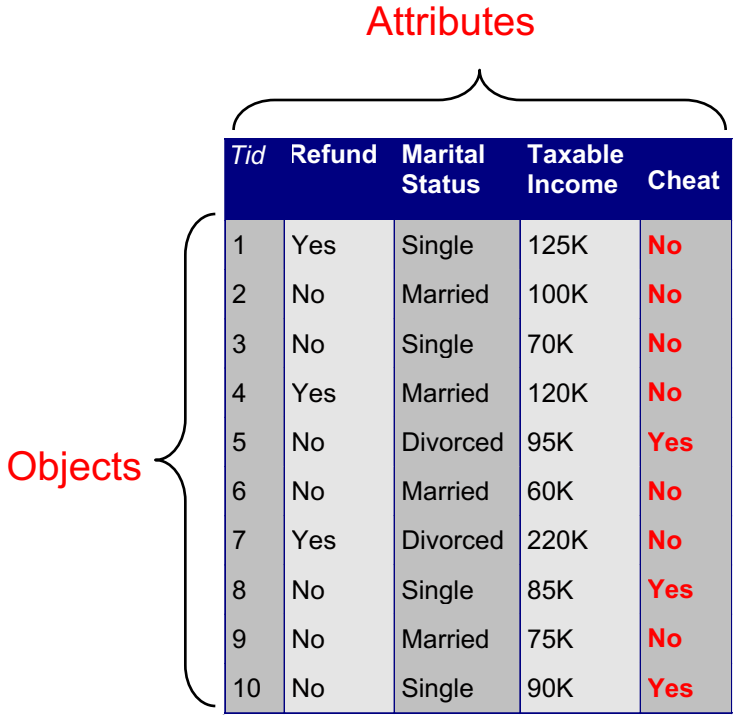
Data

데이터란?

- 데이터 객체(테이블의 행)와 데이터 속성(테이블의 열)의 한 집합

- 데이터 속성은 데이터 객체의 특성이나 성질을 나타냄.
  - 예: 사람의 눈 색깔, 온도 등
  - 데이터 속성은 변수, 필드, 특징과 흡사한 개념

- 데이터 속성의 한 집합은 하나이 객체를 기
  - 객체는 레코드, 케이스, 샘플 등으로 알려짐



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

데이터 속성의 형태

- 명목 (nominal) 고유값
  - 예: 학생번호, 눈 색깔, 우편 번호
- 순서 (ordinal)
  - 예: 순위 (신라면에 대한 맛을 1부터 10까지 나타내는 형태 포함), 성적, 키 (큰편, 중간, 작은편)
- 간격 (interval)
  - 예: 날짜, 온도 등
- 비율 (ratio)
  - 예: 길이, 시간, 횟수

데이터의 형태

- 레코드 데이터
  - 데이터 행렬
  - 문서 데이터
  - 거래 내역 데이터
- 그래프 데이터
  - 쇼셜 네트워크
  - 분자 구조 (2013년 노벨화학상 수상자 – Arieh Warshel, Michael Levitt, Martin Karplus)
- Ordered 데이터
  - DNA 염기 서열 데이터
  - 공간적 순서 데이터 (예: 위성 사진 분석 데이터)
  - 시간적 순서 데이터 (예: 주가 지수, 시간별 기온변화)
  - 순차적 데이터 (예: 거랙 내역 데이터에서 시간적 연관 관계를 보이는 데이터, A를 구매한 고객은 B를 구매)

레코드 데이터 – 테이블, 데이터 행렬

- 정해진 갯수의 데이터 속성이 정의된 데이터

정형화된 데이터

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

레코드 데이터 – 문서, 거래 내역 데이터

- 각 문서는 항목 벡터

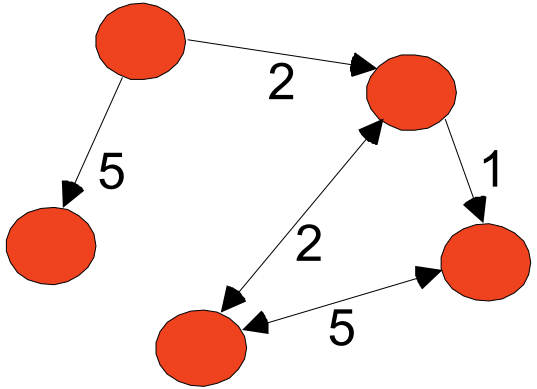
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- 거래 내역 데이터

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

그래프 데이터

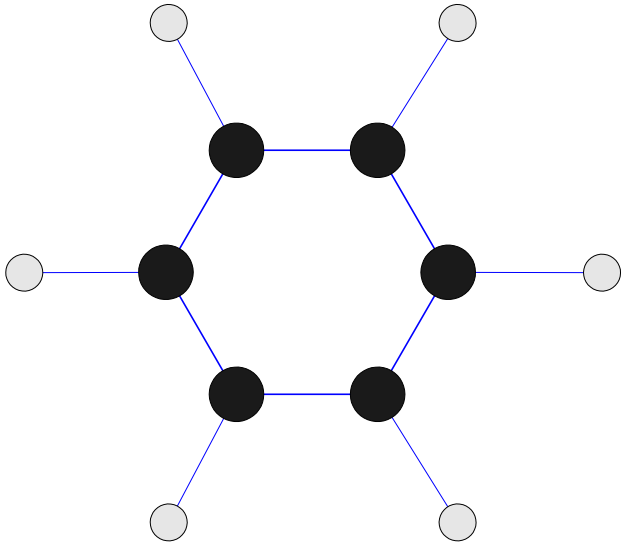
예) 일반 그래프, HTML Links



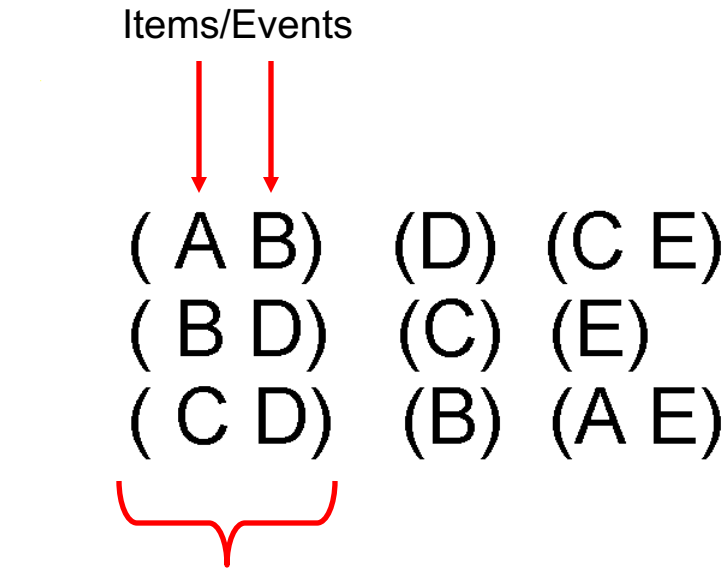

```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

그래프 데이터 – 분자 구조

예) 벤젠: C_6H_6



Ordered 데이터 – 순차적인 거래 내역 데이터



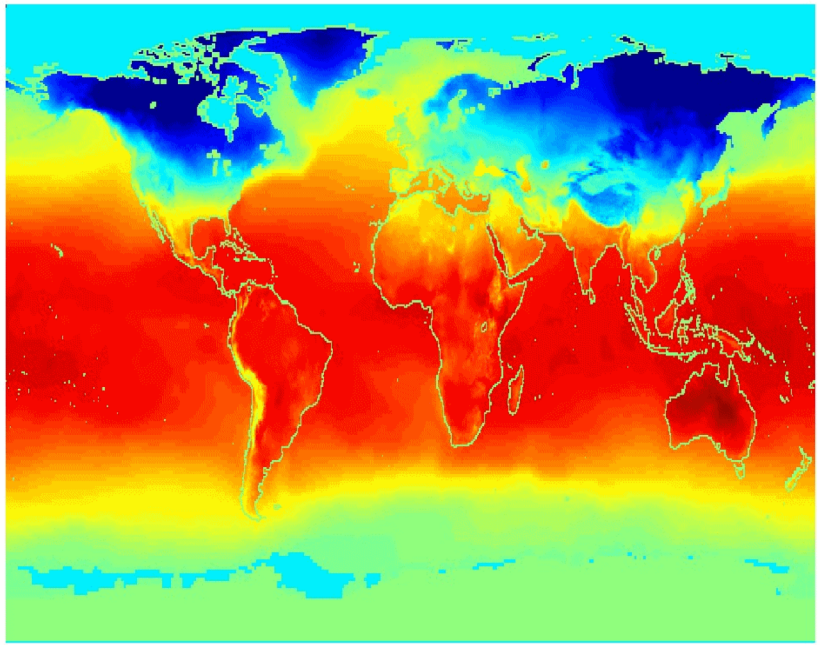
Ordered 데이터 – 유전체 염기 서열 데이터

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered 데이터 – 시공간 데이터

지구의 월평균 온도

1월



빅데이터의 특징

- Volume (데이터 양)
  - 테라바이트(TB) 단위 이상의 대량 데이터
- Velocity (속도)
  - 데이터의 수집과 분석을 실시간으로 처리해야 함
- Variety (다양성)
  - 관계 데이터베이스에 저장된 정형화된 데이터
  - 책, 의료 기록, 비디오, 오디오, 위치 정보, 로그 기록, 이메일, SNS 등 비정형 데이터

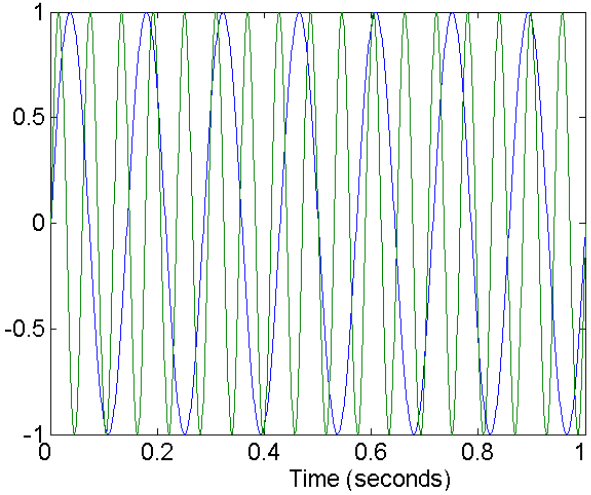
데이터 품질 (Data quality)

- 데이터 품질과 관련된 이슈
  - 어떤 종류의 데이터 품질 문제가 있는가?
  - 어떻게 데이터의 문제점을 발견해낼 수 있을까?
  - 데이터 문제점을 해결하기 위한 방안은?

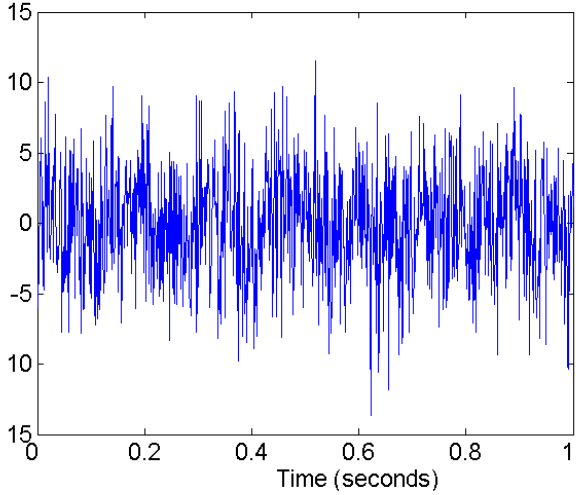
- 데이터 품질과 관련된 문제의 예
  - 데이터 노이즈 및 이상점 (outliers)
  - 손실된 데이터 (missing values)
  - 데이터 중복

데이터 노이즈

- 데이터 노이즈란 원본 데이터 값이 훼손된 경우를 의미
  - 예) 전화 녹음 상태 불량으로 통화자의 음성이 왜곡된 경우



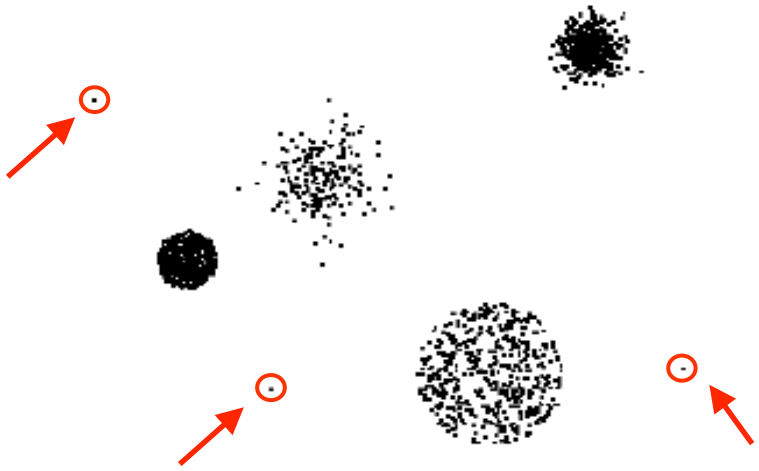
Two Sine Waves



Two Sine Waves + Noise

이상점 (Outliers)

- 이상점이란 대부분의 다른 데이터 객체와 상당히 다른 특성을 보이는 데이터 객체



데이터 값 손실 (Missing values)

- 데이터 값 손실 발생 원인
  - 모든 정보가 다 수집되는 것은 아님
    (e.g., 사람들이 개인정보 공유하는 것을 꺼림)
  - 어떤 데이터 속성은 모든 사람들에게 적용되지 않음
    (e.g., 연수입 – 아이들에게 적용하기 어려움)

- 데이터 값 손실 해결 방안
  - 해당 데이터 객체를 삭제함
  - 손실된 값을 추정함
  - 데이터 분석 시에 손실된 값은 무시함
  - 모든 가능한 경우를 다 고려함 (각 해당하는 경우에 대한 확률을 계산하여 가중치 부여)

중복 데이터

- 데이터 객체가 여러 곳에 중복되어 관리될 가능성이 있음
  - 다양한 정보소스로부터 데이터를 통합할 때 고려되어야 하는 주요 이슈

- 예:
  - 여러 개의 이메일을 가진 사람 (이메일이 주요 ID가 되는 경우)

- 데이터 클리닝
  - 중복 데이터를 다루는 과정

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Data Preprocessing

- **Aggregation**
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation (결합)

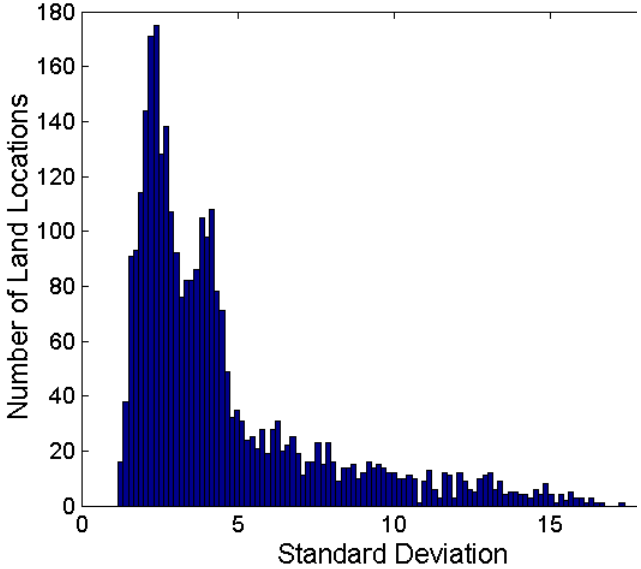
- 두 데이터 속성이나 객체를 하나로 묶음

- 목적

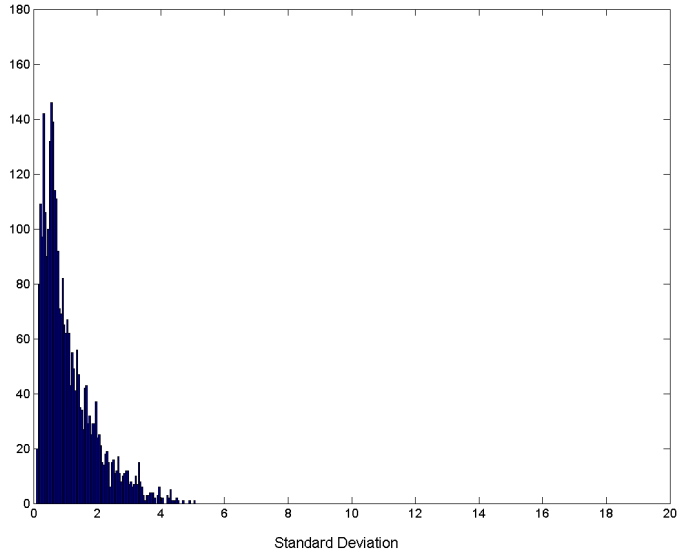
1. Data reduction
   - 속성이나 객체의 개수를 줄이고자 할 때
2. 스케일 변동
   - 도시 스케일에서 지역, 주나 국가 단위로 결합하고자 할 때
3. 데이터 안정성 확보
   - Aggregated data는 가변성이 떨어지는 경향이 있음

Aggregation – 예

- 호주 강수량 변화



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Data Preprocessing

- Aggregation
- **Sampling**
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

샘플링 (Sampling)

- 샘플링: 데이터 선택 시 적용되는 주요 기법
  - 데이터 선수 조사와 최종 데이터 분석 두 경우 모두에 자주 사용됨

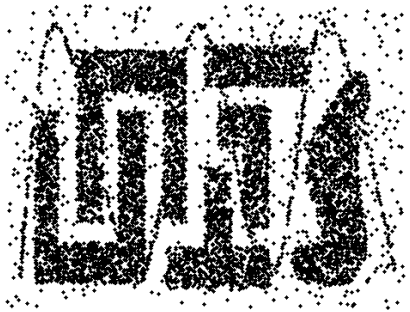
- 통계학자들이 샘플링을 하는 이유
  - 전체 데이터를 얻는 것이 시간 및 비용 측면에서 비효율적임

- 데이터 마이닝에서 샘플링을 사용하는 이유
  - 데이터 전체를 처리하는 것이 시간 및 비용 측면에서 비효율적임

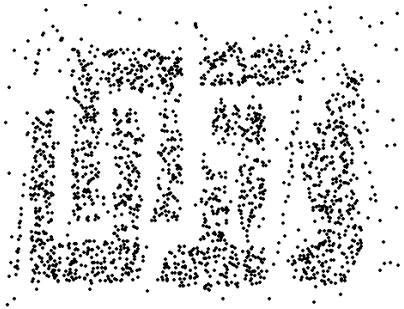
샘플링 종류

- 단순 랜덤 샘플링
  - 각 항목을 동일한 확률로 선택함
- 비복원 샘플링
  - 한번 선택되면 제외 시킴
- 복원 샘플링
  - 같은 항목이 다시 선택될 수도 있음
- 계층화 샘플링 (Stratified sampling)
  - 데이터를 파티션으로 나누고 각 파티션에서 랜덤하게 샘플링

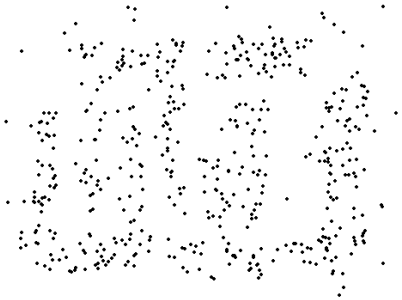
샘플 크기



8000 points



2000 Points



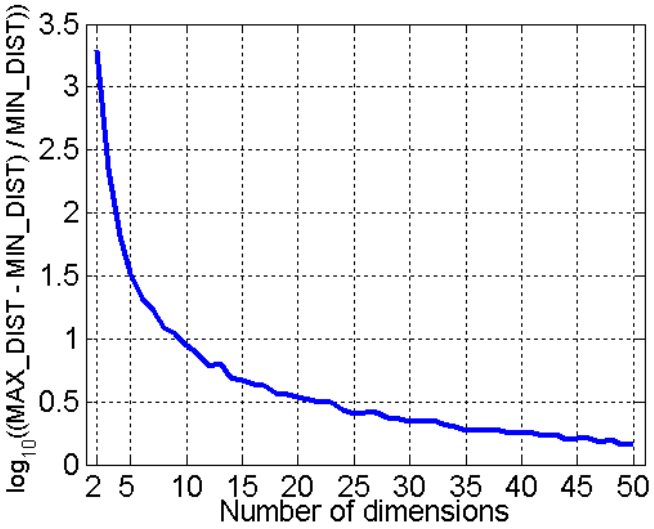
500 Points

Data Preprocessing

- Aggregation
- Sampling
- **Dimensionality Reduction**
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Dimensionality

- 차원수가 증가하게 되면 공간상에서 데이터가 성기게 분포하게 됨
- 데이터 클러스터링이나 이상점 발견 기법에서 정의되는 데이터 포인트간 밀도와 거리의 의미가 줄어들게 됨



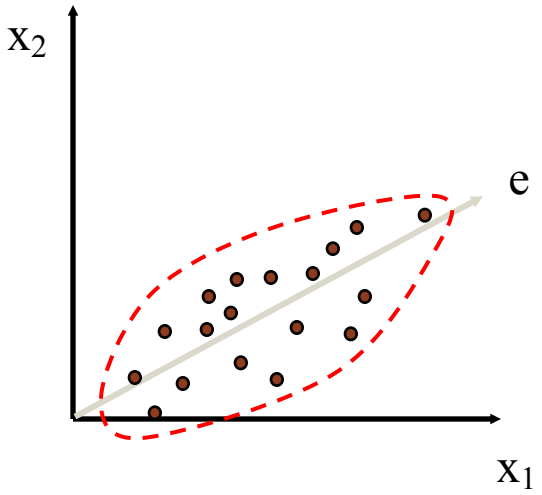
- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Dimensionality Reduction (차원 감소)

- 목적
  - 데이터 밀도를 높임
  - 데이터 마이닝 알고리즘에 요구되는 시간과 메모리 양을 감소시킴
  - 데이터를 좀 더 쉽게 시각화함
  - 불필요한 사항이나 노이즈를 제거하는데 도움이 됨
- 기법
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

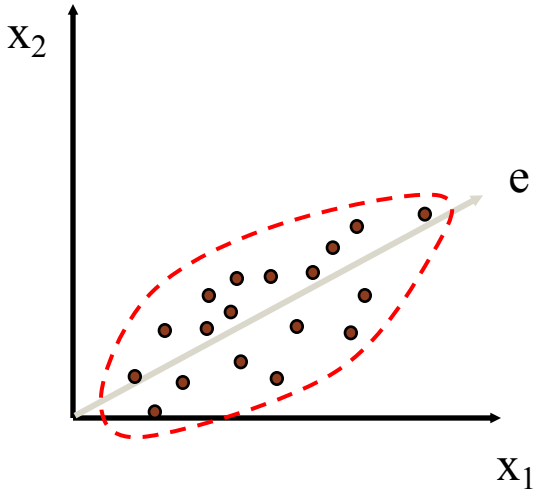
Dimensionality Reduction: PCA

- 목적: 데이터 다양성을 가장 많이 반영하는 projection을 찾는 것



Dimensionality Reduction: PCA

- Covariance matrix의 eigenvectors를 찾음
  - Covariance – 두 변수간의 correlation (연관성)을 나타냄, 두 변수가 연관성이 없으면 covariance는 0
- Eigenvectors이 새로운 공간을 정의



Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- **Feature subset selection**
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Feature Subset Selection

- Another way to reduce dimensionality of data

- 중복된 요소들
  - 많은 정보들이 하나 혹은 그 이상의 데이터 속성으로 중복되어 있음
  - 예) 한 물품에 대한 판매가와 소비세

- 연관성 없는 요소들 (Irrelevant features)
  - 데이터 마이닝에 불필요한 정보들
  - 학생들의 성적을 예측하는데 학생증 번호는 대부분의 경우 무관함

Feature Subset Selection – 기법들

- Brute-force (무차별 대입) approach:
  - 모든 가능한 경우의 subsets을 입력 데이터로 이용하여 데이터 마이닝 알고리즘을 적용해보는 방식
- Embedded approaches:
  - 데이터 마이닝 알고리즘 내에서 자연스럽게 이용될 features가 선택이 되는 방식
- Filter approaches:
  - 데이터 마이닝 알고리즘을 실행하기 전에 이용될 features를 선택하는 방식
- Wrapper approaches:
  - 가장 최적의 features를 선택하기 위해 데이터 마이닝 알고리즘을 블랙박스로 사용하는 방식

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- **Feature creation**
- Discretization and Binarization
- Attribute Transformation

Feature Creation

- 원래 attributes로 효율적으로 표현하지 못했던 중요한 정보를 표현하기 위해 새로운 attributes를 생성

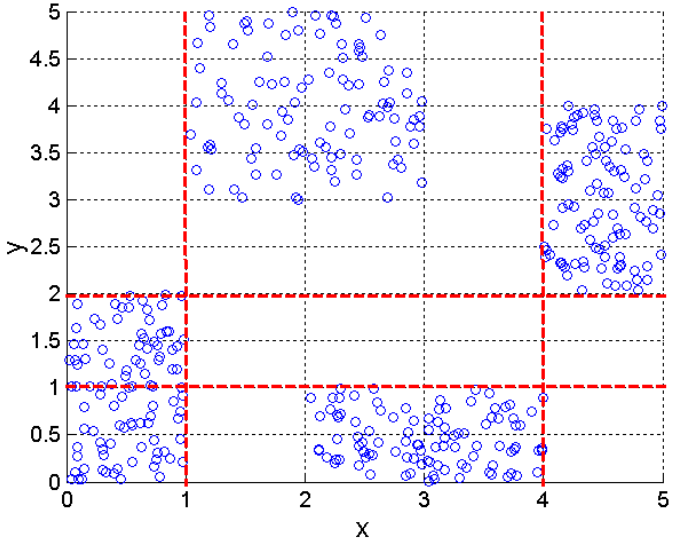
- 세 가지 방식:
  - Feature Extraction: domain-specific (특정 영역의 features를 추출)
  - Mapping Data to New Space (예: Geographic Information System)
  - Feature Construction: 여러 features를 합침

Data Preprocessing

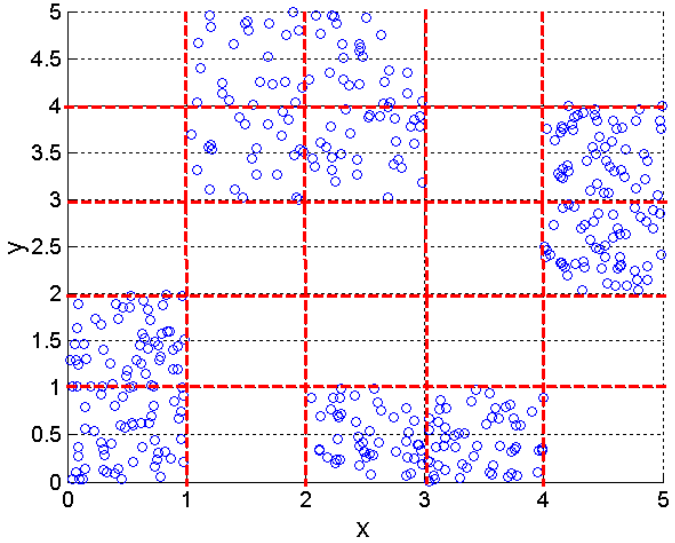
- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- **Discretization and Binarization**
- Attribute Transformation

Discretization using class labels

- 연속 변수값 ➔ 이산 변수값

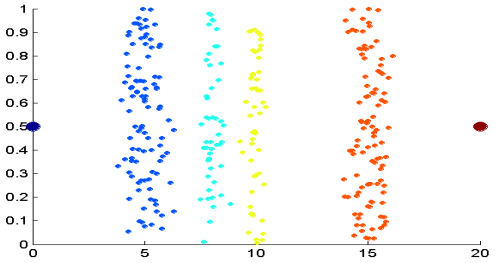


3 categories for both x and y

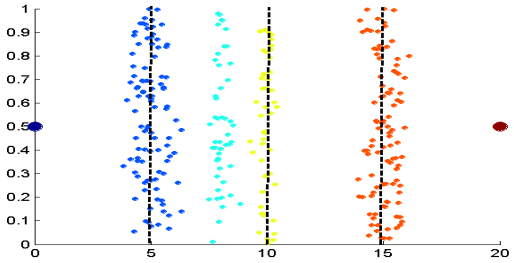


5 categories for both x and y

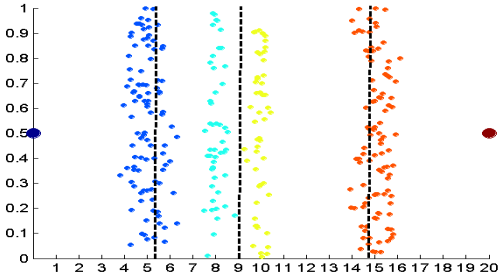
Discretization without using class labels



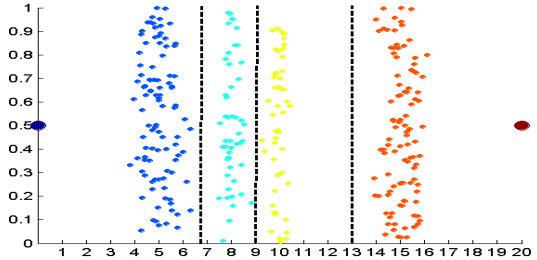
Data



Equal interval width



Equal frequency



K-means

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- **Attribute Transformation**

Attribute Transformation

A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

- Simple functions: x^k, log(x), e^x, |x|
- Standardization and Normalization

-1~1사이 값으로 변화