in statistical inference. It is the form of reasoning most relevant to data analysis, since one typically has available one set of sample data from which one intends to draw conclusions about the unknown population.

Frequentist Inference According to frequentists, inference procedures should be interpreted and evaluated in terms of their behavior in hypothetical repetitions under the same conditions. To quote David S. Moore, the frequentist consistently asks \What would happen if we did this many times?". To answer this question, the sampling distribution of a statistic is of crucial importance. The two basic types of frequentist inference are estimation and testing. In estimation one wants to come up with a plausible value or range of plausible values for an unknown population parameter. In testing one wants to decide whether a hypothesis

concerning the value of an unknown population parameter should be accepted or rejected in the light of sample data.

Point Estimation In point estimation one tries to provide an estimate for an unknown population parameter, denoted by θ, with one number: the point estimate. If G denotes the estimator of θ, then the estimation error is a random variable G -θ, which should preferably be close to zero.

Similarity and Dissimilarity: measures are essential to solve many pattern recognition problems such as classification and clustering.
Similarity measures how close two distributions are:

Similarity measure:

- Numerical measure of how alike two data objects are.
- Other falls between ‘0’(no similarity) and ‘1’(complete similarity)

Dissimilarity measures include:

- Numerical measure of how different two data objects are.
- Ranges from 0(objects are alike)to infinity(objects are different).

Properties of dissimilarity for symmetric and asymmetric data points

1. d(i,j)>=0; distance is a non-negative number.
2. d(i,j)=0; the distance of an object to itself is 0.
3. d(i,j)=d(j,i); distance is a symmetric function.
4. d(i,j)<=d(i,h)+d(h,j); directly from object to object in space is no more than making detour over any other object ‘h’(triangular inequality)

Common properties of a similarity measures:
1.s(p,q)=1(or maximum similarity) only if p=q.
2.s(p,q)=s(q,p) for all p and q where s(p,q) is the similarity between data objects p and q.
This table suits whenever all binary variables are having same weight.

A contingency table for binary variables :

		Object j		
		1	0	sum
Object i	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p

q- no. of variables that equals 1 for both the objects.
r- no. of variables that equal 1 for object ‘i’ but that are 0 for object ‘j’.

Dissimilarity measures includes:

1.Euclidean Distance

2.Manhattan Distance

3.Minkowski Distance

Common Properties of Dissimilarity Measures-for Symmetric and Asymmetric data points

1. $d(i, j) \geq 0$; Distance is a non-negative number
2. *d(i,j)=0;*The distance ofan object to itself is 0
3. *d(i,j)=d(j,i)*;Distance is a symmetric function
4. d(i,j)≤d(i,h)+d(h,j);Triangular inequality

Euclidean distance:

$d(i,j)= \sqrt{(x_{i1}-x_{j1})^2 + (X_{i2} - X_{j2})^2 + \ldots\ldots\ldots\ldots\ldots\ldots + (X_{in} - X_{jn})^2}$

Where $i=(X_{i1}, X_{i2}\ldots\ldots\ldots\ldots X_{in})$ and $j=(x_{j1}, x_{j2}\ldots\ldots\ldots\ldots x_{jn})$

Manhattan distance (city block):
$d(i,j)= |x_{i1}- x_{j1}| + | x_{i2}- x_{j2}| + \text{----------} |x_{in}- x_{jn}|$
Both Euclidean and Manhattan justify the following mathematical requirements of a distance function.

Properties of dissimilarity for symmetric and asymmetric data points

1. d(i,j)>=0; distance is a non-negative number.
2. d(i,j)=0; the distance of an object to itself is 0.
3. d(i,j)=d(j,i); distance is a symmetric function.
4. d(i,j)<=d(i,h)+d(h,j); directly from object to object in space is no more than making detour over any other object ‘h’(triangular inequality)

Minkowski distance is a generation of both Euclidean and Manhattan

$d(i,j)= (|x_{i1} - X_{j1}|^p + |x_{i2} - X_{j2}|^p + \ldots\ldots\ldots |x_{in} - X_{jn}|^p)^{1/p}$

P is a positive integer such a distance is called L p norm.

It represents the manhattan distance when p=1(L1 norm) and Euclidean distance when p=2(i.e L2 norm)

Binary variables are used to compute the dissimilarity between objects described by either symmetric or asymmetric linearly variables

- Contingency table is used to compute dissimilarity between two binary variables.

Modern Data Analytic Tools

1. Tableau Public

2. OpenRefine

3. KNIME

4. RapidMiner

5. Google Fusion Tables

6. NodeXL

7. Wolfram Alpha

8. Google Search Operators

9. Solver

10. Dataiku DSS

Data Analytics Tools and Techniques

Here are some of the useful data analytics tools and techniques that can be used for performing better:

1. Visual Analytics

There are different ways to analyze the data. One of the simplest ways to do is to create a graph or visual and look at it to spot patterns. This is an integrated method that combines data analysis with human interaction and data visualization.

2. Business Experiments

Experimental design, AB testing, and business experiments are all techniques for testing the validity of something. It is trying out something in one part of the organization and comparing it with another.

3. Regression Analysis

It is a statistical tool for investigating the relationship between variables. For instance, the cause and effect relationship between product demand and price.

4. Correlation Analysis

A statistical technique that allows you to determine whether there is a relationship between two separate variables and how strong that relationship may be. It is best to use when you know or suspect that there is a relationship between two variables and wish to test the assumption.

5. Time Series Analysis

It is the data that is collected at uniformly spaced time intervals. You can use it when you want to assess changes over time or predict future events on the basis of what happened in the past.

In statisticsthe mean squared **prediction error** of a smoothing or curve fitting procedure is the expected value of the squared difference between the fitted values implied by the predictive function and the values of the (unobservable) function g.

Quantization

The main definition of quantization is: it is the division of a large quantity into a discrete number of small parts, often assumed to be integral multiples of a commonquantity. Quantization of signals plays a major role in various applications in the areasof signal processing, speech processing and Image processing.

Quantization is required to reduce the number of bits used to represent asample of speech signal which is known as speech coding. However during thisprocess, some of the parameters like Bit-rate, complexity and memory requirementalso get reduced.Therefore Quantization of the signal results in the loss in the quality of aspeech signal, which is undesirable. Hence the researchers have to compromise eitherwith the reduction in bit-rate or with the quality of speech signal.

There are two types of quantizers. They are Non-uniform quantizer anduniform quantizer. Non-uniform quantizer is the one in which the difference betweenthe quantization levels are not uniform, while Uniform

quantizer is the one in whichthe difference between the quantization levels is uniform.

TYPES OF QUANTIZATIONS:

There are mainly two types of quantizations.

1) Scalar Quantization

2) Vector Quantization

Scalar Quantization:

Quantization is an essential component of speech coding systems. Scalarquantization is the process by which the signal samples are independently quantized (Sample by Sample basis). The quantization process is based on the probabilitydensity function of the signal samples. An N-level scalar quantizer can be treated as aone-dimensional mapping of the input range R onto an index in a mapping table(or codebook) C. Thus

$$Q:R->C \qquad C \subset R$$

The receiver uses this index to reconstruct an approximation to the input level.So, to design the Scalar quantizer, the quantizers are matched to the distribution of the

source samples, which may or may not be known in advance. If the distribution is notknown in advance, an empirical choice may be made through a Gaussian or Laplaciandistribution.

VECTOR QUANTIZATION:

Vector quantization is a process by which the elements of a vector arequantized as groups called vectors. Vector quantization increases the

optimality of aquantizer and there is an increase of computational complexity and memoryrequirements. Vector quantization is more efficient than scalar quantization in termsof error at a given bit rate. The central component of a Vector Quantizer (VQ) is acodebook C of size N x k, which maps the k-dimensional spaceonto thereproduction vectors also called code vectors or code words

$$Q:R^K \rightarrow C, \quad C=(Y_1,Y_2,\ldots\ldots Y_N)^T,\ Y_i \in R^K$$