Deepak Kumar P Honakeri
CSE – 6A

1PE17CS044

DM assignment – 3

01) What is clustering? How is it different from classification?

Ans Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar to one another and different from the objects in other group.

- Clustering analysis is related to other techniques that are used to divide data objects into groups. For instance, clustering can be regarded as a form of classification in that it creates a labelling of objects with class (cluster) labels.

However, it derives these labels only from the data In contrast, classification, new unlabelled objects are assigned a class label using a model developed from objects with known class labels. Hence, cluster analysis is sometimes referred to as unsupervised classification

1PE17CS44

When the term classification is used without any qualification within datamining, it typically refers to supervised classification

Q3 Discuss the different types of clusters.

Ans Types of clusters:

1) well-separated
2) Prototype-Based
3) Graph-Based.
4) Density-Based
5) Shared-Property.

1) well separated: A cluster is a set of objects in which each object is closer to every other object in the cluster than to any object not in the cluster. Sometimes a threshold is used to specify that all objects in a cluster must be sufficiently close to another.

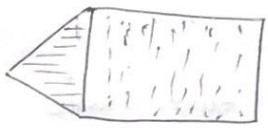
2) Prototype-Based: A cluster is a set of objects in which each object is closer to the prototype that defines the cluster than to the prototype of any cluster. For data with continuous attributes the prototype of a cluster is a centroid i.e., mean of all points in cluster

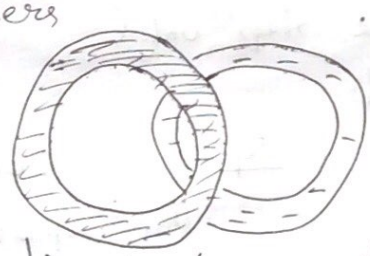
3) Graph based: If the data is represented as a graph, where the nodes are objects & the links connections among objects, then a cluster can be defined as a connected component.

Eg: Contiguity-based clusters; where two objects are connected only if they within a specified distance of each other.

4) Density Based: A cluster is a dense region of objects that is surrounded by a region of low density... This type is employed when the clusters are irregulas or intertwined & when noise and outliers are present.

5) Shared-Property: The shared property approach includes new type ~~of~~ clusters





A triangulas area is adjacent to rectangular area & the two intertwined circles, In both cases, clustering algorithm would need a very specific concept of cluster to successfully detect these clusters this is called Conceptual Clustering.

Q4) Explain K-means algorithm. what are its limitation? [IPE17CS044]

Ans Basic K-means Algorithm:

1. Select K points
2. repeat

   Form K clusters by assigning each point to its closest centroid.

   Recompute the centroid of each cluster

   until centroids do not change

- we choosese K initiall centroids, (K is user specified) namely ~~each~~, no. of clusters to be used/ desired.
- Each point is then assigned to the closest centroid & collection of points assigned to the ~~cluster~~ ~~is~~ & centroid is a cluster.
- Repeat the step until no points assigned to cluster is changed:

  Step: Centroid of each cluster is then updated based on the points assigned to the cluster.

Limitations:

- ~~K-means, although are more efficient, they are less susceptible to initialization problems.~~
- K-means is not suitable for all types of data
- It cannot handle non-globular clusters or cluster of different sizes & densities

- K-means also has trouble clustering data that contains outliers.
- K-means is restricted to data for which there is a notion of a center.

05 How are density based methods for clustering. How are parameters of the DBSCAN algo selected?

Ans) Density based clustering locates regions of high density that are separated from one another by regions of low density. DBSCAN is a simple and effective density-based algorithm that illustrates a no. of important concepts that are important for any density-based clustering approach.

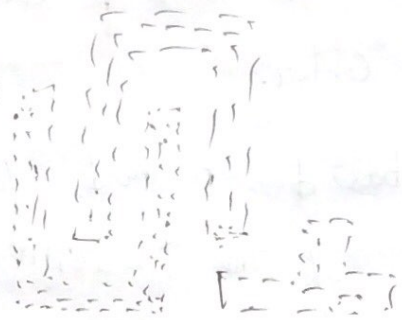
Selection of DBSCAN params:

There is, of course, the issue of how to determine the parameters EPs and MinPts. The basic approach is to look at the behaviour of the distance from a point to its k^{th} nearest neighbor. (k-dist).

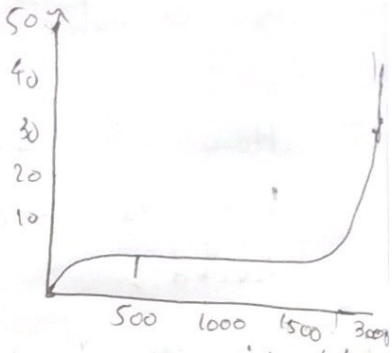
Therefore, if we compute the k-dist for all the data points for some k, sort in ascending order, then plot the sorted values, we expect to see

a sharp change at the value of k-dist that corresponds to a suitable value of Eps. [IPE17CS44]

Example :



Sample data



k-dist plot for Sample data

Q2 list and explain the desired features of Cluster analysis

Ans 1) Clustering for understanding
2) Clustering for Utility

1) Clustering for understanding: Classes or conceptually meaningful groups of objects that share common characteristics, play an important role in how people analyse & describe the world. In the context of understanding data, clusters are potential classes & cluster analysis is the study of techniques for finding classes automatically.

Eg: Biology, information retrieval, Climate, Business etc

IPE17CS44

Clustering for Utility:

Cluster analysis provides an abstraction from individual data objects to the clusters in which those data objects reside.

∴ In context of utility, Clustering analysis is the study of techniques for finding the most representative cluster prototypes.

Eg:- Summarization, Compression, Efficiently finding Nearest neighbors.

06 Explain different methods of computing distances between clusters

Ans: 1) The different methods of computing distances are

1* Single link or MIN

The proximity of 2 clusters is defined as the minimum of the distance between any two points in the two different clusters.

It is given by:

$$\text{Min}\{ d(a,b),\ a \in A,\ b \in B \}$$

a, b – points in clusters A, B

A, B – clusters A, B

2) Complete link or MAX or CLIQUE

The proximity of two clusters is defined as the maximum of the distance between any two points in the two different clusters.

Given by: $\boxed{\text{Max } \{d(a,b),\ a \in A,\ b \in B\}}$

3) Group Average:

For the group average, the proximity of two clusters is defined as the average pairwise proximity among all pairs of points in the different clusters.

Thus proximity $(C_i, C_j) = \dfrac{\sum_{\substack{x \in C_i \\ y \in C_j}} \text{proximity}(x, y)}{m_i * m_j}$

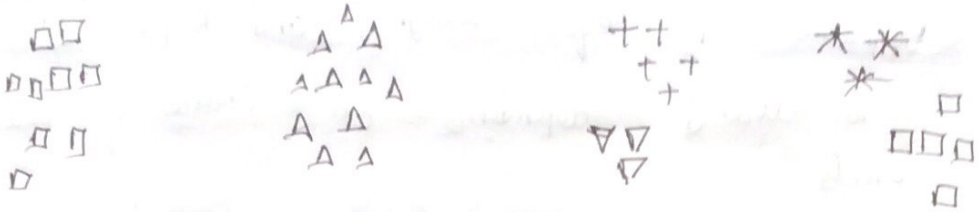
C_i, C_j — clusters of size m_i & m_j

Q7 Describe the following approaches to clustering with an example in each case

Sol a) Partitioning Method:

A partitional clustering is simply a division of the set of the data objects into non-overlapping subsets (clusters) such that each data object is exactly one subset.

Eg:



Two clusters

Four clusters

b) Hierarchial Clustering

IPE17CS044

If we permit clusters to have subclusters, then obtain a hierarchical clustering, which is a set of nested clusters that are organised as a tree. Each node (Cluster) in the tree is the the union of its children and the root of the tree is the cluster containing all the objects.

Eg

8) Briefly outline how to compute dissimalarity between objects described by

Sol) a) Interval Scaled Variable

These variables are continious measurements of a roughly linear scale. Eg: height, latitude, meters etc.

- Standardising measurements attempts to give all

variables an equal weight, and helps to avoid dependance on choice of measurement units

- Standardize using mean absolute deviation:

- Distances are normally used to measure the similarity or dissimilarity between two data objects

1) Euclidean distance, $d(i,j) = \sqrt{|x_{i_1}-x_{j_1}|^2 + |x_{i_2}-x_{j_2}| + \cdots}$

2) Manhattan distance, $d(i,j) = |x_{i_1}-x_{j_1}| + |x_{i_2}-x_{j_2}| + \cdots$

3) Minkowski distance, $d(i,j) = \sqrt[q]{|x_{i_1}-x_{j_1}|^q + |x_{i_2}-x_{j_2}|^q + \cdots}$

b) Binary Variables: These variables has only two states 0/1, 0 – variable is absent, 1 – variable is present.

One approach involves computing dissimilarity matrix or if all binary variables have same weight, we have Contingency table.

Symmetric binary dissimilarity; $d(i,j) = \frac{r+s}{q+r+s+t}$

Assymmetric binary dissimilarity; $d(i,j) = \frac{r+s}{q+r+s}$

Q4 Illustrate .. grid-based clustering algorithm. How clusters are formed from dense-grid cells. [1PE17CS044]

Ans Grid-based clustering algorithm:

1. Define a set of grid cells
2. Assign objects to the appropriate cells and compute the density of each cell.
3. Eliminate cells having a density below a specified threshold, T.
4. Form clusters from contiguous groups of dense cells.

In many cases, the data has both spatial & non-spatial attributes., In other words, some of the attributes describe the location of objects in the time or space, while other attributes describe other aspects of the objects. A common example is houses, that have both a location & characteristics such as price, space etc. Because of spatial ~~attributes~~ auto correlation, objects in a particular cell often have similar values for their other attributes. In such cases, it is possible to filter

the cells based on statistical properties of one or more non-spatial attributes, and then form clusters based on the density of the remaining points.

010) Develop DENCLUE algorithm for kernel density function.

Ans.

DENCLUE algorithm:

1. Derive a density function for the space occupied by the data points.
2. Identify the points that are local maxima.
3. Associate each point with a density attractor by moving in the direction of maximum increase in density.
4. Define clusters consisting of points associated with a particular density attractor.
5. Discard clusters whose density attractor has a density less than a user-specified threshold of ξ.
6. Combine clusters that are connected by a path of points that all have density of ξ or higher