Principles of Data Science

Vasudevan T V

Course Contents

- Module 1 - Introduction to Data Science
- Module 2 - Data Mining and Preprocessing
- Module 3 - Classification models
- Module 4 - Introduction to Association Mining , Clustering and Evaluation metrics

Module 1
Introduction to Data Science

- Data science is a collection of techniques used to extract value from data
- It has become an essential tool for any organisation that collects, stores, and processes data as part of its operations
- Data science techniques find useful patterns, connections, and relationships within data
- It can be used for decision making
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining
- However, each of these terms have a different meaning depending on the context

Importance of Data Science

- Example - Retail Clothing Shop
- Consider a retail clothing shop that wants to increase its sales.
- It can use the following types of analysis using data science techniques
- sales analysis
- By analysing past sales data, the store can identify which items were bestsellers during certain seasons
- For instance, if winter jackets sold really well last year, the store can stock more of these for the upcoming winter
- Customer Behaviour Analysis
- Data science can classify customers based on their purchasing behaviour
- For example, if data shows that younger people tend to buy casual wear more than formal wear, then the shop can start marketing campaigns suitable for the corresponding segment

Importance of Data Science

- Inventory Management
- Using predictive analytics, the store can forecast demand for different products
- If data indicates that sales of workout gear will increase in the spring, the store can ensure it has enough stock to meet that demand
- Targeted Marketing
- Data science enables personalized marketing strategies
- If a customer frequently buys shoes, the store can send them advertisements for new footwear, which is likely to increase the chances of a sale

History of Data Science

Decade	Key Development	Impact on Data Science
1950s-1960s	Emergence of Computers	Automated data processing
1970s	Development of statistical software	Broader access to data analysis tools
1980s	Introduction of Personal Computers	Spread of data skills among the populace
1990s	Introduction of World Wide Web	Massive increase in data availability
2000s	Machine Learning and Predictive Analytics	Enhanced decision-making capabilities
2010s-2020s	AI integration	Advanced predictive models

Types of Data

1. Structured data
- It is a kind of data that has a well defined structure
- It is stored in tabular form, organised into rows and columns
- Examples - Relational Databases, Spreadsheets, CSV files
2. Semi-structured data
- It is a kind of data that is less structured
- It uses tags, key-value pairs, headers etc. to separate data
- Examples - HTML, XML, JSON, Emails
3. Unstructured data
- It is a kind of data that has no fixed format
- Examples - Text documents, image files, audio files, video files

Types of Data

1. Categorical data

- It is a kind of data that can be divided into different categories or groups
- It is of two types - ordinal and nominal

(a) Ordinal data

- It is a kind of data that has a meaningful order
- Examples
- Feedback Ratings - {Poor, Satisfactory, Fair, Good, Very Good, Excellent, Outstanding}
- Education Levels - {Higher Secondary School, Bachelor's Degree, Master's Degree}
- Class Grades - {S, A+, A, B+, B, C+, C, D, P, F}

Types of Data

1. Categorical data
   - It is a kind of data that can be divided into different categories or groups
   - It is of two types - ordinal and nominal

(b) Nominal data
   - It is a kind of data that does not have any inherent order
   - Examples
   - Gender - {Male, Female, Transgender}
   - Fruits - {Mango, Apple, Banana}
   - Eye Colour - {Blue, Black, Brown}

Types of Data

2. Non-Categorical data (Numeric Data)
   - It contains numerical values that have a meaningful order
   - It is of two types - continuous and discrete

(a) Continuous data
   - It can take any value within a given range
   - Examples
   - Temperature
   - Height
   - Weight

Types of Data

2. Non-Categorical data (Numeric Data)

- It contains numerical values that have a meaningful order
- It is of two types - continuous and discrete

(a) Discrete data

- It consists of distinct values, which are often counted, not measured
- Examples
- Number of students in a class
- Number of cars in a parking lot
- Number of players in a team

Real World Applications of Data Science

1. Healthcare
   - To forecast patient admissions and optimise resource allocation
2. Finance
   - Analyse transaction patterns to identify fraudulent activities
3. Education
   - To predict the performance of a student
4. Retail Shops
   - Classify customers based on their purchasing behaviour
5. Transportation
   - To predict failures of vehicles before they occur
6. Sports
   - Analyse player statistics to optimise performance

Data Science Process

- The method of discovering useful relationships and patterns in data is called the data science process
- Steps

1. Prior Knowledge
2. Data Preparation
3. Modelling
4. Application

Data Science Process

1. Prior Knowledge

- Here we define what problem is being solved
- We find out what data is needed for solving the problem
- Example - Consumer Loan Business
- Problem - If the interest rates and credit scores of past borrowers are known, can we predict the interest rate of a new borrower?
- Data - A sample data set of 10 data points with 3 attributes : borrower id, credit score, and interest rate

Data Science Process

1. Prior knowledge

Table 2.1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

- Credit Score is a measure of the ability of a borrower to repay the loan
- Larger the credit score, greater the ability to repay the loan

Data Science Process

1. Prior knowledge

- A data set is a collection of data with a well defined structure
- Here, table is the data set
- A data point is a single instance of the data set
- Here, each record in the table is a data point
- An attribute is a single property of the data set
- Here, each column in the table is an attribute
- An identifier is a special attribute used for locating data points in a data set
- Here, Borrower Id is the identifier

Data Science Process

1. Prior knowledge

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

- A label is the special attribute to be predicted based on all the input attributes
- Here interest rate of the new borrower is to be predicted

Data Science Process

2. Data Preparation

- In this stage we prepare the whole data set needed for the data science task
- Steps

2.1 Data Exploration

- This involves in depth analysis of data to gain better understanding about it
- For this, statistical analysis and visualisation tools are used

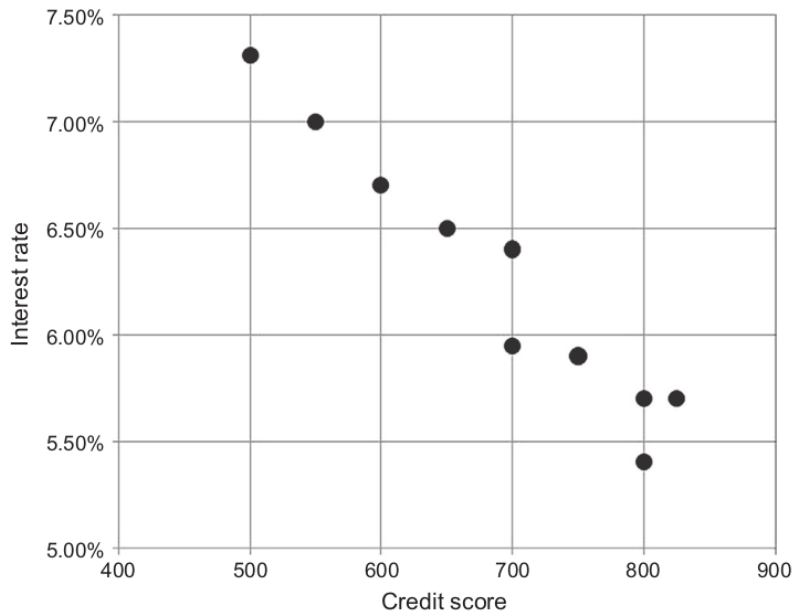


FIGURE 2.3
Scatterplot for interest rate dataset.

Data Science Process

2.2 Ensure Data Quality

- Data Cleansing techniques are used for ensuring data quality

1 Elimination of Duplicate Records

2 Dealing with Missing Values

- Example - Missing Credit Score
- It can be replaced with a credit score derived from the data set(mean)
- Alternatively, we can eliminate records with missing values

3 Data Type Conversion

- Depending on the requirement, we convert data from one type to another
- This depends on the data science algorithm we are using
- We can convert credit score to categorical values such as poor = 400, good = 600, excellent = 800

Data Science Process

2.2 Ensure Data Quality

- Data Cleansing techniques are used for ensuring data quality

4 Transformation of Attribute Ranges

- Different attributes have different ranges
- For example, range of income is larger compared to range of credit score
- For some data science algorithms, these ranges are normalised to a uniform scale from 0 to 1

5 Handling Outliers

- Outliers are anomalies that differ from majority of observations in a data set
- Example - Human height as 1.73cm instead of 1.73m in a record
- We need to correct these anomalies

Data Science Process

2.3 Feature Selection

- All the attributes in the data set may not be needed for solving the problem
- Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection
- This leads to a simplified model

2.4 Data Sampling

- It involves selecting a subset of the original data set for analysis
- It reduces the amount of data to be processed
- It can speed up the process of analysis

Data Science Process

3 Modelling

- A model is the abstract representation of the data and the relationships in a given data set
- There are 2 kinds of data sets associated with a model
- The data set used to create the model is called a training data set
- The data set used to validate the model is called a test data set
- The entire data set is split into training data set and test data set
- A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test data set

Data Science Process

Table 2.3 Training Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

Data Science Process

3 Modelling

- Now we will evaluate the model using the test data set
- We will be using simple linear regression technique for predicting interest rates of test data set

Table 2.5 Evaluation of Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)	Model Predicted (*Y*) (%)	Model Error (%)
04	700	6.40	6.11	− 0.29
07	750	5.90	5.81	− 0.09
10	825	5.70	5.37	− 0.33

4 Application

- In this stage we present our findings to the world
- We can build an application that automatically updates reports, spreadsheets and presentation slides

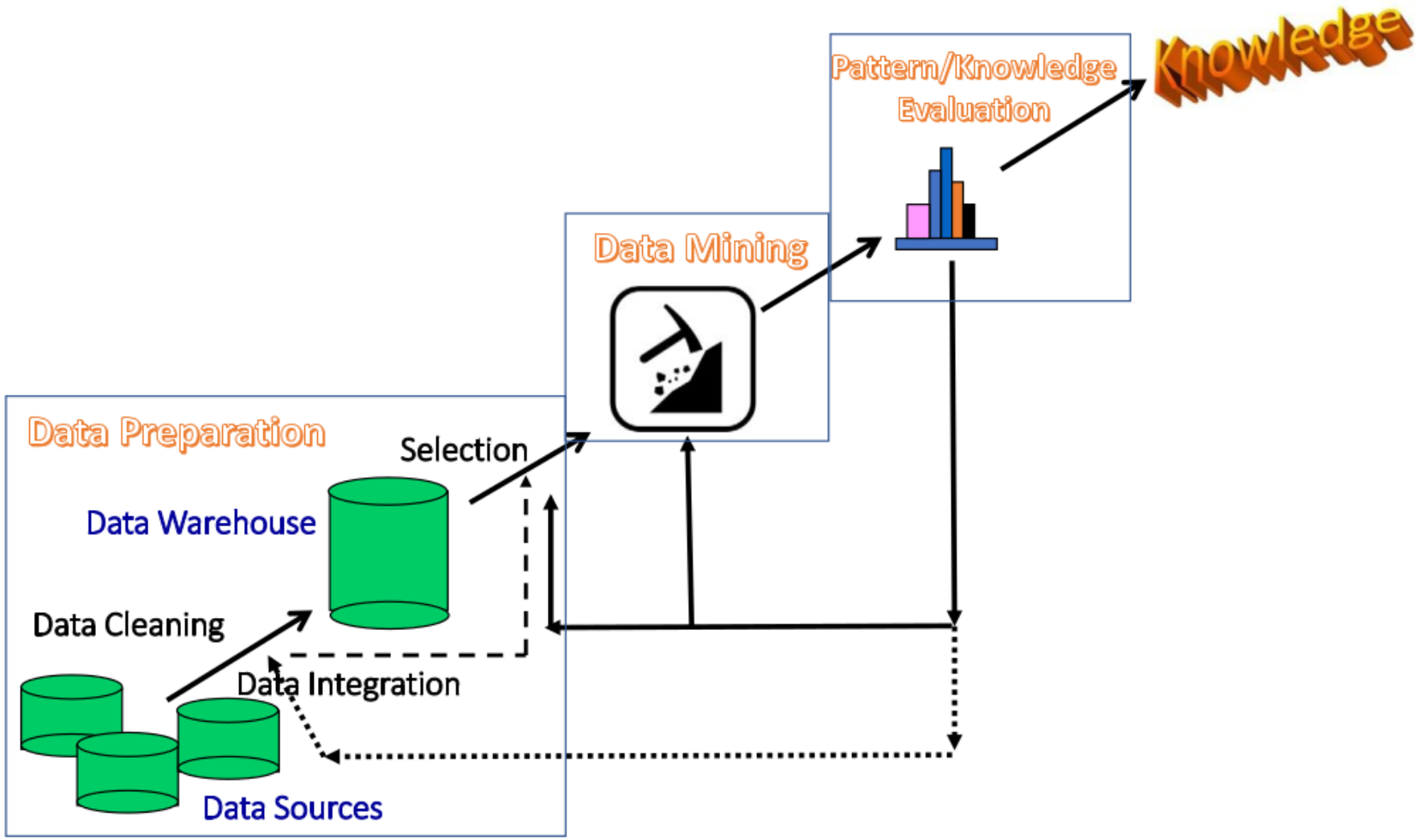
Differences between Artificial Intelligence,Machine Learning and Deep Learning

- Artificial Intelligence (AI) is the broad field focused on creating systems that perform tasks considered to require human intelligence
- Machine Learning (ML), a subset of AI, specifically concentrates on developing algorithms that allow machines to improve through experience (data)
- Deep learning (DL) is a machine learning model based on artificial neural networks with multiple hidden layers, which are inspired by the functioning of human brain

Module 2
Data Mining

- It is the process of discovering interesting patterns and hidden knowledge from large data sets that can't be retrieved using normal queries
- It is an essential step in Knowledge Discovery from Data (KDD)
- The term data mining evolved in 1990s
- Query - Example
- Is there any relationship between the performance of a student in subject A and subject B ?
- Which items are often purchased before a holiday ?
- What is the correlation between toss outcomes and match results in cricket ?
- For retrieving them, special data mining algorithms are used

Essential Steps in Knowledge Discovery



Essential Steps in Knowledge Discovery

1. Data Preparation

- In this stage we prepare the whole data set needed for the data mining task
- The following are the steps in this stage

a Data Cleaning

- It involves removing noise and inconsistent data

b Data Integration

- In this phase multiple data sources are combined

c Data Transformation

- Here data is transformed into forms appropriate for mining

d Data Selection

- Here data relevant to the analysis task are retrieved from the database

Essential Steps in Knowledge Discovery

2. Data Mining

- In this stage we use intelligent methods for retrieving interesting patterns and hidden knowledge

3. Pattern / Knowledge Evaluation

- Here we evaluate the patterns and knowledge based on interestingness measures
- In other words we select truly interesting patterns and knowledge

4. Knowledge Presentation

- Here visualisation tools and other techniques are used to present knowledge to users

Mining Various Kinds of Knowledge

- Here we will discuss about different data mining tasks

1. Multidimensional Data Summarisation

- In this process, large data sets are summarised into a comprehensible format while preserving essential information
- Here multiple dimensions or attributes of data such as time, location and type are considered
- During summarisation, data is often aggregated through various functions such as sum, average and count
- For instance, sales data might be summarised by month, region, and product category
- The output of such multidimensional summarisation can be presented in various forms, such as pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables

Mining Various Kinds of Knowledge

2. Mining frequent patterns, associations, and correlations

- Frequent patterns are patterns that occur frequently in data
- Example 1 - milk and bread, which are frequently bought together in grocery stores by many customers
- Example 2 - The pattern that customers, tend to purchase first a laptop, followed by a computer bag, and then other accessories
- Example 3 - A frequent pattern in web browsing could be users visiting a homepage, then the products page, followed by the checkout page.

Mining Various Kinds of Knowledge

- Association refers to the process of discovering interesting relationships among a set of items in large databases.
- Associations mined are represented in the form of association rules
- Example 1
- buys(X, "computer") $\Rightarrow$ buys(X, "webcam") [support = 1%, confidence = 50%]
- X is a variable representing a customer
- A 1% support means that 1% of all the transactions under analysis show that computer and webcam are purchased together
- A confidence of 50% means that if a customer buys a computer, there is a 50% chance that she will buy webcam as well

Mining Various Kinds of Knowledge

- Support and confidence are measures that indicate the interestingness of an association
- Support indicates the percentage of transactions in the transaction database that satisfy the given rule
- support($X \Rightarrow Y$) = P ($X \cup Y$)
- Confidence assesses the degree of certainty of the detected association
- confidence($X \Rightarrow Y$) = P ($Y \mid X$)

Mining Various Kinds of Knowledge

- This association rule involves a single attribute or predicate (i.e., buys)
- Association rules that contain a single predicate are referred to as single-dimensional association rules
- Example 2
- age(X, "20..29") $\wedge$ income(X, "40K..49K") $\Rightarrow$ buys(X, "laptop") [support = 0.5%, confidence = 60%]
- The rule indicates that of all its customers under study, 0.5% are 20 to 29 years old with an income of 40,000 to 49,000 and have purchased a laptop
- There is a 60% probability that a customer in this age and income group will purchase a laptop
- This is an association rule involving more than one attribute or predicate (i.e., age, income, and buys)
- Association rules that contain multiple predicates are referred to as multi-dimensional association rules

Mining Various Kinds of Knowledge

- Correlation measures the statistical relationship between attributes
- It measures the dependence of one attribute on another
- Example - There is correlation between average temperature of a day and ice cream sales
- Correlation between two attributes is commonly measured by the Pearson correlation coefficient
- It ranges from -1 to 1, where negative values indicate negative correlation, positive values indicate positive correlation and 0 indicates no correlation
- -1 and 1 indicate perfect correlation

Mining Various Kinds of Knowledge

3. Classification and regression for predictive analysis

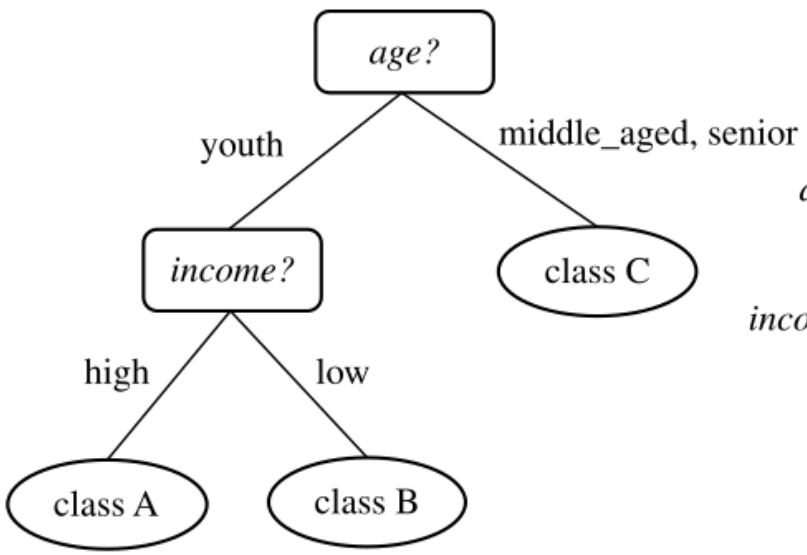
- Classification and Regression techniques predict a target variable based on input variables
- The output variable which is predicted is called a target variable
- In classification, the target variable is a category or class such as 'yes','no', 'red', 'blue' etc.
- Example - Predicting whether monsoon will be normal this year
- In regression, the target variable is a numeric value
- Example - Predicting the age of a person

Mining Various Kinds of Knowledge

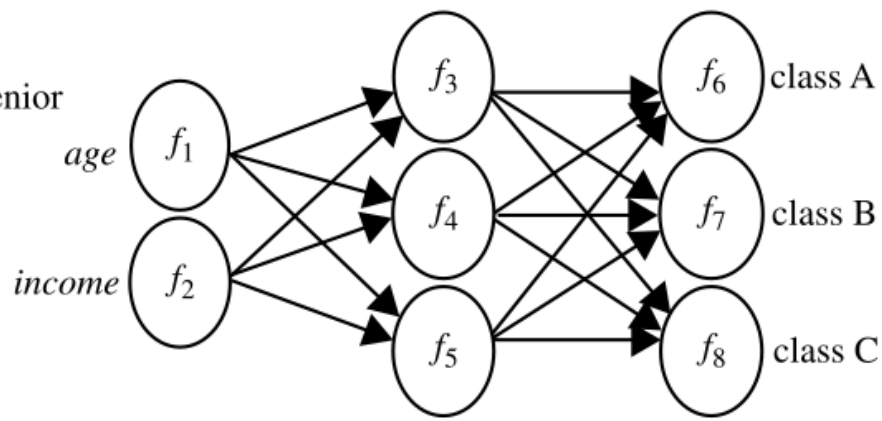
- A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network

age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*
age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*
age(X, "middle_aged") ⟶ *class(X, "C")*
age(X, "senior") ⟶ *class(X, "C")*

(a)



(b)



(c)

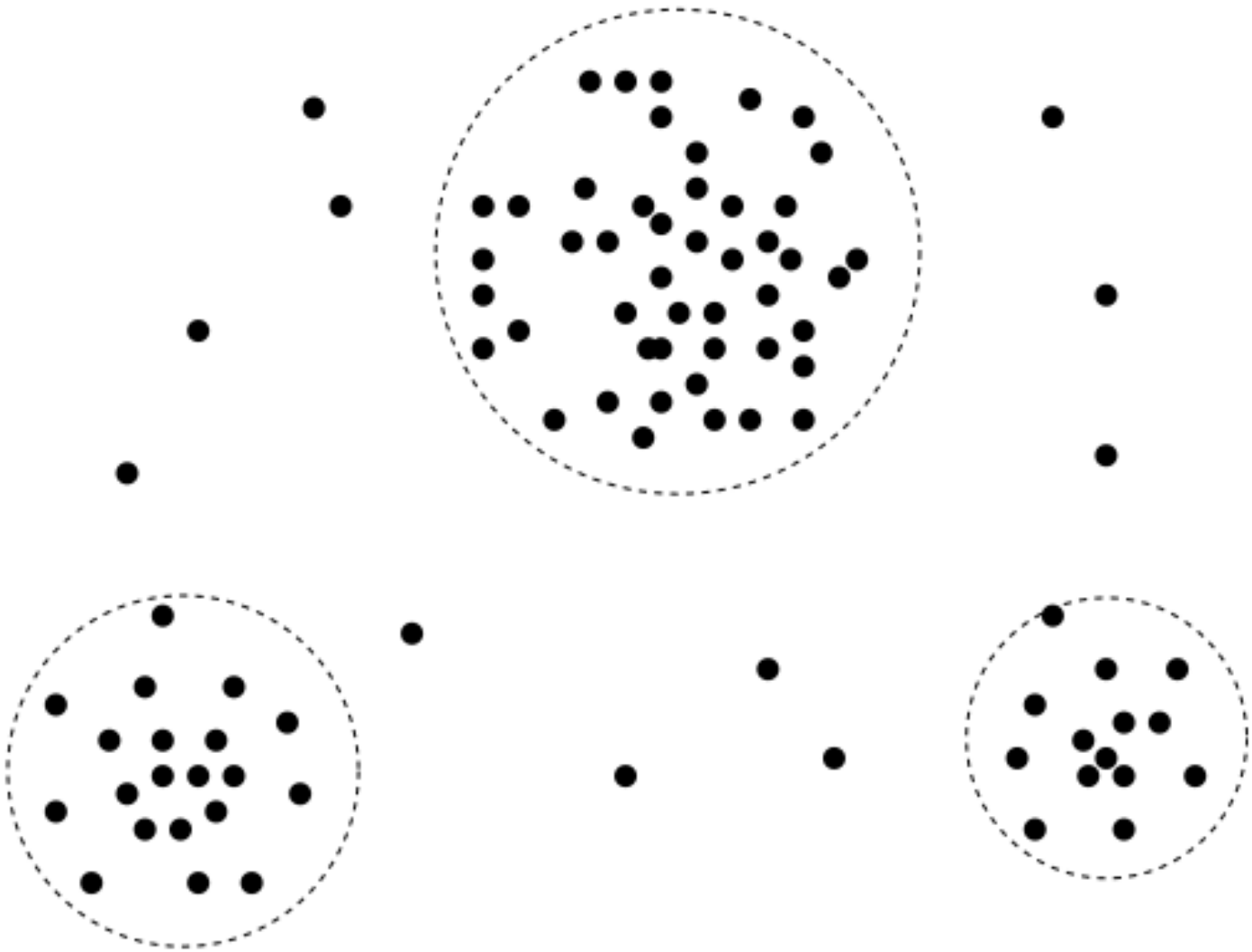
Mining Various Kinds of Knowledge

4. Cluster analysis

- Cluster analysis (Clustering) is the process of identifying natural groupings within a data set
- Example - Grouping books in a library based on topics
- The objects are clustered or grouped based on the principle of maximising the intraclass similarity and minimizing the interclass similarity
- It is used for knowledge discovery rather than prediction

Mining Various Kinds of Knowledge

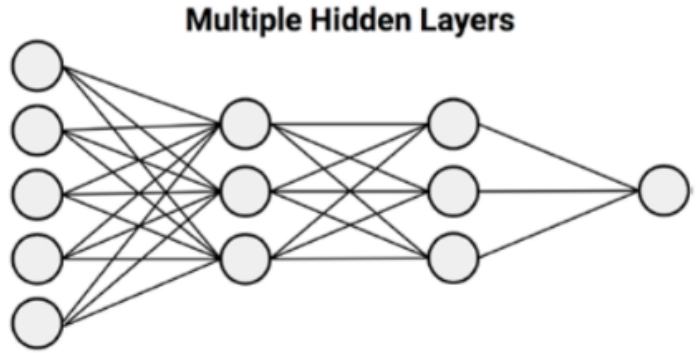
- A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters



Mining Various Kinds of Knowledge

5. Deep Learning

- Deep learning (DL) is a machine learning model based on artificial neural networks with multiple hidden layers, which are inspired by the functioning of human brain



- Deep learning has broad applications in computer vision, natural language processing, machine translation, social network analysis, and so on

Mining Various Kinds of Knowledge

6. Outlier Analysis

- A data set can contain data points which are significantly different in behaviour compared to the majority of them
- Such data points are called outliers or anomalies in the data set
- In some applications, the rare events can be more interesting than the frequently occurring ones
- The analysis of outlier data is called outlier analysis or anomaly mining
- Example 1 - A fraudulent credit card transaction may have higher spending for a customer compared to the normal limits
- Example 2 - Measurements of temperatures significantly above or below the average temperature at a particular location indicates a heat wave or a cold wave
- Example 3 - A social media post receiving an unusual number of likes can indicate a viral content or a bot activity

Mining Various Kinds of Knowledge

7. Are all mining results interesting ?

- A data mining result is interesting if the following criteria are satisfied
  - i It can be easily understood by humans
  - ii It is valid on a new or test data with a certain degree of certainty
  - iii It is potentially useful
  - iv It is novel
- Users can set a minimum threshold value for the interestingness measures support and confidence
- Example - Rules that do not satisfy a support and confidence threshold of, say, 50% can be considered uninteresting.

Data Mining and Applications

1. Education

- Student Performance Analysis - Evaluating student performance to improve teaching methods and curriculum design
- Dropout Prediction - Analysing factors that lead to student dropouts to implement retention strategies

2. Finance

- Fraud Detection - Identifying unusual patterns in transaction data that may indicate fraudulent activity
- Stock Market Analysis - Analysing trading patterns to forecast market trends

Data Mining and Applications

3. Healthcare

- Disease Prediction - Identifying patterns that may indicate the onset of diseases, improving early intervention
- Patient Care Improvement - Analysing patient data for better treatment plans and resource allocation

4. Business

- Customer Segmentation - Analysing customer data to identify distinct groups for targeted marketing
- Sales Forecasting - Predicting future sales trends based on historical data

Data Mining and Applications

5. Social Media

- Sentiment Analysis - Assessing public sentiment based on social media interactions and comments
- Content Recommendations - Suggesting relevant articles, videos, or posts to users based on their previous interactions

6. Sports

- Player Performance Analysis - Analysing player statistics to evaluate and improve individual performance
- Injury Prediction - Using historical data to identify patterns that may lead to injuries, helping in injury prevention

Data, Measurements and Data Preprocessing

- In this section, we get familiar with data
- We discuss the various attribute types associated with data
- After that we will assess the similarity and dissimilarity between data
- The measures for assessing them are called proximity measures
- Finally we look at data preprocessing, where we transform the data into a format which is suitable for analysis
- It improves the quality of data and the accuracy of mining results

Data Types

- Data sets are made up of data objects (data points)
- A data object (data point) is a single instance of the data set
- If the data set is a table, then each record in the table is a data object
- An attribute is a single property of the data set
- Each column in the table is an attribute

Data Types

- Attributes can be classified into the following types, based on the data stored in them

1. Nominal Attributes
2. Binary Attributes
3. Ordinal Attributes
4. Numeric Attributes
5. Continuous Attributes
6. Discrete Attributes

Data Types

1. Nominal Attributes

- They contain categorical data that does not have any inherent order
- Categorical data is a kind of data that can be divided into different categories or groups
- Gender - {Male, Female, Transgender}
- Fruits - {Mango, Apple, Banana}
- Eye Colour - {Blue, Black, Brown}

Data Types

2. Binary Attributes

- They are nominal attributes with only two categorical values 0(false) and 1(true)
- Survived - {0, 1}
- Passed - {True, False}

Data Types

3. Ordinal Attributes

- They contain categorical data hat has a meaningful order
- Feedback Ratings - {Poor, Satisfactory, Fair, Good, Very Good, Excellent, Outstanding}
- Education Levels - {Higher Secondary School, Bachelor's Degree, Master's Degree}
- Class Grades - {S, A+, A, B+, B, C+, C, D, P, F}

Data Types

4. Numeric Attributes

- They contain numerical values that have a meaningful order
- They are of two types - continuous and discrete

5. Continuous Attributes

- They can take any value within a given range
- Examples
- Temperature
- Height
- Weight

Data Types

6. Discrete Attributes

- It consists of distinct values, which are often counted, not measured
- Examples
- Number of students in a class
- Number of cars in a parking lot
- Number of players in a team

Similarity and Distance Measures

- In data mining applications we may need to find out how alike (similar) or unalike (dissimilar) two data objects are
- Similarity measures are used for finding out how similar different data objects are
- A similarity measure for two objects, i and j , will typically return a value 0, if the objects are completely unalike
- It will return a value 1, if the objects are identical
- Distance measures are used for finding out how dissimilar different data objects are
- A dissimilarity measure (distance measure) returns a value of 0 if the objects are the same
- The higher the dissimilarity measure, more unalike the objects are

Similarity and Distance Measures

- Euclidean Distance
- The most popular distance measure is Euclidean Distance
- Formula

$$\text{dist}(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

- Here p and q are items to be compared, having n features
- p_1 refers to the value of first feature of p
- q_1 refers to the value of first feature of q
- Let $p_1 = (1, 2)$ and $p_2 = (3, 5)$ represent two objects
- The Euclidean distance between the two is $\sqrt{(3-1)^2 + (5-2)^2} = \sqrt{4+9} = 3.61$

Similarity and Distance Measures

- Cosine Similarity
- It measures the similarity between two vectors
- It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction
- A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match
- The closer the cosine value to 1, the smaller the angle and the greater the match between vectors
- It is often used to measure document similarity in text analysis

Similarity and Distance Measures

- It can be defined as follows
- sim(x,y) = $\frac{x.y}{||x||||y||}$
- Here ||x|| is the Euclidean norm of vector x = ($x_1, x_2, ..., x_p$) defined as $\sqrt{x_1^2 + x_2^2 + ... + x_p^2}$
- Conceptually, it is the length of the vector
- Similarly, ||y|| is the Euclidean norm of vector y

Similarity and Distance Measures

- A document can be represented using a document vector or term-frequency vector, each recording the frequency of a particular word in the document
- Example - Document Vector

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss	Season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- In document 1, the word Team is repeated 5 times, the word Soccer is present 2 times, but the word Loss is not present

Similarity and Distance Measures

- Suppose that x and y are the first two term-frequency vectors in the above example
- That is, x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) and y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)
- How similar are x and y?
- We can calculate cosine similarity as follows

$$\begin{aligned}
\boldsymbol{x} \cdot \boldsymbol{y} &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 \\
&\quad + 0 \times 0 + 0 \times 1 = 25 \\
||\boldsymbol{x}|| &= \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48 \\
||\boldsymbol{y}|| &= \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12 \\
sim(\boldsymbol{x}, \boldsymbol{y}) &= 0.94.
\end{aligned}$$

- Since this value is closer to 1, the documents are considered to be quite similar

Data Quality

- The following measures are used for ensuring data quality

1. Accuracy

- Inaccurate data may occur due to human or computer errors during data entry
- Example - Human height as 1.73cm instead of 1.73m

2. Completeness

- Data can be incomplete due to non availability
- Sometimes inconsistent data might have been deleted
- Example - In a customer database, some entries may lack age information or income details

Data Quality

3. Consistency

- Inconsistent data may occur due to conflicting formats, contradictory values, variation in spellings etc.
- Example 1 - Dates recorded in different formats (e.g., "YYYY-MM-DD" vs. "DD-MM-YYYY") within the same dataset
- Example 2 - A product's price listed as both Rs.25 and Rs.30 within the same database
- Example 3 - Names recorded with different spellings (e.g., "Jon Smith" and "John Smith") referring to the same individual

Data Quality

4. Timeliness

- We need to update data in a timely fashion; Failing to do so will affect the data quality
- Example - Utilizing player statistics from several years ago for current team assessments

5. Believability

- Data need to be trusted by users
- Sometimes due to past errors in data, users will not believe in them

6. Interpretability

- Data should be easily understood by users

Data Cleaning

- Data Cleaning is the process of removing errors and inconsistencies from a data set
- The following steps are involved in this

1. Elimination of Duplicate Records
2. Dealing with Missing Values

(a) Ignore the tuple with missing value
- This method is not effective, as the tuple can contain several useful attributes

(b) Fill in the missing value manually
- This is a time consuming process

(c) Use a global constant to fill in the missing value
- Example - Fill the missing value with 'Unknown'

(d) Fill the missing value with mean, median or mode of the distribution
- Example - Missing Credit Score
- It can be replaced with a credit score derived from the data set(mean)

Data Cleaning

3 Data Type Conversion

- Depending on the requirement, we convert data from one type to another
- This depends on the data science algorithm we are using
- We can convert credit score to categorical values such as poor = 400, good = 600, excellent = 800

4 Transformation of Attribute Ranges

- Different attributes have different ranges
- For example, range of income is larger compared to range of credit score
- For some data science algorithms, these ranges are normalised to a uniform scale from 0 to 1

Data Cleaning

5 Handling Outliers

- Outliers are anomalies that differ from majority of observations in a data set
- Example - Human height as 1.73cm instead of 1.73m in a record
- We need to correct these anomalies

6. Handling Noisy Data

- A noise is a random error occurring in a measured variable
- It can be handled using a technique called binning
- Binning is the process of converting numeric data into various intervals(bins)
- Example - Replace daily sales with 7 day bin average sales data to reduce weekday noise.

Data Integration

- It is the process of merging of data from multiple data stores
- The following are the issues in it

1. Entity identification problem

- We need to identify equivalent real-world entities from multiple data sources
- Example - How can a data analyst or a computer be sure that *customer_id* in one database and *cust_number* in another refer to the same attribute?

2. Redundancy

- An attribute may be redundant if it can be derived from another attribute or set of attributes
- Example - *Annual revenue* of a company can be derived from *monthly revenue*
- Some redundancies can be detected by correlation analysis

Data Integration

3. Tuple duplication

- It is the existence of identical tuples for the same data item

Customer Id	Name	Email
1	Alice	alice@example.com
2	Bob	bob@example.com
3	Alice	alice@example.com
4	Charlie	charlie@example.com

- Here there are 2 tuples for Alice

4. Data value conflict

- This involves different data values for the same attribute from different sources
- Example - A *weight* attribute may be stored in metric units in one system and British imperial units in another

Data Transformation

- Here data is transformed into forms appropriate for mining
- The following techniques are used here

1. Data normalisation

- Here attribute data are scaled so as to fall within a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0
- One method of data normalisation is min-max normalisation
- Here the new value of X is found using the given below formula
- $X_{new} = \frac{X - min(X)}{max(X) - min(X)}$
- Example - min-max normalisation
- Suppose that the minimum and maximum values for the attribute income are Rs.12,000 and Rs.98,000, respectively
- We would like to map income to the range [0.0,1.0]
- By min-max normalization, a value of Rs.73,600 for income is transformed to $\frac{73600-12000}{98000-12000} = 0.716$

Data Transformation

2. Data Discretisation

- In this technique, the raw values of a numeric attribute are replaced by interval labels or conceptual labels
- Example
- The numeric values of attribute age are replaced by interval labels 0–10, 11–20,... or conceptual labels youth, adult, senior

3. Data Compression

- It is a data reduction technique where original data is transformed into a compressed form
- If the original data can be reconstructed from the compressed data, then the compression is called lossless compression
- If only an approximation of the original data can be reconstructed from the compressed data, then the compression is called lossy compression
- Examples - Discrete Wavelet Transformation (DWT), Discrete Fourier Transformation (DFT)

Data Transformation

4. Sampling

- It is a data reduction technique where a subset of the original data set is selected for analysis

(i) Simple random sample without replacement (SRSWOR)

- Every time a sample is drawn, it is not to be placed back to the original data set

(ii) Simple random sample with replacement (SRSWR)

- Every time a sample is drawn, it is placed back to the original data set so that it may be drawn again

Data Transformation

(iii) Cluster Sample

- Here the original data set is divided into naturally occurring groups called clusters and a random selection of clusters is chosen for analysis
- Example - Surveying all students in randomly chosen schools

(iv) Stratified Sample

- Here the original data set is divided into mutually exclusive subgroups called strata and then samples are drawn independently from each stratum
- Example - For customer data, a stratum is created for each customer age group and samples are drawn from each group

Dimensionality Reduction

- It is the process of reducing the number of attributes or features in the original data set for analysis
- The following methods are used here

1. Principal components analysis(PCA)

- In this method, original set of attributes are combined into smaller set of new attributes called principal components

2. Attribute subset selection

- In this method, a subset of the attributes from the original data set are chosen for analysis
- Example
- Attributes in original data set = {A,B,C,D,E,F,G,H,I}
- Attributes created in PCA = {PC1,PC2} which are the combinations of original attributes
- Attributes in Attributes subset selection method = {A,C,E,G,I}

Dimensionality Reduction

3. Non linear dimensionality reduction methods

- PCA can be used for linear data only ie. when the relationship between attributes can be represented using linear equations
- For nonlinear data ie. when the the relationship between attributes cannot be represented using linear equations, certain other methods are used

(i) Kernel PCA

- It extends the traditional PCA method to handle nonlinear relationships between features
- It uses a special function called kernel function for this

(ii) Stochastic neighbour embedding

- This method is particularly suited for visualising high-dimensional data in lower dimensions, like 2D or 3D

Module 3
Classification Models: Classification - Basic Concepts

- Classification technique predict a target variable based on input variables
- The output variable which is predicted is called a target variable
- In classification, the target variable is a category or class such as 'yes','no', 'red', 'blue' etc.
- Example - Predicting whether monsoon will be normal this year
- A classification model can be represented in various forms

Decision Tree

- Decision tree is a powerful machine learning tool for classification and prediction
- It is a flowchart like tree structure
- Here every non leaf node indicate a test on an attribute
- Each branch represent an outcome on the test
- Every leaf node indicate a class label

Decision Tree

Outlook	Temp	Play?
Sunny	30	Yes
Overcast	15	No
Sunny	16	Yes
Cloudy	27	Yes
Overcast	25	Yes
Overcast	17	No
Cloudy	17	No
Cloudy	35	Yes

Classification Problem

Weather -> Play (Yes, No)

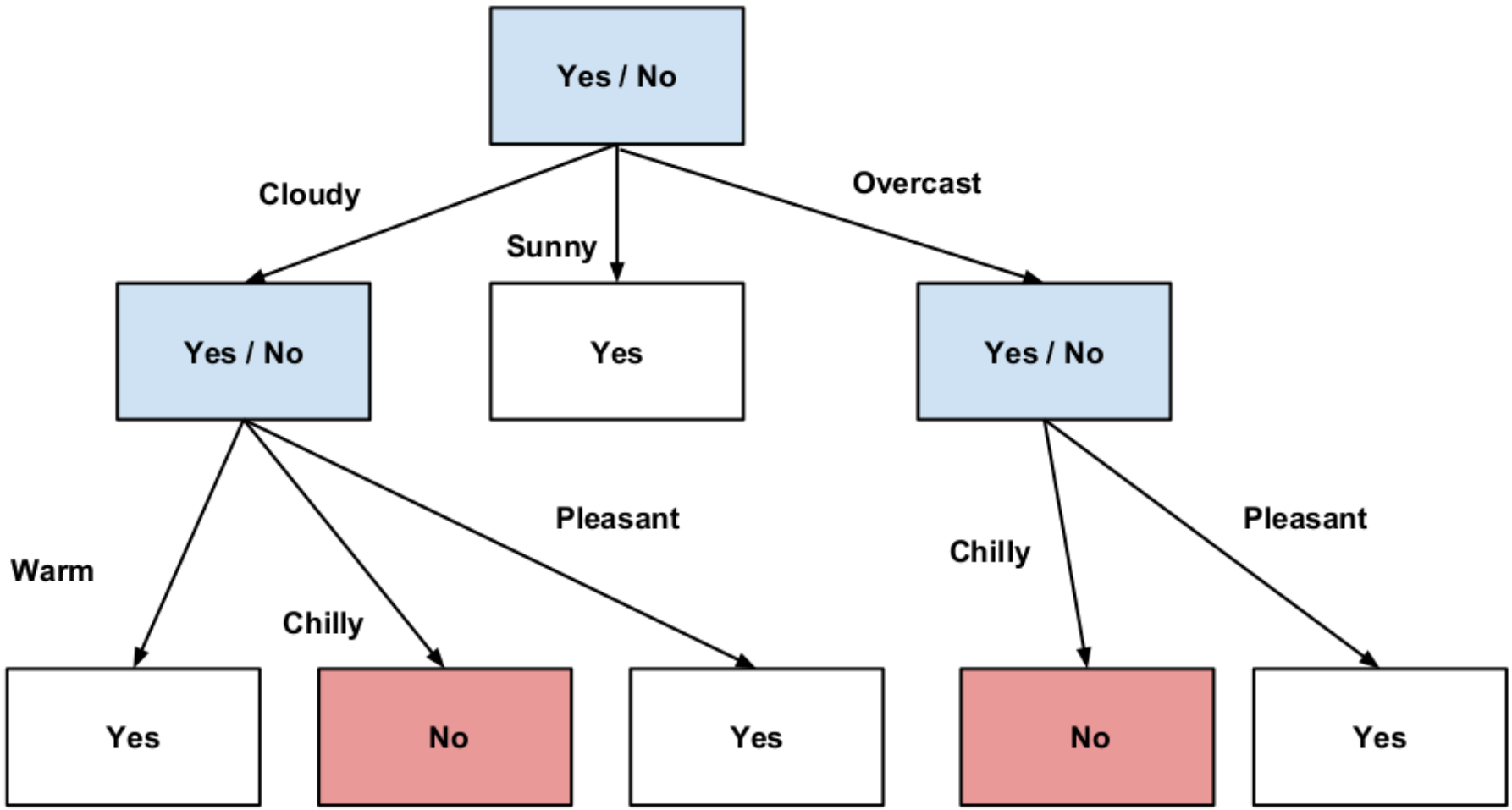
Decision Tree

- Here we convert temperature values from numeric to categorical
- Chilly: < 20
- Pleasant: 20 - 30
- Warm: > 30

Outlook	Temp	Play?
Sunny	Warm	Yes
Overcast	Chilly	No
Sunny	Chilly	Yes
Cloudy	Pleasant	Yes
Overcast	Pleasant	Yes
Overcast	Chilly	No
Cloudy	Chilly	No
Cloudy	Warm	Yes

Decision Tree

- Decision Tree for Classification



Decision Tree

- For classification, decision trees use divide and conquer approach
- Initially the whole data set is divided into several subsets
- These subsets are again divided into even smaller subsets and so on
- This process is continued until we arrive at the solution
- At first, the root node represent the entire data set
- Next the decision tree algorithm chooses a feature or attribute to split upon
- Based on the distinct values of this feature, the data set is partitioned into groups, and the first set of tree branches are formed
- Next we choose another feature to split and this process is continued till we arrive at the final solution

C5.0 Decision Tree Algorithm

- We can implement decision trees in different ways
- It is an efficient implementation developed by computer scientist J. Ross Quinlan
- Choosing the best split
- When there are several features in the problem, we have to decide which feature is to be split initially and subsequently after each iteration
- This is done based on 2 metrics called entropy and information gain
- Entropy is a measure of disorder among a set of class values

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \, log_2(p_i)$$

C5.0 Decision Tree Algorithm

- Here S is the data segment
- c refers to the number of class levels
- p_i refers to the proportion of values falling into class level i
- suppose we have a partition of data with two classes: red (60 percent) and white (40 percent)
- We can calculate the entropy as follows
- Entropy(S) = -0.60 * log2(0.60) - 0.40 * log2(0.40) = 0.9709506
- information gain = entropy(parent) - entropy(children)
- Here entropy(children) is the average entropy of child nodes
- We have to make splitting in such a way that there is more information gain

C5.0 Decision Tree Algorithm

- Pruning the decision trees
- Here the size of the decision tree is reduced by cutting of the unwanted branches
- It can be done in 2 ways
- Pre-Pruning (Early Stopping) stops the tree from growing once it reaches certain number of decisions
- Post-Pruning allows the tree to grow larger, before its leaf nodes are cut to reduce its size
- It involves the following processes
- Subtree Raising involves moving branches further up in the tree
- Subtree Replacement involves replacing branches by simpler decisions

References

1. Fundamentals of data science, Wagh, S. J., Bhende, M. S., & Thakare, A.D., Chapman and Hall/CRC, Ist Edition 2021 (Module 1)
2. Data mining: concepts and techniques, Han, J., Pei, J., & Tong, H, Morgan kaufmann, 2023