Experiment No- 1

Title - Study and use of different types of graphs and charts.

1) Line chart

In a line chart, category data is distributed evenly along the horizontal axis, and all values data is distributed evenly along the vertical axis. Line charts can show continuous data over time on an evenly scaled axis, so they're ideal for showing trends in data at equal intervals, like months, quarters or fiscal years.

2) Column chart

A column chart is a data visualization where each category is represented by a rectangle, with the height of the rectangle being proportional to the values being plotted. Column charts are also known as vertical bar charts.

3) Pie chart.

A pie chart is a pictorial representation of data in the form of a circular chart or pie where the slices of the pie show the size of the data. A list of numerical variables along with categorical variables if needed to represent data in the form of pie chart.

4) Scatter Plot.

Scatter chart are based on basic line chart with x-axis changed to a linear axis. To use a scatter chart, data must be passed as objects containing x & y properties. scatter plots / charts are used to plot data

points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another. A scatter chart uses dots to represent values for two different numeric variables.

5) Radar chart.

A radar chart is a way of showing multiple data points and the variations between them. They are often useful for comparing the points of two or more different datasets. Radar charts are used to compare two or more items or groups on various features or characteristics.

6) Histogram

A histogram is a chart that groups numeric data into bins, displaying the bins as segmented columns They're used to depict the distribution of a dataset, how often values fall into ranges.

Conclusion

Excel provides user-friendly interface for creating and customising various types of charts, including column, line, scatter, pie & radar charts. It's charting capabilities allow you to efficiently visualize & analyze data to gain insight & communicate your findings for others.

Experiment NO-2

Title - To perform Normalization of data (Min-Max) and Z-score.

Data Normalization.

The data normalization (also referred to as data pre-processing) is a basic element of data mining. It means transforming the data, namely converting the source data into another format that allows processing data efficiently. The main purpose of data normalization is to minimize or even exclude duplicated data.

① Min-max normalization.

Steps

i) Identify minimum and maximum values of dataset.
ii) Set new min and max.
iii) For each data point apply normalization formula.

$$x' = \frac{x - min_{old}}{max_{old} - min_{old}} [max_{new} - min_{new}] + min_{new}$$

Dataset - [12, 25, 18, 30, 15]

① Calculate the minimum and maximum value of dataset.

$min_{old} = 12$
$max_{old} = 30$.

② Set new min max

$min_{new} = 10$
$max_{new} = 20$.

③ Calculate normalized values.

Using formula, let's normalize the sample.

$x = 12$

$x' = \frac{12-12}{(30-12)} \times (20-10) + 10$

$\boxed{x' = 10}$

$x = 25$

$x' = \frac{25-12}{30-12} \times (20-10) + 10$

$\boxed{x' = 17.22}$

$x = 18$

$x' = \frac{18-12}{30-12} \times (20-10) + 10$

$\boxed{x' = 13.33}$

$x = 30$

$x' = \frac{30-12}{30-12} \times (10) + 10$

$\boxed{x' = 20}$

$x = 15$

$x' = \frac{(15-12)}{(30-12)} \times (20-10) + 10$

$\boxed{x' = 11.66}$

So the normalized data with min-max normalization is -

$[10, 17.22, 13.33, 20, 11.66]$

(2) z-score normalization.

1) Calculate the mean and standard deviation of the sample data.

2) For each data point, calculate the z-score using formula –

$$x' = \frac{x - \text{mean}}{\text{standard deviation}}$$

Let's perform z-score normalization.

(1) Mean $= \frac{12 + 25 + 18 + 30 + 15}{5}$ $\boxed{\bar{x} = 20}$

Standard deviation.

$$= \sqrt{\frac{(20-12)^2 + (25-20)^2 + (18-20)^2 + (30-20)^2 + (15-20)^2}{5}}$$

$$= 6.16$$

(2) Calculate z-score

$x = 12 \quad x' = \frac{12-20}{6.16} = -1.30$

$x = 25 \quad x' = \frac{25-20}{6.16} = 0.81$

$x = 18 \quad x' = \frac{18-20}{6.16} = -0.32$

$x = 30 \quad x' = \frac{30-20}{6.16} = 1.62$

$x = 15 \quad x' = \frac{15-20}{6.16} = -0.81$

z-score normalized data would be –

$[-1.30, 0.81, -0.32, 1.62, -0.81]$

Conclusion

Normalization is an important process in data preprocessing. It is used to ensure consistency in data records. In order to bring all attribute on the same scale min-max normalization is used. Standardizing scale on some scale by dividing a score deviation by standard deviation for that z-score is used. It scales a data to particular small scale which helps to allow processing data efficiency.

Experiment No- 4

Title- Find info Gain of an attribute from given data.

Theory -

Entropy - Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how the decision tree chooses to split data.
Consider a dataset of N classes, the entropy (E) can be calculated as -

$$E = -\sum_{i=1}^{N} P_i \log_2 P_i$$

where Pi - probability of randomly selecting an example in class i.

Information Gain - Information Gain measures the expected reduction in entropy caused by partitioning the dataset based on a specific attribute. It helps us decide which attribute to use as the root node of a decision tree. The attribute with the highest information gain is choosen as the root node.

Information gain = Entropy (parent) - [weighted average] * entropy (children).

Algorithm

① For calculating entropy

Input - A dataset 'D' with a target variable 'Y'

Output - Entropy of the dataset 'D'

1) Compute the frequency of each class in the dataset 'D'

2) For each class 'i', calculate the probability pi as the ratio of the frequency of class 'i' to the total no. of samples in 'D'

3) Calculate the entropy (D) using the formula -

$$\text{Entropy}(D) = -\sum_{i=1}^{N} pi \log_2(Pi)$$

4) Return the entropy value.

② For information Gain calculation.

1) Calculate the entropy of dataset 'D' using the algorithm described above.

2) For each unique value V in the domain 'A'
- Partition the dataset 'D' into subsets Dv where attribute A takes the value V.
- Calculate the proportion (Dv/D) i.e fraction of samples in Dv out of the total samples in 'D'.
- Calculate the entropy of each subset Dv using the algorithm described above.

3) Calculate the weighted sum of entropies for the subsets.

4) Calculate the information gain

Info. Gain (A, D) = Entropy (D) - $\sum_{V \in \text{values}(A)} \left|\frac{Dv}{D}\right| \cdot \text{Entropy}(Dv)$

5) Return information Gain value.

Example

For Dataset -

Day	Outlook	Temp.	Humidity	Wind	Playgame
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Steong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Steong	No
7	Overcast	Cool	Normal	Steong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Steong	Yes
12	Overcast	Mild	High	Steong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Steong	No

Expected information (T) = Enteopy $\left(\frac{9}{14}, \frac{5}{14}\right)$

$$= -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= 0.9402$$

① Consider the attribute outlook

	Yes	No	
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5
	9	5	

Enteopy (Sunny) $= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.97$

Enteopy (overcast) $= -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) = 0$

$$\text{Entropy (rain)} = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right)$$

$$= 0.9709$$

$$\text{Entropy (outlook)} = \frac{5}{14} \times (0.97) + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.9709$$

$$= 0.6931$$

$$\text{Info Gain (outlook)} = \text{Entropy (T)} - \text{Entropy (outlook)}$$

$$= 0.9402 - 0.6931$$

$$= 0.2471$$

(2) attribute - temperature.

	Yes	No	
Hot	2	2	4
mild	4	2	6
cold	3	1	4
	9	5	

$$\text{Entropy (temperature)} = \frac{4}{14}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] + \frac{6}{14}\left[-\frac{4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right)\right] + \frac{4}{14}\left[-\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)\right]$$

$$= 0.2857 + 0.3935 + 0.2317$$

$$= 0.9109$$

$$\text{Info Gain (temp)} = 0.9402 - 0.9109$$

$$= 0.0293$$

(3) attribute - Humidity

	Yes	No	
High	3	4	7
Normal	6	1	7
	9	5	

$$\text{Entropy (humidity)} = \frac{7}{14}\left[-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right] + \frac{7}{14}\left[-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right]$$

$$= 0.4926 + 0.2958$$

$$= 0.7884$$

Info Gain (humidity) = 0.9402 − 0.7884
= 0.1518

④ attribute - wind.

	Yes	No	
weak	6	2	8
steong	3	3	6
	9	5	

$$\text{Entropy (wind)} = \frac{8}{14}\left[\frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right] + \frac{6}{14}\left[\frac{-3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$$

= 0.4635 + 0.4285
= 0.892

Info Gain (wind) = 0.9402 − 0.892
= 0.0482

Result

Info gain -
1] outlook = 0.2471
2] temp = 0.0293
3] Humidity = 0.1518
4] wind = 0.0482

Conclusion

As the info-gain of attribute outlook is maximum Outlook will be the root node of the decision tree.

Experiment No - 5

Title - Find t and d weight of a data.

Theory -

To represent descriptive data mining results in the form of Rules, two weighted measures t-weight and d-weight are introduced.

T-weight - t-weight of a generalized tuple or object for a given class shows how typical the tuple is of the given class.

d-weight - The d-weight of a tuple shows how distinctive the tuple is in the given class in comparision with its rival class.

Algorithm

1) Find total count for each attribute.
2) for each value, find t-weight take row-wise count for value and find percentage.
3) for each value, find d-weight, take column wise count and find percentage.
4)

Example

Location / Item		T.V			Computer			Both-Item	
	count	t-wt	d-wt	count	t-wt	d-wt	count	t-wt	d-wt
Europe	80	25%	40%	240	75%	30%	320	100%	32%
N-Am	120	17.65%	60%	560	82.35%	70%	680	100%	68%
Both-regions	200	20%	100%	800	80%	100%	1000	100%	100%

Conclusion-

- D- weight helps to find the weight of an attribute across the class.
- T-weight helps to find the weight of an attribute within the class
- T-weight and D-weight, two weighted measures are used to present descriptive mining results.

Experiment No- 6

Title- Find 5 no. summary of the dataset.

Theory -

The five-number summary is a set of descriptive statistics that provides information about a dataset. The 5 number summary is an exploratory data analysis tool that provides insights into the distribution of values for one variable. It consists of the five most important sample percentiles:

1) The sample minimum (smallest observation)
2) the low quartile or first quartile.
3) the median (the middle value)
4) the upper quartile or third quartile.
5) the sample maximum (largest ~~quartile~~) observation)

Formula-

1) Median $= \dfrac{\begin{cases} x\left[\frac{(n+1)}{2}\right] & \text{if } n \text{ is odd} \\ x\left[\frac{n}{2}\right] + x\left[\frac{n}{2}+1\right] & \text{if } n \text{ is even.} \end{cases}}{2}$

where x = ordered list of values in the dataset.
n = number of values in dataset.

2) 1st quartile (Q_1) = Median of lower half of data.
3) 3rd quartile (Q_3) = Median of upper half of data.

Algorithm

1) Take any particular dataset and put in ascending order.
2) Find minimum and maximum for your data.
3) Find median.
4) If no. of values is odd then leave that median value and consider data before and after median to find Q_1 and Q_3
5) If total number of values is even then consider the values that are used for finding median. Before value in first dataset to find Q_1 and after value in second dataset to find Q_3. Q_1 & Q_3 are median of lower & upper dataset respectively.
6) Plot the five number summary of dataset (Boxplot)

Example

Dataset - 2, 4, 5, 8, 10, 11, 1, 1, 2, 6, 6, 7

i) Arrange the data in ascending order.
1, 1, 2, 2, 4, 5, 6, 6, 7, 8, 10, 11.

ii) Here, $n = 12$

iii) maximum = 11
minimum = 1

iv) Median = Avg. of 6th and 7th term.
$= (5+6)/2$
Median = 5.5

v) For Q_1
Take [1, 1, 2, 2, 4, 5]
$Q_1 = \frac{2+2}{2} = 2$

vi) For Q_3

Take - [6, 6, 7, 8, 10, 11]

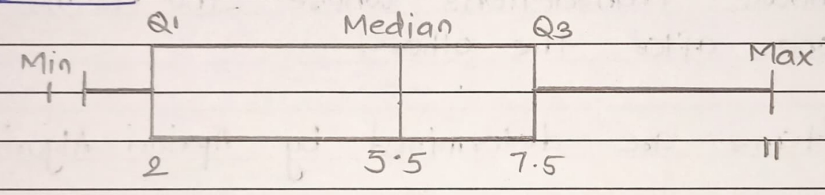
$$Q_3 = \frac{7+8}{2} = \frac{15}{2}$$

$Q_3 = 7.5$

∴ Summary

1) Minimum = 1
2) Maximum = 11
3) Median = 5.5
4) $Q_1 = 2$
5) $Q_3 = 7.5$

- BoxPlot



- Conclusion

The five number summery is an exploratory data analysis tool that provides insights into distribution of values for one variable. It is useful in discriptive analysis or during the preliminary investigation of large ~~files~~ dataset.
It gives a general sense of whether the distribution is symmetric or skewed by comparising Q_1, median and Q_3.

Experiment - 7

Title - Find frequent itemset from given transaction data.

Theory -
When items are grouped together then form an itemset. An itemset that occurs frequently is called a frequent itemset. Frequent itemset mining is a data mining technique to identify the items ~~pur~~ that are often occur together. A set of items is called frequent if it satisfies a minimum threshold value for support and confidence. Support shows transactions with items purchased together in a single transaction Confidence shows transactions where the items are purchased one after the other.

Frequent items are determined by Apriori Algorithm.

Key concepts -

① Support - It refers to the popularity of a product in a transaction. A measure of interestingness. This tells about the usefulness and certainty of rules.

$$\text{Support (A)} = \frac{\text{No. of transaction in which A appears.}}{\text{Total number of transactions.}}$$

② Confidence - Confidence shows the possibilities that the customer bought items one after another in a single transaction.

$$\text{Confidence } (A \rightarrow B) = \frac{\text{Support } (A \cup B)}{\text{Support } (A)}$$

Support_count (x) -
Number of transactions in which x appears. If x is (A ∪ B) it is the number of transactions in which A and B both are present.

* Algorithm

1) Take min_support as input from user
2) Read excel / csv file.
3) Form the itemset of size k where 'k' is the size of the candidate itemsets being considered. $1 \leq k \leq n$, n is total no. of items.
4) For each level calculate frequency of itemset and support_count.
5) If support_count of itemset is greater than minimum support, declare it as frequent itemset and print it.

Example

T_1 = { bread, milk, eggs }
T_2 = { bread, butter }
T_3 = { milk, eggs }
T_4 = { bread, milk, butter }
T_5 = { bread, milk, eggs, butter }

Let -
bread ⇒ a
milk ⇒ b
eggs ⇒ c
butter ⇒ d

Step 1 - For k = 1 Itemset

minimum_support = 75%
minimum_sup_count = $\frac{75}{100} \times 4 = 3$

∴ T_1 = { a, b, c }
T_2 = { a, d }
T_3 = { b, c }
T_4 = { a, b, d }
T_5 = { a, b, c, d }

Step 1 - k=1

Itemset	support_count
{a}	4
{b}	4
{c}	3
{d}	3

As (feequency) support-count of all itemset is greater than minimum support. All itemset are feequent for k=1

ii) For k = 2

Itemset	support_count
{a, b}	3
{a, c}	2
{a, d}	3
{b, c}	3
{b, d}	2
{c, d}	1

compare support_count with minimum support

Itemset	support_count
{a, b}	3
{a, d}	3
{b, c}	3

iii) For k = 3

Itemset	support_count
{a, b, d}	2
{a, b, c}	2

Here as count < min_sup. we will stop iterations here.

∴ {a, b}, {a, d}, {b, c} are frequent itemset for the given dataset.

Advantages and Application of finding the frequent itemset.

1) Association Rule mining -
Frequent itemsets are the basis for generating association rule, which can reveal interesting patterns and relationships in the data. These rules can be used for decision making, cross-selling and marketing strategies.

2) Market Basket Analysis. -
Frequent itemsets helps in understanding which items are often purchased together. This information is valuable for optimizing store layouts, product placements and promotional activities.

3) Product Recommendation -
Knowing which items are frequently purchased together, ~~but~~ businesses can make personalized product recommendations to customers based on their past purchasing behaviour. This can lead to increased sales and customer satisfaction.

Conclusion

Frequent itemset mining shows which items appear together in a transaction or relation frequently.

Experiment - 8

Title - Extend program 7, to find association rules.

Theory -

Association rule mining searches for interesting relationships among items in a given data set. The goal of association rule mining is to identify rules that describe how certain items tends to co-occur or be associated with each other in a given dataset.

Algorithm - Apriori Algorithm.

1) Scan the dataset to determine the support (frequency of occurance) of each item.
2) Set a minimum support threshold (min_support) to filter out infrequent items.
3) Create a list of frequent 1-itemsets based on the minimum support threshold.
4) Generate candidate itemset of size (k+1) by joining pairs of frequent k- itemsets.
5) Eliminate candidate itemsets that contain subsets that are not frequent.
6) Repeat (4) & (5) until no new frequent itemsets can be generated.
7) Use the frequent itemsets to generate association rules that satisfy a specified confidence threshold.

Example -

Consider the following dataset, find frequent itemset and generate association rules.

TID	Items
T_1	I_1, I_2, I_5
T_2	I_2, I_4
T_3	I_2, I_3
T_4	I_1, I_2, I_4
T_5	I_1, I_3
T_6	I_2, I_3
T_7	I_1, I_3
T_8	I_1, I_2, I_3, I_5
T_9	I_1, I_2, I_3

minimum_support count = 2
minimum-confidence = 50%.

Step 1 - k = 1

Itemset	support_count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

As all itemset have support_count > min_support no item is eliminated.

Step 2 - k = 2

Itemset	Support_cnt
I_1, I_2	4
I_1, I_3	4
I_1, I_4	1
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2
I_3, I_4	0
I_3, I_5	1
I_4, I_5	0

Itemset	support_count
I_1, I_2	4
I_1, I_3	4
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2

$k = 3$

Itemset	support_cnt
I_1, I_2, I_3	2
I_1, I_2, I_5	2

$k = 4$

Itemset	support_cnt
I_1, I_2, I_3, I_5	1

If 4-itemset $\{I_1, I_2, I_3, I_5\}$ the support_count < min_support.

∴ We will not consider thie.

Hence, $\{I_1, I_2, I_3\}$ and $\{I_1, I_2, I_5\}$ are feequent itemsets.

For association Rule
we need to find confidence.

$$\text{confidence}(A \rightarrow B) = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$$

So rules can be -	confidence.
$I_1, I_2 \rightarrow I_3$ —	$2/4 \times 100 = 50\%$.
$I_1, I_3 \rightarrow I_2$	$2/4 \times 100 = 50\%$.
$I_2, I_3 \rightarrow I_1$	$2/4 \times 100 = 50\%$.
$I_1 \rightarrow I_2\ I_3$	$2/6 \times 100 = 33\%$.
$I_2 \rightarrow I_1\ I_3$	$2/7 \times 100 = 28\%$.
$I_3 \rightarrow I_1\ I_2$	$2/6 \times 100 = 33\%$.

$I_1, I_2 \rightarrow I_5$	$2/4 \times 100 = 50\%$.
$I_1, I_5 \rightarrow I_2$	$2/2 \times 100 = 100\%$.
$I_2, I_5 \rightarrow I_1$	$2/2 \times 100 = 50\%$.
$I_1 \rightarrow I_2 I_5$	$2/6 \times 100 = 33\%$.
$I_2 \rightarrow I_1 I_5$	$2/7 \times 100 = 28\%$.
$I_5 \rightarrow I_1 I_2$	$2/2 \times 100 = 100\%$.

Result

The association rules whose confidence is greater than minimum_condifidence are -

$I_1, I_2 \rightarrow I_3$
$I_1, I_3 \rightarrow I_2$
$I_2, I_3 \rightarrow I_1$
$I_1, I_2 \rightarrow I_5$
$I_1, I_5 \rightarrow I_2$
$I_2, I_5 \rightarrow I_1$
$I_5 \rightarrow I_1, I_2$

Conclusion

Association Rules show how often products are purchased together. It is useful in analysing the dataset and discovering interesting relationships between entities.

Experiment 10

Title - Distance and cluster.

- To compute center of cluster assuming all multi-dimensional points belonging to one cluster.
- To find distance of all points with obtained cluster centre using suitable distance function.
- To display result in upper triangular or lower triangular matrix.

Theory

k-means clustering is an unsupervised learning algorithm which groups unlabelled dataset into different clusters here 'K' defines the number of predefined clusters that need to be created in the process.

It allows us to cluster the data into different groups & a convenient way to discover the categories or groups in ~~which~~ the unlabelled dataset on its own without the need for any training the algorithm.

It performs mainly on two tasks -

(i) Determine the best value for k-center points or centeoid by an iterative process.

(ii) Assign each data point to its closest k-center those data points which are near to the particular k-center, creates a cluster.

Algorithm

1) Choose number of clusters

11) Randomly select any 'K' data points as cluster center. Select cluster center in such a way that they are farthest.

3) Calculate distance between each point & each cluster center any distance function can be used here such as "Euclidean function".
4) Assign each data point to some cluster. A datapoint is assigned to that cluster whose center is nearest to that point.
5) Recompute center by taking mean.
6) Keep repeating steps 3 & 5 until center do not change or datapoint remain in same cluster or maximum iterations are reached.

Example

let's the points be -

Points	co-ordinates.
P_1	(10, 40)
P_2	(20, 10)
P_3	(15, 20)
P_4	(25, 30)
P_5	(15, 5)

As given in problem statement.
Consider these 5 points as a part of single cluster.
Let cluster center be c(x,y)

$$x = \frac{10+20+15+25+15}{5} = \frac{85}{5} = 17$$

$$y = \frac{40+10+20+30+5}{5} = \frac{105}{5} = 21$$

$$c(x,y) = (17,21)$$

Using Euclidians formula, finding distance.

$$d(c, P_i) = \sqrt{(x-x_i)^2 + (y-y_i)^2}$$

$d(C, P_1) = \sqrt{(17-10)^2 + (40-21)^2} = 20.24$

$d(C, P_2) = \sqrt{(17-20)^2 + (21-10)^2} = 11.40$

$d(C, P_3) = \sqrt{(17-15)^2 + (21-20)^2} = 2.23$

$d(C, P_4) = \sqrt{(17-25)^2 + (21-30)^2} = 12.04$

$d(C, P_5) = \sqrt{(17-15)^2 + (21-5)^2} = 16.12$

The nearest point from imaginary centee is $P_3(15, 20)$

Now finding distances between points considering P_3 as centee & ploting triangular mateix.

Distance mateix	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	31.62	0			
P_3	20.61	11.18	0		
P_4	18.02	20.61	14.14	0	
P_5	35.35	7.07	15	26.82	0
C	20.24	11.40	2.23	12.04	16.12

Conclusion

K-means clustering partitions dataset into k-peedefined distinct non-overlaping subgroups (clusters) where each point belongs to only one group, clustering can be done using suitable distance function and distance mateix points with closenese can be merged together to form a single clustee.

Experiment - 11

Title - Agglomerative hierarchical clustering using single linkage method.

Theory -
A hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a seperate then it repeatedly executes the subsequent steps -
Identify the clusters which can be closest together and merge the maximum comparable clusters. We need to continue these steps untill all clusters are merged together.
In hierarchical clustering, the aim is to produce a hierarchical series of nested clusters. The basic method to generate hierarchical clustering are-

Agglomerative clustering -
Initially consider every datapoint as an individual cluster & at every step merge the nearest pairs of cluster at first every dataset is considered as individual entity or cluster.
At every iteration, cluster merges with different cluster until one cluster is formed.

Algorithm
1) Read the input dataset.
2) Calculate the similarity of one cluster with all other clusters.
3) Consider every datapoint as a individual cluster.
4) Merge the clusters which are highly similar or close to each other.

5) Recalculate the aproximate mateix for each cluster.
6) Repeat step 4 & 5 until only a single cluster remains.

Example

Using single linkage

Dataset -

	A	B	C	D	E	F
A	0					
B	16	0				
C	47	37	0			
D	72	57	40	0		
E	77	65	30	31	0	
F	79	66	35	23	[10]	0

To obtain new distance mateix merge EF

	A	B	C	D	EF
A	0				
B	[16]	0			
C	47	37	0		
D	72	57	40	0	
EF	77	65	30	23	0

merge AB in this Heration.

	AB	C	D	EF
AB	0			
C	37	0		
D	57	40	0	
EF	65	30	[23]	0

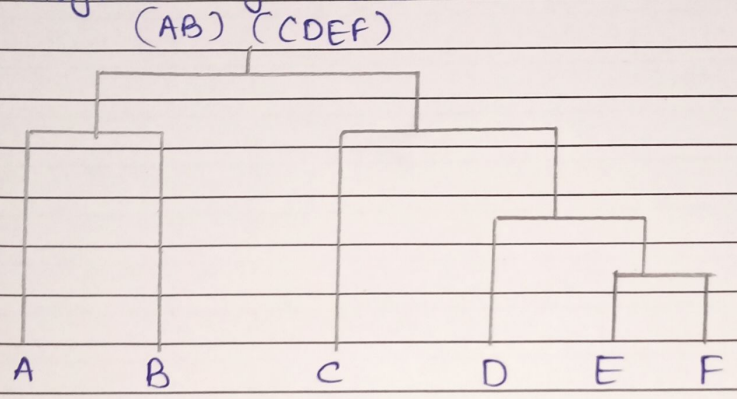
merging D & EF in next Heration.

	AB	C	DEF
AB	0		
C	37	0	
DEF	57	[30]	0

Now merging C & DEF

	AB	CDEF
AB	0	
CDEF	37	0

∴ Dendogram generated.



Conclusion

Hierarchical agglomerative clustering starts with treating each observation as an individual cluster i.e. begins with singleton sets of each point that is each data point is its own cluster & then iteratively merges cluster until all the data points are merged into a single cluster Dendogram is generated to represent hierarchical relationship between object.

Experiment 13

Title - WAP for Baye's classification.

Theory -

Baye's theorem describes the probability of an event, based on precedient knowledge of condition which might be related to the event.
In other words, Baye's theorem is the add on of conditional probability.
With the help of conditional probability one can find out probability of X given H & it's denoted by P(X|H). Baye's theorem states that if you know conditional probability, then we can find out P(X|H). Baye's theorem has two types of probabilities.

- priori probability [P(H)]
- Posterior probability [P(H|X)]

where -
X = data tuple
H = Hypothesis.

Algorithm

1) Read the dataset & take new instance from user.
2) Find the probability of each attribute & note it down.
3) Find conditional probability of new instance using Baye's classification theorem.

Example

No.	color	legs	Height	smelly	species
1	white	3	short	yes	M
2	green	2	Tall	NO	M
3	green	3	short	yes	M
4	white	3	short	yes	M
5	green	2	short	NO	H
6	white	2	Tall	No	H
7	white	2	Tall	No	H
8	white	2	short	yes	H

New Instance -
(colour = Greeen , legs = 2 , Height = Tall & smelly = no)

Here - M = 4 & H = 4

P(M) = 4/8 = 0.5
P(H) = 4/8 = 0.5

color	M	H
white	2/4	2/4
green	2/4	1/4

Height	M	H
short	3/4	2/4
Tall	1/4	2/4

legs	M	H
2	1/4	4/4
3	3/4	0

smelly	M	H
yes	3/4	1/4
NO	1/4	3/4

Now,

P(M | new instance) = P(M) * P(color = green | M) * P(legs = 2 | M) * P(Height = Tall | M) * P(smelly = No | M)

$P(M \mid \text{new\_instance}) = 0.5 \times \frac{2}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$

$= 0.0039$

$P(H \mid \text{new\_instance}) = P(H) * P(\text{color} = \text{green} \mid H) * P(\text{legs} = 2 \mid H) * P(\text{Height} = \text{Tall} \mid H) * P(\text{smelly} = \text{NO} \mid H)$

$= 0.5 \times \frac{1}{4} \times \frac{4}{4} \times \frac{2}{4} \times \frac{3}{4}$

$= 0.04687.$

Here,

$P(H \mid \text{new\_instance}) > P(M \mid \text{new\_instance})$

∴ new-instance belongs to species 'H'

Conclusion

Naive Baye's classification algorithm is a probabilistic classifier. It's useful for making predictions & forecasting data based on historical results by using bayesian classifier. We can classify unknown (new-instance) case by training over unknown data. It helps to specify the class of new instance to which it belongs to.

Experiment NO-14

Title - Implement ~~only~~ any DM concept on complex data type (image, audio, video, time series, multidimensional data).

Theory

Data mining involves exploring & analyzing large blocks of information to ~~given~~ ~~in~~ glean meaningful patterns & teends. Let's take 'Image' dataset Image classification is taken as growing field of both computer vision & data mining. DM technique is applied for image classification.

Algorithm

1) Let's take 3 images for analysis - Normal image, normal image corrupted by gaussian noisy & noisy image applied to filter.
2) This normal image is taken for training model & the other two noisy & filtered images are taken for testing.
3) Now we have added a random noise to normal image. this will used for testing.
4) Now apply data mining concept on images like smoothing to reduce adaptive noise.

Classifier used -

In this experiment, different classifiers are used - decision tree, Naive Bayes & random forest.

Result - Table of root mean square error.

Images	Image Type	Naive Bayes	Random forest
cancer image	Normal	0.216	0.0258
	noisy	0.4351	0.2818
	filtered	0.4282	0.2827
satellite image	Normal	0.1205	0.0099
	noisy	0.5229	0.3539
	filtered	0.5073	0.3371

Conclusion

It's observed that DM concept can be applied on any complex data type. These classifier technique applied to classify region of interest from images in order to get meaningful observation. The filtered image enhances the classification accuracy. Random forest is better for normal & filtered while Naive Bayes for noisy image.

Experiment No - 3

Title - Perform binning of data.

Binning -

Binning is a data pre-processing technique used to categorize or group continuous numerical data into discrete intervals or 'bins'. It can help simply complex data distributions, provide insights and make data visualization easier.

Steps

1. Choose the number of bins.
2. Calculate bin width by dividing the range of your data by the no. of bins.
3. Create Bins - Start with minimum value of data. Then, for each subsequent bin, add the bin width to lower bound of the previous bin.
4. Assign Data points - For each data point, find the bin whose interval range it falls into, & assign the data point to that bin.

Example

In the below example data is partitioned into equidepth bins of depth 3. (1) In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

(2) In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as a bin boundaries. Each bin value is then replaced by the closest boundary value.

The categorization of data into bins follows two primary methods

1) Equal frequency Binning
   Each bin has equal no. of observations.

2) Equal width binning
   width of bin is uniform.

Formula

For equal freq. binning
bin_size = no. of data points / no. of bins.

for equal width binning
$$\text{bin\_width} = \frac{\text{max\_element} - \text{min\_element}}{\text{no. of bins}}$$

Example

Dataset - [5, 10, 11, 13, 15, 35]

Here - total data points = 6
NO. of bins = 3

1) For equal freq. ⇒ bin_size = 6/3 = 2

∴ Bin 1 = [5, 10]
Bin2 = [11, 13]
Bin3 = [15, 35]

2) For equal width
$$\text{bin\_width} = \frac{35-5}{3} = \frac{30}{3} = 10.$$

∴ Bin 1 = 5, 10, 11, 13
Bin 2 = 15
Bin 3 = 35

Conclusion - Binning is useful technique for transforming continuous data into discrete categories, making it easier to analyze & visualize.

Experiment No- 9

Title - Find correlation between items / entities.

Theory

Correlation coefficient are used in statistics to measure how steong a relationship is between two variables.

The occurance of itemset A is independent of the occurance of itemset B if $P(A \cup B) = P(A) \cdot P(B)$

Otherwise, itemsets A and B are dependent and correlated as events.

The correlation between the occurance of A and B can be measured by computing

$$corr_{A,B} = \frac{P(A \cup B)}{P(A) \cdot P(B)}$$

- If resulting value of $corr_{A,B}$ is less than 1, then the occurance of A is negatively co-related with occurance of B

- If resulting value is >1, A & B are tvely correlated

- If resulting value is equal to 1, A & B are independent and there is no co-rrelation.

Algorithm

1) Take input dataset
2) If needed convert data enteies into binary
3) To find co-relation coefficient, take input from user that in which two entities user wish to find co-rrelation.
4) Find '1' count for first entity as well as second entity & count the events at which both the

entities are '1'

5) Apply formula and find correlation coefficient.

Example

Tid.	M	T	W	Th	F	S
1	Y	Y	N	N	Y	N
2	N	Y	Y	N	N	N
3	Y	Y	Y	N	Y	Y
4	N	N	N	Y	Y	Y

for 'y' value

Correlation ratio (1-2) = $\frac{3}{3 \times 3} = \frac{1}{9}$

(1-3) = $\frac{3}{3 \times 5} = \frac{1}{5}$

(1-4) = $\frac{1}{3 \times 3} = \frac{1}{9}$

(2-3) = $\frac{3}{3 \times 5} = \frac{1}{5}$

(2-4) = $\frac{1}{3 \times 3} = \frac{1}{9}$

(3-4) = $\frac{2}{5 \times 3} = \frac{2}{15}$

Conclusion

By using correlation the study of closeness of relationship bet^n different entiies i.e. degree to which variables are associated can be carried out, to find correlation statistical measure, correlation coefficient used which describe the steength & direction of an association bet^n variables.