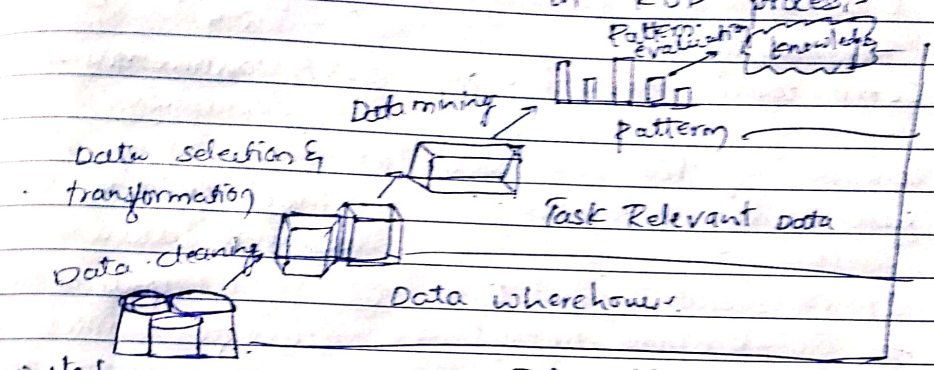
2017

Describe the steps involved data mining when viewed as a process of knowledge discover.

=> steps involved in KDD process:-



(1) Data Cleaning: The noise & inconsistent data is removed.

(2) Data Integration: multiple Data Source are ~~combid~~ Combined

(3) Data selection: Data relevant to the analysis task are retrived from database

(4) Data transformation:- Data is transformed or consolidated into forms appropriate for mining by performing Summary or aggregating opreation

(5) Data mining:- Intelligent methods are applied in order to extract data patterns

(6) Pattern Evaluation: Data patterns are evaluated

(7) Knowledge prentation: visualization & Knowledge.

2020-05-15 09:53

Is representation techniques are used to present mined knowledge to users.

5 Discuss the distinguishing features of OLTP & OLAP

OLAP	OLTP
OLAP online analytical processing	OLTP online transaction processing.
1) consists of historical data from various databases.	1) consists of operational oriented used for business current data
2) It is subject oriented. Used for data mining, analytics, decision making etc.	2) It is applicationed used for business tasks.
3) The data is used in planing problem solving & decision making	3) The data is used to perform day to day fundamental operations.
4) It reveals a snapshot of present business tasks	4) It provides a multi-dimensional view of different business task.
5) Large amount of data is stored typically in TB. PB.	5) The size of the data is relatively small as the historical data is advised for Ex: GB, MB.
6) Relatively slow as the amount of data involved is large. Queries may take hours.	6) very fast as the queries operate on 5% of data
7) It only need backup from time to time as compared to OLTP	7) Backup & recovery process is maintained religously.

2020-5-5 08:52

⑧ This data is generally managed by CEO, MG. Gm.	⑧ This data is managed by clerks, manages
⑨ Only read & rarely write operation.	⑨ Some read & write operations.

⑤ Explain star schema with an example of sales data warehouse

—> In the star schema, the center of the star can have fact table & a number of associated dimension tables. It is known as star schema as its structure resembles a star. The star schema is the simplest type of Data ware-house schema. It is also known as star join schema & is optimized for querying large data sets.

* Each dimension in a star schema is represented with only one dimension table.
* This dimension table contains the set of attributes
* The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, & location
* There is a fact table ~~that~~ at the center. It contains the keys to each of dimensions
* The fact table also contains the attribute namely dollars sold & units sold.

Branch.
Dimension table
• branch : key
branch-name
branch-type

• location
Dimension - key
location - key
street
city
province - or - state
country.

2020-5-5 08:52

4) write a note on meta-rule guided mining association rule.

→ "How are metarules useful?" metarules allow user to specify the syntactic form of rules that they are interested in mining. The rules forms can be as constraints to help improve the efficiency of the mining process metarule less may be based on the anaysis experience excepectations, or intertion regarding the data or may be automating generatted based on the database schema.

Meta-rule-guided Mining.

Suppose that as a market analysis for All Electronics, you have access to the data describing customers (such as customer age, adaress, & credit rating) as well as the list of customer transactions. You are intersted in finding associations b/w customer traits & the items that customers buy. However, rather than finding all of the association rules reflecting these relationship

Scanned with CamScanner

2020-5-5 08:52

* Information Gain: This measure is based on pioneering work by Claude Shannon on Information theory, which studied the value or Information content of message. Let node N represents or hold the tuple of partition D. The attribute with the highest information gain is choosen as the splitting attribute for the node N. The expected information Needed to classify a type in D is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i.$$

where p_i is the probability that an arbitrary tuple in D belongs to class C_i & is estimated by $|C_i, D| / |D|$ Info(D) is the average amount of information needed to identify the class label of a tuple in D. Info(D) is also known as the entropy of D The expected information required to classify a tuple from D. based on the prititioning by attribute A is calculate by.

$$Info(D) = \sum_{j}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Information gain is defined as the difference b/w the original information requirment (i.e obtained after partitioning on A)

$$Gain(A) = Info(D) - Info_A(D)$$

Q. What is clustering? Explain K means clustering technique.

Clustering is the process of making a group of abstract objects into classes of similar objects.

K means clustering Algorithm:

K means simplest is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple & easy way to classify a given data set through a certain number of clusters (assume k-cluster) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set & associate it to the nearest center. When no point is pending, the 1st step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the cluster resulting from the previous step. After we have these k new centroids a new binding has to be done between the same data set points & the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step

until no more changes are done or in other words centers do not move any more. finally, this algorithm aims at minimizing an objective function known as squared error function given by

$$J(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\| x_i - v_j \|)^2$$

where '$\| x_i - v_j \|$' is the Euclidean distance b/w x_i & v_j

'c_i' is the no of data points in i^{th} cluster

'c' is the no of cluster centers.

Algorithmic steps for k-means clustering:-

Let $X = \{x_1, x_2, x_3 \dots x_n\}$ be the set of data points and $V = \{v_1, v_2 \dots v_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.

2) ~~Then~~ calculate the distance b/w each data point and clusters centers.

3) Assign the data point to the cluster center whose distance from the cluster is minimum of ~~the~~ all the cluster centers.

4) Recalculate the new cluster center using

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_j$$

where c_i represents the no of data points in cluster

2020-5-5 08:51

(c) If no data point was reassigned then stop, otherwise repeat from step 3.

(3) Explain the various methods for obtaining reliable classifier accuracy estimates.

=> Random partition into train and test parts:

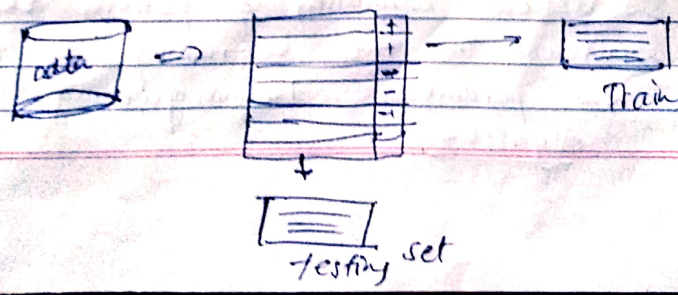
(a) Hold out
* use two independent data sets, eg. training set (2/3) test set (1/3) random sampling
* repeated hold out.

(b) k-fold cross-validation
* randomly divide the data set into k subsamples
* use k-1 subsamples as training data & one sub sample as test data - repeat k times

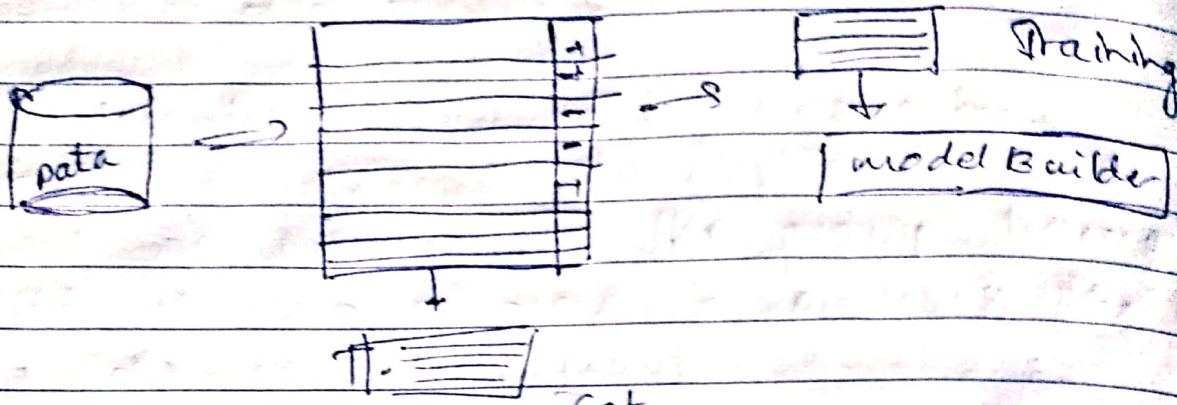
(c) leave-one-out for small size data.

Evaluation on "LARGE" data, hold-out
* A simple evaluation is sufficient.
* Randomly split data into training & test sets (usually 2/3 for train, 1/3 for test)
* Build a classifier using the train set & evaluate it using the test set.

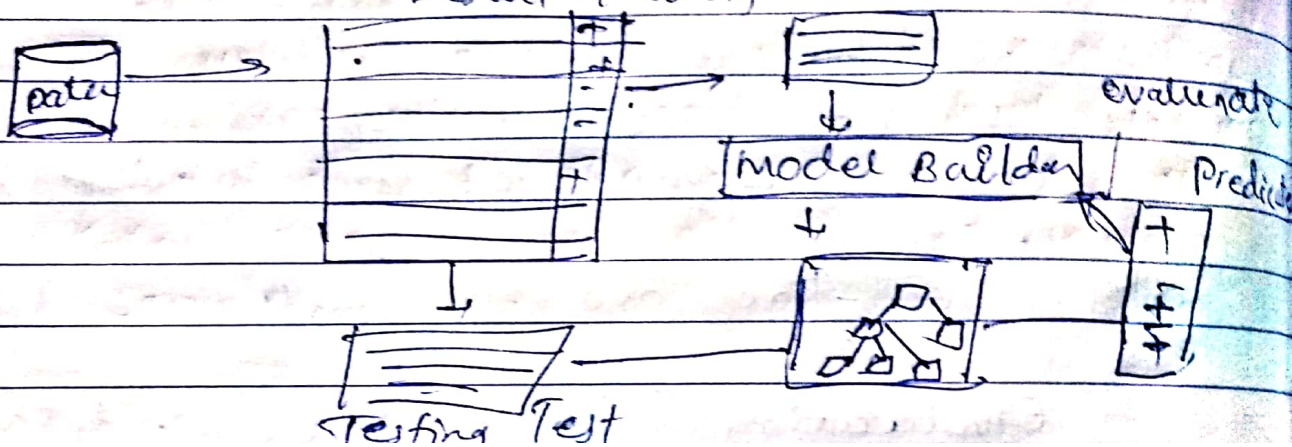
Steps: Split data into train and test sets



Step 2 :- Build a model on a training set



Step 3: Evaluate on test set.



A query is classified using majority vote of its neighbors being assigned to the most common class among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the case is simply assigned to the class of its nearest neighbor.

Euclidean distance:

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Manhattan distance

$$\sum_{i=1}^{k} |x_i - y_i|$$

Minkowski:

$$\left[\sum_{i=1}^{p} (|x_i - y_i|)^q \right]^{1/q}$$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 & 1 when there is a mixture of numerical and categorical variables in the data set.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

→ Always needs to determine the value of K which may be complex same time

→ The computation cost is high because of calculating the distance of the data points for all the training samples.