A quality product by
Brainheaters^{TM}

Bh.Notes: DMBI

IT Semester 6

A series of Important Concepts/Questions highly recommended for MU Exam

'C' SCHEME - 2019-2020

Q1.What is data mining? Explain the KDD process with a diagram.

(P4-Appeared 1 time)(3-7M)

Ans : Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

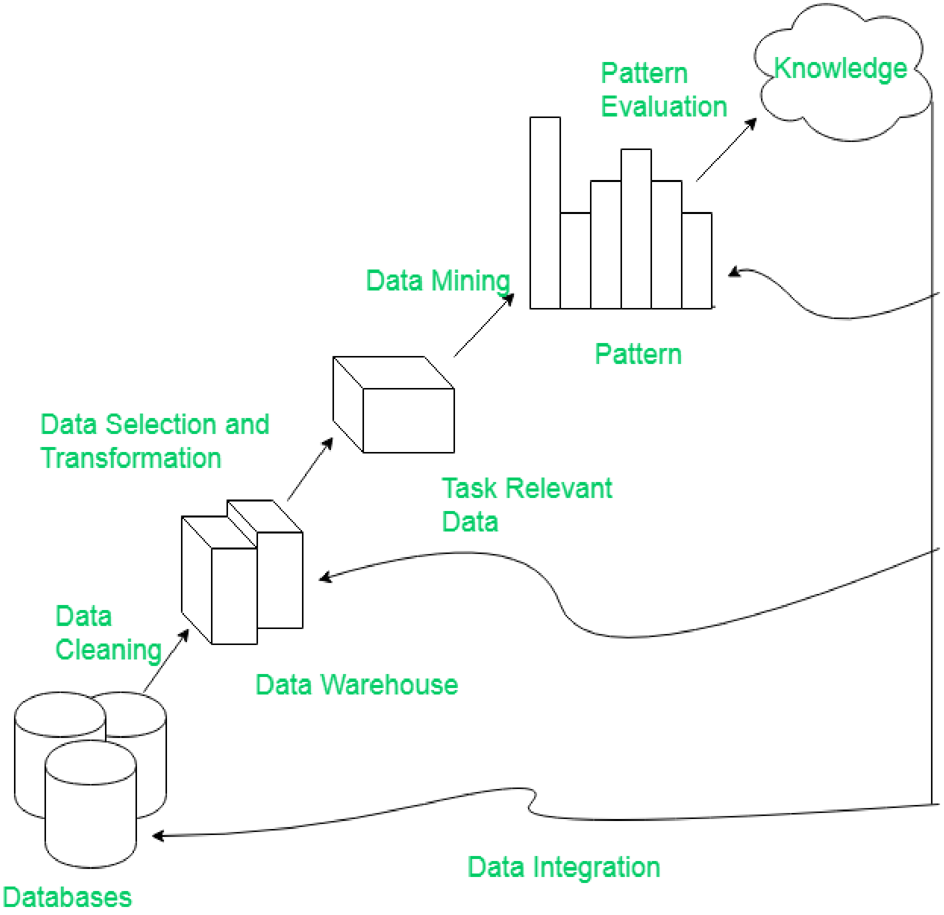
- Data Mining, also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

KDD process

- Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.
  - Cleaning in case of Missing values.
  - Cleaning noisy data, where noise is a random or variance error.
  - Cleaning with Data discrepancy detection and Data transformation tools.
- Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
  - Data integration using Data Migration tools.
  - Data integration using Data Synchronization tools.
  - Data integration using ETL(Extract-Load-Transformation) process.
- Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
  - Data selection using Neural network.

    - Data selection using Decision Trees.
    - Data selection using Naive bayes.
    - Data selection using Clustering, Regression, etc.
- Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.
  Data Transformation is a two step process:
    - Data Mapping: Assigning elements from source base to destination to capture transformations.
    - Code generation: Creation of the actual transformation program.
- Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
    - Transforms task relevant data into patterns.
    - Decides purpose of model using classification or characterization.
- Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
    - Find interestingness score of each pattern.
    - Uses summarization and Visualization to make data understandable by user.
- Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
    - Generate reports.
    - Generate tables.
    - Generate discriminant rules, classification rules, characterization rules, etc.

Steps Involved in KDD Process:



Q2.What is data visualization? Explain any 3 visualization techniques with examples.(P4-Appeared 1 time)(3-7M)

Ans :Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

- In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Common general types of data visualization:
Line chart

- A line chart illustrates changes over time. The x-axis is usually a period of time, while the y-axis is quantity. So, this could illustrate a company's sales for the year broken down by month or how many units a factory produced each day for the past week.

Area chart

- An area chart is an adaptation of a line chart where the area under the line is filled in to emphasize its significance. The color fill for the area under each line should be somewhat transparent so that overlapping areas can be discerned.

Bar chart

- A bar chart also illustrates changes over time. But if there is more than one variable, a bar chart can make it easier to compare the data for each variable at each moment in time. For example, a bar chart could compare the company's sales from this year to last year.

Histogram

- A histogram looks like a bar chart, but measures frequency rather than trends over time. The x-axis of a histogram lists the "bins" or intervals of the variable, and the y-axis is frequency, so each bar represents the frequency of that bin. For example, you could measure the frequencies of each answer to a survey question. The bins would be the answer: "unsatisfactory," "neutral," and "satisfactory." This would tell you how many people gave each answer.

Scatter plot

- Scatter plots are used to find correlations. Each point on a scatter plot means "when x = this, then y equals this.
- " That way, if the points trend a certain way (upward to the left, downward to the right, etc.) there is a relationship between them. If the plot is truly scattered with no trend at all, then the variables do not affect each other at all.

Bubble chart

- A bubble chart is an adaptation of a scatter plot, where each point is illustrated as a bubble whose area has meaning in addition to its placement on the axes. A pain point associated with bubble charts is the limitations on sizes of bubbles due to the limited space within the axes. So, not all data will fit effectively in this type of visualization.

Pie chart

- A pie chart is the best option for illustrating percentages, because it shows each element as part of a whole. So, if your data explains a breakdown in percentages, a pie chart will clearly present the pieces in the proper proportions.

Gauge

- A gauge can be used to illustrate the distance between intervals. This can be presented as a round clock-like gauge or as a tube type gauge resembling a liquid thermometer. Multiple gauges can be shown next to each other to illustrate the difference between multiple intervals.

Q3.Short note on Data transformation and Data(P4-Appeared 1 time)(3-7M)

Ans: Data transformation is the process of changing the format, structure, or values of data. For data analytics projects, data may be transformed at two stages of the data pipeline.

- Organizations that use on-premises data warehouses generally use an ETL (extract, transform, load) process, in which data transformation is the middle step.

Benefits and challenges of data transformation

Transforming data yields several benefits:

- Data is transformed to make it better organized. Transformed data may be easier for both humans and computers to use.
- Properly formatted and validated data improves the data quality and protects applications from potential landmines such as null values, unexpected duplicates, incorrect indexing, and incompatible formats.
- Data transformation facilitates compatibility between applications, systems, and types of data. Data used for multiple purposes may need to be transformed in different ways..

However, there are challenges to transforming data effectively:

- Data transformation can be expensive. The cost is dependent on the specific infrastructure, software, and tools used to process data. Expenses may include those related to licensing, computing resources, and hiring necessary personnel.
- Data transformation processes can be resource-intensive. Performing transformations in an on-premises data warehouse after loading, or transforming data before feeding it into applications, can create a computational burden that slows down other operations. If you use a cloud-based data warehouse, you can do the transformations after loading because the platform can scale up to meet demand.

- Lack of expertise and carelessness can introduce problems during transformation. Data analysts without appropriate subject matter expertise are less likely to notice typos or incorrect data because they are less familiar with the range of accurate and permissible values. For example, someone working on medical data who is unfamiliar with relevant terms might fail to flag disease names that should be mapped to a singular value or notice misspellings.
- Enterprises can perform transformations that don't suit their needs. A business might change information to a specific format for one application only to then revert the information back to its prior format for a different application.

How to transform data

- Data transformation can increase the efficiency of analytic and business processes and enable better data-driven decision-making.
- The first phase of data transformations should include things like data type conversion and flattening of hierarchical data. These operations shape data to increase compatibility with analytics systems.
- Data analysts and data scientists can implement further transformations additively as necessary as individual layers of processing.
- Each layer of processing should be designed to perform a specific set of tasks that meet a known business or technical requirement.

Q4. Explain Regression. Explain linear regression with an example.

(P4-Appeared 1 time)(3-7M)

Ans: Regression is a statistical method used in finance, investing, and other disciplines that attempt to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

- Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.
- The two basic types of regression are simple linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis.
- Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome.
- Linear Regression is a predictive model used for finding the linear relationship between a dependent variable and one or more independent variables.
- Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables, in particular, are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?
- These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c =

constant, b = regression coefficient, and x = score on the independent variable.

- Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.
- Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

Q5.Explain the DBSCAN clustering algorithm with an example (P4-Appeared 1 time)(3-7M)

Ans :DBSCAN is a clustering algorithm that defines clusters as continuous regions of high density and works well if all the clusters are dense enough and well separated by low-density regions.

- In the case of DBSCAN, instead of guessing the number of clusters, will define two hyperparameters: epsilon and minPoints to arrive at clusters.
    1. Epsilon (ε): A distance measure that will be used to locate the points/to check the density in the neighbourhood of any point.
    2. minPoints(n): The minimum number of points (a threshold) clustered together for a region to be considered dense.
- Algorithms start by picking a point(one record) x from your dataset at random and assign it to a cluster 1. Then it counts how many points are located within the ε (epsilon) distance from x. If this quantity is greater than or equal to minPoints (n), then considers it

as core point, then it will pull out all these ε-neighbours to the same cluster 1.

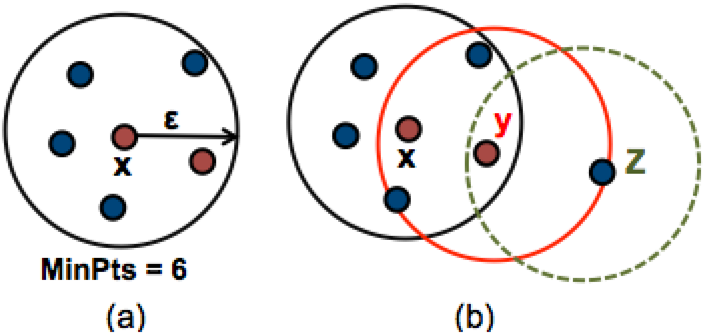
- It will then examine each member of cluster 1 and find their respective ε -neighbours. If some member of cluster 1 has n or moreε-neighbours, it will expand cluster 1 by putting those ε-neighbours to the cluster. It will continue expanding cluster 1 until there are no more examples to put in it.

DBSCAN Parameter Selection

- DBSCAN is very sensitive to the values of epsilon and minPoints. Therefore, it is important to understand how to select the values of epsilon and minPoints. A slight variation in these values can significantly change the results produced by the DBSCAN algorithm.
  - minPoints(n):
    - As a starting point, a minimum n can be derived from the number of dimensions D in the data set, as n ≥ D + 1. For data sets with noise, larger values are usually better and will yield more significant clusters.
    - Hence, n = 2·D can be evaluated, but it may even be necessary to choose larger values for very large data.
  - Epsilon(ε):
    - If a small epsilon is chosen, a large part of the data will not be clustered.
    - Whereas, for a too high value of ε, clusters will merge and the majority of objects will be in the same cluster.
    - Hence, the value for ε can then be chosen by using a k-graph, plotting the distance to the k = minPoints-1 nearest neighbour ordered from the largest to the smallest value. Good values of ε are where this plot shows an “elbow”:
  - Distance Function:

- By default, DBSCAN uses Euclidean distance, although other methods can also be used (like great circle distance for geographical data).
- The choice of distance function is tightly linked to the choice of epsilon (ε) value and has a major impact on the outcomes. Hence, the distance function needs to be chosen appropriately based on the nature of the data set.

- Core Point(x): Data point that has at least minPoints (n) within epsilon (ε) distance.
- Border Point(y): Data point that has at least one core point within epsilon (ε) distance and lower than minPoints (n) within epsilon (ε) distance from it.
- Noise Point(z): Data point that has no core points within epsilon (ε) distance.



Q6.Explain Market Basket Analysis(P4-Appeared 1 time)(3-7M)

Ans :Market basket analysis is an analysis conducted to determine which products the customer purchases together.

- Market basket analysis is a mathematical modeling technique based upon the theory that if u buy a certain group of items you are likely to buy another group of items.
- Frequent item set mining leads to the discovery of associations and correlations among items in large transactional or relational data sets.
- With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases.
- The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes.
- A typical example of frequent item set mining is market basket analysis.
- This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets"
- The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.
- market basket analysis may help you design different store layouts.
- In one strategy, items that are frequently purchased together can be placed in proximity in order to further encourage the sale of such items together.
- In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way
- Market basket analysis can also help retailers plan which items to put on sale at reduced prices.

Eg:

Tid	Items
01	a c d f g i m p
02	a b c f l m o
03	b f h j o
04	b c k p s
05	a c e f l m n p

FREQUENT ITEMSET

Given a support threshold of 3 frequent itemsets are :

A	3
B	3
C	4
F	4
M	3
P	3
Ac	3
Af	3
Am	3
Cf	3

Cm	3
Cp	3
Fm	3
Acf	3
Acm	3
Afm	3
Cfm	3

Q7.Explain Business Intelligence issues.(P3-Appeared 2 time)(3-7M)

Ans :There are three major issues in Business Intelligence:

Data mining methodology:

- Mining different kinds of knowledge from diverse data types.
- Performance: efficiency, effectiveness and scalability.
- Pattern evaluation: The interestingness pattern.
- Incorporation of background pattern.
- Handling noise and incomplete data.
- Parallel, distributed and incremental mining methods.
- Integration of the discovered knowledge with existing one : Knowledge fusion.

User interaction:

- Data mining query languages and ad-hoc mining.
- Expression and visualization of resultant knowledge.
- Interactive mining of knowledge at multiple levels of abstraction.

Applications and social impact:

- Domain specific data mining and invisible data mining.
- Protection of data security, integrity and privacy.

Q8. more questions are available in Brainheaters app….

.

.

.

.