Page No.
Date

Data Science

It is an interdisciplinary field that uses algorithms, scientific tools and systems to extract knowledge and insights from structured and unstructured data.

Interdisciplinary Fields

1. Statistics → Inference, probability hypothesis testing.
2. Computer Science → Data structures, algorithms, programming
3. Machine learning :- Automated pattern detection, prediction
4. Domain Knowledge :- domain specific insights (e.g. healthcare, finance).

GOAL → Build predictive / descriptive models.
→ Drive decision-making based on data.
→ Discover meaningful pattern.
→ Generate actionable insights.

Why do we need.
→ Explosive Growth → 1. Data collection and data availability.
→ Automated data collection tools.
2. Data storage
→ terabytes to petabytes
→ Cheap, reliable and fast.
3. Physical to Digital Experience.

"Abundance of data but lack of valuable knowledge."

→ Data driven thinking. (enables)
→ Improve efficiency
→ Make better Decision
→ Predict better future
→ Understand customer behaviour.

} Business and govt. need to derive meaning from data to

→ Personalise products and services
→ Detect fraud or anomalies.
→ Traditional statistical method are insufficient for big, fast and varied data.

REAL WORLD TRENDS

1. Healthcare :- Predicts disease risk, automate diagnosis, optimise treatment plans.

2. Finance : Fraud detection, credit scoring, algorithmic trading.

3. Marketing : Customer segmentation, recommendation engines, sentiment analysis.

4. Transportation :- Predicts traffic patterns, optimise delivery logistics.

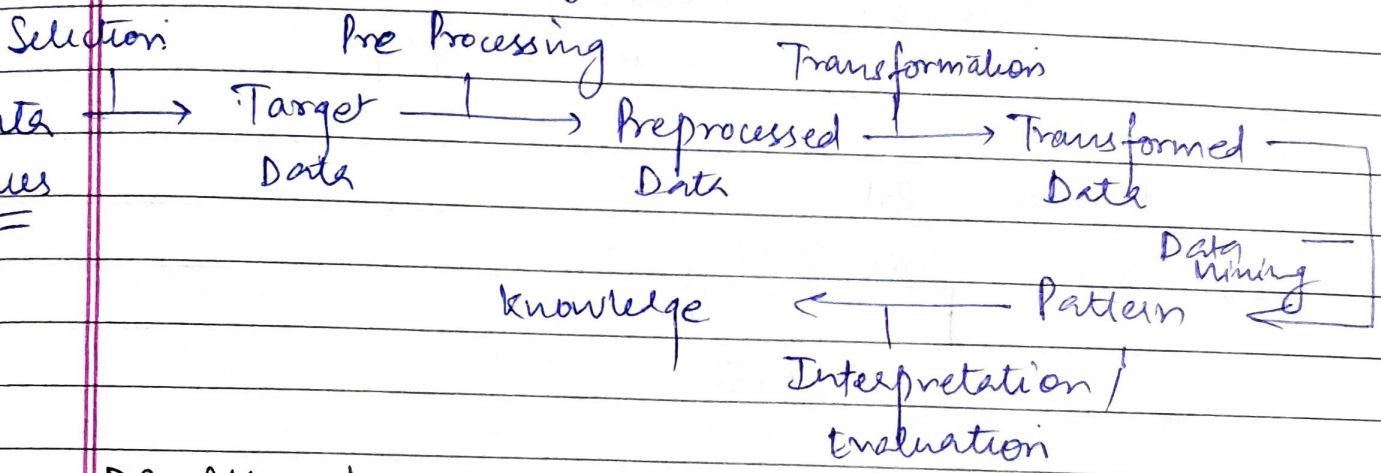
5. Environment : Climate modelling, air pollution forecasting

6. Public Health :- Google Flu Trend - predicted flu outbreaks using search data patterns, often earlier than traditional CDC methods.

Era	Paradigm	Description
Pre-1600	Empirical	Observation, Trial & Error.
1600 - 1950s	Theoretical	Laws of physics, mathematical modelling
1950s - 1990s	Computational	Simulation of complex models, numerical methods.
1990s - present	Data-Driven Science	Extracting patterns from dataset.

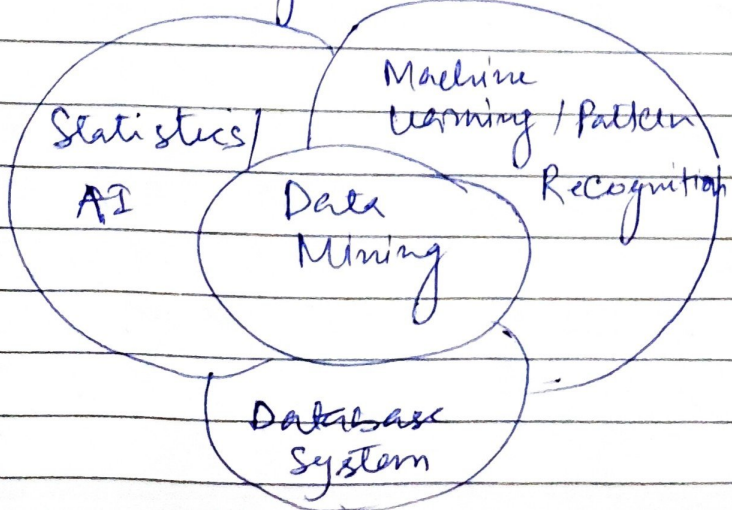
(Data Science) →

- Extraction of non-trivial, implicit, previously unknown and potentially useful patterns / structure / knowledge from huge amounts of data.
- Exploration and analysis, by automatic or semi-auto. means, of large quantities of data in order to discover meaningful patterns.



DS Alternative Names

→ Knowledge discovery (mining) in databases (KDD)
→ Knowledge extraction
→ Data / pattern analysis
→ Data archaeology
→ Data dredging.
→ Information harvesting.
→ Business intelligence etc.



Page No.
Date

Input Data → Data Pre Processing → Data Mining → Post Processing → Pattern/knowledge

Data integration
Normalization
Feature selection
Dimension Reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis

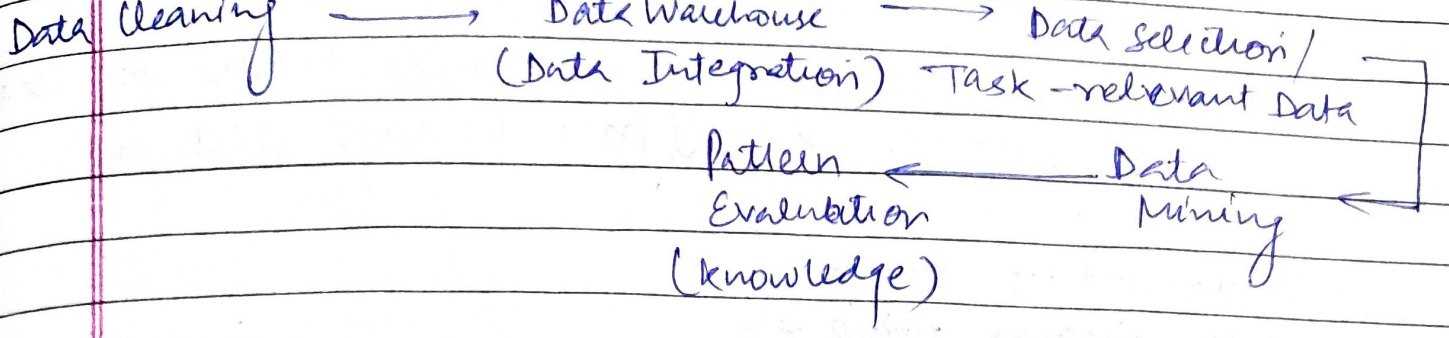
Pattern discovery ← Same

DS Process in BI

Decision ~~Data~~ Making
↕
Data Presentation
↑
Data Mining
↑
Data Exploration (EDA)
↑
Data Pre processing | Integration, Data Warehouses
↑
~~Data~~base Sources.

Page No.
Date

DS Process in KDD.



Data Science Methods.

1. Predictive / Supervised
2. Descriptive / Unsupervised

1. PREDICTIVE / SUPERVISED Method

→ Predict unknown values or labels.

GOAL → Classification, Regression, Deviation Detection
↳ Data + Class label.

Classification Example — Finite (2) Outcomes.
Cheat Detection System, Rain / No Rain, Cancer / No Cancer.
Fraud detection, Direct Marketing, Churn Detection (OLA, UBER)

Regression → Infinite Outcomes; Continuous Valued outcomes
Petrol Price ~~detection~~ detection system → 100, 100.5, 10.25 etc.
Stock, Gold Price, Temp, Time Series prediction of Stock market.

2. DESCRIPTIVE / UNSUPERVISED Method.

→ Predicts pattern or summaries.

GOAL → Clustering, Association

Clustering → Division of states on the basis of language.
Find a pattern in data and putting them under one attribute.

Page No.
Date

Distance $\propto \frac{1}{\text{Similarity}}$ = Euclidean Distance Based Clustering.

Association Rule Mining → Thanda Matlab Coca-Cola.
{Milk} → {Bread} Dahi ke aage Jeet hai.

Data and its Sources.

- Data stream: Audio data
- Time series data :- Temporal data, sequence data.
- Graph data :- www, Social Network.
  Spatial data and spatiotemporal data
  ↳ Time and location.
- Heterogenous databases and legacy databases.

Dataset → Collection of data objects and their attributes.
→ Contains pieces of raw information we want to analyze.

Real life Example
→ Bank Database → Collection of customer accounts, transaction and statement.

→ University Database :- A collection of data on students, professors and courses.

→ Sales Database → Medical Database

Objects → collection of attributes.
→ Record, point, case, sample, entity, or instance

Attribute → Property or characteristic of an object.
e.g. Refund, Marital status etc.
→ Characteristic, feature, dimension, variable or field.

TYPES OF ATTRIBUTES

- NOMIAL (Categorical Qualitative) = Relating to Name (=, ≠)
→ Symbol / Identifiers.
→ Examples : Name, ID NO., eye colours, zip code.

- ORDINAL (Categorical Qualitative) = Involving Order (=, ≠, <>)
→ Example :- Ranking (e.g. taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}.
→ Can be arranged in ascending or descending order.

- INTERVAL (Numeric ~~Qualitative~~ Quantitative) = Space in between
- Example :- Calendar dates, temp in Celsius.
- (+, −, =, ≠, <, >)

RATIO (Numeric Quantitative)
- Example : length, time, counts, electric current.
- (+, −, ÷, ×, =, ≠, <, >)

Discrete Attribute Vs ~~Non~~ Continuous Attribute
→ Has a finite or countably infinite set of values.
→ Often represented as integers.

Continuous Attribute
→ Has real numbers as values
→ Temp, Height, Weight.

Data Set
★ Record Data → That contains of collection of records, each of which consists of ~~mixed~~ a fixed set of attributes.
1) → Data Matrix → If data objects have the same fixed set of attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dim represents a distinct attributes.

→ such that is represented in m×n matrix.

2→ Document Data → eg. Google News
→ Each document becomes a 'term' vector.

3. Transaction Data – Each record (transaction) involves a set of items. (ARMS)

★ Graphical Data – Used when the relationships and connections between objects are as important as the obj themselves.
→ Objects are represented as nodes (vertices).
→ Relationships are represented as link (edges).

★ Ordered Data → Data where the sequence or ordering of values is a critical part of its meaning.
eg. Stock also known as Temporal Data (Time-Series)
→ Genomic Sequence Data
→ Measurement taken over time

DATA QUALITY
Refers to the overall quality of datasets(s) as a function of its ability to be easily processed and analyzed for ~~storage~~ usage

Dimensions to measure data quality
1. Accuracy :- Correct or wrong, accurate or not
2. Completeness :- not recorded, unavailable
3. Consistency :- some modified but some not, dangling
4. Timeliness :- Timely update.
5. Believability :- How trustable the data are correct?
6. Interpretability :- How easily the data can be understood

Page No.
Date

* Data has quality if it satisfy the requirements of intended users.

BAD DATA ⟶ BAD DATA MINING

Data Quality Issue

1) Incomplete Data (Reasons of missing values)
→ Information is not collected. e.g. decline to provide info.
→ Attributes not be applicable to all cases. e.g. children's income.
→ Incomplete vs Incorrect

2) Noisy Data :- Noise refers to modification of original values.
→ Random error or variance in a variable.
  • Distortion of a person's voice when talking on a poor signal range.
  • Error due to sensor malfunction.
→ Noise should be removed before detecting the outliers.

3) Inconsistent Data - discrepancies with in the dataset. A common problem when integrating data from multiple sources that use different standards. e.g. DD/MM/YYYY, MM/DD/YYYY, different rating etc.

GARBAGE IN, GARBAGE OUT

Detecting Data Quality Problems.
→ Detecting these issues requires a combination of automated and manual methods.
→ We can use meta data (data about data) to check for violations of domain, range or uniqueness rules.

Page No.
Date

Common Sources of Noise

1. Faulty Data Collections Instruments.
   A malfunction sensor or measuring device.
2. Data Entry Problems:
   Human error during manual data input.
3. Data Transmission Problems.
   Errors introduced as data is sent from one system to another.
4. Technology limitations
   Inherent imprecision in measurement technology.

Outliers

→ They are legitimate data objects whose characteristics are considerably different from most of the other data objects in the dataset.

→ They are not errors or corruptions; they are real but unusual data points.

Disguised Missing Data

→ Sometimes, missing data isn't represented by an empty field (like NULL or NA). Instead, it is "disguised" by a default or placeholder value.

→ This is particular dangerous because it can be mistaken for real data, leading to incorrect analysis.

→ A classic example is using Jan 1 as a default birthday for any record where the true birthday is unknown.

Challenges Arising from Duplicate Data

1) Entity Identification Problem
2) Tuple Duplication
3) Redundant Attributes

Page No.
Date

Detection of Duplicate Data.

1. Correlation Analysis :- Redundant attributes can often be detected by analyzing the correlation between them. If two attributes are highly correlated, one may be redundant.

2. Entity Identification Ambiguities :- Use rules and algo to identify records that likely refer to same entity.

3. Removal / Merging

Main Objective of Data Pre-processing

- To improve data quality
- To modify the data to better fit the requirements of a data science technique.

Data Pre-processing Task

→ Data Cleaning
→ Data Transformation and Aggregation
→ Data Reduction
→ Proximity Analysis.

1. Data Cleaning (Ignore,
   → Missing Values (Global Constant, local constant, Use of CLT
   → Noisy Data (Redundant Data, Unstructured Data, poorly formatted data, outliers, Misclassified or Mislabelled data)
   → Outlier Analysis
   → Inconsistent Data.

2. Data Transformation and Aggregation Task
   → Data Transformation
   → Data Normalisation Techniques
   → Data Aggregation
   → Data Integration

Precision - Measured by Std deviation
Bias - Diff b/w the mean of set of values and the known value of the quantity being measured.

Page No.
Date

3. Data Reduction Task
   → Reducing Data Volume
   → Data Sampling
   → Reducing Data Attributes
   → Attributes Subset Selection

4. Proximity Measures Task
   → Proximity Analysis for Nomial Attributes
   → Proximity Analysis for Binary Attributes.
   → Proximity Measures for Numerical Attributes.
   → Proximity Measures for Ordinal and Mixed Attributes.

★ DATA CLEANING

How to handle Noisy Data

1. Binning
   → First sort the data and partition into bins (by frequency)
   → then one can "smooth by bins mean", "smooth by bin median", "smooth by bin boundaries" etc.

2. Regression
   → Smooth by filtering the data into regression functions.

3. Clustering
   → Detect and remove outliers.
   → Combined computer and human inspection: detect suspicious values and check by human.

Methods to Identify Outliers.
1. Statistical Method :- Z-score.
2. Visualization Techniques
   1. Box Plot
   2. Scatter Plot

Page No.
Date

J. ~~DBSCAN~~ Clustering and Density-Based Methods.

DBSCAN :- Points far from any cluster or low-density region all outliers.

Approaches to Remove Outliers

1. Direct Removal
   → Exclude outlier records from the dataset ("trimming").
   → Suitable when outliers are due to errors or are not meaningful.

2. Transformation or Replacement (log Transformation)
   → Replace by the closest non-outlier value or with the mean/media of the group.
   → Preserves dataset size and structure.

3. Automated Detection
   → Use algo like Isolation Forest or One Class SVM for large or high-dimensional data.

Iterative Process :- Outlier detection and remove may require several rounds of analysis and validation.

Cautions :- Removing outliers can sometimes discard valuable info - always assess the impact on your analysis

Sources of Inconsistencies.

1. Sources - Human error during data entry or manipulation
   → System glitches or integration issues.
   → Merging datasets with diff standards
   → Lack of ~~go~~ data governance & standardization

Page No.
Date

Type of Inconsistencies.

Syntactic ~~Syntatic~~ :- Format or type mismatches (e.g. text in numeric fie

2. Semantic :- Contradictory meanings (e.g. "Active" vs "Yes" for status)

3. Structural :- Diff schema or data organisation across sources. Such as having different columns names for the same data in different tables.

Challenges Posed by Inconsistent Data.

1. Data Compromise (Quality)
   - Reduces accuracy and trustworthiness of insights

2. Biased Analysis
   Leads to incorrect influences and skewed results

3. Time - Consuming cleaning
   Requires significant effort to identify and resolve issues.

4. Operational and Compliance Risks
   Can result in inefficiencies, regulatory issues and reputational damages.

Identify Inconsistencies.

1. Data Profiling :- Analyze data distribution formats and pattern to spot anomilies.

2. Automated Tools :- Use data quality software for pattern recognition, rule-based checks, and cross-fie validation.

3. Visualisation :- Box plot, histograms and scatter plots Can reveal outliers and format issues.

Date

4. Manual Review : Domain expert can spot subtle or context-specific inconsistencies.

Resolving Inconsistencies

1. Data cleaning :- Correct or remove inconsistent entires (e.g. fix typos, standardize formats, remove duplicates).

2. Standardistion :- Convert data to common formats and units (e.g. ISO 8601 for dates, unified abbreviations).

3. Data ~~Validati~~ Validation :- ~~Apply~~ Rules for range, completeness, and logical consistency.

4. Integration Tools :- Use ETL (Extract, transform, load) and data integration solutions ~~for~~ to harmonize data from multiple sources.

Outliers

1. Errors :- They are illegitimate data points that are fundamentally incorrect.
   1. Data Entry Error :- A person's age is entered 150 instead of 50.
   2. Measurement Errors :- A faculty sensor temporarily reports a temperature of -100° C.
   3. Data Processing Errors :- An error in an ETL scripts corrupts a sets of records. The outliers are noise and should be corrected or removed.

2. Novelty / Interesting Events :- These are legitimate but rare data points that are correct.
   1. Fraudulent Transaction :- A credit card transaction that is genuinely fraudulent will have very different

characteristics from normal transaction.

2. System failure :- A spike in server errors indicating a real system outrage.
3. The "Bill Gates" :- In a dataset of regular people's income, Bill Gate's Income is a valid but extreme data point.

Why does outliers matter? The Malign Influence.

1. Influence on Mean and Std Deviation
2. Linear Models (e.g. Linear Regression) :- These models work by minimizing the sum of squared errors. An outlier creates a very large error, and the model will twist itself significantly to reduce this singe error, often at the expense of fitting the rest of the data well.
3. Clustering Algorithm

✱ DATA TRANSFORMATION

It is a process of converting raw data into a format that is more suitable and useful for analysis and modeling data science problem.

Purpose

It enhances data consistency, relevance, and quality by organizing, standardizing and simplifying data from various sources.

→ Improve Data Quality
→ Reduces Complexity
→ Enable Standardization
→ Enhances Analytical Accuracy.

Page No.
Date

Advantages and Benefits
→ Better Data Quality
→ Reduced Data Volume
→ Enhance Efficiency
→ Facilitates Integration

Challenges and Limitation
→ Information loss
→ Resource Intensive (can be costly and time consuming)
→ Complexity (Poor choice can introduce bias or overfitting)
→ Interpretability (High transformed data might be hard to interpret)

Data Normalization
A function that maps the entire set of values of a given attribute to a new set of replacement values s.t each old value can be identified with one of the new values.

Normalization
- Min-Max
- Z-score
- Normalization by decimal scaling

Min-Max
- Performs a linear transformation on the original data:

$$v' = \frac{v - min_A}{max_A - min_A} (new\ max_A - new\ min_A) + new\_min_A$$

- Preserves the relationships among the original data values.

Z-Score
- aka z-mean normalization.

$$v' = \frac{v - mean_A}{Standard\ A}$$

Decimal Scaling Normalization

- By moving the decimal points of values of attribute A.
- The number of decimal points moved depends on the maximum absolute value of A.
- A value, v_i of A is normalised to v_i' by computing

$$v_i' = \frac{v_i}{10^j} \qquad j = \text{No. of decimals.}$$

Data Aggregation

- It is a process of collecting and combine raw data from multiple sources to create a summarized, unified dataset for analysis and reporting.
- Aggregation transforms large, complex datasets into concise, manageable forms, making it ~~easy~~ easier to identify patterns, trends, and key metrics.
- This step is fundamental in data preprocessing, supporting business intelligence, analytics, and decision-making by providing a high-level overview of the data.

Purpose

- Data Reduction
- Change of scales (Cities aggregated into regions, states etc)
- More "stable" data
- Simplifies Analysis
- Improves Data Quality.
- Enhance Efficiency
- Support Better Decision Making

P. No.
Date

☆ DATA REDUCTION

It is a process of minimizing the amount of data while retaining it's essential information and integrity.

Purpose
→ Simplifies data
→ Improves computational efficiency
→ Enhances the quality of data analysis.

IMPORTANCE

- Storage Optimization :- Reduces the vol of data stored, saved space and costs.
- Processing Speed :- Accelerates data mining and analysis by reducing computational load.
- Enhanced Analysis :- Makes data more manageable and easier to interpret, supporting better decision-making.
- Noise Reduction :- Helps eliminate redundant or irrelevant information, improving data quality.

TECHNIQUES

- Tuple Reduction :- Reduces the sample by using sampling & clustering.
- Dimensionality Reduction :- Reduces the no. of attributes like PCA (Principle Component Analysis) or feature selection.
- Data cube Aggression :- (e.g. daily to monthly sales)
- Data Compression :- Uses encoding methods (e.g. wavelet transform, run-length encoding) to reduce data size.

Date

KEY CONCEPTS

1. Dimensionality → ↑ dimens = ↑ complexity (Curse of Dimensionality)
2. Data Sparsity :- High-dimensional data contains many empty or irrelevant values, complicating analysis.
3. Information Preservation :- The goal is to reduce data size while retaining the most important information.

CHALLENGES

1. Risk of Information loss
2. Complexity :- Some techniques require advanced algorithms and domain expertise.

Reducing Data Volume

Shrinking the amount of data stored or processed, while retaining its essential information and integrity

PURPOSE

- Improves computational efficiency.
- Reduces storage requirements.
- Enhances data manageability and analysis speed.

- Lossless :- If the original data can be reconstructed from the compressed data without any loss into loss
  e.g. ZIP, PNG, Huffman coding, Run-length encoding.
- Lossy :- If we can only reconstruct an approximation of the original data e.g: JPEG, MP3, MPEG.

CONSIDERATION

1. Lossy vs lossless :- Choose based on importance of data integrity
2. Compression :- Balance between file size reduction and quality
3. Processing Overhead :- Compression and decompression requires computational resources.
4. File Suitability :- Not all files benefit from compression (e.g. already compressed or small files).

FEATURE SELECTION / ATTRIBUTE-SUBSELECTION

Process of identifying and retaining only the most relevant & informative attributes.

- Forward Selection (stepwise) (accuracy)
- Forward Selection (Stagewise) (error)
- Backward Selection
- Decision Tree induction (Flowcharts)

PROXIMITY ANALYSIS FOR NOMIAL ATTRIBUTES

Defination :- Measures how similar or dissimilar data objects are, based on their attributes.

TYPES :-
- Similarity :- High value means objects are alike.
- Dissimilarity (Distance) :- High value means object are different.

APPLICATIONS
- Clustering
- Classification
- Outlier detection
- Recommendation Systems

* Distance $\propto \frac{1}{\text{Similarity}}$

Similarity Measure :- Numerical measure of how ~~alike~~ alike two data objects are.

→ Is higher when objects are like.
→ Often falls in the range [0, 1].
0 ↙ Not same
1 ↘ Same

Dissimilarity Measure :- Numerical measure of how different two data objects are.
→ Lower when objects are more alike.
→ Minimum dissimilarity is often 0, upper limit varies
→ The term distance is used as a synonym for dissimilarity.

* **Proximity** refers to similarity or dissimilarity.

DISSIMILARITY MEASURE :-

Simple Matching

$$d(A, B) = \frac{p - m}{p}$$

p :- Number of attributes
m :- Number of matches.

Page No.
Date

LIMITATIONS AND EDGE CASES

1. **Missing values** :- Requires special handling.
2. **High-Cardinality Attributes** :- Many categories can make similarity / dissimilarity values less meaningful.
3. **Data Quality** :- Errors in data entry can affects results.
4. **Domain knowledge** :- Understanding context is crucial for correct interpretation.

BINARY ATTRIBUTES

TYPES :-
- **Symmetric Binary** :- Both states (0 and 1) are equally important and interchangeable.
  e.g. Gender (Male / Female).
- **Assymmetric Binary** :- One state (usually 1) is more significant or rare than the other (0).
  e.g. Disease presence (1: Present, 0: Absent)

CONSIDERATIONS

* Choice of similarity measure depends on attribute type
* Asymmetric attributes requires special handling in similarity calculations.

f_{00} = the number of attributes where x is 0 and y is 0.
f_{01} = " " " " " " " " 0 and " " 1.

f_{10} = " " " " " " " " 1 and " " 0.
f_{11} = " " " " " " " " 1 and " " 1.

Page No.
Date

Proximity Measures for Binary Attributes.

Simple Matching Coefficient (SMC) :- For symmetric attri.

Jaccard Coefficient :- For assymmetric binary attri.

Interpretation
- SMC :- Both matches (1-1, 0-0) are considered.
- Jaccard :- Only positive matches (1-1) are considered ignoring negative matches (0-0).

Use
SMC :- One commonly used similarity coefficient.

$$= \frac{\text{Number of matching attributes values}}{\text{No. of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{00} + f_{11}}$$

- This measures counts both presences and absence equally.

Jaccard Similarity :- Frequently used to handle objects consisting of assymmetric binary attributes.

$$j = \frac{\text{No. of matching presence}}{\text{No. of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{11} + f_{10}}$$

- This measures count both presences and absences equally.

The measures count both presences and absences equally.

Page No.
Date

Consequently, the SMC could be used to find students who had answered questions similarly on test that considered only of true / false questions.

Proximity Analysis for Numerical Attributes

Numerical attributes are continuous or discrete numbers (e.g., age, temperature, income).

Common Measures

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance (Generalisation)

Euclidean Distance

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

* Standardization is necessary, if scales differ.

Minkowski Distance

It is a generalization of Euclidean Distance, and is given by

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

r = parameter

- The following are the three most common examples of Minkowski distances.
- $r=1$, City Block (Manhattan, Taxicab, L_1 norm) distance.
- A common example of this for binary vectors is the Hamming distance, which is just the number of bits that are different between two binary vectors.
- $r=2$, Euclidean distance (L_2 norm)
- $r=\infty$, Supremum (L_{max} norm, L_∞ norm) distance
  → This is the maximum difference between any components of the vectors!
- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

Proximity Analysis for Ordinal Attributes

1. Mapping to Ranks
   Assign integer ranks to categories (e.g., Poor = 1, Fair = 2, Good = 3, Excellent = 4)

2. Normalization
   Convert ranks to a scale using $Z_i = \frac{r_i - 1}{M - 1}$
   - r_i = Rank of category i
   - M :- Maximum rank

3. Proximity Calculation

Page No.
Date

Using normalised ranks with numerical distance measures (e.g. Manhattan, Euclidean)

Mixed - Type Attributes

This contains a combination of nomial, ordinal, binary and numerical attributes.

Challenges

Need to handle each attribute type appropriately in proximity analysis.

Solution

Use weighted combination of normalized dissimilarities for each attributes type.

Application

Customer ~~Apps~~ Profiling, medical diagonsis, multi-source data integration.

Weighted Combination

$$d(A, B) = \frac{\sum_{k=1}^{p} w_k \, \delta_k(A_k, B_k)}{\sum_{k=1}^{p} w_k}$$

w_k = weight for attribute k

δ_k = dissimilarity for attribute k.

Visualization of Mixed - Type Data

- Heatmap or parallel coordinates plot showing values for each attributes type (nomial, ordinal, numerical).

Page No.
Date

Interpretation

- Helps identify patterns and relationships across mixed attribute types.

Data Distribution

It is a function that shows all the possible values for a variable and how often each value occurs. It can be called as data summary.

Why understanding the Data Distribution Crucial?

1. Foundation for statistical Inference :- Most statistical tests and assumptions (like t-test or ANOVA) are based on the underlying distribution of the data. Using the wrong test might leads to invalid conclusions.

2. Informs model Selection :- The nature of the target variable's distribution can guide to the right machine learning model. A normally distributed target might work well with linear Regression, while a count-based, skewed distribution might require a Binomial model.

3. Guides Data Pre-processing :- Identifying skewed or ~~unusual~~ unusual distributions is the first step toward effective feature engineering. Techniques like log transformations, normalization, or standarization are applied specifically to handle certain distribution characteristics and improve model performance.

4. Helps in Anomly Detection :- An outlier is, by definition, a point that lies far from the central mass of the ~~education~~ distribution. You cannot

identify what is "unusual" without first understanding what is "usual".

Why use of Statistics to Describe data?
- Better understand the data :- (Central Tendency, σ)
- Identify potential issues (Outliers, skewed data)
- Guide further analysis

CENTRAL TENDENCY
- Tells where the "center" or "typical" value of the data lies.
- Mean :- The arithmetic average. Sensitive to Outliers
- Median :- The middle value when the data is sorted. Robust to outliers.
- Mode :- The most frequently occuring value.
- Range :- The difference b/w the max and min value
- Variance / Standard Deviation :- Measures the average distance of data from mean.

Measuring the spread :- Data Dispersion
It describes how spread out or varied the data is.

- Quartiles and Interquartile Range (IQR) :-
  - Quartile :- divide the data into four equal parts. Q1 (25th percentile), Q2 (50th percentile / Median), Q3 (75th percentile).
  - IQR = Q3 − Q1. It represents the spread of middle 50% of the data and is robust to outliers.

Page No.
Date

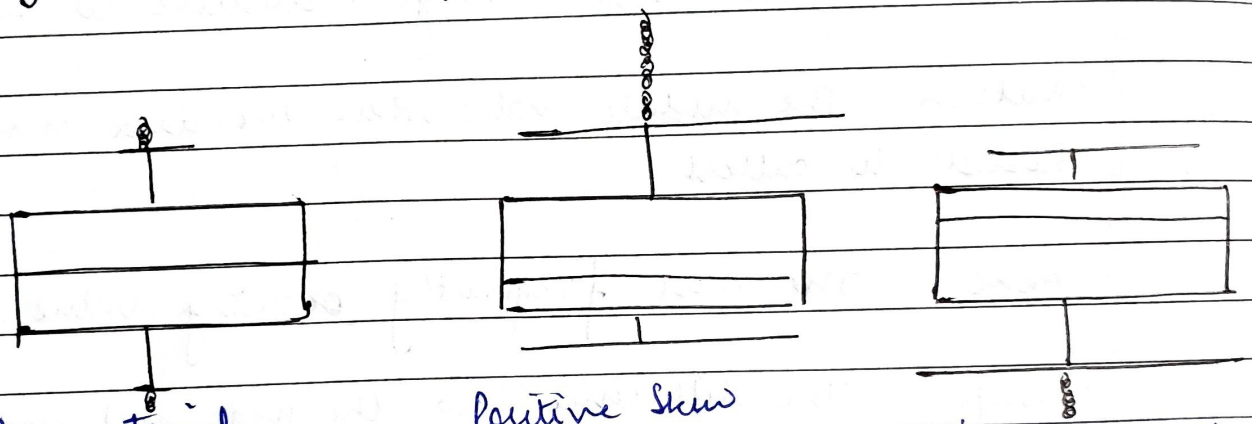
Central Tendency

- Symmetry :- Is the distribution balanced on both sides of the center?
- Skewness :- Is the data lopsided to one side?

Symmetric (No Skew) Mean ≈ Median ≈ Mode

Left Skew (Negatively Skewed) Mean < Median < Mode

Right Skew (Positively Skewed) Mean > Median > Mode



Symmetrical | Positive Skew Distribution | Negative Skew

The "Black Box Problem"

Modern machine learning has given us incredibly powerful and accurate models like Gradient Boosting, Deep Neural Network and complex Ensembles.

These model can learn intricate, non-linear patterns in data that simpler models ~~can~~ like linear Regression cannot.

This comes with a cost :- Interpretability.

The inner working of these models are complex that it's nearly impossible for a human to understand exactly how they arrived at a specific decision. They are often referred to as "black box" models. We can see the inputs and the outputs, but the reasoning inside is opaque.

SAMPLE EXPLAINABILITY (Local Explanations)