Chapter 2

Data

Terminology

Student ID	Year	GPA	…
1042129	Junior	3.85	…
1034262	Senior	3.24	…
1052663	Sophomore	3.51	…
1082246	Freshman	3.62	…

Attribute

Data object

Data set

- **Data object**: an entity with measurable properties
  - Also called *record*, *point*, *vector*, *case*, *sample*, *instance*, *observation*, …
- **Attribute**: a property or characteristic of a data object
  - Also called *variable*, *field*, *feature*, *dimension*, …
- **Data set**: a collection of data objects
  - Commonly stored in flat files or database tables

Data-Related Issues for Data Mining (1/2)

1. The types of data

- The attributes can be of ***different*** types
  - (ex) categorical (city, gender, genre, …), numeric (temperature, age, price, …)
- Data sets often have ***different*** characteristics
  - (ex) record data, graph data (social network), ordered data (time series), …
- <u>The type of data determines which methods and techniques can be used</u>

2. The quality of the data

- Data is often ***far*** from perfect
  - (ex) noise, outliers, missing data, inconsistent data, duplicate data
  - (ex) biased or unrepresentative data
- Understanding and improving data quality typically ***improves*** the quality of the resulting analysis

Data-Related Issues for Data Mining (2/2)

3. Preprocessing
   - Often, the raw data must be processed to make it ***suitable*** for analysis
     - (ex) continuous attribute (e.g., length) → categorical attribute (e.g., S/M/L)
     - (ex) dimensionality reduction (e.g., 100 attributes → 10 attributes)
   - The goal is to modify the data so that it ***better fits*** a specific technique

4. Measures of similarity
   - Data mining tasks often need to measure the ***similarity*** between objects
     - (ex) clustering, classification, or anomaly detection
   - There are ***many*** similarity or distance measures
     - The proper choice depends on the type of data and the particular application

Types of Data

1. Types of Attributes

- ***Categorical*** (qualitative) attribute
  - An attribute that can take on one of a ***limited*** number of possible values
    - (ex) zip code, student ID, city
  - Lacks most of the properties of numbers and should be treated as ***symbols***
    - (ex) ‘Junior’ + ‘Senior’ (X)
  - However, the values may have an ***order*** relationship (e.g., ‘S’ < ‘M’ < ‘L’)

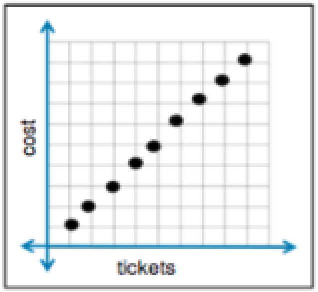
- ***Numeric*** (quantitative) attribute
  - An attribute whose value can be ***any*** number from a defined range
    - (ex) temperature, age, mass, length, counts
  - Has most of the properties of numbers (e.g., 35.1°C < 40.2°C (O))
  - Associated with a ***measurement scale*** (e.g., °C, °F, cm, kg, GB)

Different Attribute Types

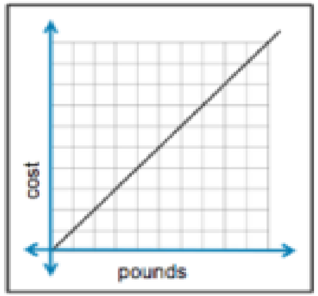
Attribute Type		Description	Examples
Categorical (qualitative)	**Nominal**	The values are just different names ($=$, $\neq$)	zip codes, employee IDs, eye color, gender
	Ordinal	The values provide enough information to order objects ($<$, $>$)	hardness of minerals, *{good, better, best}*, grades, street numbers
Numeric (quantitative)		The values are represented by numbers (e.g., real numbers, integers) ($+$, $-$, $\times$, $/$)	temperature, monetary quantities, counts, age, mass, length, electrical current

Another Way to Distinguish Attributes

- ***Discrete*** attribute
  - Has a ***finite*** or ***countably infinite*** set of values (e.g., 1, 2, 3, …)
  - Categorical (e.g., zip codes) or numeric (e.g., counts)
  - Often represented using ***integer*** variables
  - Binary attribute: a special case with only two values
    - (ex) true/false, yes/no, 0/1



- ***Continuous*** attribute
  - One whose values are real numbers (i.e., can take ***any*** value)
    - (ex) temperature, height, weight
  - Typically represented as ***floating-point*** variables



2. Types of Data Sets

- There are ***many*** types of data sets
  - As the field of data mining develops and matures, a greater variety of data sets become available for analysis

- We focus on some of the most common types:

  (1) Record data

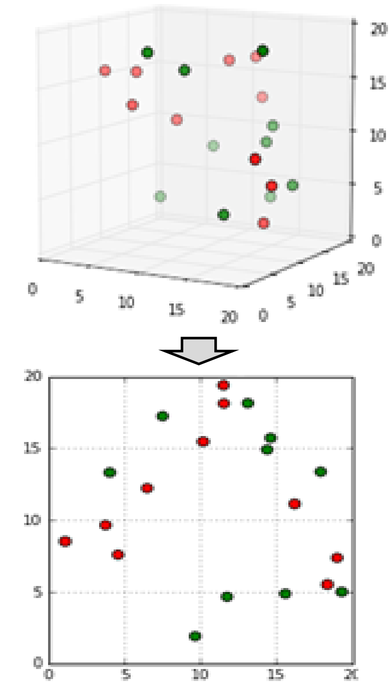
  (2) Graph-based data

  (3) Ordered data

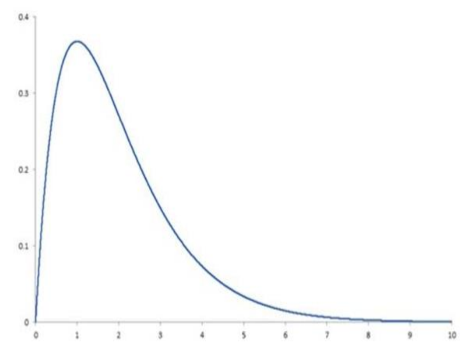
- However, these categories do not cover all possibilities and other types are certainly possible

General Characteristics of Data Sets

- Dimensionality
  - The number of attributes in the data set
  - ***The curse of dimensionality***
    - The difficulties associated with high-dimensional data
    - Because of this, ***dimensionality reduction*** is often used



- Distribution
  - The frequency of various values for the attributes
    - (ex) Gaussian (normal) distribution
  - However, many data sets have distributions that are ***not*** well captured by standard statistical distributions
  - ***Skewness*** in the distribution can make mining difficult
    - (ex) Male : Female = 5 : 95



(1) Record Data

- The data set is a collection of ***records*** (data objects)
  - Each record consists of a fixed set of fields (attributes)
- There is ***no*** explicit relationship among records or fields
- Usually stored either in flat files or in relational databases
  - However, data mining often does ***not*** use any of the additional information available in a relational database
  - Rather, the database serves as a convenient place to find records

ID	artistName	albumTitle	genre	releaseDate	rating	length	label
1	Bach, J.S	6 Favorite Cantatas	Classical	14-Oct-07	9.5	75:15	L'Oiseau Lyre
2	Rush	Moving Pictures [Remastered]	Rock	03-Jun-97	9.75	45:32	Mercury
3	Wild Pink Puppies	Tales from Beyond	Punk	15-May-03	3	32:15	Orange Goblin
4	Mr Mister	Welcome to the Real World [Re-Release]	Rock	08-Jun-11	8.5	74:43	RCA
5	Anwynn	Epic	Gothic	03-Apr-09	7.75	65:54	Relativity
6	Novembre	Blue	Rock	23-Jan-11	8.5	56:55	Azure Records

(Ex) Record Data

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
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

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Soda, Diapers, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	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

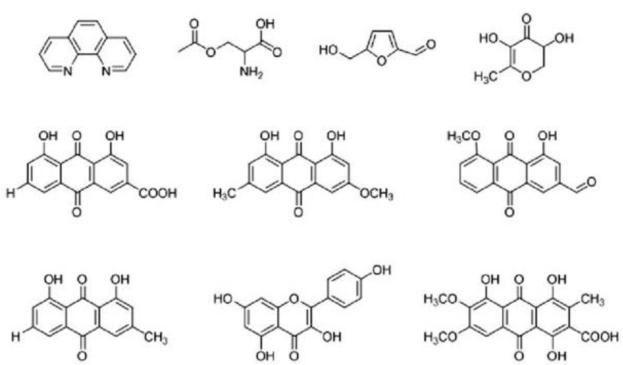
(d) Document-term matrix.

(2) Graph-Based Data

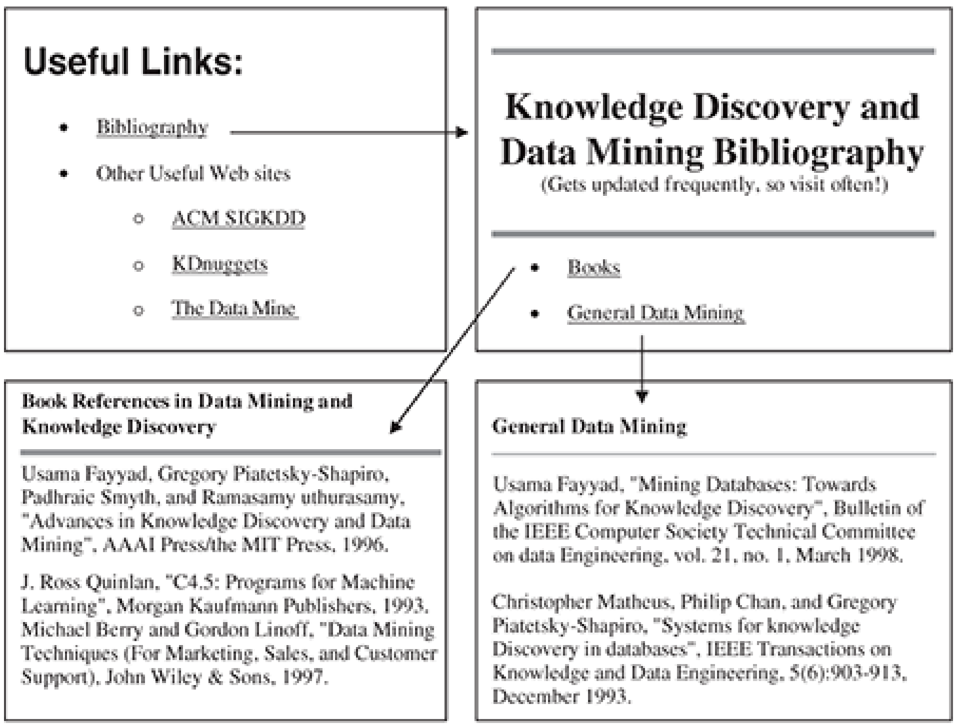
- The data is represented as one or more ***graphs***

- **(Case 1)** Data with relationships among objects
  - The graph captures relationships among data objects
    - Nodes: data objects
    - Links: the relationships among objects
  - (ex) World Wide Web, social networks

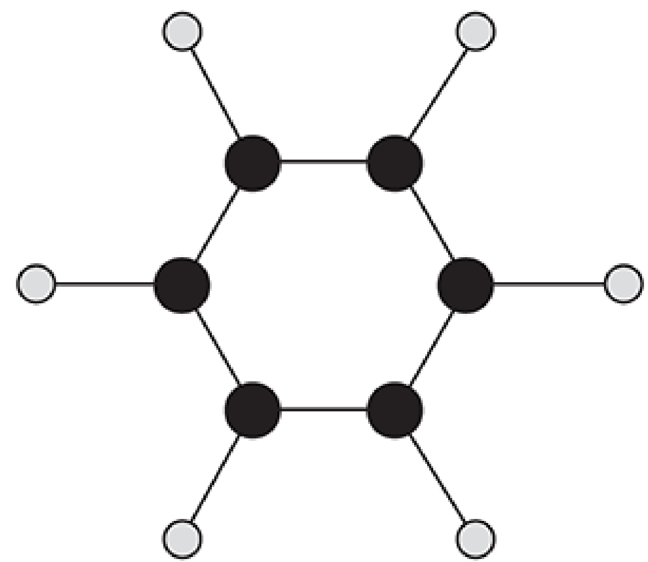
- **(Case 2)** Data with objects that are graphs
  - Each data object is represented as a graph
  - (ex) chemical compounds



(Ex) Graph-Based Data



(a) Linked web pages.



(b) Benzene molecule.

(3) Ordered Data (1/2)

- The attribute values have ***order*** relationships in time or space

- **(Case 1)** Sequential transaction data
  - Each transaction has a timestamp associated with it
  - It is possible to find sequential patterns
    - (ex) people who buy DVD players tend to buy DVDs
  - (ex) retail transaction data, purchase history

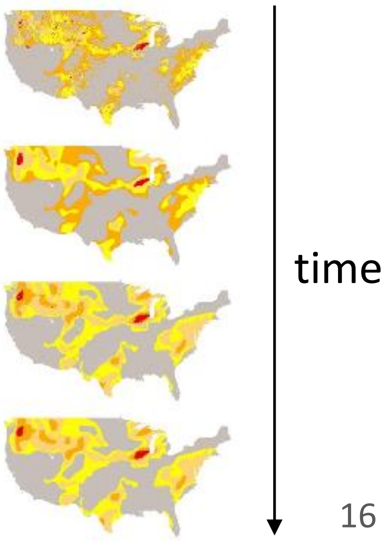
TID	Date	Items Purchased
101	01/01/2001	Cheese, Wine, Bread
102	01/02/2001	Bread, Water, Milk
103	01/03/2001	Milk, Cheese, Magazine
104	01/03/2001	Cheese, Wine, Bread, Milk
105	01/04/2001	Milk, Bread

- **(Case 2)** Time series data
  - Each record is a time series (i.e., a series of measurements taken over time)
  - It is important to consider ***temporal*** autocorrelation
    - i.e., two values close in time are often very similar
  - (ex) the daily prices of stocks, temperature data

(3) Ordered Data (2/2)

- **(Case 3)** Sequence data
  - A data set is a sequence of individual entities
  - There are ***no*** time stamps
    - Instead, there are positions in a sequence
  - Many problems involve finding similar sequences
  - (ex) sequences of words, genetic sequence data

- **(Case 4)** Spatial and spatio-temporal data
  - The data consists of time series at various locations
  - A more complete analysis requires consideration of both the spatial and temporal aspects of the data
  - It is important to consider ***spatial*** autocorrelation
  - (ex) Earth science data sets, gas flow simulation data



(Ex) Ordered Data

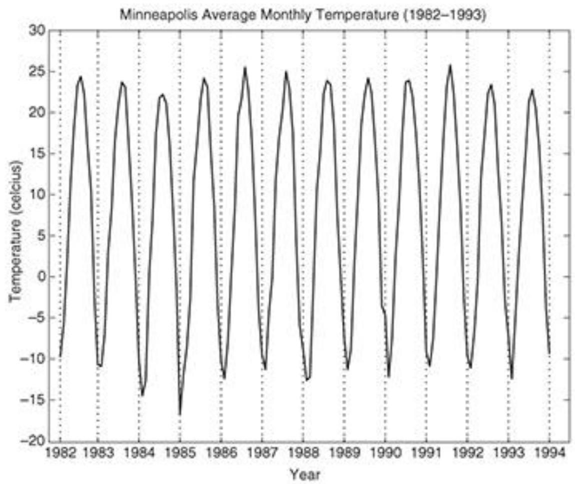
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

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)

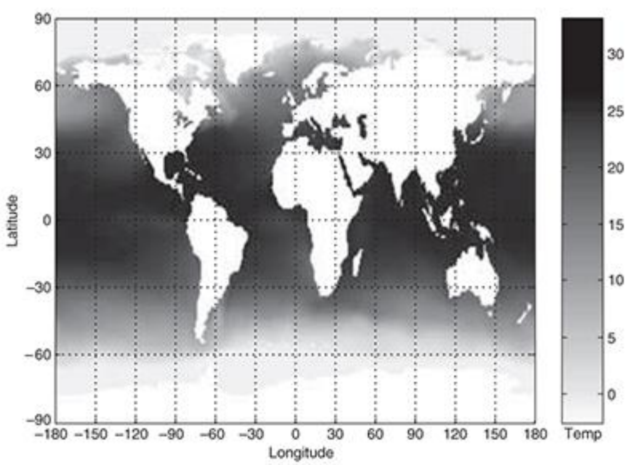
(a) Sequential transaction data.

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

(b) Genomic sequence data.



(c) Temperature time series.



(d) Spatial temperature data.

Data Quality

Data Quality

- It is ***unrealistic*** to expect that data will be perfect
  - Human error
  - Limitations of measuring devices
  - Flaws in the data collection process, etc.

- Examples: data quality problems
  - Values or even entire data objects can be missing
  - Spurious or duplicate objects (e.g., multiple records for a single person)
  - Inconsistencies (e.g., a person has a height of $2\ \mathrm{m}$, but weights only $2\ \mathrm{kg}$)

- To prevent data quality problems, data mining focuses on
  ① The detection and correction of data quality problems → ***data cleaning***
  ② The use of algorithms that can tolerate poor data quality

Measurement and Data Collection Errors

- Measurement error
  - Any problem resulting from the measurement process
    - (ex) the numerical difference of the measured and true value (i.e., error)
- Data collection error
  - Errors such as omitting data objects or attribute values, or inappropriately including a data object
    - (ex) including similar but unrelated data objects

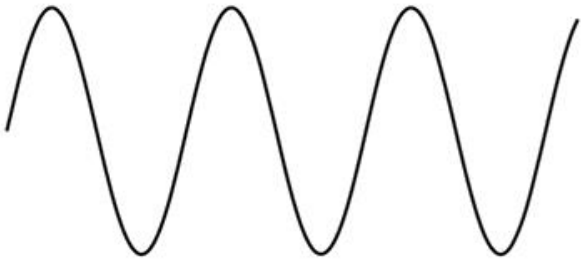
? ? ?

ID	Last Name	First Name	Street	City	State	Zip	Phone	Fax	E-mail
113	Smith		123 S. Main	Denver	CO	80210	(303) 777-1258	(303) 777-5544	ssmith@aol.com
114	Jones	Jeff	12A	Denver	CO	80224	(303) 666-6868	(303) 666-6868	(303) 666-6868
115	Roberts	Jenny	1244 Colfax	Denver	CO	85231	759-5654	853-6584	jr@msn.com
116	Robert	Jenny	1244 Colfax	Denver	CO	85231	759-5654	853-6584	jr@msn.com

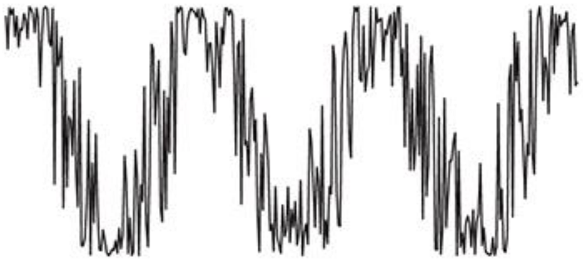
? ? ?

Noise and Artifacts

- Noise
  - The ***random*** component of a measurement error
    - Typically involves the distortion of a value or the addition of spurious values
  - Because its elimination is difficult, much work focuses on ***robust algorithms***
    - They produce acceptable results even when noise is present

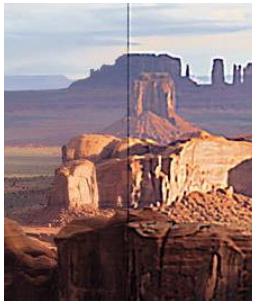


(a) Time series.



(b) Time series with noise.

- Artifacts
  - ***Deterministic*** distortions of data
    - (ex) a streak in the same place on a set of photographs

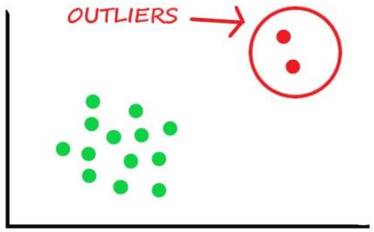


Outliers

- Data objects that have characteristics that are ***different*** from most of the other data objects in the data set

- Or, values that are ***unusual*** with respect to the typical values

- Also referred to as ***anomalous*** objects or values



- Many different definitions have been proposed by the statisticians and data mining communities

- It is important to distinguish between noise and outliers
  - Outliers can be ***legitimate*** data objects or values

Missing Values

- The information was not collected or not applicable
- Several strategies for dealing with missing data
  - Eliminate data objects or attributes
  - Estimate missing values (e.g., average, interpolation)
  - Ignore the missing value during analysis

181 cm

Name	Height	Weight
Lee	172 cm	65 kg
Kim	X	74 kg
Park	182 cm	75 kg
Seo	169 cm	62 kg
Jeong	178 cm	69 kg

Inconsistent Values

- Values that ***violate*** given consistency constraints

- Examples
  - Different zip codes for the same area
  - A person's height is negative
  - Nonexistent name
  - 6-digit telephone number

Name	City	Tel
Lee	Seoul	031-710-4112
Kim	Daejeon	042-270-4615
Park	Busan	051-200-1679

- It is important to detect and, if possible, correct such problems
  - The correction may require ***additional*** or ***external*** information

Duplicate Data

- Data objects that are ***duplicates*** of one another

- Two main issues
  1. If there are two objects that actually represent a single object, then it is important to resolve ***inconsistent*** values
  2. Care needs to be taken to avoid accidentally combining data objects that are similar, but ***not*** duplicates (e.g., two people with identical names)

- Deduplication
  – The process of dealing with these issues

	A	B	C
1	Name	Gender	Age
2	ABC	Male	25
3	DEF	Male	28
4	GHI	Female	27
5	JKL	Female	22
6	MNO	Female	31
7	PQR	Male	30
8	STU	Male	24
9	XYZ	Female	19
10	JKL	Female	35
11	BCD	Male	32
12	RST	Male	18
13	VWX	Female	21

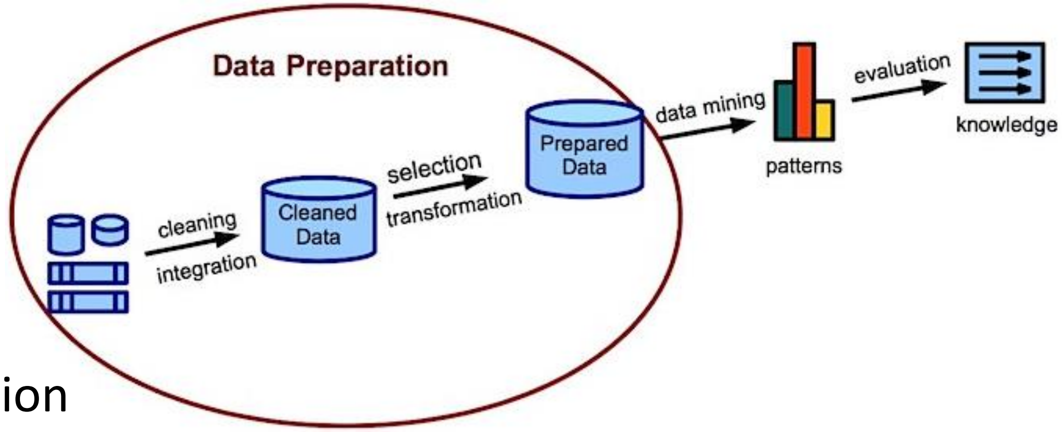
Data Preprocessing

Data Preprocessing

- Additional steps to make the data ***more suitable*** for data mining

- A broad area and consists of a number of different strategies and techniques that are interrelated in complex ways

- We will discuss the following topics:
  - Aggregation
  - Sampling
  - Dimensionality reduction
  - Feature selection
  - Feature creation
  - Discretization and binarization
  - Variable transformation



1. Aggregation

- Combine two or more objects into a single object
  - Because sometimes "***less is more***"

- Example: customer purchase data set
  ① Replace all the transactions of a single store location with a single object
  ② Reduce the possible values for *Date* from 365 days to 12 months

Transaction ID	Item	Store Location	Date	Price	…
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	…
101123	Battery	Chicago	09/06/04	$5.99	…
101124	Shoes	Minneapolis	09/06/04	$75.00	…

①

Items	Store Location	Total Price	…
:	:	:	:
Watch, Battery, …	Chicago	$428.98	…
Shoes, …	Minneapolis	$195.02	…

②

Items	Date	Total Price	…
:	:	:	:
Watch, Battery, Shoes, …	09/06	$1523.75	…

Motivations for Aggregation

1. The smaller data sets require ***less*** memory and processing time
   – Hence, it enables the use of more expensive data mining algorithms

2. Aggregation can provide a ***high-level*** view of the data
   – (ex) each store's sales → each location's sales

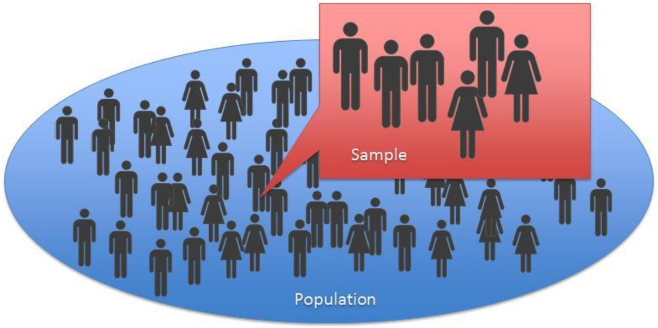
3. The behavior of groups of objects is often ***more stable*** than that of individual objects
   – (ex) hourly temperature → daily temperature (on average)

- **Disadvantage**: the potential loss of interesting details
  – (ex) aggregating over months → which day has the highest sales?

2. Sampling

- Select a ***subset*** of the data objects to be analyzed

- Motivations for sampling
  - Statisticians: obtaining the entire data set is too expensive
  - Data miner: ***processing*** the entire data set is too expensive
    - In terms of memory or processing time



- Key principle for effective sampling
  - Use a ***representative*** sample
    - It should have approximately the same property as the original data set
    - (ex) the mean of a sample ≈ the mean of the original data set

Sampling Approaches

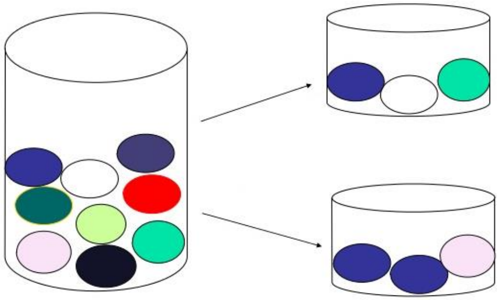
① Simple random sampling

– There is an equal probability of selecting any particular object

– Two variations on random sampling

  • Sampling without replacement

  • Sampling with replacement



② Stratified sampling

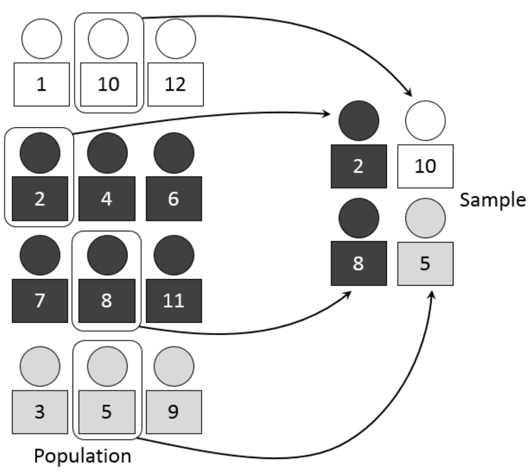
– When the population consists of different types of objects, simple random sampling can fail

  • (ex) A: 10000, B: 10 → A: 100, B: **0**

– Select objects from ***each*** group

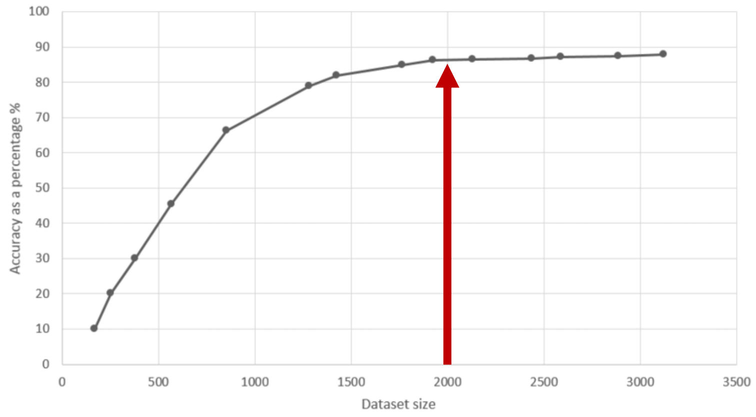
  • Equal numbers of objects

  • The number proportional to the size of that group



Progressive (or Adaptive) Sampling

- Used when the proper sample size can be ***difficult*** to determine
- Basic technique
  - Starts with a small sample
  - Increase the sample size until a sample of sufficient size has been obtained
- Important point
  - There must be a way to evaluate the sample to judge if it is large enough
    - (ex) Stop increasing the sample size if the increase in accuracy ***levels off***



3. Dimensionality Reduction

- The process of ***reducing*** the number of attributes in the data set
  - Dimensionality = the number of attributes

	sepal length	sepal width	petal length	petal width
0	-0.900681	1.032057	-1.341272	-1.312977
1	-1.143017	-0.124958	-1.341272	-1.312977
2	-1.385353	0.337848	-1.398138	-1.312977
3	-1.506521	0.106445	-1.284407	-1.312977
4	-1.021849	1.263460	-1.341272	-1.312977

PCA
(2 components)

	principal component 1	princial component 2
0	-2.264542	0.505704
1	-2.086426	-0.655405
2	-2.367950	-0.318477
3	-2.304197	-0.575368
4	-2.388777	0.674767

- Key benefits

  ① Many data mining algorithms work ***better*** if the dimensionality is lower
  - Partly because irrelevant features are eliminated and noise is reduced
  - Partly because the ***curse of dimensionality***

  ② A more understandable model can be obtained
  - Because the model involves fewer attributes
  - (ex) $y = x_1 + 5.1x_2 + 4.2x_3 + 8.7x_4 + 7.4x_5 + 2.9x_6 + 10x_7 \rightarrow y = z_1 + 7.2z_2$

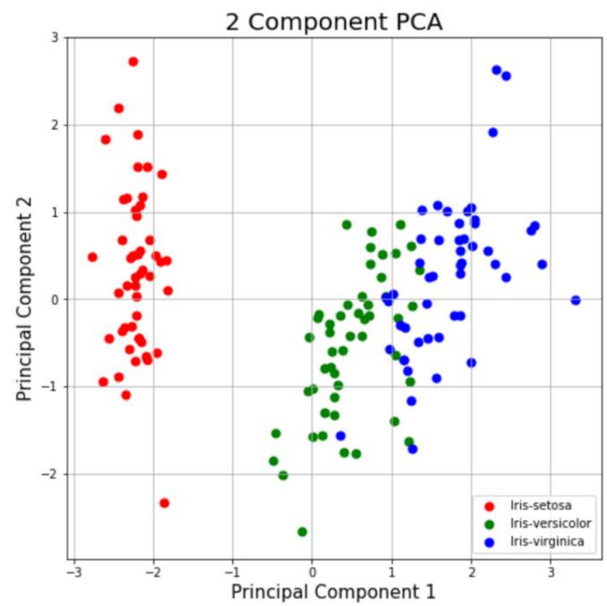
3. Dimensionality Reduction

- Key benefits (cont'd)
  - ③ The data can be more easily visualized
    - Because the data can be reduced to two or three dimensions
  - ④ The amount of time and memory required by the algorithm is reduced
    - Because the size of the data is reduced

	sepal length	sepal width	petal length	petal width
0	-0.900681	1.032057	-1.341272	-1.312977
1	-1.143017	-0.124958	-1.341272	-1.312977
2	-1.385353	0.337848	-1.398138	-1.312977
3	-1.506521	0.106445	-1.284407	-1.312977
4	-1.021849	1.263460	-1.341272	-1.312977

PCA
(2 components)

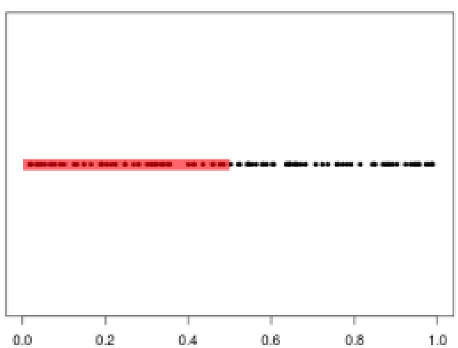
	principal component 1	princial component 2
0	-2.264542	0.505704
1	-2.086426	-0.655405
2	-2.367950	-0.318477
3	-2.304197	-0.575368
4	-2.388777	0.674767



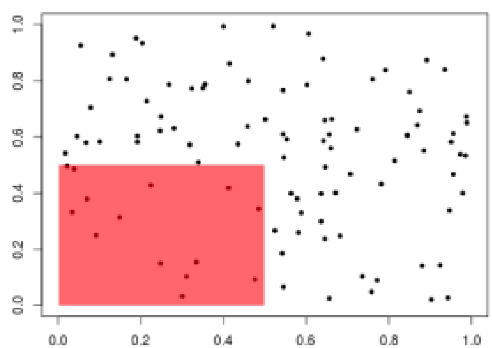
The Curse of Dimensionality (1/2)

- The phenomenon that data analysis becomes ***significantly harder*** as the dimensionality of the data increases

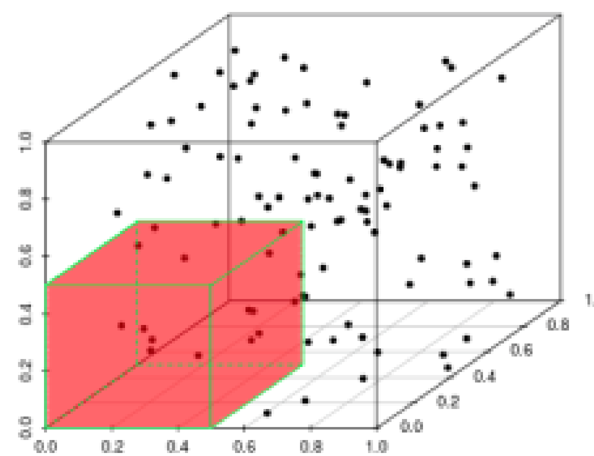
- Because, as dimensionality increases, the data becomes increasingly ***sparse*** in the space
  - Also the distances between objects become very ***large***
  - Eventually the distances between objects become almost ***the same***



1D (50% of data)



2D (25% of data)

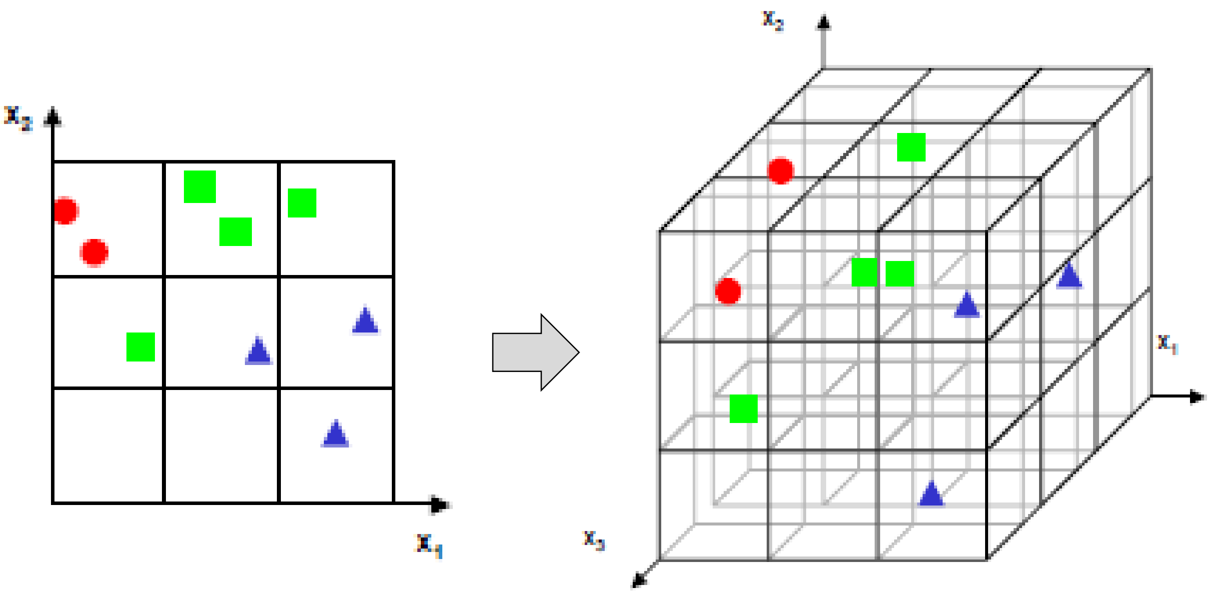


3D (12.5% of data)

The Curse of Dimensionality (2/2)

- Problem for ***classification***
  - There are ***not enough*** data objects to allow the creation of a model that reliably assigns a class to all possible objects

- Problem for ***clustering***
  - The distance between objects become ***less meaningful***



1. The amount of training data needed to cover 80% of the space grows exponentially
2. Nearly all objects become far from each other

4. Feature Selection

- Another way to reduce the dimensionality is to use only a ***subset*** of the features
  - We would not lose information if ***redundant*** and ***irrelevant*** features are present

- Redundant features
  - Duplicate much or all of the information contained in other attributes
  - (ex) the price of a product $\leftrightarrow$ the amount of sales tax

- Irrelevant features
  - Contain almost no useful information for the data mining task
  - (ex) 'student ID' for the task of predicting students' GPA

Approaches to Feature Selection

1. Use common sense or domain knowledge

2. Embedded approaches
   - The data mining algorithm ***itself*** decides which attributes to use
   - (ex) decision trees

3. Filter approaches
   - Features are selected ***before*** the data mining algorithm is run
   - (ex) select attributes whose pairwise correlation is as low as possible

4. Wrapper approaches
   - Use the target data mining algorithm as a black box to find the ***best*** subset of attributes
   - (ex) add attributes one by one as long as the performance improves

Feature Weighting

- An alternative to keeping or eliminating features
  - Assign more important features a ***higher*** weight, while giving less important features a ***lower*** weight

- Two approaches
  - Use domain knowledge about the relative importance of features
  - The data mining algorithm determines the weights automatically

- (ex) support vector machine (SVM)
  - Produces classification models in which each feature is given a weight
  - (ex) $y = 100x_1 + 0.01x_2 + 20x_3 + 4$
    - x_1 is the most important feature, while x_2 is the least important feature

5. Feature Creation

- It is frequently possible to create, from the original attributes, ***new*** attributes
  - That captures the important information in a data set much more ***effectively***

transaction_ID	user_home_country	transaction_country
01	US	US
02	Canada	Canada
03	Canada	Spain
04	US	US
05	US	Japan

➡

transaction_ID	user_home_country	transaction_country	in_foreign_country
01	US	US	False
02	Canada	Canada	False
03	Canada	Spain	True
04	US	US	False
05	US	Japan	True

- Two related methodologies
  ① Feature extraction
  ② Mapping the data to a new space

Feature Extraction

- The creation of a new set of features from the original raw data

- Example
  - We want to classify historical artifacts with respect to their ***materials***
    - (ex) wood, clay, bronze, gold
  - In this case, a ***density*** feature constructed from the mass and volume features would most directly yield an accurate classification

Artifact	Mass	Volume

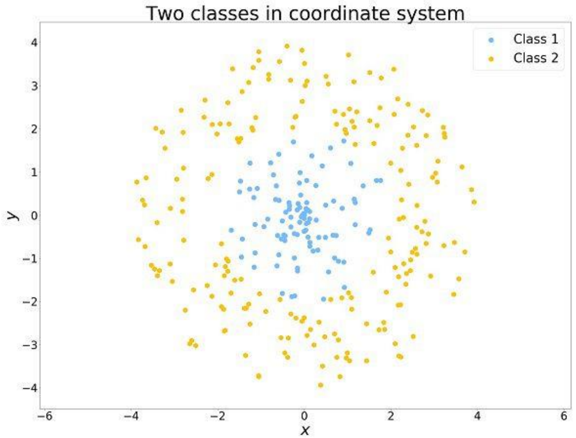
Artifact	Mass	Volume	Density (Mass/Volume)

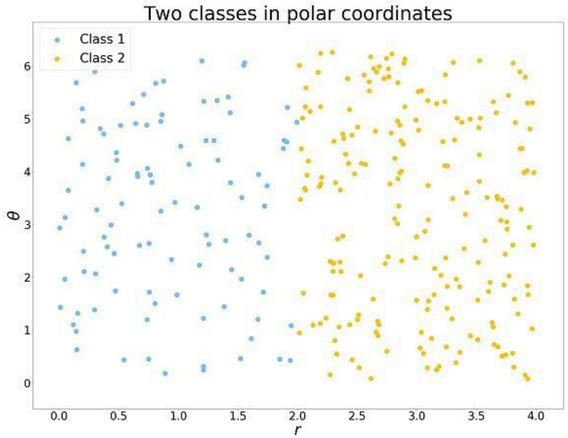
- Unfortunately, the most common approach is to use domain expertise

Mapping the Data to a New Space

- A totally different view of the data can reveal important and interesting features

- Example
  - The following points represented in the Euclidean space (x, y) are difficult for decision trees to classify
  - However, if we represent the points in the polar coordinate system (r, θ), it is easy for decision trees to classify the points





6. Discretization and Binarization

- Discretization
  - Transform a ***continuous*** attribute into a ***categorical*** attribute

Humidity
85.1
78.2
62.6

Humidity
High
Normal
Low

- Binarization
  - Transform an attribute into one or more ***binary*** attributes

Name	Gender	Age
Lee	Male	24
Kim	Female	17

Name	Male	Female	Age
Lee	1	0	24
Kim	0	1	17

- Why?
  - Some data mining algorithms require categorical or binary attributes
    - (ex) certain classification algorithms, association rule mining algorithms

Binarization

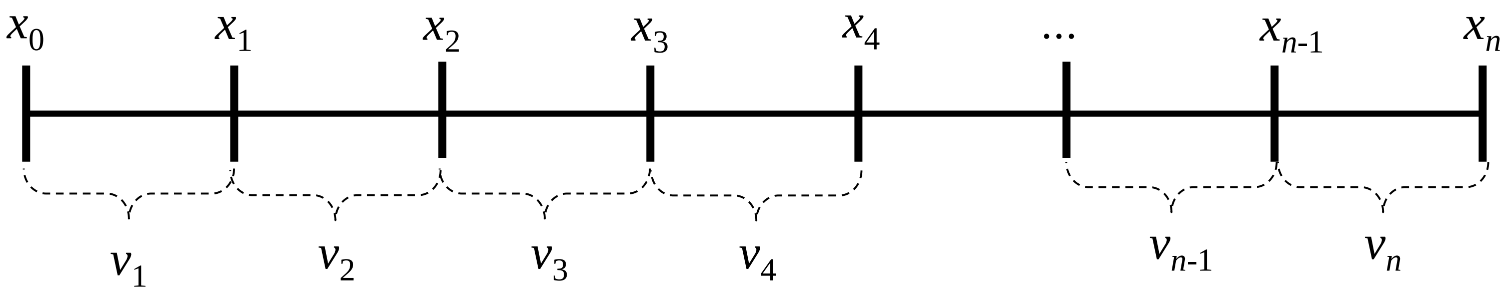
- Simple technique
  - Suppose there are m categorical values
  - Introduce one binary attribute for ***each*** categorical value
  - For each of the m binary attributes
    - Assign 1, if the binary attribute represents the categorical value of the object
    - Assign 0, otherwise

Categorical Value
awful
poor
OK
good
great

awful	poor	OK	good	great
x_1	x_2	x_3	x_4	x_5
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

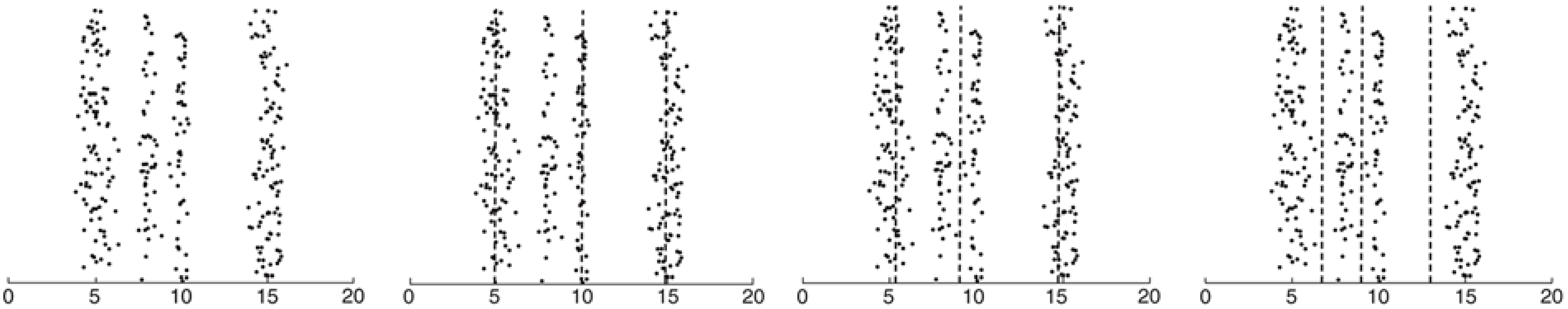
Discretization

- Basic steps
  - Decide how many categories, n, to have
  - Divide the values of the continuous attribute into n intervals
  - Map all the values in one interval to the same categorical value



- Several simple approaches
  - Equal width discretization
  - Equal frequency discretization
  - Clustering-based discretization (e.g., k-means)

Several Approaches to Discretization



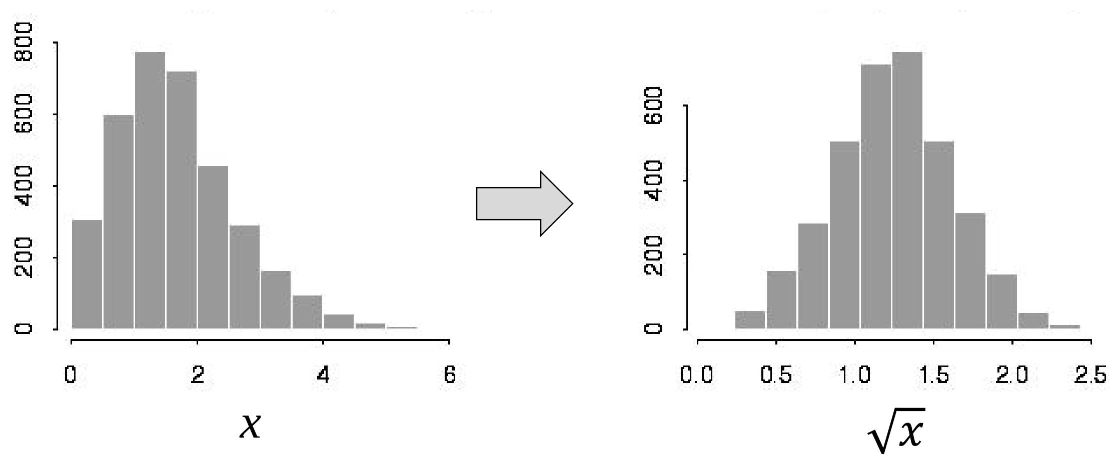
(a) Original data. (b) Equal width discretization. (c) Equal frequency discretization. (d) K-means discretization.

- Equal width discretization
  - Divide the range into a number of intervals each having the ***same width***
- Equal frequency discretization
  - Try to put the ***same number of objects*** into each interval
- Clustering-based discretization (e.g., k-means)
  - Find clusters of objects and divide the range according to the ***clusters***

7. Variable Transformation (1/2)

- Apply a transformation to all the values of a variable (attribute)

- **(Type 1)** Simple functions
  - Apply a simple mathematical function to each value individually
    - (ex) x^k, $\log x$, e^x, $\sqrt{x}$, $1/x$, $\sin x$, or $|x|$
  - Examples
    - $\log_{10} x$ is used when the range of values is very huge (e.g., 10^8, 10^9 → 8, 9)
    - $\sqrt{x}$, $\log x$, and $1/x$ are often used to transform data into a ***normal distribution***



7. Variable Transformation (2/2)

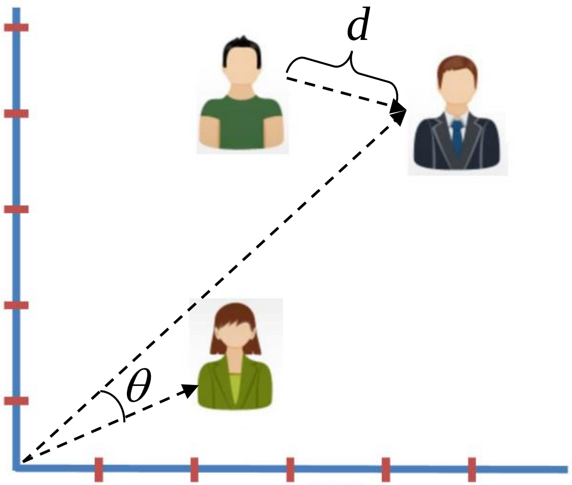
- **(Type 2)** Normalization or standardization
  - Make an entire set of values have a particular property
  - (ex) "Standardizing a variable" in statistics $x' = (x - \bar{x})/s_x$
    - $\bar{x}$: the mean of the attribute values
    - s_x: the standard deviation of the attribute values
    - Creates a new variable that has a ***mean of* 0** and a ***standard deviation of* 1**
  - Often used to avoid having a variable with large values dominate the results of analysis
    - (ex) Consider comparing people based on *age* and *income*
    - The comparison between people will be ***dominated*** by differences in *income*
      - (ex) $person_1 = (24, 25000)$, $person_2 = (67, 25050)$, $person_3 = (25, 25100)$

Measures of Similarity

Measures of Similarity and Dissimilarity

- Similarity and dissimilarity between objects are ***important***
  - Because they are used by a number of data mining techniques
  - (ex) clustering, nearest neighbor classification, and anomaly detection

- Proximity
  - Used to refer to either similarity or dissimilarity
  - There are many ***proximity measures*** for objects
    - Euclidean distance
    - Jaccard coefficient
    - Cosine similarity
    - …

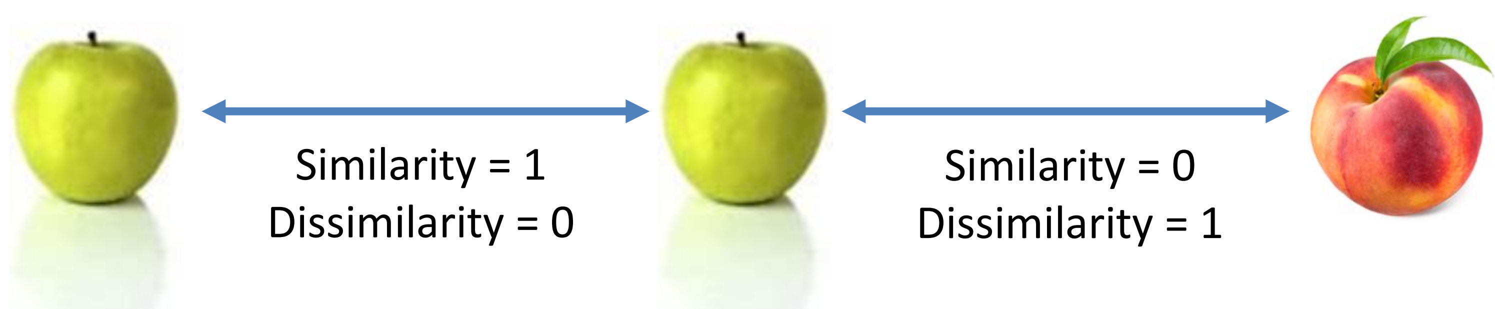


- We will discuss ***various proximity measures*** for objects

Definitions

- **Similarity**: a numerical measure of the degree to which two objects are ***alike***
  - The more alike, the ***higher*** the similarity is
    - (ex) 0 (no similarity) → 1 (complete similarity)

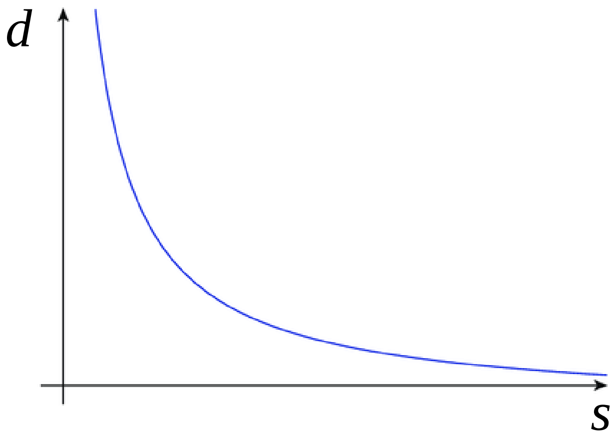
- **Dissimilarity**: a numerical measure of the degree to which two objects are ***different***
  - The more similar, the ***lower*** the dissimilarity is
    - (ex) 0 (complete similarity) → 1 or ∞ (no similarity)



Transformations

- A similarity can be ***converted*** to a dissimilarity, or vice versa

- Examples
  - Let $s \in [0, 1]$ and $d \in [0, 1]$ be a similarity and a dissimilarity, respectively
  - s can be converted to d as follows:
  - **Subtract**: $d = 1 - s$
    - (ex) $s = 0 \rightarrow d = 1$, $s = 1 \rightarrow d = 0$
  - **Reciprocal**: $d = 2/(s + 1) - 1$
    - (ex) $s = 0 \rightarrow d = 1$, $s = 1 \rightarrow d = 0$
  - **Exponent**: $d = e^{-s}$
    - (ex) $s = 0 \rightarrow d = 1$, $s = 1 \rightarrow d = 0.37$



- In general, any monotonic decreasing function can be used

Examples of Proximity Measures

1. [Dissimilarity] Distances
   - Manhattan distance (L_1 distance)
   - Euclidean distance (L_2 distance)
   - Supremum distance (L_∞ distance)
2. [Similarity] Similarity coefficients
   - Simple matching coefficient
   - Jaccard coefficient
3. [Similarity] Cosine similarity
4. [Similarity] Correlation
5. [Similarity] Mutual information

$\mathbf{x}$ = (1, 3, 6, 2, 9, 7, 5, 4, 10, 8)

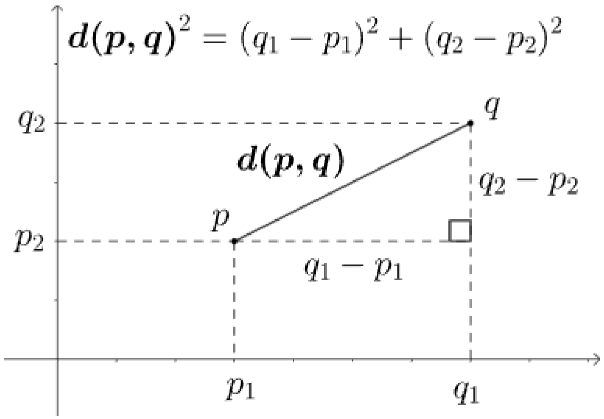
$\updownarrow$ proximity?

$\mathbf{y}$ = (7, 2, 5, 8, 1, 6, 10, 4, 3, 9)

1. Distances (1/2)

- Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be two data objects
  - n: the number of dimensions
  - x_k and y_k: the kth attributes of $\mathbf{x}$ and $\mathbf{y}$, respectively
- ***Distances***: dissimilarities with certain properties
- Euclidean distance
  - The straight-line distance between two points

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$



1. Distances (2/2)

- Generalization of the Euclidean distance (***Minkowski distance***)

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{1/r}$$

– $r = 1$: Manhattan distance (L_1 norm)

  - $d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$

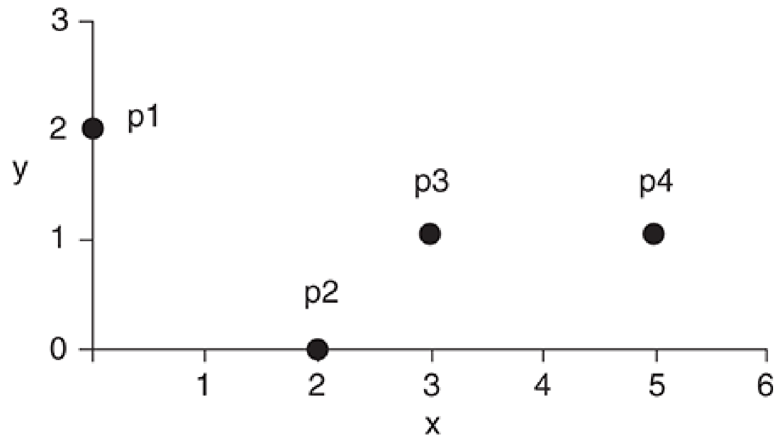
– $r = 2$: Euclidean distance (L_2 norm)

– $r = \infty$: Supremum distance (L_{max} or L_∞ norm)

  - $d(x,\ y) = \lim_{r\to\infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{1/r} = \max_k(|x_k - y_k|)$

(Ex) Minkowski Distance

- Consider the following four two-dimensional points



L_1	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

L_1 distance matrix

L_2	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

L_2 distance matrix

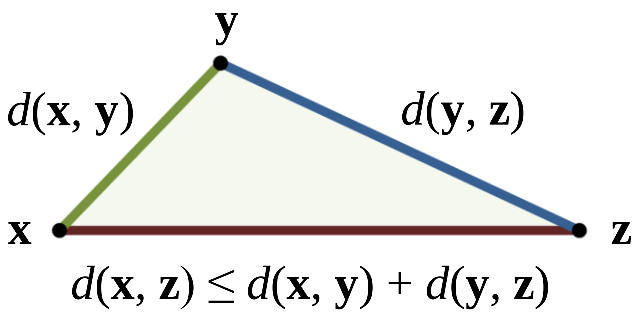
L_∞	p1	p2	p3	p4
p1	0.0	2.0	3.0	5.0
p2	2.0	0.0	1.0	3.0
p3	3.0	1.0	0.0	2.0
p4	5.0	3.0	2.0	0.0

L_∞ distance matrix

The Properties of Distances

- If $d(\mathbf{x}, \mathbf{y})$ is a ***distance***, the following properties hold:

1. Positivity
   - $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$
   - $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$
2. Symmetry
   - $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$
3. Triangle inequality
   - $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$



- These properties are useful because they express our intuition about a distance well

2. Similarity Coefficients (1/2)

- Similarity measures between objects that contain only ***binary*** attributes
  - Typically have values between 0 and 1
    - 0: the objects are not at all similar
    - 1: the objects are completely similar

$\mathbf{x} = (1, 0, 0, 1, 0, 1, 0, 0, 0, 1)$

$\Updownarrow$ Similarity?

$\mathbf{y} = (0, 1, 0, 1, 1, 0, 0, 0, 1, 0)$

- Notations
  - Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be two objects
    - where x_k and y_k are binary attributes ($k = 1, 2, \ldots, n$)
  - f_{00} = the number of attributes where $\mathbf{x}$ is 0 and $\mathbf{y}$ is 0
  - f_{01} = the number of attributes where $\mathbf{x}$ is 0 and $\mathbf{y}$ is 1
  - f_{10} = the number of attributes where $\mathbf{x}$ is 1 and $\mathbf{y}$ is 0
  - f_{11} = the number of attributes where $\mathbf{x}$ is 1 and $\mathbf{y}$ is 1

2. Similarity Coefficients (2/2)

- Simple matching coefficient (*SMC*)
  - Counts both presences and absences equally

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

- Jaccard coefficient (*J*)
  - Counts only presences (e.g., items purchased by both customers)

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

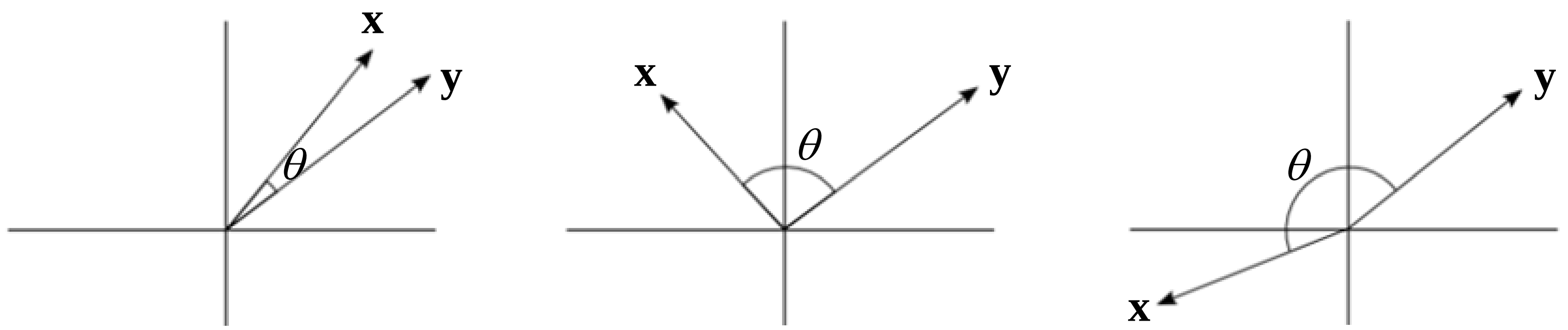
- Example
  - $\mathbf{x} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
  - $\mathbf{y} = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$
  - $SMC = (f_{11} + f_{00})/(f_{01} + f_{10} + f_{11} + f_{00}) = 0.7$, $J = f_{11}/(f_{01} + f_{10} + f_{11}) = 0$

3. Cosine Similarity (1/2)

- Measure the (cosine of the) ***angle*** between two vectors $\mathbf{x}$ and $\mathbf{y}$

$$\cos(\mathbf{x},\ \mathbf{y}) = \frac{\langle \mathbf{x},\ \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

  - $<\mathbf{x}, \mathbf{y}>$: the inner product of $\mathbf{x}$ and $\mathbf{y}$, i.e., $\langle \mathbf{x},\ \mathbf{y} \rangle = \sum_{k=1}^{n} x_k y_k$
  - $\|\mathbf{x}\|$: the length of vector $\mathbf{x}$, i.e., $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2}$



$\theta \approx 0° \rightarrow \cos(\mathbf{x}, \mathbf{y}) \approx 1$ $\theta \approx 90° \rightarrow \cos(\mathbf{x}, \mathbf{y}) \approx 0$ $\theta \approx 180° \rightarrow \cos(\mathbf{x}, \mathbf{y}) \approx -1$

3. Cosine Similarity (2/2)

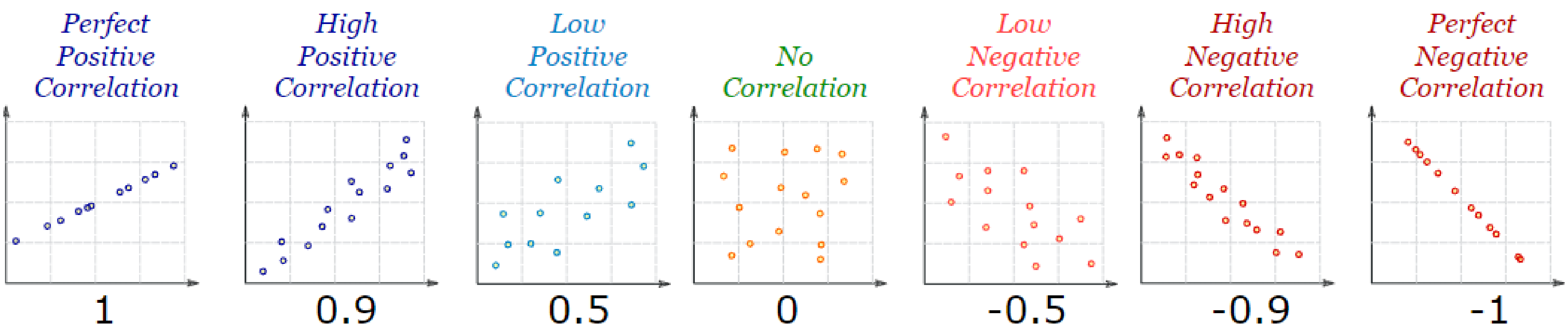
- Useful for measuring the similarity between ***documents***
  - Documents are often represented as ***vectors***
    - Each component represents the frequency of a particular term (word)
  - 0-0 matches are ignored (i.e., words that do not appear in both)
    - If 0-0 matches are counted, most documents will be similar to each other
  - Depends only upon the words that appear in both documents

- (ex) Cosine similarity between two document vectors
  - $\mathbf{x} = (3, 2, 0, 5, 0, 0, 0, 0, 2, 0, 0)$
  - $\mathbf{y} = (1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 2)$
  - $\cos(\mathbf{x}, \mathbf{y}) = \langle\mathbf{x}, \mathbf{y}\rangle/(||\mathbf{x}||\cdot||\mathbf{y}||) = 5/(6.48\cdot2.45) = 0.31$

- Note that the lengths of $\mathbf{x}$ and $\mathbf{y}$ are ***not*** important in $\cos(\mathbf{x}, \mathbf{y})$

4. Correlation

- Measure the ***linear relationship*** between two sets of values
  - Examples
    - **x** = (1, 2, 3, 4, 5), **y** = (2, 4, 6, 8, 10) → perfect positive correlation (= 1)
    - **x** = (1, 2, 3, 4, 5), **y** = (5, 4, 3, 2, 1) → perfect negative correlation (= –1)



- There are many types of correlation
  - In this course, we focus on ***Pearson's correlation***

Pearson's Correlation

- Definition

$$\mathrm{corr}(\mathbf{x},\ \mathbf{y}) = \frac{\mathrm{covariance}(\mathbf{x},\ \mathbf{y})}{\mathrm{standard\_deviation}(\mathbf{x}) \times \mathrm{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x\ s_y}$$

  – where we use the following standard statistical notation and definitions:

$$\mathrm{covariance}(\mathbf{x},\ \mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$\mathrm{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

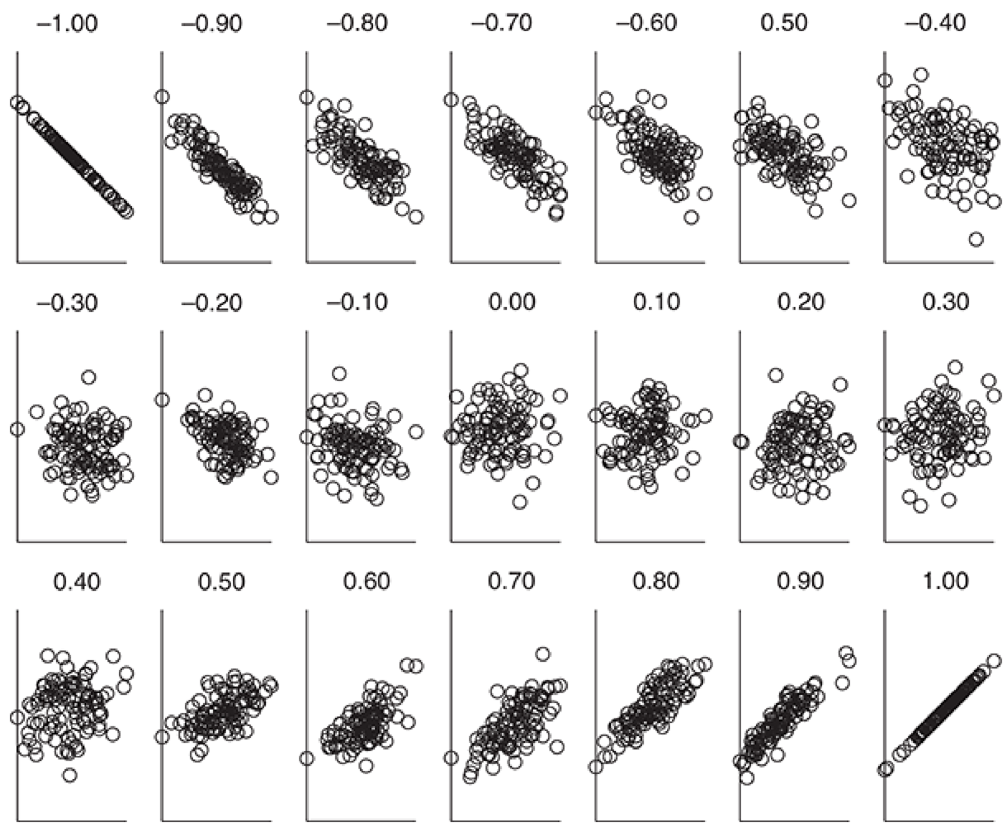
$$\mathrm{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

(Ex) Pearson's Correlation

- Perfect negative correlation
  - $\mathbf{x} = (-3, 6, 0, 3, -6)$
  - $\mathbf{y} = (1, -2, 0, -1, 2)$
  - $\mathrm{corr}(\mathbf{x}, \mathbf{y}) = -1$ ($\because x_k = -3y_k$)
- Perfect positive correlation
  - $\mathbf{x} = (3, 6, 0, 3, 6)$
  - $\mathbf{y} = (1, 2, 0, 1, 2)$
  - $\mathrm{corr}(\mathbf{x}, \mathbf{y}) = 1$ ($\because x_k = 3y_k$)
- No linear correlation
  - $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
  - $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$
  - $\mathrm{corr}(\mathbf{x}, \mathbf{y}) = 0$ ($\because y_k = x_k^2$)



Correlations from –1 to 1

(Ex) Comparing Proximity Measures

Objects / Measure	$\mathbf{x}$ = (1, 2, 4, 3, 0, 0, 0) $\mathbf{y}$ = (1, 2, 3, 4, 0, 0, 0)	$\mathbf{x}$ = (1, 2, 4, 3, 0, 0, 0) $\mathbf{y}_s$ = (2, 4, 6, 8, 0, 0, 0) ($\mathbf{y}_s = 2\mathbf{y}$)	$\mathbf{x}$ = (1, 2, 4, 3, 0, 0, 0) $\mathbf{y}_t$ = (6, 7, 8, 9, 5, 5, 5) ($\mathbf{y}_t = \mathbf{y} + 5$)
cos($\mathbf{x}$, $\mathbf{y}$)	0.9667	0.9667	0.7940
corr($\mathbf{x}$, $\mathbf{y}$)	0.9429	0.9429	0.9429
Euclidean distance($\mathbf{x}$, $\mathbf{y}$)	1.4142	5.8310	14.2127

Mutual Information (1/2)

- Measure the similarity between two sets of ***paired values***
  - Particularly when a ***nonlinear*** relationship is suspected
- Measure how much ***information*** one set of values provides about another
  - Given that the values in pairs (e.g., height and weight)
- Intuitive example (0: head, 1: tail)

x	y	Mutual information
(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)	(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)	**1**
(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)	(1, 1, 1, 1, 1, 0, 0, 0, 0, 0)	**1**
(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)	(0, 0, 1, 0, 0, 0, 1, 0, 1, 1)	**0.1535**

Mutual Information (2/2)

- If the two sets of values are completely ***independent***
  - i.e., the value of one tells us ***nothing*** about the other
  - Then their mutual information is 0
- If the two sets of values are completely ***dependent***
  - i.e., knowing the value of one ***tells*** us the value of the other
  - Then they have maximum mutual information

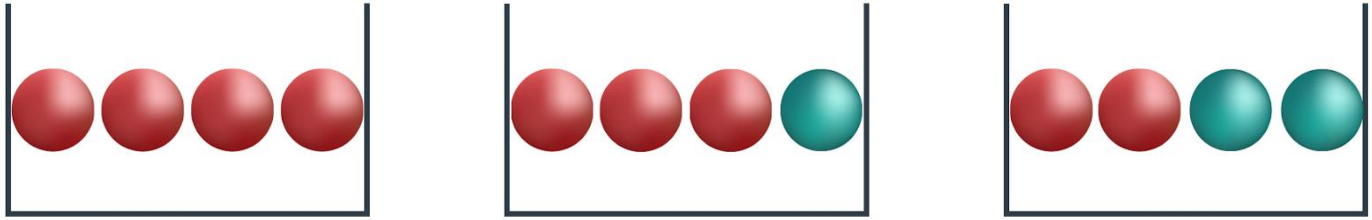
x	y	Mutual information
(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	**0**
(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)	(10, 9, 8, 7, 6, 5, 4, 3, 2, 1)	**3.322**

Entropy

- Measure the ***average information*** in a single set of values

$$H(X) = \sum_{j=1}^{m} P(X = u_j) I(X = u_j) = -\sum_{j=1}^{m} P(X = u_j) \log_2 P(X = u_j)$$

  - X: a set of values with m distinct values $u_1, u_2, \ldots, u_m$
  - $H(X)$: the entropy of X
  - $P(X = u_j)$: the probability of u_j in X
  - $I(X = u_j)$: the amount of information acquired through observing u_j
    - $I(X = u_j) = \log_2(1/P(X = u_j)) = -\log_2 P(X = u_j)$
    - As $P(X = u_j)$ increases, $I(X = u_j)$ decreases, and vice versa



Entropy = 0 Entropy = 0.81 Entropy = 1

Definition: Mutual Information (1/2)

- Consider two sets of values, X and Y, which occur in pairs (X, Y)

X	(1, 2, 3, 1, 3)
Y	(2, 3, 1, 2, 2)
(X, Y)	((1, 2), (2, 3), (3, 1), (1, 2), (3, 2))

- First, we measure the ***average information*** **(*i.e., entropy*)** of X, Y, and (X, Y), respectively

$$H(X) = -\sum_{j=1}^{m} P(X = u_j) \log_2 P(X = u_j)$$

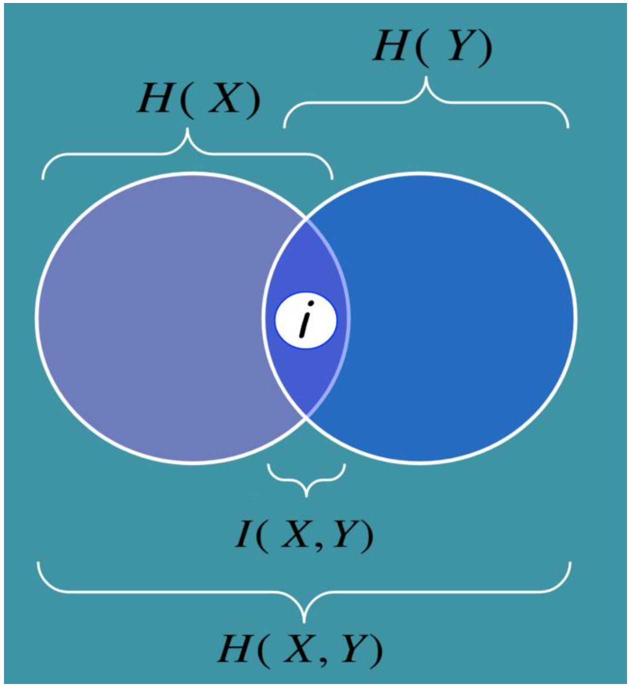
$$H(Y) = -\sum_{k=1}^{n} P(Y = v_k) \log_2 P(Y = v_k)$$

$$H(X, Y) = -\sum_{j=1}^{m} \sum_{k=1}^{n} P(X = u_j, Y = v_k) \log_2 P(X = u_j, Y = v_k)$$

Definition: Mutual Information (2/2)

- Finally, we obtain the mutual information of X and Y as follows:

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$



- The mutual information quantifies the "***amount of information***" obtained about X by observing Y, and vise versa
  - Note that $I(X, Y)$ is symmetric, i.e., $I(X, Y) = I(Y, X)$

(Ex) Mutual Information

- Suppose we have two sets of values $\mathbf{x}$ and $\mathbf{y}$ to compare
  - $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$, $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$
  - Although there is a relationship $y_k = x_k^2$, their correlation is 0

- However, their mutual information $I(\mathbf{x}, \mathbf{y}) = 1.9502$
  - $I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) = 2.8074 + 1.9502 - 2.8074$

x_j	$P(\mathbf{x} = x_j)$	$-P(\mathbf{x} = x_j)\log_2 P(\mathbf{x} = x_j)$
-3	1/7	0.4011
-2	1/7	0.4011
-1	1/7	0.4011
0	1/7	0.4011
1	1/7	0.4011
2	1/7	0.4011
3	1/7	0.4011
$H(\mathbf{x})$		2.8074

y_k	$P(\mathbf{y} = y_k)$	$-P(\mathbf{y} = y_k)\log_2 P(\mathbf{y} = y_k)$
9	2/7	0.5164
4	2/7	0.5164
1	2/7	0.5164
0	1/7	0.4011
$H(\mathbf{y})$		1.9502

x_j	y_k	$P(\mathbf{x} = x_j,\ \mathbf{y} = x_k)$	$-P(\mathbf{x} = x_j,\ \mathbf{y} = x_k)\log_2 P(\mathbf{x} = x_j,\ \mathbf{y} = x_k)$
-3	9	1/7	0.4011
-2	4	1/7	0.4011
-1	1	1/7	0.4011
0	0	1/7	0.4011
1	1	1/7	0.4011
2	4	1/7	0.4011
3	9	1/7	0.4011
$H(\mathbf{x}, \mathbf{y})$			2.8074

Issues in Proximity Calculation (1/2)

- **(Issue 1)** Standardization
  - If attributes have different scales, ***standardize*** them to avoid being dominated by attributes with large values
  - **Rescaling**
    - Rescale the range of attributes to be $[0, 1]$

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  - **Mean normalization**
    - Rescale based on the distance from the mean

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

  - **Standardization (in statistics)**
    - Make attributes have 0-mean and 1-variance

$$x' = \frac{x - \bar{x}}{\sigma}$$

Issues in Proximity Calculation (2/2)

- **(Issue 2)** Using weights
  - In some cases, some attributes are more important than others
    - (ex) When comparing two people, *Age* may be more important than *Height*
  - We can assign each attribute a ***different*** weight w_k

Attribute	Age	Height	Weight	Salary
Weight	0.5	0.2	0.2	0.1

  - The definition of the Minkowski distance can also be modified as follows:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} \underline{w_k} |x_k - y_k|^r \right)^{1/r}$$