DATA

Content

- Attributes and Objects
- Types of Attributes
- Types of Data
- Data Quality
- Similarity and Distance
- Data Preprocessing

Data

Database: collection of **data objects**

record, *point*, *vector*, *instance*,, *point event*, *case*, ***sample, observation***, *entity*

data objects are described by a number of **attributes** capture the basic characteristics of an object

variable, *characteristic*, *field*, ***feature***, ***dimension***

Data

Attributes

Objects

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**

Different Types of Attributes

Nominal

The values of a nominal attribute are just different Names

Examples: ID numbers, eye color

Ordinal

The values of an ordinal attribute provide enough information to order objects.

Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}

Interval

For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists.

Examples: calendar dates, temperatures in Celsius or Fahrenheit.

Ratio

Both differences and ratios are meaningful.

Examples: temperature in Kelvin, length, counts

Different Types of Attributes

The type of an attribute depends on which of the following properties/operations it possesses:

Distinctness:	$= \quad \neq$
Order:	$< \quad >$
Differences are meaningful :	$+ \quad -$
Ratios are meaningful	$* \quad /$

Nominal attribute: distinctness
Ordinal attribute: distinctness & order
Interval attribute: distinctness, order & meaningful differences
Ratio attribute: all 4 properties/operations

Different Types of Attributes

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	employee ID numbers, eye color, {*male, female*}	mode, entropy, contingency correlation, χ2 test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, length	geometric mean, harmonic mean, percent variation

Different Types of Attributes

Discrete Attribute

Has only a finite or countably infinite set of values

Examples: counts, or the set of words in a collection of documents

Often represented as integer variables.

Note: binary attributes are a special case of discrete attributes

Continuous Attribute

Has real numbers as attribute values

Examples: temperature, height, or weight.

Data

Asymmetric Attributes

only presence—a non-zero attribute value—is regarded as important

- Students and courses
- Words present in documents
- Items present in customer transactions

Asymmetric attributes typically arise from objects that are sets

General Characteristics of Data Sets

- **Dimensionality**
- **Sparsity**
- **Resolution**

Types of Data Sets

Types of Data Sets

1. Record Data

2. Graph-Based Data

3. Ordered Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes

1.1 Data Matrix

1.2 Document Data

1.3 Transaction or Market Basket Data

Types of Data Sets

1.1 Data Matrix

- data objects have the same fixed set of numeric attributes
- data objects are points in a multi-dimensional space
- each dimension represents a distinct attribute
- represented by an *m* by *n* matrix

Data matrix or pattern matrix

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

Types of Data Sets

1.2 Document Data

Each document becomes a 'term' vector

- Each term is a component (attribute) of the vector
- The value of each component is the number of times the corresponding term occurs in the document.

	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

Types of Data Sets

1.3 Transaction or Market Basket Data

A special type of record data, where Each record (transaction) involves a set of items.

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

Types of Data Sets

2. Graph-Based Data

2.1 graph captures relationships among data objects
2.2 data objects themselves are represented as graphs.

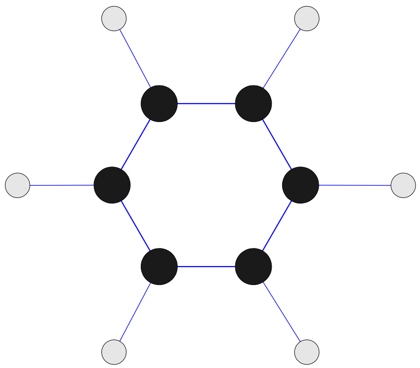
2.1 Data with Relationships among Objects

- relationships among objects frequently convey important information.
- data objects are mapped to nodes
- relationships among objects are captured by the links between objects

Types of Data Sets

2.2 Data with Objects That Are Graphs

objects contain subobjects that have relationships



Types of Data Sets

3. Ordered Data

the attributes have relationships that involve order in time or space

3.1 Sequential (Temporal) Data

each record has a time associated with it

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Types of Data Sets

3.2 Sequence Data

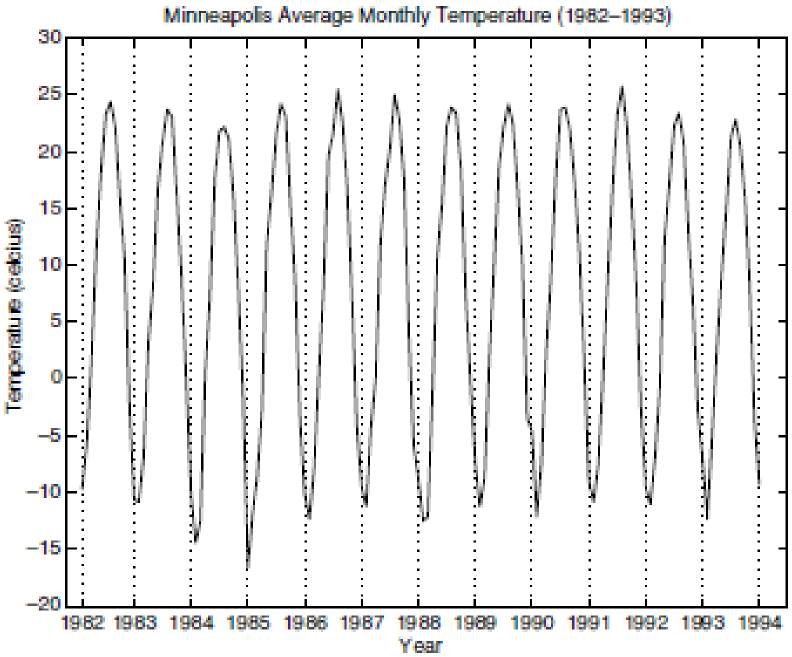
- sequence of words or letters
- no time stamps
- Positions in an ordered sequence.

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

Types of Data Sets

3.3 Time Series Data

- each record is a **time series**
- a series of measurements taken over time



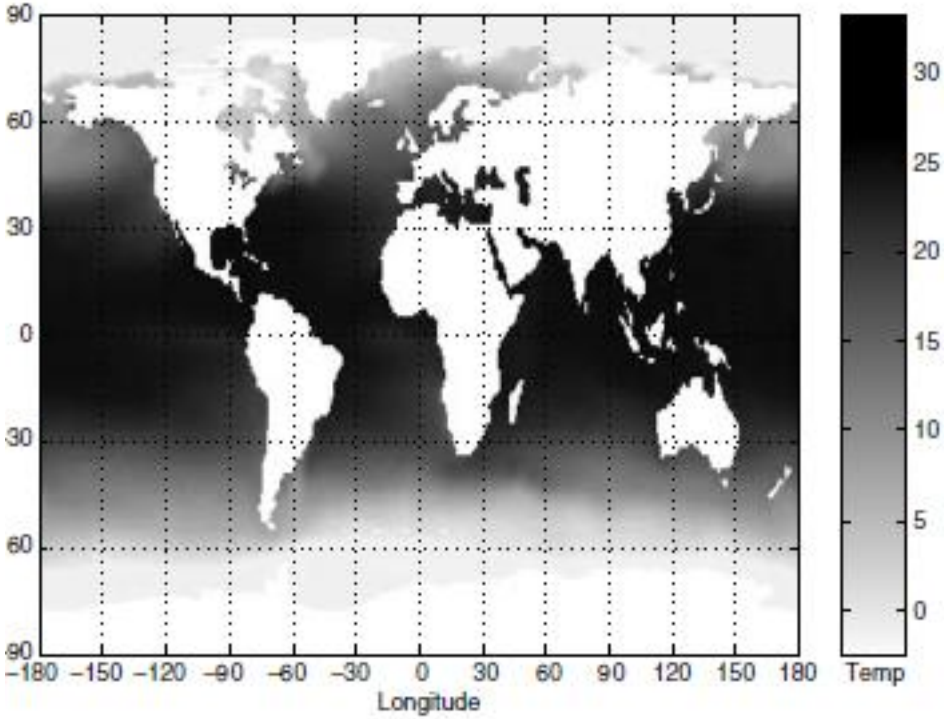
temporal autocorrelation

Types of Data Sets

3.4 Spatial Data

spatial autocorrelation

spatio-temporal data



Data Quality

Data quality

Poor data quality negatively affects many data processing efforts

(1) correction of data quality
(2) use of algorithms that can tolerate poor data quality

data cleaning

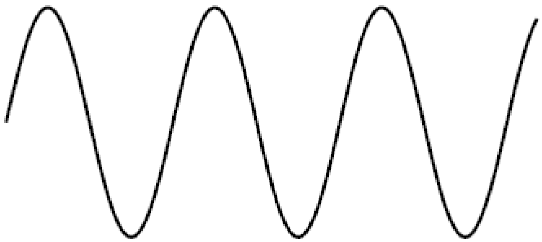
Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data
- Wrong data

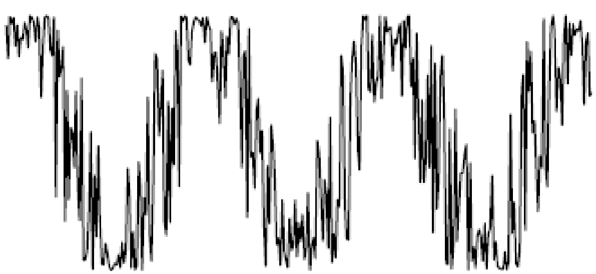
Noise

Noise is the random component of a measurement error

- ❖ For objects, noise is an extraneous object
- ❖ For attributes, noise refers to modification of original values



(a) Time series.



(b) Time series with noise.

(a) Three groups of points.

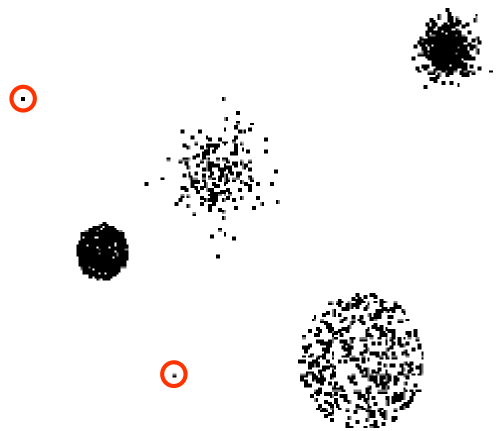
(b) With noise points (+) added.

signal processing can frequently be used to reduce noise

Outlier

(1) data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set
(2) values of an attribute that are unusual with respect to the typical values for that attribute.

anomalous objects or values



Outlier

Case 1: Outliers are noise that interferes with data analysis

Case 2: Outliers are the goal of our analysis

- Credit card fraud
- Intrusion detection

Missing Values

Reasons for missing values

- Information is not collected (e.g., people decline to give their age and weight)
- Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

Handling missing values

- Eliminate data objects or variables
- Estimate missing values
  - Example: time series of temperature
  - Example: similar data points
- Ignore the missing value during analysis

Data Preprocessing

Data Preprocessing

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

Aggregation

Combining two or more attributes (or objects) into a single attribute (or object)

- Data reduction
  - Reduce the number of attributes or objects
- Change of scale
  - Cities aggregated into regions, states, countries, etc.
  - Days aggregated into weeks, months, or years
- More “stable” data
  - Aggregated data tends to have less variability

Sampling

- Sampling is a commonly used approach for selecting a subset of the data objects to be analyzed
- Processing the entire set of data of interest is too expensive or time consuming.

Effective Sampling

- ✓ Using a sample will work almost as well as using the entire data set, if the sample is representative
- ✓ A sample is representative if it has approximately the same properties (of interest) as the original set of data

Simple Random Sampling

There is an equal probability of selecting any particular item

Sampling Method & size?

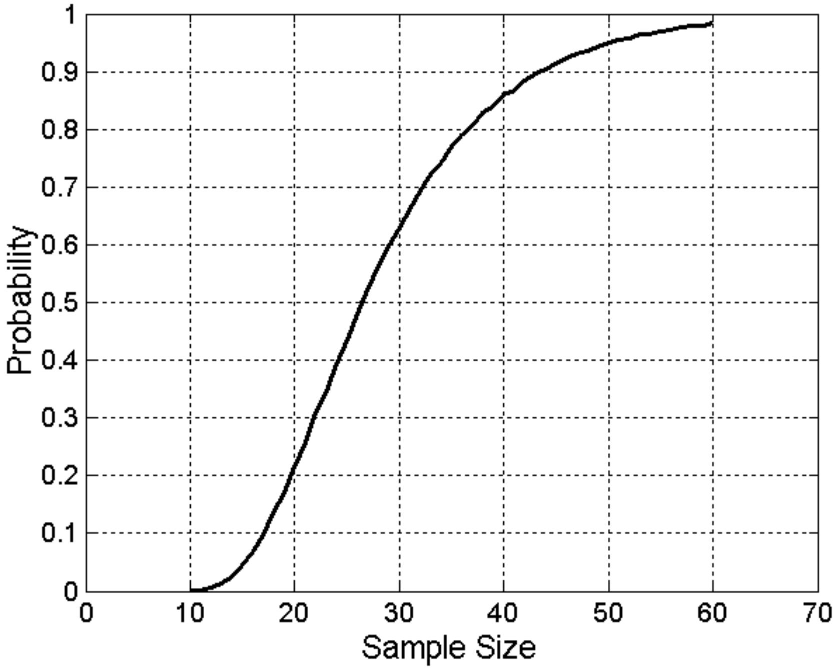
(a) 8000 points (b) 2000 points (c) 500 points

Sampling

- ❖ Sampling without replacement
  - ✓ As each item is selected, it is removed from the population
- ❖ Sampling with replacement
  - ✓ Objects are not removed from the population as they are selected for the sample.
  - ✓ In sampling with replacement, the same object can be picked up more than once

Sampling

- What sample size is necessary to get at least one object from each of 10 equal-sized groups.



- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

Dimensionality Reduction

many data mining algorithms work better if the dimensionality is lower

- eliminate irrelevant features and reduce noise
- curse of dimensionality
- more easily visualized
- Reduce amount of time and memory required by data mining algorithms

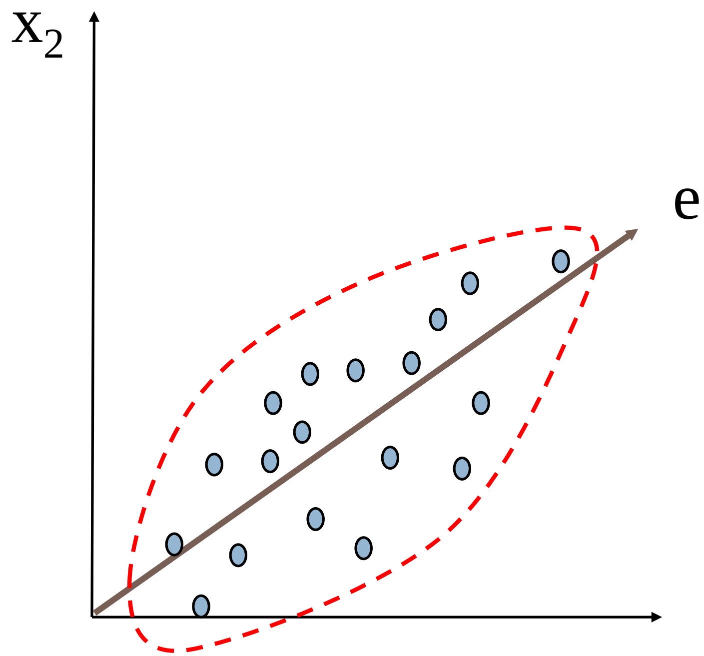
1) creating new features that are a combination of the old attributes
2) selecting features **feature subset selection** or **feature selection**

Curse of Dimensionality

- many types of data analysis become significantly harder as the dimensionality of the data increases
- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

Dimensionality Reduction: PCA

finds new features (principal components) that
(1) linear combinations of the original attributes
(2) are orthogonal to each other
(3) capture the maximum amount of variation
in the data



Feature Selection

Another way to reduce dimensionality of data

Redundant features

- Duplicate much or all of the information contained in one or more other attributes
- Example: purchase price of a product and the amount of sales tax paid

Irrelevant features

- Contain no information that is useful for the data mining task at hand
- Example: students' ID is often irrelevant to the task of predicting students' GPA

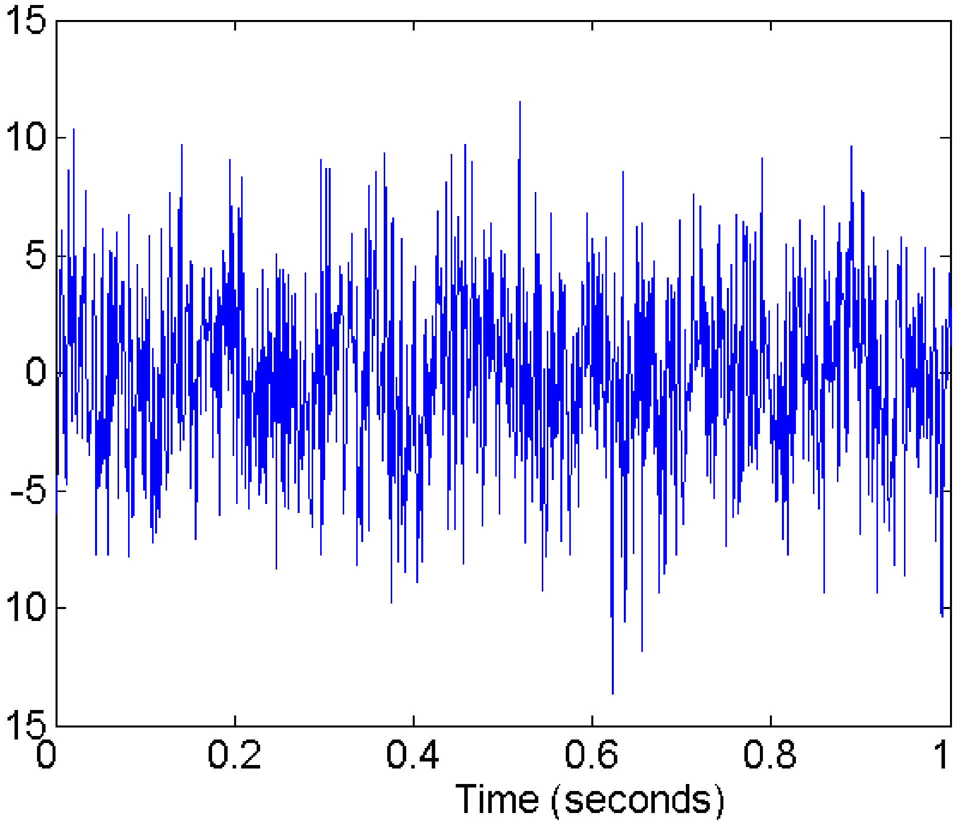
Feature Creation

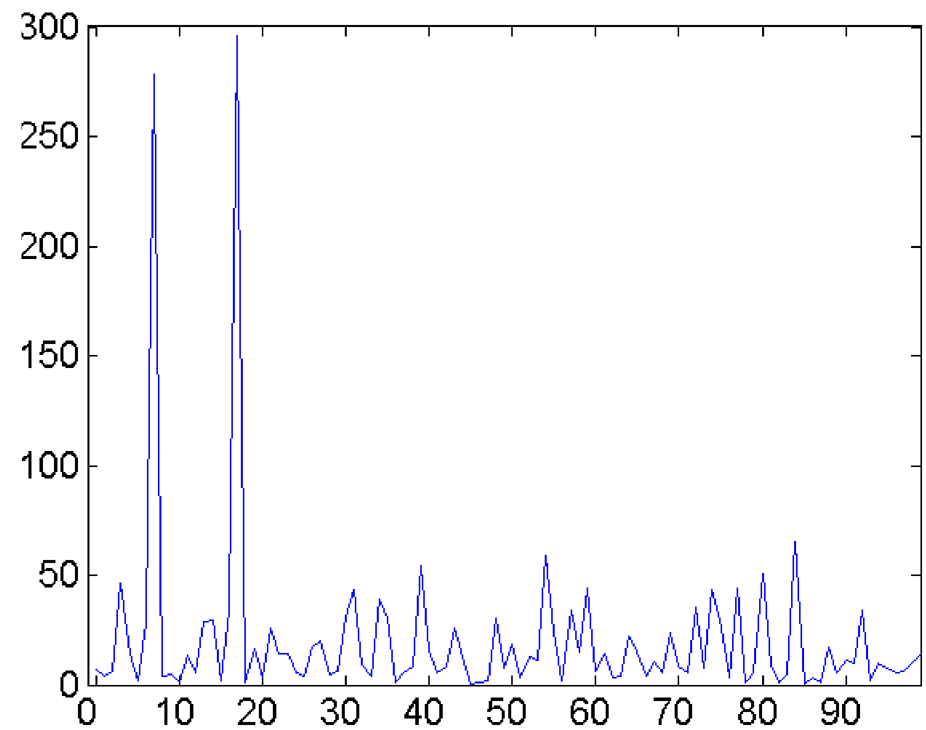
Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

Feature extraction/construction

Mapping data to new space : different view of the data

Fourier and wavelet transform





Two Sine Waves + Noise

Frequency

Discretization

Discretization is the process of converting a continuous attribute into an ordinal attribute

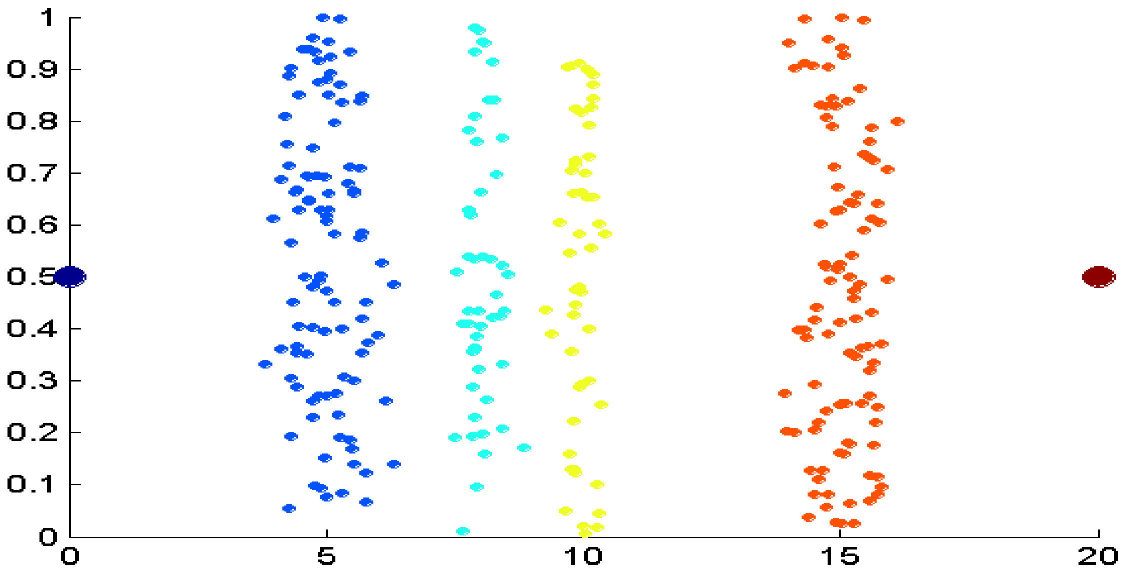
Binarization

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Discretization

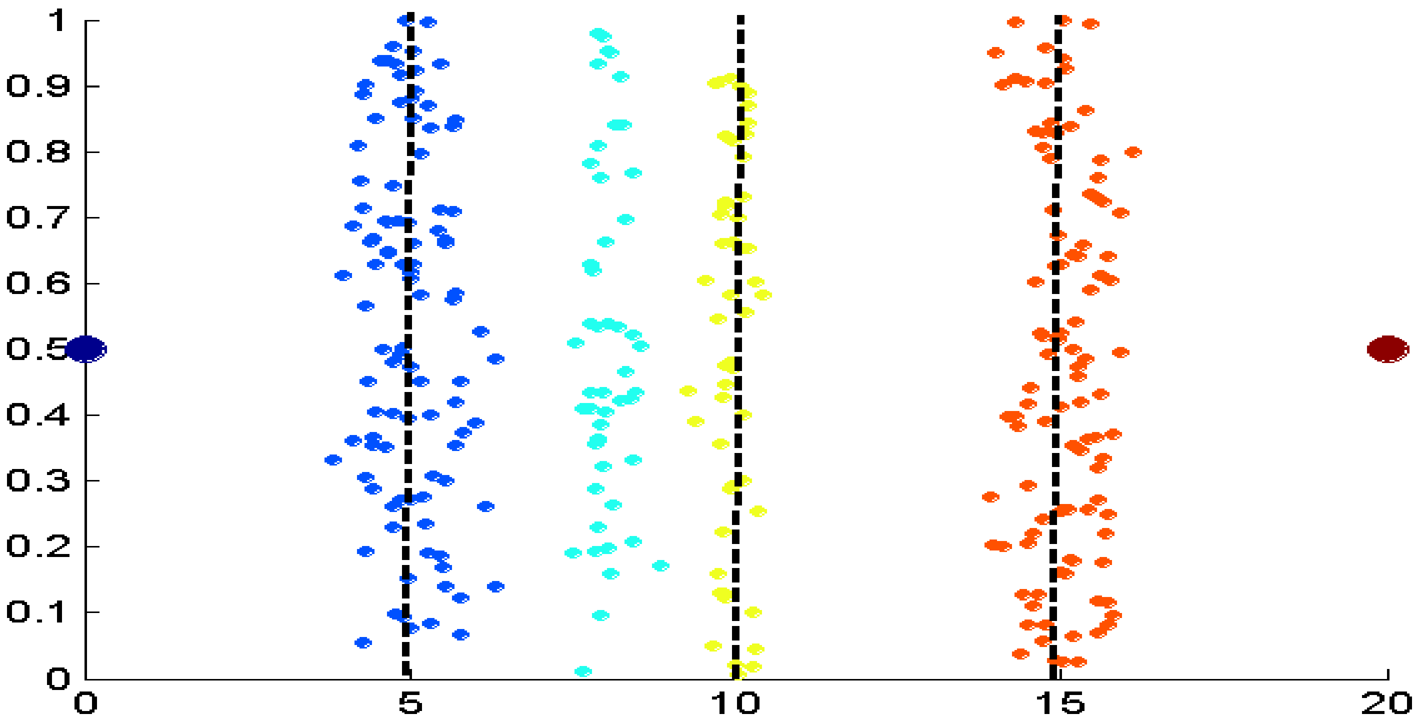
Unsupervised discretization: find breaks in the data values
Supervised discretization: Use class labels to find breaks

Example: Discretization Without Using Class Labels



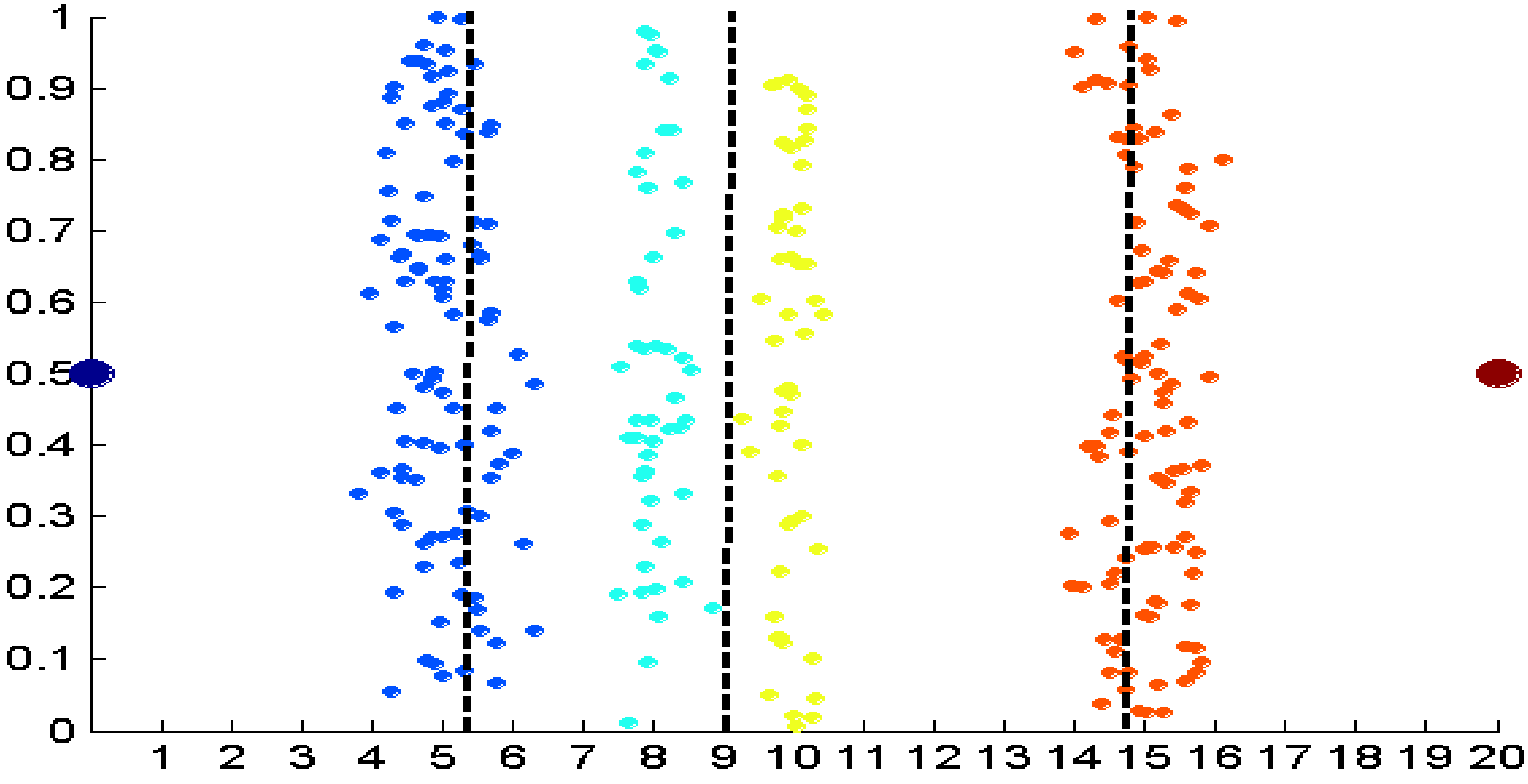
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap

Discretization



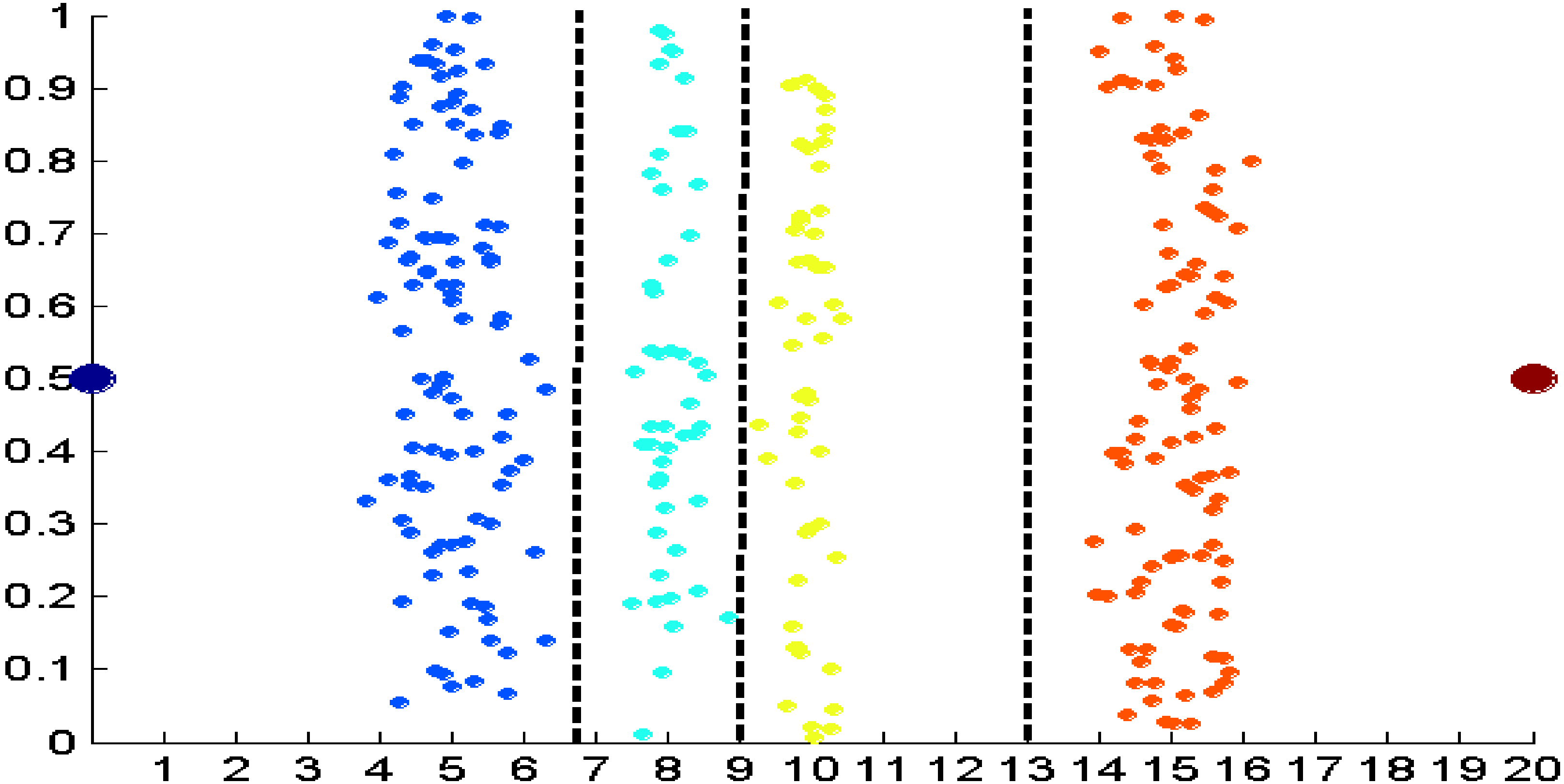
Equal interval width **approach used to obtain 4 values**

Discretization



Equal frequency approach used to obtain 4 values

Discretization



K-means approach to obtain 4 values.

Attribute Transformation

- An attribute transform is 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, $|x|$
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation