04/01/2022

Data warehousing

Collecting data (raw facts) from multiple sources available in multiple geographical location and organizing them properly in one place by gathering them to use according to need.

06/01/2022

MODULE - 2

DATA PREPROCESSING

* The process of making the data more suitable for data mining.
* The tasks employed in this process are informed by the process of data Understanding
* Data Mining → process of getting useful insights from the data
  → classification + prediction + clustering + Association Rule Mining

What is Data?

* Collection of data objects and their attributes
* An attribute is a property (or) characteristic of an object
  → Eg: Eye colour of person, Temperature, etc
  → also known as variable, field, characteristic (or) feature
* A collection of attribute describe an object
  → object is also known as record, point, case, sample, entity (or) instance

Missing values

* Absence of information in instances, which brings harmful consequences to the validity of the subsequent analyzes.

* No missing value → complete cases
  Missing values → incomplete cases

* If value is not missing, the checking its validity / correctness is an important preprocessing task which will cause crucial role in Decision making

Types of Attribute

Nominal : Eg: ID numbers, eye colour, zip codes
No order

Ordinal : Eg: Rankings, grade, size
order exist

Interval : Eg: Calender dates, temperature
range of values / duration

Ratio : Eg: Temperature in kelvin, length, time, counts

Properties of Attribute values

* Nominal → distinctness
* Ordinal → distinctness & order
* Interval → distinction, order & addition
* Ratio → all 4 properties

distinctness	= ≠
order	< >
Addition	+ −
Ratio	* /

Types of Data Set

* Records
  -
  -
  -
*
  -
  -
*
  -
  -

Document Data

Each document becomes a 'term' vector,

- each term is a component (attribute) of the vector
- the value of each component is the no. of times the corresponding term occurs

Transaction Data

A special type of data, where (record)

- each record

Data Quality:

Examples of data Quality problems

- Noise and outliers
- Missing values
- Duplicate Data

11/01/2022

DATA REDUCTION

1. Sampling: selecting a subset of the data objects to be analyzed
2. Feature Selection: selecting a subset of the features to be analyzed. (important features)
3. Dimensionality Reduction: Creating new features that are a combination of old feature

DATA SMOOTHING

Sorted data for price ($) : 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* partition into equal-frequency (equi-depth) bins:
  - Bin 1 : 4, 8, 9, 15
  - Bin 2 : 21, 21, 24, 25
  - Bin 3 : 26, 28, 29, 34

* Smoothing by bin means:
  - Bin 1 : 9, 9, 9, 9
  - Bin 2 : 23, 23, 23, 23
  - Bin 3 : 29, 29, 29, 29

mean

* Smoothing by bin boundaries:
  - Bin 1 : (4), 4, 4, (15)
  - Bin 2 : (21), 21, 25, (25)
  - Bin 3 : (26), 26, 26, (34)

* Boundaries are fixed
* Middle numbers are replaced with boundary values to which they are close to

DATA PREPROCESSING

1. Aggregation
2. sampling
3. Feature selection

1. Aggregation

* combining two (or) more attribute into a single attribute (or) object
* purpose
  - Data reduction
  - Change of scale
  - More stable data
* ~~Ex~~ Register number (VIT)
  19MID0031

2. Sampling

* main technique employed for data selection
* often used for both preliminary investigation of data and the final data analysis
* The key principle for effective sampling
  - using a sample will work almost as well as using the entire data sets

* Types of sampling

- Simple Random sampling → cannot ensure

  - There is an equal probability of selecting any particular item

  - sampling without replacement

    As each item is selected, it is removed from the population.

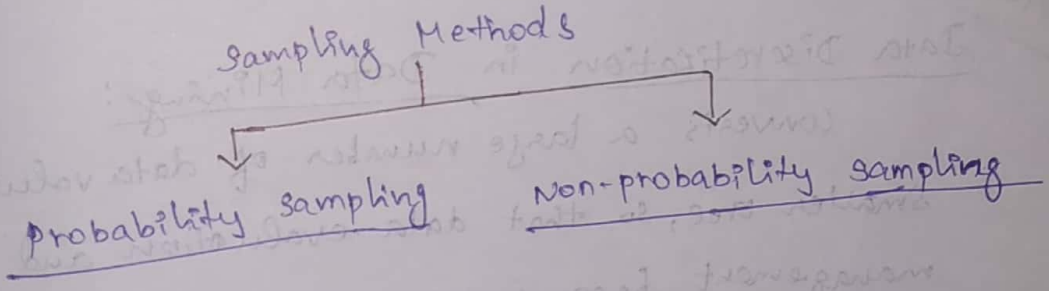
  - sampling with replacement

    • objects are not removed from population as they are selected for the sample

    • objects can be picked more than once

- stratified sampling

  • split the data into several partition



3. Feature Selection

- Select a minimal set of features such that the probability distribution of the class is close to the one obtained by all the features.
- A good feature vector is defined by its capacity to discriminate between examples
  - Maximize the inter-class separation and minimize the intra class separation

13/01/2022

Missing values

Equivalent ratio

3	10
6	x
9	30
y	40

$\frac{3}{10} = \frac{6}{x} \Rightarrow x = 20$

$\frac{9}{30} = \frac{y}{40} \Rightarrow y = 12$

17/01/2022

Data Discretization in Data Mining:

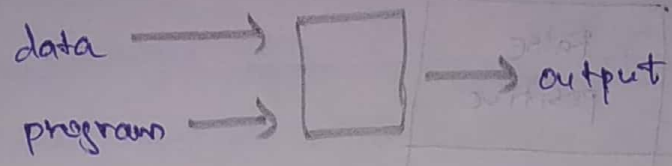
converts a large number of data values into smaller one, so that data evaluation and data management becomes very easy

Eg: converting continuous attribute "Age" into discrete attribute "Age category" consisting {"Young", "Mature", "Old"}

Binarization:

Mapping a categorical attribute to a set of attributes that are binary

Traditional Learning



Machine Learning:

Herbert Simon

Learning is any process by which a system improves performance from experience.

Samuel's checkers-player

01/02/2022

~~DECISION TREE - ID3~~

CONFUSION MATRIX:

Prediction \ Actual		
	True positive	False positive
	False Negative	True Negative

True positive rate = $\frac{TP}{TP+FN}$

False positive rate =

23/02/2022

NAIVE BAYES ANALYSIS

* Supervised learning technique that uses probability theory based analysis

* Machine learning Technique that computes the probabilities of an instance belonging to each one of many target classes, given the prior probabilities of classification using individual factors

* Used often in classifying text document into one of multiple predefined categories.

* The posterior probability (of belonging to class K) is calculated as a function of prior probabilities and ~~calculated~~ current likelihood value, as shown in the equation

$$P(C_k|x) = \frac{P(C_k) * P(x|C_k)}{P(x)}$$

$P(C_k|x)$ → posterior probability of class k, given predictor x.

$P(C_k)$ → prior probability of class k.

$P(x)$ → prior probability of predictor

$P(x|C_k)$ → current likelihood of predictor given class

Step-1

$P(\text{play tennis} = \text{Yes}) = \frac{9}{14} = 0.64$

$P(\text{play tennis} = \text{No}) = \frac{5}{14} = 0.36$

Step-2

OUTLOOK	Y	N
Sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5

HUMIDITY	Y	N
High	3/9	4/5
Normal	6/9	1/5

TEMPERATURE	Y	N
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

WINDY	Y	N
strong	3/9	3/5
weak	6/9	2/5

Step-3

$$V_{NB} = \underset{v_j \in [Yes, No]}{\text{argmax}} \; P(v_j) \prod_i P(a_i | v_j)$$

$$= \underset{v_j \in [Yes, No]}{\text{argmax}} \; P(v_j) \Big[P(\text{Outlook} = \text{sunny} | v_j) \times P(\text{Temperature} = \text{cool} | v_j) \times P(\text{Humidity} = \text{high} | v_j) \times P(\text{wind} = \text{strong} | v_j) \Big]$$

$$V_{NB}(Yes) = P(Yes) \cdot P(\text{Sunny} | Yes) \cdot P(\text{Cool} | Yes) \cdot P(\text{high} | Yes) \cdot P(\text{strong} | Yes)$$

$$= \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = 0.0053$$

$$V_{NB}(No) = P(No) \cdot P(\text{Sunny} | No) \cdot P(\text{Cool} | No) \cdot P(\text{High} | No) \cdot P(\text{Strong} | No)$$

$$= \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} =$$

$$V_{NB}(Yes) = \frac{V_{NB}(Yes)}{V_{NB}(Yes) + V_{NB}(NO)}$$

$$V_{NB}(NO) = \frac{V_{NB}(NO)}{V_{NB}(Yes) + V_{NB}(NO)}$$

Scanned with CamScanner