MODULE-2: Mathematical Models, Data Mining Data Preparation

Compiled by: Prof. Hrishikesh Tendulkar
hrishikesh.tendulkar@vsit.edu.in

VSIT | Vidyalankar School of Information Technology
NAAC Accredited College

Vidyalankar School of Information Technology
Wadala (E), Mumbai
www.vsit.edu.in

VSIT | Vidyalankar School of Information Technology

NAAC ACCREDITED COLLEGE

Certificate

This is to certify that the e-book titled “BUSINESS INTELLIGENCE” comprises all elementary learning tools for a better understating of the relevant concepts. This e-book is comprehensively compiled as per the predefined eight parameters and guidelines.

H M Tendulkar

Signature
Mr. Hrishikesh Tendulkar
Assistant Professor
Department of IT

Date: 21-01-2022

⚠ *DISCLAIMER: The information contained in this e-book is compiled and distributed for educational purposes only. This e-book has been designed to help learners understand relevant concepts with a more dynamic interface. The compiler of this e-book and Vidyalankar Institute of Technology give full and due credit to the authors of the contents, developers and all websites from wherever information has been sourced. We acknowledge our gratitude towards the websites YouTube, Wikipedia, and Google search engine. No commercial benefits are being drawn from this project.*

Unit II

- **Contents**

- **Mathematical models for decision making:** Structure of mathematical models, Development of a model, Classes of models
- **Data mining:** Definition of data mining, Representation of input data, Data mining process, Analysis methodologies
- **Data preparation**: Data validation, Data transformation, Data reduction

- **Recommended Books**
- **Text Books And Reference Books**
- Business Intelligence: Data Mining and Optimization for Decision Making, Carlo Vercellis, Wiley, First, 2009
- Decision support and Business Intelligence Systems , Efraim Turban, Ramesh Sharda, Dursun Delen , Pearson , Ninth, 2011
- Fundamental of Business Intelligence, Grossmann W, Rinderle-Ma, Springer, First, 2015

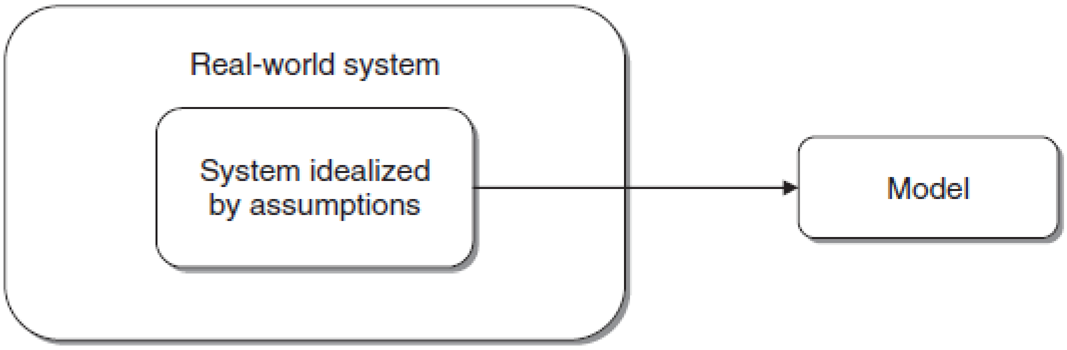
Unit II	Prerequisites						Linkage
	Sem. I	Sem. II	Sem. III	Sem. IV	Sem. V	Sem. VI	MSc IT Sem. I
Mathematical Models For Decision Making, Data Mining, Data Preparation	Discrete Mathematics	-	Data Structures	Computer Oriented Statistical Technique	Artificial Intelligence	-	Data Science (Sem I), Big Data Analytics (Sem II), Machine Learning, Robotic Process Automation (Sem III), Deep Learning (Sem IV)

MATHEMATICAL MODELS AND METHODS

In the previous chapters we have emphasized the critical role played by mathematical models in the development of business intelligence environments and decision support systems aimed at providing *active* support for knowledge workers. In this chapter we will focus on the main characteristics shared by different mathematical models embedded into business intelligence systems. We will also develop taxonomy of the most common classes of models, identifying for each of them the prevailing application domain.

Structure of mathematical models

Mathematical models have been developed and used in many application domains, ranging from physics to architecture, from engineering to economics. The models adopted in the various contexts differ substantially in terms of their mathematical structure. However, it is possible to identify a few fundamental features shared by most models. a model is a selective abstraction of a real system. In other words, a model is designed to analyse and understand from an abstract point of view the operating behaviour of a real system, regarding which it only includes those elements deemed relevant for the purpose of the investigation carried out. The following figure expresses in graphical terms the definition of a model.



Scientific and technological development has turned to mathematical models of various types for the abstract representation of real systems. According to their characteristics, models can be divided into iconic, analogical and symbolic.

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

Analogical: An analogical model is also a material representation, although it imitates the real behaviour by analogy rather than by replication. A wind tunnel built to investigate the aerodynamic properties of a motor vehicle is an example of an analogical model intended to represent the actual progression of a vehicle on the road.

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.

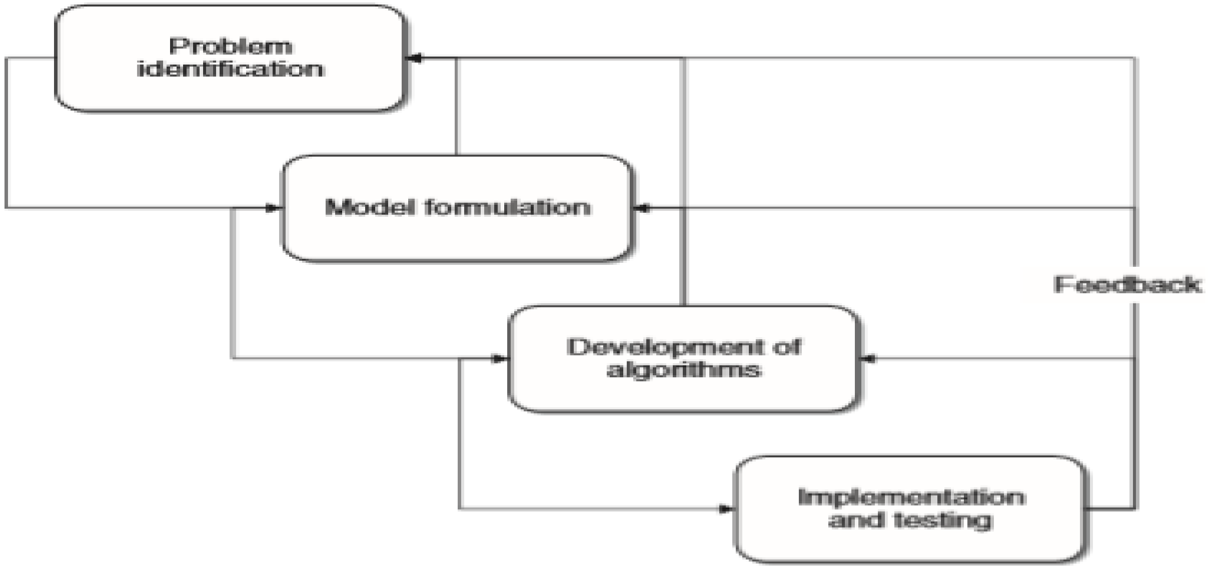
A further relevant distinction concerns the probabilistic nature of models, which can be either stochastic or deterministic.

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. Notice, however, that even for deterministic models the hypothesis of knowing the data with certainty may be relaxed. Sensitivity and scenario analyses, as well as what-if analysis, allow one to assess the robustness of optimal decisions to variations in the input parameters.

A further distinction concerns the temporal dimension in a mathematical model, which can be either static or dynamic.
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. Discrete-time dynamic models, which largely prevail in business intelligence applications, observe the status of a system only at the beginning or at the end of discrete intervals. Continuous-time dynamic models consider a continuous sequence of periods on the time axis.

Development of a model
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. For example, too high a stock level, corresponding to an excessive stock turnover rate, may possibly represent a symptom for a company manufacturing consumable goods. It is therefore necessary to understand what caused the problem, based on the opinion of the production managers. In this case, an ineffective production plan may be the cause of the stock accumulation.

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.

Time horizon – Usually a model includes a temporal dimension. For example, to formulate a tactical production plan over the medium term it is necessary to specify the production rate for each week in a year, whereas to derive an operational schedule it is required to assign the tasks to each production line for each day of the week. As we can see, the time span considered in a model, as well as the length of the base intervals, may vary depending on the specific problem considered.

Evaluation criteria: Appropriate measurable performance indicators should be defined in order to establish a criterion for the evaluation and comparison of the alternative decisions. These indicators may assume various forms in each different application, and may include the following factors:

- monetary costs and payoffs;
- effectiveness and level of service;
- quality of products and services;
- flexibility of the operating conditions;
- reliability in achieving the objectives.

1. **Decision variables**: Symbolic variables representing alternative decisions should then be defined. For example, if a problem consists of the formulation of a tactical production plan over the medium term, decision variables should express production volumes for each product, for each process and for each period of the planning horizon.

2. **Numerical parameters**: It is also necessary to accurately identify and estimate all numerical parameters required by the model. In the production planning example, the available capacity should be known in advance for each process, as well as the capacity absorption coefficients for each combination of products and processes.

3. **Mathematical relationships**: The final step in the formulation of a model is the identification of mathematical relationships among the decision variables, the numerical parameters and the performance indicators defined during the previous phases. Sometimes these relationships may be exclusively deterministic, while in other instances it is necessary to introduce probabilistic relationships. In this phase, the trade-off between the accuracy of the representation achieved through the model and its solution complexity should be carefully considered. It may turn out more helpful at a practical level to adopt a model that sacrifices some marginal aspects of reality in the representation of the system but allows an efficient solution and greater flexibility in view of possible future developments.

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. In other words, a solution algorithm should be identified and a software tool that incorporates the solution method should be developed or acquired. An analyst in charge of model formulation should possess a thorough knowledge of current solution methods and their characteristics.

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. Once the first numerical results have been obtained using the solution procedure devised, the model must be validated by submitting its conclusions to the opinion of decision makers and other experts in the application domain. A number of factors should be taken into account at this stage:

- the plausibility and likelihood of the conclusions achieved;
- the consistency of the results at extreme values of the numerical parameters;
- the stability of the results when minor changes in the input parameters are introduced.

Classes of models

There are several classes of mathematical models for decision making, which in turn can be solved by a number of alternative solution techniques. Each model class is better suited to represent certain types of decision-making processes. In this section we will cover the main categories of mathematical models for decision making:

Predictive models: A significant proportion of the models used in business intelligence systems, such as optimization models, require input data concerned with future events. For example, the results of random events determine the future demand for a product or service, the development of new scenarios of technological innovation and the level of prices and costs. As a consequence, ***predictive*** models play a primary role in business intelligence systems, since they are logically placed upstream with respect to other mathematical models and, more generally, to the whole decision-making process.

Predictions allow input information to be fed into different decision-making processes, arising in strategy, research and development, administration and control, marketing, production and logistics. Basically, all departmental functions of an enterprise make some use of predictive information to develop decision making, even though they pursue different objectives
Topic - Why Predictive Models.?

Source - https://www.youtube.com/watch?v=n5yKDyedUes

Predictive models can be subdivided into two main categories.
Explanatory **models** – To functionally identify a possible relationship between a dependent variable and a set of independent attributes. *Regression* models and *classification* models belong to this category.

Time series **models** – To functionally identify any temporal pattern expressed by a time series of observations referred to the same numerical variable.

Pattern recognition and machine learning models: In a broad sense, the purpose of pattern recognition and learning theory is to understand the mechanisms that regulate the development of intelligence, understood as the ability to extract knowledge from past experience in order to apply it in the future. Mathematical models for learning can be used to develop efficient algorithms that can perform such task. This has led to intelligent machines capable of learning from past observations and deriving new rules for the future, just like the human mind is able to do with great effectiveness due to the sophisticated mechanisms developed and fine-tuned in the course of evolution. Besides an intrinsic theoretical interest, mathematical methods for learning are applied in several domains, such as recognition of images, sounds and texts; biogenetic and medical diagnosis; relational marketing, for segmenting and profiling customers; manufacturing process control; identification of anomalies and fraud detection.

Optimization models: Many decision-making processes faced by companies or complex organizations can be cast according to the following framework: given the problem at hand, the decision maker defines a set of *feasible* decisions and establishes a criterion for the evaluation and comparison of alternative choices, such as monetary costs or payoffs. At this point, the decision maker must identify the *optimal* decision according to the evaluation criterion defined, that is, the choice corresponding to the minimum cost or to the maximum payoff. In general, optimization models arise naturally in decision-making processes where a set of limited resources must be allocated in the most effective way to different entities. These resources may be personnel, production processes, raw materials, components or financial factors. Among the main application domains requiring an optimal allocation of the resources we find:

- logistics and production planning;
- financial planning;
- work shift planning;
- marketing campaign planning;
- price determination.

Mathematical optimization models represent a fairly substantial class of optimization problems that are derived when the objective of the decision-making process is a function of the decision variables, and the criteria describing feasible decisions can be expressed by a set of mathematical equalities and inequalities in the decision variables. In light of the structure of the objective function and of the constraints, optimization models may assume different forms:

- linear optimization;
- integer optimization;
- convex optimization;
- network optimization;
- multiple-objective optimization.

Project management models: A *project* is a complex set of interrelated activities carried out in pursuit of a specific goal, which may represent an industrial plant, a building, an information system, a new product or a new organizational structure, depending on the different application domains. The execution of the project requires a planning and control process for the interdependent activities as well as the human, technical and financial resources necessary to achieve the final goal. *Project management* methods are based on the contributions of various disciplines, such as business organization, behavioural psychology and operations research. Mathematical models for decision making play an important role in project management methods. In particular, network models are used to represent the component activities of a project and the precedence relationships among them. These models allow the overall project execution time to be determined, assuming a deterministic knowledge of the duration of each activity. Stochastic models, on the other hand, usually referred to as *project evaluation and review techniques* (*PERT*), are used to derive the execution times when stochastic assumptions are made regarding the duration of the activities, represented by random variables. Finally, different classes of optimization models allow the analysis to be extended to the complex problem of optimally allocating a set of limited resources among the project activities in view of execution costs and times.

Risk analysis models: Some decision problems can be described according to the following conceptual paradigm: the decision maker is required to choose among a number of available alternatives, having uncertain information regarding the effects that these options may have in the future. For example, assume that senior management wishes to evaluate different alternatives in order to increase the company’s production capacity.

- On the one hand, the company may build a new plant providing a high operating efficiency and requiring a high investment cost.
- On the other hand, it may expand an existing plant with a lower investment but with higher operating costs.
- Finally, it may subcontract to external third parties part of its production: in this case, the investment cost is minimized but the operating costs are the highest among the available alternatives.

Clearly, in this situation the effects of the different options are strongly influenced by future stochastic events. In particular, a high level of future demand makes the construction of a new plant advantageous, while low demand levels tend to favour the subcontracting option. At an intermediate level of demand, the expansion of an existing plant may be convenient. However, the decision maker is forced to make a choice *before* knowing with absolute certainty the level of future demand. At best, she may obtain some stochastic information regarding the likelihood of occurrence of future events by carrying out some market research. In situations of this type, the methodological support offered by risk analysis models, may prove quite helpful

Waiting line models: The purpose of *waiting line* theory is to investigate congestion phenomena occurring when the demand for and provision of a service are stochastic in nature. If the arrival times of the customers and the duration of the service are not known beforehand in a deterministic way, conflicts may arise between customers in the use of limited shared resources. As a consequence, some customers are forced to wait in a line. Schematically, a waiting line system is made up of three main components:

- a ***source*** generating a stochastic process in which entities, also referred to as customers, arrive at a given location to obtain a service;
- a set of ***resources*** providing the service;
- a ***waiting area*** able to receive the entities whose requests cannot immediately be satisfied.

Waiting line models allow the performance of a system to be evaluated once its structure has been defined, and therefore are mostly useful within the system design phase. Indeed, in order to determine the appropriate values for the parameters that characterize a new system, relevant economic factors are considered, which depend on the service level that the system should guarantee when operating in optimal conditions. The main components of a waiting line system are

- The **population**, which can be finite or infinite, represents the source from which potential customers are drawn and to which they return once the requested service has been received.
- The **arrivals** process describes how customers arrive at the system entry point.
- The **service** process describes how the providers meet the requests of the customers waiting in line.
- The **number of existing stations and the number of providers** assigned to each station are additional relevant parameters of the waiting line system.
- The **waiting rules** describe the order in which customers are extracted from the line to be admitted to the service.

A primary role is finally played by priority schemes in which a level of priority is assigned to each customer. The customer with the highest priority is then served before all the other customers waiting in line.

DATA MINING

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.

Definition of data mining

Data mining activities constitute an iterative process aimed at the analysis of large databases, with the purpose of extracting information and knowledge that may prove accurate and potentially useful for knowledge workers engaged in decision making and problem solving. The term data mining refers therefore to the overall process consisting of data gathering and analysis, development of inductive learning models and adoption of practical decisions and consequent actions based on the knowledge acquired.

The term mathematical learning theory is reserved for the variety of mathematical models and methods that can be found at the core of each data mining analysis and that are used to generate new knowledge.

The data mining process is based on inductive learning methods, whose main purpose is to derive general rules starting from a set of available examples, consisting of past observations recorded in one or more databases. In other words, the purpose of a data mining analysis is to draw some conclusions starting from a sample of past observations and to generalize these conclusions with reference to the entire population, in such a way that they are as accurate as possible.

A further characteristic of data mining depends on the procedure for collecting past observations and inserting them into a database. Indeed, these records are usually stored for purposes that are not primarily driven by data mining analysis. For instance, information on purchases from a retail company, or on the usage of each telephone number stored by a mobile phone provider, will basically be recorded for administrative purposes, even if the data may be later used to perform some useful data mining analysis. The data gathering procedure is therefore largely independent and unaware of the data mining objectives, so that it substantially differs from data gathering activities carried out according to predetermined sampling schemes, typical of classical statistics. In this respect, data mining represents a secondary form of data analysis. Data mining activities can be subdivided into two major investigation streams, according to the main purpose of the analysis:

Interpretation – The purpose of interpretation is to identify regular patterns in the data and to express them through rules and criteria that can be easily understood by experts in the application domain The rules generated must be original and insignificant in order to actually increase the level of knowledge and understanding of the system of interest. For example, for a company in the retail industry it might be advantageous to cluster those customers who have taken out loyalty cards according to their purchasing profile. The segments generated in this way might prove useful in identifying new market places and directing future marketing campaigns.

Prediction – The purpose of prediction is to anticipate the value that a random variable will assume in the future or to estimate the likelihood of future events. For example, a mobile phone provider may develop a data mining analysis to estimate for its customers the probability of collaborating in favour of some competitor. Most data mining techniques derive their predictions from the value of a set of variables associated with the entities in a database. For example, a data mining model may indicate that the likelihood of future churning for a customer depends on features such as age, duration of the contract and percentage of calls to subscribers of other phone providers

Models and methods for data mining

A number of techniques originated in the field of computer science, such as classification trees or association rules, and are referred to as machine learning or knowledge discovery in databases. Other methods belong to multivariate (involving two or more variable quantities) statistics, such as regression or Bayesian classifiers, and are often parametric in nature but appear more theoretically grounded.

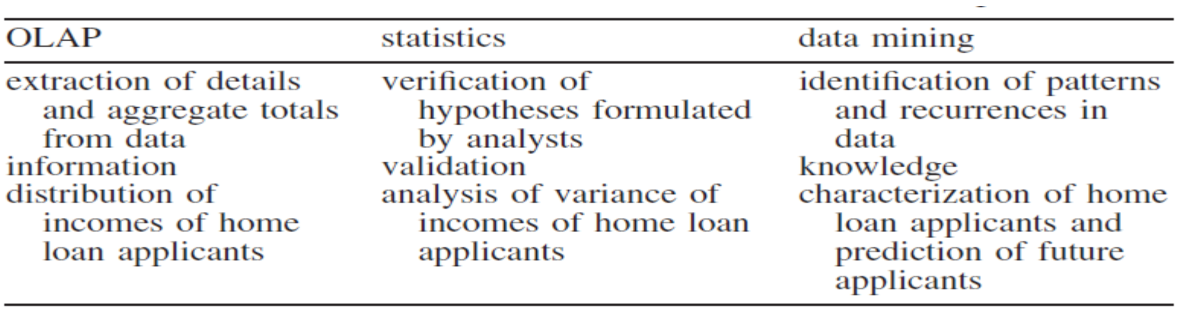
Example 2.1 – Linear regression. Linear regression models, are one of the best-known learning and predictive methodologies in classical statistics. In its simplest form, linear regression is used to relate a dependent response variable Y to an independent predictor X through a linear regression in the form Y = aX + b, where a and b are parameters to be determined using the available past observations. For example, Y may represent the sales of a mass consumption product during a week and X the total advertisement cost during the same week. With respect to the development phases of a model, the selection of a linear function determines the type of relationship between the predictor and the response variable. A reasonable evaluation metric is the sum of the squared differences between the values of Y actually observed in the past and the values predicted by the linear model. An appropriate optimization algorithm calculates the value of the parameters a and b in order to minimize the sum of squared errors.

Irrespective of the specific learning method that one wishes to adopt, there are other recurrent steps in the development of a data mining model, as shown example above

- The selection of a class of models to be used for learning from the past and of a specific form for representing patterns in the data
- The definition of a metric for evaluating the effectiveness and accuracy of the models being generated
- The design of a computational algorithm in order to generate the models by optimizing the evaluation metric

Data mining, classical statistics and OLAP

Data mining projects differ in many respects from both classical statistics and OLAP analyses. Such differences are shown in table below



OLAP	statistics	data mining
extraction of details and aggregate totals from data	verification of hypotheses formulated by analysts	identification of patterns and recurrences in data
information	validation	knowledge
distribution of incomes of home loan applicants	analysis of variance of incomes of home loan applicants	characterization of home loan applicants and prediction of future applicants

The main difference consists of the active orientation offered by inductive learning models, compared with the passive nature of statistical techniques and OLAP. In statistical analyses decision makers formulate a hypothesis that then has to be confirmed on the basis of sample evidence. Similarly, in OLAP analyses knowledge workers express some intuition on which they base extraction, reporting and visualization criteria. Both methods – on one hand statistical validation techniques and on the other hand information tools to navigate through data cubes – only provide elements to confirm or disprove the hypotheses formulated by the decision maker, according to a top-down analysis flow. Conversely, learning models, which represent the core of data mining projects, are capable of playing an active role by generating predictions and interpretations which actually represent new knowledge available to the users. The analysis flow in the latter case has a bottom-up structure. In particular, when faced with large amounts of data, the use of models capable of playing an active role becomes a critical success factor, since it is hard for knowledge workers to formulate a priori meaningful and well-founded hypotheses.

Applications of data mining

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 behaviour 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 both for 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 behaviours 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.
Topic – Application of Data Mining

Source - https://www.youtube.com/watch?v=ZrnyplW94-I

Representation of input data
The input to a data mining analysis takes the form of a two-dimensional table, called a *dataset*, irrespective of the actual logic and material representation adopted to store the information in files, databases, data warehouses and data marts used as data sources. The rows in the dataset correspond to the *observations* recorded in the past and are also called *examples, cases, instances or records.* The columns represent the information available for each

observation and are termed *attributes, variables, characteristics or features.* The attributes contained in a dataset can be categorized as *categorical* or *numerical*, depending on the type of values they take on.

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 the coding of their values is expressed by integer numbers.

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

$$X = [x_{ij}],\ i \in M = \{1, 2, \ldots, m\},\ j \in N = \{1, 2, \ldots, n\}, \quad (5.1)$$

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

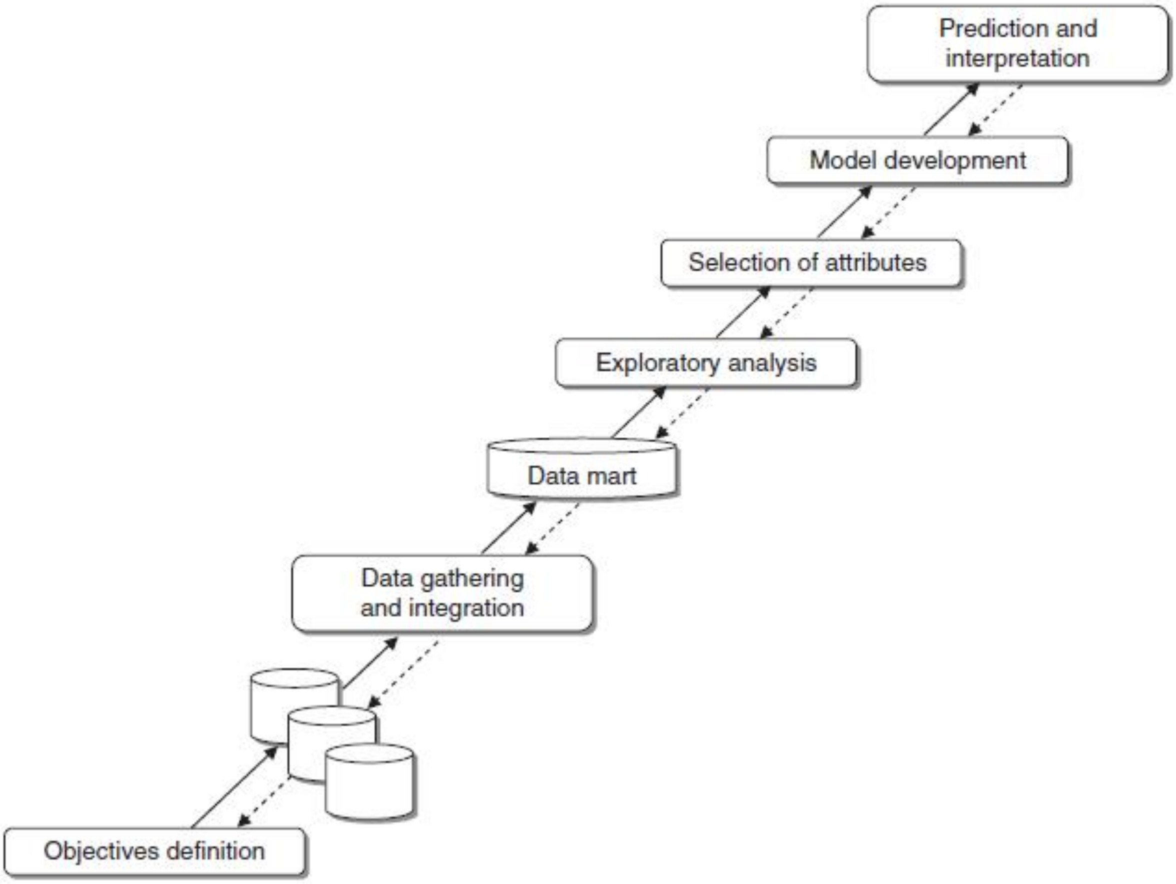
$$x_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \quad (5.2)$$

$$a_j = (x_{1j}, x_{2j}, \ldots, x_{mj}) \quad (5.3)$$

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.

Data mining process

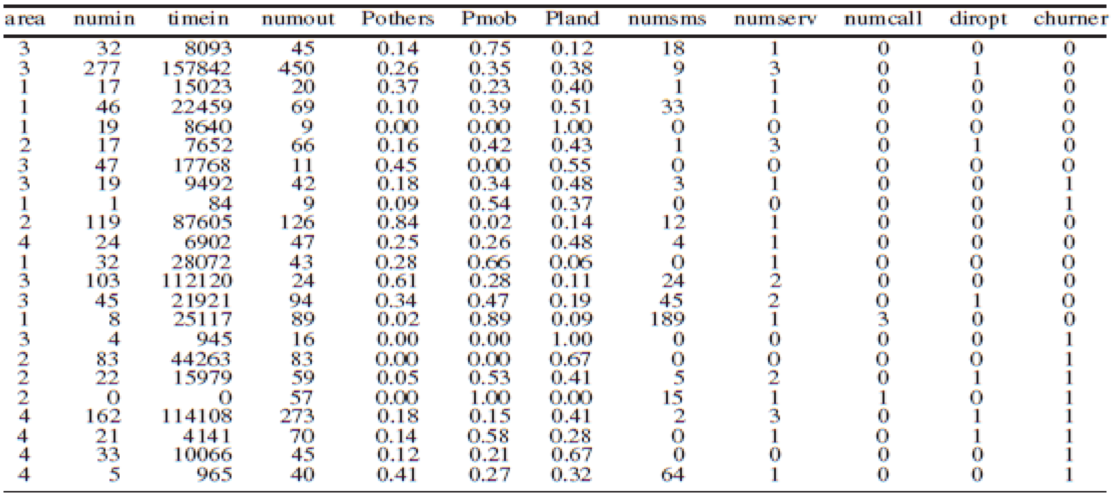
Data mining analyses are carried out in specific application domains and are intended to provide decision makers with useful knowledge. As a consequence, intuition and competence are required by the domain experts in order to formulate plausible and well-defined investigation objectives. With reference to example given below it is possible to define the problem and the goals of the investigation as the analysis of past data and identification of a model so as to express the propensity of customers to leave the service (churn) based on their characteristics, in order to understand the reasons for such disloyalty and predict the probability of churn.



Data mining process

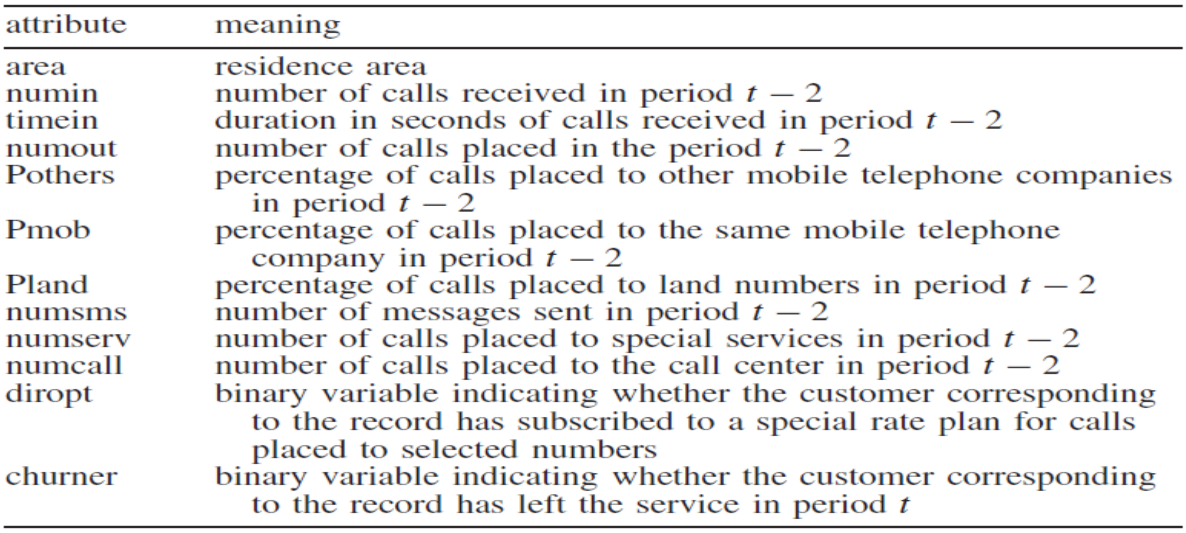
Example – Retention in the mobile phone industry – Table 2.2 shows the two-dimensional structure of input data from an example of the analysis of customer loyalty. Suppose that a mobile phone company carries out a data mining analysis with both prediction and interpretation goals. On the one hand, the company wishes to assess the likelihood of future churning by each customer, in order to target marketing actions for retention purposes. On the other hand, the intent is to understand the reasons why customers churn, with the purpose of improving the service level and reducing future churning.

Table 2.2 contains 2.3 observations and 12 attributes, whose meaning is indicated in Table 2.3. The first 11 attributes represent explanatory variables, while the last attribute represents the target variable, expressing the class of each record in relation to the objectives of the data mining analysis. The first explanatory variable gives personal demographic information while the rest refer to the use of the service. Observed values are relative to time period of index t − 2 for the explanatory attributes, whereas for the target variable they refer to period t. The difference in time placement is required in order to use the model for predictive purposes. It is necessary to predict during the current period which customers will leave the service within 2 periods, based on the available information, in order to develop timely and effective retention actions.



area	numin	timein	numout	Pothers	Pmob	Pland	numsms	numserv	numcall	diropt	churner
3	32	8093	45	0.14	0.75	0.12	18	1	0	0	0
3	277	157842	450	0.26	0.35	0.38	9	3	0	1	0
1	17	15023	20	0.37	0.23	0.40	1	1	0	0	0
1	46	22459	69	0.10	0.39	0.51	33	1	0	0	0
1	19	8640	9	0.00	0.00	1.00	0	0	0	0	0
2	17	7652	66	0.16	0.42	0.43	1	3	0	1	0
3	47	17768	11	0.45	0.00	0.55	0	0	0	0	0
3	19	9492	42	0.18	0.34	0.48	3	1	0	0	1
1	1	84	9	0.09	0.54	0.37	0	0	0	0	1
2	119	87605	126	0.84	0.02	0.14	12	1	0	0	0
4	24	6902	47	0.25	0.26	0.48	4	1	0	0	0
1	32	28072	43	0.28	0.66	0.06	0	1	0	0	0
3	103	112120	24	0.61	0.28	0.11	24	2	0	0	0
3	45	21921	94	0.34	0.47	0.19	45	2	0	1	0
1	8	25117	89	0.02	0.89	0.09	189	1	3	0	0
3	4	945	16	0.00	0.00	1.00	0	0	0	0	1
2	83	44263	83	0.00	0.00	0.67	0	0	0	0	1
2	22	15979	59	0.05	0.53	0.41	5	2	0	1	1
2	0	0	57	0.00	1.00	0.00	15	1	1	0	1
4	162	114108	273	0.18	0.15	0.41	2	3	0	1	1
4	21	4141	70	0.14	0.58	0.28	0	1	0	1	1
4	33	10066	45	0.12	0.21	0.67	0	0	0	0	1
4	5	965	40	0.41	0.27	0.32	64	1	0	0	1

Table 2.3 Meaning of the attributes in Table 2.2



attribute	meaning
area	residence area
numin	number of calls received in period $t-2$
timein	duration in seconds of calls received in period $t-2$
numout	number of calls placed in the period $t-2$
Pothers	percentage of calls placed to other mobile telephone companies in period $t-2$
Pmob	percentage of calls placed to the same mobile telephone company in period $t-2$
Pland	percentage of calls placed to land numbers in period $t-2$
numsms	number of messages sent in period $t-2$
numserv	number of calls placed to special services in period $t-2$
numcall	number of calls placed to the call center in period $t-2$
diropt	binary variable indicating whether the customer corresponding to the record has subscribed to a special rate plan for calls placed to selected numbers
churner	binary variable indicating whether the customer corresponding to the record has left the service in period t

Analysis methodologies

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.

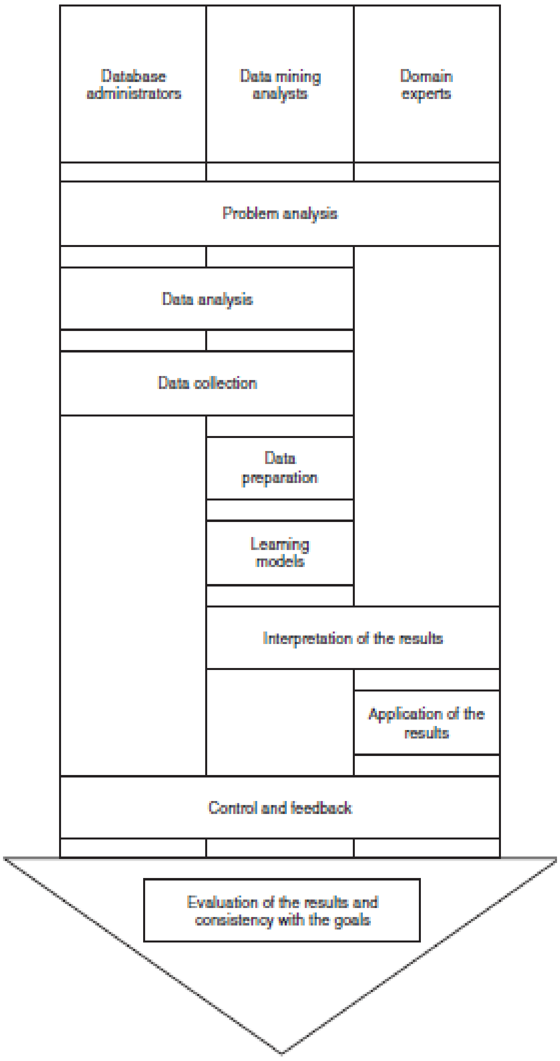


Figure 2.2 Actors and roles in data mining process

Characterization and discrimination – Where a categorical target attribute exists, before starting to develop a classification model, it is often useful to carry out an exploratory analysis whose purpose is twofold. On the one hand, the aim is to achieve a characterization by comparing the distribution of the values of the attributes for the records belonging to the same class. On the other hand, the purpose is to detect a difference, through a comparison between the distribution of the values of the attributes for the records of a given class and the records of a different class, or between the records of a given class and all remaining records. The information so acquired is usually presented to users in the form of histograms and other types of charts. The value of the information generated is, however, remarkable and may often direct the subsequent phase of attribute selection.

Classification – In a classification problem a set of observations is available, usually represented by the records of a dataset, whose target class is known. Observations may correspond, for instance, to mobile phone customers and the binary class may indicate whether a given customer is still active or has mixed. Each observation is described by a given number of attributes whose value is known; in the previous example, the attributes may correspond to age, customer seniority and outgoing telephone traffic distinguished by destinationThe categorical nature of the target determines the distinction between classification and regression.

Regression – Unlike classification, which is intended for discrete targets, regression is used when the target variable takes on continuous values. Based on the available explanatory attributes, the goal is to predict the value of the target variable for each observation. If one wishes to predict the sales of a product based on the promotional campaigns mounted and the sale price, the target variable may take on a very high number of discrete values and can be treated as a continuous variable. A classification problem may be turned into a regression problem, and vice versa. To see this, a mobile phone company interested in the classification of customers based on their loyalty, may come up with a regression problem by predicting the probability of each customer remaining loyal.

Time series – Sometimes the target attribute evolves over time and is therefore associated with adjacent periods on the time axis. In this case, the sequence of values of the target variable is said to represent a time series. For instance, the weekly sales of a given product observed over 2 years represent a time series containing 104 observations. Models for time series analysis investigate data characterized by a temporal dynamics and are aimed at predicting the value of the target variable for one or more future periods.

Association rules – Association rules, also known as affinity groupings, are used to identify interesting and recurring associations between groups of records of a dataset. For example, it is possible to determine which products are purchased together in a single transaction and how frequently. Companies in the retail industry resort to association rules to design the arrangement of products on shelves or in catalogues. Groupings by related elements are also used to promote cross-selling or to devise and promote combinations of products and services.

Clustering – The term cluster refers to a homogeneous subgroup existing within a population. Clustering techniques are therefore aimed at segmenting a heterogeneous population into a given number of subgroups composed of observations that share similar characteristics; observations included in different clusters have distinctive features. Unlike classification, in clustering there are no predefined classes or reference examples indicating the target class, so that the objects are grouped together based on their mutual homogeneity. Sometimes, the identification of clusters represents a preliminary stage in the data mining process, within exploratory data analysis. It may allow homogeneous data to be processed with the most appropriate rules and techniques and the size of the original dataset to be reduced, since the subsequent data mining activities can be developed autonomously on each cluster identified.

Description and visualization – The purpose of a data mining process is sometimes to provide a simple and concise representation of the information stored in a large dataset. Although, in contrast to clustering and association rules, descriptive analysis does not pursue any particular grouping or partition of the records in the dataset, an effective and concise description of information is very helpful, since it may suggest possible explanations of hidden patterns in the data and lead to a better understanding the phenomena to which the data refer. Notice that it is not always easy to obtain a meaningful visualization of the data. However, the effort of representation is justified by the remarkable conciseness of the information achieved through a well-designed chart.

DATA PREPARATION

Business intelligence systems and mathematical models for decision making can achieve accurate and effective results only when the input data are highly reliable. However, the data extracted from the available primary sources and gathered into a data mart may have several anomalies which analysts must identify and correct

Data validation

The quality of input data may prove unsatisfactory due to incompleteness, noise and inconsistency.

Incompleteness – Some records may contain missing values corresponding to one or more attributes, and there may be a variety of reasons for this. It may be that some data were not recorded at the source in a systematic way, or that they were not available when the transactions associated with a record took place. In other instances, data may be missing because of malfunctioning recording devices. It is also possible that some data were deliberately removed during previous stages of the gathering process because they were deemed incorrect. Incompleteness may also derive from a failure to transfer data from the operational databases to a data mart used for a specific business intelligence analysis.

Noise – Data may contain erroneous or anomalous values, which are usually referred to as outliers. Other possible causes of noise are to be sought in malfunctioning devices for data measurement, recording and transmission. The presence of data expressed in heterogeneous measurement units, which therefore require conversion, may in turn cause anomalies and inaccuracies.

Inconsistency – Sometimes data contain discrepancies due to changes in the coding system used for their representation, and therefore may appear inconsistent. For example, the coding of the products manufactured by a company may be subject to a revision taking effect on a given date, without the data recorded in previous periods being subject to the necessary transformations in order to adapt them to the revised encoding scheme. The purpose of data validation techniques is to identify and implement corrective actions in case of incomplete and inconsistent data or data affected by noise.

Incomplete data

To partially correct incomplete data one may adopt several techniques.

Elimination – It is possible to discard all records for which the values of one or more attributes are missing. In the case of a supervised data mining analysis, it is essential to eliminate a record if the value of the target attribute is missing. A policy based on systematic elimination of records may be ineffective when the distribution of missing values varies in an irregular way across the different attributes, since one may run the risk of incurring a substantial loss of information.

Inspection – Alternatively, one may opt for an inspection of each missing value, carried out by experts in the application domain, in order to obtain recommendations on possible substitute values. Obviously, this approach suffers from a high degree of arbitrariness and subjectivity, and is rather burdensome and time-consuming for large datasets. On the other hand, experience indicates that it is one of the most accurate corrective actions if skilfully exercised.

Identification – As a third possibility, a conventional value might be used to encode and identify missing values, making it unnecessary to remove entire records from the given dataset. For example, for a continuous attribute that assumes only positive values it is possible to assign the value $\{-1\}$ to all missing data. By the same token, for a categorical attribute one might replace missing values with a new value that differs from all those assumed by the attribute.

Substitution – Several criteria exist for the automatic replacement of missing data, although most of them appear somehow arbitrary. For instance, missing values of an attribute may be replaced with the mean of the attribute calculated for the remaining observations. This technique can only be applied to numerical attributes, but it will clearly be ineffective in the case of an asymmetric distribution of values. In a supervised analysis it is also possible to replace missing values by calculating the mean of the attribute only for those records having the same target class

Data affected by noise

The term noise refers to a random perturbation within the values of a numerical attribute, usually resulting in noticeable anomalies. First, the outliers in a dataset need to be identified, so that subsequently either they can be corrected and regularized or entire records containing them are eliminated.

The easiest way to identify outliers is based on the statistical concept of dispersion. The sample mean μ¯ j and the sample variance σ¯ 2j of the numerical attribute aj are calculated. If the attribute follows a distribution that is not too far from normal, the μ¯ j values falling outside an appropriate interval cantered around the mean value are identified as outliers. With a confidence of 100(1 − α) % (approximately 96% for α = 0.05) it is possible to consider as outliers those values that fall outside the interval

(μ¯ j − zα/2σ¯j, μ¯ j + zα/2 ¯ σj), (3.1)

where zα/2 is the α/2 quantile of the standard normal distribution. This technique is simple to use, although it has the drawback of relying on the critical assumption that the distribution of the values of the attribute is bell-shaped and roughly normal. However, it is possible to obtain analogous bounds independent of the distribution, with intervals that are only slightly less stringent. Once the outliers have been identified, it is possible to correct them with values that are deemed more plausible, or to remove an entire record containing them.

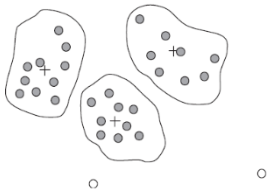


Figure 3.1 Identifiers of outliers using cluster analysis

An alternative technique, illustrated in the above figure is based on the distance between observations and the use of clustering methods. Once the clusters have been identified, representing sets of records having a mutual distance that is less than the distance from the records included in other groups, the observations that are not placed in any of the clusters are identified as outliers.

Clustering techniques offer the advantage of simultaneously considering several attributes, while methods based on dispersion can only take into account each single attribute separately. A variant of clustering methods, also based on the distances between the observations, detects the outliers through two parametric values, p and d, to be assigned by the user. An observation $\mathbf{x}_i$ is identified as an outlier if at least a percentage p of the observations in the dataset are found at a distance greater than d from $\mathbf{x}_i$.

There are also other methods which correct anomalous data there exist regularization method which automatically corrects the anomalous data. For example, simple or multiple regression models predict the value of the attribute a_j that one wishes to regularize based on other variables existing in the dataset.

Data transformation

In most data mining analyses it is appropriate to apply a few transformations to the dataset in order to improve the accuracy of the learning models subsequently developed.

Standardization

Most learning models benefit from a preventive standardization of the data, also called normalization. The most popular standardization techniques include the decimal scaling method, the min-max method and the z-index method.

Decimal Scaling – Decimal scaling is based on the transformation

$$x'_{ij} = \frac{x_{ij}}{10^h}, \qquad (6.2)$$

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

$$x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}(x'_{\max,j} - x'_{\min,j}) + x'_{\min,j}, \qquad (6.3)$$

Where

$$x_{\min,j} = \min_i x_{ij}, \quad x_{\max,j} = \max_i x_{ij}, \qquad (6.4)$$

are the minimum and maximum values of the attribute **a**$_j$ $x'_{\min,j}$ before transformation, while $x'_{\max,j}$ and 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 $\bar{\mu}_j$ $\bar{\sigma}_j$ and are respectively the sample mean and sample standard deviation of the attribute **a**$_j$. If the distribution of values of the attribute **a**$_j$ is roughly normal, the *z*-index based transformation generates values that are almost certainly within the range *(*−3*,* 3*)*.

Topic - Z-score and Decimal Scaling

Source – https://www.youtube.com/watch?v=pHdiuG920lM

Feature extraction

Attribute extraction may also consist of the creation of new variables that summarize within themselves the relevant information contained in a subset of the original attributes. For example, in the context of image recognition one is often interested in identifying the existence of a face within a digitalized photograph. There are different indicators intended for the synthesis of each piece of information contained in a group of adjacent pixels, which make it easier for classification algorithms to detect faces.

Data reduction

When dealing with a small dataset, the transformations described above are usually adequate to prepare input data for a data mining analysis. However, when facing a large dataset it is also appropriate to reduce its size, in order to make learning algorithms more efficient, without sacrificing the quality of the results obtained. There are three main criteria to determine whether a data reduction technique should be used: *efficiency*, *accuracy* and *simplicity* of the models generated.

Efficiency – The application of learning algorithms to a dataset smaller than the original one usually means a shorter computation time. If the complexity of the algorithm is a super linear function, as is the case for most known methods, the improvement in efficiency resulting from a reduction in the dataset size may be dramatic. a reduction in processing times allows the analyses to be carried out more quickly.

Accuracy – In most applications, the accuracy of the models generated represents a critical success factor, and it is therefore the main criterion followed in order to select one class of learning methods over another. Therefore, data reduction techniques should not significantly compromise the accuracy of the model generated. it may also be the case that some data reduction techniques, based on attribute selection, will lead to models with a higher generalization capability on future records.

Simplicity – In some data mining applications, concerned more with interpretation than with prediction, it is important that the models generated be easily translated into simple rules that can be understood by experts in the application domain. As a trade-off for achieving simpler rules, decision makers are sometimes willing to allow a slight decrease in accuracy. Data reduction often represents an effective technique for deriving models that are more easily interpretable.

Sampling

A further reduction in the size of the original dataset can be achieved by extracting a sample of observations that is significant from a statistical standpoint. This type of reduction is based on classical inferential reasoning. It is therefore necessary to determine the size of the sample that guarantees the level of accuracy required by the subsequent learning algorithms and to define an adequate sampling procedure. Sampling may be *simple* or *stratified* depending on whether one wishes to preserve in the sample the percentages of the original dataset with respect to a categorical attribute that is considered critical. It is also useful to set up several independent samples, each of a predetermined size, to which learning algorithms should be applied.

Feature Selection

The purpose of *feature selection*, also called *feature reduction*, is to eliminate from the dataset a subset of variables that are not deemed relevant for the purpose of the data mining activities. Feature reduction has several potential advantages. Due to the presence of fewer columns, learning algorithms can be run more quickly on the reduced dataset than on the original one.

Filter methods – Filter methods select the relevant attributes before moving on to the subsequent learning phase, and are therefore independent of the specific algorithm being used. The attributes deemed most significant are selected for learning, while the rest are excluded. The simplest filter method to apply for supervised learning involves the assessment of each single attribute based on its level of correlation with the target. Consequently, this lead to the selection of the attributes that appear mostly correlated with the target.

Wrapper methods – If the purpose of the data mining investigation is classification or regression, and consequently performances are assessed mainly in terms of accuracy, the selection of predictive variables should be based not only on the level of relevance of each single attribute but also on the specific learning algorithm being utilized. Wrapper methods are able to meet this need, since they assess a group of variables using the same classification or regression algorithm used to predict the value of the target variable. Wrapper methods are usually burdensome from a computational standpoint, since the assessment of every possible combination identified by the search engine requires one to deal with the entire training phase of the learning algorithm.

Embedded methods – For the embedded methods, the attribute selection process lies *inside* the learning algorithm, so that the selection of the optimal set of attributes is directly made during the phase of model generation. At each tree node, they use an evaluation function that estimates the predictive value of a single attribute or a linear combination of variables. In this way, the relevant attributes are automatically selected, and they determine the rule for splitting the records in the corresponding node.

Filter methods are the best choice when dealing with very large datasets, whose observations are described by a large number of attributes. In these cases, the application of wrapper methods is inappropriate due to very long computation times. Moreover, filter methods are flexible and in principle can be associated with any learning algorithm. However, when the size of the problem at hand is moderate, it is preferable to turn to wrapper or embedded methods, which afford in most cases accuracy levels that are higher compared to filter methods. Wrapper methods select the attributes according to a search scheme that inspects in sequence several subsets of attributes and applies the learning algorithm to each subset in order to assess the resulting accuracy of the corresponding model.

If a dataset contains *n* attributes, there are 2*n* possible subsets and therefore an exhaustive search procedure would require excessive computation times even for moderate values of *n*. Therefore, the procedure for selecting the attributes for wrapper methods is usually of a heuristic nature, based in most cases on a *greedy* logic that evaluates for each attribute a relevance indicator adequately defined and then selects the attributes based on their level of relevance. In particular, three distinct schemes can be followed

Forward – According to the forward search scheme, also referred to as *bottom-up* search, the exploration starts with an empty set of attributes and subsequently introduces the attributes one at a time based on the ranking induced by the relevance indicator. The algorithm stops when the relevance index of all the attributes still excluded is lower than a prefixed threshold.

Backward – The backward search scheme, also referred to as *top-down* search, begins the exploration by selecting all the attributes and then eliminates them one at a time based on the preferred relevance indicator. The algorithm stops when the relevance index of all the attributes still included in the model is higher than a prefixed threshold.

Forward–backward – The forward–backward method represents a trade-off between the previous schemes, in the sense that at each step the best attribute among those excluded is introduced and the worst attribute among those included is eliminated. Also in this case, threshold values for the included and excluded attributes determine the stopping criterion.

Principal component analysis

Principal component analysis (PCA) is the most widely known technique of attribute reduction by means of projection. The purpose of this method is to obtain a projective transformation that replaces a subset of the original numerical attributes with a lower number of new attributes obtained as their linear combination, without this change causing a loss of information. Before applying the principal component method, it is expedient to standardize the data, so as to obtain for all the attributes the same range of values, usually represented by the interval [−1, 1]. Moreover, the mean of each attribute **a***j* is made equal to 0 by applying the transformation

$$\tilde{x}_{ij} = x_{ij} - \frac{1}{m}\sum_{i=1}^{m} x_{ij}. \tag{6.6}$$

Let **X** denote the matrix resulting from applying the transformation (6.6) to the original data, and let **V** = **X_X** be the covariance matrix of the attributes. Principal components are better suited than the original attributes to explain fluctuations in the data, in the sense that usually a subset consisting of *q*

Principal components, with *q* < *n*, has an information content that is almost equivalent to that of the original dataset. Principal components are generated in sequence by means of an iterative algorithm. The first component is determined by solving an appropriate optimization problem, in order to explain the highest percentage of variation in the data. At each iteration the next principal component is selected, among those vectors that are orthogonal to all components already determined, as the one which explains the maximum percentage of variance not yet explained by the previously generated components. At the end of the procedure the principal components are ranked in non-increasing order with respect to the amount of variance that they are able to explain. Example of PCA

Step 1: Get some data
Step 2: Calculate the mean
Step 3: Calculate the covariance matrix
Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix
Step 5: Choosing components and forming a feature vector
Step 6: Deriving the new data set

Step 1: Get some data

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9
Mean: 1.81	Mean: 1.91

Step 2: Calculate the mean

Mean of X = 1.81
Mean of Y = 1.91

Step 3: Calculate the covariance matrix

C = $\begin{bmatrix} cov\ (x,x) & cov\ (x,y) \\ cov\ (y,x) & cov\ (y,y) \end{bmatrix}$

Cov (x,x) = $\sum_{i=1}^{n} \frac{(xi-\overline{x})\ (xi-\overline{x})}{n-1}$

Cov (x,y) = $\sum_{i=1}^{n} \frac{(xi-\overline{x})\ (yi-\overline{y})}{n-1}$

Cov (y,y) = $\sum_{i=1}^{n} \frac{(yi-\overline{y})\ (yi-\overline{y})}{n-1}$

Cov (x, x)

X-X'	(X-X') (X-X')
0.69	0.4761
1.31	1.7161
0.39	0.1521
0.09	0.0081
1.29	1.6641
0.49	0.2401
0.19	0.0361
0.81	0.6561
0.31	0.0961
0.71	0.5041

Sum = 5.549

Cov (x,x) =5.549/ 9 = 0.61655556

Cov (y, y):

Y-Y'	(Y-Y') (Y-Y')
0.49	0.2401
1.21	1.4641
0.99	0.9801
0.29	0.0841
1.09	1.1881
0.79	0.6241
0.31	0.0961
0.81	0.6561
0.31	0.0961
1.01	1.0201

Sum = 6.449

Cov (y, y) = 6.449 / 9 = 0.7165556

Cov (x, y) (y, x) :

(X-X') (Y-Y')
0.3381
1.5851
0.3861
0.0261
1.4061
0.3871
0.0589
0.6561
0.0961
0.7171

Sum = 5.539

Cov (x, y) = 5.539/ 9 = 0.61544444

So, Covariance matrix will become:

C = $\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix

C – λ I = 0

$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$ - λ $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ = 0

$$\begin{bmatrix} 0.6165 - \lambda & 0.6154 \\ 0.6154 & 0.7165 - \lambda \end{bmatrix}$$

Use matrix determinant formula to solve this:

$(A * D) - (B * C)$

= $[(0.6165 - \lambda) * (0.7165 - \lambda)] - [0.6154 * \ 0.6154]$ = 0

= [(0.6165 * 0.7165) -0.6165 $\lambda - 0.7165\lambda + \ \lambda * \ \lambda]$ – [0.378717] = 0

= [λ^2 - 1333 λ + 0.0630] = 0

Roots are: 1.2840, 0.0490 They are called as λ1, and λ2.

Condition to be satisfied as: λ1 > λ2

So, λ1 = 1.2840 & λ2 = 0.0490

Step 5: Choosing components and forming a feature vector

C V= λ V

Use λ1 first:

$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{matrix} X1 \\ Y1 \end{matrix}$ = 1.2840 $\begin{matrix} X1 \\ Y1 \end{matrix}$

(1) 0.6165X1 + 0.6154 Y1 = 1.2840 X1

0.6165X1 – 1.2840X1 = -0.6154Y1

-0.6675X1 = -0.6154Y1

X1 = $\frac{0.6154}{0.6675}$ Y1

X1 = 0.92194 Y1

$$\begin{bmatrix} 0.92194 \\ 1 \end{bmatrix}$$

Square them, and add them and then square root it

$$\begin{bmatrix} 0.8499 \\ 1 \end{bmatrix}$$

= 1.8499

= 1.3601

Now to create PCA components Divide them by this value

PC1 =

$$\begin{bmatrix} 0.92194/1.3601 \\ 1/1.3601 \end{bmatrix} = \begin{bmatrix} 0.6778 \\ 0.7352 \end{bmatrix}$$

(2) 0.6154X1 + 0.7165 Y1 = 1.2840 Y1

Solve for it.

Similarly, do for λ2

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{matrix} X2 \\ Y2 \end{matrix} = 0.0490 \begin{matrix} X2 \\ Y2 \end{matrix}$$

(1) 0.6165X2 + 0.6154Y2 = 0.0490 X2

(2) 0.6154X2 + 0.7165Y2 = 0.0490Y2

(1) 0.6165X2 + 0.6154Y2 = 0.0490 X2

0.6165X2-0.0490X2 = -0.6154Y2

0.5675X2 = -0.6154Y2

X2 = $\frac{-0.6154}{0.5675}$ Y2

X2 = -1.0845 Y2

$$\begin{bmatrix} -1.0845 \\ 1 \end{bmatrix}$$

Square them, and add them and then square root it

$$\begin{bmatrix} 0.8499 \\ 1 \end{bmatrix}$$

= 2.1761

= 1.4751

Now to create PCA components Divide them by this value

PC2 =

$$\begin{bmatrix} -1.0845/1.47517 \\ 1/1.47517 \end{bmatrix} = \begin{bmatrix} -0.7351 \\ 0.6778 \end{bmatrix}$$

Step 6: Deriving the new data set

Now, we got two principal components

PC1 = $\begin{bmatrix} 0.6778 \\ 0.7352 \end{bmatrix}$ PC2= $\begin{bmatrix} -0.7351 \\ 0.6778 \end{bmatrix}$

Let’s check which components are relevant:

Eigen value 1 = λ1 = 1.2840

Eigen value 2= λ2 = 0.0490

To compute the percentage of variance (information) accounted for by each component, we divide the eigenvalue of each component by the sum of eigenvalues.

Percentage variance of λ1= 1.2840/(1.2840+0.0490)*100 = 96%

Percentage variance of λ2= 0.0490/(1.2840+0.0490)*100 = 4%

If we apply this on the example above, we find that PC1 and PC2 carry respectively 96% and 4% of the variance of the data.

Continuing with the example from the previous step, we can either form a feature vector with both of the eigenvectors *v*1 and *v*2:

$$\begin{bmatrix} 0.67787 & -0.7352 \\ 0.7352 & 0.6778 \end{bmatrix}$$

Or discard the eigenvector *v*2, which is the one of lesser significance, and form a feature vector with *v*1 only:

$$\begin{bmatrix} 0.67787 \\ 0.7352 \end{bmatrix}$$

Discarding the eigenvector *v2* will reduce dimensionality by 1, and will consequently cause a loss of information in the final data set. But given that *v*2 was carrying only 4% of the information, the loss will be therefore not important and we will still have 96% of the information that is carried by *v*1.

Final dataset = FeatureVector ^{T} * StandardizedOriginalDataSet ^{T}

Data discretization

The general purpose of data reduction methods is to obtain a decrease in the number of distinct values assumed by one or more attributes. Data discretization is the primary reduction method. On the one hand, it reduces continuous attributes to categorical attributes characterized by a limited number of distinct values. On the other hand, its aim is to significantly reduce the number of distinct values assumed by the categorical attributes.

Subjective subdivision. Subjective subdivision is the most popular and intuitive method. Classes are defined based on the experience and judgment of experts in the application domain.

Subdivision into classes. Subdivision into categorical classes may be achieved in an automated way using the techniques described below. In particular, the subdivision can be based on classes of equal size or equal width.

Hierarchical discretization. The third type of discretization is based on hierarchical relationships between concepts and may be applied to categorical attributes, just as for the hierarchical relationships between provinces and regions. In general, given a hierarchical relationship of the one-to-many kind, it is possible to replace each value of an attribute with the corresponding value found at a higher level in the hierarchy of concepts.

Example of Data Discretization

Consider the following dataset
3, 4, 4, 7, 12, 15, 21, 23, 27
Number of classes are 3. Therefore, Subdivision into equal size classes will give
Class 1 3,4,4
Class 2 7,12,15
Class 3 21,23,27

Once the classes has been created Regularization by mean value has to be done. Hence the mean values will be

Mean of (3,4,4) - 3.66
Mean of (7,12,15) - 11.33
Mean of (21,23,27) - 23.66

Therefore, Regularization by mean value will give

Class 1 - 3.66,3.66,3.66
Class 2 -11.33,11.33,11.33
Class 3 - 23.66,23.66,23.66

Once the mean values have been found Regularization by boundary values has to be done. Here the boundary values are 3 4 and 7 15 and 21 27. Middle values must be changed to the nearby boundary values. Therefore, Regularization by boundary values will give

Class 1 - 3,4,4
Class 2 -7,15,15
Class 3 - 21,21,27

- **Graded Questions**
  1. What are the different types of mathematical models?
  2. Explain in brief different phases of mathematical models
  3. Write a short note on Predictive models
  4. Write a short note on Project management models
  5. What are the different factors used for model formulation?
  6. Explain in detail the purpose of Pattern & Machine Learning model
  7. What are the different data streams in Data Mining?
  8. Explain the differences between OLAP, statistics and data mining
  9. What are the different types of attributes contained in a dataset?
  10. Write a short note on exploratory analysis in data mining process
  11. What do you mean by Supervised Learning and Un-Supervised Learning?
  12. Explain in detail the classification task of data mining
  13. Write a note on clustering task of data mining
  14. What are the techniques to partially correct the incomplete data?
  15. What are the different methods of standardization?
  16. Write a short note on Data Reduction
  17. Write a short note on sampling
  18. Explain what Principal Component Analysis is
  19. Write a note on Feature selection methods
  20. Explain in short Data discretization

- **Multiple Choice Questions**

1	________ model is a material representation of a real system whose behaviour is imitated for the purpose of analysis		
a	Analogical	b	Symbolic
c	**Iconic**	d	Dynamic

2	_______ model is an abstract representation of a real system		
a	Analogical	b	Dynamic
c	Iconic	d	**Symbolic**

3	________ model consider a given system and the related decision making process within one single temporal stage		
a	Dynamic	b	Analogical
c	**Static**	d	Deterministic

4	________ models arise naturally in decision making process where a set of limited resources must be allocated in most effective way to different entities		
a	**Optimization**	b	Project Management
c	Predictive	d	Risk Analysis

5	_______ models allow the performance of the system to be evaluated once its structure has been defined		
a	Project Management	b	Risk Analysis
c	Predictive	d	**Waiting Line**

6	________indicates the process of exploration and analysis of a dataset to find regular patterns		
a	Data Exploration	b	Knowledge Pattern
c	**Data Mining**	d	Data Warehousing

7	_______ term refers to the overall process consisting of data gathering and analysis, development of inductive models		
a	**Data Mining**	B	Knowledge Pattern
c	Data Exploration	d	Data Warehousing

8	_________ term is reserved for the variety of mathematical models and methods that can be found at the core of each data		
a	Knowledge Pattern	b	**Mathematical Learning Theory**
c	Data Warehousing	d	Data Mining

9	The purpose of _____ is to identify regular patterns in the data and to express them through certain rules and criteria that can be easily understood by the experts		
a	**Interpretation**	b	Classification
c	Prediction	d	Regression

10	The purpose of _______ is to anticipate the value that a random variable will assume in future		
a	Interpretation	b	Classification
c	**Prediction**	d	Regression

11	Identification of customer segments that are most likely to respond to marketing campaigns is an application of _____________		
a	Fraud Detection	b	Text Mining

c	Risk Evaluation	d	**Relational Marketing**

12	________ is another field of application of data mining which deals with illegal use of credit cards, false claims etc		
a	**Fraud Detection**	b	Text Mining
c	Risk Evaluation	d	Relational Marketing

13	The purpose of _________ is to estimate the risk connected with future decisions which often assume dichotomous form		
a	Fraud Detection	b	Text Mining
c	**Risk Evaluation**	d	Relational Marketing

14	________application of data mining can be applied to different kinds of texts which represent unstructured data in order to classify articles, books, documents		
a	Fraud Detection	b	**Text Mining**
c	Risk Evaluation	d	Relational Marketing

15	_________ application of data mining deals with treatment and classification of digital images		
a	Text Mining	b	Web Mining
c	**Image Recognition**	d	Medical diagnosis

16	________ applications are intended for the analysis of e-commerce sites or evaluating e-learning training course		
a	**Web Mining**	b	Text Mining
c	Image Recognition	d	Medical Diagnosis

17	The input to the data mining analysis takes the form of a two-dimensional table called ___________		
a	Data Table	b	Data Row
c	**Dataset**	d	Data Column

18	_______assume a finite number of distinct values representing a qualitative property of an entity to which they refer		
a	Numerical	b	Nominal
c	Ordinal	d	**Categorical**

19	_________attributes are categorical attributes without a natural ordering, such as the province of residence		
a	Numerical	b	Ordinal
c	Categorical	d	**Nominal**

20	_________attributes are numerical attributes that assume a finite number or a countable infinity of values		
a	Categorical	b	Counts
c	**Discrete**	d	Continuous

21	________attributes are numerical attributes that assume an uncountable infinity of values		
a	**Continuous**	b	Nominal
c	Discrete	d	Ordinal

22	________also known as affinity groupings, are used to identify interesting and recurring associations between groups of records of a dataset.		
a	**Association Rules**	b	Regression
c	Time Series	d	Classification

23	Data may contain erroneous or anomalous values, which are usually referred to as outliers or ________		
a	Inconsistency	b	Incompleteness
c	**Noise**	d	All of these

24	________is a technique to discard all records for which the values of one or more attributes are missing		
a	**Elimination**	b	Identification
c	Inspection	d	Substitution

25	In _________technique missing values of an attribute may be replaced with the mean of the attribute calculated for the remaining observations.		
a	Identification	b	**Substitution**
c	Elimination	d	Inspection