Clustering Methods

Clustering Methods

- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Constraint-Based Methods
- Outlier Analysis

Density Based Methods

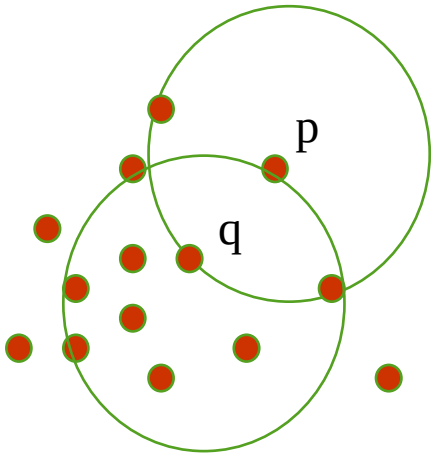
- Based on the notion of density
- Regard clusters as dense regions of objects in the data space that are separated by regions of low density (representing noise)
- Discover clusters with arbitrary shape
- Several interesting algorithms:
  - DBSCAN (Density Based Spatial Clustering of Applications with Noise): Ester, et al.
  - OPTICS (Ordering Points To Identify the Clustering Structure): Ankerst, et al .
  - DENCLUE (DENsity-based CLUstEring): Hinneburg & D. Keim .

Density-Based Clustering: Background

- Two important parameters:
  - ***ϵ - neighbourhood*** :Objects within a radius ϵ from an object.

$$N_{\epsilon}(p) : \{ (q \mid d (p, q) <= \epsilon \}$$

  - ***MinPts***: Minimum number of points in an ϵ -neighbourhood of that point (object).



MinPts = 5

ϵ = 1 cm

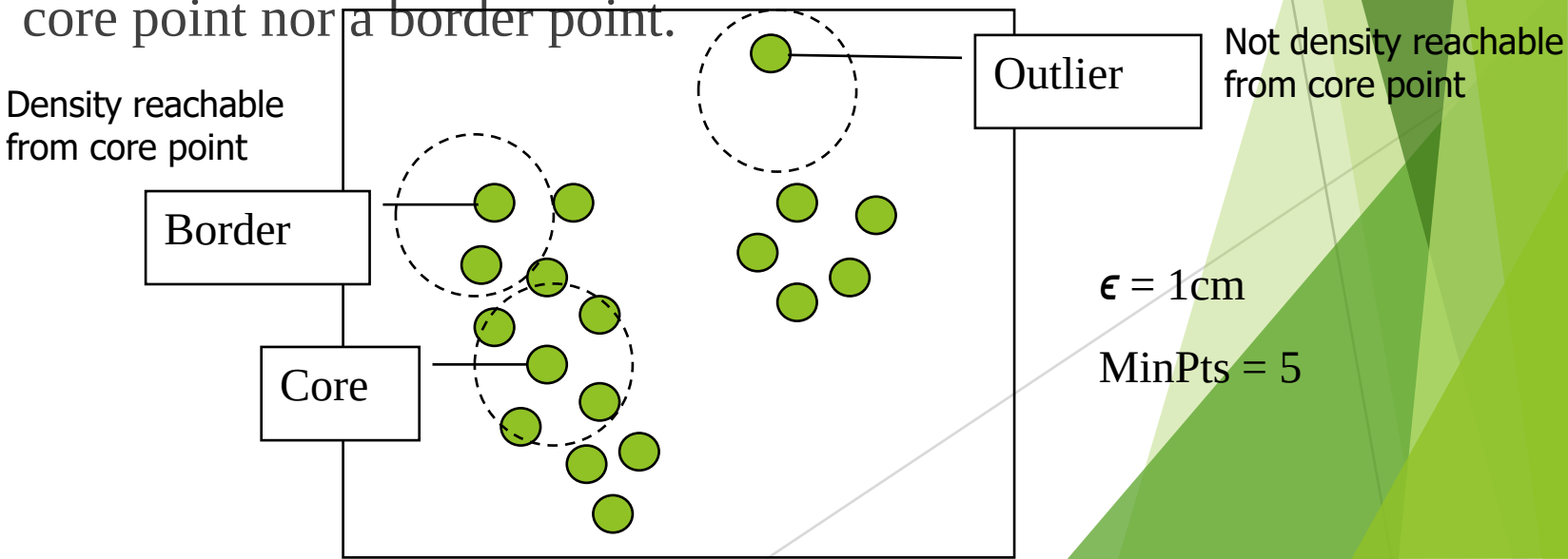
Density of q is "high" (MinPts = 5)

Density of p is "low" (MinPts = 5)

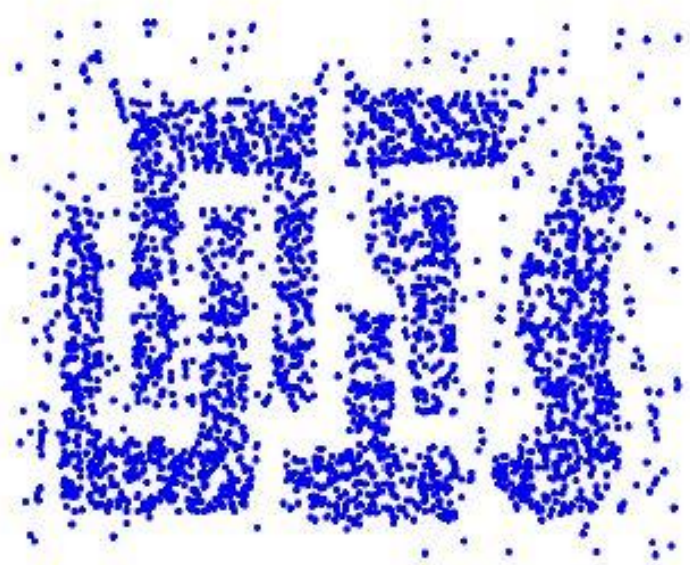
Core, Border & Outlier

Given ϵ and *MinPts*, categorize the objects into three exclusive groups :

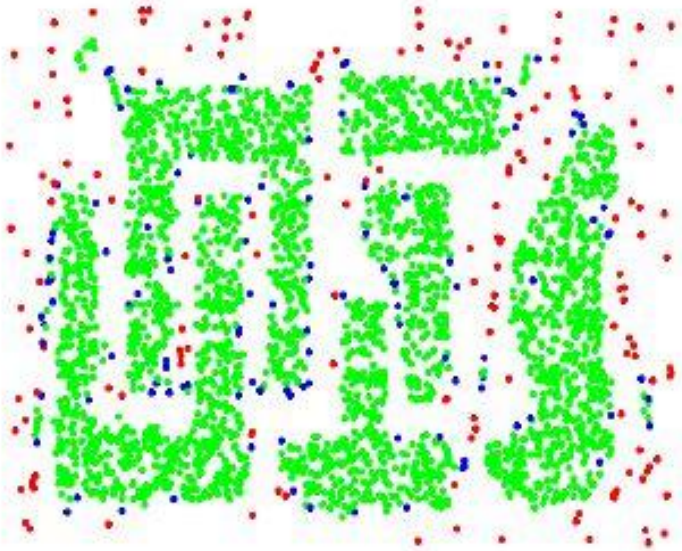
- **Core** : A point is a core point if it has more than a specified number of points (MinPts) within ϵ - neighborhood. These are points that are at the interior of a cluster.
- **Border** : A border point has fewer than MinPts within ϵ, but is in the neighbourhood of a core point.
- **Outlier** : A noise point (outlier) is any point that is neither a core point nor a border point.



Example



Original Points

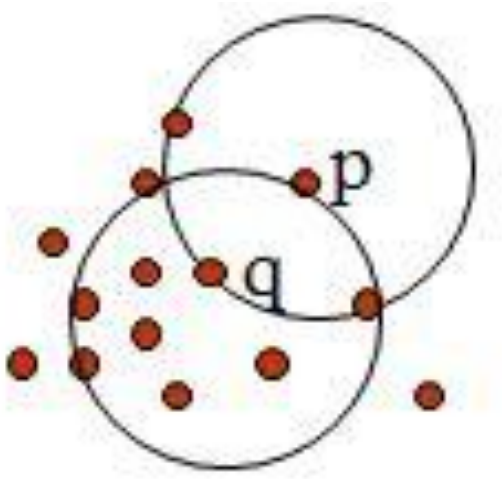


Point types: core, border and outliers

ε = 10, MinPts = 4

Concepts: Reachability

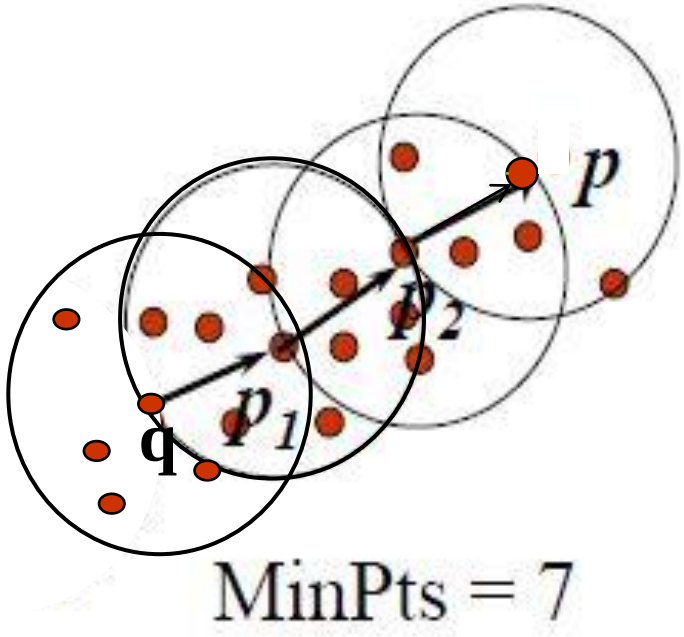
- **Directly density-reachable**
  - An object p is directly density-reachable from object q if p is within the ε-Neighborhood of q and q is a core object.



- p is directly density reachable from q
- q is not directly density reachable from p?

Concepts: Reachability

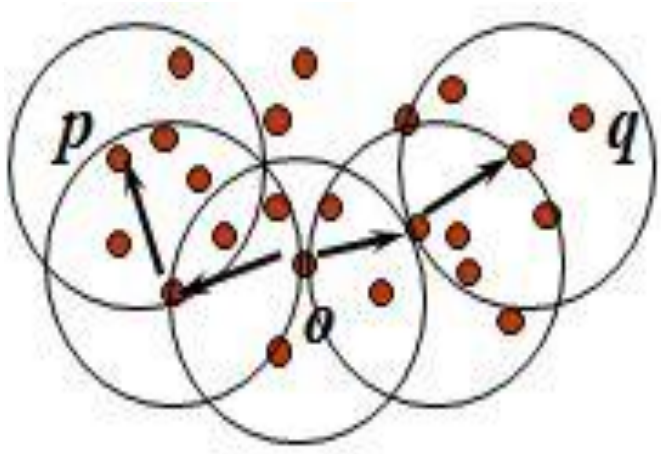
- **Density-reachable:**
  - An object p is density-reachable from q w.r.t ε and ***MinPts*** if there is a chain of objects $p_1,\ldots,p_n$, with $p_1=q$, $p_n=p$ such that p_{i+1} is directly density-reachable from p_i for all $1 <= i <= n$



- p is density-reachable from q
- q is not density- reachable from p?
- Transitive closure of direct density-reachability, asymmetric

Concepts: Connectivity

- **Density-connectivity**
  - Object *p* is density-connected to object *q* w.r.t ε and *MinPts* if there is an object *o* such that both *p* and *q* are density-reachable from *o* w.r.t ε and *MinPts*



- *P* and *q* are density-connected to each other by *o*
- Density - connectivity is a symmetric relation

DBSCAN: Density Based Spatial Clustering of Applications with Noise

- A density-based *cluster* is a set of **density-connected** objects that is **maximal** w.r.t density-reachability:

- A cluster ***C*** in a set of objects ***D*** w.r.t ε and *MinPts* is a non empty subset of ***D*** satisfying
  - **Maximality**: For all *p*, *q* if $p \in C$ and if *q* is density-reachable from *p* w.r.t ε and *MinPts*, then also $q \in C$.
  - **Connectivity**: For all $p, q \in C$, *p* is density-connected to *q* w.r.t ε and *MinPts* in ***D*.**

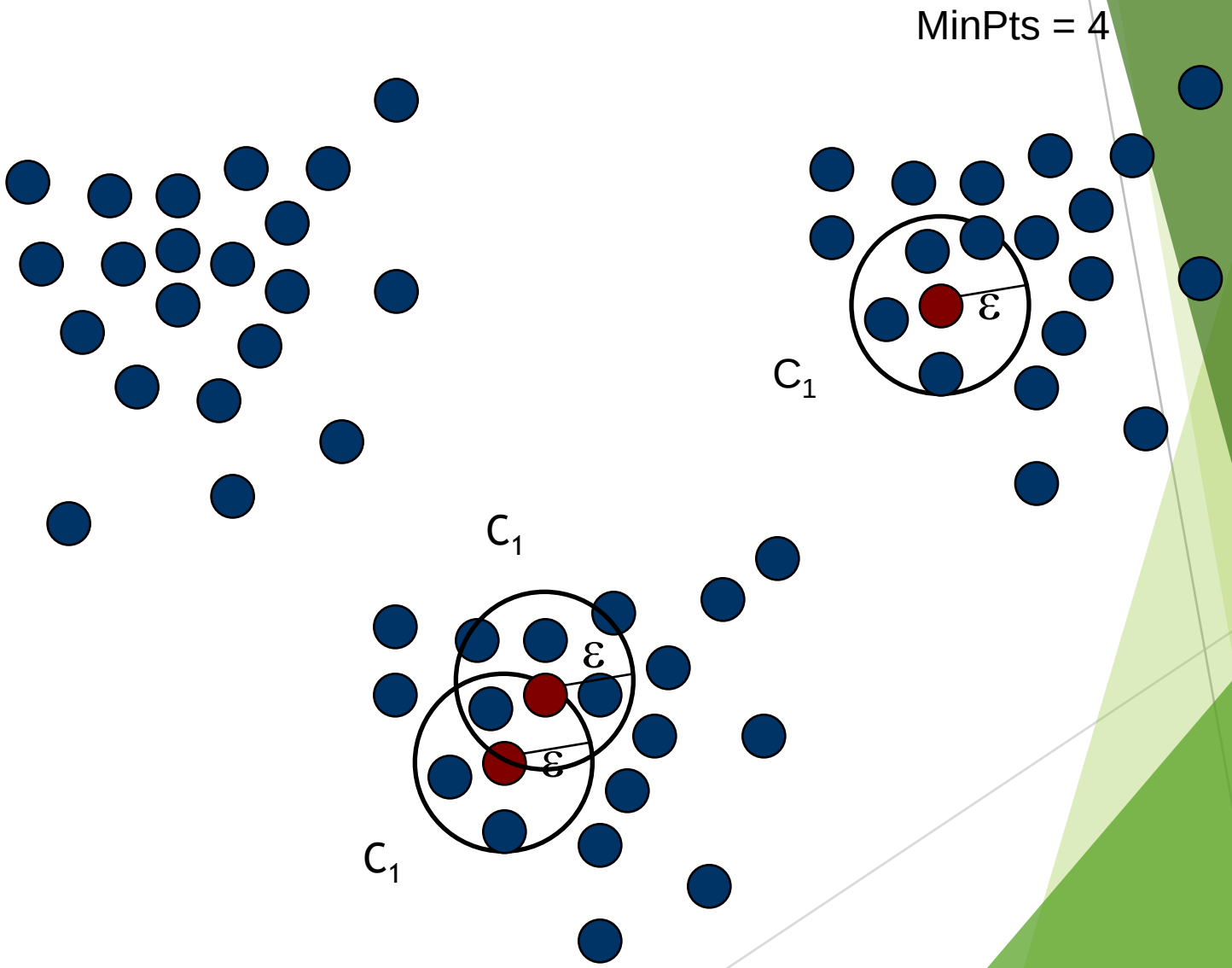
- **Noise:** Objects which are not directly density-reachable from at least one core object.

DBSCAN: The Algorithm

- Arbitrarily select a point ***p***
- Retrieve all points density-reachable from ***p*** wrt ε and ***MinPts***.
- If ***p*** is a core point, a cluster is formed.
- If ***p*** is not a core point i.e., no points are density-reachable from ***p***, then DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

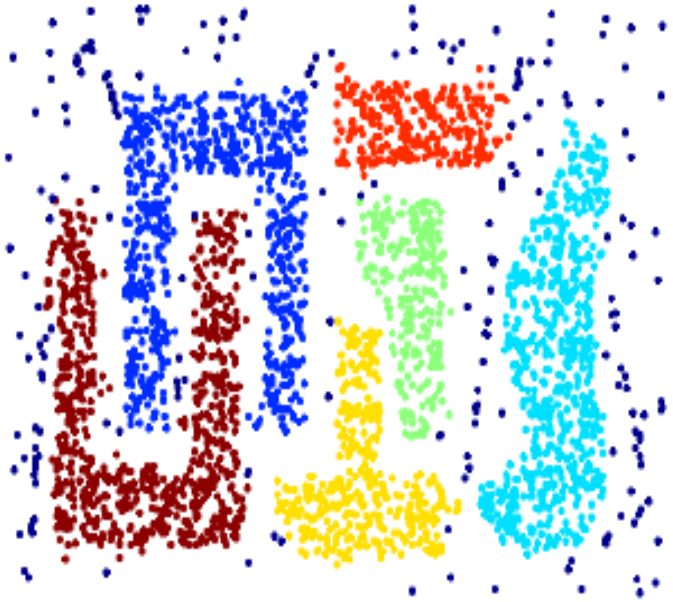
for each $o \in D$ **do**
  if o is not yet classified **then**
    if o is a core-object **then**
      collect all objects density-reachable from o and assign them to a new cluster.
    else
      assign o to NOISE

An Example



When DBSCAN Works Well

Original Points



Clusters

Grid-Based Methods

- Uses multi-resolution grid data structure
- The object space is quantized into finite number of cells that form a grid structure on which all of the operations for clustering are performed.
- Clustering complexity depends on the number of populated grid cells and not on the number of objects in the dataset

- Several interesting methods:
  - STING (a STatistical INformation Grid approach) : Wang, Yang and Muntz (1997)
  - CLIQUE (Clustering In Quest): Agrawal, et al.
  - WaveCluster : Sheikholeslami, Chatterjee, and Zhang

Steps of Grid-based Clustering Algorithms

Suppose we have a set of records & we want to cluster w.r.t any two attributes, then we divide the related space (plane) into a grid structure and then we find the clusters.

Basic Grid-based Algorithm

1. Define a set of grid-cells
2. Assign objects to the appropriate grid cell and compute the density of each cell.
3. Eliminate cells, whose density is below a certain threshold t.
4. Form clusters from contiguous (adjacent) groups of dense cells

Advantages of Grid-based Clustering Algorithms

- Fast:
  - No distance computations
  - Clustering is performed on summaries and not individual objects; complexity is usually O(#-populated-grid-cells) and not O(#objects)
  - Easy to determine which clusters are neighboring
- Shapes are limited to union of grid-cells

Algorithm: CLIQUE (Clustering In Quest)

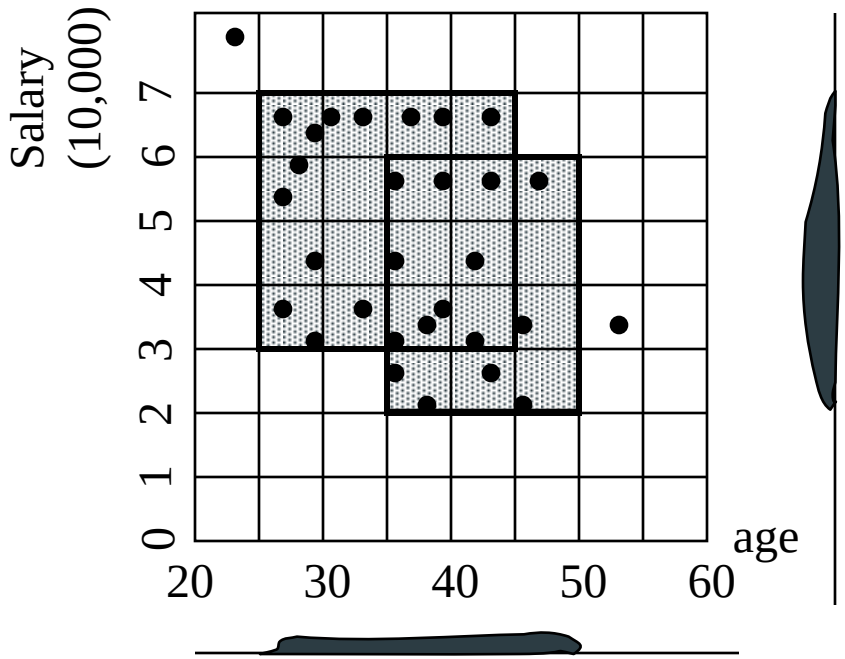
- Integrates density-based and grid-based clustering
- Useful for clustering high-dimensional data in large databases
- Based on the following:
  - Discovering overall distribution patterns (i.e., identification of the sparse and crowded areas (units) in space) of the given large data set which is not usually uniformly occupied by the data points.
  - A unit is dense if the fraction of total data points contained in it exceeds an input model parameter.

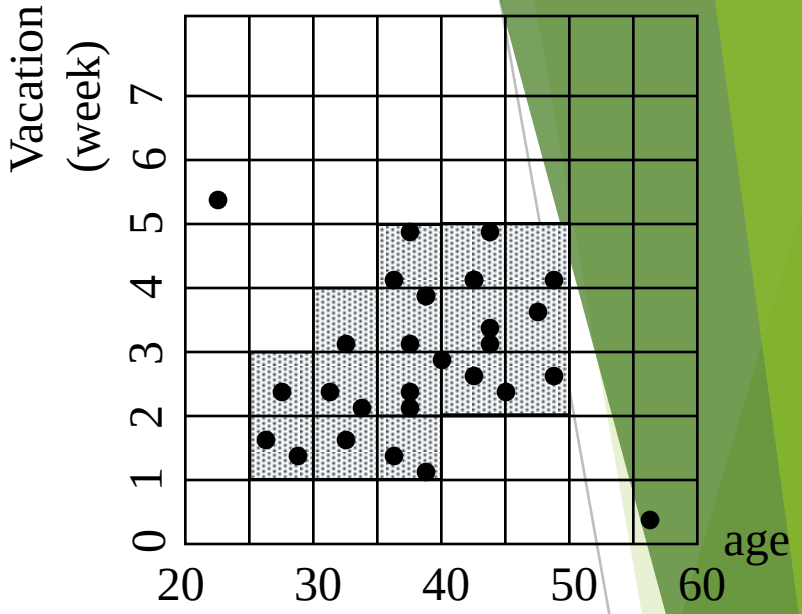
In CLIQUE, a cluster is defined as a maximal set of dense units

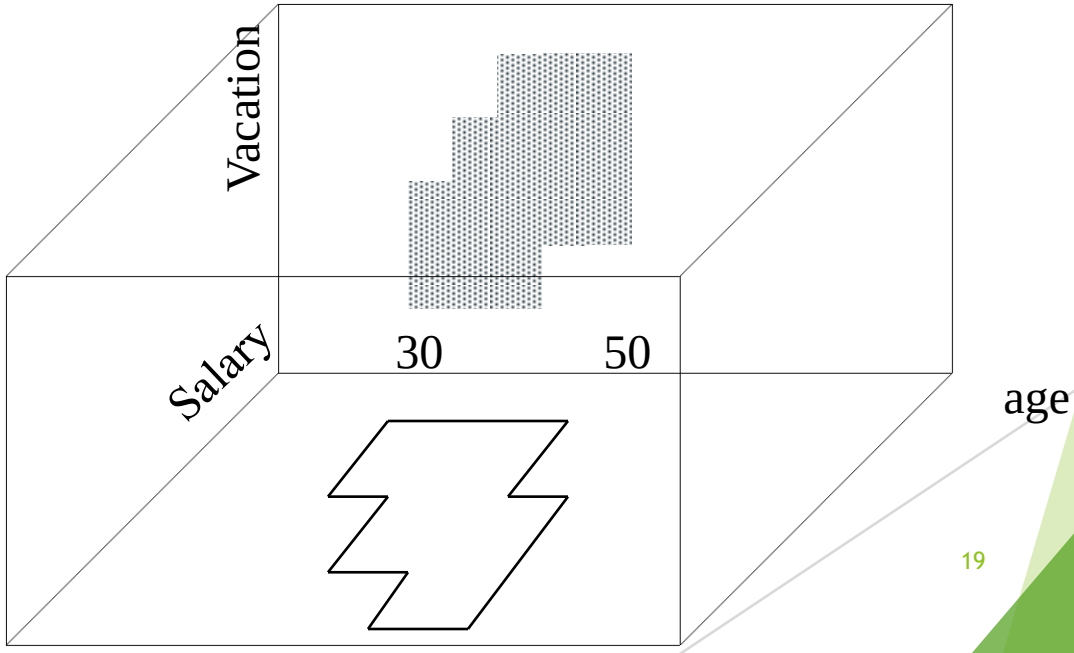
CLIQUE(cont…)

CLIQUE performs multidimensional clustering in two steps:

1. Partitioning the n-dimensional data space into non-overlapping rectangular units, identifying the dense units among these.
2. Generates a minimal description for each cluster – determines the maximal region that covers the cluster of connected dense units & then a minimal cover for each cluster.


Salary
(10,000)
0 1 2 3 4 5 6 7
20 30 40 50 60
age


Vacation
(week)
0 1 2 3 4 5 6 7
20 30 40 50 60
age


Vacation
Salary
30
50
age

Model-Based Clustering Methods

- Attempt to optimize the fit between the given data and some mathematical model
- Based on the assumption that data are generated by a mixture of underlying probability distributions.

Two major approaches:

1) Statistical Approach

2) AI (neural network) Approach

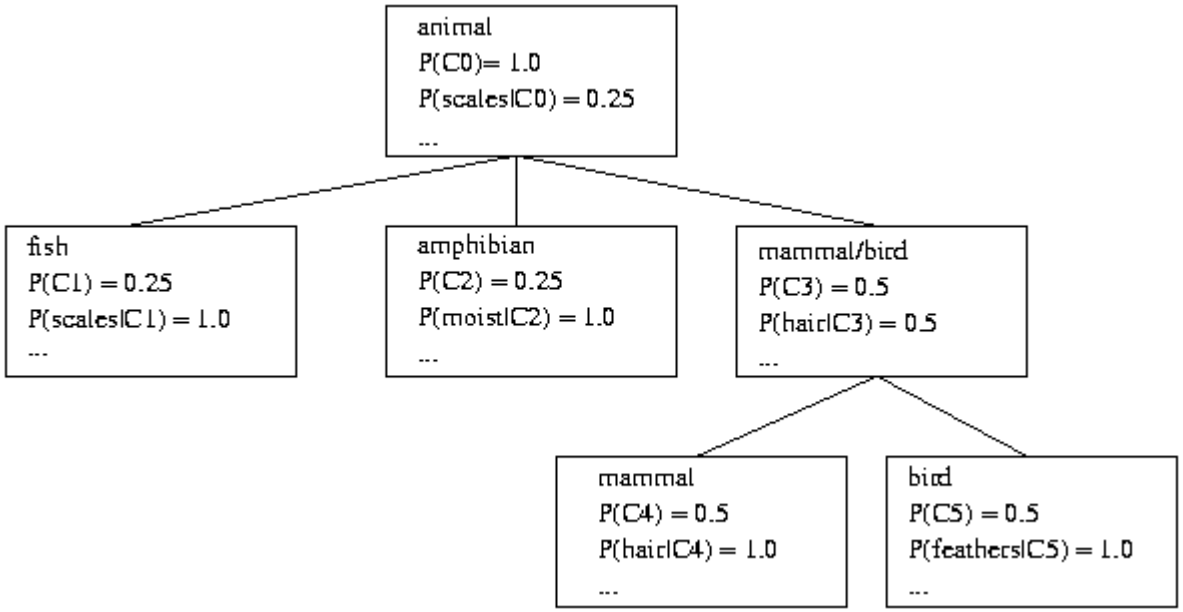
Statistical Approach

- Conceptual clustering is
  - A form of clustering in machine learning
  - Produces a classification scheme for a set of unlabeled objects
  - Finds characteristic description for each group (concept /class)
  - Usually adopts a statistical approach that uses probability measurements in determining the concepts/clusters

- COBWEB (Fisher'87)
  - A popular and simple method of incremental conceptual learning
  - Creates a hierarchical clustering in the form of a **classification tree**
  - Each node refers to a concept and contains a probabilistic

COBWEB Clustering Method

A classification tree



- **Neural network (AI) approaches**
  - Represent each cluster as an *exemplar*, acting as a "prototype" of the cluster
  - New objects are distributed to the cluster whose exemplar is the most similar according to some distance measure
  - Uses two prominent clustering methods:

1) *Competitive learning*

  - Involves a hierarchical architecture of several units (neurons)
  - These units are said to be in competition for input patterns
  - The output unit that provides the highest activation to a given input pattern is declared the winner and is moved closer to the input pattern, whereas the rest of the neurons are left unchanged

- Thus, neurons compete in a "winner-takes-all" fashion for the object currently being presented since only the winning neuron is updated

2) *Self-organizing feature maps (SOMs)*

- Clustering is also performed by having several units competing for the current object
- The unit whose weight vector is closest to the current object wins
- The winner and its neighbors learn by having their weights adjusted
- SOMs are believed to resemble processing that can occur in the brain
- Useful for visualizing high-dimensional data in 2- or 3-D space

Constraint – Based Clustering

- Class of semi-supervised learning algorithms.
- Clustering under various kinds of constraints
- Constraints guide a clustering algorithm to find clusters in a data set which satisfy the specified constraints
- Used in real world applications (e.g., GPS)
- Incorporates ***two*** types of constraints with a Data Clustering Algorithm which define a relationship between two data instances:
    1) Must-link constraint
    2) Cannot-link constraint
- A cluster which conforms to both of these constraints is termed as a ***chunklet***
- What if there is constraint violation?

Example Algorithm :

Cop k-means

- Initialize k cluster centres
- Assign Phase

  – objects are assigned to closest cluster centre *without violating constraints*

  For all objects try to assign it to closest k

  - 1. No constraint broken: –

    Assign object o to cluster k.

  - 2. Broken → is there a next closest cluster ?

    – Yes → Back to 1.

    – No → 3.

  - 3. fail

- Update Cluster Centres

  – update the cluster centres to the mean of constituent objects

What Is Outlier Discovery?

- What are outliers?
  - The set of objects are considerably dissimilar from the remainder of the data
  - Can be caused by measurement or execution error

- Outlier mining concept and problem

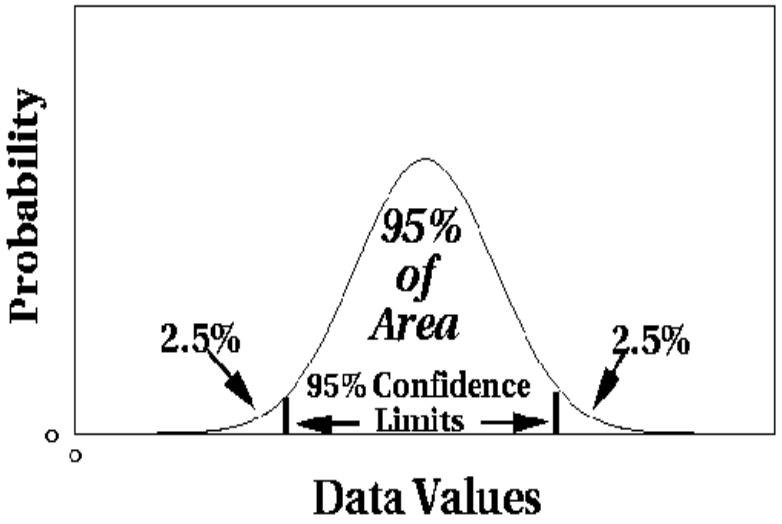
- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis

Outlier Discovery: Statistical Approaches

- Assume a probability or distribution model for the given data set (e.g. normal distribution) & then identifies outliers w.r.t model using a discordancy tests

- Use discordancy tests depending on
  - data distribution
  - distribution parameter

    (e.g., mean, variance)
  - number of expected outliers



- Drawbacks
  - most tests are for single attribute
  - In many cases, data distribution may not be known

Outlier Discovery: Distance-Based Approach

- Introduced to counter the main limitations imposed by statistical methods
- Distance-based outlier: A DB(p, D)-outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from O
- Avoids the excessive computations…

- Algorithms for mining distance-based outliers :
  - Index-based algorithm
  - Nested-loop algorithm
  - Cell-based algorithm

Outlier Discovery: Deviation-Based Approach

- Identifies outliers by examining the main characteristics of objects in a group
- Objects that “deviate” from this description are considered outliers
- Uses two main techniques:
- **Sequential exception technique**
  - simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects
- **OLAP data cube technique**
  - uses data cubes to identify regions of anomalies in large multidimensional data