clastering

→ un supervised

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis (공간데이터)
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research) 유사성향 고객군집화
- WWW 웹문서
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost 보험청구多 → 보험사기 감시
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults 진앙지 표시 → 군집 (지진분포)

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity 높은 군집내 유사성
  - low inter-class similarity 낮은 군집간 유사성

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*: minkowski sum

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

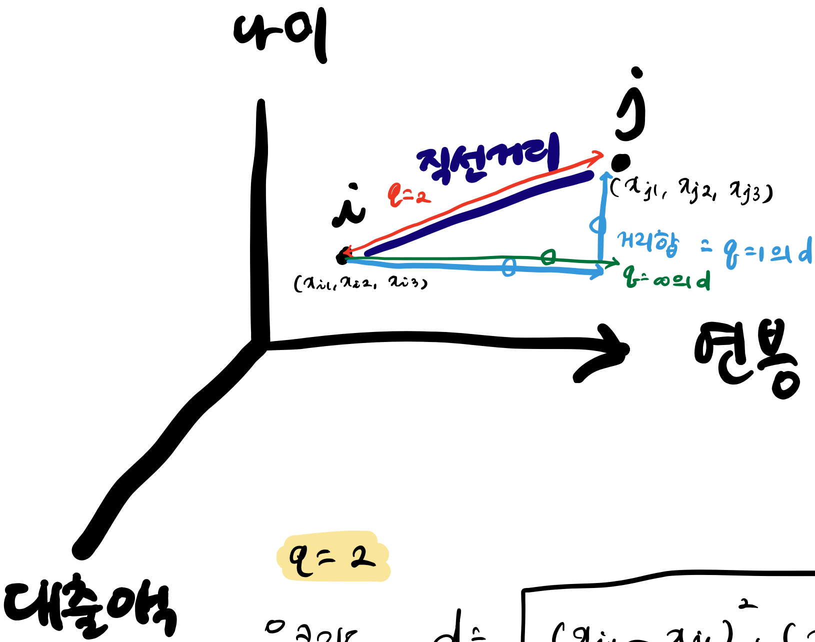
q=2

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

Minkowski distance



직선거리 간격에 따라 유사도 측정 [0, 1]

아주 다르다 / 동일

q=2

유클리드 $d = \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+(x_{i3}-x_{j3})^2}$

q=1

$d = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + |x_{i3}-x_{j3}|$

q=∞

$d = \max(|x_{i1}-x_{j1}|, |x_{i2}-x_{j2}|, |x_{i3}-x_{j3}|)$ $\boxed{3^{\infty}}$ vs $\cancel{2^{\infty}}$

거억,

中 제일 큰 값이 d가 됨.

Similarity and Dissimilarity Between Objects (Cont.)

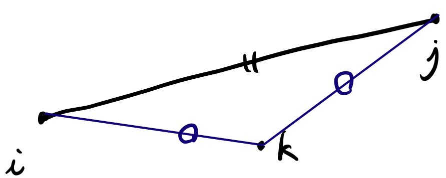
- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- Properties

4가지 특징

- $d(i,j) \geq 0$
- $d(i,i) = 0$ 같은점거리
- $d(i,j) = d(j,i)$ 거리는 방향성X
- $d(i,j) \leq d(i,k) + d(k,j)$

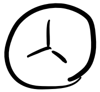


triangular inequality

Major Clustering Approaches

클러스터링 방법

- <u>Partitioning algorithms</u>: Construct various partitions and then evaluate them by some criterion 공간을 쪼개면서

  계층형
- <u>Hierarchy algorithms</u>: Create a hierarchical decomposition of the set of data (or objects) using some criterion 

  밀도기반
- <u>Density-based</u>: based on connectivity and density functions

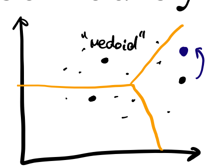
  격자기반 (밀도 높~낮)
- <u>Grid-based</u>: based on a multiple-level granularity structure

  데이터 분포에 적합한 모델
- <u>Model-based</u>: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Partitioning Algorithms: Basic Concept

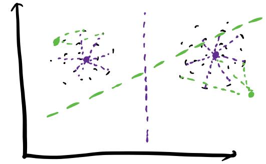
단점) k가 주어져야 한다는 것

- Partitioning method: Construct a partition of a database ***D*** of ***n*** data item objects into a set of (주어진 값) ***k*** 개의 군집 clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion 나누는 최적 k를 찾기
  - Global optimal: exhaustively enumerate all partitions 어려움
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - (spss에 있음) *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster 임의의 중심점 선택
  - 더 좋다 — *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster 실제로 있는 임의의 대표점 선택.
    → 대용량 데이터 처리에 부적합



바꿔보기
→ 해쳐모여
중심과의 거리가 작아질수록 Good!!

k= 작 → 큰
군집의 품질을 고려해서
k값 정하기

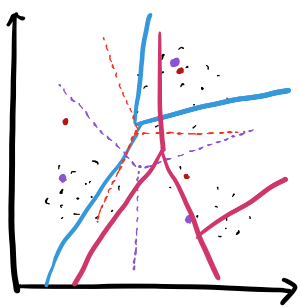


The *K-Means* Clustering Method

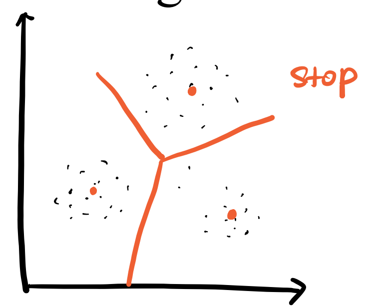
품질계산) 거리 계산으로 계산.
중점 (seed)까지

주어진 k

- Given *k*, the *k-means* algorithm is implemented in 4 steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
  - Assign each object to the cluster with the nearest seed point.
  - Go back to Step 2, stop when no more new assignment.



k=3
① 임의로 3개의 공간으로 나눔
② 중심좌표 (seed) 찾기
③ 중심좌표와 가까운 쪽으로 assign
④ 2 - 3 반복 → change 가 없을때 stop
IF k=4, 이상한 모양

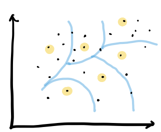


SPSS k-means 확인
군집화 PPT에서

The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - *PAM* works effectively for small data sets, but does not scale well for large data sets
- *CLARA* (Kaufmann & Rousseeuw, 1990)

→ sample 추출. ⇒ sample만 가지고 PAM 알고리즘 돌리기

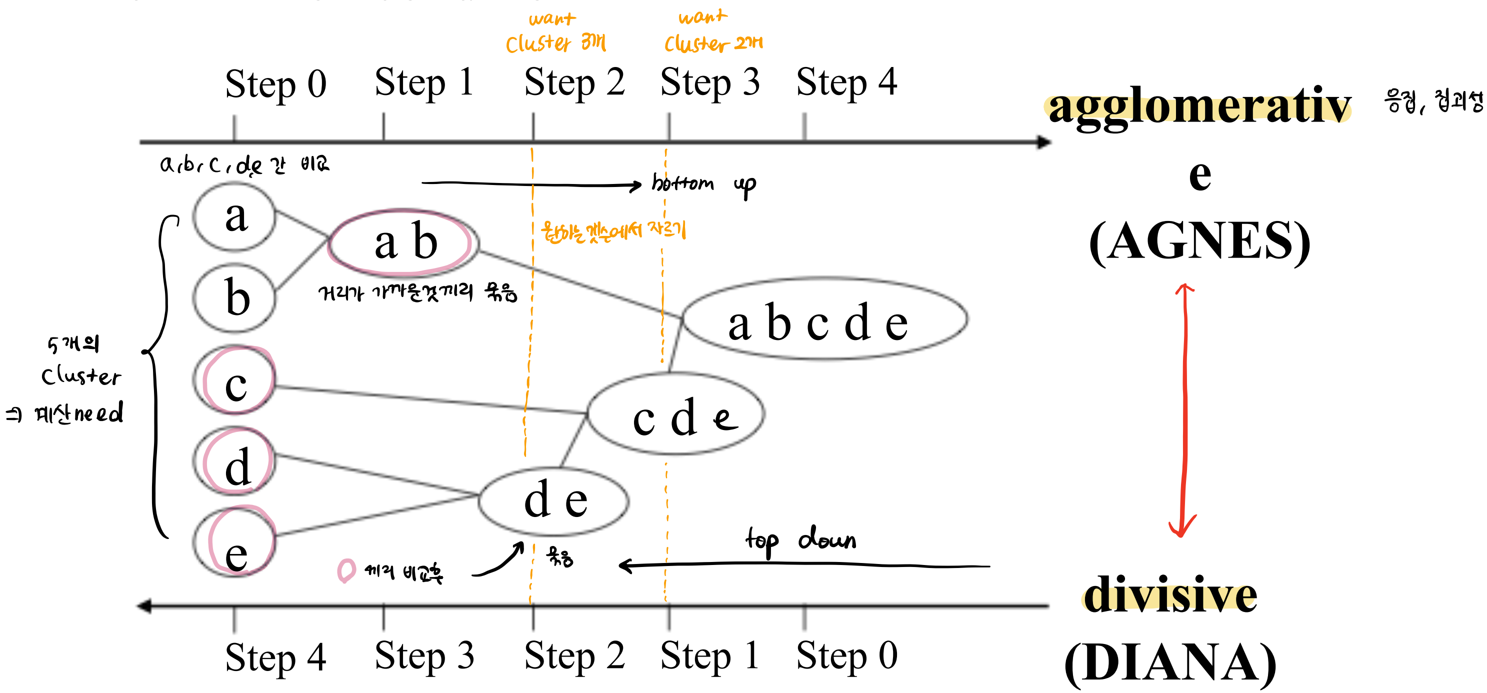


파티셔닝
k값 필요 (군집의 수)
임의로 나눈고 → 데이터 할당 → 중심구하고 → 데이터할당

Hierarchical Clustering

계층화 장점 (k값 필요X) 단점 데이터 多 → 효율성↓

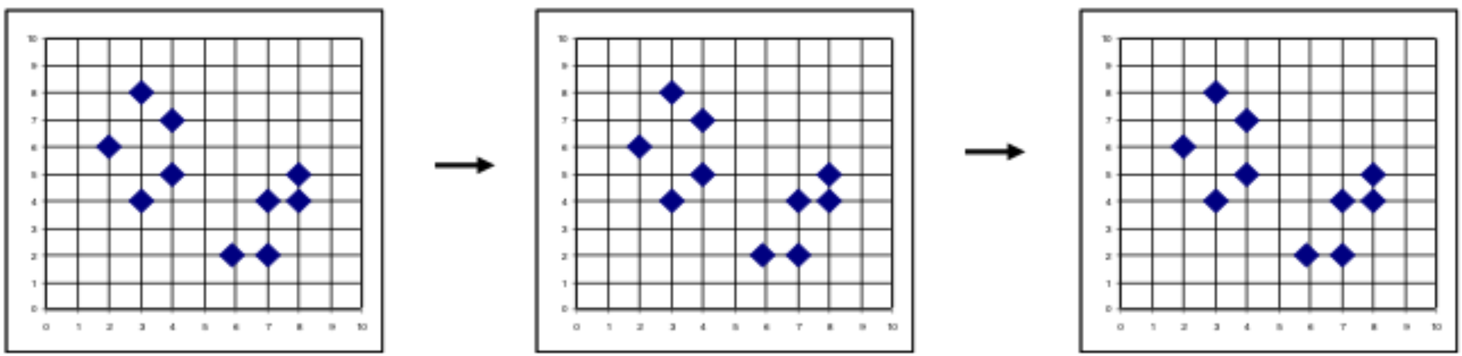
- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



AGNES (Agglomerative Nesting)

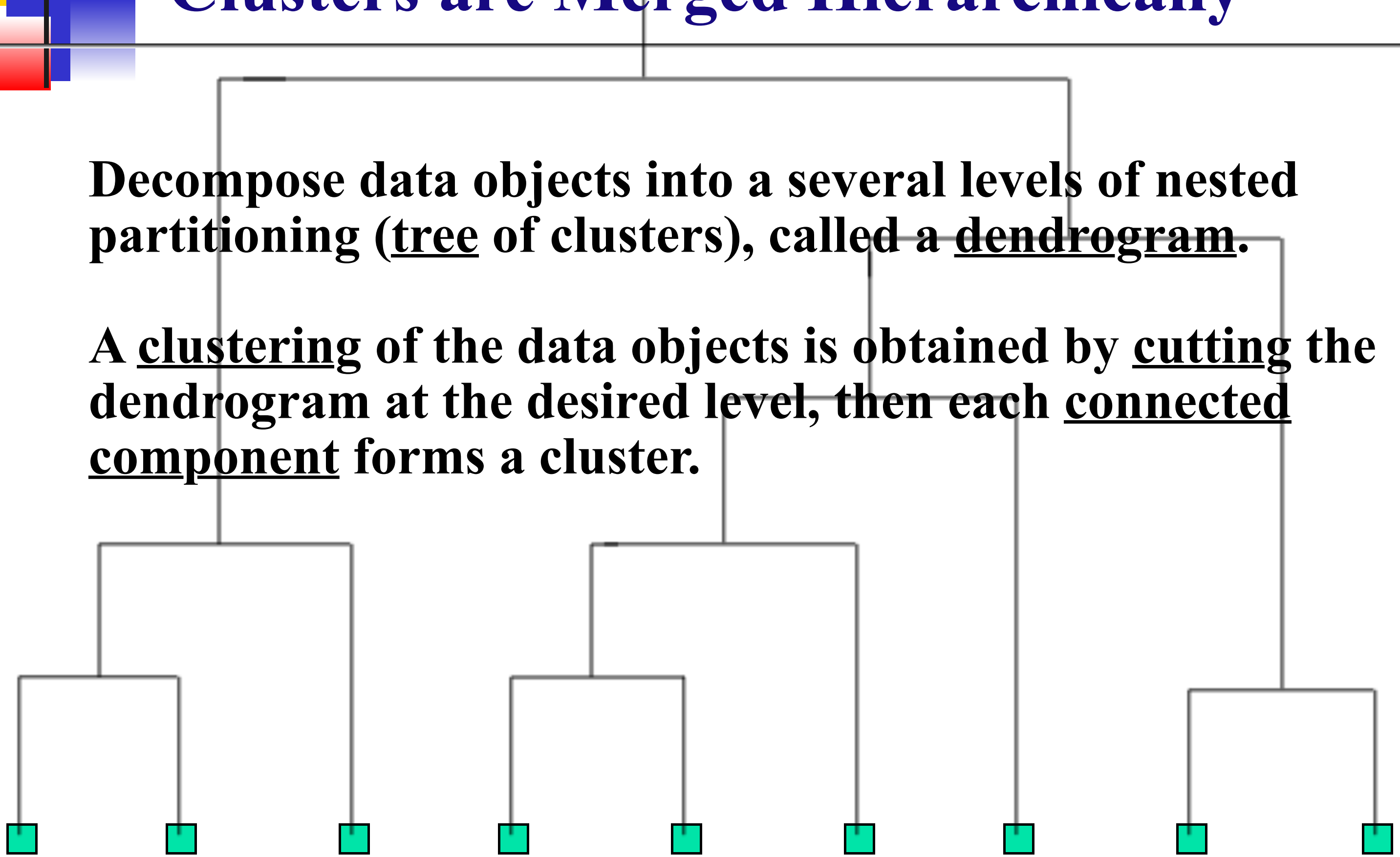
- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

참고만



PPT확인 need

A *Dendrogram* Shows How the Clusters are Merged Hierarchically



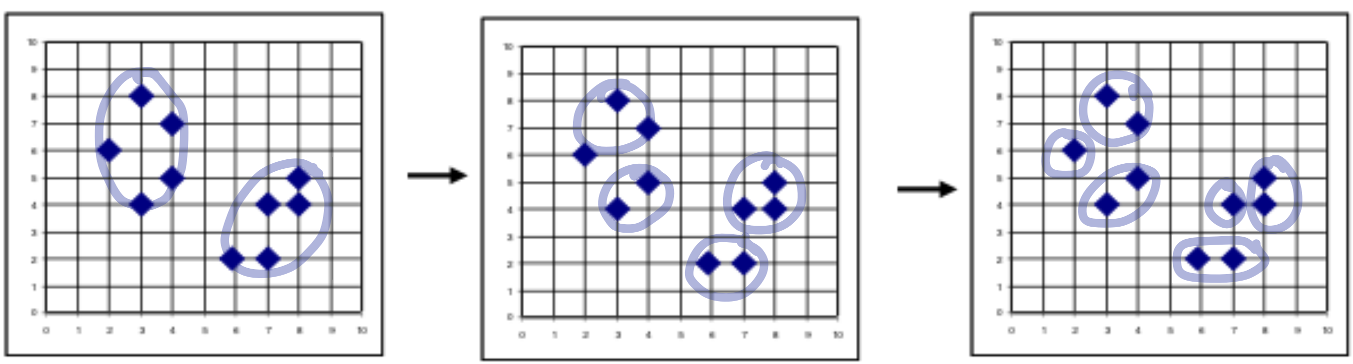
Decompose data objects into a several levels of nested partitioning (<u>tree</u> of clusters), called a <u>dendrogram</u>.

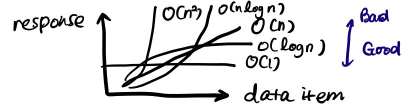
A <u>clustering</u> of the data objects is obtained by <u>cutting</u> the dendrogram at the desired level, then each <u>connected component</u> forms a cluster.

DIANA (Divisive Analysis)

이게 더 어렵죠.

- 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





More on Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods
  - do not scale well: time complexity of at least $O(n2)$, where n is the number of total objects
  - can never undo what was done previously 번복 불가능
- Integration of hierarchical with distance-based clustering
  - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
  - CURE (1998): selects well-scattered points from the cluster and then shrinks them towards the center of the cluster by a specified fraction
  - CHAMELEON (1999): hierarchical clustering using dynamic modeling

이런게 있다 (개선안)

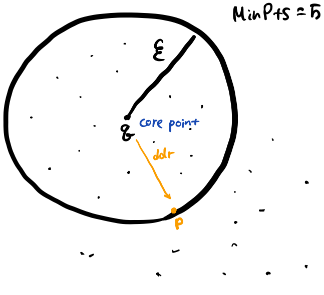
Density-Based Clustering Methods

밀도 기반 클러스터링

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape 임의의 모양의 클러스터를 만들수 있음
  - Handle noise
  - One scan 한번 스캔만으로 가능
  - Need density parameters as termination condition ε MinPts
- Several interesting studies:
  - <u>DBSCAN</u>: Ester, et al. (KDD'96)
  - <u>OPTICS</u>: Ankerst, et al (SIGMOD'99).
  - <u>DENCLUE</u>: Hinneburg & D. Keim (KDD'98)
  - <u>CLIQUE</u>: Agrawal, et al. (SIGMOD'98)

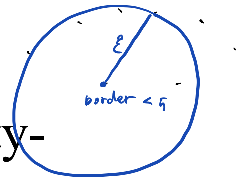
Density-Based Clustering: Background

- Two parameters**:**
  - ***Eps***: Maximum radius of the neighbourhood
  - ***MinPts***: Minimum number of points in an Eps-neighbourhood of that point



- ***NEps(p)***: ***{q belongs to D | dist(p,q) <= Eps}***

p로부터 입실론 반경안의 이웃들.

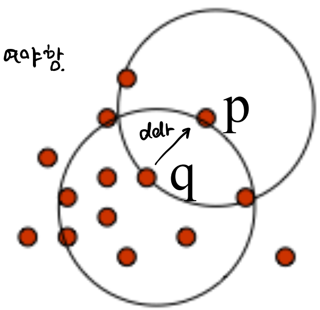


- Directly density-reachable**:** A point ***p*** is directly density-reachable from a point ***q*** wrt. ***Eps***, ***MinPts*** if

ddr 조건

  - 1) ***p*** belongs to ***NEps(q)*** q의 이웃이여야하고
  - 2) core point condition: core point 여야함.

    |***NEps (q)***| >= ***MinPts***

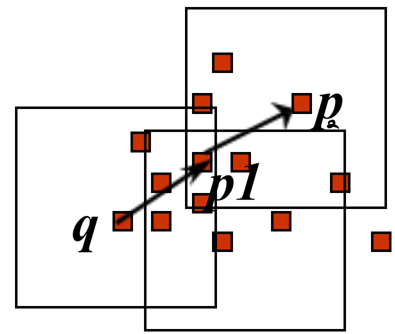


MinPts = 5

Eps = 1 cm

Density-Based Clustering: Background (II)

- Density-reachable:
  - A point *p* is density-reachable from a point *q* wrt. *Eps*, *MinPts* if there is a chain of points *p1*, …, *pn*, *p1* = *q*, *pn* = *p* such that *pi+1* is directly density-reachable from *pi*



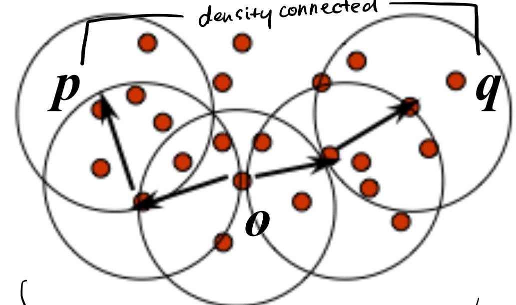
한번 방문시,
재 방문X

$q \xrightarrow{ddr} p_1 \xrightarrow{ddr} p_2 \xrightarrow{ddr} p_3$

dr

- Density-connected
  - A point *p* is density-connected to a point *q* wrt. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* wrt. *Eps* and *MinPts*.

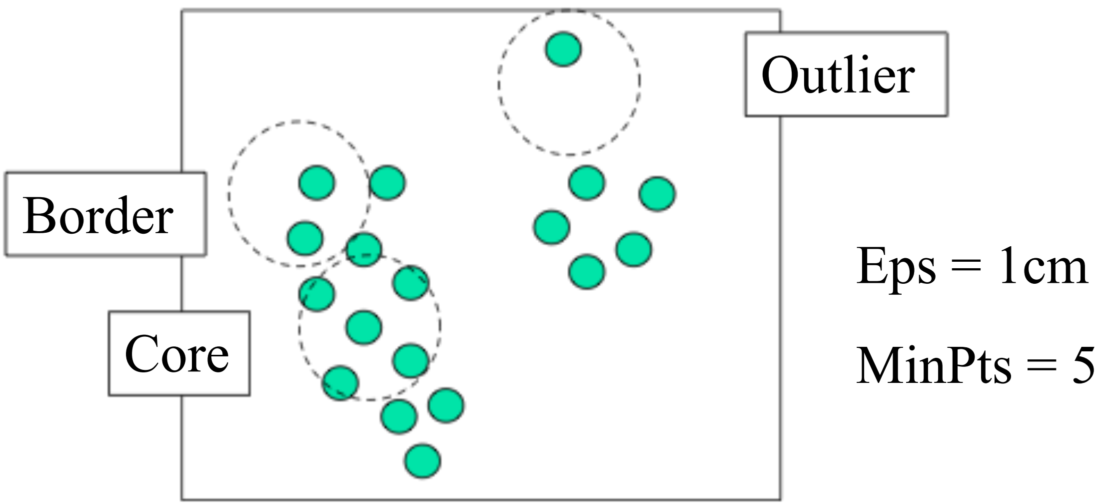
density connected



더이상 ddr 연결X 그룹 — cluster

DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise

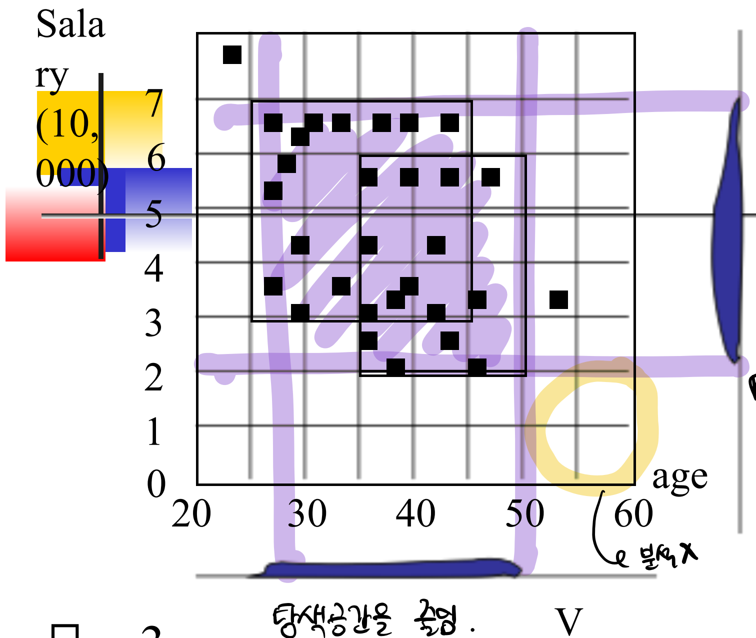


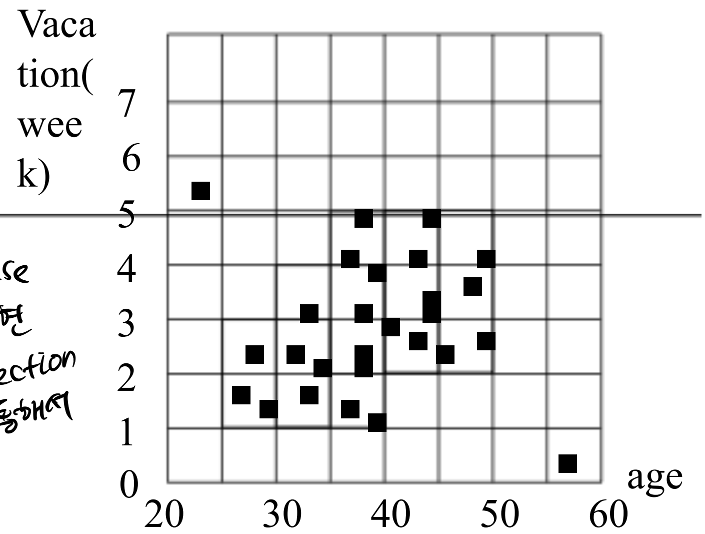
Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
  - STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - A multi-resolution clustering approach using wavelet method
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

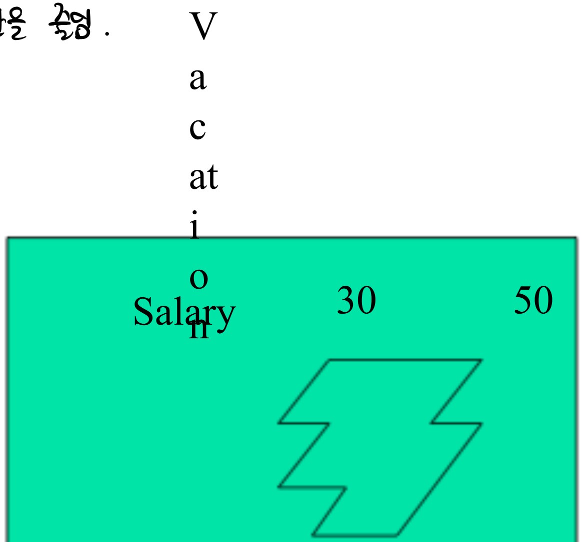
CLIQUE: The Major Steps

- Partition the data space and find the number of points that lie inside each cell of the partition.
- Identify the subspaces that contain clusters using the Apriori principle
- Identify clusters:
  - Determine dense units in all subspaces of interests
  - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
  - Determine maximal regions that cover a cluster of connected dense units for each cluster
  - Determination of minimal cover for each cluster


Salary (10,000)
7
6
5
4
3
2
1
0
20
30
40
50
60
age
dense 표현
projection 통해서
분석X
탐색공간을 줄임.


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

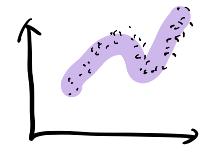
□ = 3


Vacation
Salary
30
50
age

수학적인 모델링 사용. (제일 이상적 but 어려움, 느림)

Model-Based Clustering Methods

- Attempt to optimize the fit between the data and some mathematical model



$y = ax^3 + bx^2 + cx + d$

- Statistical and AI approach
  - Conceptual clustering
    - A form of clustering in machine learning
    - Produces a classification scheme for a set of unlabeled objects
    - Finds characteristic description for each concept (class)
  - COBWEB (Fisher'87)
    - A popular a 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 description of that concept

신경망분석

PPT 추가하기