BSc.(Information Technology)
(Semester VI)
2018-19

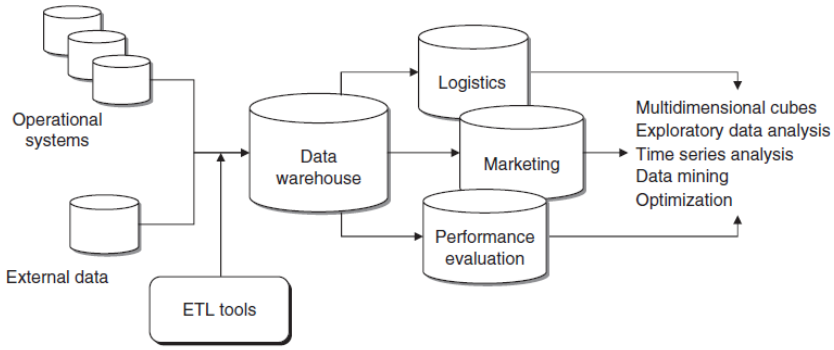
Business Intelligence
(USIT 603 Core)
University Paper Solution

By
Hrishikesh Tendulkar

Question 1

Q1a. What is business intelligence? Explain architecture of the business intelligence

Ans: Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes.



Data sources: In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type. The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers.

Data warehouses and data marts: Using extraction and transformation tools known as extract, transform, load (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses.

Data exploration: The tools for performing a passive business intelligence analysis, which consist of query and reporting systems, as well as statistical methods. These are referred to as passive methodologies because decision makers are requested to generate prior hypotheses or define data extraction criteria, and then use the analysis tools to find answers and confirm their original insight.

Data mining: Purpose is the extraction of information and knowledge from data. Their purpose is instead to expand the decision makers' knowledge. These include mathematical models for pattern recognition, machine learning and data mining techniques

Optimization: By moving up one level in the pyramid we find optimization models that allow us to determine the best solution out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite.

Decisions: Finally, the top of the pyramid corresponds to the choice and the actual adoption of a specific decision and in some way represents the natural conclusion of the decision-making process. Even when business intelligence methodologies are available and successfully adopted, the choice of a decision pertains to the decision makers, who may also take advantage of informal and unstructured information available to adapt and modify the recommendations and the conclusions achieved through the use of mathematical models.

Q1b. Explain different phases in development business intelligence system

Ans: **Analysis:** This preliminary phase is generally conducted through a series of interviews of knowledge workers performing different roles and activities within the organization. It is necessary to clearly describe the general objectives and priorities of the project, as well as to set out the costs and benefits deriving from the development of the business intelligence system.

Design: The second phase includes two sub-phases and is aimed at deriving a provisional plan of the overall architecture, taking into account any development in the near future and the evolution of the system in the mid term. First, it is necessary to make an assessment of the

existing information infrastructures. Moreover, the main decision-making processes that are to be supported by the business intelligence system should be examined, in order to adequately determine the information requirements. Later on, using classical project management methodologies, the project plan will be laid down, identifying development phases, priorities, expected execution times and costs, together with the required roles and resources.

Planning: The planning stage includes a sub-phase where the functions of the business intelligence system are defined and described in greater detail. Subsequently, existing data as well as other data that might be retrieved externally are assessed. This allows the information structures of the business intelligence architecture, which consist of a central data warehouse and possibly some satellite data marts, to be designed. Simultaneously with the recognition of the available data, the mathematical models to be adopted should be defined, ensuring the availability of the data required to feed each model and verifying that the efficiency of the algorithms to be utilized will be adequate for the magnitude of the resulting problems. Finally, it is appropriate to create a system prototype, at low cost and with limited capabilities, in order to uncover beforehand any discrepancy between actual needs and project specifications.

Implementation and control: The last phase consists of five main sub-phases. First, the data warehouse and each specific data mart are developed. These represent the information infrastructures that will feed the business intelligence system. Moreover, ETL procedures are set out to extract and transform the data existing in the primary sources, loading them into the data warehouse and the data marts. The next step is aimed at developing the core business intelligence applications that allow the planned analyses to be carried out. Finally, the system is released for test and usage.

Q 1c. What is decision support system (DSS)? What are the factors that affect the degree of success of the DSS?

Ans: A decision support system (DSS) is an interactive computer-based application that combines data and mathematical models to help decision makers solve complex problems faced in managing the public and private enterprises and organizations. Factors that affect the degree of success of DSS

Economic: Economic factors are the most influential in decision-making processes, and are often aimed at the minimization of costs or the maximization of profits. For example, an annual logistic plan may be preferred over alternative plans if it achieves a reduction in total costs.

Technical: Options that are not technically feasible must be discarded. For instance, a production plan that exceeds the maximum capacity of a plant cannot be regarded as a feasible option.

Legal: Legal rationality implies that before adopting any choice the decision makers should verify whether it is compatible with the legislation in force within the application domain.

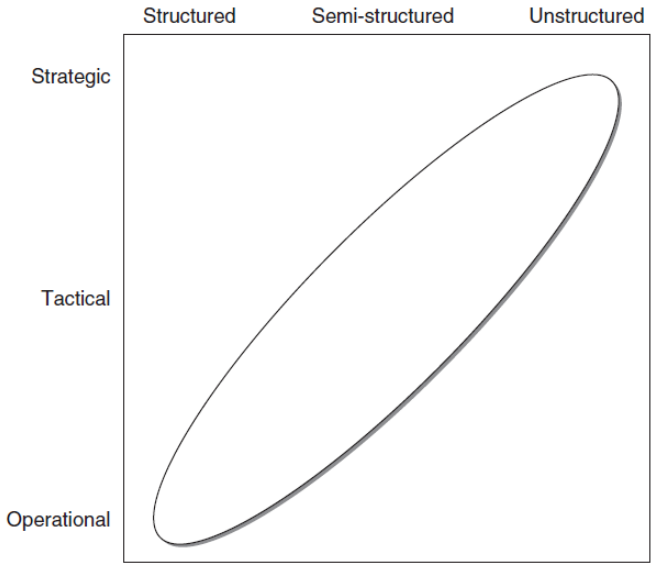
Ethical: Besides being compliant with the law, a decision should abide by the ethical principles and social rules of the community to which the system belongs.

Procedural: A decision may be considered ideal from an economic, legal and social standpoint, but it may be unworkable due to cultural limitations of the organization in terms of prevailing procedures and common practice.

Political: The decision maker must also assess the political consequences of a specific decision among individuals, departments and organizations.

Q1d. Explain classification of decisions according to their nature and scope

Ans:



Structured decisions: A decision is structured if it is based on a well-defined and recurring decision-making procedure. In most cases structured decisions can be traced back to an algorithm, which may be more or less explicit for decision makers, and are therefore better suited for automation. More specifically, we have a structured decision if input flows, output flows and the transformations performed by the system can be clearly described in the three phases of intelligence, design and choice.

Unstructured decisions: A decision is said to be unstructured if the three phases of intelligence, design and choice are also unstructured. This means that for each phase there is at least one element in the system (input flows, output flows and the transformation processes) that cannot be described in detail and reduced to a predefined sequence of steps. Such an event may occur when a decision-making process is faced for the first time or if it happens very seldom. In this type of decisions the role of knowledge workers is fundamental, and business intelligence systems may provide support to decision makers through timely and versatile access to information.

Semi-structured decisions: A decision is semi-structured when some phases are structured and others are not. Most decisions faced by knowledge workers in managing public or private enterprises or organizations are semi-structured.

Depending on their scope, decisions can be classified as strategic, tactical and operational.

Strategic decisions: Decisions are strategic when they affect the entire organization or at least a substantial part of it for a long period of time. Strategic decisions strongly influence the general objectives and policies of an enterprise. As a consequence, strategic decisions are taken at a higher organizational level, usually by the company top management.

Tactical decisions: Tactical decisions affect only parts of an enterprise and are usually restricted to a single department. The time span is limited to a medium-term horizon, typically up to a year. Tactical decisions place themselves within the context determined by strategic decisions. In a company hierarchy, tactical decisions are made by middle managers, such as the heads of the company departments.

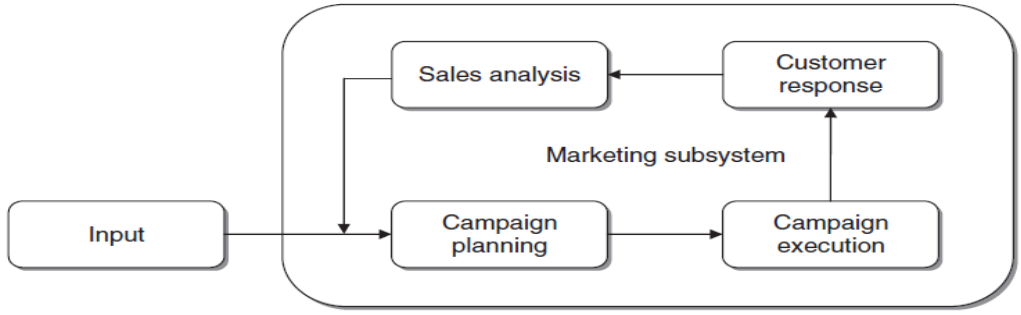
Operational decisions: Operational decisions refer to specific activities carried out within an organization and have a modest impact on the future. Operational decisions are framed within the elements and conditions determined by strategic and tactical decisions. Therefore, they are usually made at a lower organizational level, by knowledge workers responsible for a single activity or task such as sub-department heads, workshop foremen, back-office heads.

Q1e. Define system. Explain closed cycle and open cycle system with suitable example

Ans: System is made up of a set of components that are in some way connected to each other so as to provide a single collective result and a common purpose. Every system is characterized by boundaries that separate its internal components from the external environment.

Open Cycle System - A system is said to be open if its boundaries can be crossed in both directions by flows of materials and information. In general terms, any given system receives specific input flows, carries out an internal transformation process and generates observable output flows.

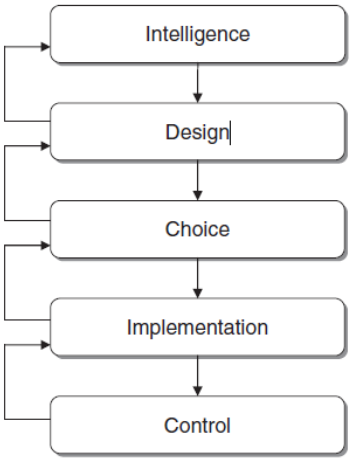
Closed Cycle System – Systems that are able to modify their own output flows based on feedback are called *closed cycle systems*



For example, the closed cycle system outlined in the above figure. It describes the development of a sequence of marketing campaigns. The sales results for each campaign are gathered and become available as feedback input so as to design subsequent marketing promotions.

Q1f. Describe different phases in the development of a decision support system (DSS)

Ans:



Phases of decision – making process:

Intelligence: In the intelligence phase the task of the decision maker is to identify, circumscribe and explicitly define the problem that emerges in the system under study. The analysis of the context and all the available information may allow decision makers to quickly grasp the signals and symptoms pointing to a corrective action to improve the system performance.

Design: In the design phase actions aimed at solving the identified problem should be developed and planned. At this level, the experience and creativity of the decision makers play

a critical role, as they are asked to devise viable solutions that ultimately allow the intended purpose to be achieved.

Choice: Once the alternative actions have been identified, it is necessary to evaluate them on the basis of the performance criteria deemed significant. Mathematical models and the corresponding solution methods usually play a valuable role during the choice phase. For example, optimization models and methods allow the best solution to be found in very complex situations involving countless or even infinite feasible solutions. On the other hand, decision trees can be used to handle decision-making processes influenced by stochastic events.

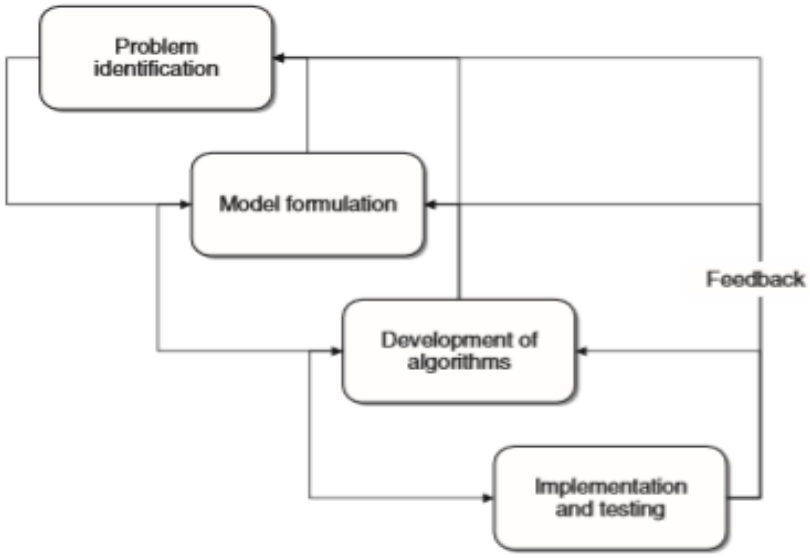
Implementation: When the best alternative has been selected by the decision maker, it is transformed into actions by means of an implementation plan. This involves assigning responsibilities and roles to all those involved into the action plan.

Control: Once the action has been implemented, it is finally necessary to verify and check that the original expectations have been satisfied and the effects of the action match the original intentions. In particular, the differences between the values of the performance indicators identified in the choice phase and the values actually observed at the end of the implementation plan should be measured.

Question 2

Q2a. What are the phases in the development of mathematical models for decision making?

Ans: It is possible to break down the development of a mathematical model for decision making into four primary phases, shown in the following figure. The figure also includes a feedback mechanism which takes into account the possibility of changes and revisions of the model.



Problem identification - First of all, the problem at hand must be correctly identified. The observed critical symptoms must be analysed and interpreted in order to formulate hypotheses for investigation.

Model formulation – Once the problem to be analyzed has been properly identified, effort should be directed toward defining an appropriate mathematical model to represent the system. A number of factors affect and influence the choice of model, such as the time horizon, the decision variables, the evaluation criteria, the numerical parameters and the mathematical relationships.

1. **Time horizon** – Usually a model includes a temporal dimension.
2. **Evaluation criteria**: Appropriate measurable performance indicators should be defined in order to establish a criterion for the evaluation and comparison of the alternative decisions.

Development of algorithms

Once a mathematical model has been defined, one will naturally wish to proceed with its solution to assess decisions and to select the best alternative.

Implementation and test

When a model is fully developed, then it is finally implemented, tested and utilized in the application domain. It is also necessary that the correctness of the data and the numerical parameters entered in the model be preliminarily assessed. These data usually come from a data warehouse or a data mart previously set up.

Q2b. Explain the divisions of mathematical models according to their characteristics, probabilistic nature, temporal dimension.

Ans: According to their characteristics, models can be divided into

Iconic: An iconic model is a material representation of a real system, whose behavior is imitated for the purpose of the analysis. A miniaturized model of a new city neighborhood is an example of iconic model.

Analogical: An analogical model is also a material representation, although it imitates the real behavior by analogy rather than by replication.

Symbolic: A symbolic model, such as a mathematical model, is an abstract representation of a real system. It is intended to describe the behaviour of the system through a series of symbolic variables, numerical parameters and mathematical relationships.

A further relevant distinction concerns the probabilistic nature of models, which can be

Stochastic: In a stochastic model some input information represents random events and is therefore characterized by a probability distribution, which in turn can be assigned or unknown. Predictive models, waiting line models are examples of stochastic models.

Deterministic: A model is called deterministic when all input data are supposed to be known a priori and with certainty. Since this assumption is rarely fulfilled in real systems, one resorts to deterministic models when the problem at hand is sufficiently complex and any stochastic elements are of limited relevance.

A further distinction concerns the temporal dimension in a mathematical model, which can be

Static: Static models consider a given system and the related decision-making process within one single temporal stage.

Dynamic: Dynamic models consider a given system through several temporal stages, corresponding to a sequence of decisions. In many instances the temporal dimension is subdivided into discrete intervals of a previously fixed span: minutes, hours, days, weeks, months and years are examples of discrete subdivisions of the time axis.

Q2c. What is data mining? List the real life applications of data mining

Ans: The term data mining indicates the process of exploration and analysis of a dataset, usually of large size, in order to find regular patterns, to extract relevant knowledge and to obtain meaningful recurring rules. Data mining plays an ever-growing role in both theoretical studies and applications. Data mining methodologies can be applied to a variety of domains, from marketing and manufacturing process control to the study of risk factors in medical diagnosis, from the evaluation of the effectiveness of new drugs to fraud detection.

Relational marketing – Data mining applications in the field of relational marketing, have significantly contributed to the increase in the popularity of these methodologies. Some

relevant applications within relational marketing are:

- Identification of customer segments that are most likely to respond to targeted marketing campaigns, such as cross-selling and up-selling
- Identification of target customer segments for retention campaigns
- Prediction of the rate of positive responses to marketing campaigns
- Interpretation and understanding of the buying behavior of the customers
- Analysis of the products jointly purchased by customers, known as market basket analysis.

Fraud detection – Fraud detection is another relevant field of application of data mining. Fraud may affect different industries such as telephony, insurance (false claims) and banking (illegal use of credit cards and bank checks; illegal monetary transactions).

Risk evaluation – The purpose of risk analysis is to estimate the risk connected with future decisions, which often assume a dichotomous form. For example, using the past observations available, a bank may develop a predictive model to establish if it is appropriate to grant a monetary loan or a home loan, based on the characteristics of the applicant.

Text mining – Data mining can be applied to different kinds of texts, which represent unstructured data, in order to classify articles, books, documents, emails and web pages. Examples are web search engines or the automatic classification of press releases for storing purposes. Other text mining applications include the generation of filters for email messages and newsgroups.

Image recognition – The treatment and classification of digital images, both static and dynamic, is an exciting subject for both its theoretical interest and the great number of applications it offers. It is useful to recognize written characters, compare and identify human faces, apply correction filters to photographic equipment and detect suspicious behaviors through surveillance video cameras.

Web mining – Web mining applications, are intended for the analysis of so-called clickstreams – the sequences of pages visited and the choices made by a web surfer. They may prove useful for the analysis of e-commerce sites, in offering flexible and customized pages to surfers, in caching the most popular pages or in evaluating the effectiveness of an e-learning training course.

Medical diagnosis – Learning models are an invaluable tool within the medical field for the early detection of diseases using clinical test results. Image analysis for diagnostic purpose is another field of investigation that is currently burgeoning.

Q2d. Explain categorical and numerical attributes with proper example.

Ans: **Categorical** – Categorical attributes assume a finite number of distinct values, in most cases limited to less than a hundred, representing a qualitative property of an entity to which they refer. Examples of categorical attributes are the province of residence of an individual (which takes as values a series of names, which in turn may be represented by integers) or whether a customer has abandoned her service provider (expressed by the value 1) or remained loyal to it (expressed by the value 0). Arithmetic operations cannot be applied to categorical attributes even when integer numbers expresses the coding of their values.

Counts – Counts are categorical attributes in relation to which a specific property can be true or false. These attributes can therefore be represented using Boolean variables {true, false} or binary variables {0,1}. For example, a bank's customers may or may not be holders of a credit card issued by the bank.

Nominal – Nominal attributes are categorical attributes without a natural ordering, such as the province of residence.

Ordinal – Ordinal attributes, such as education level, are categorical attributes that lend themselves to a natural ordering but for which it makes no sense to calculate differences or

ratios between the values.

Numerical – Numerical attributes assume a finite or infinite number of values and lend themselves to subtraction or division operations. For example, the amount of outgoing phone calls during a month for a generic customer represents a numerical variable. Regarding two customers A and B making phone calls in a week for 27 and 36 respectively, it makes sense to claim that the difference between the amounts spent by the two customers is equal to 9 and that A has spent three fourths of the amount spent by B.

Discrete – Discrete attributes are numerical attributes that assume a finite number or a countable infinity of values.

Continuous – Continuous attributes are numerical attributes that assume an uncountable infinity of values.

To represent a generic dataset D, we will denote by m the number of observations, or rows, in the two-dimensional table containing the data and by n the number of attributes, or columns. Furthermore, we will denote by

$$X = [x_{ij}],\ i \in M = \{1, 2, \ldots, m\},\ j \in N = \{1, 2, \ldots, n\},$$

the matrix of dimensions m × n that corresponds to the entries in the dataset D. We will write

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$$

$$a_j = (x_{1j}, x_{2j}, \ldots, x_{mj})$$

for the n-dimensional row vector associated with the ith record of the dataset and the m-dimensional column vector representing the j^{th} attribute in D, respectively

Q2e. Differentiate between supervised and unsupervised learning.

Ans: Data mining activities can be subdivided into a few major categories, based on the tasks and the objectives of the analysis.

Supervised learning – In a supervised (or direct) learning analysis, a target attribute either represents the class to which each record belongs, For example on loyalty in the mobile phone industry, a measurable quantity, such as the total value of calls that will be placed by a customer in a future period. As a second example of the supervised perspective, consider an investment management company wishing to predict the balance sheet of its customers based on their demographic characteristics and past investment transactions. Supervised learning processes are therefore oriented toward prediction and interpretation with respect to a target attribute.

Unsupervised learning – Unsupervised (or indirect) learning analyses are not guided by a target attribute. Therefore, data mining tasks in this case are aimed at discovering recurring patterns and affinities in the dataset. As an example, consider an investment management company wishing to identify clusters of customers who exhibit homogeneous investment behaviour, based on data on past transactions. In most unsupervised learning analyses, one is interested in identifying clusters of records that are similar within each cluster and different from members of other clusters.

Q2f. Explain the following normalization techniques:

i) Decimal scaling ii) Min-max

Ans: **Decimal Scaling** – Decimal scaling is based on the transformation

$$x'_{ij} = \frac{x_{ij}}{10^h},$$

where *h* is a given parameter which determines the scaling intensity. In practice, decimal scaling corresponds to shifting the decimal point by *h* positions toward the left. In general, *h* is fixed at a value that gives transformed values in the range [−1, 1].

Min-Max. Min-max standardization is achieved through the transformation

$$x'_{ij} = \frac{}{x_{\max,j} - x_{\min,j}}(x_{\max,j} - x_{\min,j}) + x'_{\min,j},$$

Where

$$x_{\min,j} = \min_i x_{ij}, \quad x_{\max,j} = \max_i x_{ij},$$

are the minimum and maximum values of the attribute $\mathbf{a}_j$ before transformation, while $x'_{\min,j}$ and $x'_{\max,j}$ are the minimum and maximum values that we wish to obtain after transformation.

In general, the extreme values of the range are defined so that

$$x'_{\min,j} = -1 \text{ and } x'_{\max,j} = 1 \text{ or } x'_{\min,j} = 0 \text{ and } x'_{\max,j} = 1.$$

z-index – z-index based standardization uses the transformation

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j},$$

Where $\bar{\mu}_j$ and $\bar{\sigma}_j$ are respectively the sample mean and sample standard deviation of the attribute $\mathbf{a}_j$. If the distribution of values of the attribute $\mathbf{a}_j$ is roughly normal, the *z*-index based transformation generates values that are almost certainly within the range *(*−3*,* 3*)*.

Question 3

Q3a. What are the criteria used to evaluate classification methods?

Ans: Classification methods can be evaluated based on several criteria, as follows

Accuracy. Evaluating the accuracy of a classification model is crucial for two main reasons. First, the accuracy of a model is an indicator of its ability to predict the target class for future observations. Based on their accuracy values, it is also possible to compare different models in order to select the classifier associated with the best performance

Speed. Some methods require shorter computation times than others and can handle larger problems. However, classification methods characterized by longer computation times may be applied to a small-size training set obtained from a large number of observations by means of random sampling schemes. It is not uncommon to obtain more accurate classification rules in this way.

Robustness. A classification method is *robust* if the classification rules generated, as well as the corresponding accuracy, do not vary significantly as the choice of the training set and the test set varies, and if it is able to handle missing data and outliers.

Scalability. The scalability of a classifier refers to its ability to learn from large datasets, and it is inevitably related to its computation speed. Therefore, the remarks made in connection with sampling techniques for data reduction, which often result in rules having better generalization capability, also apply in this case.

Interpretability. If the aim of a classification analysis is to interpret as well as predict, then the rules generated should be simple and easily understood by knowledge workers and experts in the application domain.

Q3b. Explain top-down induction of decision tree. Examine the components of the top-down induction of decision trees procedure

Ans: The development of a classification tree corresponds to the training phase of the model and is regulated by a recursive procedure of heuristic nature, based on a divide-and-conquer partitioning scheme referred to as top-down induction of decision trees

Procedure For Top-down induction of decision trees

1. In the initialization phase, each observation is placed in the root node of the tree. The root is included in the list *L* of active nodes.
2. If the list *L* is empty the procedure is stopped, otherwise a node *J* belonging to the list *L* is selected, is removed from the list and is used as the node for analysis.
3. The optimal rule to split the observations contained in *J* is then determined, based on an appropriate preset criterion. The splitting rule generated in this way is then applied, and descendant nodes are constructed by subdividing the observations contained in *J* . For each descendant node the conditions for stopping the subdivision are verified. If these are met, node *J* becomes a leaf, to which the target class is assigned according to the majority of the observations contained in *J* . Otherwise, the descendant nodes are added to the list *L*. Finally, step 2 is repeated.

Components of the top-down induction of decision trees procedure.

Splitting rules. For each node of the tree it is necessary to specify the criteria used to identify the optimal rule for splitting the observations and for creating the descendant nodes. As shown in the next section, there are several alternative criteria, which differ in the number of descendants, the number of attributes and the evaluation metrics.

Stopping criteria. At each node of the tree different *stopping* criteria are applied to establish whether the development should be continued recursively or the node should be considered as a leaf. In this case too, various criteria have been proposed, which result in quite different topologies of the generated trees, all other elements being equal.

Pruning criteria. Finally, it is appropriate to apply a few *pruning* criteria, first to avoid excessive growth of the tree during the development phase (*pre-pruning*), and then to reduce the number of nodes after the tree has been generated (*post-pruning*).

Q3c. Write a short note on naïve Bayesian classifiers

Ans: Naive Bayesian classifiers are based on the assumption that the explanatory variables are conditionally independent given the target class. This hypothesis allows us to express the probability *P(***x***|y)* as

$$P(\mathbf{x}|y) = P(x_1|y) \times P(x_2|y) \times \cdots \times P(x_n|y) = \prod_{j=1}^{n} P(x_j|y)$$

The probabilities *P(xj |y)*, *j* ∈ *N*, can be estimated using the examples from the training set, depending on the nature of the attribute considered

Categorical or discrete numerical attributes. For a categorical or discrete numerical attribute **a***j* which may take the values {*rj*1, *rj*2, . . . , *rjKj*}, the probability *P(xj |y) = P(xj = rjk|y = vh)* is evaluated as the ratio between the number *sjhk* of instances of class *vh* for which the attribute **a***j* takes the value *rjk*, and the total number *mh* of instances of class *vh* in the dataset *D*, that is,

$$P(x_j|y) = P(x_j = r_{jk}|y = v_h) = \frac{s_{jhk}}{m_h}$$

Numerical attributes. For a numerical attribute **a***j* , the probability *P(xj |y)* is estimated assuming that the examples follow a given distribution. For example, one may consider a Gaussian density function, for which

$$P(x_j | y = v_h) = \frac{1}{\sqrt{2\pi}\sigma_{jh}} e^{-\frac{(x_j - \mu_{jh})^2}{2\sigma_{jh}^2}},$$

Where *μjh* and *σjh* respectively denote the mean and standard deviation of the variable *Xj* for the examples of class *vh*, and may be estimated based on the examples contained in *D*.

Q3d. Write k-means algorithm for clustering

Ans The *K-means* algorithm receives as input a dataset *D*, a number *K* of clusters to be generated and a function dist*(***x***i , **x***k)* that expresses the inhomogeneity between each pair of observations, or equivalently the matrix **D** of distances between observations. Given a cluster *Ch*, *h* = 1, 2, . . . *K*, the *centroid* of the cluster is defined as the point **z***h* having coordinates equal to the mean value of each attribute for the observations belonging to that cluster, that is,

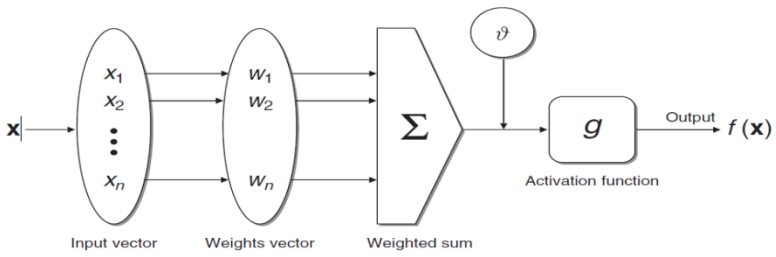
$$z_{hj} = \frac{\sum_{\mathbf{x}_i \in C_h} x_{ij}}{\text{card}\{C_h\}}$$

***K*-means algorithm**

1. During the initialization phase, *K* observations are arbitrarily chosen in *D* as the centroids of the clusters.
2. Each observation is iteratively assigned to the cluster whose centroid is the most similar to the observation, in the sense that it minimizes the distance from the record.
3. If no observation is assigned to a different cluster with respect to the previous iteration, the algorithm stops.
4. For each cluster, the new centroid is computed as the mean of the values of the observations belonging to the cluster, and then the algorithm returns to step 2.

Q3e. Explain the 'Rosenblatt perceptron' form of neural network with diagram.

Ans: The *perceptron*, shown in figure is the simplest form of neural network and corresponds to a single neuron that receives as input the values *(x*1*, x*2*, . . . , xn)* along the incoming connections, and returns an output value *f (***x***)*. The input values coincide with the values of the explanatory attributes, while the output value determines the prediction of the response variable *y*. Each of the *n* input connections is associated with a weight *wj* . An activation function *g* and a constant *ϑ*, called the *distortion*, are also assigned. Suppose that the values of the weights and the distortion have already been determined during the training phase. The prediction for a new observation **x** is then derived by performing the following steps.



For binary classification problems it is possible to give a geometrical interpretation of the prediction obtained using a Rosenblatt perceptron. Indeed, if we place the *m* observations of the training dataset in the space R*n*, the weighted linear combination calculated for **x***i* expresses the slack between the observation and the hyperplane:

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n - \vartheta = \mathbf{w}'\mathbf{x} - \vartheta.$$

The purpose of the activation function $g(\cdot) = \text{sgn}(\cdot)$ is therefore to establish if the point associated with the example $\mathbf{x}i$ is placed in the lower or upper halfspace with respect to the separating hyperplane. Hence, the Rosenblatt perceptron corresponds to a linear separation of the observations based on the target class. The aim of the iterative procedure is therefore to determine the coefficients of the separating hyperplane

Q3f. Write a short note on confusion matrix.

Ans: The accuracy measurement methods described above are not always adequate for discriminating among models, and in some instances they may even yield paradoxical results it is useful to resort to decision tables, usually called confusion matrices, which for the sake of simplicity we will only describe in connection with binary classification, though they can be easily extended to multicategory classification. Let us assume that we wish to analyze a binary classification problem where the values taken by the target class are {−1, 1}. We can then consider a 2 × 2 matrix whose rows correspond to the observed values and whose columns are associated with the values predicted using a classification model, The elements of the confusion matrix have the following meanings: p is the number of correct predictions for the negative examples, called true negatives; u is the number of incorrect predictions for the positive examples, called false negatives; q is the number of incorrect predictions for the negative examples, called false positives; and v is the number of correct predictions for the positive examples, called true positives. Using these elements, further indicators useful for validating a classification algorithm can be defined

		predictions		
		−1 (negative)	+1 (positive)	total
examples	−1 (negative)	p	q	$p+q$
	+1 (positive)	u	v	$u+v$
	total	$p+u$	$q+v$	m

Accuracy. The accuracy of a classifier may be expressed as

$$\text{acc} = \frac{p+v}{p+q+u+v} = \frac{p+v}{m}$$

True negatives rate. The true negatives rate is defined as

$$\text{tn} = \frac{p}{p+q}.$$

False negatives rate. The false negatives rate is defined as

$$\text{fn} = \frac{u}{u+v}.$$

False positives rate. The false positives rate is defined as

$$\text{fp} = \frac{q}{p+q}$$

True positives rate. The true positives rate, also known as *recall*, is defined as

$$\text{tp} = \frac{v}{u+v}.$$

Precision. The precision is the proportion of correctly classified positive examples, and is given by

$$\text{prc} = \frac{v}{q + v}$$

Geometric mean. The geometric mean is defined as

$$\text{gm} = \sqrt{\text{tp} \times \text{prc}},$$

and sometimes also as

$$\text{gm} = \sqrt{\text{tp} \times \text{tn}}.$$

***F*-measure.** The *F*-measure is defined as

$$F = \frac{(\beta^2 - 1)\,\text{tp} \times \text{prc}}{\beta^2\,\text{prc} + \text{tp}}$$

where $\beta \in [0,\infty)$ regulates the relative importance of the precision with respect to the true positives rate. The *F*-measure is also equal to 0 if all the predictions are incorrect

Question 4

Q4a. Write a short note on market basket analysis

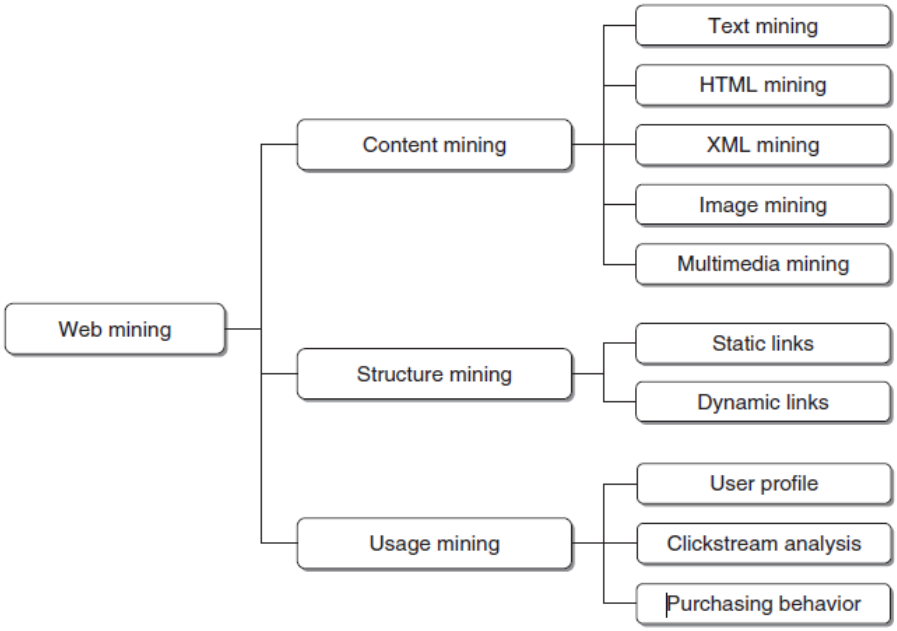
Ans: The purpose of *market basket analysis* is to gain insight from the purchases made by customers in order to extract useful knowledge to plan marketing actions. It is mostly used to analyze purchases in the retail industry and in e-commerce activities, and is generally amenable to unsupervised learning problems. It may also be applied in other domains to analyze the purchases made using credit cards, the complementary services activated by mobile or fixed telephone customers, the policies or the checking accounts acquired by a same household. The data used for this purpose mostly refer to purchase transactions, and can be associated with the time dimension if the purchaser can be tracked through a loyalty card or the issue of an invoice. Each transaction consists of a list of purchased items. This list is called a *basket*, just like the baskets available at retail points of sale. If transactions cannot be connected to one another, say because the purchaser is unknown, one may then apply association rules, The rules extracted in this way can then be used to support different decision-making processes, such as assigning the location of the items on the shelves, determining the layout of a point of sale, identifying which items should be included in promotional flyers, advertisements or coupons distributed to customers. If customers are individually identified and traced, besides the above techniques it is also possible to develop further analyses that take into account the time dimension of the purchases

Q4b. What is use of web mining methods? What are the different purposes of web mining?

Ans: It is natural to use *web mining* methods in order to analyze data on the activities carried out by the visitors to a website. Web mining methods are mostly used for three main purposes, as shown in the figure below: *content mining*, *structure mining* and *usage mining*.

Content mining. Content mining involves the analysis of the content of web pages to extract useful information. Search engines primarily perform content mining activities to provide the links deemed interesting in relation to keywords supplied by users. Content mining methods can be traced back to data mining problems for the analysis of texts, both in free format or

HTML and XML formats, images and multimedia content.



Structure mining. The aim of this type of analysis is to explore and understand the topological structure of the web. Using the links presented in the various pages, it is possible to create graphs where the nodes correspond to the web pages and the oriented arcs are associated with links to other pages. Results and algorithms from graph theory are used to characterize the structure of the web, that is, to identify areas with a higher density of connections, areas disconnected from others and maximal *cliques*, which are groups of pages with reciprocal links. In this way, it is possible to pinpoint the most popular sites, or to measure the distance between two sites, expressed in terms of the lowest number of arcs along the paths that connect them in the links graph. For example, a page whose purpose is to direct navigation on the site should be viewed by each user only briefly. Should this not be the case, the page has a problem due to a possible ambiguity in the articulation of the links offered.

Usage mining. Analyses aimed at *usage mining* are certainly the most relevant from a relational marketing standpoint, since they explore the paths followed by navigators and their behaviors during a visit to a company website. Methods for the extraction of association rules are useful in obtaining correlations between the different pages visited during a session. In some instances, it is possible to identify a visitor and recognize her during subsequent sessions. This happens if an identification key is required to access a web page, or if a cookie-enabling mechanism is used to keep track of the sequence of visits. Sequential association rules or time series models can be used to analyze the data on the use of a site according to a temporal dynamic. Usage mining analysis is mostly concerned with *clickstreams* – the sequences of pages visited during a given session. For E-commerce sites, information on the purchase behavior of a visitor is also available

Q4c. Explain "tactical planning" optimization model for logistics planning

Ans: The aim of tactical planning is to determine the production volumes for each product over the *T* periods included in the medium-term planning horizon in such a way as to satisfy the given demand and capacity limits for a single resource, and also to minimize the total cost, defined as the um of manufacturing production costs and inventory costs. We therefore

consider the decision variables

Pit = units of product *i* to be manufactured in period *t,*

Iit = units of product *i* in inventory at the end of period *t,*

and the parameters

dit = demand for product *i* in period *t,*

cit = unit manufacturing cost for product *i* in period *t,*

hit = unit inventory cost for product *i* in period *t,*

ei = capacity absorption to manufacture a unit of product *i,*

bt = capacity available in period *t.*

The resulting optimization problem is formulated as follows:

$$\begin{aligned} \min \quad & \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} (c_{it}P_{it} + h_{it}I_{it}) \\ \text{s.to} \quad & P_{it} + I_{i,t-1} - I_{it} = d_{it}, \quad i \in \mathcal{I}, t \in \mathcal{T}, \\ & \sum_{i\in\mathcal{I}} e_i P_{it} \le b_t, \qquad t \in \mathcal{T}, \\ & P_{it}, I_{it} \ge 0, \qquad i \in \mathcal{I}, t \in \mathcal{T}. \end{aligned}$$

Q4d. Explain the Charnes-Cooper-Rhodes (CCR) model

Ans: Using data envelopment analysis, the choice of the optimal system of weights for a generic DMU*j* involves solving a mathematical optimization model whose decision variables are represented by the weights *ur* , *r* ∈ *K*, and *vi* , *i* ∈ *H*, associated with each output and input. The best-known of which is probably the Charnes–Cooper–Rhodes (CCR) model. The CCR model formulated for DMU*j* takes the form

$$\begin{aligned} \max \quad & \vartheta = \frac{\sum_{r\in\mathcal{K}} u_r y_{rj}}{\sum_{i\in\mathcal{H}} v_i x_{ij}}, \\ \text{s.to} \quad & \frac{\sum_{r\in\mathcal{K}} u_r y_{rj}}{\sum_{i\in\mathcal{H}} v_i x_{ij}} \le 1, \qquad j \in \mathcal{N}, \\ & u_r, v_i \ge 0, \qquad r \in \mathcal{K}, i \in \mathcal{H}. \end{aligned}$$

The objective function involves the maximization of the efficiency measure for DMU*j*..Constraints require that the efficiency values of all the units, calculated by means of the weights system for the unit being examined, be lower than one. Finally, conditions guarantee that the weights associated with the inputs and the outputs are non-negative. In place of these conditions, sometimes the constraints *ur, vi* ≥ *δ*, *r* ∈ *K*, *i* ∈ *H* may be applied, where *δ* > 0, preventing the unit from assigning a null weight to an input or output. This condition leads to an alternative optimization problem, the *input-oriented* CCR model, where the objective function consists of the maximization of the weighted sum of the outputs

$$\begin{aligned} \max \quad & \vartheta = \sum_{r\in\mathcal{K}} u_r y_{rj}, \\ \text{s.to} \quad & \sum_{i\in\mathcal{H}} v_i x_{ij} = 1, \\ & \sum_{r\in\mathcal{K}} u_r y_{rj} - \sum_{i\in\mathcal{H}} v_i x_{ij} \le 0, \qquad j \in \mathcal{N}, \\ & u_r, v_i \ge 0, \qquad r \in \mathcal{K}, i \in \mathcal{H}. \end{aligned}$$

Let ϑ^* be the optimum value of the objective function corresponding to the optimal solution $(\mathbf{v}^*,\mathbf{u}^*)$ of problem. DMU*j* is said to be *efficient* if $\vartheta^* = 1$ and if there exists at least one optimal solution $(\mathbf{v}^*,\mathbf{u}^*)$ such that $\mathbf{v}^* > \mathbf{0}$ and $\mathbf{u}^* > \mathbf{0}$. By solving a similar optimization model for each of the *n* units being compared, one obtains *n* systems of weights. The flexibility enjoyed by the units in choosing the weights represents an undisputed advantage, in that if a unit turns out to be inefficient based on the most favorable system of weights, its inefficiency cannot be traced back to an inappropriate evaluation process.

Q4e. Write a short note on efficient frontier.

Ans: The *efficient frontier*, also known as *production function*, expresses the relationship between the inputs utilized and the outputs produced. It indicates the maximum quantity of outputs that can be obtained from a given combination of inputs. At the same time, it also expresses the minimum quantity of inputs that must be used to achieve a given output level. Hence, the efficient frontier corresponds to *technically efficient* operating methods. The efficient frontier may be empirically obtained based on a set of observations that express the output level obtained by applying a specific combination of input production factors. In the context of data envelopment analysis, the observations correspond to the units being evaluated. Data envelopment analysis forgoes any assumptions on the functional form of the efficient frontier, and is therefore nonparametric in character. It only requires that the units being compared are not placed above the production function, depending on their efficiency value. A possible alternative to the efficient frontier is the regression line that can be obtained based on the available observations. In this case, the units that fall above the regression line may be deemed excellent, and the degree of excellence of each unit could be expressed by its distance from the line. The regression line reflects the average behavior of the units being compared, while the efficient frontier identifies the best behavior, and measures the inefficiency of a unit based on the distance from the frontier itself. The *output-oriented* efficiency θ_j^O *is* defined as the ratio between the quantity of output *yj* actually produced by the unit and the ideal quantity *y* that it should produce in conditions of efficiency:

$$\theta_j^O = \frac{y_j}{y^*}$$

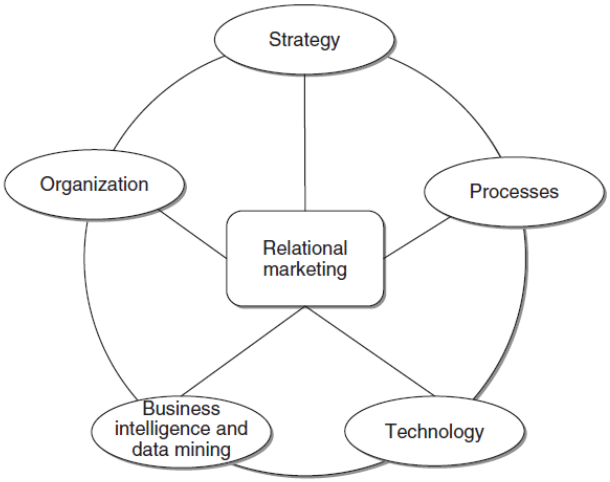
Q4f. What is relational marketing? What are the data mining applications in the field of relational marketing?

Ans: The aim of a *relational marketing* strategy is to initiate, strengthen, intensify and preserve over time the relationships between a company and its stakeholders, represented primarily by its customers, and involves the analysis, planning, execution and evaluation of the activities carried out to pursue these objectives. The reasons for the spread of relational marketing strategies are complex and interconnected. Some of them are listed below

- The increasing concentration of companies in large enterprises and the resulting growth in the number of customers have led to greater complexity in the markets.
- Since the 1980s, the innovation–production–obsolescence cycle has progressively shortened, causing a growth in the number of customized options on the part of customers, and an acceleration of marketing activities by enterprises.
- The increased flow of information and the introduction of e-commerce have enabled global comparisons. Customers can use the Internet to compare features, prices and opinions on products and services offered by the various competitors.
- Customer loyalty has become more uncertain, primarily in the service industries, where often filling out an on-line form is all one has to do to change service provider.

- In many industries a progressive commoditization of products and services is taking place, since their quality is perceived by consumers as equivalent, so that differentiation is mainly due to levels of service.
- The systematic gathering of sales transactions, largely automated in most businesses, has made available large amounts of data that can be transformed into knowledge and then into effective and targeted marketing actions.
- The number of competitors using advanced techniques for the analysis of marketing data has increased.

  Relational marketing strategies revolve around the choices shown in the following figure, which can be effectively summarized as formulating for each segment.



The essential components of relational marketing are a well-designed and correctly fed marketing data mart, a collection of business intelligence and data mining analytical tools, and, most of all, and the cultural education of the decision makers. These tools will enable companies to carry out the required analyses and translate the knowledge acquired into targeted marketing actions

Question 5

Q5a. Define knowledge management. What are data, information and knowledge?

Ans: Knowledge management is the systematic and active management of ideas, information and knowledge residing in an organization's employees. The structuring knowledge enables effective and efficient problem solving, dynamic learning, strategic planning and decision making. KM initiatives focus on identifying knowledge, explicating it in such a way that it can be shared in a formal manner and leveraging its value through reuse.

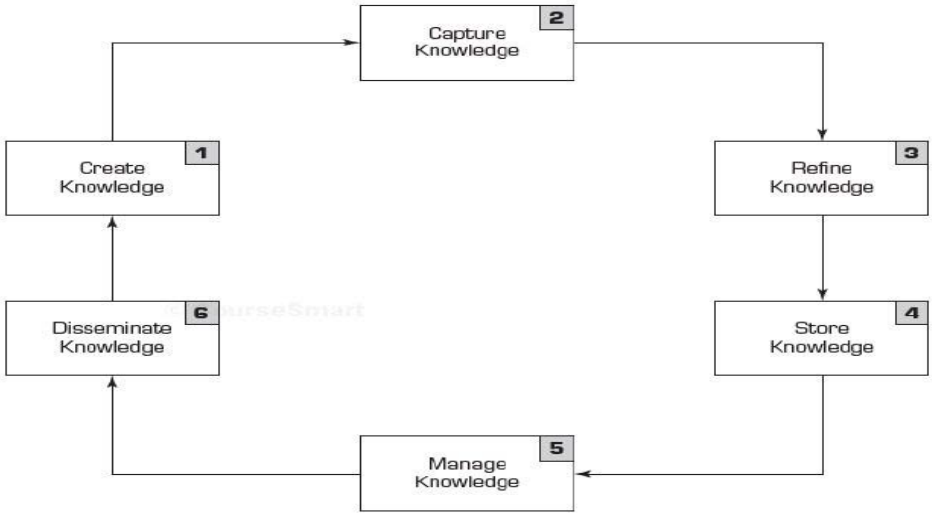
Data – Data are facts, measurements and statistics. Data compiled in the meaningful context provides information. It is expressed in the form of numbers, letters or symbols.

Information – Information is refined form of data which can be used in the process of decision making. The main characteristics of information are accuracy, relevance completeness and availability. It can be communicated in the form of content of a message or through observation

Knowledge – Knowledge is information that is contextual, relevant and actionable. Having knowledge implies that it can be exercised to solve a problem. Knowledge has extra ordinary leverage and increasing returns. It is always needed to be refreshed as it is fragmented as it grows. It is also difficult to estimate the impact of an investment in knowledge and its value of sharing.

Q5b. Describe the knowledge management (KMS) cycle.

Ans:



A functioning KMS cycle is shown in the above figure. The knowledge in good KMS is never finished because the environment changes over time and the knowledge must be updated to reflect the changes. The KMS cycle works as follows

Create Knowledge – Knowledge is created as people determine new ways of doing things or develop somehow. Sometimes external knowledge is brought in.

Capture Knowledge – New knowledge must be identified as valuable and be represented in reasonable way.

Refine Knowledge – New knowledge must be placed in context so that it is actionable. This is where human insights must be captured along with explicit facts.

Store Knowledge – Useful knowledge must be stored in a reasonable format in a knowledge repository so that others in the organization can access it.

Manage Knowledge – A repository must be kept current. It must be reviewed to verify that it is relevant and accurate

Disseminate Knowledge – Knowledge must be made available in a useful format to anyone in the organization who needs it, anywhere and anytime

Q5c.* Describe how AI and intelligent agents support knowledge management. *Relate XML to knowledge management and knowledge portals.

Ans: AI methods and tools are embedded in number of KMS either by vendors or by system developers. AI methods can assist in identifying expertise, eliciting knowledge automatically and semi-automatically interfacing through natural language processing and intelligently searching through intelligent agents. AI methods notably expert systems, neural networks, fuzzy logic, and intelligent agents are used in KMS to do the following

- Assist in and enhance searching knowledge
- Help establish knowledge profiles of individuals and groups
- Help determine the relative importance of knowledge when it is contributed to and accessed from the knowledge repository
- Scan e-mail, documents and databases to perform knowledge discovery, determine, meaningful relationships or induce rules for expert systems
- Identify patterns in data
- Forecast future results by using existing knowledge
- Provide advice directly from knowledge by suing neural networks or expert systems
- Provide a natural language or voice command-driven user interface for a KMS

Extensible Markup Language (XML) enables standardized representations of data structures so that data can be processed appropriately by heterogeneous systems without case by case programming. This method suits e-commerce applications and supply-chain management (SCM) systems that operate enterprise boundaries. XML not only automate processes and reduce paperwork. A portal that uses XML allows the company to communicate better with its customers, linking them in a virtual demand chain where changes in consumer requirements are immediately reflected in production plans. XML also solves the problem of integrating the data from separate sources

Q5d. List and explain characteristics of artificial intelligence.

Ans: Artificial Intelligence is an area of computer science that is concerned with two basic ideas the study of human thought processes and the representation and duplication of those thought processes in machines. Following are the characteristics of AI

Symbolic Processing – AI is a branch of computer science that deals with symbolic, non-algorithmic, methods of problem solving which focuses on numeric versus symbolic characteristics and algorithmic versus heuristics processing

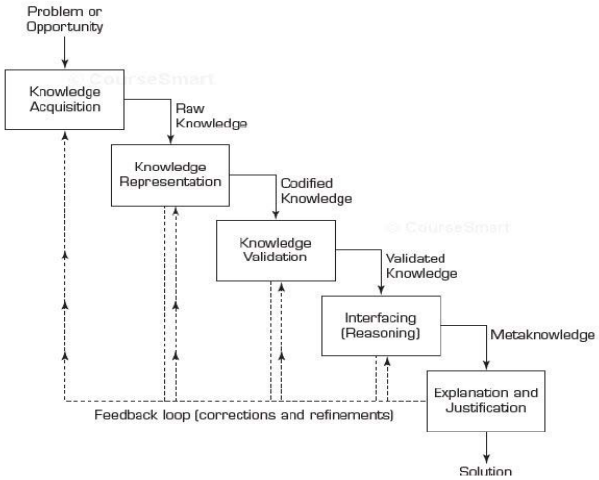
Heuristics – These are intuitive knowledge or rules of thumb learned from experience. AI deals with ways of representing knowledge using symbols with heuristics methods for processing information. Many AI methods use heuristics to reduce the complexity of problem solving.

Inferencing – AI also includes reasoning capabilities that can build higher level knowledge using existing knowledge represented as heuristics in the form of rules. Inferencing is the process of deriving a logical outcome from a given set of facts and rules

Machine Learning – AI systems have simplistic learning capabilities methods called Machine Learning. It allows the computer system to monitor and sense their environmental factors and adjust their behavior to react to changes. It is a scientific discipline that is concerned with the design and the development of the algorithms that allow computers to learn based on data coming from servers or databases

Q5e. What is knowledge engineering? Explain the process of knowledge engineering.

Ans: The collection of intensive activities encompassing the acquisition of knowledge from human experts and conversion of this knowledge into a repository are called knowledge engineering. Knowledge engineering requires cooperation and close communication between the human experts and the knowledge engineer to successfully codify and explicitly represent the rules used to solve a problem within a specific domain. The figure below shows the process of knowledge engineering



Following are the five major activities in knowledge engineering:

Knowledge Acquisition – It involves the acquisition of knowledge from human experts, books, documents, sensors or computer files. The knowledge may be specific to the problem domain or to the problem solving procedures.

Knowledge Representation – Acquired knowledge is organized so that it will be ready for use in activity called knowledge representation. This activity involves preparation of a knowledge map and encoding of knowledge in the knowledge base

Knowledge Validation – It involves validating and verifying the knowledge until its quality is acceptable

Inferencing – This activity involves the design of software to enable the computer to make inferences based on the stored knowledge and specifics of the problem. The system can then provide advice to no expert users

Explanation and Justification – This step involves the design and programming of an explanation capability

Q5f. What are the areas for expert system applications?

Ans: **Interpretation Systems** – Systems that infer situation descriptions from observations. This category includes surveillance, speech understanding, image analysis, signal interpretation and many kinds of intelligence analysis. An interpretation system explains observed data by assigning them symbolic meanings that describe the situation

Prediction Systems – These systems include weather forecasting, demographic predictions, economic forecasting, traffic predictions, crop estimates, and military, marketing and financial forecasting

Diagnostic Systems – These systems include medical, electronic, mechanical and software diagnoses. Diagnostic systems typically relate observed behavioral irregularities to underlying causes

Design Systems – These systems develop configurations of objects that satisfy the constraints of the design problem. Such problems include circuit layout, building design and plant layout. Design systems construct descriptions of objects in various relationships with one another and verify that these configurations conform to stated constraints.

Planning Systems – These systems specialize in planning problems such as automatic programming. They also deal with short and long term planning areas such as project management, routing, communications, product development, military applications and financial planning

Monitoring Systems – These systems compare observations of system behavior with standards that seem crucial for successful goal attainment. These crucial features correspond to potential flaws in the plan.

Debugging Systems – These systems rely on planning, design and prediction capabilities for creating specifications or recommendations to correct a diagnosed problem

Repair Systems – These systems develop and execute plans to administer a remedy for certain diagnosed problem. Such systems incorporate debugging, planning and execution capabilities.

Instruction Systems – Systems that incorporate diagnosis and debugging subsystems that specifically address students need. These systems begin by constructing a hypothetical description of the student that interprets his behavior. They then diagnosis his weakness and

identify appropriate remedies to overcome the deficiencies. Finally they plan a tutorial interaction intended to deliver remedial knowledge to the student

Control Systems – Systems that adaptively govern the overall behavior of the system. To do this a control system must repeatedly interpret the current situation predict the future, diagnose the cause of anticipated problem, formulate a remedial plan and monitor its execution to ensure success.