Informed Search Strategies

Module 2

Informed (Heuristic) Search strategy

- Informed search strategy—one that uses problem-specific knowledge beyond the definition of the problem itself—can find solutions more efficiently than can an uninformed strategy.
- The general approach we consider is called best-first search.
  - Best-first search is an instance of the general TREE-SEARCH or GRAPH-SEARCH algorithm in which a node is selected for expansion based on an evaluation function, f(n).
  - The evaluation function is construed as a cost estimate, so the node with the lowest evaluation is expanded first.
  - The implementation of best-first graph search is identical to that for uniform-cost search, except for the use of f instead of g to order the priority queue.

Heuristic function

- The choice of f determines the search strategy. Most best-first algorithms include as a component of f a heuristic function, denoted h(n):

  h(n) = estimated cost of the cheapest path from the state at node n to a goal state

- Heuristic functions are the most common form in which additional knowledge of the problem is imparted to the search algorithm. For now, we consider them to be arbitrary, nonnegative, problem-specific functions, with one constraint:

  if n is a goal node, then h(n)=0.

- The remainder of this section covers two ways to use heuristic information to guide search.

Example

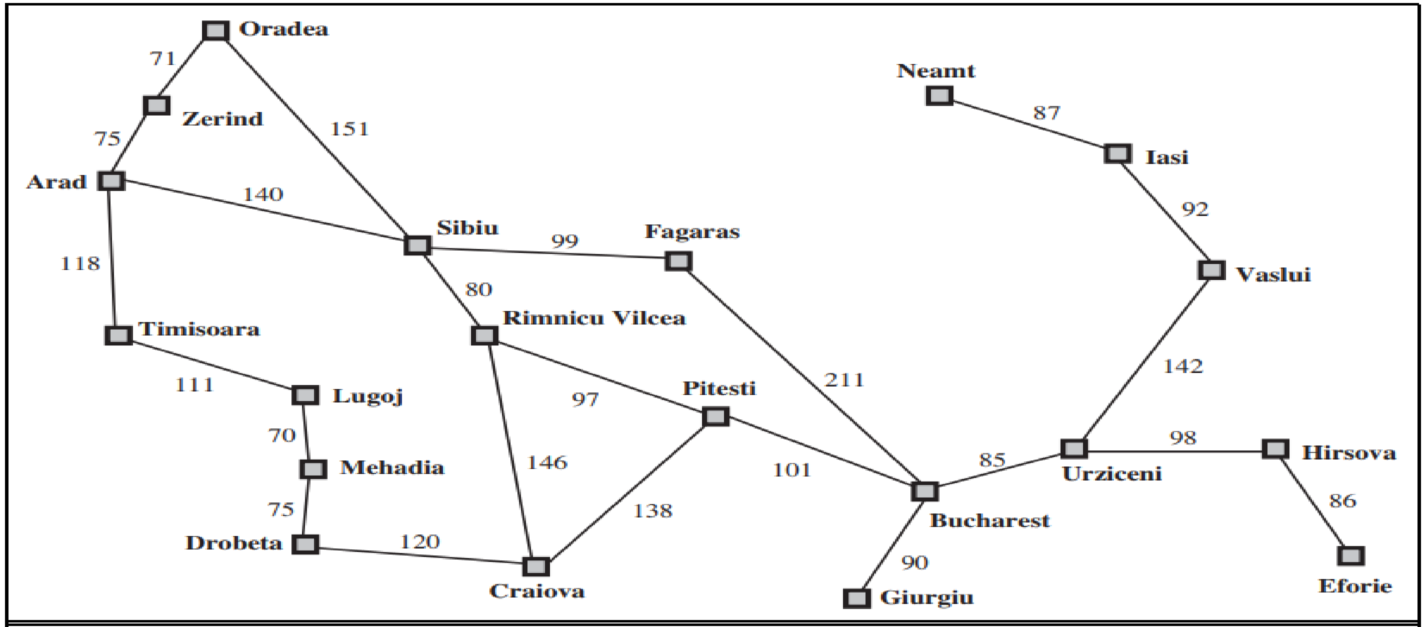
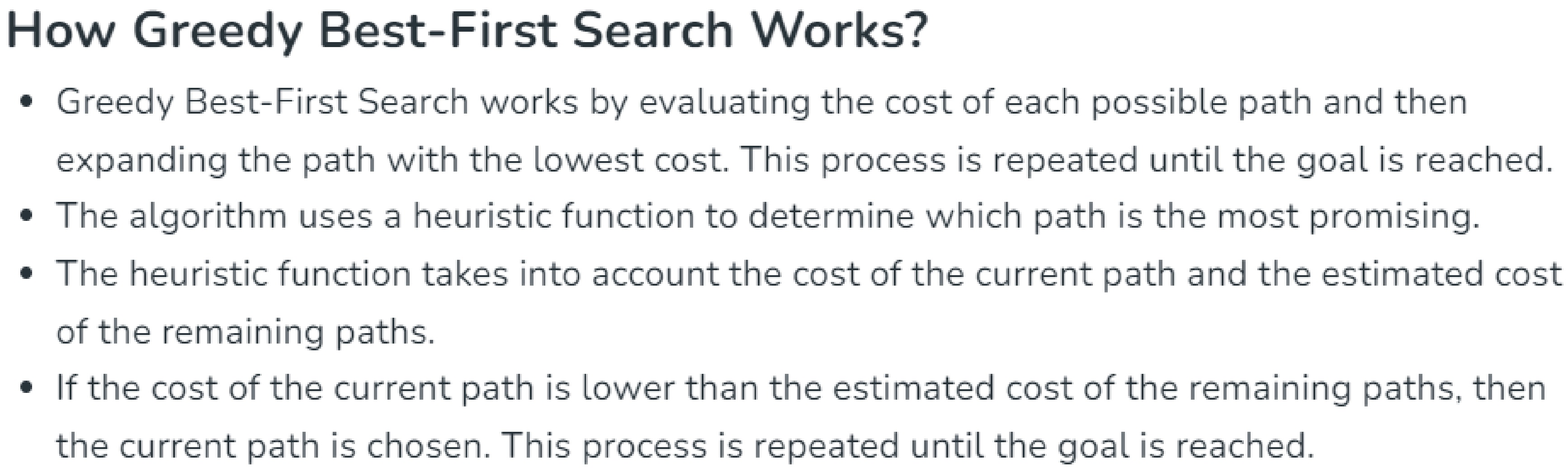


Figure 3.2 A simplified road map of part of Romania.



How Greedy Best-First Search Works?

- Greedy Best-First Search works by evaluating the cost of each possible path and then expanding the path with the lowest cost. This process is repeated until the goal is reached.
- The algorithm uses a heuristic function to determine which path is the most promising.
- The heuristic function takes into account the cost of the current path and the estimated cost of the remaining paths.
- If the cost of the current path is lower than the estimated cost of the remaining paths, then the current path is chosen. This process is repeated until the goal is reached.

Greedy Best First Search

- Greedy best-first search tries to expand the node that is closest to the goal, that this is likely to lead to a solution quickly.
- Thus, it evaluates nodes by using just the heuristic function; that is,

$f(n) = h(n)$.

Working:

- we use the straight- line distance heuristic, which we will call h_{SLD} . If the goal is Bucharest, we need to know the straight-line distances to Bucharest
- h_{SLD} (In(Arad)) = 366

Example of Greedy Best First Search

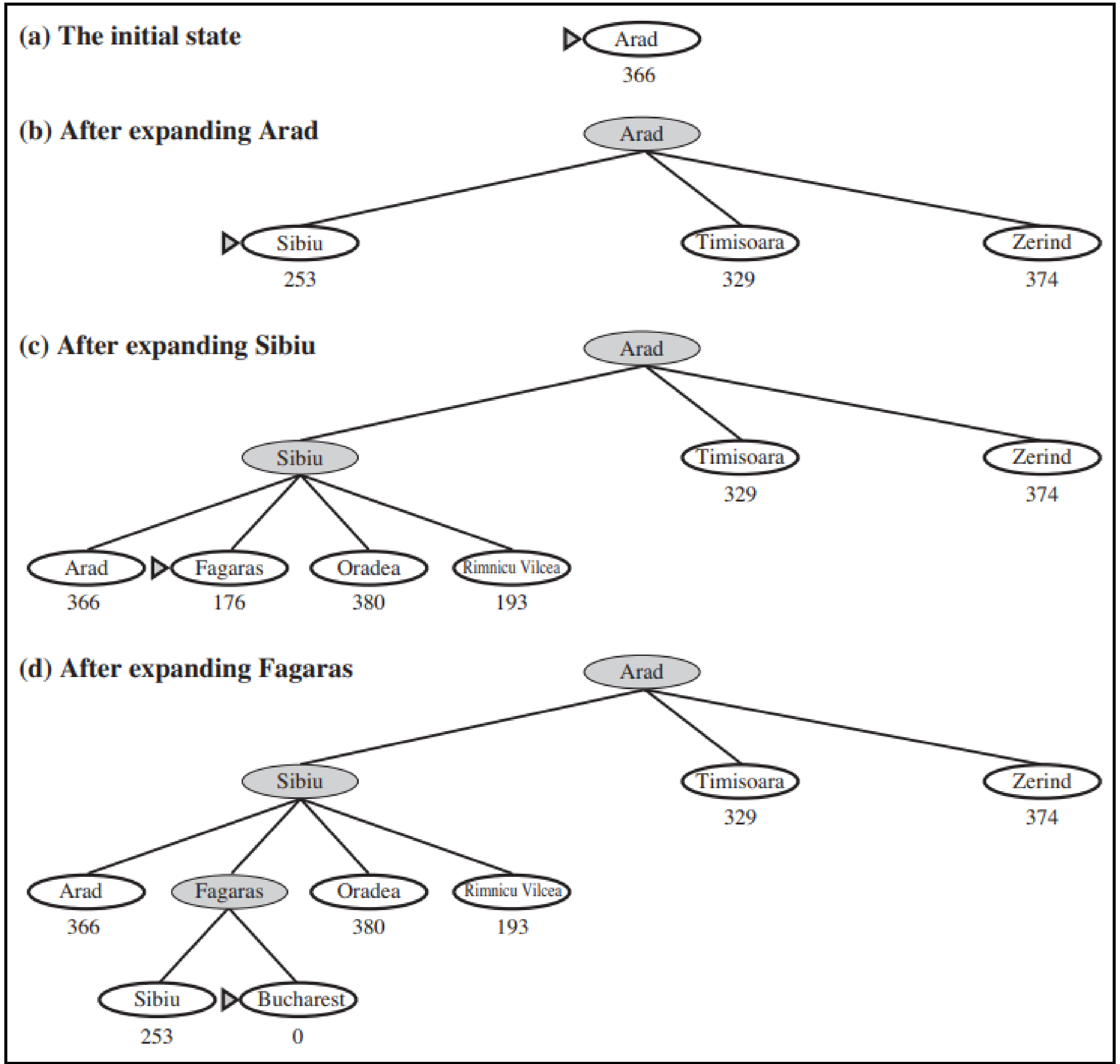
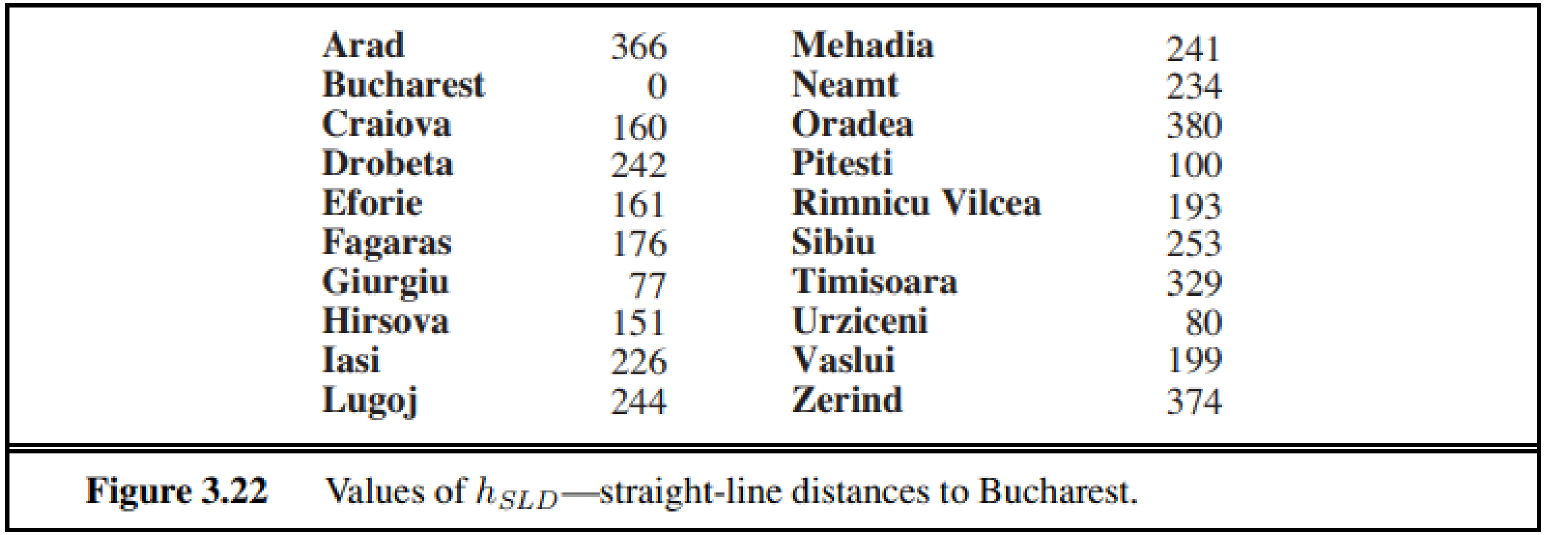


Figure 3.23 Stages in a greedy best-first tree search for Bucharest with the straight-line distance heuristic h_{SLD}. Nodes are labeled with their h-values.

Greedy Best First Search



Arad	366	**Mehadia**	241
Bucharest	0	**Neamt**	234
Craiova	160	**Oradea**	380
Drobeta	242	**Pitesti**	100
Eforie	161	**Rimnicu Vilcea**	193
Fagaras	176	**Sibiu**	253
Giurgiu	77	**Timisoara**	329
Hirsova	151	**Urziceni**	80
Iasi	226	**Vaslui**	199
Lugoj	244	**Zerind**	374

Figure 3.22 Values of h_{SLD}—straight-line distances to Bucharest.

- Figure 3.23 shows the progress of a greedy best-first search using h_{SLD} to find a path from Arad to Bucharest.
- The first node to be expanded from Arad will be Sibiu because it is closer to Bucharest than either Zerind or Timisoara.
- The next node to be expanded will be Fagaras because it is closest. Fagaras in turn generates Bucharest, which is the goal.
- For this particular problem, greedy best-first search using h_{SLD} finds a solution without ever expanding a node that is not on the solution path; hence, its search cost is minimal

Greedy Best First Search

- It is not optimal,
- However: the path via Sibiu and Fagaras to Bucharest is 32 kilometers longer than the path through Rimnicu Vilcea and Pitesti.
- This shows why the algorithm is called “greedy”—at each step it tries to get as close to the goal as it can.

- Greedy best-first tree search is also incomplete even in a finite state space, much like depth-first search.

Drawback Greedy Best First Search

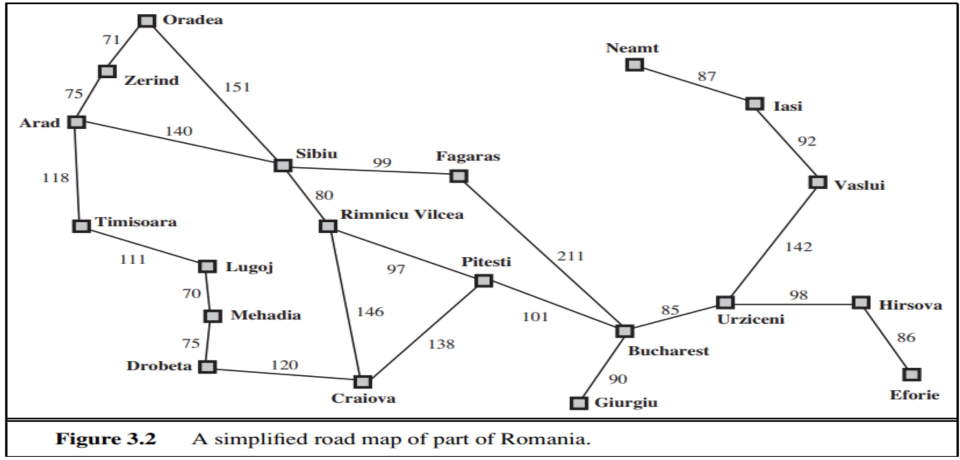


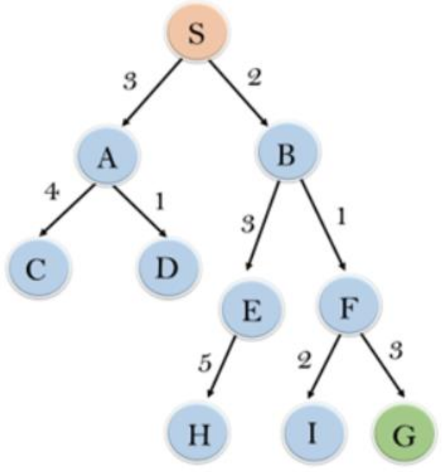
Figure 3.2 A simplified road map of part of Romania.

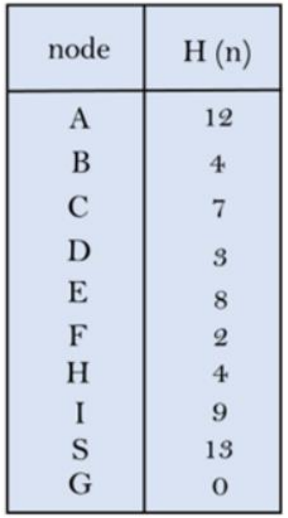
- Consider the problem of getting from Iasi to Fagaras
- The heuristic suggests that Neamt be expanded first because it is closest to Fagaras, but it is a dead end.
- The solution is to go first to Vaslui—a step that is actually farther from the goal according to the heuristic—and then to continue to Urziceni, Bucharest, and Fagaras.
- The algorithm will never find this solution, however, because expanding Neamt puts Iasi back into the frontier, Iasi is closer to Fagaras than Vaslui is, and so Iasi will be expanded again, leading to an infinite loop.
- The worst-case time and space complexity for the tree version is O(b^{m}), where m is the maximum depth of the search space.
- With a good heuristic function, however, the complexity can be reduced substantially. The amount of the reduction depends on the particular problem and on the quality of the heuristic

General Example to understand Greedy best first search

Example:

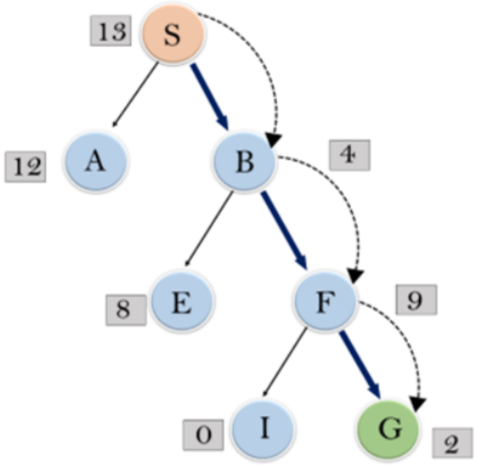
Consider the below search problem, and we will traverse it using greedy best-first search. At each iteration, each node is expanded using evaluation function f(n)=h(n) , which is given in the below table.





node	H (n)
A	12
B	4
C	7
D	3
E	8
F	2
H	4
I	9
S	13
G	0

In this search example, we are using two lists which are **OPEN** and **CLOSED** Lists. Following are the iteration for traversing the above example.



Expand the nodes of S and put in the CLOSED list

Initialization: Open [A, B], Closed [S]

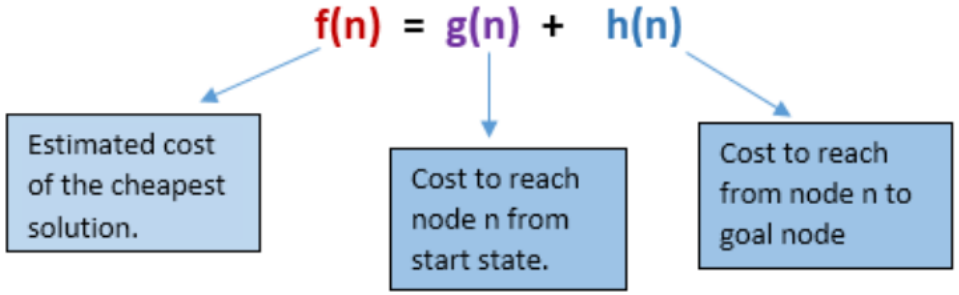
Iteration 1: Open [A], Closed [S, B]

Iteration 2: Open [E, F, A], Closed [S, B]
: Open [E, A], Closed [S, B, F]

Iteration 3: Open [I, G, E, A], Closed [S, B, F]
: Open [I, E, A], Closed [S, B, F, G]

Hence the final solution path will be: **S----> B----->F----> G**

A* search: Minimizing the total estimated solution cost



f(n) = g(n) + h(n)

Estimated cost of the cheapest solution.

Cost to reach node n from start state.

Cost to reach from node n to goal node

- The most widely known form of best-first search is called A∗ A search (pronounced "A-star search").
- It evaluates nodes by combining g(n), the cost to reach the node, and h(n), the cost to get from the node to the goal:

$$f(n) = g(n) + h(n)$$

- Since g(n) gives the path cost from the start node to node n, and h(n) is the estimated cost of the cheapest path from n to the goal, we have f(n) = estimated cost of the cheapest solution through n
- Thus, if we are trying to find the cheapest solution, a reasonable thing to try first is the node with the lowest value of g(n) + h(n).

Conditions for optimality: Admissibility and consistency

- The first condition we require for optimality is that h(n) be an admissible heuristic.
- An admissible heuristic is one that never overestimates the cost to reach the goal.
- Because g(n) is the actual cost to reach n along the current path, and f(n) = g(n) + h(n), we have as an immediate consequence that f(n) never overestimates the true cost of a solution along the current path through n.
- Admissible heuristics are by nature optimistic because they think the cost of solving the problem is less than it actually is.
  - An obvious example of an admissible heuristic is the straight-line distance hSLD that we used in getting to Bucharest.
  - Straight-line distance is admissible because the shortest path between any two points is a straight line, so the straight line cannot be an overestimate.

A* Search

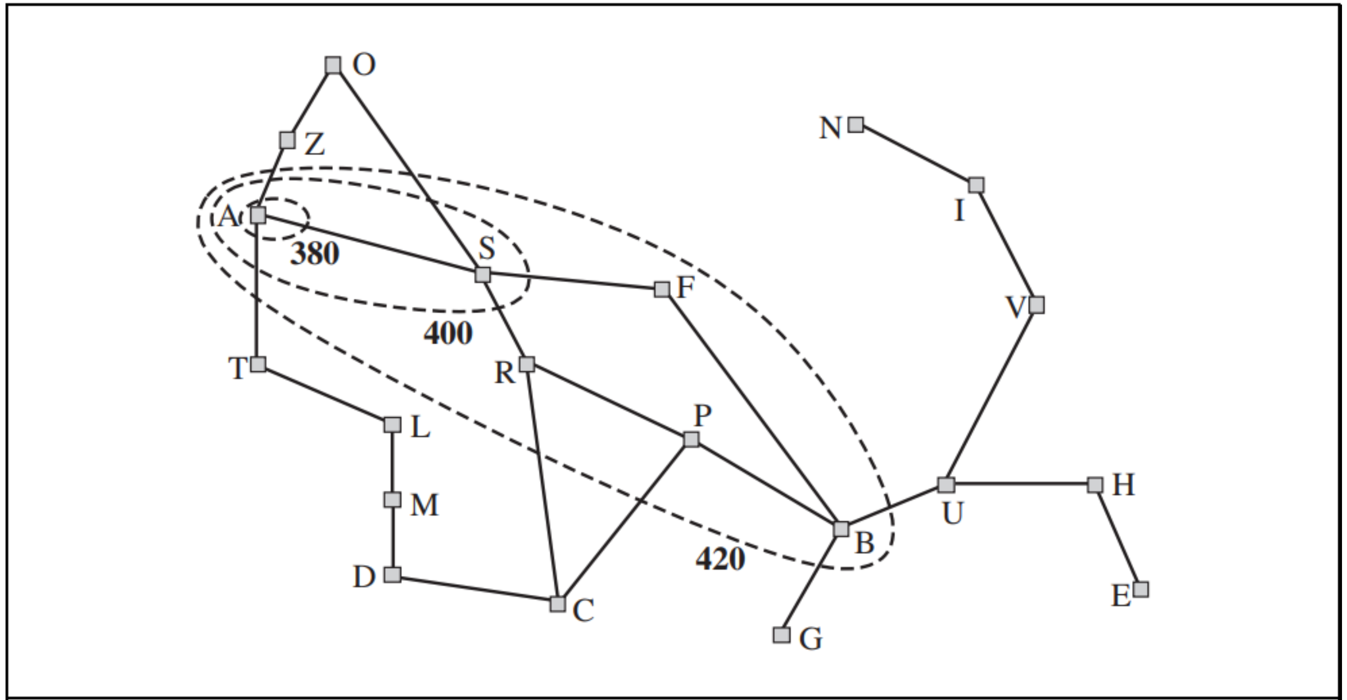
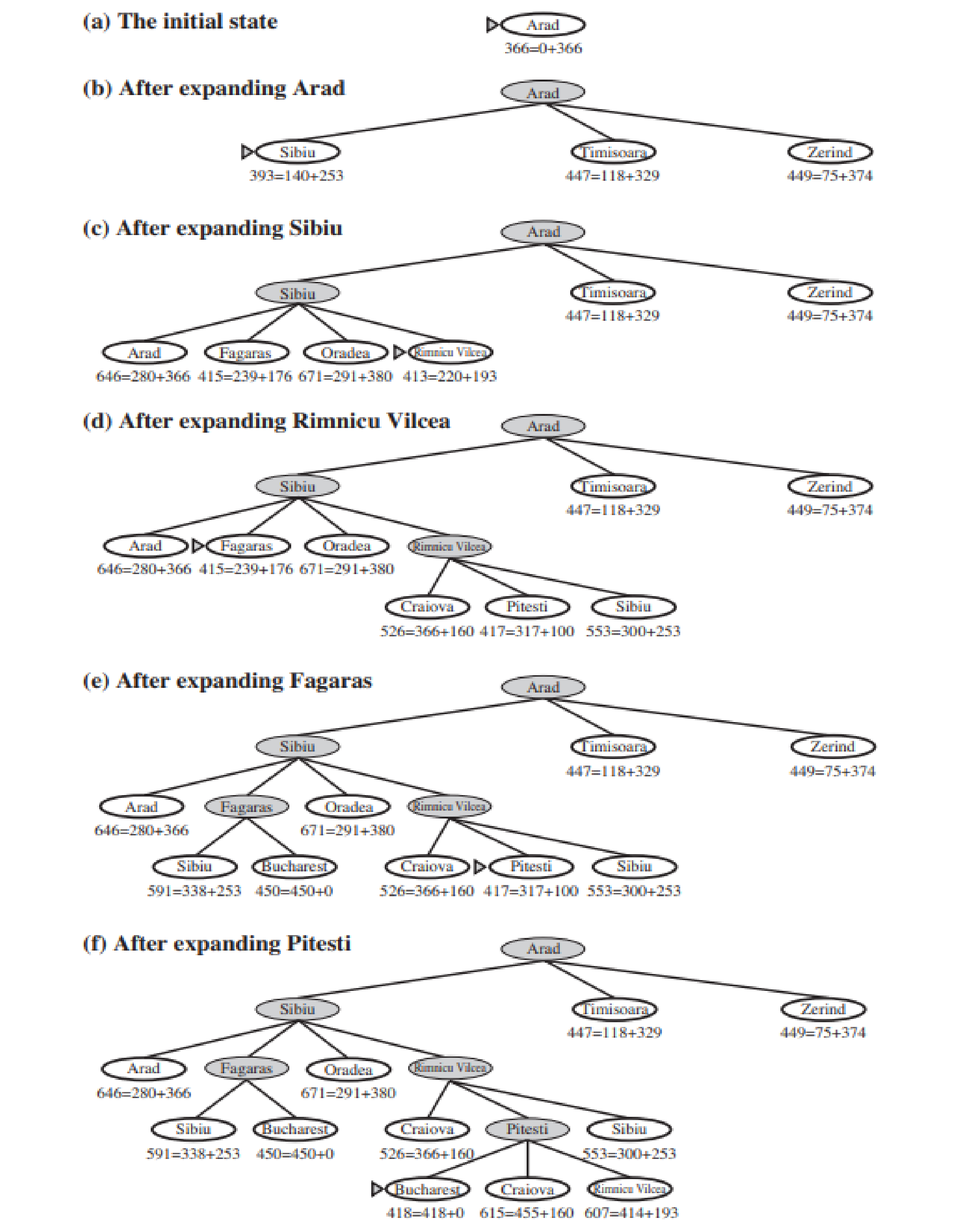


Figure 3.25 Map of Romania showing contours at $f = 380$, $f = 400$, and $f = 420$, with Arad as the start state. Nodes inside a given contour have f-costs less than or equal to the contour value.



(a) The initial state

(b) After expanding Arad

(c) After expanding Sibiu

(d) After expanding Rimnicu Vilcea

(e) After expanding Fagaras

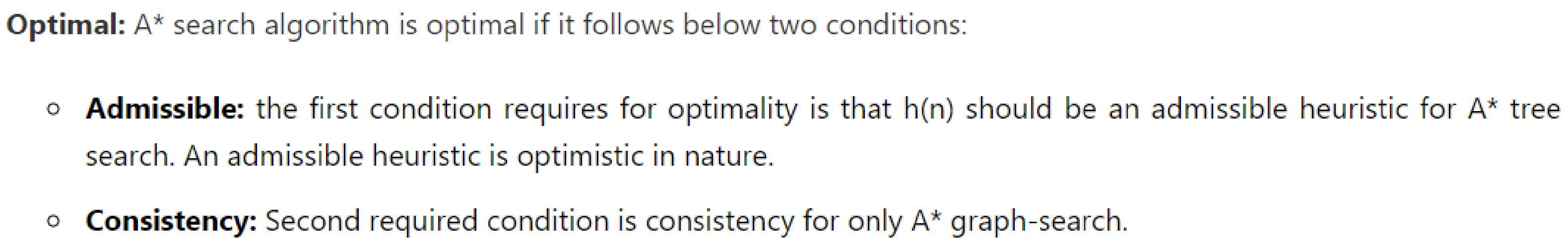
(f) After expanding Pitesti

Figure 3.24 Stages in an A* search for Bucharest. Nodes are labeled with $f = g + h$. The h values are the straight-line distances to Bucharest taken from Figure 3.22.

Drawback A* Search

- In Figure 3.24, we show the progress of an A∗ tree search for Bucharest
- Notice in particular that Bucharest first appears on the frontier at step (e), but it is not selected for expansion because its f-cost (450) is higher than that of Pitesti (417).
- Another way to say this is that there might be a solution through Pitesti whose cost is as low as 417, so the algorithm will not settle for a solution that costs 450.

Optimality of A*



Optimal: A* search algorithm is optimal if it follows below two conditions:

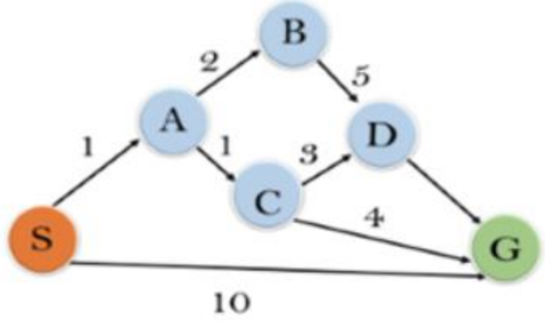
- **Admissible:** the first condition requires for optimality is that h(n) should be an admissible heuristic for A* tree search. An admissible heuristic is optimistic in nature.
- **Consistency:** Second required condition is consistency for only A* graph-search.

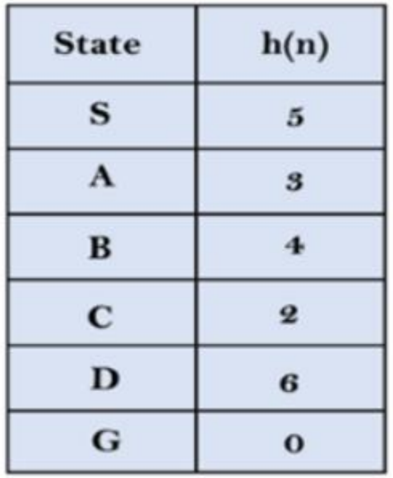
- A∗ has the following properties:
  - Tree-search version of A∗ is optimal if h(n) is admissible, while the graph-search version is optimal if h(n) is consistent.
  - We show the second of these two claims since it is more useful.
  - The argument essentially mirrors the argument for the optimality of uniform-cost search, with g replaced by f—just as in the A∗ algorithm itself.
  - The first step is to establish the following: if h(n) is consistent, then the values of f(n) along any path are nondecreasing.
  - The proof follows directly from the definition of consistency.
  - Suppose n is a successor of n; then g(n) = g(n) + c(n, a, n) for some action a, and we have

$$f(n\ ) = g(n\ ) + h(n\ ) = g(n) + c(n, a, n\ ) + h(n\ ) \geq g(n) + h(n) = f(n)$$

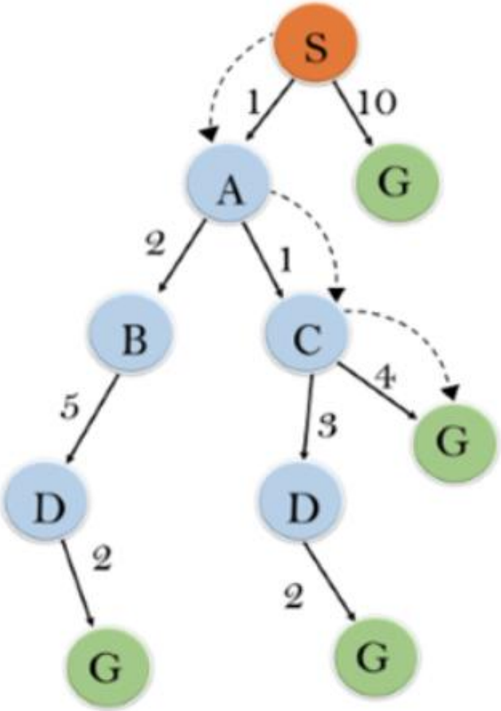
The next step is to prove that whenever A∗ selects a node n for expansion, the optimal path to that node has been found.

Example – A*





State	h(n)
S	5
A	3
B	4
C	2
D	6
G	0



Initialization: {(S, 5)}

Iteration1: {(S--> A, 4), (S-->G, 10)}

Iteration2: {(S--> A-->C, 4), (S--> A-->B, 7), (S-->G, 10)}

Iteration3: {(S--> A-->C--->G, 6), (S--> A-->C--->D, 11), (S--> A-->B, 7), (S-->G, 10)}

Iteration 4 will give the final result, as **S--->A--->C--->G** it provides the optimal path with cost 6.

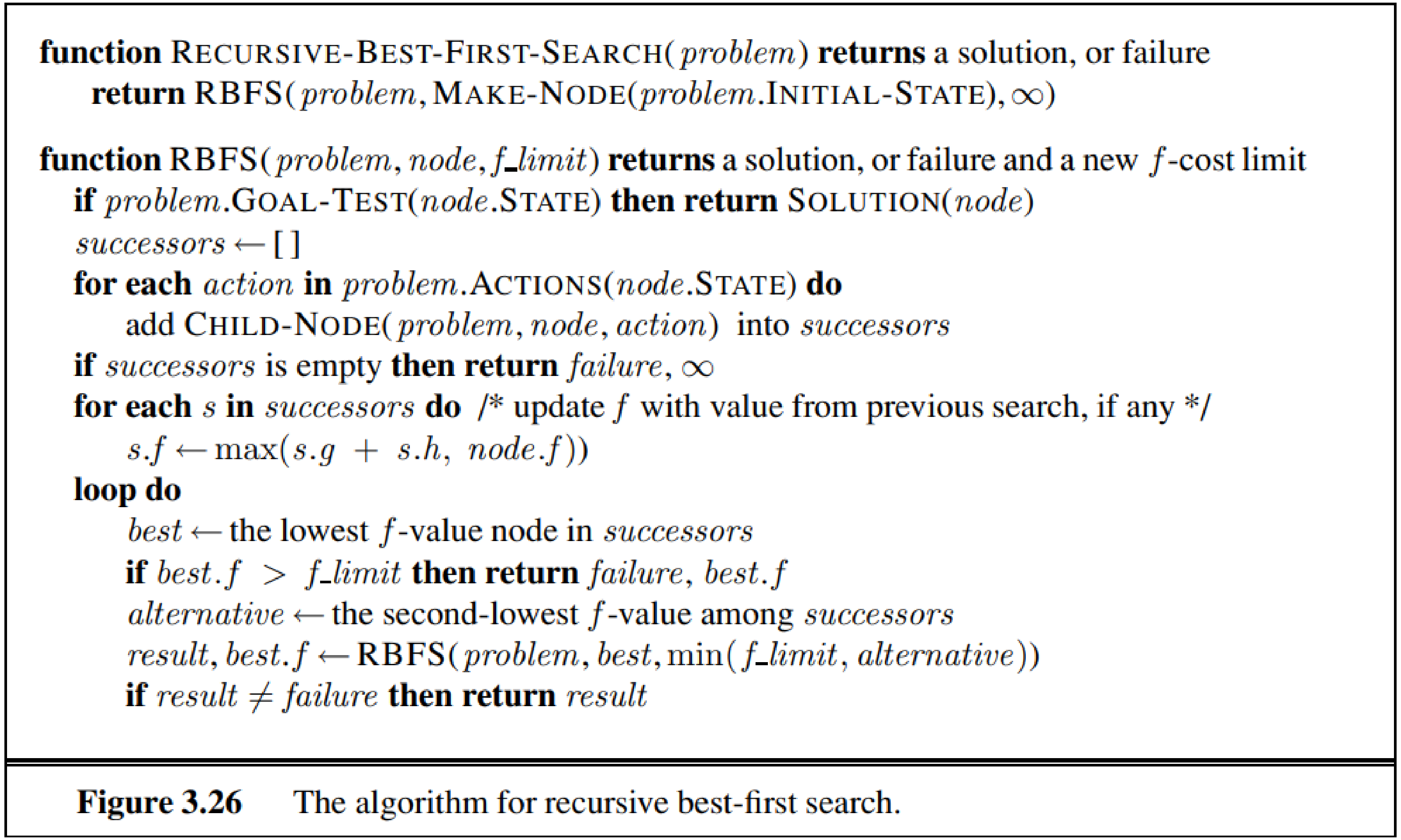
Memory-bounded heuristic search

- The simplest way to reduce memory requirements for A∗ is to adapt the idea of iterative deepening to the heuristic search context, resulting in the iterative-deepening A∗ (IDA∗) algorithm.
- The main difference between IDA∗ and standard iterative deepening is that the cutoff used is the f-cost (g +h) rather than the depth; at each iteration, the cutoff value is the smallest f-cost of any node that exceeded the cutoff on the previous iteration.
- IDA∗ is practical for many problems with unit step costs and avoids the substantial overhead associated with keeping a sorted queue of nodes.

Recursive best-first search

- Is a simple recursive algorithm that attempts to mimic the operation of standard best-first search, but using only linear space.
- The algorithm is shown in Figure 3.26.
- Its structure is similar to that of a recursive depth-first search, but rather than continuing indefinitely down the current path, it uses the f limit variable to keep track of the f-value of the best alternative path available from any ancestor of the current node.
- If the current node exceeds this limit, the recursion unwinds back to the alternative path.
- As the recursion unwinds, RBFS replaces the f-value of each node along the path with a backed-up value—the best f-value of its children.
- In this way, RBFS remembers the f-value of the best leaf in the forgotten subtree and can therefore decide whether it’s worth re-expanding the subtree at some later time.

Recursive Best First Search



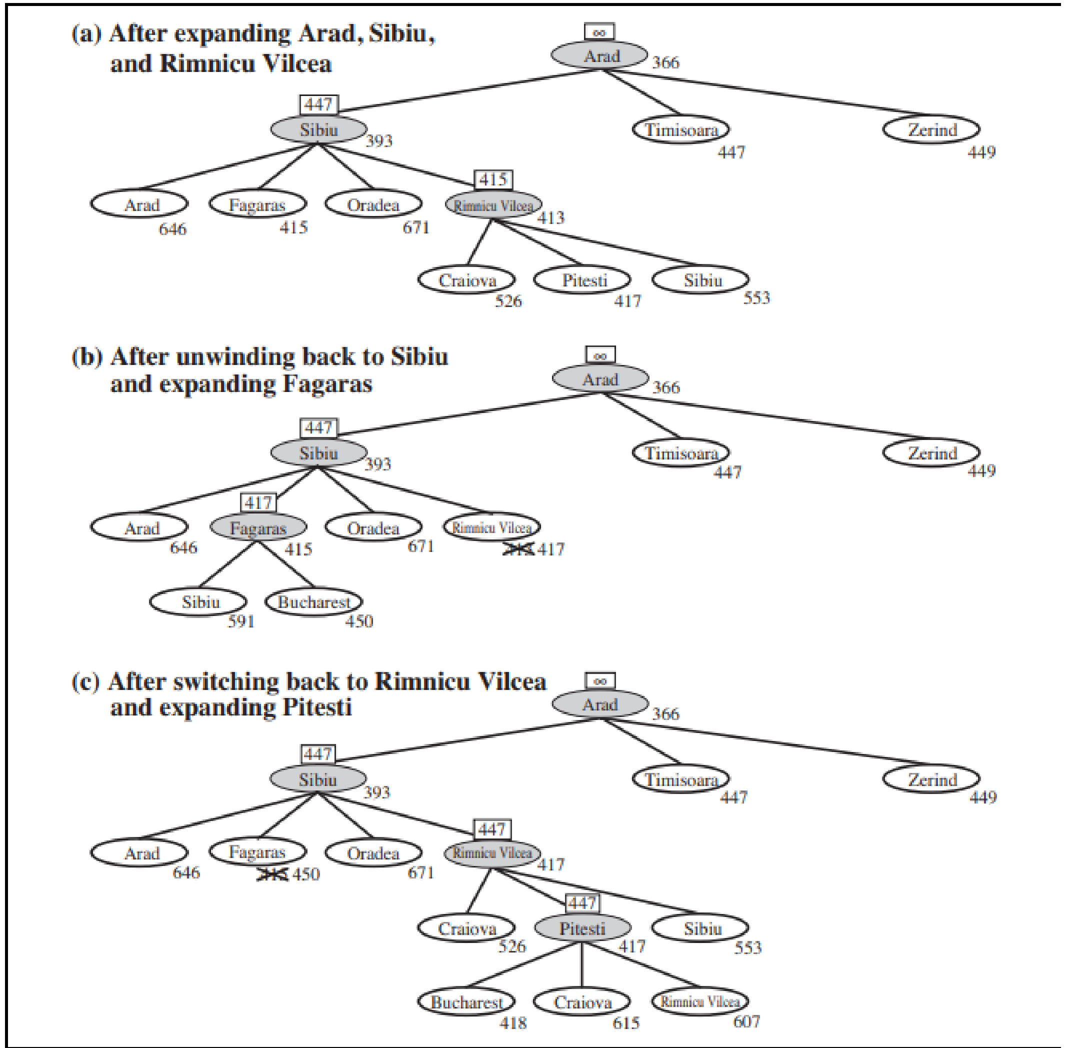
```
function RECURSIVE-BEST-FIRST-SEARCH(problem) returns a solution, or failure
   return RBFS(problem, MAKE-NODE(problem.INITIAL-STATE), $\infty$)

function RBFS(problem, node, f_limit) returns a solution, or failure and a new $f$-cost limit
  if problem.GOAL-TEST(node.STATE) then return SOLUTION(node)
  successors $\leftarrow$ [ ]
  for each action in problem.ACTIONS(node.STATE) do
     add CHILD-NODE(problem, node, action) into successors
  if successors is empty then return failure, $\infty$
  for each s in successors do  /* update $f$ with value from previous search, if any */
     $s.f \leftarrow \max(s.g + s.h,\ node.f)$
  loop do
     best $\leftarrow$ the lowest $f$-value node in successors
     if $best.f > f\_limit$ then return failure, $best.f$
     alternative $\leftarrow$ the second-lowest $f$-value among successors
     result, $best.f \leftarrow$ RBFS(problem, best, min(f_limit, alternative))
     if result $\neq$ failure then return result
```

Figure 3.26 The algorithm for recursive best-first search.

RBFS

- Figure 3.27 shows how RBFS reaches Bucharest
- RBFS is somewhat more efficient than IDA∗, but still suffers from excessive node regeneration.
- In the example in Figure 3.27, RBFS follows the path via Rimnicu Vilcea, then “changes its mind” and tries Fagaras, and then changes its mind back again.
- These mind changes occur because every time the current best path is extended, its f-value is likely to increase—h is usually less optimistic for nodes closer to the goal.



(a) After expanding Arad, Sibiu, and Rimnicu Vilcea

(b) After unwinding back to Sibiu and expanding Fagaras

(c) After switching back to Rimnicu Vilcea and expanding Pitesti

Figure 3.27 Stages in an RBFS search for the shortest route to Bucharest. The *f*-limit value for each recursive call is shown on top of each current node, and every node is labeled with its *f*-cost. (a) The path via Rimnicu Vilcea is followed until the current best leaf (Pitesti) has a value that is worse than the best alternative path (Fagaras). (b) The recursion unwinds and the best leaf value of the forgotten subtree (417) is backed up to Rimnicu Vilcea; then Fagaras is expanded, revealing a best leaf value of 450. (c) The recursion unwinds and the best leaf value of the forgotten subtree (450) is backed up to Fagaras; then Rimnicu Vilcea is expanded. This time, because the best alternative path (through Timisoara) costs at least 447, the expansion continues to Bucharest.

Learning to search better

- We have presented several fixed strategies—breadth-first, greedy best-first, and so on—that have been designed by computer scientists.
- Could an agent learn how to search better?
- The answer is yes, and the method rests on an important concept called the metalevel state space.
- Each state in a metalevel state space captures the internal (computational) state of a program that is searching in an object-level state space such as Romania

Learning to search better

- For example, the internal state of the A∗ algorithm consists of the current search tree. Each action in the metalevel state space is a computation step that alters the internal state;
  - for example, each computation step in A∗ expands a leaf node and adds its successors to the tree.
  - Thus, Figure 3.24, which shows a sequence of larger and larger search trees, can be seen as depicting a path in the metalevel state space where each state on the path is an object-level search tree.

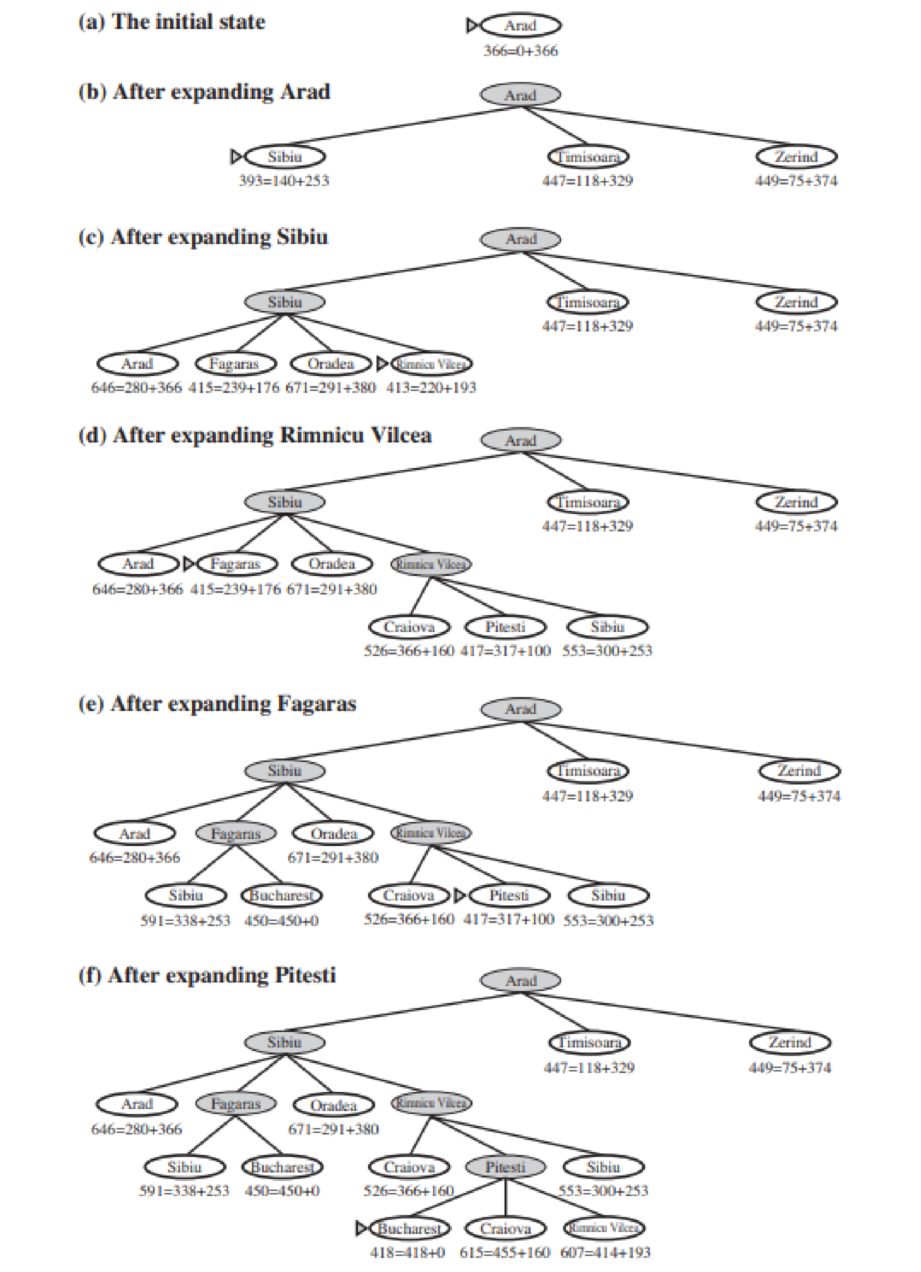


Figure 3.24 Stages in an A* search for Bucharest. Nodes are labeled with $f = g + h$. The h values are the straight-line distances to Bucharest taken from Figure 3.22.

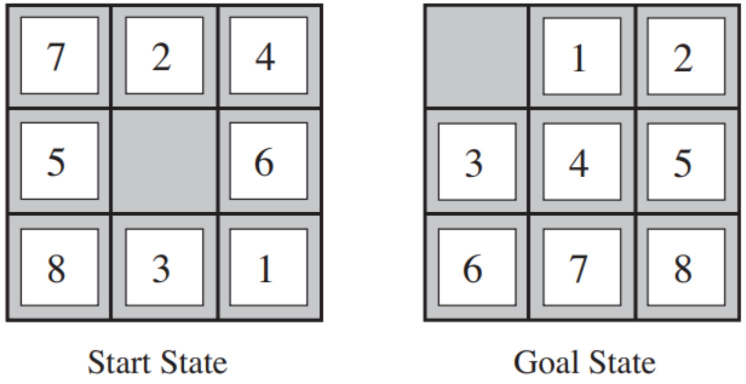
Learning to search better

- Now, the path in Figure 3.24 has five steps, including one step, the expansion of Fagaras, that is not especially helpful.
- For harder problems, there will be many such missteps, and a metalevel learning algorithm can learn from these experiences to avoid exploring unpromising subtrees.
- The goal of learning is to minimize the total cost of problem solving, trading off computational expense and path cost.

Heuristic Functions

- Let us consider an 8-puzzle problem, the objective of the puzzle is to slide the tiles horizontally or vertically into the empty space until the configuration matches the goal configuration
- The average solution cost for a randomly generated 8-puzzle instance is about 22 steps. The branching factor is about 3
  - When the empty tile is in the middle, four moves are possible; when it is in a corner, two; and when it is along an edge, three.
  - This means that an exhaustive tree search to depth 22 would look at about 322 ≈ 3.1 × 10^{10} states
  - A graph search would cut this down by a factor of about 170,000 because only 9!/2 = 181, 440 distinct states are reachable. This is a manageable number, but the corresponding number for the 15-puzzle is roughly 10^{13}

Heuristic Function



7	2	4
5		6
8	3	1

	1	2
3	4	5
6	7	8

Figure 3.28 A typical instance of the 8-puzzle. The solution is 26 steps long.

- If we want to find the shortest solutions by using A∗, we need a heuristic function that never overestimates the number of steps to the goal.
- There is a long history of such heuristics for the 8-puzzle; here are two commonly used candidates:
  - h1 = the number of misplaced tiles. For Figure 3.28, all of the eight tiles are out of position, so the start state would have h1 = 8. h1 is an admissible heuristic because it is clear that any tile that is out of place must be moved at least once.
  - h2 = the sum of the distances of the tiles from their goal positions. Because tiles cannot move along diagonals, the distance we will count is the sum of the horizontal and vertical distances. This is sometimes called the city block distance or Manhattan distance

Heuristic Function

- h2 is also admissible because all any move can do is move one tile one step closer to the goal.
- Tiles 1 to 8 in the start state give a Manhattan distance of

$$h2 = 3 + 1 + 2 + 2 + 2 + 3 + 3 + 2 = 18$$

- As expected, neither of these overestimates the true solution cost, which is 26

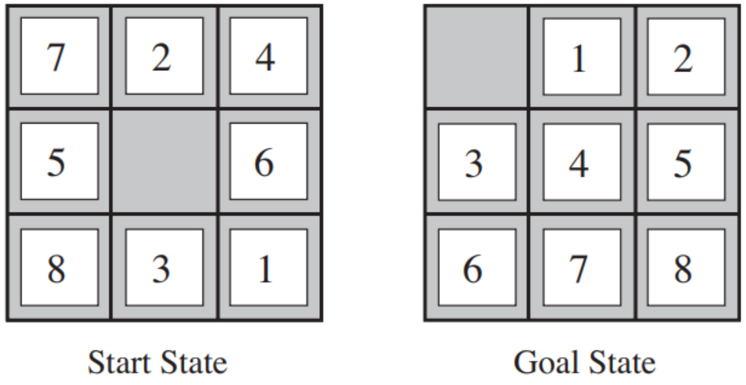


Figure 3.28 A typical instance of the 8-puzzle. The solution is 26 steps long.

The effect of heuristic accuracy on performance

- One way to characterize the quality of a heuristic is the effective branching factor b∗.
- If the total number of nodes generated by A∗ for a particular problem is N and the solution depth is d, then b∗ is the branching factor that a uniform tree of depth d would have to have in order to contain N + 1 nodes.
- Thus, $N + 1 = 1 + b* + (b*)^2 + \cdots + (b*)^d$
- For example, if A∗ finds a solution at depth 5 using 52 nodes, then the effective branching factor is 1.92.

Introduction to Machine Learning

Module 2

Textbook 2

Need for Machine Learning

- Use of huge data in business organization
- Lack of awareness about software tool to help extract information from data
- Business organizations has now started to use Machine Learning for these purposes

Popularity of Machine learning

- High Volume of data available to manage
- Cost of storage has reduced
- Availability of complex algorithm

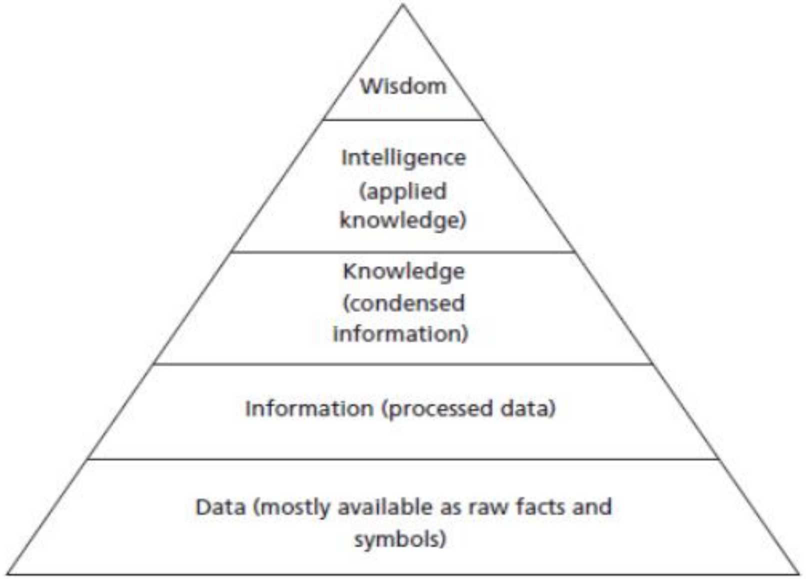

Wisdom
Intelligence (applied knowledge)
Knowledge (condensed information)
Information (processed data)
Data (mostly available as raw facts and symbols)

Figure 1.1: The Knowledge Pyramid

What is data?

- Data are facts
- Data can be number or text that can be processed by a computer
  - Organizations are accumulating vast and growing data wit data sources such as flat files, databases or data warehouse in different storage formats
- Processed data is called information
  - This includes patterns, associations or relationships among data
  - Eg: sales data can be analyzed to extract information like which is fast selling, which products are brought together
- Historical patterns and future trends obtained from the sales data is called knowledge
- Unless knowledge is extracted data is of no use, also knowledge is not useful if it is not put into action

Need for Machine Learning

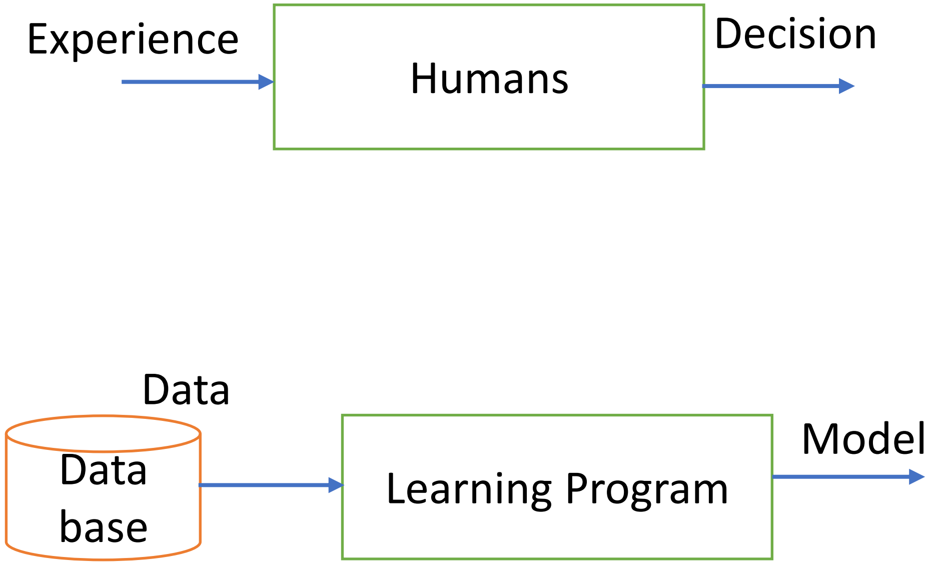
- The objective of machine learning is to process these data for organizations to take better decisions in
  - designing new products
  - improve the business processes and
  - to develop effective decision support system

Machine Learning Explained

- It is an important sub-branch of AI
- One of the definition:
  - Machine learning is the field of study that gives the computer ability to learn without being explicitly programmed
- The systems should learn by itself without explicit programming
  - In conventional programming, after understanding the problem, a detailed design of the program such as flowchart or an algorithm, which is then converted to into programs using suitable language
  - Initially, AI aims to understand these problems and develop general purpose rules, these rules are formulated into logic and reasoning by converting an expert's knowledge into a set of rules and program is called an expert system

Machine Learning Explained

- Humans takes decision based on experience, computer make models based on extracted patterns in the input data then data-filled models for predictions to take decision
- Quality of data determines the quality of experience which inturn gives the Quality Learning System



Statistical Learning

- The relationship between input x and output y is modeled as a function in the form y = f(x)
    - f – is the learning function that maps the input x to output y
    - It is simply called as mapping of input to output
- A model is an explicit description of patterns within the data in the form of
    - Mathematical equation
    - Relational diagrams like trees/graphs
    - Logical if/else rules or
    - Grouping called clusters

Another definition

- Tom Mitchell’s definition on ML
  - *A computer program is said to learn from experience E with respect to task T and some performance measure P improves with experience E*

- Models of computer systems are equivalent to human experience
- Experience is based on data
- Where as human gains experience by various means- they observe by trial and error
- Once knowledge is gained, when a new problem is encountered, humans search similar past situation and then formulate the heuristics and use that for prediction

ML Explained

- Experience is gathered by the following steps:
  - Collection of data
  - Abstract concepts are formed out of data
    - Abstraction is used to generate concepts(it is equivalent human’s idea of object)
  - Generalization converts the abstraction into an actionable form of intelligence
    - It can be viewed as ordering of all possible concepts
    - Generalization involves ranking of concepts, inferencing and forming heuristics,
  - Heuristics normally works

MACHINE LEARNING IN RELATION TO OTHER FIELDS

- Machine learning uses the concepts of Artificial Intelligence, Data Science, and Statistics primarily. It is the resultant of combined ideas of diverse fields.

Machine Learning and AI

- Machine learning is an important branch of AI, which is a much broader subject.
- The aim of AI is to develop intelligent agents.
  - An agent can be a robot, humans, or any autonomous systems.
  - Initially, the idea of AI was ambitious, that is, to develop intelligent systems like human beings.
  - The focus was on logic and logical inferences.
  - It had seen many ups and downs.
  - These down periods were called AI winters.

Machine Learning and AI

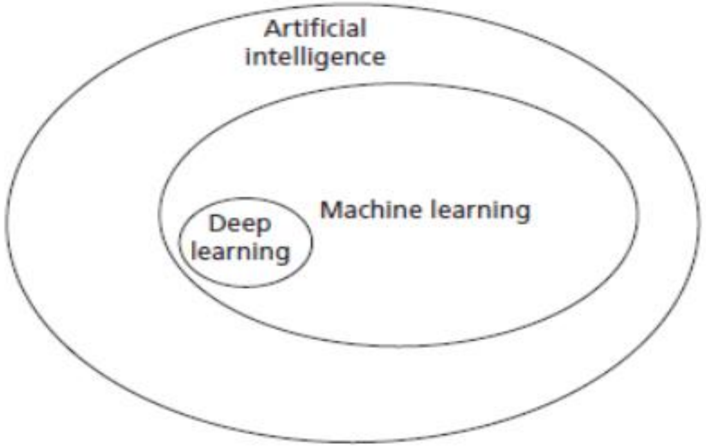


Figure 1.3: Relationship of AI with Machine Learning

- The popularity of AI stated again due to development of data driven systems.
- The aim is to find relations and regularities present in the data. Machine learning is the subbranch of AI, whose aim is to extract the patterns for prediction.
- It is a broad field that includes learning from examples and other areas like reinforcement learning.
- The relationship of AI and machine learning is shown in Figure 1.3. The model can take an unknown instance and generate results.
- Deep learning is a subbranch of machine learning. In deep learning, the models are constructed using neural network technology
- Neural networks are based on the human neuron models. Many neurons form a network connected with the activation functions that trigger further neurons to perform tasks.

Machine Learning, Data Science, Data Mining, and Data Analytics

- Data science is an 'Umbrella' term that encompasses many fields. Machine learning starts with data.
- Therefore, data science and machine learning are interlinked. Machine learning is a branch of data science.
- Data science deals with gathering of data for analysis.
- It is a broad field that includes:

Big Data

- Data science concerns about collection of data.
- Big data is a field of data science that deals with data's following characteristics:
  - Volume: Huge amount of data is generated by big companies like Facebook, Twitter, You Tube.
  - Variety: Data is available in variety of forms like images, videos, and in different formats.
  - Velocity: It refers to the speed at which the data is generated and processed.
- Big data is used by many machine learning algorithms for applications such as language translation and image recognition.
- Big data influences the growth of subjects like Deep learning.
- Deep learning is a branch of machine learning that deals with constructing models using neural networks.

Data Mining

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

Data Analytics

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

Pattern Recognition

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

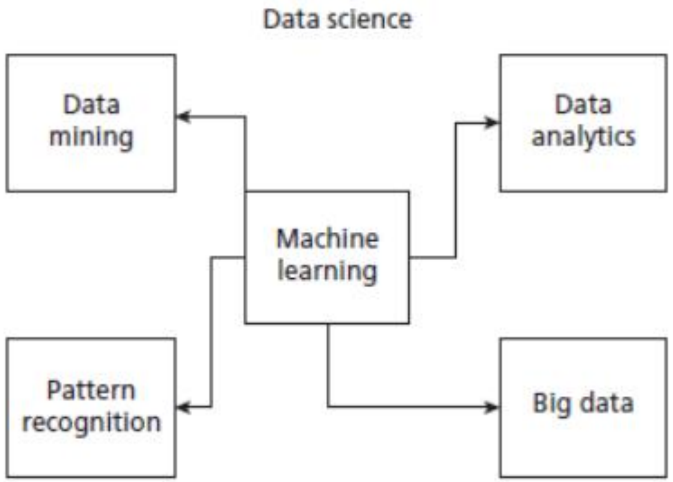


Figure 1.4: Relationship of Machine Learning with Other Major Fields

Machine Learning and Statistics

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

Machine Learning and Statistics

- It is mathematics intensive and models are often complicated equations and involve many assumptions.
- Statistical methods are developed in relation to the data being analysed. In addition, statistical methods are coherent and rigorous. It has strong theoretical foundations and interpretations that require a strong statistical knowledge.
- Machine learning, comparatively, has less assumptions and requires less statistical knowledge. But, it often requires interaction with various tools to automate the process of learning.
- Machine learning is just the latest version of'old Statistics' and hence this relationship should be recognized.

TYPES OF MACHINE LEARNING

- What does the word 'learn' mean?
- Learning, like adaptation, occurs as the result of interaction of the program with its environment.
- It can be compared with the interaction between a teacher and a student.
- There are four types of machine learning as shown in Figure 1.5.

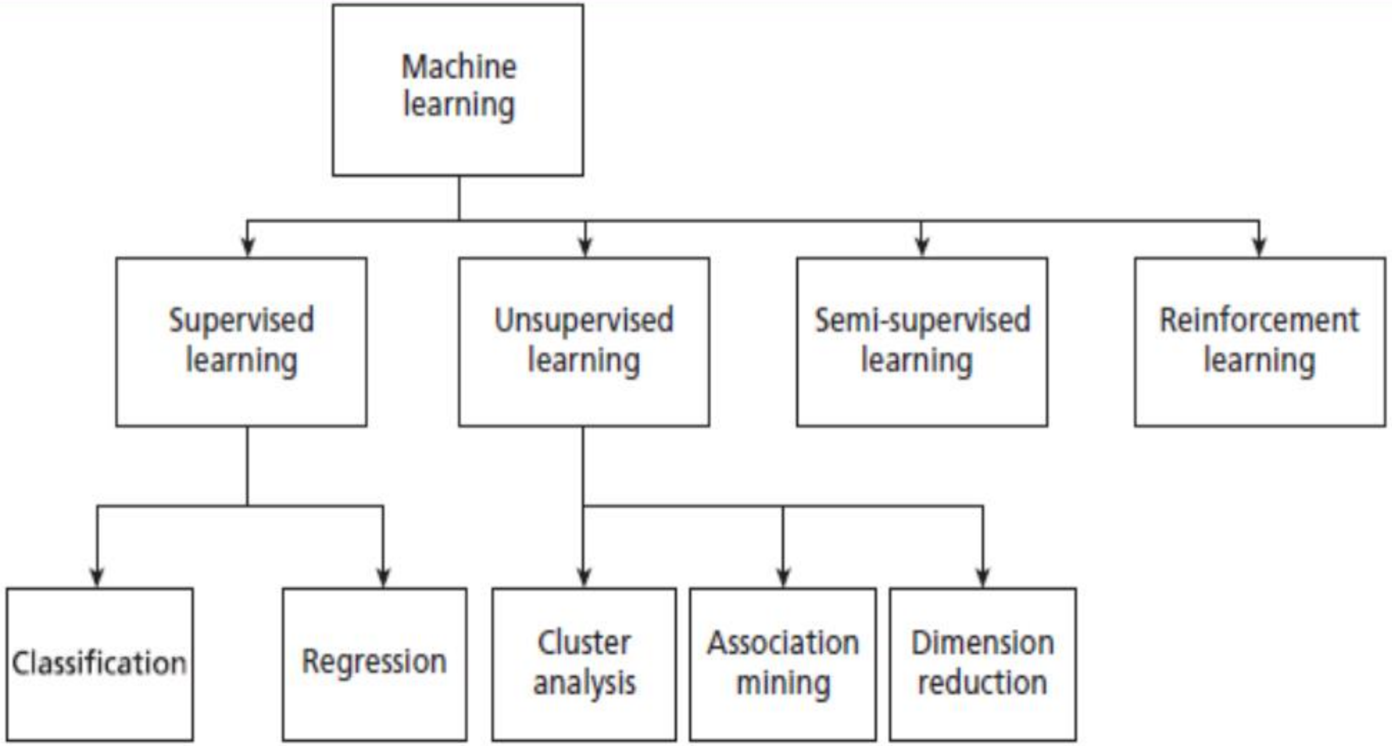


Figure 1.5: Types of Machine Learning

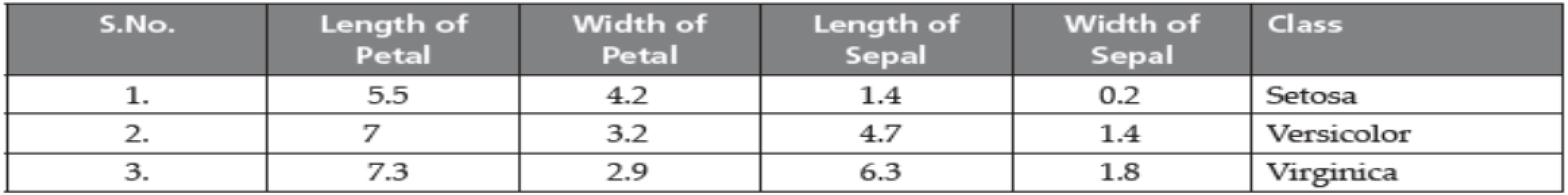
Types of ML

- Before learning types, it is necessary to discuss about data
- Labelled and Unlabelled Data - Data is a raw fact. Normally, data is represented in the form of a table. Data also can be referred to as a data point, sample, or an example.
- Each row of the table represents a data point. Features are attributes or characteristics of an object.
- Normally, the columns of the table are attributes. Out of all attributes, one attribute is important and is called a label. Label is the feature that we aim to predict. Thus, there are two types of data - labelled and unlabelled.

Types of ML

- Labelled Data To illustrate labelled data, let us take one example dataset called Iris flower dataset or Fisher's Iris dataset.
- The dataset has 50 samples of Iris - with four attributes, length and width of sepals and petals.
- The target variable is called class. There are three classes - Iris setosa, Iris virginica, and Iris versicolor.

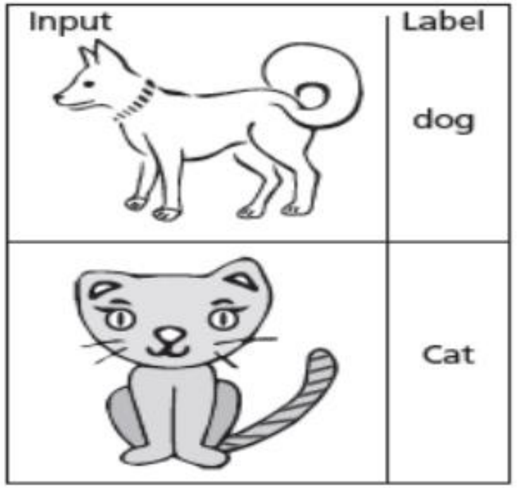
Table 1.1: Iris Flower Dataset



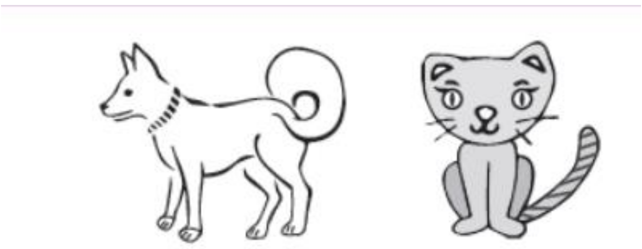
S.No.	Length of Petal	Width of Petal	Length of Sepal	Width of Sepal	Class
1.	5.5	4.2	1.4	0.2	Setosa
2.	7	3.2	4.7	1.4	Versicolor
3.	7.3	2.9	6.3	1.8	Virginica

Types of ML

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



a)Labelled Dataset



b)Unlabelled Dataset

Types of ML- Supervised

- Supervised algorithms use labelled dataset.
- As the name suggests, there is a supervisor or teacher component in supervised learning. A supervisor provides labelled data so that the model is constructed and generates test data.
- In supervised learning algorithms, learning takes place in two stages. In layman terms, during the first stage, the teacher communicates the information to the student that the student is supposed to master.
- The student receives the information and understands it. During this stage, the teacher has no knowledge of whether the information is grasped by the student.
- This leads to the second stage of learning. The teacher then asks the student a set of questions to find out how much information has been grasped by the student. Based on these questions, the student is tested, and the teacher informs the student about his assessment. This kind of learning is typically called supervised learning.

Types of ML

- Supervised has 2 methods:
  - Classification
  - Regression

- Classification
  - Classification is a supervised learning method.
  - The input attributes of the classification algorithms are called independent variables.
  - The target attribute is called label or dependent variable.
  - The relationship between the input and target variable is represented in the form of a structure which is called a classification model.
  - So, the focus of classification is to predict the 'label' that is in a discrete form (a value from the set of finite values).

Types of ML

- Classification
  - An example is shown in Figure 1.7 where a classification algorithm takes a set of labelled data images such as dogs and cats to construct a model that can later be used to classify an unknown test image data.

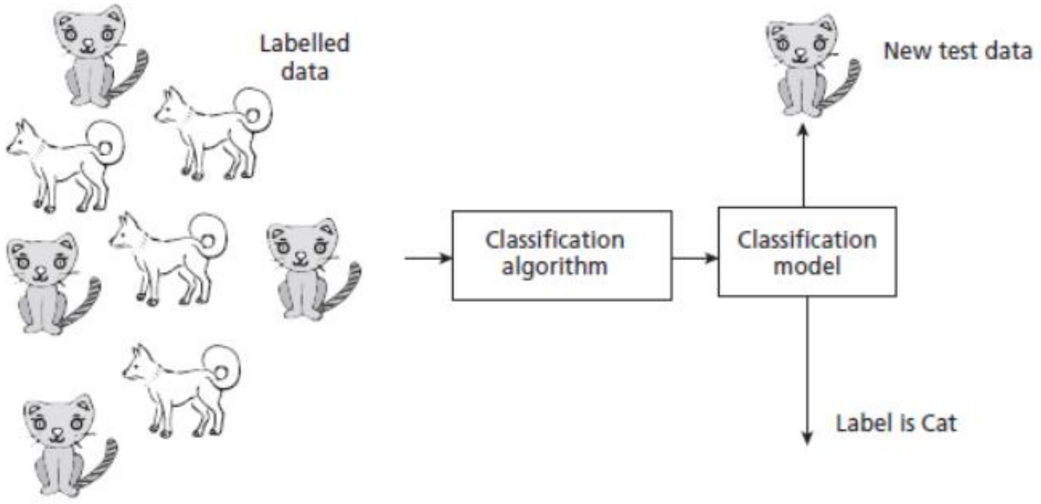


Figure 1.7: An Example Classification System

ML Types

- Classification
- In classification, learning takes place in two stages.
- During the first stage, called training stage, the learning algorithm takes a labelled dataset and starts learning.
- After the training set, samples are processed and the model is generated.
- In the second stage, the constructed model is tested with test or unknown sample and assigned a label. This is the classification process.

ML types

- Classification:
- The classification models can be categorized based on the implementation technology like decision trees, probabilistic methods, distance measures, and soft computing methods.
- Classification models can also be classified as generative models and discriminative models. Generative models deal with the process of data generation and its distribution. Probabilistic models are examples of generative models. Discriminative models do not care about the generation of data. Instead, they simply concentrate on classifying the given data.
- Some of the key algorithms of classification are:
  - Decision Tree, Random Forest, Support Vector Machines, Naive Bayes, Artificial Neural Network and Deep Learning networks like CNN

ML Types

- Regression
- Regression models, unlike classification algorithms, predict continuous variables like price. In other words, it is a number. A fitted regression model is shown in Figure 1.8 for a dataset that represent weeks input x and product sales y.

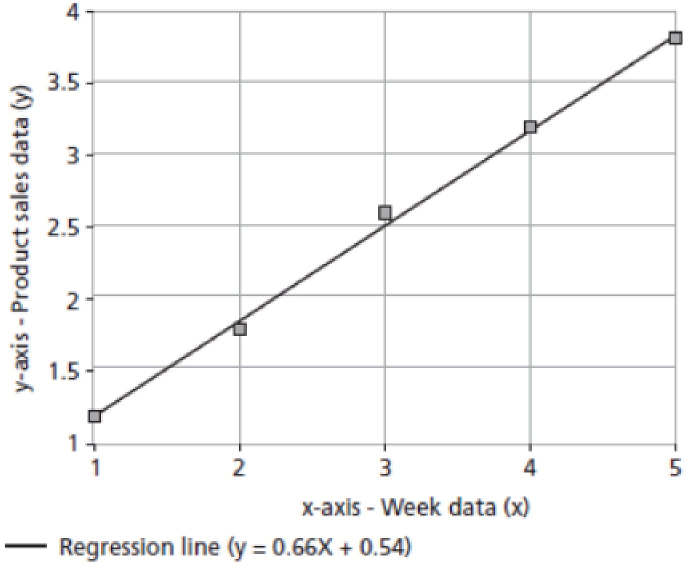


Figure 1.8: A Regression Model of the Form *y* = *ax* + *b*

- The regression model takes input x and generates a model in the form of a fitted line of the form y = f(x).
  - Here, x is the independent variable that may be one or more attributes and y is the dependent variable.
  - In Figure 1.8, linear regression takes the training set and tries to fit it with a line - product sales = 0.66 ? Week + 0.54. Here, 0.66 and 0.54 are all regression coefficients that are learnt from data.
  - The advantage of this model is that prediction for product sales (y) can be made for unknown week data (x). For example, the prediction for unknown eighth week can be made by substituting x as 8 in that regression formula to get y.

ML Types- regression

- One of the most important regression algorithms is linear regression that is explained in the next section.
- Both regression and classification models are supervised algorithms. Both have a supervisor and the concepts of training and testing are applicable to both.
- What is the difference between classification and regression models?
- The main difference is that regression models predict continuous variables such as product price, while classification concentrates on assigning discrete labels such as class.

Unsupervised Learning

- The second kind of learning is by self-instruction.
- As the name suggests, there are no supervisor or teacher components. In the absence of a supervisor or teacher, self-instruction is the most common kind of learning process.
- This process of self-instruction is based on the concept of trial and error. Here, the program is supplied with objects, but no labels are defined.
- The algorithm itself observes the examples and recognizes patterns based on the principles of grouping. Grouping is done in ways that similar objects form the same group.
- The program is supplied with objects, but no labels are defined. The algorithm itself observes the examples and recognizes patterns based on the principles of grouping. Grouping is done in ways that similar objects form the same group

Unsupervised Learning- Cluster Analysis

- Cluster analysis is an example of unsupervised learning.
- It aims to group objects into disjoint clusters or groups.
- Cluster analysis clusters objects based on its attributes.
- All the data objects of the partitions are similar in some aspect and vary from the data objects in the other partitions significantly.
- Some of the examples of clustering processes are - segmentation of a region of interest in an image, detection of abnormal growth in a medical image, and determining clusters of signatures in a gene database.

Unsupervised Learning- Cluster Analysis

- An example of clustering scheme is shown in Figure 1.9 where the clustering algorithm takes a set of dogs and cats images and groups it as two clusters-dogs and cats. It can be observed that the samples belonging to a cluster are similar and samples are different radically across clusters.

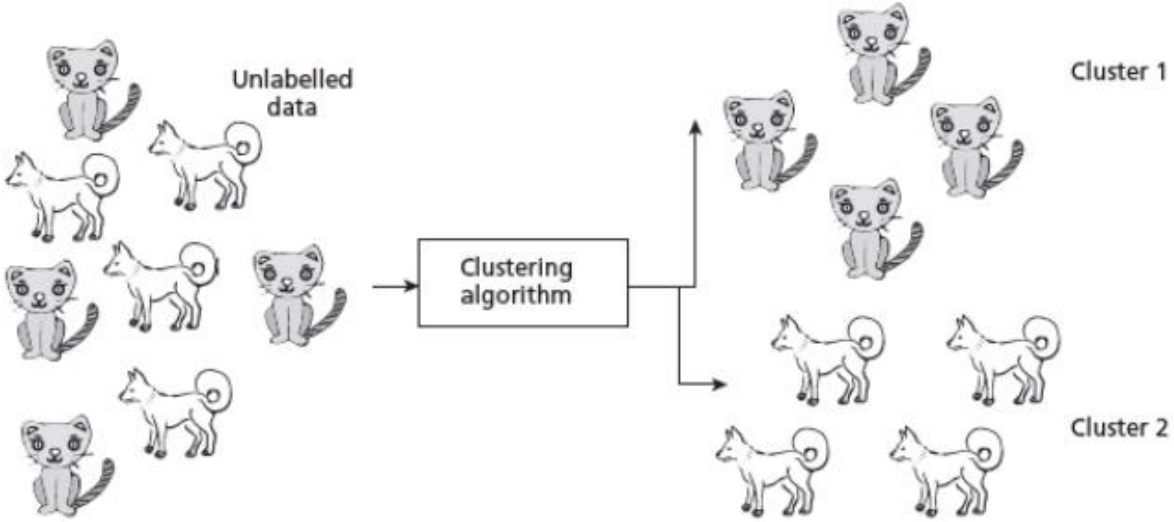
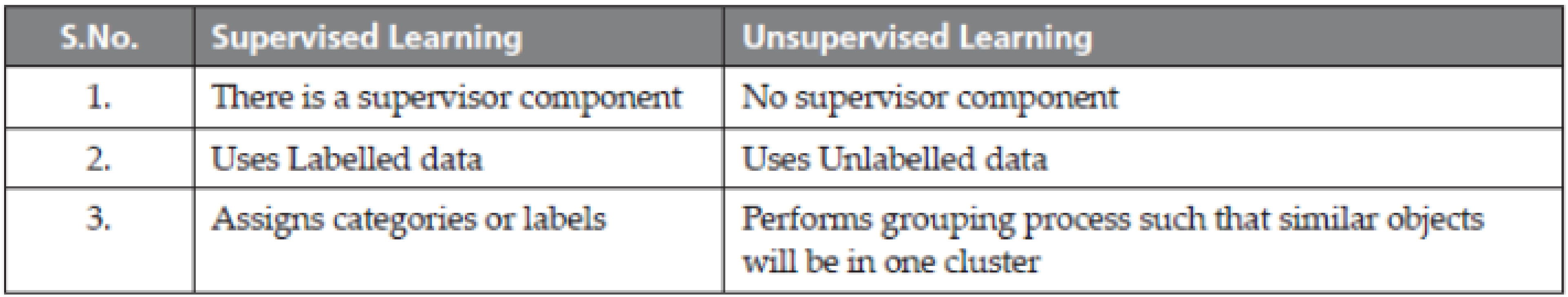


Figure 1.9: An Example Clustering Scheme

Unsupervised Learning- Dimensionality Reduction

Supervised Vs Unsupervised



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

Semi-supervised Learning

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

Reinforcement Learning

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

as shown in Figure 1.10.

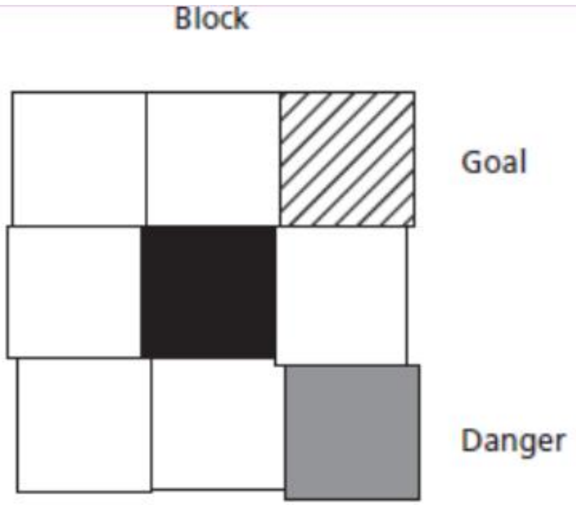


Figure 1.10: A Grid game

Grid Game Explanation

- In this grid game, the gray tile indicates the danger, black is a block, and the tile with diagonal lines is the goal. The aim is to start, say from bottom-left grid, using the actions left, right, top and bottom to reach the goal state.
- To solve this sort of problem, there is no data. The agent interacts with the environment to get experience, In the above case, the agent tries to create a model by simulating many paths and finding rewarding paths. This experience helps in constructing a model.

CHALLENGES OF MACHINE LEARNING

- Ill-posed problems – problems whose specifications are not clear
- Huge data- Availability of a quality data is a challenge.
- Huge computation power- With the availability of Big Data, the computational resource requirement has also increased.
- Complexity of algorithms- The selection of algorithms, describing the algorithms, application of algorithms to solve machine learning task, and comparison of algorithms have become necessary for machine learning or data scientists now
- Bias-variance- Variance is the error of the model. This leads to a problem called bias/ variance tradeoff. A model that fits the training data correctly but fails for test data, in general lacks generalization, is called overfitting.

MACHINE LEARNING PROCESS

- The emerging process model for the data mining solutions for business organizations is CRISP-DM. Since machine learning is like data mining, except for the aim, this process can be used for machine learning.
- This process involves six steps. The steps are listed below in Figure 1.11

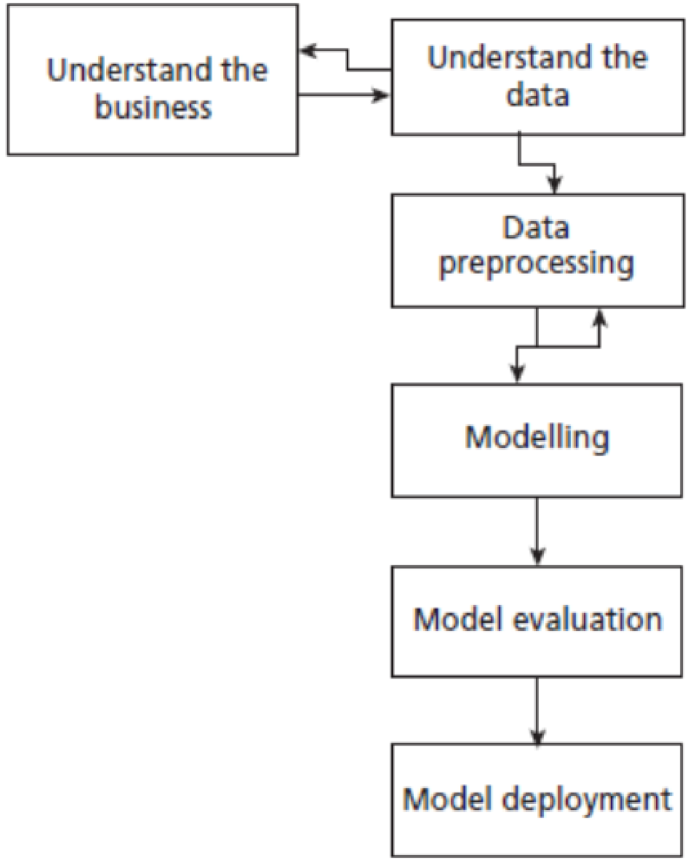


Figure 1.11: A Machine Learning/Data Mining Process

Steps

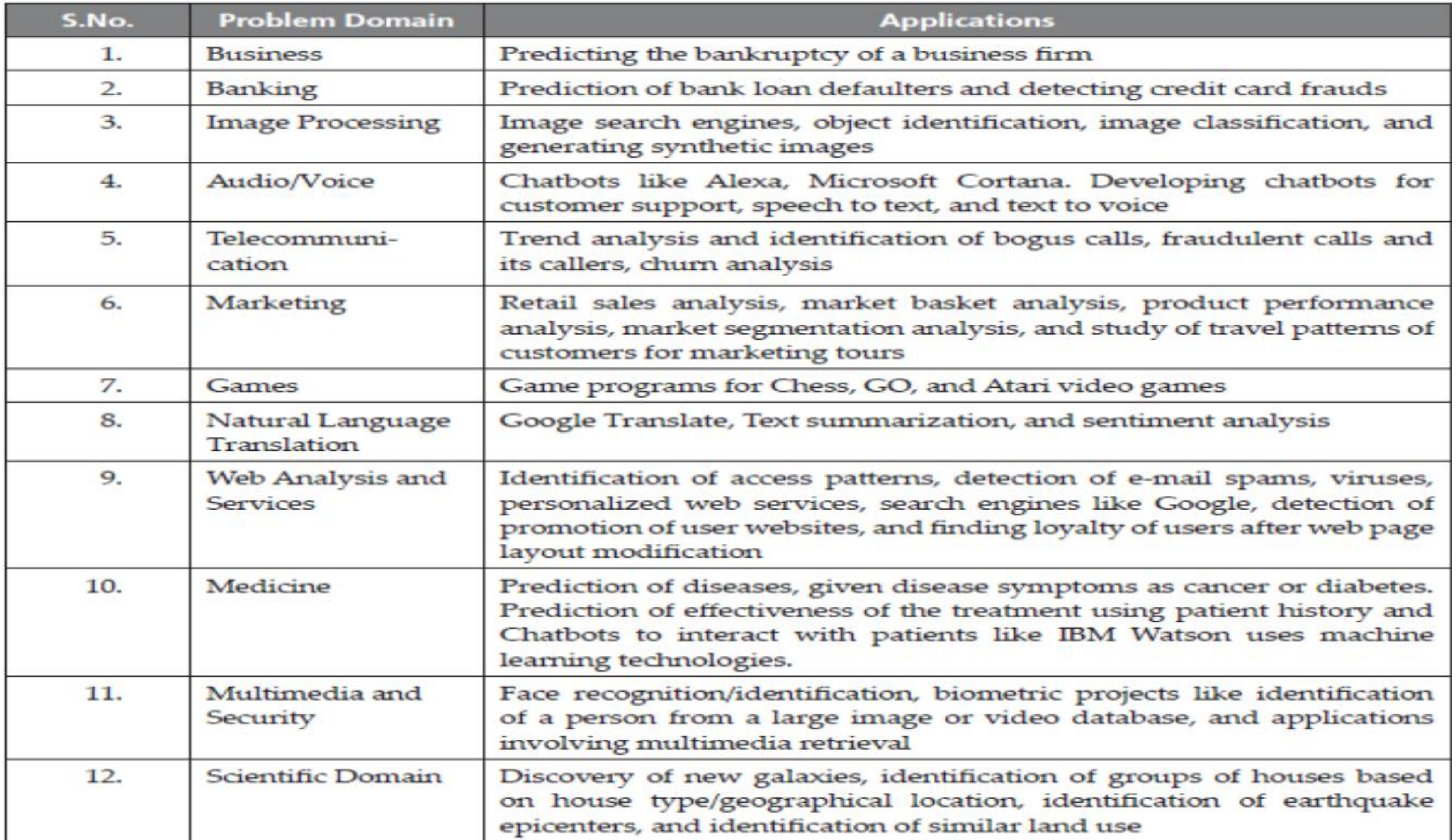
- Understanding the business - This step involves understanding the objectives and requirements of the business organization. Generally, a single data mining algorithm is enough for giving the solution. This step also involves the formulation of the problem statement for the data mining process.
- Understanding the data - It involves the steps like data collection, study of the characteristics of the data, formulation of hypothesis, and matching of patterns to the selected hypothesis.
- Preparation of data - This step involves producing the final dataset by cleaning the raw data and preparation of data for the data mining process. The missing values may cause problems during both training and testing phases. Missing data forces classifiers to produce inaccurate results. This is a perennial problem for the classification models. Hence, suitable strategies should be adopted to handle the missing data.

Steps

- Modelling - This step plays a role in the application of data mining algorithm for the data to obtain a model or pattern.
- Evaluate - This step involves the evaluation of the data mining results using statistical analysis and visualization methods. The performance of the classifier is determined by evaluating the accuracy of the classifier. The process of classification is a fuzzy issue. For example, classification of emails requires extensive domain knowledge and requires domain experts. Hence, performance of the classifier is very crucial.
- Deployment - This step involves the deployment of results of the data mining algorithm to improve the existing process or for a new situation.

MACHINE LEARNING APPLICATIONS

- Sentiment analysis
- Recommendation systems
- Voice assistants



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

Understanding Data

Module 2

What is data

- All facts are data
- In computer system, bits encode facts present in human interpretable (such as numbers or can be interpreted only by a computer.
- Today, growing amounts of data of the order of either 0 or 1. A kilo byte (KB) is 1024 bytes, one giga byte is approximately 1,000,000 KB, is one Exa byte.
- Data is available in different data sources like flat files, databases, or data warehouses.

What is data ?

- It can either be an operational data or a non-operational data.
  - Operational data is the one that is encountered in normal business procedures and processes. For example, daily sales data is operational data
  - Non-operational data is the kind of data that is used for decision making.
- Data by itself is meaningless.
- It has to be processed to generate any information.
  - A string of bytes is meaningless. Only when a label is attached like height of students of a class, the data becomes meaningful.
  - Processed data is called information that includes patterns, associations, or relationships among data. For example, sales data can be analyzed to extract information like which product was sold larger in the last quarter of the year.

Elements of Big Data

- Data whose volume is less and can be stored and processed by a small-scale computer is called 'small data’.
- These data are collected from several sources, and integrated and processed by a small-scale computer.
- Big data, on the other hand, is a larger data whose volume is much larger than 'small data' and is characterized as follows:
- Volume - Since there is a reduction in the cost of storing devices, there has been a tremendous growth of data.
  - Small traditional data is measured in terms of gigabytes (GB) and terabytes (TB)
  - Big Data is measured in terms of petabytes (PB) and exabytes (EB). One exabyte is 1 million terabytes.

Elements of data

- Velocity - The fast arrival speed of data and its increase in data volume is noted as velocity. The availability of IoT devices and Internet power ensures that the data is arriving at a faster rate. Velocity helps to understand the relative growth of big data and its accessibility by users, systems and applications.
- Variety - The variety of Big Data includes:
  - Form - There are many forms of data. Data types range from text, graph, audio, video, to maps. There can be composite data too, where one media can have many other sources of data, for example, a video can have an audio song.
  - Function - These are data from various sources like human conversations, transaction records, and old archive data.
  - Source of data - This is the third aspect of variety. There are many sources of data. Broadly, the data source can be classified as open/public data, social media data and multimodal data.

Elements of data

- Some of the other forms of Vs that are often quoted in the literature as characteristics of Big data are:
  - Veracity of data - Veracity of data deals with aspects like conformity to the facts, truthfulness, believability, and confidence in data. There may be many sources of error such as technical errors, typographical errors, and human errors. So, veracity is one of the most important aspects of data.
  - Validity - Validity is the accuracy of the data for taking decisions or for any other goals that are needed by the given problem.
  - Value - Value is the characteristic of big data that indicates the value of the information that is extracted from the data and its influence on the decisions that are taken based on it.

Types of Data

- Structured Data
- Unstructured Data
- Semi Structured Data

Structured Data

- In structured data, data is stored in an organized manner such as a database where it is available in the form of a table. The data can also be retrieved in an organized manner using tools like SQL.
- The structured data frequently encountered in machine learning are listed below:
  - Record Data- A dataset is a collection of measurements taken from a process. We have a collection of objects in a dataset and each object has a set of measurements. The measurements can be arranged in the form of a matrix.
  - Rows in the matrix represent an object and can be called as entities, cases, or records. The columns of the dataset are called attributes, features, or fields. The table is filled with observed data. Also, it is better to note the general jargons that are associated with the dataset. Label is the term that is used to describe the individual observations.

Structured Data

- The structured data frequently encountered in machine learning are listed below:
  - Data Matrix It is a variation of the record type because it consists of numeric attributes. The standard matrix operations can be applied on these data. The data is thought of as points or vectors in the multidimensional space where every attribute data is a dimension describing the objects
  - Graph data: It involves the relationships among objects.
    - Example: a web page can refer to another web page. This can be modeled as a graph i.e., nodes are web pages and hyperlinks as edge that connects the node
  - Ordered Data : Ordered data objects involve attributes that have an implicit order among them.
    - Examples of ordered data are:

Structured Data

- Ordered Data : Ordered data objects involve attributes that have an implicit order among them.
  - Examples of ordered data are:
  - Temporal data - It is the data whose attributes are associated with time. For example the customer purchasing patterns during festival time is sequential data. Time series data is a special type of sequence data where the data is a series of measurements over time.
  - Sequence data - It is like sequential data but does not have time stamps. This data involves the sequence of words or letters. For example, DNA data is a sequence of four characters -ATGC.
  - Spatial data - It has attributes such as positions or areas. For example, maps are spatial data where the points are related by location.

Unstructured Data

- Unstructured data includes video, images and audio
- It is also textual documents, programs and blog data
- It is estimated that 80% of the data are unstructured

Semi structured

- Semi-structured data are partially structured and partially unstructured.
- These include data like XML/JSON data, RSS feeds, and hierarchical data.

Data Storage and Representation

- Once the dataset is assembled, it must be stored in a structure that is suitable for data analysis.
- The goal of data storage management is to make data available for analysis.
- There are different approaches to organize and manage data in storage files and systems from flat file to data warehouses.

Flat Files

- Some of them are listed below:
  - Flat Files These are the simplest and most commonly available data source. It is also the cheapest way of organizing the data. These flat files are the files where data is stored in plain ASCII or EBCDIC format. Minor changes of data in flat files affect the results of the data mining algorithms. Hence, flat file is suitable only for storing small dataset and not desirable if the dataset becomes larger.
- Some of the popular spreadsheet formats are listed below:
  - CSV files – comma separated files where the values are separated by commas. These are used by spreadsheets and database applications- first row may have attributes and rest of the rows represent data
  - TSV files – tab separated files where the values are separated by tab

Both CSV AND TSV are generic in nature and can be shared, there are many tools like google sheets and Microsoft excel to process these files

Database System

- Normally consists of database files and a database management system(DBMS)
    - Database files contain original data and metadata.
    - DBMS aims to manage data and improve operator performance by including various tools like database administrator, query processing, and transaction manager.
    - A relational database consists of sets of tables. The tables have rows and columns.
    - The columns represent the attributes and rows represent tuples.
    - A tuple corresponds to either an object or a relationship between objects.
    - A user can access and manipulate the data in the database using SQL.

Types of Database

- Transactional Database:
  - A transactional database is a collection of transactional records.
  - Each record is a transaction.
  - A transaction may have a time stamp, identifier and a set of items, which may have links to other tables.
  - Normally, transactional databases are created for performing associational analysis that indicates the correlation among the items.
- Time Series Database:
  - Time-series database stores time related information like log files where data is associated with a time stamp.
  - This data represents the sequences of data, which represent values or events obtained over a period (for example, hourly, weekly or yearly) or repeated time span.
  - Observing sales of product continuously may yield a time-series data.

Types of Database

- Spatial Database:
  - Spatial are either databases bitmaps contain or pixel spatial maps. Information
  - For example, in a images can be stored as a raster data
  - On other hand, the vector format can be used to store maps use basiv=c geometric primitives like points, lines, polygons and so on.

Types of data storage and representation

- World wide web (WWW)- It provides a diverse, worldwide online information source. The objective of data mining algorithms is to mine interesting patterns of information present in WWW.
- XML(eXtensible Markup Language)- It is both human and machine interpretable data format that can be used to represent data that needs to be shared across the platforms.
- Data Stream - It is dynamic data, which flows in and out of the observing environment. Typical characteristics of data stream are huge volume of data, dynamic, fixed order movement, and real-time constraints.
- RSS(Really Simple Syndication)- It is a format for sharing instant feeds across services.
- JSON(JavaScript Object Notation)- It is another useful data interchange format that is often used for many machine learning algorithms.

Big data analytics and types of analytics

- The primary aim of data analysis is to assist business organizations to take decisions.
    - For example a business organization may want to know which is the fastest selling product, in order for them to market activities.
- Data analysis is an activity that takes the data and generates useful information and insights for assisting the organizations.
- Data analysis and data analytics are terms that are used interchangeably to refer to the same concept. However, there is a subtle difference.
- Data analytics is a general term and data analysis is a part of it.
    - Data analytics refers to the process of data collection, preprocessing and analysis. It deals with the complete cycle of data management.
    - Data analysis is just analysis and is a part of data analytics. It takes historical data and does the analysis. Data analytics, instead, concentrates more on future and helps in prediction.

Types of data analytics

- Descriptive analytics
- Diagnostic analytics
- Predictive analytics
- Prescriptive analytics

Types of Data Analytics

- Descriptive Analytics
  - It is about describing the main features of the data. After data collection is done, descriptive analytics deals with the collected data and quantifies it. It is often stated that analytics is essentially statistics.
  - There are two aspects of statistics - Descriptive and Inference
  - Descriptive analytics only focuses on the description part of the data and not the inference part.
- Diagnostic Analytics
  - It deals with the question - 'Why?’.
  - This is also known as causal analysis, as it aims to find out the cause and effect of the events.
  - For example, if a product is not selling, diagnostic analytics aims to find out the reason. There may be multiple reasons and associated effects are analyzed as part of it.

Types of data analytics

- Predictive Analytics
  - It deals with the future. It deals with the question - 'What will happen in future given this data?’,
  - This involves the application of algorithms to identify the patterns t predict the future.
  - The entire course of machine learning is mostly about predictive analytics
- Prescriptive Analytics
  - It is about the finding the best course of action for the business organizations.
  - Prescriptive analytics goes beyond prediction and helps in decision making by giving a set of actions.
  - It helps the organizations to plan better for the future and to mitigate the risks that are involved.

Big data analysis framework

- For performing data analytics, many frameworks are proposed. All proposed analytics frameworks have some common factors. Big data framework is a layered architecture. Such an architecture has many advantages such as genericness.
- A 4-layer architecture has the following layers:
  - Data connection layer
  - Data management layer
  - Data analytics later
  - Presentation layer

Big data analysis framework

- Data Connection Layer-
  - It has data ingestion mechanisms and data connectors. Data ingestion means taking raw data and importing it into appropriate data structures.
  - It performs the tasks of ETL process. By ETL, it means extract, transform and load operations.
- Data Management Layer-
  - It performs preprocessing of data.
  - The purpose of this layer is to allow parallel execution of queries, and read, write and data management tasks.
  - There may be many schemes that can be implemented by this layer such as data-in-place, where the data is not moved at all, or constructing data repositories such as data warehouses and pull data on-demand mechanisms.

Big data analysis framework

- Data Analytic Layer
  - It has many functionalities such as statistical tests, machine learning algorithms to understand, and construction of machine learning models.
  - This layer implements many model validation mechanisms too.
  - Types of processing
  - Cloud computing
  - Grid Computing
  - H-computing (high performance computing or HPC)

Cloud computing

- Cloud computing is an emerging technology which is basically a business service model or simply called as pay-per-use model
- Cloud refers to the internet that provides sharing of processing power, applications, storage, and services
- It offers different services such as Iaas, Paas, Saas
    - SAAS(software as a Service)-enables user to access software applications from the cloud
    - PAAS(platform as a Service)- provides user platform to develop and run their applications
    - IAAS(Infrastructure as a Service)- enables users to access the infrastructure required to run their applications, storage, OS etc.

Cloud Computing

- The cloud services can be deployed in four most commonly used deployment model
  - Such as Public Cloud, Private Cloud, Community Cloud, and Hybrid Cloud based on the service model, organization, geographic location, etc.
- The Public Cloud is accessible to the public and is owned by a vendor, who offers the services of the cloud to the users.
- Private Cloud is ; privately-owned cloud where the user or an organization owns the cloud and only the user or employees of that organization have access to the cloud, thereby making data and transactions secure.
- In Community Cloud, the infrastructure is owned jointly by different organizations
- The Hybrid Cloud is the combination of two or more cloud types.

Characteristics of cloud computing

- The characteristics of cloud computing are:
  - Shared Infrastructure - Sharing of physical services, storage, and networking capabilities
  - Dynamic Provisioning - Resources assigned dynamically, based on demands
  - Dynamic Scaling - Expansion and contraction of service capability
  - Network Access - Needs to be accessed across the internet
  - Utility-based Metering - Uses metering to provide reporting and billing information
  - Multitenancy - Serves multiple customers
  - Reliability - Customer reliable service

Grid computing

- Grid Computing is a parallel and distributed computing framework consisting of a network of computers offering a super computing service as a single virtual supercomputer.
- This high-performance computing is required to perform specialized tasks that require a high computing power and a single computer cannot provide enough computing resources.
- The grid computing model forms a grid by connecting tens of thousands of nodes as a cluster that runs on an operating system.
- In this model, the resources are pooled together and the load is shared across multiple nodes to accomplish a task more quickly.
- This grid is constructed by middleware software that evenly distributes the task to several nodes connected in the grid.
- The individual nodes perform the task independently and in parallel which are then integrated to complete the large-scale task.
- This model of computing is best suited for applications that are complex and can be computed in parallel.

H-Computing (High Performance Computing or HPC)

- It enables to perform complex tasks at high speed.
- It aggregates computing power in such a way that provides much higher performance to solve complex problems in science, engineering, research or business.
- It leverages parallel processing techniques for solving complex computational problems.
- HPC system achieves this sustained performance through concurrent use of computing resources.
- An HPC system combines the computing power of thousands of compute nodes that work in parallel to complete tasks faster.
- The system comprises three key components called compute, network and storage.
- The architecture of HPC consists of compute servers that are networked together to form a cluster.
- Software programs are run in parallel on the servers in the cluster and are networked to the data storage to capture the output.
- These components work together to complete a task.

Big data analysis framework

- Presentation Layer
  - It has mechanisms such as dashboards, and applications that display the results of analytical engines and machine learning algorithms.
  - Thus, the Big Data processing cycle involves data management that consists of the following steps.
    - Data collection
    - Data preprocessing
    - Applications of machine learning algorithm
    - Interpretation of results and visualization of machine learning algorithm
  - This is an iterative process and is carried out on a permanent basis to ensure that data is suitable for data mining.

Data Collection

- The first task of gathering datasets are the collection of data.
- It is often estimated that most of the time is spent for collection of good quality data.
- A good quality data yields a better result.
- It is often difficult to characterize a 'Good data'. 'Good data' is one that has the following properties:
  - Timeliness - The data should be relevant and not stale or obsolete data.
  - Relevancy - The data should be relevant and ready for the machine learning or data mining algorithms.
  - All the necessary information should be available and there should be no bias in the data.
  - Knowledge about the data - The data should be understandable and interpretable, and should be self-sufficient for the required application as desired by the domain knowledge engineer.

Data Sources

- Broadly, the data source can be classified as open/public data, social media data and multimodal data.
- Open or public data source - It is a data source that does not have any stringent copyright rules or restrictions. Its data can be primarily used for many purposes. Government census data are good examples of open data:
  - Digital libraries that have huge amount of text data as well as document images
  - Scientific domains with a huge collection of experimental data like genomic data and biological data
  - Healthcare systems that use extensive databases like patient databases, health insurance data, doctors' information, and bioinformatics information
- Social media - It is the data that is generated by various social media platforms like Twitter, Facebook, YouTube, and Instagram. An enormous amount of data is generated by these platforms.
- Multimodal data - It includes data that involves many modes such as text, video, audio and mixed types. Some of them are listed below:
  - Image archives contain larger image databases along with numeric and text data
  - The World Wide Web (WWW) has huge amount of data that is distributed on the Internet

  These data are heterogeneous in nature.

Attendance 22-1-24

- Kiran K M
- Srujan A
- Nishwamm
- Kiran U
- Dalio
- Yuvalakshmi
- Tejaswini y k
- Afia fareen

Data Preprocessing

- In real world, the available data is 'dirty'. By this word 'dirty', it means:
    - Incomplete data
    - Inaccurate data
    - Outlier data
    - Data with missing values
    - Data with inconsistent values
    - Duplicate data
- Data preprocessing improves the quality of the data mining techniques.
- The raw data must be preprocessed to give accurate results.

Data Preprocessing

- The process of detection and removal of errors in data is called data cleaning.
- Data wrangling means making the data processable for machine learning algorithms.
- Some of the data errors include human errors such as typographical errors or incorrect measurement and structural errors like improper data formats.
- Data errors can also arise from omission and duplication of attributes.
- Noise is a random component and involves distortion of a value or introduction of spurious objects.
- Often, the noise is used if the data is a spatial or temporal component. Certain deterministic distortions in the form of a streak are known as artifacts.

Data Preprocessing

- Consider, for example, the following patient Table 2.1. The 'bad' or 'dirty' data can be observed in this table.

Patient ID	Name	Age	Date of Birth	Fever	Salary
1	John	21		low	-1500
2	Andre	36		High	Yes
3	David	5	10/10/1980	Low	“ “
4	Raju	132		High	Yes

- It can be observed that data like Salary = ' ' is incomplete data. The DoB of patients, John, Andre, and Raju, is the missing data. The age of David is recorded as '5' but his DoB indicates it is 10/10/1980. This is called inconsistent data.

Data Preprocessing

- Inconsistent data occurs due to problems in conversions, inconsistent formats, and difference in units.
- The Salary for John is - 1500 it cannot be less than 0 – instance of noisy data
- Outliers are the data exhibit the characteristics that are different from other data and have unusual values
  - The age of Raju cannot be 136- it might be a typographical.
  - It is required to differentiate between Noise and Outliers
- These Outliers may be legitimate data and sometimes are of interest to the data mining algorithms
  - These errors often come during data collection stage
  - These must be removed so that machine learning algorithms yields better results as the quality as the quality of results is determined by the quality input data. This removal process is called data cleaning.

Missing Data Analysis

- The primary data cleaning process is missing data analysis.
- Data cleaning routines attempt to fill up the missing values, smoothen the noise while identifying the outliers and correct the inconsistencies of the data.
- This enables data mining to avoid overfitting of the models.
- The procedures that are given below can solve the problem of missing data:
  - Ignore the tuple - A tuple with missing data, especially the class label, is ignored. This method is not effective when the percentage of the missing values increases.
  - Fill in the values manually - Here, the domain expert can analyse the data tables and carry out the analysis and fill in the values manually. But, this is time consuming and may not be feasible for larger sets.
  - A global constant can be used to fill in the missing attributes. The missing values may be 'Unknown' or be 'Infinity'. But, some data mining results may give spurious results by analysing these labels.
  - The attribute value may be filled by the attribute average. Say, the average income can replace a missing value.
  - Use the attribute mean for all samples belonging to the same class. Here, the average value replaces the missing values of all tuples that fall in this group.
  - Use the most possible value to fill in the missing value. The most probable value can be obtained from other methods like classification and decision tree prediction.
- Some of these methods introduce bias in the data. The filled value may not be correct and could be just an estimated value. Hence, the difference between the estimated and the original value is called an error or bias.

Removal of Noisy or Outlier Data

- Noise is a random error or variance in a measured value. It can be removed by using binning, which is a method where the given data values are sorted and distributed into equal frequency bins. The bins are also called as buckets. The binning method then uses the neighbor values to smooth the noisy data. Some of the techniques commonly used are 'smoothing by means' where the mean of the bin removes the values of the bins, 'smoothing by bin medians' where the bin median replaces the bin values, and 'smoothing by bin boundaries' where the bin value is replaced by the closest bin boundary. The maximum and minimum values are called bin boundaries. Binning methods may be used as a discretization technique.

Removal of Noisy or Outlier Data- Example

- Consider the following set: S = {12, 14, 19, 22, 24, 26, 28, 31, 32}
- Apply various binning techniques and show the result.
- Solution: By equal-frequency bin method, the data should be distributed across bins
- Let us assume bins of size 3, then the above data is distributed across the bins as shown below:

Bin 1	12,14,19
Bin 2	22,24,26
Bin 3	28, 31,32

Bin Mean value

Bin 1	15,15,15
Bin 2	22,24,24
Bin 3	30.3, 30.3, 30.3

Smoothing by bin boundaries Method

Bin 1	12,12,19
Bin 2	22,22,26
Bin 3	28, 32,32

- As per the method, the minimum and maximum values of the bin are determined, and it serves as bin boundary and does not change. Rest of the values are transformed to the nearest value. It can be observed in Bin 1, the middle value 14 is compared with the boundary values 12 and 19 and changed to the closest value, that is 12. This process is repeated for all bins.

Data Integration and Data Transformations

- Data integration involves routines that merge data from multiple sources into a single data source. So, this may lead to redundant data.
- The main goal of data integration is to detect and remove redundancies that arise from integration.
- Data transformation routines perform operations like normalization to improve the performance of the data mining algorithms.
- It is necessary to transform data so that it can be processed. This can be considered as a preliminary stage of data conditioning.
- Normalization is one such technique. In normalization, the attribute values are scaled to fit in a range (say 0-1) to improve the performance of the data mining algorithm.
- Often, in neural networks, these techniques are used. Some of the normalization procedures used are:
    - Min-Max
    - Z-Score

Min-Max Procedure

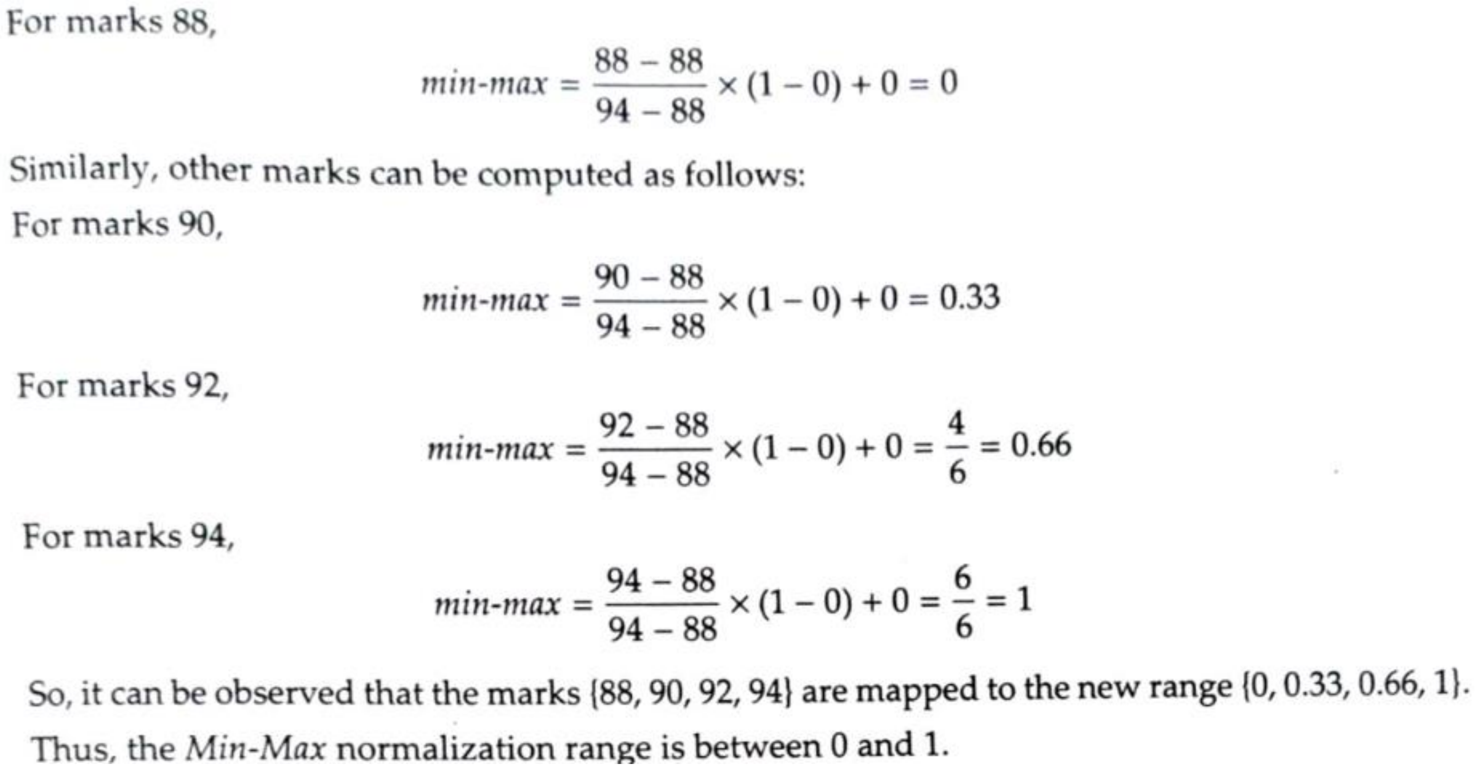
- It is a normalization technique where each variable V is normalized by its difference with the minimum value divided by the range to a new range, say 0-1. Often, neural networks require this kind of normalization. The formula to implement this normalization is given as:

min-max =$\frac{V-min}{max-min} * (new \max - new\ min) + new\ min$

- Here max-min is the range. Min and max are the minimum and maximum of the given data, new max and new min are the minimum and maximum of the target range, say 0 and 1.

Min-Max Example

- Consider the set: V = (88, 90, 92, 94). Apply Min-Max procedure and map the marks to a new range 0-1.
- Solution: The minimum of the list V is 88 and maximum is 94. The new min and new max are 0 and 1, respectively. The mapping can be done using Equation as:



For marks 88,

$$\textit{min-max} = \frac{88-88}{94-88} \times (1-0) + 0 = 0$$

Similarly, other marks can be computed as follows:

For marks 90,

$$\textit{min-max} = \frac{90-88}{94-88} \times (1-0) + 0 = 0.33$$

For marks 92,

$$\textit{min-max} = \frac{92-88}{94-88} \times (1-0) + 0 = \frac{4}{6} = 0.66$$

For marks 94,

$$\textit{min-max} = \frac{94-88}{94-88} \times (1-0) + 0 = \frac{6}{6} = 1$$

So, it can be observed that the marks {88, 90, 92, 94} are mapped to the new range {0, 0.33, 0.66, 1}.

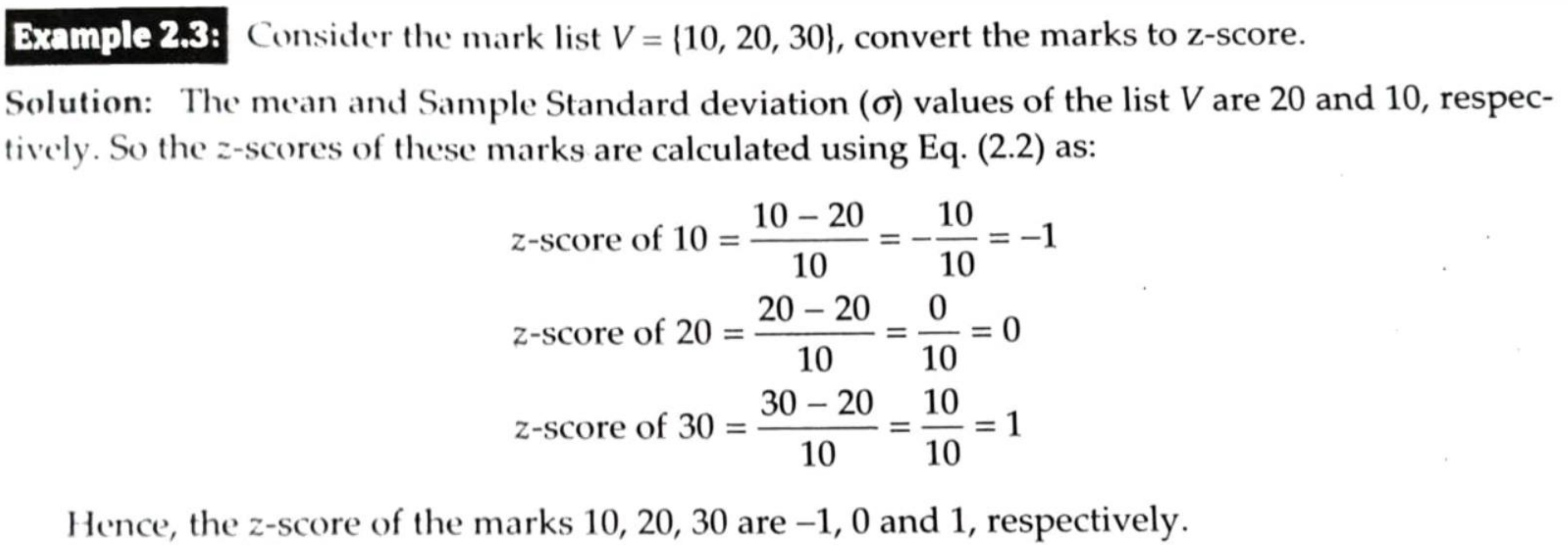
Thus, the *Min-Max* normalization range is between 0 and 1.

z- Score

- Z-Score Normalization procedure works by taking the difference between the field value and mean value, and by scaling this difference by standard deviation of the attribute.

$$V* = V \; - \; \mu/\sigma$$

- Here, σ is the standard deviation of the list V and µ is the mean of the list V.



Example 2.3: Consider the mark list $V = \{10, 20, 30\}$, convert the marks to z-score.

Solution: The mean and Sample Standard deviation (σ) values of the list V are 20 and 10, respectively. So the z-scores of these marks are calculated using Eq. (2.2) as:

$$\text{z-score of } 10 = \frac{10-20}{10} = -\frac{10}{10} = -1$$

$$\text{z-score of } 20 = \frac{20-20}{10} = \frac{0}{10} = 0$$

$$\text{z-score of } 30 = \frac{30-20}{10} = \frac{10}{10} = 1$$

Hence, the z-score of the marks 10, 20, 30 are –1, 0 and 1, respectively.

Data reduction

- Data reduction reduces data size but produces the same results. There are different ways in which data reduction can be carried out such as data aggregation, feature selection, and dimensionality reduction.

Attendance – 5/2/24

- Swetha
- Chethana
- Yuva Lakshmi
- Lokitha
- Keerthana k
- Sumathi
- Sumalatha
- Keerthi k
- Afia
- Tejaswini y k
- Harshitha b
- Sonal K J
- Ragashree g

- Nishwal M M
- Karthik Kumar A
- Yogesh M
- Kiran K M
- Surjan
- Gangadhar K S
- Gagan Raj
- Prasanna

Descriptive Statistics

- Descriptive statistics is a branch of statistics that does dataset summarization.
- It is used to summarize and describe data.
- Descriptive statistics are just descriptive and do not go beyond that
- In other words, descriptive statistics do not bother too much about machine learning algorithms and its functioning.
- Data visualization is a branch of study that is useful for investigating the given data. Mainly the plots are useful to explain and present data to customers.

Descriptive Statistics

- Descriptive analytics and data visualization techniques help to understand the nature of the data, which further helps to determine the kinds of machine learning or data mining tasks that can be applied to the data.
- This step is often known as Exploratory Data Analysis (EDA).
- The focus of EDA is to understand the given data and to prepare it for machine learning algorithms.
- EDA includes descriptive statistics and data visualization.

Dataset and Data Types

- A dataset can be assumed to be a collection of data objects.
- The data objects may be records, points, vectors, patterns, events, cases, samples or observations.
- These records contain many attributes.
- An attribute can be defined as the property or characteristics of an object.
- Every attribute should be associated with a value.
- This process is called measurement.
- The type of attribute determines the data types, often referred to as measurement scale types.

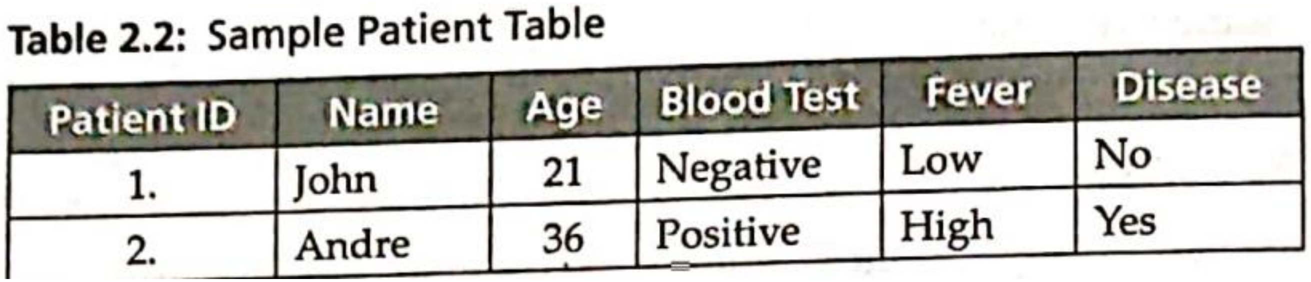
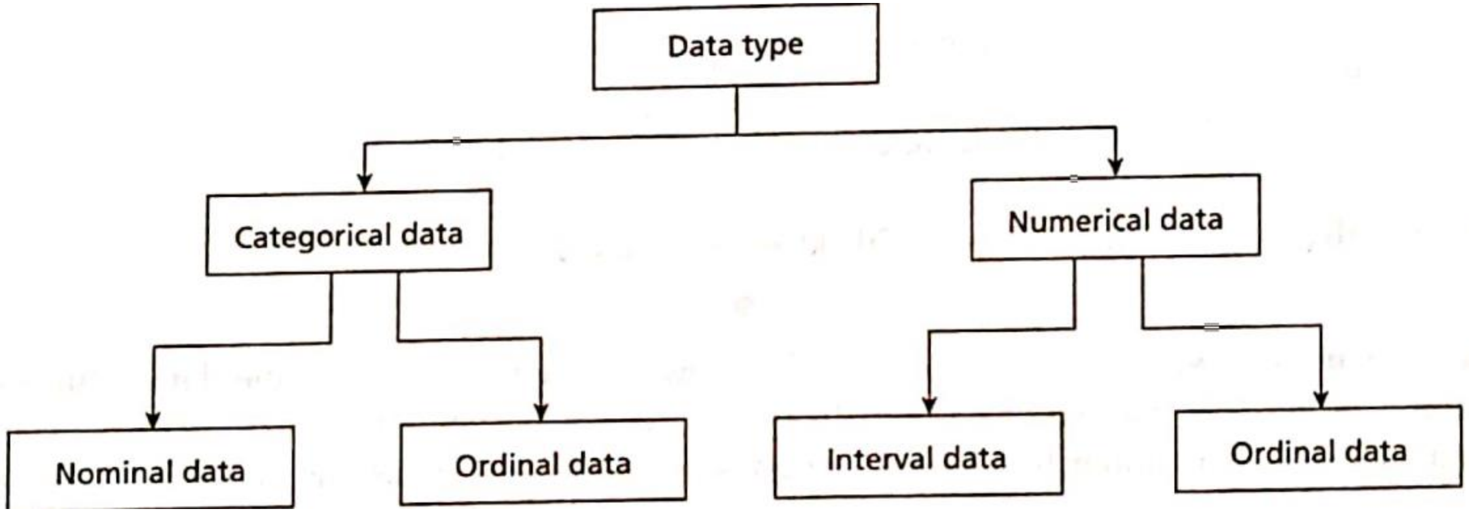


Table 2.2: Sample Patient Table

Patient ID	Name	Age	Blood Test	Fever	Disease
1.	John	21	Negative	Low	No
2.	Andre	36	Positive	High	Yes

Data Types



- **Categorical or Qualitative Data**
  - The categorical data can be divided into two types. They are nominal type and ordinal type.
  - Nominal Data - In Table 2.2, patient ID is nominal data.
  - Nominal data are symbols and cannot be processed like a number.
  - For example, the average of a patient ID does not make any statistical sense.
  - Nominal data type provides only information but has no ordering among data.
  - Only operations like (=, #) are meaningful for these data. For example, the patient ID can be checked for equality and nothing else..
  - Ordinal Data - It provides enough information and has natural order.
  - For example, Fever= {Low, Medium, High) is an ordinal data. Certainly, low is less than medium and medium is less than high, irrespective of the value. Any transformation can be applied to these data to get a new value.

Data Types

- **Numeric or Qualitative**
  - Data It can be divided into two categories. They are interval type and ratio type.
  - Interval Data - Interval data is a numeric data for which the differences between values are meaningful. For example, there is a difference between 30 degree and 40 degree. Only the permissible operations are + and -.
  - Ratio Data - For ratio data, both differences and ratio are meaningful. The difference between the ratio and interval data is the position of zero in the scale. For example, take the Centigrade-Fahrenheit conversion. The zeroes of both scales do not match. Hence, these are interval data.

Data types

- Another way of classifying the data is to classify it as:
    - Discrete value data
    - Continuous data
- Discrete Data
    - This kind of data is recorded as integers. For example, the responses of the survey can be discrete data. Employee identification number such as 10001 is discrete data.
- Continuous Data
    - It can be fitted into a range and includes decimal point. For example, age is a continuous data. Though age appears to be discrete data, one may be 12.5 years old and it makes sense. Patient height and weight are all continuous data.

Data Types

- Third way of classifying the data is based on the number of variables used in the dataset.

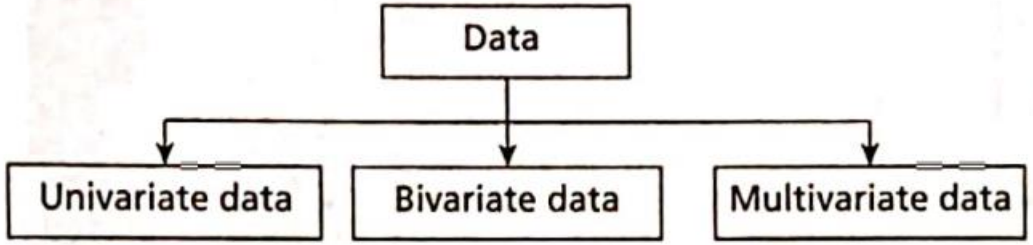


Figure 2.2: Types of Data Based on Variables

- In case of univariate data, the dataset has only one variable.

A variable is also called as category.

- Bivariate data indicates that the number of variables used are two
- Multivariate data uses three or more variables.

Univariate Data Analysis And Visualization

- Univariate analysis is the simplest form of statistical analysis.
- As the name indicates, the dataset has only one variable.
- A variable can be called as a category.
- Univariate does not deal with cause or relationships.
- The aim of univariate analysis is to describe data and find patterns.
- Univariate data description involves finding the frequency distributions, central tendency measures, dispersion or variation, and shape of the data.

Data Visualization

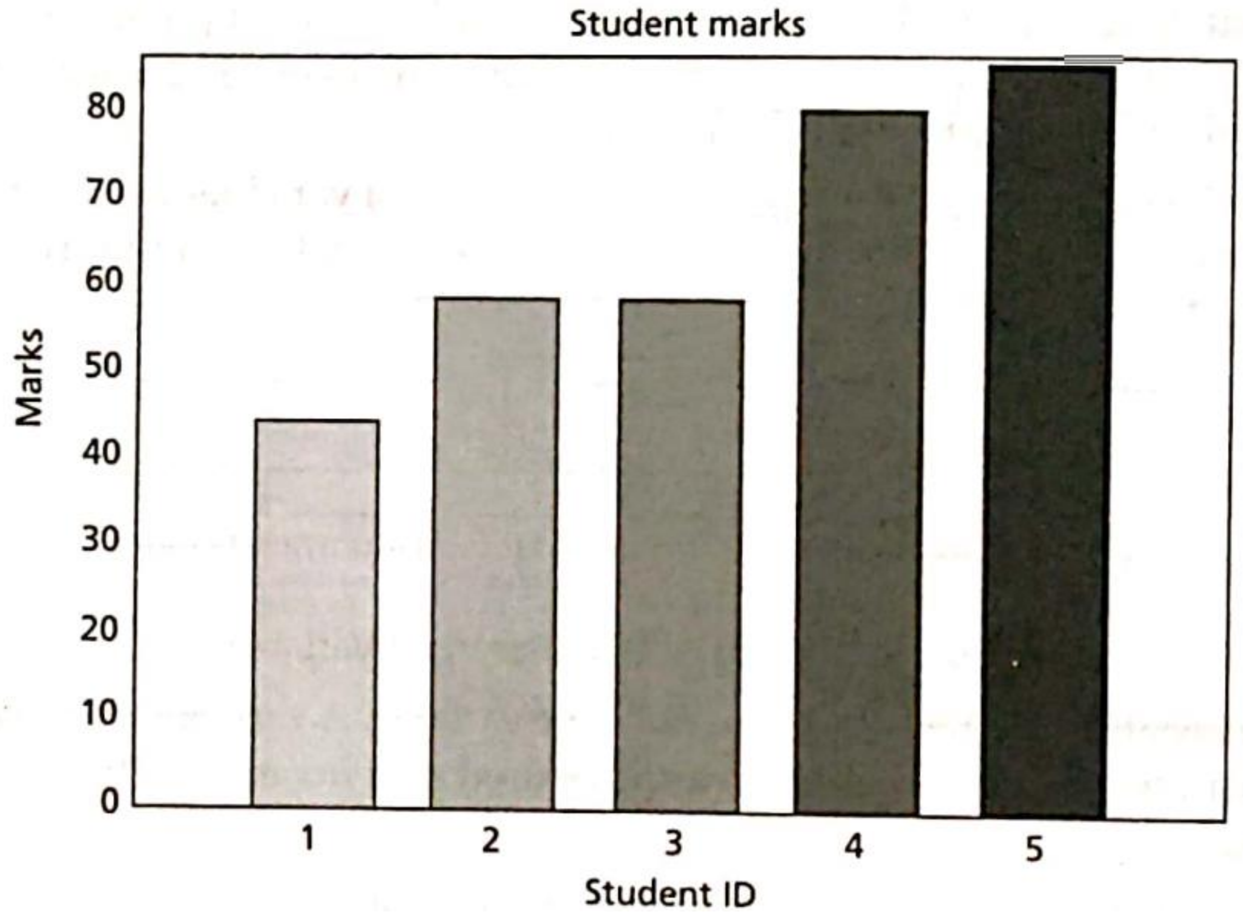
- To understand data, graph visualization is must. Data visualization helps to understand data.
- It helps to present information and data to customers. Some of the graphs that are used in univariate data analysis are bar charts, histograms, frequency polygons and pie charts.
- The advantages of the graphs are presentation of data, summarization of data, description of data, exploration of data, and to make comparisons of data.

Data Visualization - Bar Chart

- Let us consider some forms of graphs now:

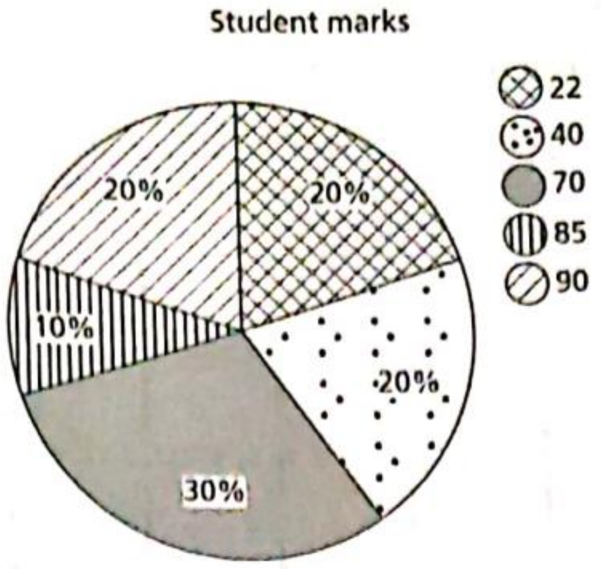
Bar Chart

  - A Bar chart (or Bar graph) is used to display the frequency distribution for variables. Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups.
  - The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown below in Figure



Data Visualization - Bar Chart

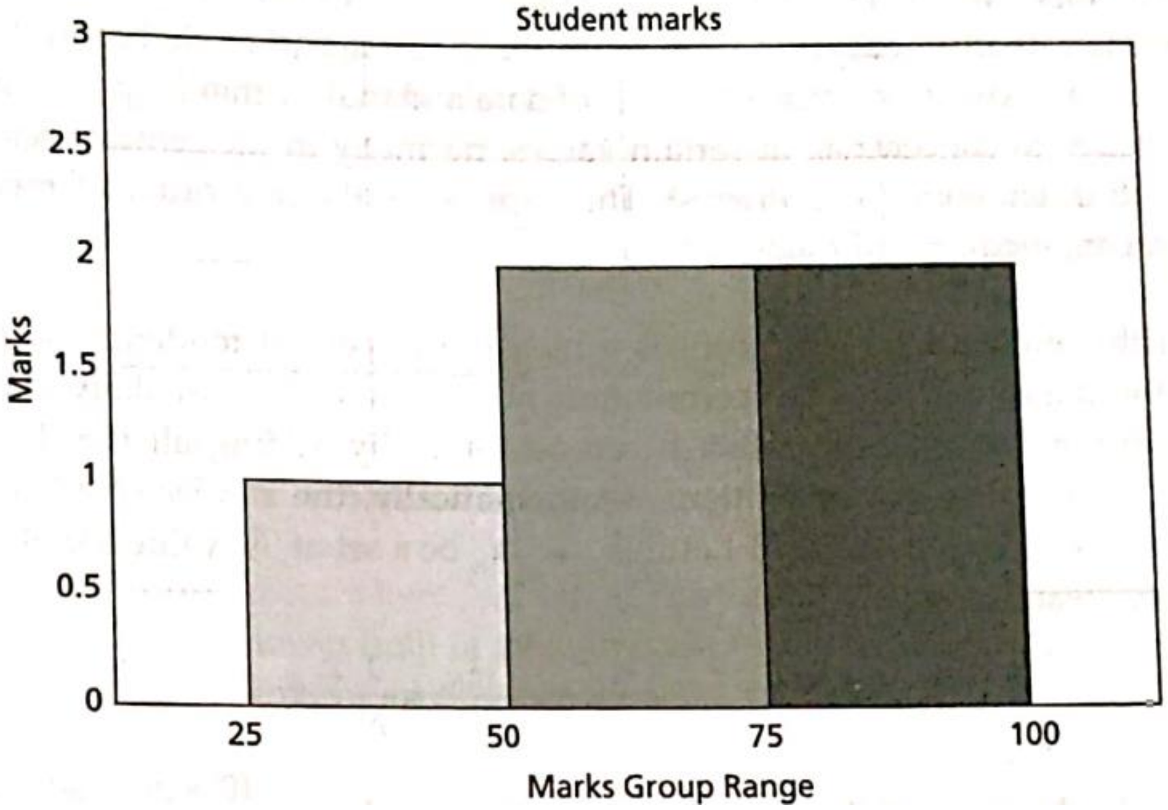
- Pie Chart
  - These are equally helpful in illustrating the univariate data. The percentage frequency distribution of students' marks {22, 22, 40, 40, 70, 70, 70, 85, 90, 90} is below in Figure
  - It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, 2/10 x 100 = 20% space in a pie of 100% is allotted for marks 22 in Figure 2.4.



Data Visualization - Histogram

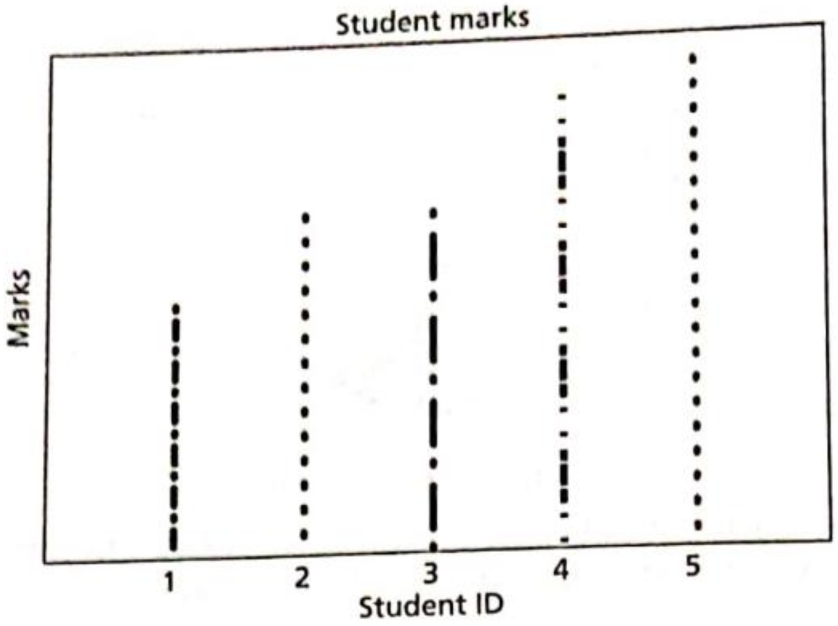
Histogram

- It plays an important role in data mining for showing frequency distributions. The histogram for students' marks {45, 60, 60, 80, 85} in the group range of 0-25, 26-50, 51-75,76-100 is given below in Figure 2.5. One can visually inspect from Figure 2.5 that the number of students in the range 76-100 is 2.
- Histogram conveys useful information like nature of data and its mode. Mode indicates the peak of dataset. In other words, histograms can be used as charts to show frequency, skewness present in the data, and shape.



Data Visualization - Dot plots

- Dot plots
  - These are similar to bar charts. They are less clustered as compared to bar charts as they illustrate the bars only with single points. The dot plot of English marks for five students with ID as {1, 2, 3, 4, 5) and marks (45, 60, 60, 80, 85) is given in Figure 2.6. The advantage is that by visual inspection one can find out who got more marks.



Central Tendency

- One cannot remember all the data. Therefore, a condensation or summary of the data is necessary.
- This makes the data analysis easy and simple. One such summary is called central tendency.
- Thus, central tendency can explain the characteristics of data and that further helps in comparison.
- Mass data have tendency to concentrate at certain values, normally in the central location. It is called measure of central tendency (or averages). This represents the first order of measures. Popular measures are mean, median and mode.

Mean

- Arithmetic average (or mean) is a measure of central tendency that represents the 'center' of the dataset.
- This is the commonest measure used in our daily conversation suchas average income or average traffic.
- It can be found by adding all the data and dividing the sum by the number of observations. Mathematically, the average of all the values in the sample (population) is denoted as x. Let x1, x2, ... , xN be a set of 'N' values or observations, then the arithmetic mean is given as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- For example, the mean of the three numbers 10, 20, and 30 is

(10+20 + 30)/ 3 = 60/3=20

Weighted mean

- Unlike arithmetic mean that gives the weightage of all items equally, weighted mean gives different importance to all items as the item importance varies. Hence, different weightage can be given to items.
- In case of frequency distribution, mid values of the range are taken for computation.
- In weighted mean, the mean is computed by adding the product of proportion and group mean. It is mostly used when the sample sizes are unequal.

Geometric mean

- Let x1, x2, ... , xy be a set of 'N' values or observations. Geometric mean is the Nth root of the product of N items. The formula for computing geometric mean is given as follows:

$$\text{Geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{N}} = \sqrt[N]{x_1 \times x_2 \times \cdots \times x_N}$$

- Here, n is the number of items and x are values. For example, if the values are 6 and 8,the geometric mean is given as $\sqrt[2]{6 * 8} = \sqrt{48}$.
- In larger cases, computing geometric mean is difficult. Hence, it is usually calculated as:

$$\text{Anti-log of } \frac{\log(x_1) + \log(x_2) + \cdots + \log(x_N)}{N}$$

$$= \text{anti-log } \frac{\sum_{i=1}^{n} \log(x_i)}{N}$$

- The problem of mean is its extreme sensitiveness to noise. Even small changes in the input affect the mean drastically. Hence, often the top 2% is chopped off and then the mean is calculated for a larger dataset.

Median

- The middle value in the distribution is called median.
- If the total number of items in the distribution is odd, then the middle value is called median.
- If the numbers are even, then the average value of two items in the center is the median.
    - It can be observed that the median is the value where x, is divided into two equal halves, with half of the values being lower than the median and half higher than the median.
    - A median class is that class where (N/2)th item is present.
- In the continuous case, the median is given by the formula: $\text{Median} = L_1 + \dfrac{\frac{N}{2} - cf}{f} \times i$
- Median class is that class where N/2th item is present. Here, i is the class interval of the median class and L is the lower limit of median class, f is the frequency of the median class, and cf is the cumulative frequency of all classes preceding median.

Mode

- Mode is the value that occurs more frequently in the dataset.
- In other words, the value that has the highest frequency is called mode.
- Mode is only for discrete data and is not applicable for continuous data as there are no repeated values in continuous data.
- The procedure for finding the mode is to calculate the frequencies for all the values in the data, and mode is the value (or values) with the highest frequency.
- Normally, the dataset is classified as unimodal, bimodal and trimodal with modes 1, 2 and 3, respectively

$22, 22, 40, 40, \underbrace{70, 70, 70}_{\text{Mode - highest No. of times appearing}}, 85, 90, 90$

Dispersion

- The spread out of a set of data around the central tendency (mean, median or mode) is called dispersion.
- Dispersion is represented by various ways such as range, variance, standard deviation and standard error. These are second order measures. The most common measures of the dispersion data are listed below:
  - Range: Range is the difference between the maximum and minimum of values of the given list of data.
  - Standard Deviation: The mean does not convey much more than a middle point.
    - For example, the following datasets {10, 20, 30} and {10, 50, 0} both have a mean of 20. The difference between these two sets is the spread of data.
    - Standard deviation is the average distance from the mean of the dataset to each point. The formula for sample standard deviation is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

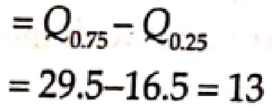
Quartiles and Inter Quartile Range

- It is sometimes convenient to subdivide the dataset using coordinates.
- Percentiles are about data that are less than the coordinates by some percentage of the total value.
- kth percentile is the property that the k% of the data lies at or below X,.
  - For example, median is 50th percentile and can be denoted as Q050.
  - The 25th percentile is called first quartile (Q1)and the 75th percentile is called third quartile (Q3).
  - Another measure that is useful to measure dispersion is Inter Quartile Range (IQR). The IQR is the difference between Q3 and Qr.

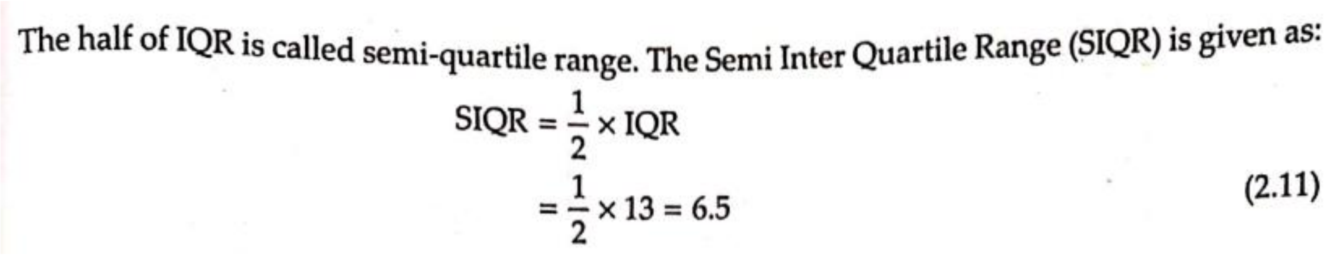
Interquartile percentile = Q3 - Q1

Example - IQR

- For patients' age list {12, 14, 19, 22, 24, 26, 28, 31, 34}, find the IQR.
- **Solution:** The median is in the fifth position. In this case, 24 is the median.
  - The first quartile is median of the scores below the mean i.e., {12, 14, 19, 22}.
  - Hence, it's the median of the list below 24.In this case, the median is the average of the second and third values, that is, Q025=16.5.
  - Similarly, the third quartile is the median of the values above the median, that is {26, 28, 31, 34). So, Qoz, is the average of the seventh and eighth score. In this case, it is 28 + 31/2 = 59/2 = 29.5.Hence, the IQR using Eq. (2.10) is:



$$= Q_{0.75} - Q_{0.25}$$
$$= 29.5–16.5 = 13$$



The half of IQR is called semi-quartile range. The Semi Inter Quartile Range (SIQR) is given as:

$$\begin{aligned} \text{SIQR} &= \frac{1}{2} \times \text{IQR} \\ &= \frac{1}{2} \times 13 = 6.5 \end{aligned} \qquad (2.11)$$

Five-point Summary and Box Plots

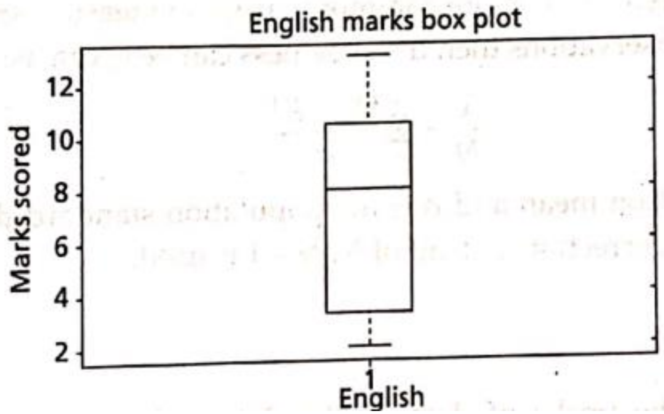
- The median, quartiles Q1 and Q3 and minimum and maximum written in the order < Minimum, Q1, Median, Q3 Maximum > is known as five-point summary.
- Box plots are suitable for continuous variables and a nominal variable. Box plots can be used to illustrate data distributions and summary of data. It is the popular way for plotting five number summaries. A Box plot is also known as a Box and whisker plot.
- The box contains bulk of the data. These data are between first and third quartiles. The line inside the box indicates location - mostly median of the data. If the median is not equidistant, then the data is skewed. The whiskers that project from the ends of the box indicate the spread of the tails and the maximum and minimum of the data value.

Five-point Summary and Box Plots

- The box contains bulk of the data. These data are between first and third quartiles. The line inside the box indicates location - mostly median of the data. If the median is not equidistant, then the data is skewed. The whiskers that project from the ends of the box indicate the spread of the tails and the maximum and minimum of the data value.

Example

- Find the 5-point summary of the list {13, 11, 2, 3, 4, 8, 9}.
- **Solution:** The minimum is 2 and the maximum is 13. The Qy Q2 and Q, are 3, 8 and 11, respectively. Hence, 5-point summary is {2, 3, 8, 11, 13}, that is, {minimum, Q1, median, Q3, maximum}. Box plots are useful for describing 5-point summary.

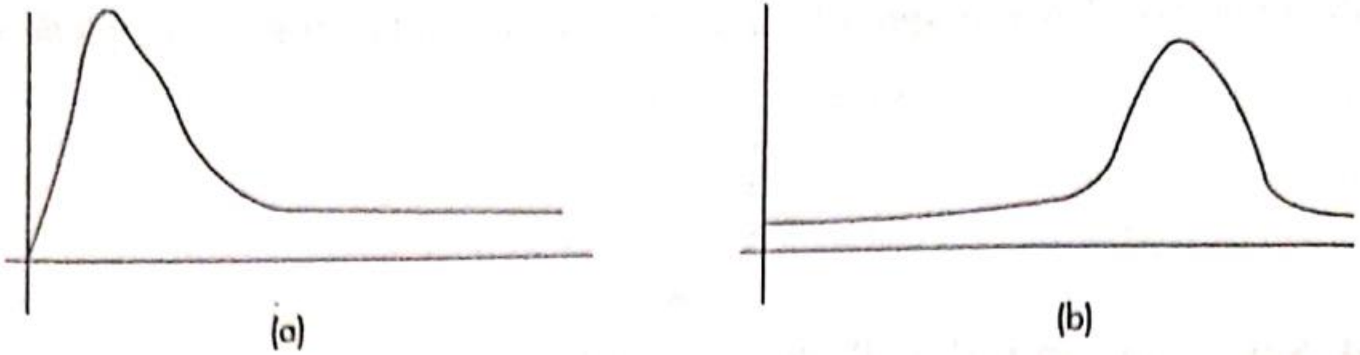


Ascending order = {2,3,4,8,9,11,13}
Median = 8
Min = 2
Max = 13
Q.25 = 3
Q.27= 11

Shape

- Skewness and Kurtosis (called moments) indicate the symmetry/asymmetry and peak location of the dataset.
- Skewness
    - The measures of direction and degree of symmetry are called measures of third order. Ideally, skewness should be zero as in ideal normal distribution. More often, the given dataset may not have perfect symmetry
    - The dataset may also either have very high values or extremely low values.
    - If the dataset has far higher values, then it is said to be skewed to the right. On the other hand, if the dataset has far more low values then it is said to be skewed towards left.

Shape



- Skewness
  - If the tail is longer on the left-hand side and hump on the right-hand side, it is called positive skew. Otherwise, it is called negative skew.
  - The given dataset may have an equal distribution of data.
  - The implication of this is that if the data is skewed, then there is a greater chance of outliers in the dataset.
  - This affects the mean and median. Hence, this may affect the performance of the data mining algorithm. A perfect symmetry means the skewness is zero.
  - In the case of skew, the median is greater than the mean. In positive skew, the mean is greater than the median.
  - For negatively skewed distribution, the median is more than the mean. The relationship between skew and the relative size of the mean and median can be summarized by a convenient numerical skew index known as Pearson 2 skewness coefficient.

$$\frac{3 \times (\mu - median)}{\sigma}$$

Shape

- Skewness
  - Also, the following measure is more commonly used to measure skewness. Let X_1, X_2 ·.. , X_N be a set of 'N' values or observations then the skewness can be given as:

$$\frac{1}{N} \times \sum_{i=1}^{N} \frac{(x_i - \mu)^3}{\sigma^3}$$

  - Here, u is the population mean and o is the population standard deviation of the univariate data. Sometimes, for bias correction instead of N, N - 1 is used.

Shape

- Kurtosis
  - Kurtosis also indicates the peaks of data. If the data is high peak, then it indicates higher kurtosis and vice versa.
  - Kurtosis is the measure of whether the data is heavy tailed or light tailed relative to normal distribution.
  - It can be observed that normal distribution has bell-shaped curve with no long tails.
  - Low kurtosis tends to have light tails.
  - The implication is that there is no outlier data.
  - Let x_1, x_2, .·· ,x_N be a set of 'N' values or observations.
  - Then, kurtosis is measured using the formula given by:

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^4/N}{\sigma^4}$$

  - It can be observed that N - 1 is used instead of N in the numerator of Eq. (2.14) for bias correction.
  - Here, x and o are the mean and standard deviation of the univariate data, respectively.

Shape

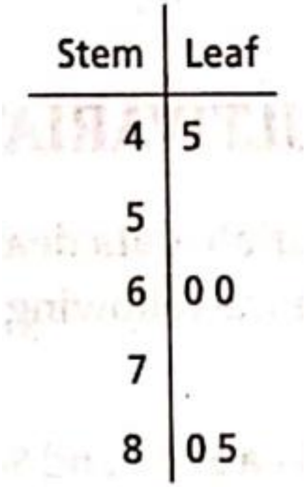
- Some of the other useful measures for finding the shape of the univariate dataset are mean absolute deviation (MAD) and coefficient of variation (CV).
  - Mean Absolute Deviation (MAD)
    - MAD is another dispersion measure and is robust to outliers. Normally, the outlier point is detected by computing the deviation from median and by dividing it by MAD. Here, the absolute deviation between the data and mean is taken. Thus, the absolute deviation is given as: lx -µl
    - The sum of the absolute deviations is given as: Σ lx - µl
    - Therefore, the mean absolute deviation is given as: $\Sigma|x – \mu|/N$

Shape

- Coefficient of Variation (CV)
  - Coefficient of variation is used to compare datasets with different units. CV is the ratio of standard deviation and mean, and %CV is the percentage of coefficient of variations.

Special Univariate Plots

- The ideal way to check the shape of the dataset is a stem and leaf plot.
- A stem and leaf plot are a display that help us to know the shape and distribution of the data. In this method, each value is split into a 'stem' and a 'leaf'.
- The last digit is usually the leaf and digits to the left of the leaf mostly form the stem. For example, marks 45 are divided into stem 4 and leaf 5 in Figure 2.9.
- It can be seen from Figure 2.9 that the first column is stem and the second column is leaf.
- For the given English marks, two students with 60 marks are shown in stem and leaf plot as stem-6 with 2 leaves with 0.



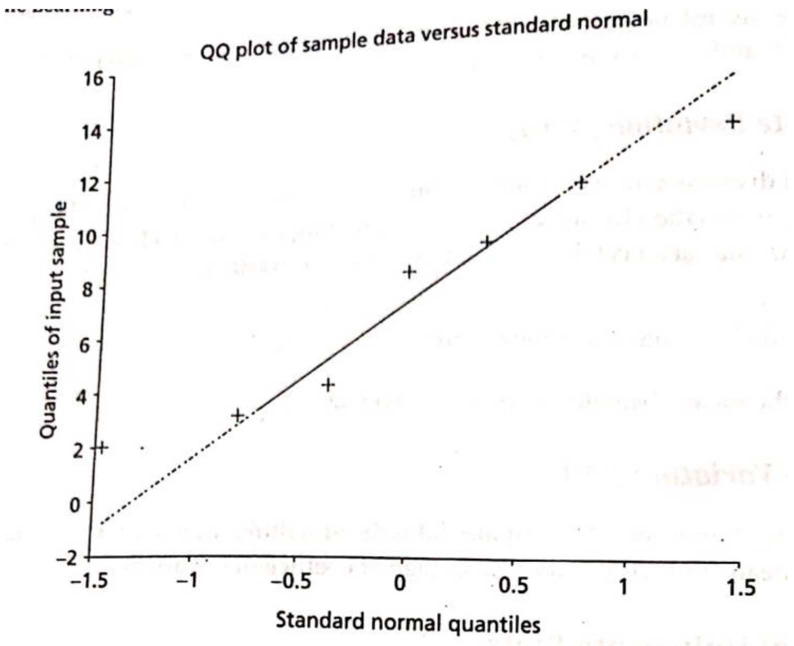
Stem	Leaf
4	5
5	
6	0 0
7	
8	0 5

Special Univariate Plots

- It can be seen from Figure 2.9 that the first column is stem and the second column is leaf. For the given English marks, two students with 60 marks are shown in stem and leaf plot as stem-6 with 2 leaves with 0.
- As discussed earlier, the ideal shape of the dataset is a bell-shaped curve.
- This corresponds to normality. Most of the statistical tests are designed only for normal distribution of data.
- A Q-Q plot can be used to assess the shape of the dataset.
- The Q-Q plot is a 2D scatter plot of an univariate data against theoretical normal distribution data or of two datasets - the quartiles of the first and second datasets.
- The normal Q-Q plot for marks x = [13 11 2 3 4 8 9] is given below in Figure 2.10.

Special Univariate Plots

- Ideally, the points fall along the reference line (45 Degree) if the data follows normal distribution.
- If the deviation is more, then there is greater evidence that the datasets follow some different distribution, that is, other than the normal distribution shape.
- In such a case, careful analysis of the statistical investigations should be carried out before interpretation.
- This skewness, kurtosis, mean absolute deviation and coefficient of variation help in assessing the univariate data.



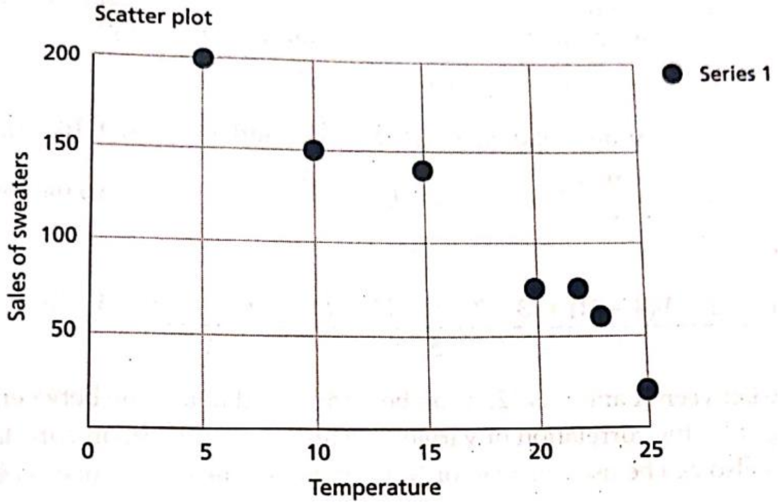
BIVARIATE DATA AND MULTIVARIATE DATA

- Bivariate Data involves two variables. Bivariate data deals with causes of relationships.
- The aim is to find relationships among data. Consider the following Table 2.3, with data of the temperature in a shop and sales of sweaters.
- Temperature in a Shop and Sales Data

Temperature	Sales of Sweater
5	200
10	150
15	140
20	75
22	60
23	55
25	20

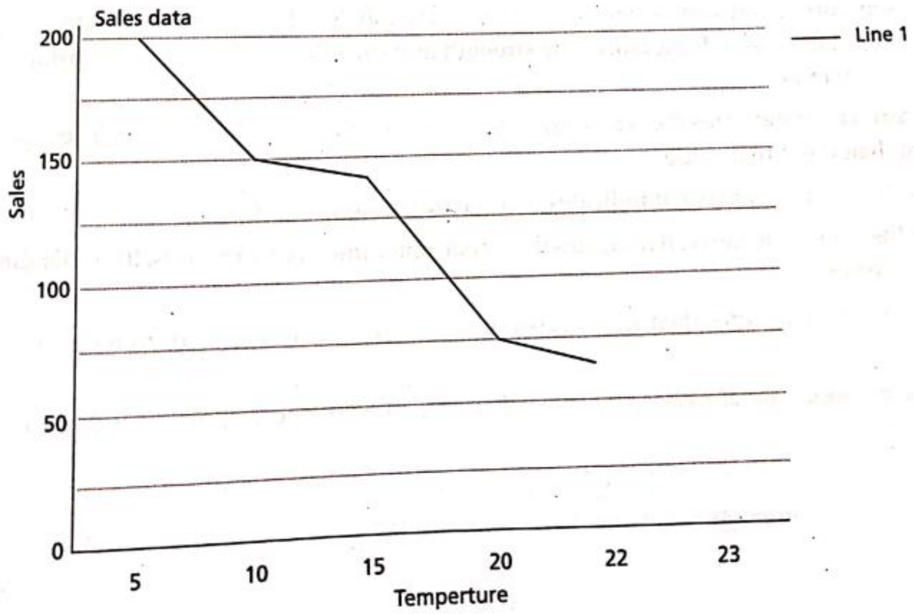
BIVARIATE DATA AND MULTIVARIATE DATA

- Here, the aim of bivariate analysis is to find relationships among variables. The relationships can then be used in comparisons, finding causes, and in further explorations. To do that, graphical display of the data is necessary. One such graph method is called scatter plot.
- Scatter plot is used to visualize bivariate data. It is useful to plot two variables with or without nominal variables, to illustrate the trends, and also to show differences. It is a plot between explanatory and response variables. It is a 2D graph showing the relationship between two variables.
- The scatter plot (Refer Figure 2.11) indicates strength, shape, direction and the presence of Outliers. It is useful in exploratory data before calculating a correlation coefficient or fitting regression curve.



BIVARIATE DATA AND MULTIVARIATE DATA

- Line graphs are similar to scatter plots. The Line Chart for sales data is shown in Figure 2.12.



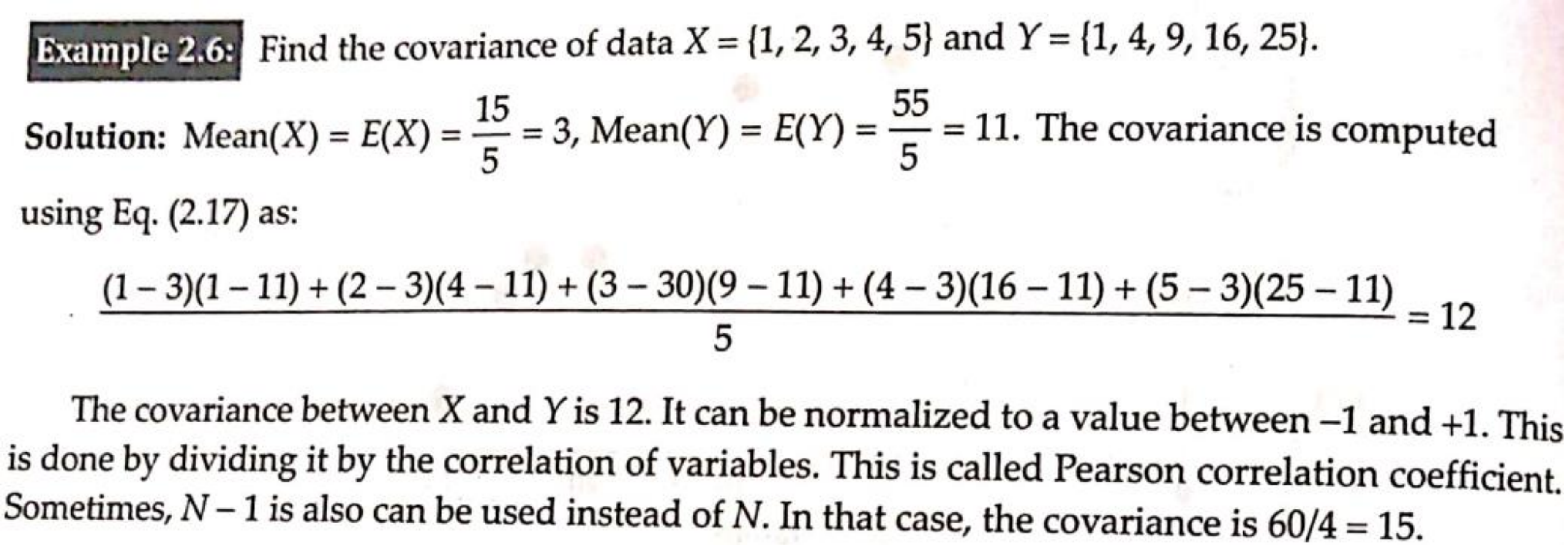
Bivariate Statistics

- Covariance and Correlation are examples of bivariate statistics. Covariance is a measure of joint probability of random variables, say X and Y.
- Generally, random variables are represented in capital letters. It is defined as covariance(X, Y) or COV(X, Y) and is used to measure the variance between two dimensions.
- The formula for finding co-variance for specific x, and y are:

$$\text{cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(X))(y_i - E(Y))$$

- Here, x, and y, are data values from X and Y.
- E(X) and E(Y) are the mean values of x, and y. N is the number of given data. Also, the COV(X, Y) is same as COV(Y, X).

Example



Example 2.6: Find the covariance of data $X = \{1, 2, 3, 4, 5\}$ and $Y = \{1, 4, 9, 16, 25\}$.

Solution: $\text{Mean}(X) = E(X) = \frac{15}{5} = 3$, $\text{Mean}(Y) = E(Y) = \frac{55}{5} = 11$. The covariance is computed using Eq. (2.17) as:

$$\frac{(1-3)(1-11)+(2-3)(4-11)+(3-30)(9-11)+(4-3)(16-11)+(5-3)(25-11)}{5} = 12$$

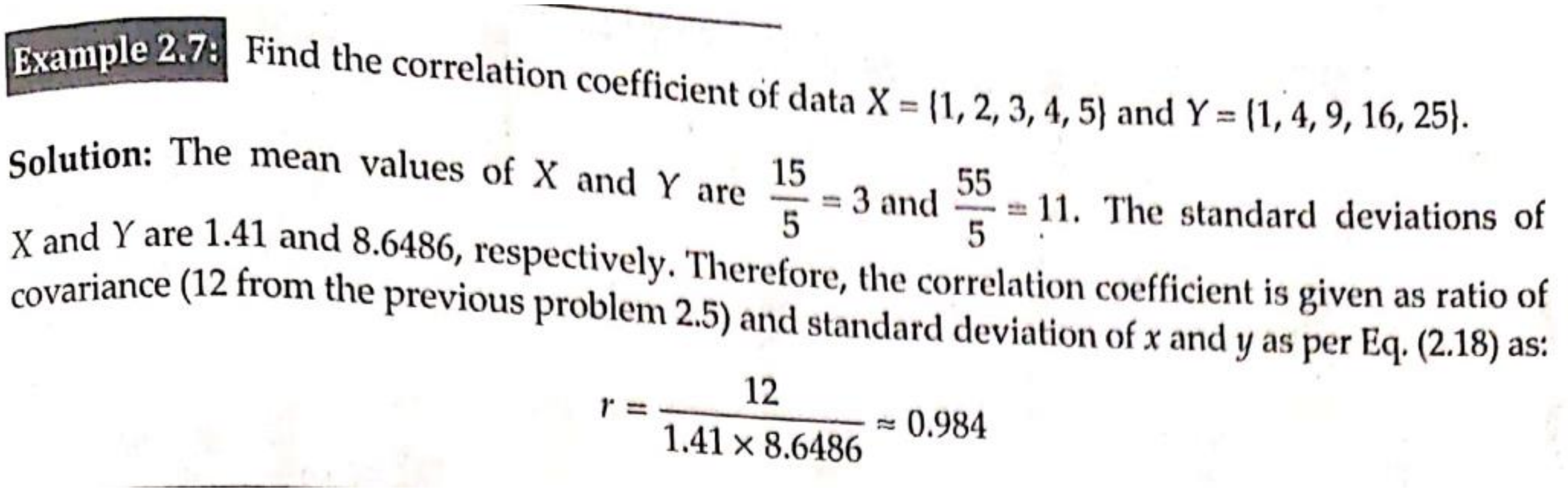
The covariance between X and Y is 12. It can be normalized to a value between –1 and +1. This is done by dividing it by the correlation of variables. This is called Pearson correlation coefficient. Sometimes, $N - 1$ is also can be used instead of N. In that case, the covariance is 60/4 = 15.

Correlation

- The Pearson correlation coefficient is the most common test for determining any association between two phenomena.
- It measures the strength and direction of a linear relationship between the x and y variables.
- The correlation indicates the relationship between dimensions using its sign. The sign is more. important than the actual value.
  1. If the value is positive, it indicates that the dimensions increase together.
  2. If the value is negative, it indicates that while one-dimension increases, the other dimension decreases.
  3. If the value is zero, then it indicates that both the dimensions are independent of each other.

If the dimensions are correlated, then it is better to remove one dimension as it is a redundant dimension.

Correlation Example



Example 2.7: Find the correlation coefficient of data $X = \{1, 2, 3, 4, 5\}$ and $Y = \{1, 4, 9, 16, 25\}$.

Solution: The mean values of X and Y are $\frac{15}{5} = 3$ and $\frac{55}{5} = 11$. The standard deviations of X and Y are 1.41 and 8.6486, respectively. Therefore, the correlation coefficient is given as ratio of covariance (12 from the previous problem 2.5) and standard deviation of x and y as per Eq. (2.18) as:

$$r = \frac{12}{1.41 \times 8.6486} \approx 0.984$$

Machine Learning and Importance of Probability and Statistics

- Machine learning is linked with statistics and probability. Like linear algebra, statistics is the heart of machine learning.
- The importance of statistics needs to be stressed as without statistics analysis of data is difficult.
- Probability is especially important for machine learning. Any data can be assumed to be generated by a probability distribution.
- Machine learning datasets have multiple data that are generated by multiple distributions.
- So, a knowledge of probability distribution and random variables are must for better understanding of the machine learning concepts.

Probability Distributions

- A probability distribution of a variable, say X, summarizes the probability associated with X's events. Distribution is a parameterized mathematical function. In other words, distribution is a function that describes the relationship between the observations in a sample space.
- Consider a set of data. The data is said to follow a distribution if it obeys a mathematical function that characterizes that distribution. The function can be used to calculate the probability of individual observations.
- Probability distributions are of two types:
  - Discrete Probability Distribution
  - Continuous Probability Distribution

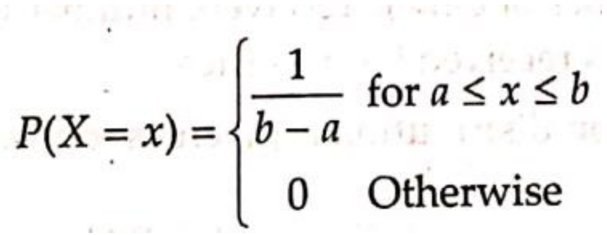
Continuous Probability Distributions

Normal, Rectangular, and Exponential distributions fall under this category.

- Normal Distribution:
  - Normal distribution is a continuous probability distribution.
  - This is also known as gaussian distribution or bell-shaped curve distribution. It is the most common distribution function.
  - The shape of this distribution is a typical bell-shaped curve.
  - In normal distribution, data tends to be around a central value with no bias on left or right.
  - The heights of the students, blood pressure of a population, and marks scored in a class can be approximated using normal distribution.

- Normal-
  - PDF of the normal distribution is given as: $f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Here, u is mean and o is the standard deviation.
  - Normal distribution is characterized by two parameters - mean and variance.
  - Mostly, one uses the normal distribution curve of mean 0 and a SD of 1. In normal distribution, mean, median and mode are same. The distribution extends from - ∞ to + ∞.
  - Standard deviation is how the data is spread out.
  - One important concept associated with normal distribution is z-score. It can be computed as: $z = \frac{x - \mu}{\sigma}$
  - When μ is zero and σ is 1, z-score is same as x. This is useful to normalize the data.

- Rectangular Distribution
  - This is also known as uniform distribution.
  - It has equal probabilities for all values in the range a, b. The uniform distribution is given as follows:



$$P(X = x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{Otherwise} \end{cases}$$

- Exponential Distribution –
  - This is a continuous uniform distribution. This probability distribution is used to describe the time between events in a Poisson process.
  - Exponential distribution is another special case of Gamma distribution with a fixed parameter of 1.
  - This distribution is helpful in modelling of time until an event occurs.
  - The PDF is given as follows:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \qquad (\lambda > 0)$$

  - Here, x is a random variable λ and 1 is called rate parameter. The mean and standard deviation of exponential distribution is given as ß, where, ß = 1/ λ

Discrete Distribution

Binomial, Poisson, and Bernoulli distributions fall under this category.

- Binomial Distribution –
  - Binomial distribution is another distribution that is often encountered in machine learning. It has only two outcomes: success or failure. This is also called Bernoulli trial.
  - The objective of this distribution is to find probability of getting success k out of n trial. The way to get success out of k out of n number of trials is given as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

  - The binomial distribution function is given as follows, where p is the probability of success and probability of failure is (1 - p). The probability of success in a certain number of trials is given as: $p^k(1-p)^{n-k}$ or $p^k q^{n-k}$

Discrete Distribution

- Combining both, one gets PDF of binomial distribution as: $\binom{n}{k} p^k (1-p)^{n-k}$
- Here, p is the probability of each choice, k is the number of choices, and n is the total number of choices. The mean of binomial distribution is given below: $\mu = n \times p$
- And the variance is given as: $\sigma^2 = np(1-p)$
- Hence, the standard deviation is given as: $\sigma = \sqrt{np(1-p)}$

Discrete Distribution

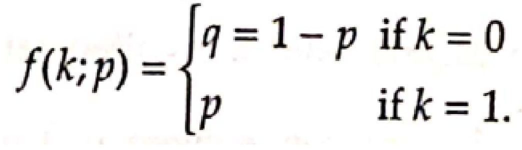
- Poisson Distribution –
  - It is another important distribution that is quite useful. Given an interval of time, this distribution is used to model the probability of a given number of events k.
  - The mean rule A is inclusive of previous events. Some of the examples of Poisson distribution are number of emails received, number of customers visiting a shop and the number of phone calls received by the office.
  - The PDF of Poisson distribution is given as follows:

$$f(X = x; \lambda) = Pr[X = x] = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

  - Here, x is the number of times the event occurs and 1 is the mean number of times an event occurs.
  - The mean is the population mean at number of emails received and the standard deviation is $\sqrt{\lambda}$.

Discrete Distribution

- Bernoulli Distribution –
  - This distribution models an experiment whose outcome is binary. The outcome is positive with p and negative with 1 - p.
  - The PMF of this distribution is given as:



$$f(k;p) = \begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1. \end{cases}$$

  - The mean is p and variance is p(1 - p) = q

OVERVIEW OF HYPOTHESIS

- Data collection alone is not enough.
- Data must be interpreted to give a conclusion.
- The conclusion should be a structured outcome.
- This assumption of the outcome is called a hypothesis.
- Statistical methods are used to confirm or reject the hypothesis.
- The assumption of the statistical test is called null hypothesis. Or hypothesis zero (H0).
- In other words, hypothesis is the existing belief. The violation of this hypothesis is called first hypothesis (H1) or hypothesis one.
- This is the hypothesis the researcher is trying to establish.

OVERVIEW OF HYPOTHESIS

- There are two types of hypothesis tests, parametric and non-parametric. Parametric tests are based on parameters such as mean and standard deviation. Non-parametric tests are dependent on characteristics such as independence of events or data following certain distribution.
- Statistical tests help to:
  - Define null and alternate hypothesis
  - Describe the hypothesis using parameters
  - Identify the statistical test and statistics
  - Decide the criteria called significance value α
  - Compute p-value (probability value)
  - Take the final decision of accepting or rejecting the hypothesis based on the parameters

- No matter how effective the statistical tests are, two kinds of errors are involved, that are Type I and Type II.
  - Type I error is the incorrect rejection of a true null hypothesis and is called false positive.
  - Type II error is the incomplete failure of rejecting a false hypothesis and is called false negative.