Theoritical Examination - 2022

RONE MONDAL

SEMESTER - 5th

CU ROLL NO - 193114-21-0007

Date - 17/01/2022

CU REGD. NO - 314-1111-0229-19

~~PAPER - CC-12~~

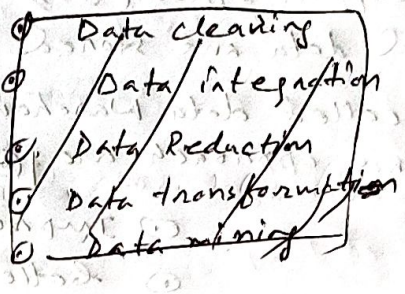
PAPER : DSE-A-2

Subject - Data Mining

193314-21-0007

1) a) Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems.

To 6 Data mining process includes:




- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

1) b) Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate the model's accuracy.

1) c) A classifier is a supervised function (machine learning where the learned (target) attribute is categorical ("nominal") in order to classify. It is used after the learning process to classify new records (data) by giving them the best target attribute (prediction). Rows are classified into buckets.

1) d) Web mining is the application of data mining techniques to discover patterns from the worldwide web. It uses automa methods to extract both structured and unstructured data from web pages, server logs and link structures. Web content mining extracts information from within a page.

1) e) Data integration is the process of combining data from multiple different sources in order to extract additional value. ~~Think for example of a situation~~ It refers to the technical and business processes used to combine da from multiple sources to provide a unified, single view the data. Think for example of a situation where you ne to combine various Excel spreadsheets with the informati held in an Access database - this is a simple data integr task.

8) a) Data warehouse is one of the most rapidly growing areas in management information system. With this approach, data for Executive Information System(EIS) and Decision Support System(DSS) applications are separated from operational data and stored in a seperate database. This process is called data warehousing. The major advantages of this approach are:

- Improved in performance
- better data quality
- the ability to consolidate and summarize data from heterogeneous systems.

A data warehouse is a part of a larger infrastructure that includes legacy data sources, external data sources, a repository, data acquisition software, and user interface and related analytical tools. The aim of this research work is to elaborate that how the textile industry can manage and improve their production capacity and resources at optimum level to produce a good quality result using data warehousing and data mining techniques.

8) b) The principal components of Autocorrelation Matrix are -

- Adaptive Filters
- Eigenvalue
- Eigenvector
- Least Mean Square
- Mean Square Error
- Filter Co-efficient
- Reference Signal

7) a) Pre-processing of data is mainly to check the data quality. The quality can be checked by the following.

- Accuracy: To check whether the data entered is correct or not
- Completeness: To check whether the data is available or not recorded
- Consistency: To check whether the same data is kept in all the places the do or do not match.
- Timeliness: The data should be updated correctly
- Believability: The data should be trustable

- some major tasks in Data-preprocessing:

- Data cleaning: The process to remove incorrect, incomplete, inaccurate data.
- Data integration: The process of combining multiple sources into a single dataset.
- Data Transformation: The change made in the format or the structure of the data is called Data Transformation.
- Data reduction: This process helps in the reduction of the volume of the data which makes the analysis easier.

2) e) Utility of data cleaning:

Data cleaning is very essential process in data pre-processing because having clean data will ultimately increase overall productivity and allow for the highest quality information in our decision making. It includes:

i) Removal of errors when multiple sources of data are at play.

ii) Fewer errors make for happier client and less-frustrated employees.

iii) Ability to map the different function and what our data is intendend to do.

2) c) Data Extraction:

It is the process of obtaining data from a database platform so that it can be replicated to a destination such as data warehouse.

Data ~~extract~~ extraction is the first step in a data ingestion process.

Data extraction can be done in the following ways:

- update notification: It is the easiest way to extract data from a source system is ~~has~~ to have that system issue a notification when a record has been changed.

- ~~Incre~~

• **Incremental extraction:** Some data sources are unable to provide notification that an update has occurred, but they are able to identify which records have been modified and provide an extract of those records.

• **Full extraction:** Full extraction involves high data transfer volumes, which can put a load on the network.

**Minimum distance classifier:**

6) a) The minimum distance classifier is used to classify unknown image data to classes which minimize the distance between the image data and the classes in multi-feature space. The distance is defined as an index of similarity so that the minimum distance is identical to the maximum similarity.

**Euclidean distance:**

$$d_k^2 = (x - \mu_k)^t (x - \mu_k)$$

6) b) **similarity measure:**

It is an algorithm that calculates the degree of some aspect of similarity between two entities. It is measure of how much alike two data objects are. In data mining, it is distance with dimension representing features of the objects. similarly, method measure is a method to calculate the degree of similarity between mapping sources.

6) c) **Nearest Neighbour Technique:**

~~It is very~~

6) c) Nearest Neighbour Technique (NN):

It is very simple, highly effecient and effective in the field of pattern recognition, text categorization object recognition etc.

It's simplicity is the main advantage!

This technique is broadly classified into -

- ⊙ structure less technique
- ⊙ structure based technique
- ⊙

- structure less technique is involved in weight KNN, modelbased KNN, condensed NN, reduced NN.

- structure base technique is involved in k-d tree, ball tree, principal axis tree etc.

7) b) Data Warehousing Data Mining

3) a) Need of feature selection:

Feature selection helps in solving two major problems -

» Having too much data that is of little value or having to little data is of high value. Feature selection identify the minimum numbers of colounms from the data sources that are significant in building a model

» It removes the irrelevant data, improves learning accuracy, reduces the computation time and facilitate an enhances the Learning of model.

3) a) Process of Feature Selection:

- **Filter methods:** They collects the fundamental properties of the features that one measured through statistics instead of using cross-validation performance. Dealing with higher dimensional data, is computationally cheaper to use filter methods.

- **supervised method** These methods are used for labeled data, and one also used to classify the relevant features for increasing the effeciency of supervised models such as classification and regression.

- **Embedded method:** They cover the advantages of both filtering and wrapper method by not only compromising interaction of features but also by retaining a re-assemble computational cost.

- **Wrapper method** : This method is used to search the space of all possible subsets of features, accessing a classifier with that subset feature subset and evaluating their quality by learning, the feature selection process is based on Machine Learning algorithm that one fines to fit on a given database.