26/8/21

DWDM

Tutorial - 2

Sahil Bondre - U18CO021

Q1 Needs of Preprocessing Data:

i Incomplete Data: Lacking attribute values, lacking attributes of interest or containing errors or outliers.

Eg: 'salary' = NA

ii Noisy: Data containing errors or outliers.

Eg: 'age' = -40 ; 'salary' = True

iii Inconsistent: Data containing discrepencies in codes & data

Eg: PersonA.age = 45 & PersonB.dob = 1997
CarA.state = 'GUJARAT' &
CarB.state = 'GJ'

Q2 Factors compromising data quality:

i Accuracy
ii Completness
iii Consistency
iv Believability
v Interpretability
vi Accessibility
vii Timeliness.

Q3 Tasks in Data Preprocessing

1 Data Cleaning: This task involves filling of missing values, smoothing or removing noisy data & outliers along with inconsistencies.

Eg: Data collected from sensors can contain no noise which needs to be cleaned.

2 Data Integration: This task involves integrating data from various sources.

Eg: Data can be present in sensors, databases, third party APIs which needs to be combined

3 Data Transformation: This involves normalisation & aggregation of data according to the needs of dataset

Eg: Many analytical modes perform better with normalised data

4 Data Reduction: In this steps, the irrelevant records, attributes & dimensions can be removed

Eg: color attribute in housing dataset will not have much effect on final price of the house & can be removed

5 Data Discretization: Numerical attributes are replaced with normal one

Eg: Exact longitude & latitude coordinates can be replaced by binning into regions.

Q4 Data Cleaning Methods:

1 Listwise Deletion: Entire record is excluded if even one attribute is missing

2 Pairwise Deletion: Pairwise columns are taken & only if attributes from these columns are absent then the record is excluded

3 Single Value Imputation: Missing value is replaced by a single value

4 Hot Deck Imputation: Find all samples similar to other variables then randomly choose one of their values

5 Cold Deck Imputation: Systematically choose the values from an individual with similar values on other variables

6 KNN based Imputation : Finding nearest cluster using KNN & using the value of the cluster

7 Regression Based Imputation : Regression is used to predict the missing value

8 Multiple Imputation : Imputation technique that assigns several imputed values to each missing value & take mean of the result

Q5

i Mean : It is the distributive measure of data

$$\mu = \frac{\sum x_i}{N}$$

Weighted Mean : $\bar{x} = \frac{\sum \omega_i x_i}{\sum \omega_i}$

Eg: $S = \{1, 2, 2, 2, 3, 7, 8\}$

$$\mu = \frac{\sum s_i}{n} = \frac{25}{7} = 3.57$$

ii Median : It is the midpoint of the list when the distribution is in sorted order

Eg: $S = \{1, 2, 2, 3, 7, 8\} \Rightarrow m = \frac{2+3}{2} = 2.5$

iii. **Mode**: It is defined as the most frequently occuring score

Eg: $S = \{1, 2, 2, 2, 3, 7, 8, 8\}$

$M = 2$