Data Mining and Business Intelligence

Scheme and Solution

PART I

1. Write notes on Classification and Regression.
    a. Classification – Decision Tree, Naïve Bayes, Rule Based Classification(5m)
    b. Regression – Linear Regression, Logistic Regression(5m)
2. Explain CART Decision tree induction classification method with example.
    a. Finding Error (5m)
    b. Formation of Decision Tree (5m)

PART II

3. Explain Linear and Logistic Regression Prediction methods with example.
    a. Linear Regression(5m)
    b. Logistic Regression(5m)
4. Explain ID3 Decision tree induction classification method.
    a. Calculation of Entropy(3m)
    b. Calculation of Information gain(3m)
    c. Formation of Decision Tree (4m)

PART III

5. How are Rule Based Classifiers used for Classification?
    a. Concept(5m)
    b. Example(5m)
6. What is an Artificial Neuron? Explain multilayer neural Networks.
    a. Basic artificial neuron(5m)
    b. Multilayer neural networks(5m)

PART IV

7. Write a note on Naive Bayes Classifier.
    a. Concept(3m)
    b. Formula and Example(7m)
8. Write a note on data mining tools DB Miner, WEKA and DTREG.
    a. DB Miner(3m)
    b. WEKA(4m)
    c. DTREG(3m)

PART V

9. Explain the main phases of Data Analytics Life Cycle in detail.
    a. Phases(7m)
    b. Diagram(3m)
10. Discuss the key stakeholders of analytics project.
    a. Each 2 m

1. **Write notes on Classification and Regression.**

There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends.

 These two forms are as follows

o Classification

o Prediction

 Classification models predict categorical class labels.

 Prediction models predict continuous valued functions. For example,

 We can build a classification model to categorize bank loan applications as either safe or risky.

 Prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

Following are the examples of cases where the data analysis task is Classification −

o A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.

o A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

 In both of the above examples, a model or classifier is constructed to predict the categorical labels.

These labels are risky or safe for loan application data and yes or no for marketing data.

Here is the criteria for comparing the methods of Classification and Prediction −

 Accuracy − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

 Speed − this refers to the computational cost in generating and using the classifier or predictor.

 Robustness − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

 Scalability − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.

 Interpretability − It refers to what extent the classifier or predictor understands.

Regression is a data mining function that predicts a number.

 Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques.

 For example, a regression model could be used to predict children's height, given their age, weight, and other factors.

 A regression task begins with a data set in which the target values are known.

 For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time.

 The data might track age, height, weight, developmental milestones, family history, and so on.

 Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.

 Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values.

 It is required to understand the mathematics used in regression analysis to develop quality regression models for data mining.

 The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide.

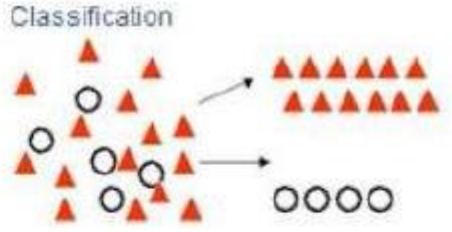
2. Explain CART Decision tree induction classification method with example.

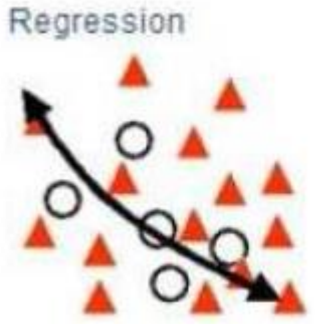
Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of a target (or dependent variable) based on the values of several input (or independent variables).

 The CART or Classification & Regression Trees methodology was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone.

 Classification Trees: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.

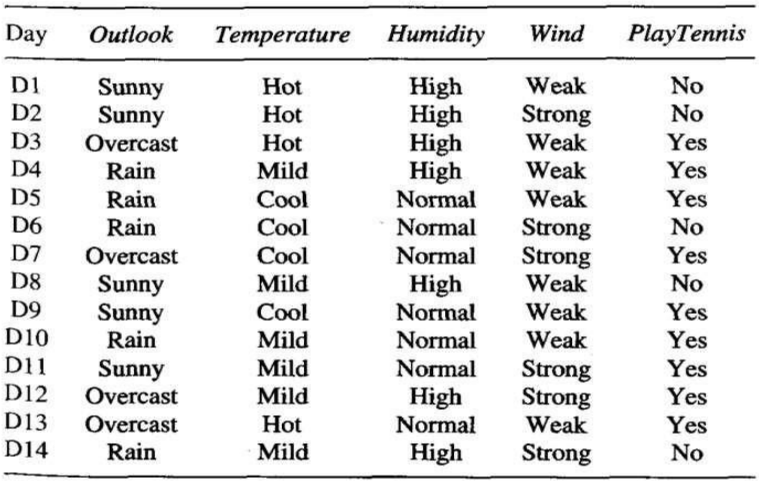
 Regression Trees: where the target variable is continuous and tree is used to predict its value.



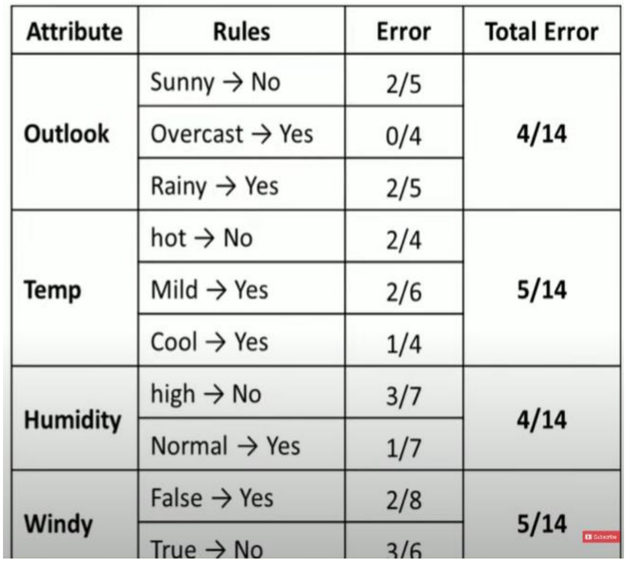


- **Classification and Regression Tree(CART)** is a **dynamic learning algorithm** that can produce a regression tree as well as a classification tree depending upon the dependent variable.
- Used for Classification when the target variable is continuous.
- Here, **Minimum Error** attribute is selected.

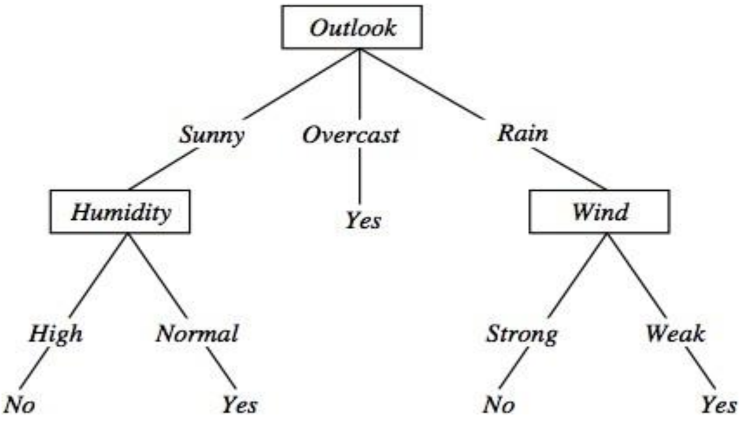
Eg:



Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Attribute	Rules	Error	Total Error
Outlook	Sunny → No	2/5	**4/14**
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	hot → No	2/4	**5/14**
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	high → No	3/7	**4/14**
	Normal → Yes	1/7	
Windy	False → Yes	2/8	**5/14**
	True → No	3/6	



3. Explain Linear and Logistic Regression Prediction methods with example.

Regression is a data mining function that predicts a number.

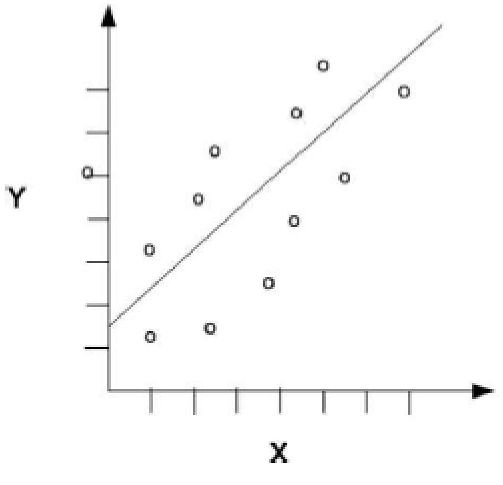
- Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques.

 For example, a regression model could be used to predict children's height, given their age, weight, and other factors.
 A regression task begins with a data set in which the target values are known.
 For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time.
 The data might track age, height, weight, developmental milestones, family history, and so on.
 Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.
 Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values.
 It is required to understand the mathematics used in regression analysis to develop quality regression models for data mining.
 The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide.

Linear Regression

 The simplest form of regression to visualize is linear regression with a single predictor.
 A linear regression technique can be used if the relationship between x and y can be approximated with a straight line.
 Linear regression with a single predictor can be expressed with the following equation.
 $y = \theta2x + \theta1 + e$
 The regression parameters in simple linear regression are:
 The slope of the line (θ) — the angle between a data point and the regression line
 The y intercept (θ) — the point where x crosses the y axis (x = 0)

Figure 4-1 Linear Relationship Between x and y

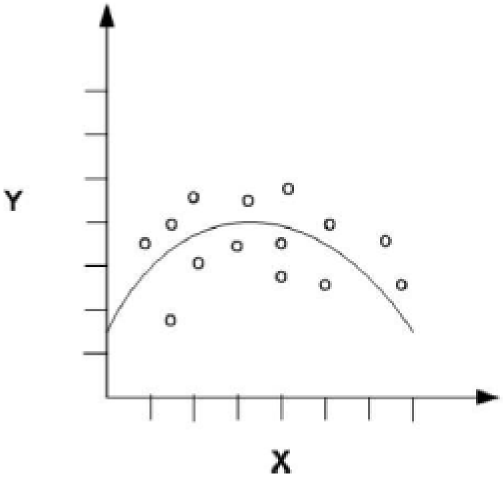


Nonlinear Regression

 Often the relationship between x and y cannot be approximated with a straight line.

 In this case, a nonlinear regression technique may be used. Alternatively, the data could be preprocessed to make the relationship linear

Figure 4-2 Nonlinear Relationship Between x and y



Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy).

 The prediction is based on the use of one or several predictors (numerical and categorical).

 A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

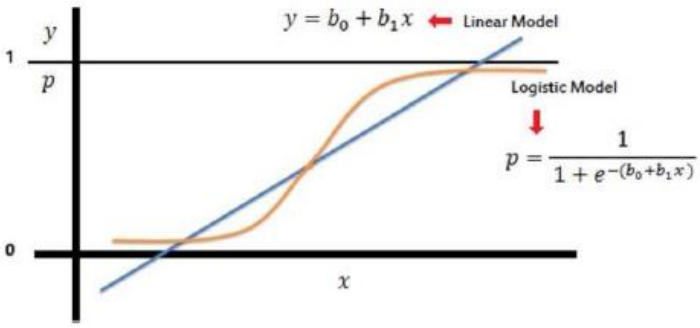
o A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1).

o Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

 A logistic regression produces a logistic curve, which is limited to values between 0 and 1.

 Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm
of the “odds” of the target variable, rather than the probability.

 Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



4. Explain ID3 Decision tree induction classification method.

ENTROPY MEASURES HOMOGENEITY OF EXAMPLES

- S is a collection of training examples
  - $p_{\oplus}$ the proportion of positive examples in S $\boxed{\left(\frac{P}{P+N}\right)}$
  - $p_{\ominus}$ the proportion of negative examples in S $\boxed{\left(\frac{N}{P+N}\right)}$
- ***Entropy is 0 if all members of S belong to the same class.(all are either +ve or –ve)***
- ***Entropy is 1 when the collection contains equal number of +ve and –ve examples.***
- ***If the collection contains unequal number of +ve and –ve examples, entropy is between 0 and 1.***

ENTROPY MEASURES HOMOGENEITY OF EXAMPLES

- ***Entropy,*** characterizes the **impurity of an arbitrary collection** of examples.
- Given a collection S, containing positive and negative examples of some target concept, the entropy of S relative to this boolean classification is

$$\boxed{Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}}$$

- where p+, is the proportion of positive examples in S and p-, is the proportion of negative examples in S.

- Information gain measures the expected reduction in Entropy.
- It is simply the expected reduction in entropy caused by partitioning the examples according to an attribute.

  (Selects Node with min impurity)

$$\boxed{Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)}$$

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

$\boldsymbol{Values\ (a1) = T, F}$

$\boldsymbol{S\ =\ [3+, 3-]}$ $\boldsymbol{Entropy(S) = 1.0}$

$\boldsymbol{S_T\ =\ [2+,\ 1-]}$ $\boldsymbol{Entropy(S_T) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183}$

$\boldsymbol{S_F \leftarrow [1+,\ 2-]}$ $\boldsymbol{Entropy(S_F) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9183}$

$\boldsymbol{Gain\,(S, a1)}$ =Entropy (S) $\sum_{v\in\{T,F\}} \frac{|S_v|}{|S|}Entropy(S_v)$

$$Gain(S, a1) = Entropy(S) - \frac{3}{6}Entropy(S_T) - \frac{3}{6}Entropy(S_F)$$

$$Gain(S, a1) = 1.0 - \frac{3}{6} * 0.9183 - \frac{3}{6} * 0.9183 = 0.0817$$

$Values\ (a2) = T, F$

$S\ =\ [3+, 3-]$ $\qquad$ $Entropy(S) = 1.0$

$S_T\ =\ [2+,\ 2-]$ $\qquad$ $Entropy(S_T) = 1.0$

$S_F \leftarrow [1+, 1-]$ $\qquad$ $Entropy(S_F) = 1.0$

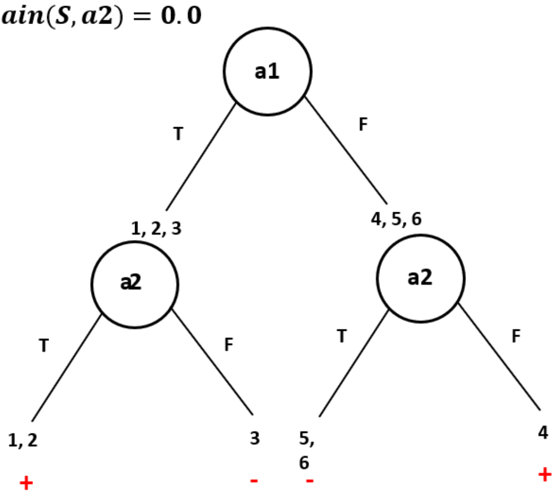
$$Gain\ (S, a2) = Entropy(S) - \sum_{v \in \{T,F\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, a2) = Entropy(S) - \frac{4}{6} Entropy(S_T) - \frac{2}{6} Entropy(S_F)$$

$$Gain(S, a2) = 1.0 - \frac{4}{6} * 1.0 - \frac{2}{6} * 1.0\ = 0.0$$

$Gain(S, a1) = 0.0817\ \ -\ Maximum\ Gain$

$Gain(S, a2) = 0.0$



5. How are Rule Based Classifiers used for Classification?
   Rule-based classifier makes use of a set of IF-THEN rules for classification.
    We can express a rule in the following from
    Let us consider a rule
   R1,
   IF condition THEN conclusion
   R1: IF age=youth AND
   student=yes THEN
   buy_computer=yes
    The IF part of the rule is called rule antecedent or precondition.
    The THEN part of the rule is called rule consequent.

- The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.
- We can also write rule R1 as follows:

R1: (age = youth) ^ (student = yes))(buys_computer = yes)

- If the condition (that is, all of the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied (or simply, that the rule is satisfied) and that the rule covers the tuple.
- A rule R can be assessed by its coverage and accuracy.
- Given a tuple, X, from a class labeled data set D, let ncovers be the number of tuples covered by R; ncorrect be the number of tuples correctly classified by R; and |D| be the number of tuples in D.
- That is, a rule's coverage is the percentage of tuples that are covered by the rule (i.e. whose attribute values hold true for the rule's antecedent).
- For a rule's accuracy, we look at the tuples that it covers and see what percentage of them the rule can correctly classify.
- We can use rule-based classification to predict the class label of a given tuple X.
- If a rule is satisfied by X, the rule is said to be triggered.
- For example, suppose we have

X= (age = youth, income = medium, student = yes, credit rating = fair)

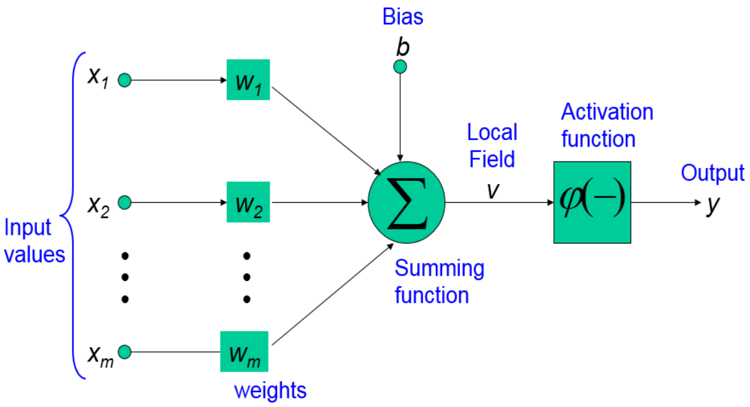
- We would like to classify X according to buys_computer. X satisfies R1, which triggers the rule.
- If R1 is the only rule satisfied, then the rule fires by returning the class prediction for X.
- If more than one rule is triggered, we need a conflict resolution strategy to figure out which rule gets to fire and assign its class prediction to X.
- There are many possible strategies. We look at two, namely size ordering and rule ordering.
- **Size ordering**
  - The size ordering scheme assigns the highest priority to the triggering rule that has the "toughest" requirements, where toughness is measured by the rule antecedent size.
  - That is, the triggering rule with the most attribute tests is fired.
- **Rule ordering**
  - The rule ordering scheme prioritizes the rules beforehand. The ordering may be class based or rule- based.

$$coverage(R) = \frac{n_{covers}}{|D|}$$

$$accuracy(R) = \frac{n_{correct}}{n_{covers}}.$$

o With class-based ordering, the classes are sorted in order of decreasing "importance," such as by decreasing order of prevalence.
o That is, all of the rules for the most prevalent (or most frequent) class come first, the rules for the next prevalent class come next, and so on.
o With rule-based ordering, the rules are organized into one long priority list, according to some measure of rule quality such as accuracy, coverage, or size (number of attribute tests in the rule antecedent), or based on advice from domain experts.
o When rule ordering is used, the rule set is known as a decision list.
o With rule ordering, the triggering rule that appears earliest in the list has highest priority, and so it gets to fire its class prediction.
o Any other rule that satisfies X is ignored. Most rule-based classification systems use a class based rule-ordering strategy.

6. What is an Artificial Neuron? Explain multilayer neural Networks.

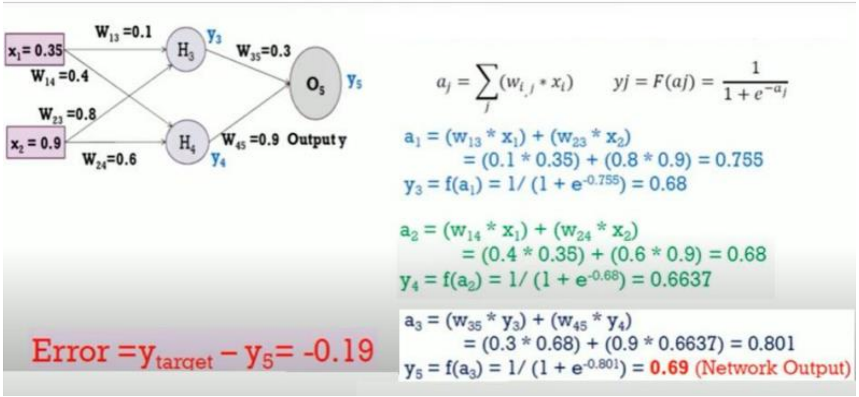


Neural Network is a set of connected INPUT/OUTPUT UNITS, where each connection has a WEIGHT associated with it.
⍰ Neural Network learning is also called CONNECTIONIST learning due to the connections between units.
⍰ It is a case of SUPERVISED, INDUCTIVE or CLASSIFICATION learning.
⍰ Neural Network learns by adjusting the weights so as to be able to correctly classify the training data and hence, after testing phase, to classify unknown data.
Strengths of Neural Network:
⍰ It can handle against complex data. (i.e., problems with many parameters)
⍰ It can handle noise in the training data.
⍰ The Prediction accuracy is generally high.
⍰ Neural Networks are robust, work well even when training examples contain errors.
⍰ Neural Networks can handle missing data well.
⍰ The greatest power of Neural Networks is that it is endowed with a finite number of hidden units, can yet approximate any continuous function to any desired degree

of accuracy. This has been commonly referred to as the property of universal approximation.

- No prior knowledge of the data generating process is needed for implementing Neural Network.
- Problem of model misspecification does not occur.
- In case of Neural Network since no specifications are used as the network merely learns the hidden relationship in the data.

Example of Multilayer Network



$$a_j = \sum_j (w_{i,j} * x_i) \qquad yj = F(aj) = \frac{1}{1+e^{-a_j}}$$

$a_1 = (w_{13} * x_1) + (w_{23} * x_2)$
$= (0.1 * 0.35) + (0.8 * 0.9) = 0.755$
$y_3 = f(a_1) = 1/(1 + e^{-0.755}) = 0.68$

$a_2 = (w_{14} * x_1) + (w_{24} * x_2)$
$= (0.4 * 0.35) + (0.6 * 0.9) = 0.68$
$y_4 = f(a_2) = 1/(1 + e^{-0.68}) = 0.6637$

$a_3 = (w_{35} * y_3) + (w_{45} * y_4)$
$= (0.3 * 0.68) + (0.9 * 0.6637) = 0.801$
$y_5 = f(a_3) = 1/(1 + e^{-0.801}) = 0.69$ (Network Output)

Error $= y_{target} - y_5 = -0.19$

7. Write a note on Naive Bayes Classifier.

Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

- Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes.
- This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve”.
- Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes.
- Bayesian belief networks can also be used for classification.
- In Bayesian terms, X is considered “evidence”.
- As usual, it is described by measurements made on a set of n attributes.
- Let H be some hypothesis, such as that the data tuple X belongs to a specified class C.
- For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds given the “evidence” or observed data tuple X.
- In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

Posterior probability:

- P(H|X) is the posterior probability, or a posterior probability, of H conditioned on X.

 For example, suppose our world of data tuples is confined to customers described by the attributes age and income, respectively, and that X is a 35-year-old customer with an income of $40,000.

 Suppose that H is the hypothesis that our customer will buy a computer.
Then P(H|X) reflects the probability that customer X will buy a computer given that we know the customer's age and income.

Prior probability:

 P(H) is the prior probability, or a priori probability, of H.

 For example, this is the probability that any given customer will buy a computer, regardless of age, income, or any other information, for that matter.

 The posterior probability, P(H|X), is based on more information (e.g., customer information) than the prior probability, P(H), which is

 Bayes' theorem is useful in that it provides a way of calculating the posterior probability, P(H|X), from P(H), P(X|H), and P(X).

Eg.

- Here there are 14 training examples of the target concept PlayTennis, where each day is described by the attributes Outlook, Temperature, Humidity, and Wind.
- Here we use the naive Bayes classifier and the training data from this table to classify the following novel instance:
-

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$$

$$\begin{aligned} v_{NB} &= \underset{v_j \in \{yes, no\}}{\operatorname{argmax}} \; P(v_j) \prod_i \; P(a_i|v_j) \\ &= \underset{v_j \in \{yes, no\}}{\operatorname{argmax}} \; P(v_j) \quad P(Outlook = sunny|v_j) P(Temperature = cool|v_j) \\ &\qquad\qquad\qquad \cdot P(Humidity = high|v_j) P(Wind = strong|v_j \end{aligned}$$

$$P(yes)\; P(sunny|yes)\; P(cool|yes)\; P(high|yes)\; P(strong|yes) = .0053$$

$$P(no)\; P(sunny|no)\; P(cool|no)\; P(high|no)\; P(strong|no) \quad = .0206$$

- Thus, the naive Bayes classifier assigns the target value ***PlayTennis = no*** *to this* new instance.

8. Write a note on data mining tools DB Miner, WEKA and DTREG.

DB Miner

 DBMiner, a data mining system for interactive mining of multiple-level knowledge in large relational databases, has been developed based on our years-of-research.

 The system implements a wide spectrum of data mining functions, including generalization, characterization, discrimination, association, classification, and prediction.

 By incorporation of several interesting data mining techniques, including attribute-oriented induction, progressive deepening for mining multiple-level rules, and meta-rule guided knowledge mining, the system provides a user-friendly, interactive data mining environment with good performance.

WEKA

 Weka is a collection of machine learning algorithms for data mining tasks.

 The algorithms can either be applied directly to a dataset or called from your own Java code.

 Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

 It is also well-suited for developing new machine learning schemes.

DTREG

 It is a robust application that is installed easily on any Windows system.

 DTREG reads Comma Separated Value (CSV) data files that are easily created from almost any data source. Once you create your data file, just feed it into DTREG, and let DTREG do all of the work of creating a decision tree, Support Vector Machine, K-Means clustering, Linear Discriminant Function, Linear Regression or Logistic Regression model. Even complex analyses can be set up in minutes.

 Classification and Regression Trees. DTREG can build Classification Trees where the target variable being predicted is categorical and Regression Trees where the target variable is continuous like income or sales volume.

9. Explain the main phases of Data Analytics Life Cycle in detail.

Data Identification

□ The Data Identification stage is dedicated to identifying the datasets required for the analysis project and their sources.

□ Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for.

□ Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.

- The Data analytics lifecycle is designed for Big Data problems and data science projects.
- The cycle is iterative to represent real project.
- To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

Phase 1: Discovery

- The data science team learn and investigate the problem.
- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates initial hypothesis that can be later tested with data.

Phase 2: Data Preparation

- Steps to explore, preprocess, and condition data prior to modeling and analysis.
- It requires the presence of an analytic sandbox (testing environment), the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.
- Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning

- Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- In this phase, data science team develop data sets for training, testing, and production purposes.
- Team builds and executes models based on the work done in the model planning phase.
- Several tools commonly used for this phase are – Matlab, STASTICA.

Phase 4: Model Building

- Team develops datasets for testing, training, and production purposes.
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – R and PL/R, Octave, WEKA.
- Commercial tools – Matlab , STASTICA.

Phase 5: Communication Results

- After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

Phase 6: Operationalize

- The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes.
- Free or open source tools – Octave, WEKA, SQL, MADlib.

10. Discuss the key stakeholders of analytics project.

- There are certain key roles that are required for the complete and fulfilled functioning of the data science team to execute projects on analytics successfully.
- The key roles are seven in number.
- Each key plays a crucial role in developing a successful analytics project. There is no hard and fast rule for considering the listed seven roles, they can be used fewer or more depending on the scope of the project, skills of the participants, and organizational structure.
- **Example –**
  For a small, versatile team, these listed seven roles may be fulfilled by only three to four people but a large project on the contrary may require 20 or more people for fulfilling the listed roles.

Business User :

- The business user is the one who understands the main area of the project and is also basically benefited from the results.
- This user gives advice and consult the team working on the project about the value of the results obtained and how the operations on the outputs are done.
- The business manager, line manager, or deep subject matter expert in the project mains fulfills this role.

- **Project Sponsor :**

- The Project Sponsor is the one who is responsible to initiate the project. Project Sponsor provides the actual requirements for the project and presents the basic business issue.
- He generally provides the funds and measures the degree of value from the final output of the team working on the project.
- This person introduce the prime concern and brooms the desired output.

Project Manager :

- This person ensures that key milestone and purpose of the project is met on time and of the expected quality.

Business Intelligence Analyst :

- Business Intelligence Analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators (KPIs), key matrix, and business intelligence from a reporting point of view.
- This person generally creates outlook and reports and knows about the data feeds and sources.

Data Scientist :

- Data scientist facilitates with the subject matter expertise for analytical techniques, data modelling, and applying correct analytical techniques for a given business issues.
- He ensures overall analytical objectives are met.
- Data scientists outline and apply analytical methods and proceed towards the data available for the concerned project.

Database Administrator (DBA) :

- DBA facilitates and arrange the database environment to support the analytics need of the team working on a project.
- His responsibilities may include providing permission to key databases or tables and making sure that the appropriate security stages are in their correct places related to the data repositories or not.

Data Engineer :

- Data engineer grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.
- The data engineer works jointly with the data scientist to help build data in correct ways for analysis.