Unit-4.

Association Rule Mining-

Association Rule mining finds interesting associations and/or correlation relationship among large set of data items.
Association rule show attribute value conditions that occur frequently together in a given dataset.
A typical & widely used example of association rule mining is Market Basket Analysis.

Market Basket Analysis :-

Market Basket Analysis (Association Analysis) is a mathematical modeling technique based upon the theory that if you buy a certain group of items, you are likely to buy another group of items.
The set of items a customer buy is referred ~~to~~ as an itemset, and market basket analysis seeks to find relationship between purchases.
It is used to analyze the customer purchasing behavior and helps in increasing the sales and maintain inventory by focusing on the point of sale transaction data.

→ If we think the universe as the set of items available in the store, then each item has a Boolean variable representing the presence and absence of that item. Each basket can then be represented by a boolean vector of value assigned to this variable. The boolean vectors can be analyzed for buying patterns which recent items that are frequent associated or purchased together. These patterns can be presented in the form of association rules.

Association Rules are of the form 'if X then Y'.

Examples of areas in which association rule have been used include :-

- Credit card transactions :- item purchased by credit card give insight into other products the customer is likely to purchase.
- Supermarket purchases - common combination of products can be used to inform product placement on supermarket shelves.
- Banking services :- the pattern of services used by retail customers are used to identify other services they may wish to purchase.
- Medical patient histories :- certain combinations of conditions can indicate increased risk of various complications.

Consider a market with the collection of huge customer transactions. An association rule is $X \rightarrow Y$, where X is called the antecedent and Y is the consequent.

X and Y are set of items and the rule means that customers who buy X are likely to buy Y with probability % c where c is the confidence.

If(bread) THEN (butter)

This above condition extracts the hiddon information i.e., if a customer used to buy bread, he will also buy butter as side dish.

There are two types of Association rule levels.

1. Support level.

Measurement of items that appear or purchased together.

$$\text{Support } (A \rightarrow B) = P(A \text{ and } B)$$

2. Confidence

Measurement of tendency to buy item after another one.

$$\text{Confidence } (A \rightarrow B) = P(A|B)$$

In terms of support :-

$$\text{confidence } (A \rightarrow B) = P(A|B)$$

$$= \frac{P(A \text{ and } B)}{P(A)}$$

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

Hence,

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

APRIORI Algorithm :-

It is a dominant algorithm by R. Agarwal and R. Srikant in 1994 for mining frequent itemsets.

- The name of the algo is based on the fact that the algorithm uses prior knowledge of fuquent itemset properties.
- It employs an iterative approach known as level-wise search, where k itemsets are used to explore (k+1) itemsets.

The algorithm consist of namely, two steps.

(1.) The Join step :-

To find L_k, a set of candidate k items is generated by joining L_{k-1} with itself. This set of candidates is denoted by C_k.

2. The Prune Step :-

C_k is a superset of L_k, that is, its member may or may not be frequent but all of all the frequent k-itemsets are included in C_k.
A scan of the database to determine the count of each candidate in would result in the determination of L_k (i.e. all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k)

Now let us take an example to understand the algorithm.

Transaction-Id	Item bought
T1	{Mango, Onion, Noodles, Key chains, eggs, books}
T2	{Doll, Onion, Noodles, Key chains, eggs, books}
T3	{Mango, Apple, Key chains, eggs}
T4	{Mango, Umbrella, Corn, Key chains, books}
T5	{Corn, Onion, Key-chain, Ice-cream, eggs}.

where, Support minimum count = 3.

i.e Sup min = 3.

for simplicity

M = Mango,
O = Onion, and so on.

So the table becomes :-

Transaction_id	Item Bought
T1	{M, O, N, K, E, B}
T2	{D, O, N, K, E, B}
T3	{M, A, K, E}
T4	{M, U, C, K, B}
T5	{C, O, K, I, E}

Step-1:
Count the number of transactions in which each item occurs.

$C_1 \rightarrow$

Item	Count
M	3
O	3
N	2
K	5
E	4
B	3
D	1
A	1
U	1
C	2
I	1

$L_1 \rightarrow$

Step 2.
As the support minimum count is 3, so eliminate all the items which have occured less than 3 times. So we are left with

Item	Count
M	3
O	3
K	5
E	4
B	3

$L_1 \rightarrow$

Step-3.

We will make pair of the item.

Item set
{M, O}
{M, K}
{M, E}
{M, B}
{O, K}
{O, E}
{O, B}
{K, E}
{K, B}
{E, B}.

C_3 →

Step -4.

Count the number of times these itemset bought together.

Item set	Count
{M, O}	1
{M, K}	3
{M, E}	2
{M, B}	2
{O, K}	3
{O, E}	3
{O, B}	2
{K, E}	4
{K, B}	3
{E, B}	2

C_4 →

Step-5

As Sup min = 3. So eliminate the element with count less than 3. So we are left with.

Item pairs	count
{M, K}	3
{O, K}	3
{O, E}	3
{K, E}	4
{K, B}	3

L_2 →

Step-6

To make the set of 3 items we need, one more rule (i.e self join).

It simply means, from the item pairs in the above table, we find two pair with the same first alphabet, So. we get

- OK and OE, this gives {O, K, E}
- KE and KB, this gives {K, E, B}

Now the table will be.

Item Set	Count
{O, K, E}	3
{K, E, B}	2

Step-7 :-

As the Sup min = 3, So we will eliminate {K, E, B}, which leaves us with {O, K, E}.

O, K, E

Market - Basket Model

The market-basket model of data is used to describe a common form of many-many relationship between two kinds of objects. On the one hand, we have *items*, and on the other we have *baskets*, sometimes called "*transaction*". Each basket consists of a set items (an itemset), and usually we assume that the number of items in a basket is small - much smaller than the total number of items. The number of baskets is usually assumed to be very large, bigger than what can fit in main memory. The data is assumed to be represented in a file consisting of a sequence of basket.

Frequent Itemsets

Intuitively, a set of items that appear in many baskets is said to be "*frequent*". We assume there is a number s, called the *support threshold*. If I is a set of items, the support for I is the number of baskets for which I is a subset. We say I is frequent if its support is s or more.

Let us take an example of sets of words. Each set is a basket, and the words are items.

1. { cat, and, dog, bites }
2. { Yahoo, news, claims, a, cat, mated, with, a, dog, and, produced, viable, offspring }
3. { cat, killer, likely, is, a, big, dog }
4. { Professional, free, advice, on, dog, training, puppy, training }
5. { cat, and, kitten, training, and, behavior }
6. { dog, &, cat, provides, dog, training, in, Eugene, Oregon }

fuzzy logi-
s..

7. {"Dog, and, cat", is, a, slang, term, used
by, police, officers, for, a, male-female,
relationship}.
8. {shop, for, your, show, dog, grooming, and,
pet, supplies}

Here are eight baskets, each consisting of items that are words.

Since the empty set is a subset of any set, the support for ϕ is 8. ~~However~~

Among the singleton sets, obviously {cat} and {dog} are quite frequent. "Dog" appear in all but bucket (5), so its support is 7, while cat appears in all but (4) and (8), so its support is 6. The word "and" is also quite frequent; it appears in (1), (2), (5), (7) and (8), so its support is 5. The words "a" and "training" appear in three sets, while "for" and "is" appear in two each. No other word appears more than once.

Suppose we set our threshold at $s = 3$. Then there are five frequent singleton it omsets :-
{dog}, {cat}, {and}, {a}, and, {training}.

Now let us look at the doubletons.
A doubleton cannot be frequent unless both items in the set are frequent by themselves.
Thus, there are only ten possible frequent doubletons.
The table below is ~~intating~~ indicating which baskets contain which doubletons:

	training	a	and	cat
dog	4, 6	2, 3, 7	1, 2, 8	1, 2, 3, 6, 7
cat	5, 6	2, 7	1, 2, 5	
and	5	2, 3		
a	none			

we can see that doubleton {dog, training} appears only in basket (4) and (6). Therefor, its support is 2 and it is not frequent.

There are four frequent doubletons if $s = 3$;

they are

1. {dog, a}
2. {dog, and}
3. {cat, and}
4. {dog, cat}

Each appear exactly 3 times, except for {dog, cat} which appears five times.

Now, let us see if there are frequent triples. In order to be a frequent triple, each pair of elements in the set must be a frequent doubleton.

For example, {dog, a, and} cannot be a frequent itemset, because if it were, then surely {a, and} would be frequent, but it is not.

The triple "{dog, cat, and} might be frequent, because each of its doubleton setbset is frequent.

Unfortunately, the three words appear together only in basket (1) and (2), so there are in fact no frequent triplets.

If there are no frequent triples, then there surely are no frequent quadruples or larger sets.

Applications of frequent Itemsets :-

The original application of the market-basket model was in the analysis of true market basket. That is, supermarkets and chain stores record the contents of every market basket (physical shopping cart) brought to the register for checkouts. Here the "items" are the different products that the store sells, and the "baskets" are the sets of items in a single market basket. A major chain might sell 10,000 different items and collect data about million of market baskets.

By finding frequent itemsets, a retailer can learn what is commonly bought together. Especially important are pairs or larger sets of items that occur much more frequently than would be expected were the items bought independently.

Supervised learning vs unsupervised learning

- **Supervised learning :-**

Discover patterns in the data that relate data attributes with a target (class) attribute

- These patterns are then utilized to predict the values of the target attribute in future data instances.

- **Unsupervised learning :-**

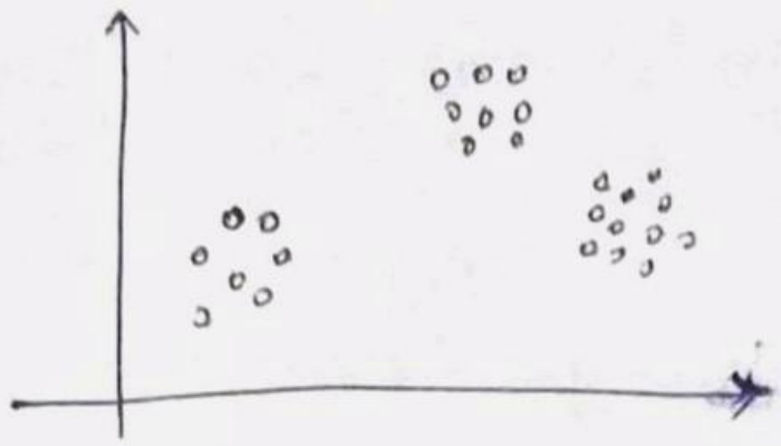
The data have no target attribute.

- we want to explore the data to find some intrinsic structures in them.

Clustering :-

- Clustering is a technique for finding similarity groups in data, called clusters, i.e.,
  - it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.
- clustering is often called unsupervised learning task as no class values denoting a priori grouping of the data instances are given, which is the case in supervised learning.
- Due to historical reasons, clustering is often considered synonyms with unsupervised learning.
  - in fact, association rule mining is also unsupervised.

- The data set has 3 natural groups of data points i.e., 3 natural clusters.



Clustering Quality →

→ Inter-clusters distance → maximized.
→ Intra-clusters distance → minimized.

The quality of a clustering results depend on the algorithm, the distance function, and the application.

K-means clustering

K-means is a partitioning clustering algorithm.

- Let the set of data points (or instances) D be $\{x_1, x_2, \ldots, x_n\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $X \subseteq R^r$, and r is the number of attributes (dimensions) in the data.
- The K-mean algorithm partitions the given data into K-clusters
  - Each cluster has a cluster center called centroids.
  - K is specified by the user.

K-mean Algorithm.

- Given K, the K-means algorithm works as follows:-

(1.) Randomly choose K data points to be the initial centroids, cluster centers.

(2.) Assign each data point to the closest centroid.

(3) Re-compute the centroids using the current cluster memberships.

(4) If a convergence criterion is not met, go to (2).

Convergence Criterion

1. no (or minimum) re-assignments of data points to different clusters.
2. no (or minimum) change of centroids, or
3. minimum decrease in the sum of squared error (SSE),

$$SSE - \sum_{j=1}^{K} \sum_{x \in C_j} dis(x, m_j)^2$$

- C_j is the jth cluster, m_j is the centroid of cluster C_j (the mean vector of all the data points in C_j), and $dis(x, m_j)$ is the distance between data point x and centroid m_j.

Let us see with an example :-

Data set K = {2, 3, 4, 10, 11, 12, 20, 25, 30}

for K = 2 i.e. 2 clusters / centroids.

Randomly select two centroids

e.g - 4 and 12.

So

$M_1 = 4$ $M_2 = 12$.

Now calculate the distance of all data set points, whichever centroid has less distance from these point will lie in that cluster.

e.g - 2 → |4 - 2| = 2 — ✓ closer to m,
|12 - 2| = 10

3 → |4 - 3| = 1 — ✓ closer to m,
|12 - 3| = 9

4 → |4 - 4| = 0 — ✓ closer to m,
|12 - 4| = 8

10 → |10 - 4| = 6
|12 - 10| = 2 — ✓ closer to m_2

and so on . .

So the final cluster for m_1 and m_2 are :-

$m_1 = 4$ $m_2 = 12$

$K_1 = \{2, 3, 4\}$ $K_2 = \{10, 11, 12, 20, 25, 30\}$

Now compute the next mean.

$M_1 = \frac{2+3+4}{3} = 3$ $M_2 = \frac{10+11+12+20+25+30}{12} = 18$

.

$M_1 = 3$ $M_2 = 18$

Now repeat the same process for these means.

$M1 = 3$ $\qquad$ $M2 = 18$

$K1 = \{2, 3, 4, 10\}$ $\qquad$ $K2 = \{11, 12, 20, 25, 30\}$.

compute the mean again.

$M1 = \frac{2+3+4+10}{4} = 4.75 \simeq 5$ $\qquad$ $M2 = \frac{11+12+20+25+30}{5} = 19.6 \simeq 20$.

$M1 = 5$ $\qquad$ $M2 = 20$

$K1 = \{2, 3, 4, 10, 11, 12\}$ $\qquad$ $K2 = \{20, 25, 30\}$.

Compute the mean again.

$M1 = \frac{2+3+4+10+11+12}{6} = 7$ $\qquad$ $M2 = \frac{20+25+30}{3} = 25$.

$M1 = 7$ $\qquad$ $M2 = 25$

$K1 = \{2, 3, 4, 10, 11, 12\}$ $\qquad$ $K2 = \{20, 25, 30\}$

Compute the mean again.

$M1 = \frac{2+3+4+10+11+12}{6}$ $\qquad$ $M2 = \frac{20+25+30}{3}$

$M1 = 7$ $\qquad$ $M2 = 25$

There is no change in the mean, hence the clusters for the final means $m_1 = 7$ and $M_2 = 25$ will be

$K1 = \{2, 3, 4, 10, 11, 12\}$

$K2 = \{20, 25, 30\}$.

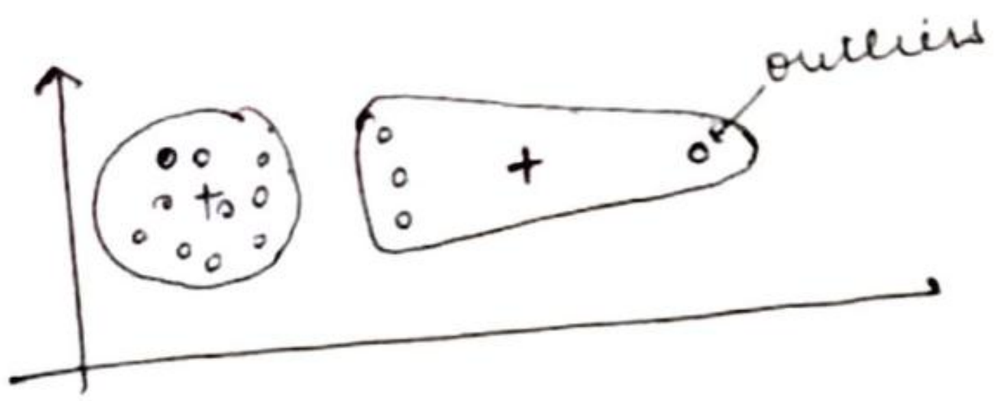
K-means Clustering

Strengths :-

1. Simple - easy to understand & to implement
2. Efficient - Time complexity : $O(tKn)$,
   where n is the number of data points,
   K is number of clusters, and
   t is the number of iterations.
3. Since both K and t are small. K-means is considered a linear algorithm.

weakness of K-mean

- The algorithm is only applicable if the mean is defined.
- The user need to specify K.
- The algorithm is sensitive to outliers
  - outliers are data points that are far away from other data points.
  - Outliers could be errors in the data recording or some special data points with very different values.



Undesirable Cluster

Ideal Cluster

Major Clustering Approaches / Types.

1. Grid-based approch.
   - based on multiple-level granularity structure
   - Typical methods - STING, CLIQUE.

2. Partitioning approach.
   - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors.
   - Typical methods - K-mean, K-medoids, CLARANS.

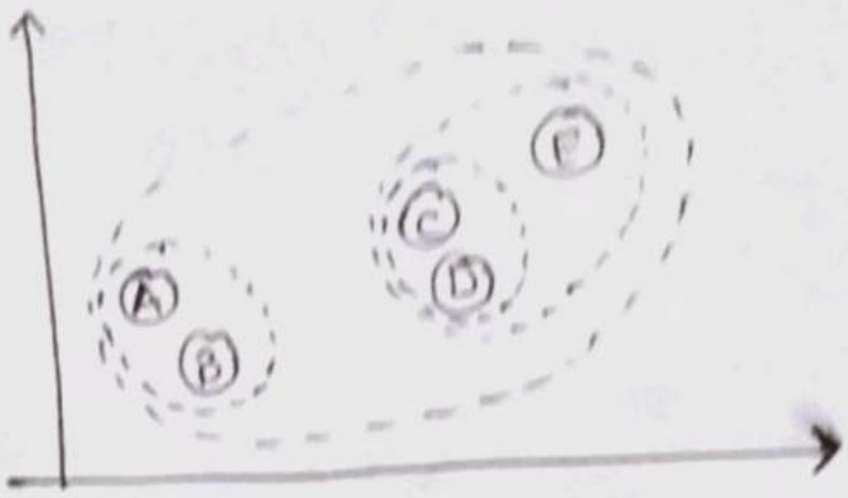
3. Hierarchical Approach.
   - Create a hierarchical decomposition of the set of data (or objects) using some criterion.
   - Typical methods: Diana, Agnes, BIRCH, ROCK,

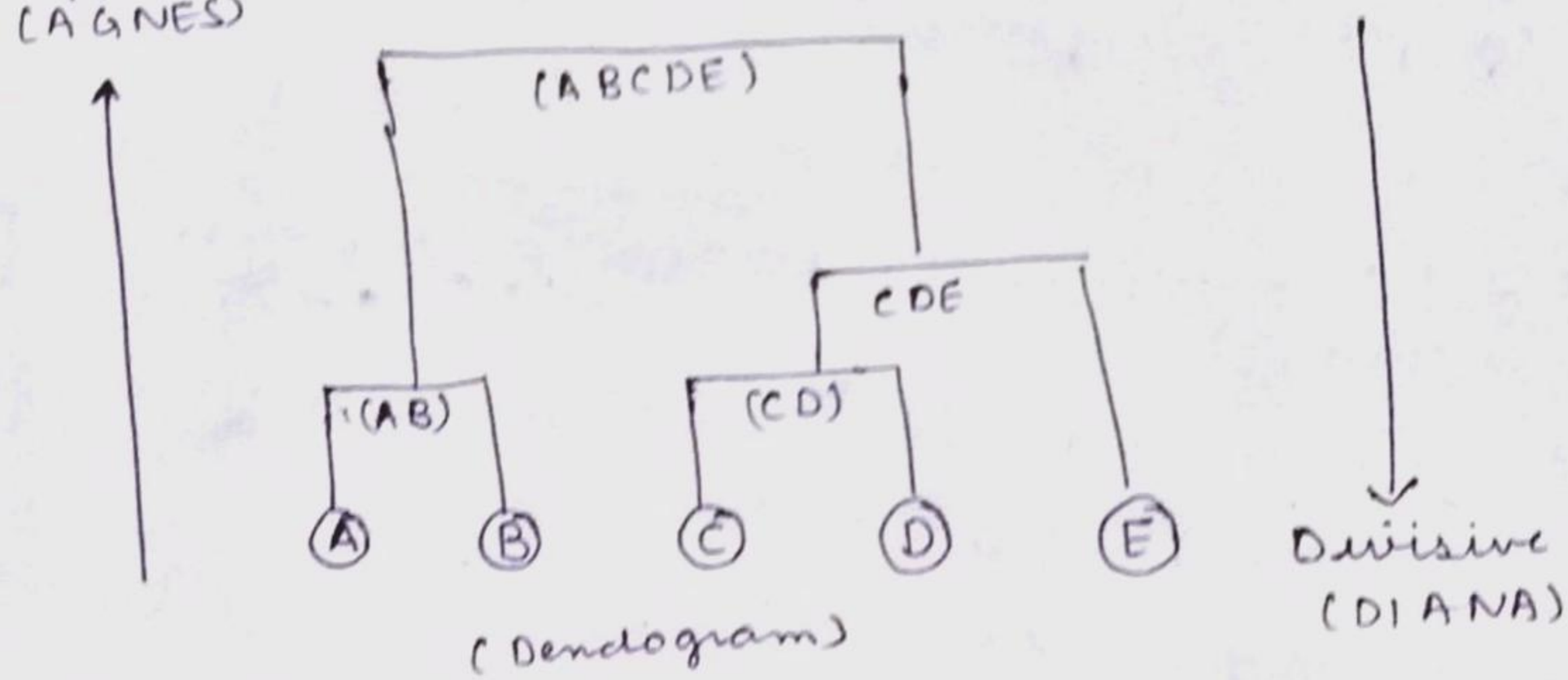
4. Density-based Approach.
   - Based on connectivity and density function.
   - Typical methods : DBSCAN, OPTICS.

5. Frequent pattern-based :-
   - Based on the analysis of frequent patterns
   - Typical methods - pCluster.

Hierarchical Clustering

Use distance matrix as clustering criteria. This method does not require the number of clusters k as an input, but needs a termination condition.





(1) AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- → Implemented in statistical analysis packages, e.g., Splus.
- → Use the single-link method and the dissimilarity matrix.
- → It builds the dendogram (from) the bottom level, i.e. bottom-up approach.

→ Merge nodes that have the least dissimi-
- larity.
→ Go on in a non-descending fashion.
→ Eventually all nodes belong to the same cluster.

Dendogram.

Shows how the clusters are merged.

→ Decompose data objects into several levels of nested partitioning tree of clusters, called a dendogram.

→ A clustering of the data objects is obtained by cutting the dendogram at the desired level, then each connected component forms a cluster.

(2) DIANA (Divisive Analysis)

→ Introduced in Kaufmann and Rousseeuw (1990)

→ Implemented in statistical analysis packages, e.g., Splus

→ Inverse order of AGNES

→ Eventually each node forms a cluster on its own.

→ It is a top-down approach.

Clustering High-Dimensional Data

As dimensionality increases
- number of irrelevant dimensions may produce noise and mask real clusters.
- data become sparse
- Distance measure - meaningless.

Feature transformation methods.

- PCA, SVD. Summarize data by creating linear combinations of attributes.
- But do not remove any attributes, transformed attributes - complex to interpret.

Feature Selection method.

- Most relevant set of attribute with respect to class labels.
- Entropy Analysis.
- Subspace clustering - searches for groups of clusters within different subspace of the same data set.

Clustering High Dimensional Data

There are many approaches for high dimensional data clustering, namely.

1. Subspace Clustering

- These clustering algorithm localize the search for relevant dimensions allowing them to find clusters that exist in multiple, and possibly overlapping subspaces.
- This technique is an extension of feature selection that attempts to find clusters in different subspaces of the same dataset
- Subspace clustering requires a search method and a evaluation criteria.
- It limits the scope of the evaluation criteria so as to consider different subspaces for each different cluster

2. Projected clustering

- In H-D spaces, even though a good partition cannot be defined on all the dimensions because of the sparsity of the data, some subset of the dimensions can always be obtained on which some subsets of data from high quality and significant cluster.
- Projected clustering methods aims to find cluster specific to a particular group of dimensions. Each cluster may refer to different subsets of dimensions.
- The output of a typical projected clustering algorithm, searching for k-clusers in subspaces of dimension l, is two fold;

(i) A partition of data of K+1 different clusters, where the first K clusters are well shaped, while the (K+1)th cluster elements are outliers, which by definition donot cluster well.

(ii) A possible different set of 1D for each of the fiist K clusters, such that the points in eacs of those clusters are well clustered in the subspaces defined by these vectors.

3. Bi clustering

- Biclustering (or two-way clustering) is a methodo-logy allowing for features set and data points clustering simuntaneously, i.e., to find clusters of samples possessing similar characterestics together with features creating these similarities.

- The o/p of biclustering is not a partition or hierarchy of partitions of either rows or columns, but a partition of the whole matrix into sub-matrices or patches.

- The goal of biclustering is to find many patches as possible, and to have as large as possible, while maintaining strong homogeneity within pateches.

CLIQUE - Clustering In Quest

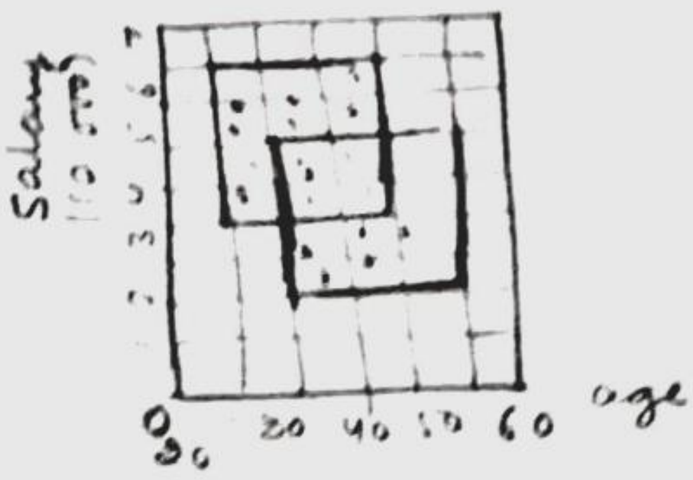
- Agarwal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- Automatically identifying subspaces of a H-D data space that allow better clustering than original space.
- Clique can be considered as both density-based and grid-based.
  - It partition each dimension into a small number of equal length interval.
  - It partitions an m-dimensional data space into non-overlapping rectangular units.
  - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameters.
  - A cluster is a maximal set of connected dense units within a subspace.

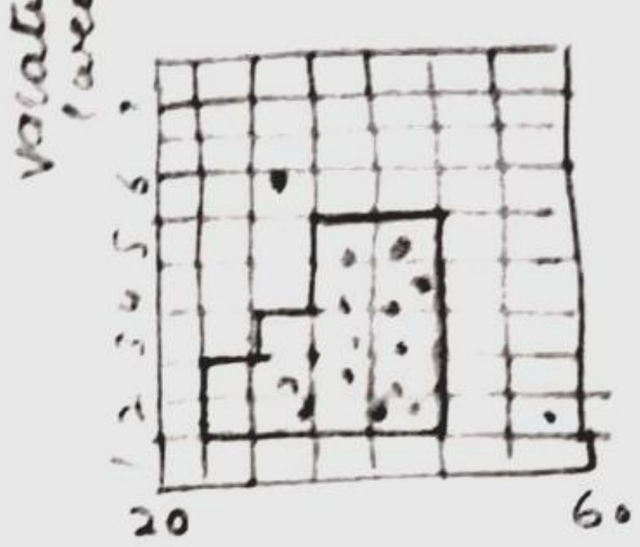
Clique :- Major Steps

1. Partition the data space and find the number of points that lie inside each cell of the partition.
2. Identify the subspaces that contain clusters using the Apriori Principle.
3. Identify clusters
   - Determine dense units in all subspaces of interests.
   - Determine connected dense units in all subspace of interests.

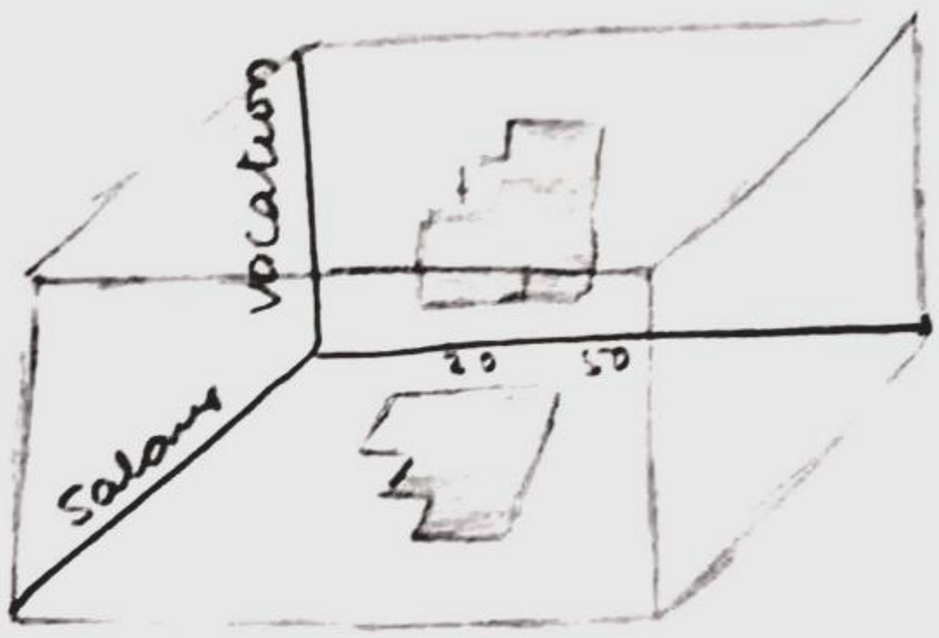
4 Generate minimal description for the clusters.

- Determine maximal region that cover a cluster of connected dense units for each cluster
- Determination of minimal cover for each cluster.





$\tau = 3$



Strengths of Clique

- automatically find subspaces of the highest dimensionality such that the high density clusters exist in those subspaces.
- insensitive to the orders of records in input & does not presume some canonical data distribution.

- Scales linearly with the size of i/p and has good scalability as the number of dimensions in the data increases.

Weakness of Clique

- The accuracy of the clustering results may be degraded at the expense of simplicity of the method.

PROCLUS - PRojected CLUStering

- Dimension reduction subspace clustering technique.
- Finds initial approximation of clusters in high dimensional space
- Avoid generation of large number of overlapped clusters of low dimensionality.
- Finds best set of medoids by hill-climbing process.
- Manhattan segmental distance measure.

Algorithm

1. Initialization Phase.
   - Greedy algorithm to select a set of initial medoids that are far apart.
2. Iteration Phase.
   - Select a random set of K-medoids.
   - Replace bad medoids
   - For each medoid a set of dimensions is chosen whose average distance are small.

3. Refinement Phase.

- Compute new dimensions for each medoid based on clusters found. reasigns points to medoids and remove outliers.

Frequent Pattern based clustering

- Frequent pattern may also form clusters
- Instead of growing clusters dimension by dimension sets of frequent itemset are determined.
- Two common techniques
  (1) Frequent term-based text clustering
  (2) Clustering by Pattern Similarity.

(1.) Frequent term-based text clustering

- Text documents are clusters based on frequent terms they contain.
- Documents - terms
- Dimensionality is very high
- Frequent terms based analysis.
  - Well selected subset of set of all frequent terms must be discovered.
  - Fi - set of frequent term sets.
  - Cov (F_i) - set of documents covered by F_i
  - $\sum i = 1^{k} cov(F_i) = D$ and overlap beltween F_i and F_j must be minimized.
  - Description of clusters - their frequent term sets

(2) Clustering by Pattern Similarity

- P-cluster on micro array data analysis.
- DNA micro-array analysis - expression levels of two genes may rise and fall synchronously in response to stimuli
- Two objects are similar if they exhibit a coherent pattern on a subset of dimensions

P-cluster :-

- Shift pattern discovery.
  - → Euclidean distance - not suitable,
  - → Derive new attributes.
  - - Bi-clustering based on mean squared residual score.
- P cluster
  - → object - x, y ; attribute - a, b.

$$pScore\left(\begin{bmatrix} d_{xa} & d_{xb} \\ d_{ya} & d_{yb} \end{bmatrix}\right) = |(d_{xa} - d_{xb}) - (d_{ya} - d_{yb})|,$$

- A pair (O, T) forms a δ-p cluster if for any 2×2 matrix X in (O, T), p Score(X) <= δ.
- Each pair of objects and their features must satisfy threshold.
- Scaling Patterns

$$\frac{d_{xa}/d_{ya}}{d_{xb}/d_{yb}} \leq \delta'.$$