

Data Mining and Analytics Part-2

Implement algorithms for the following using any programming languages of your choice out of C, C++, Java, Python

NDEX

SN		
1		Computing dissimilarity (distance) between objects described by numeric data in one-, two-, three-, or higher dimensional space. Han's book (page no. 72-74) Tan's book (page no. 96-99)
		Hamming distance
		Euclidean distance
		Manhattan distance
		Minkowski distance
		Supermum distance
2		Proximity (closeness) measure for binary attributes
		Simple Matching Coefficient (SCM) between two objects described by binary attributes
		Dissimilarity between two objects described by binary attributes
		Dissimilarity between two objects described by asymmetric binary attributes
		Jaccard Coefficient
		Cosine Similarity

2	About Data
	Write an algorithm to accept age data from the monitor. The data should be integer values representing age of all groups of people present in the society from 5 to 95 years. Take about 50 data values. (Q. No. 2.2 on page no. 80, Han book)
	a Arrange data in ascending order
	b Compute mean, median and mode of the data.
	c Compute percentile of a data item from the formed dataset.
	d Compute 70 and 80 percentiles.
	e Compute 70 th and 80 th percentiles.
	f Compute midrange of the data.
	g Compute 1 st and 3 rd quartiles of the dataset.
	h Find 1 st and 3 rd quartiles of the dataset.
	i Give five number summary of the data
	j Repeat experiment from (a) to (i) at least five times by taking different data values.
3	Association Rule Mining Han's book (page no. 248 to 253)
	<p>a Objective: To compute maximal frequent itemset.</p> <p>Compute candidate 3-itemsets from frequent 2-itemsets using join $C_3 = L_2 \times L_2$. (Han's book example)</p> <p>Generalize the algorithm for generating candidate C_{i+1} itemsets from frequent L_i itemsets $C_{i+1} = L_i \times L_i$</p>
	<p>b Objective: To develop prune operation using apriory property.</p> <p>Prune unnecessary 3-itemsets from the set of generated 3-itemsets C_3 to make C_3 to set of frequent 3-itemsets L_3. (Han book example)</p> <p>Generalize the algorithm for pruning unnecessary i-itemsets from the set of generated i-itemsets C_i to make C_i to set of frequent i-itemsets L_i.</p>

	c	Write Apriori algorithm using the above join and prune procedures.
4		FP Growth algorithm.....
	a	Objective: To compute maximal frequent itemset. Compute item set of frequent items (1-itemsets) and their support counts from a given transactional dataset. Sort frequent itemsets and generate them in L order (descending order of support counts).
	b	Sort items in the transactions of the dataset in L-order (descending order of support counts).
	c	Construct FP tree using the Han's book example. Display FP tree using appropriate notation/representation.
	d	Using FP tree, construct pattern bases and conditional FP trees.
	e	Generate frequent patterns.

Distances

Variable transformation

Kernal function matrix

Central tendency measures

Smoothing

χ^2 correlation test for normal data

Correlation coefficient for numeric data

Covariance of numeric data

Transformation by normalization