Data Mining: Data

Dr. Meng Qu
Rutgers University

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Outline

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

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

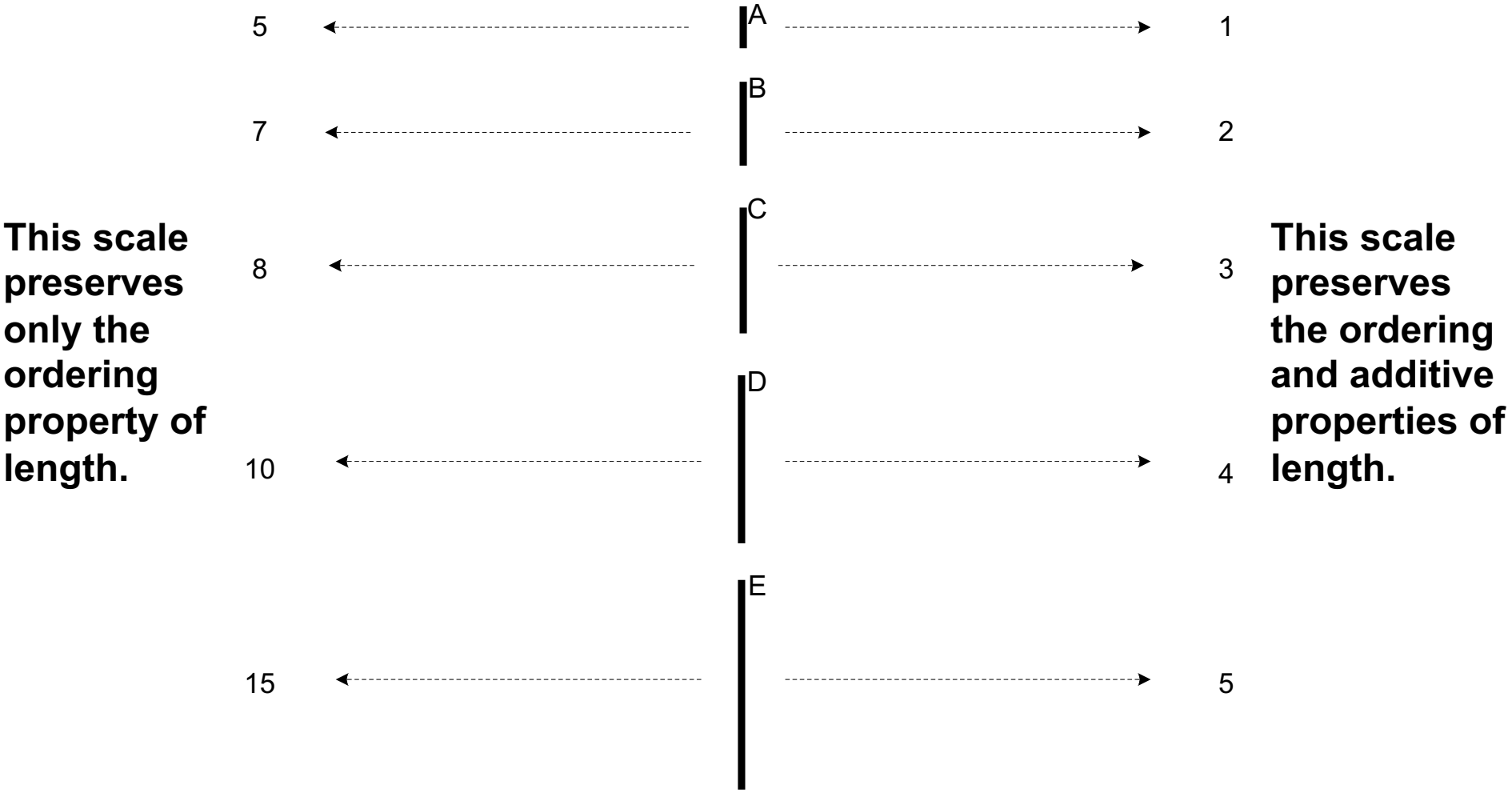
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different

Measurement of Length

- The way you measure an attribute may not match the attributes properties.



This scale preserves only the ordering property of length.

This scale preserves the ordering and additive properties of length.

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
    - Nominal values provide only enough information to distinguish one object from another.
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
    - The values of an ordinal attribute provide enough information to order objects.

Types of Attributes

- There are different types of attributes
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
    - For interval attributes, the differences between values are meaningful.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts
    - For ratio variables, both differences and ratios are meaningful.

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10 ° degrees twice that of 5° on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?

- Consider measuring the height above average
  - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?
  - Is this situation analogous to that of temperature?

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ 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, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 0}.
Numeric Quantitative	Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

This categorization of attributes is due to S. S. Stevens

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, 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.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Asymmetric Attributes

- Only presence (a non-zero attribute value) is regarded as important
- Examples:
  - Words present in documents
  - Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

  "I see our purchases are very similar since we didn't buy most of the same things."

- We need two asymmetric binary attributes to represent one ordinary binary attribute
  - Association analysis uses asymmetric attributes

Asymmetric Attributes

- Consider a data set where each object is a student and each attribute records whether or not a student took a particular course at a university.
- For a specific student, an attribute has a value of 1 is the student took the course associated with that attribute and a value of 0 otherwise.
- Students take only a small fraction of all available courses, most of the data set would be 0.
- It is more meaningful and more efficient to focus on the non-zero values.

Types of data sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Important Characteristics of Structured Data

- Dimensionality
  - The number of attributes that the objects in the data set possess.
  - Curse of Dimensionality
- Sparsity
  - Only presence counts
  - In many cases, fewer than 1% of the entries are non-zero.
- Resolution
  - Patterns depend on the scale
  - The surface of the Earth seems very uneven at a resolution of a few meters, but is relatively smooth at a resolution of tens of kilometers.

Record Data

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

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

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

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

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

Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

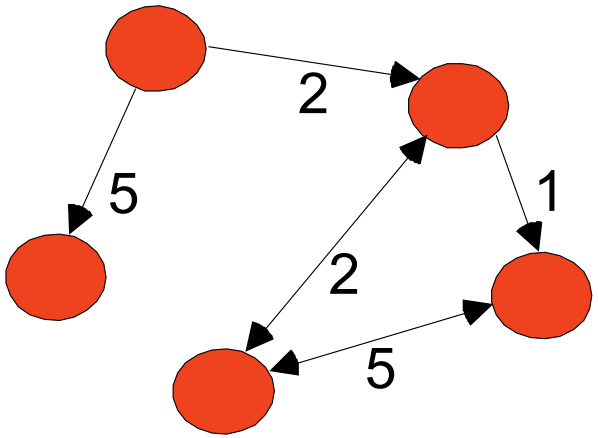
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

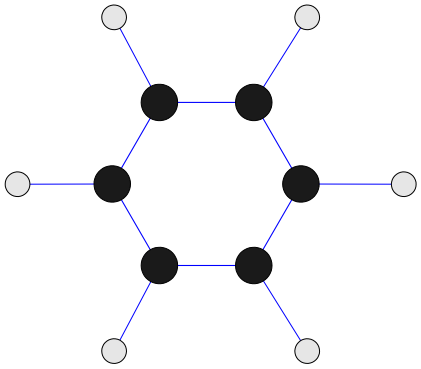
Transaction Data

- A special type of record data
  - Also been called market basket data (the items in each record are the person's "market basket")
  - Can be viewed as a set of records whose fields are asymmetric attributes.
  - The attributes can be binary, indicating weather or not an item was purchased.

Graph Data

- Examples: Generic graph, a Moleule, and Webpages





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

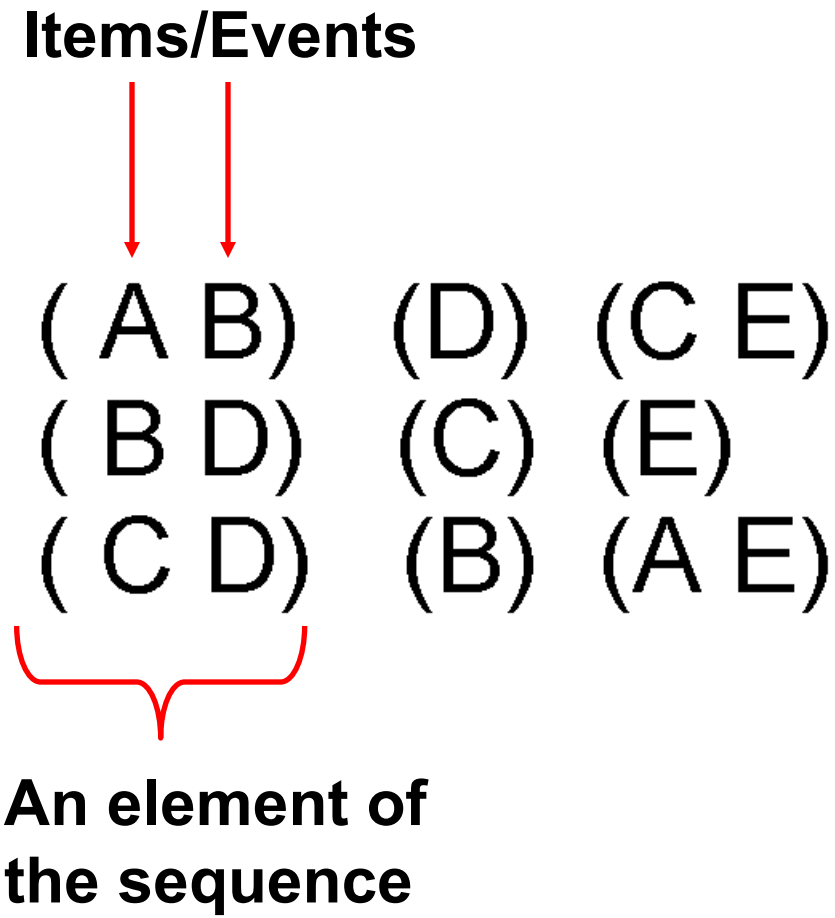
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- Sequential data of transactions



Ordered Data

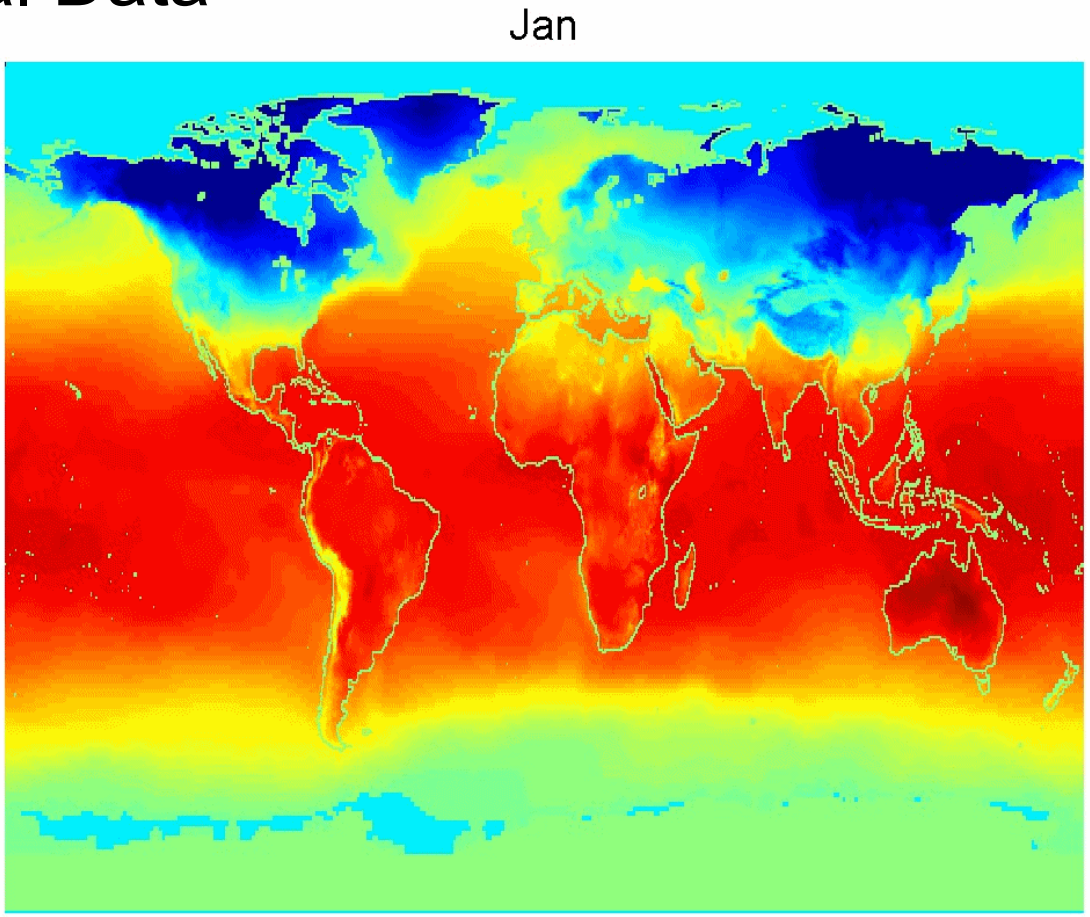
- Genomic sequence data (no time stamps)

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

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Data Quality

- Poor data quality negatively affects many data processing efforts

"The most important point is that poor data quality is an unfolding disaster.

  - Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate."

Thomas C. Redman, DM Review, August 2004

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

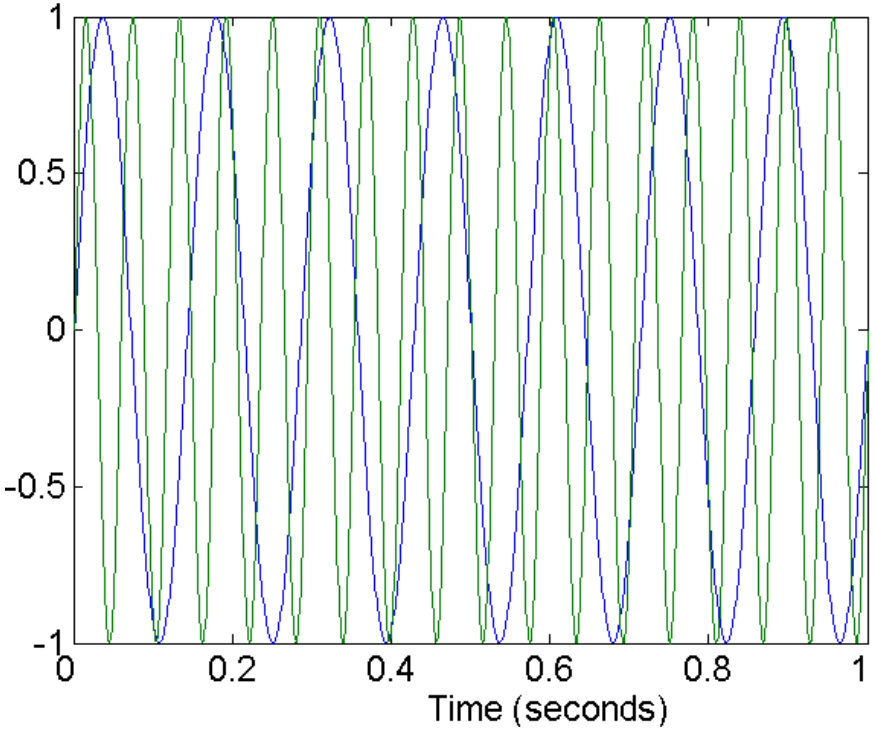
Data Quality …

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

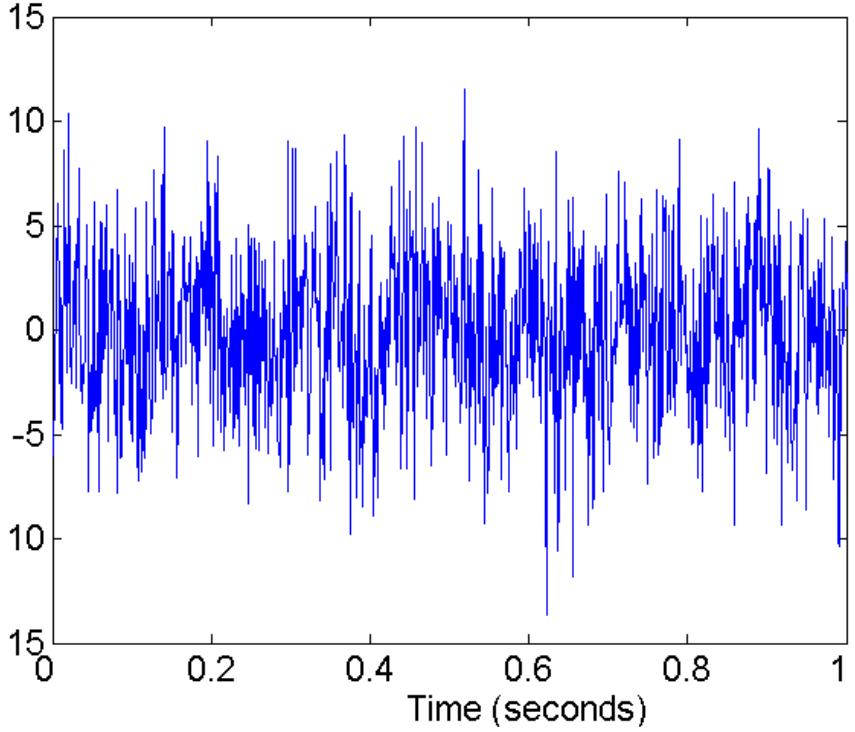
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



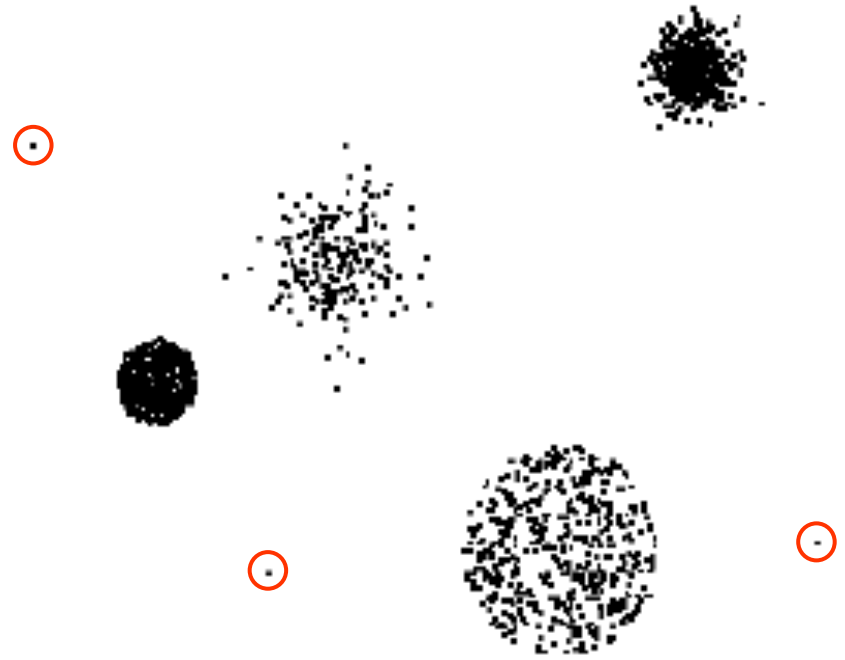
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **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
  - Estimate missing values
    - Example: time series of temperature
    - Example: census results
  - Ignore the missing value during analysis

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

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)

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

Example: Precipitation in Australia

- This example is based on precipitation in Australia from the period 1982 to 1993.

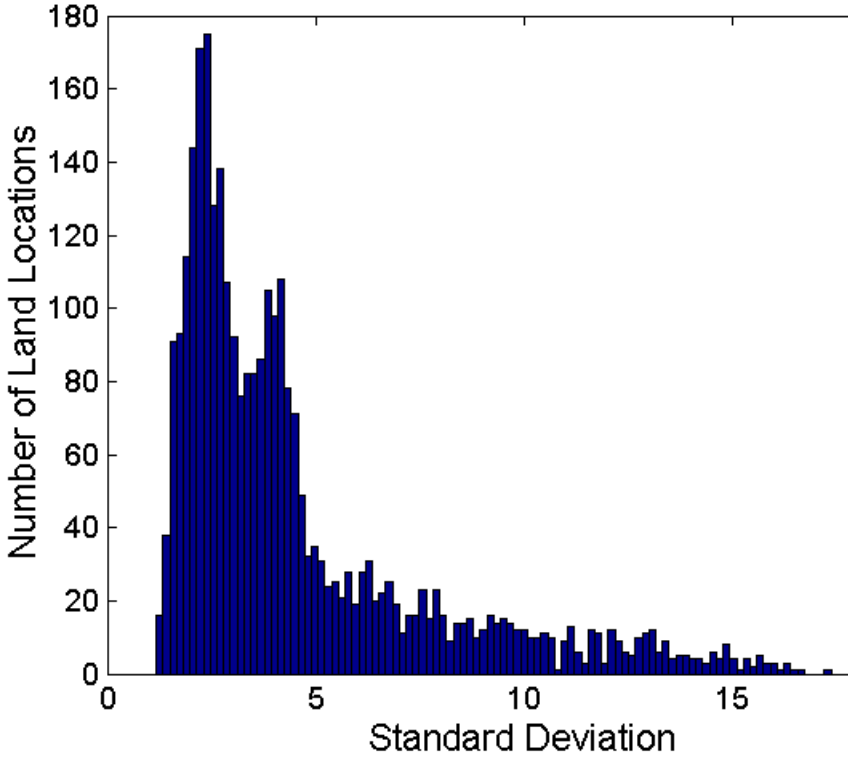
  The next slide shows

  - A histogram for the standard deviation of average monthly precipitation for 3,030 0.5◦ by 0.5◦ grid cells in Australia, and
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.

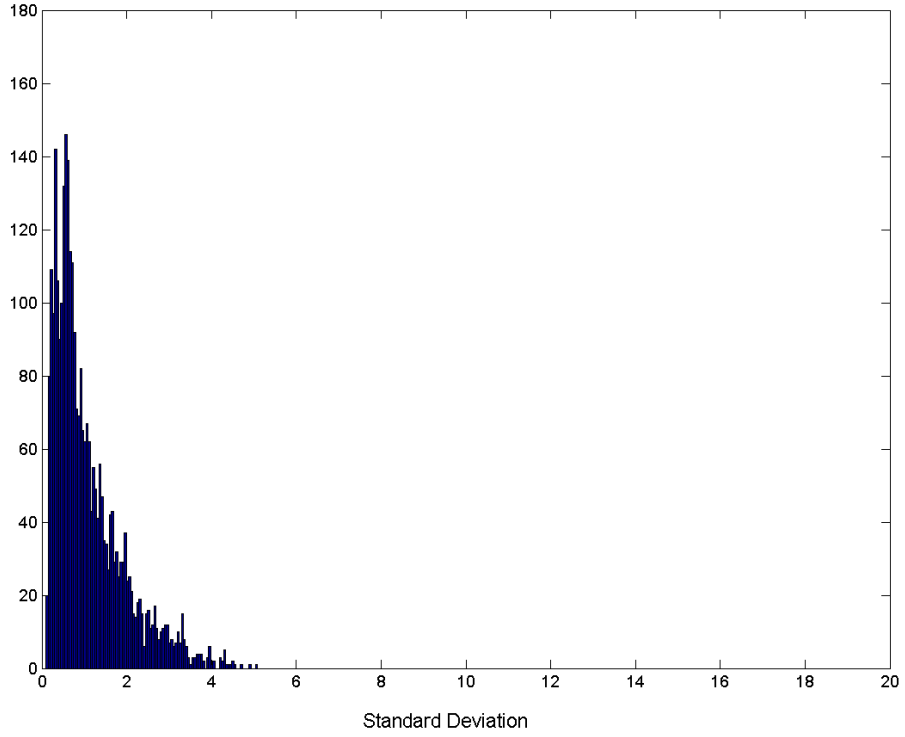
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia …

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

Sampling …

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

How to Sample Badly…

- The design of a sample is **biased** if it systematically favors certain outcomes.

Biased Sampling Methods

Selection of whichever individuals are easiest to reach is called convenience sampling.

A voluntary response sample chooses itself by responding to a general appeal. Write-in or call-in opinion polls are examples of voluntary response samples.

Convenience samples and voluntary response samples are often biased.

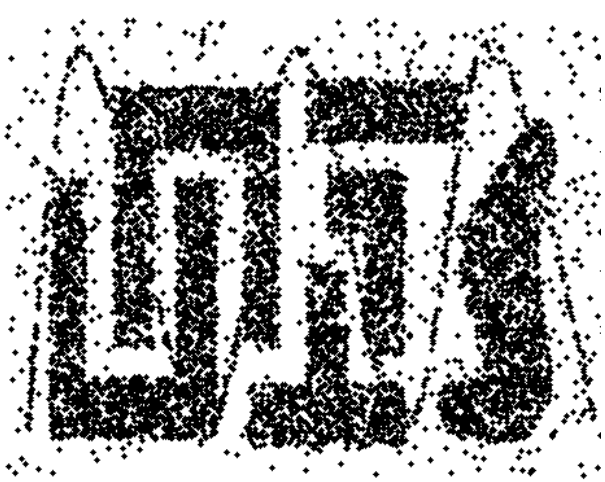
How to Sample Badly…

EXAMPLE 2 Write-in opinion polls

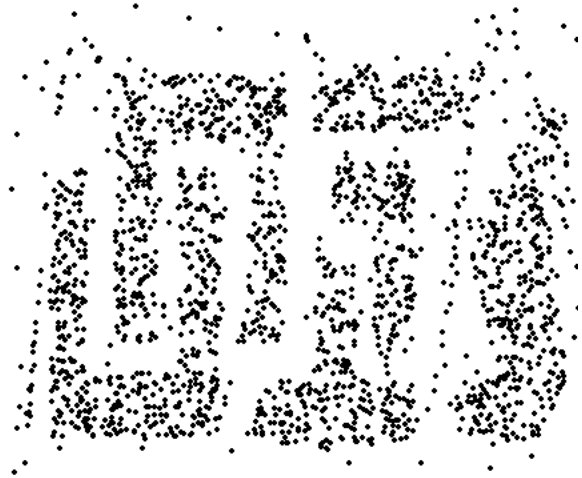
Ann Landers once asked the readers of her advice column, "If you had it to do over again, would you have children?" She received nearly 10,000 responses, almost 70% saying "NO!" Can it be true that 70% of parents regret having children? Not at all. This is a voluntary response sample. People who feel strongly about an issue, particularly people with strong negative feelings, are more likely to take the trouble to respond. Ann Landers's results are strongly biased—the percentage of parents who would not have children again is much higher in her sample than in the population of all parents.

On August 24, 2011, Abigail Van Buren (the niece of Ann Landers) revisited this question in her column "Dear Abby." A reader asked, "Many years ago an advice columnist (your mother?) posed the question to her readers, 'If you had it to do over again, would you still have children?' I'm wondering when the information was collected and what the results of that inquiry were, and if you asked the same question today, what the majority of your readers would answer."

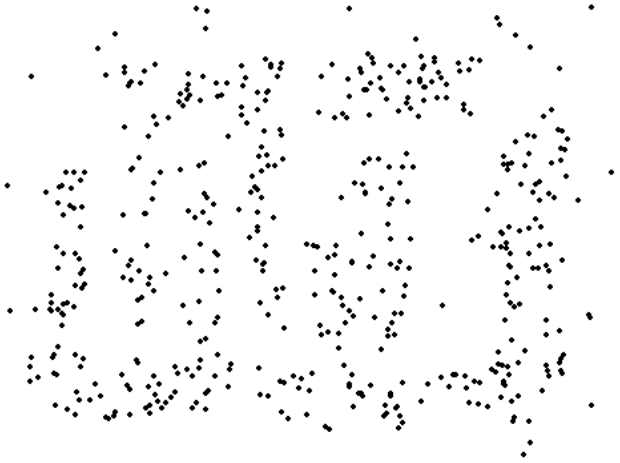
Sample Size



8000 points



2000 Points



500 Points

Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - 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
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

Cautions About Sample Surveys

- Good sampling technique includes the art of reducing all sources of error.

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.

A systematic pattern of incorrect responses in a sample survey leads to **response bias**.

The **wording of questions** is the most important influence on the answers given to a sample survey.

Case Study: The 1936 Literary Digest Poll

The presidential election of 1936 pitted Alfred Landon, the Republican governor of Kansas, against the incumbent President, Franklin D. Roosevelt. The year 1936 marked the end of the Great Depression, and economic issues such as unemployment and government spending were the dominant themes of the campaign. The *Literary Digest* was one of the most respected magazines of the time and had a history of accurately predicting the winners of presidential elections that dated back to 1916. For the 1936 election, the *Literary Digest* prediction was that Landon would get 57% of the vote against Roosevelt's 43% (these are the *statistics* that the poll measured). The actual results of the election were 62% for Roosevelt against 38% for Landon (these were the *parameters* the poll was trying to measure). The sampling error in the *Literary Digest* poll was a whopping 19%, the largest ever in a major public opinion poll. Practically all of the sampling error was the result of sample bias.

Case Study: The 1936 Literary Digest Poll

The irony of the situation was that the *Literary Digest* poll was also one of the largest and most expensive polls ever conducted, with a sample size of around 2.4 million people! At the same time the *Literary Digest* was making its fateful mistake, George Gallup was able to predict a victory for Roosevelt using a much smaller sample of about 50,000 people.

This illustrates the fact that bad sampling methods cannot be cured by increasing the size of the sample, which in fact just compounds the mistakes. The critical issue in sampling is not sample size but how best to reduce sample bias.

Case Study: The 1936 Literary Digest Poll

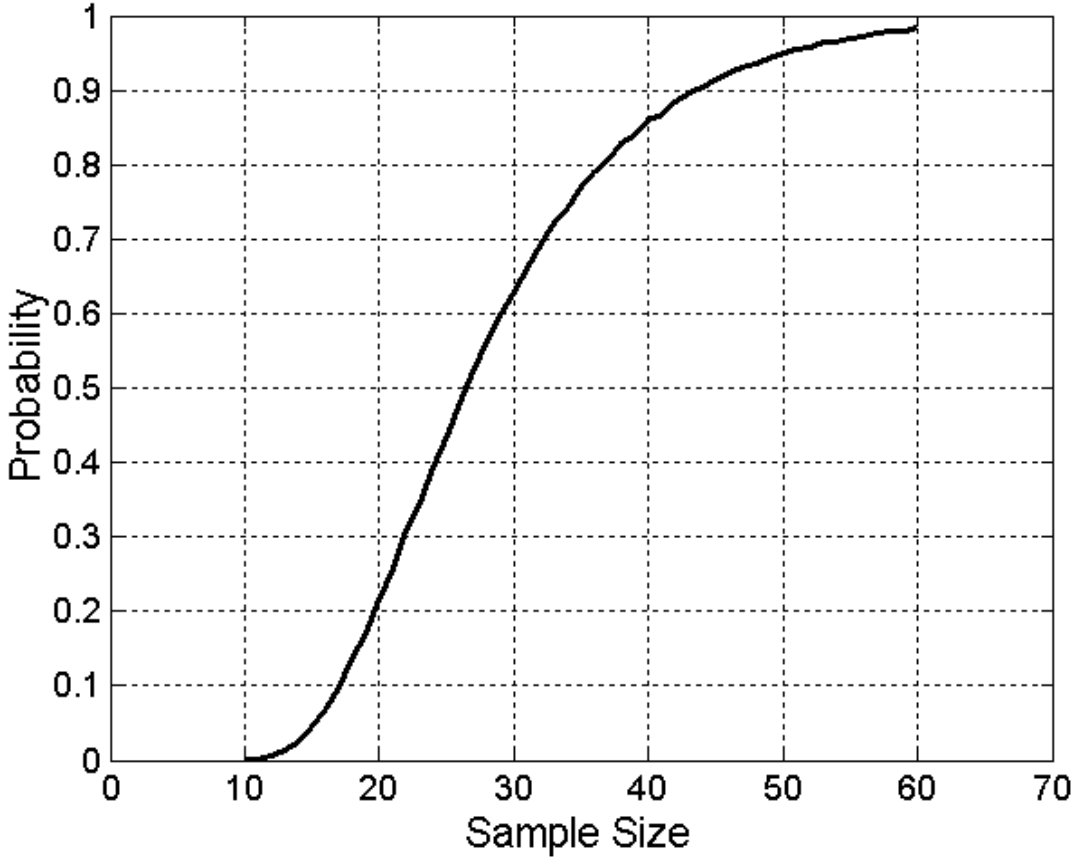
The first major problem with the poll was in the selection process for the names on the mailing list, which were taken from telephone directories, club membership lists, lists of magazine subscibers, etc. Such a list is guaranteed to be slanted toward middle- and upper-class voters, and by default to exclude lower-income voters. One must remember that in 1936, telephones were much more of a luxury than they are today. Furthermore, at a time when there were still 9 million people unemployed, the names of a significant segment of the population would not show up on lists of club memberships and magazine subscribers. At least with regard to economic status, the *Literary Digest* mailing list was far from being a representative cross-seciton of the population. This is always a critical problem because voters are generally known to vote their pocketbooks, and it was magnified in the 1936 election when economic issues were preeminent in the minds of the voters. This sort of sample bias is called *selection bias*.

Case Study: The 1936 Literary Digest Poll

The second problem with the *Literary Digest* poll was that out of the 10 million people whose names were on the original mailing list, only about 2.4 million responded to the survey. Thus, the size of the sample was about one-fourth of what was originally intended. People who respond to surveys are different from people who don't, not only in the obvious way (their attitude toward surveys) but also in more subtle and significant ways. When the response rate is low (as it was in this case, 0.24), a survey is said to suffer from *nonresponse bias*. This is a special type of selection bias where reluctant and nonresponsive people are excluded from the sample.

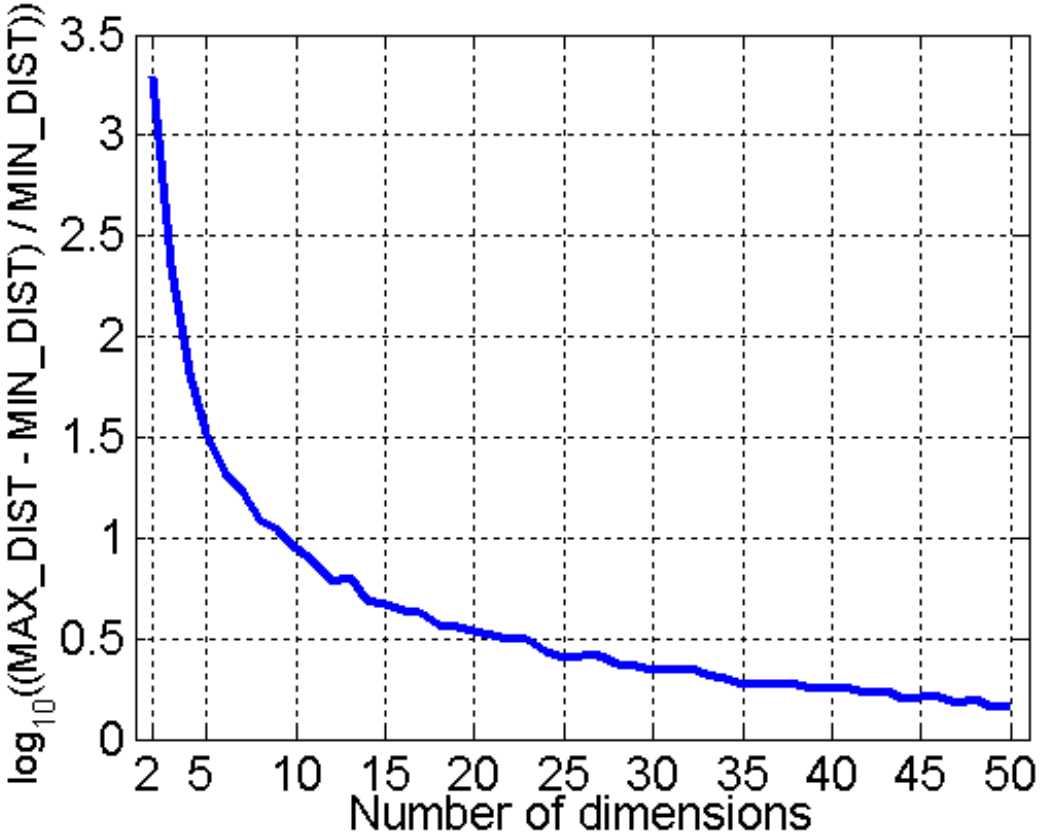
Sample Size

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



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
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



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

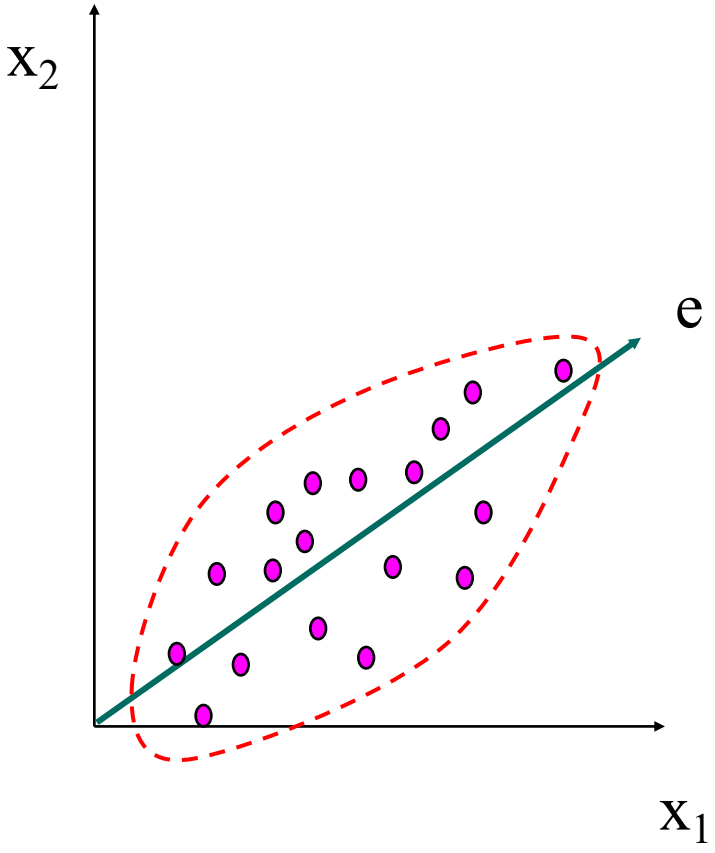
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Dimensionality Reduction: PCA

Feature Subset 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
- Many techniques developed, especially for classification

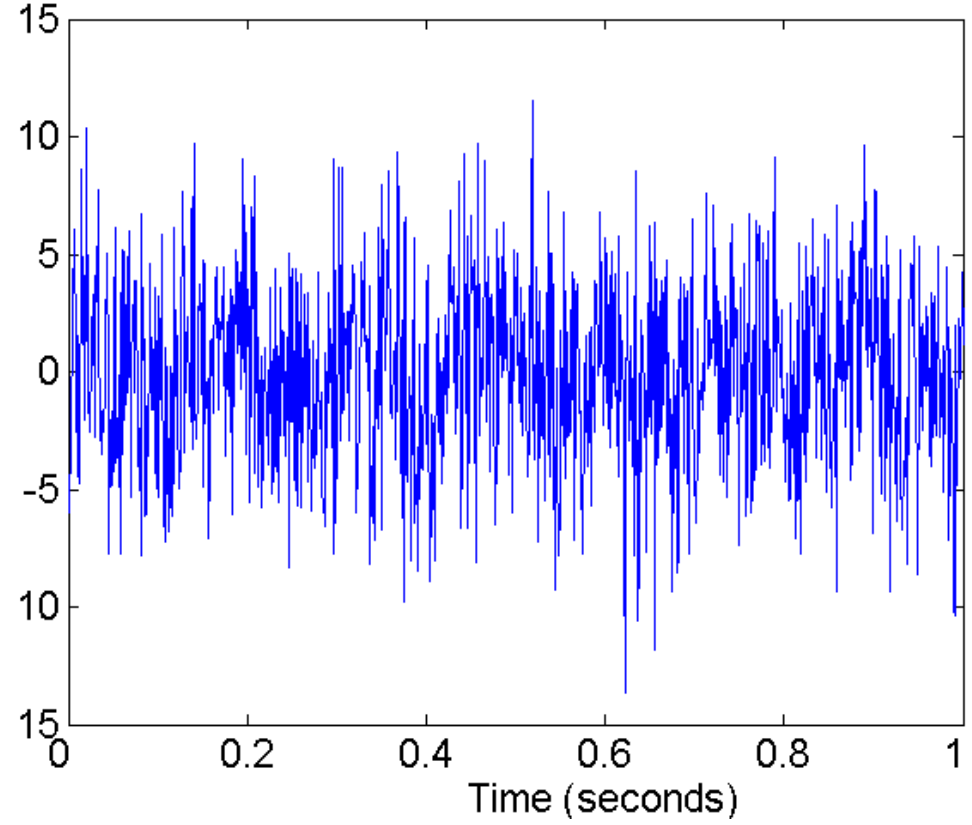
Feature Creation

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

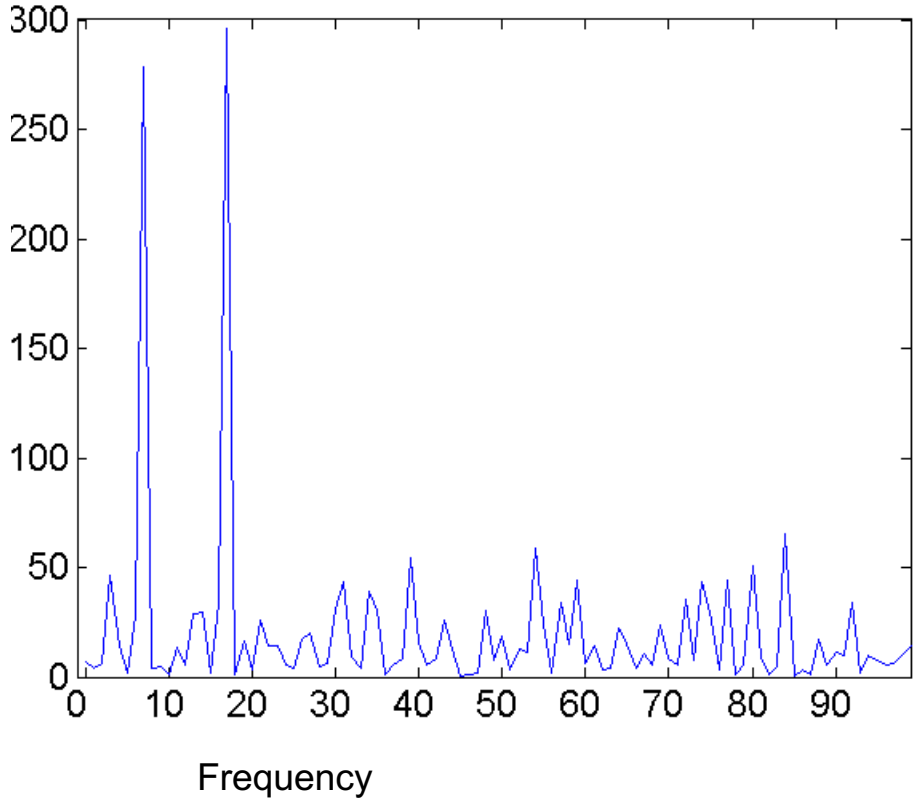
- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Mapping Data to a New Space

- Fourier and wavelet transform



Two Sine Waves + Noise



Frequency

While **dimensionality reduction** is an important tool in machine learning/data mining, we must always be aware that it can distort the data in misleading ways.

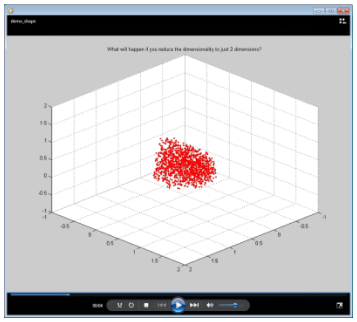
Above is a two dimensional projection of an intrinsically three dimensional world….

Original photographer unknown

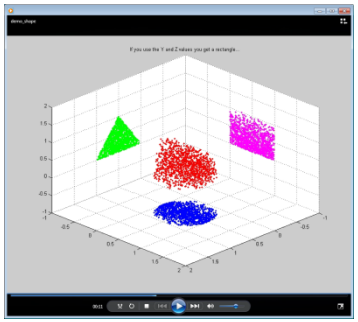
See also www.cs.gmu.edu/~jessica/DimReducDanger.htm

Screen dumps of a short video from www.cs.gmu.edu/~jessica/DimReducDanger.htm
I recommend you imbed the original video instead

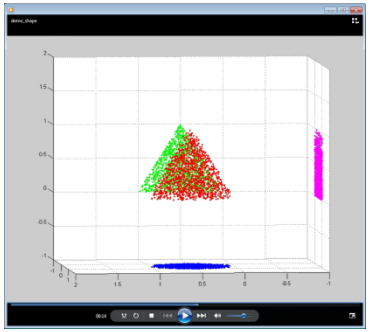
A cloud of points in 3D



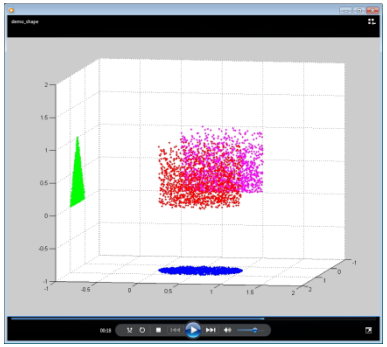
Can be projected into 2D
XY or XZ or YZ



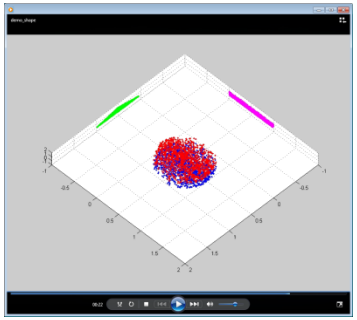
In 2D XZ we see
a triangle



In 2D YZ we see
a square



In 2D XY we see
a circle



Discretization

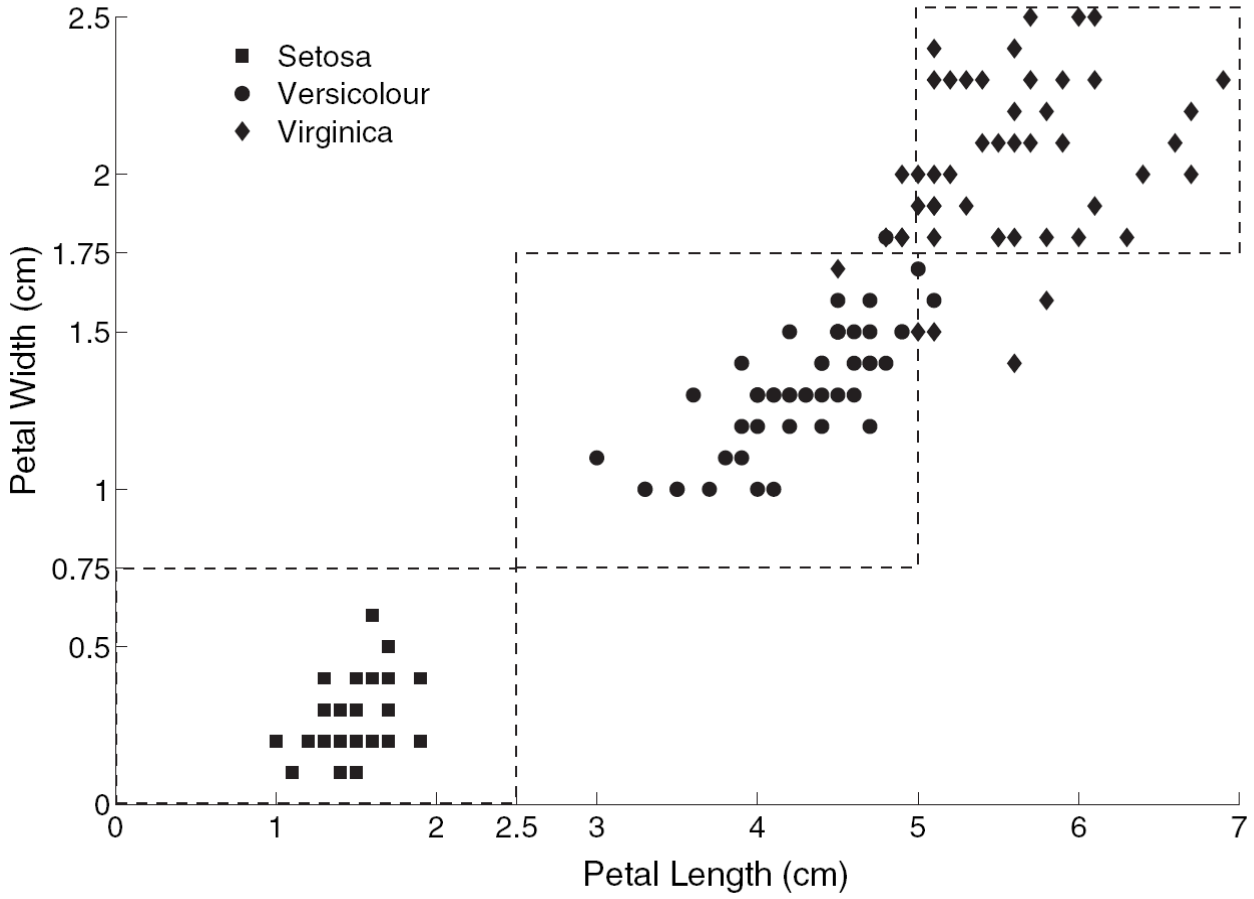
- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values
  - We give an illustration of the usefulness of discretization using the Iris data set

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

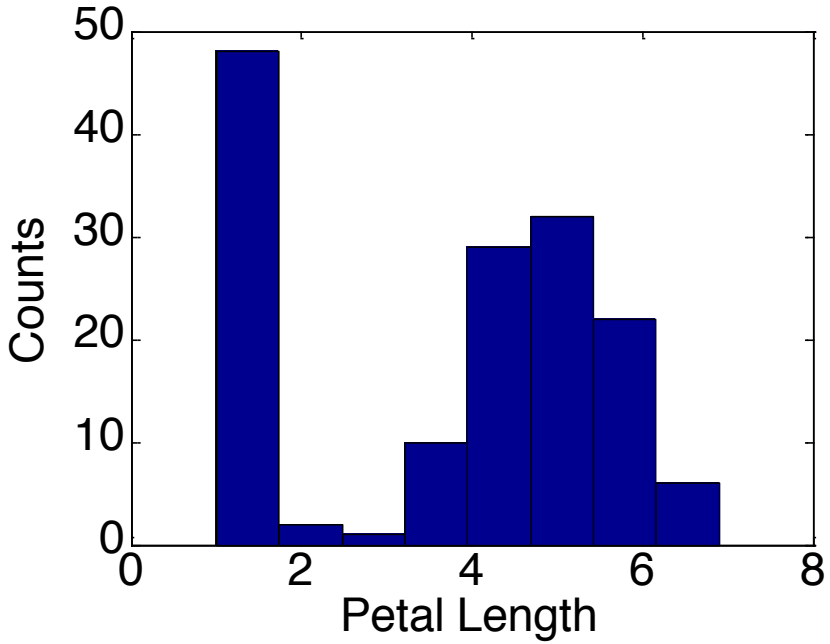
Discretization: Iris Example



Petal width low or petal length low implies Setosa.
Petal width medium or petal length medium implies Versicolour.
Petal width high or petal length high implies Virginica.

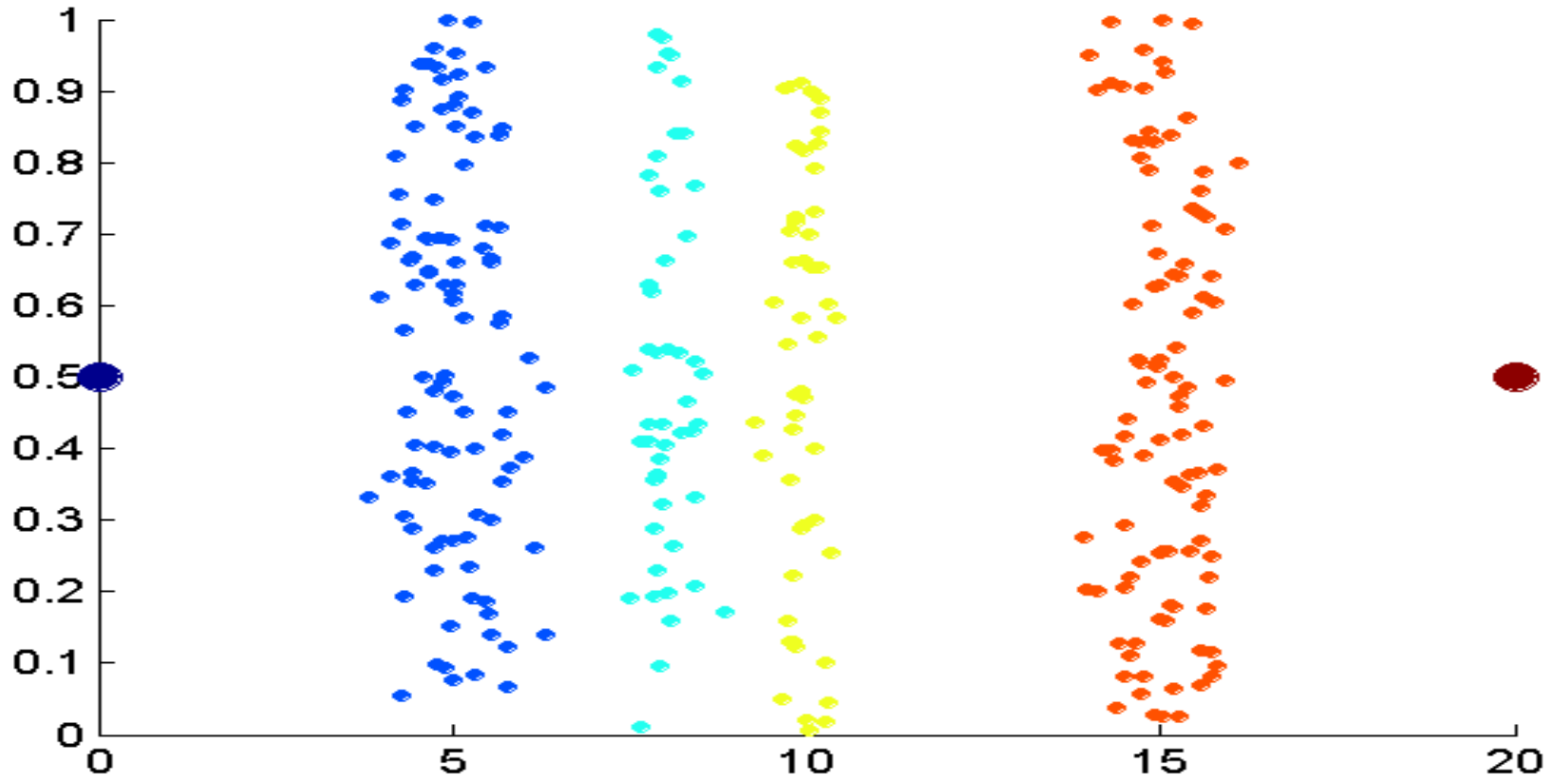
Discretization: Iris Example …

- How can we tell what the best discretization is?
  - Unsupervised discretization: find breaks in the data values
    - Example: Petal Length



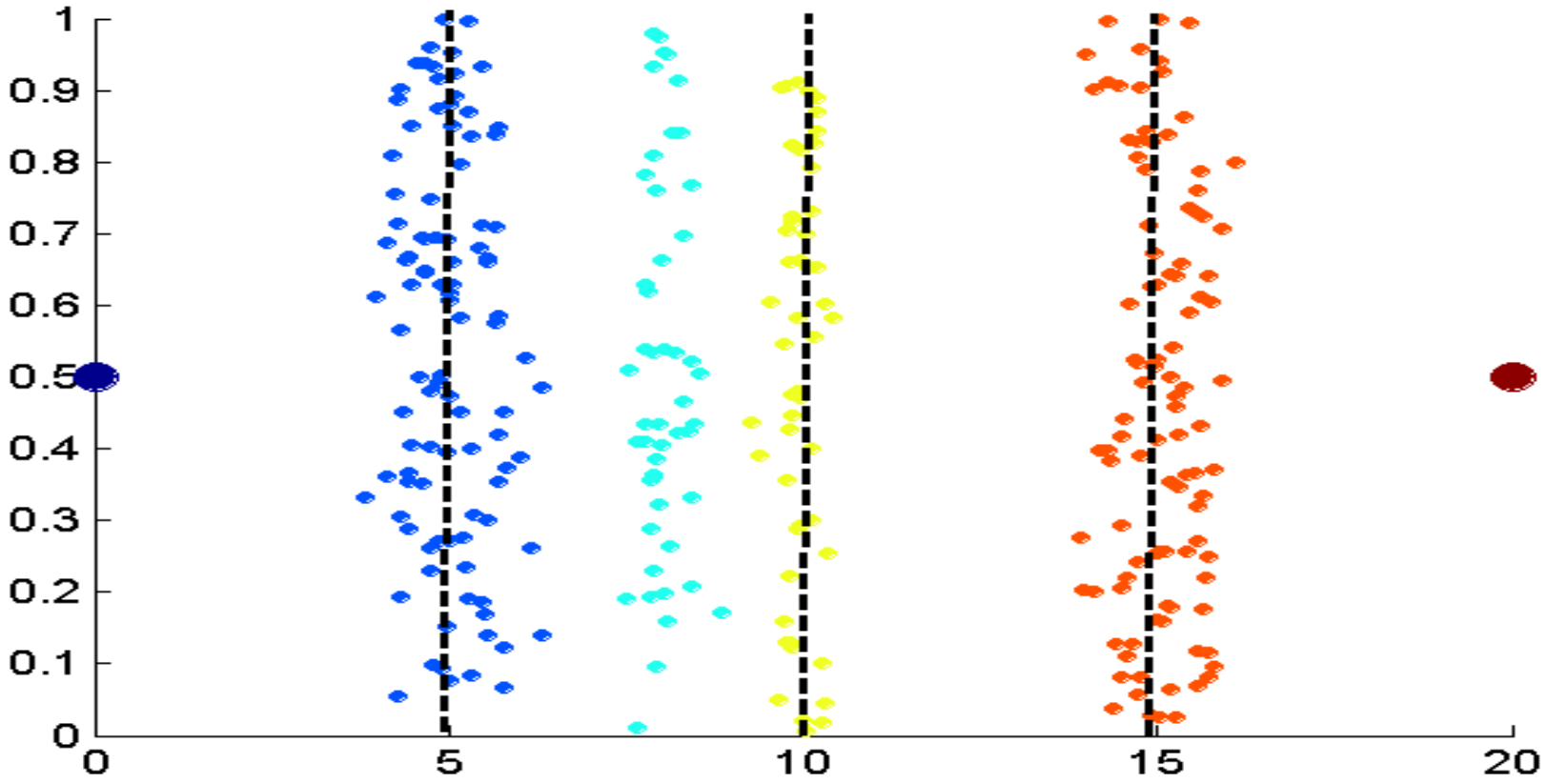
  - Supervised discretization: Use class labels to find breaks

Discretization Without Using Class Labels



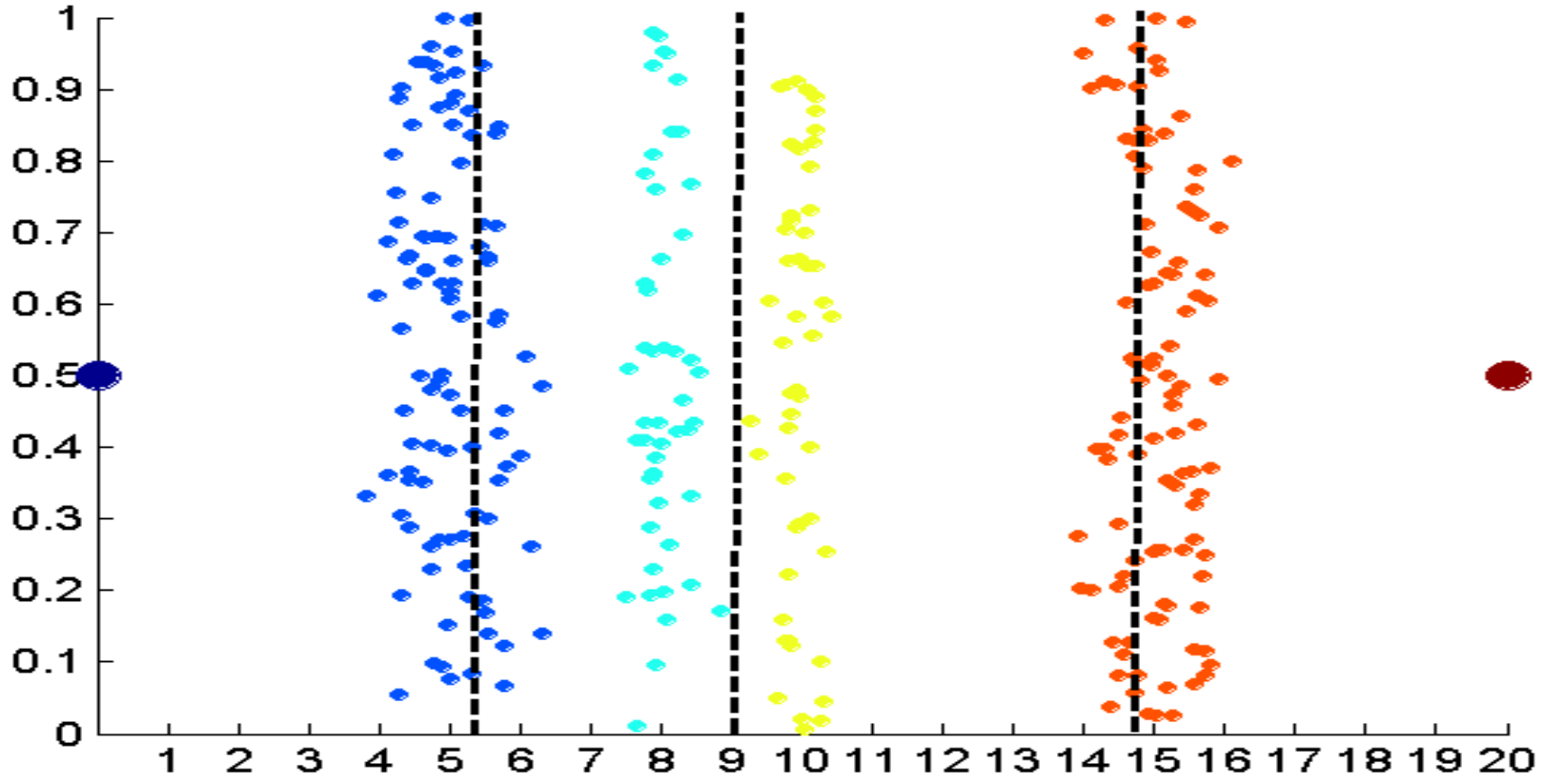
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Discretization Without Using Class Labels



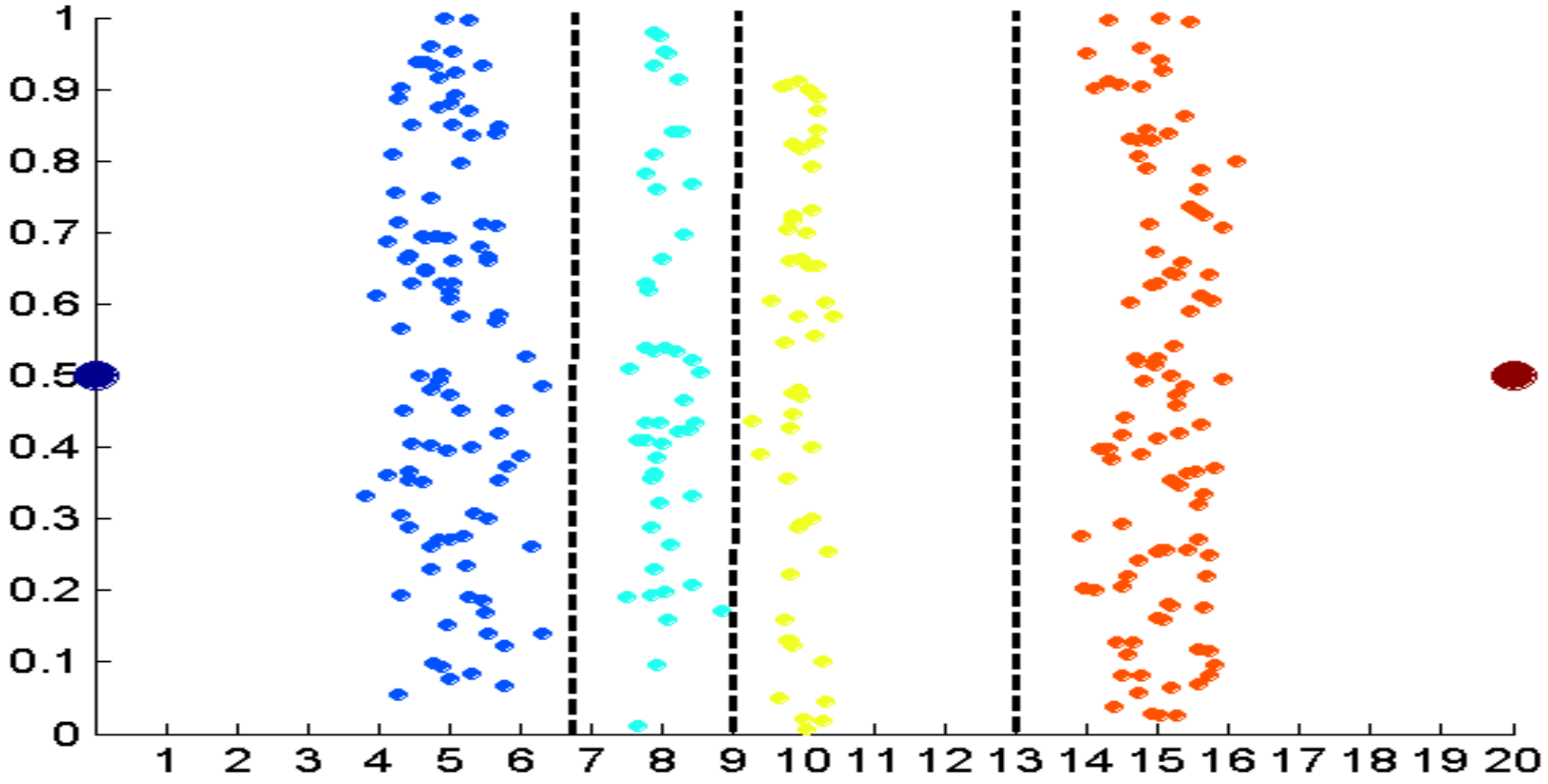
Equal interval width approach used to obtain 4 values.

Discretization Without Using Class Labels



Equal frequency approach used to obtain 4 values.

Discretization Without Using Class Labels



K-means approach to obtain 4 values.

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

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 frequency of occurrence, mean, variance, magnitude
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Standard Scores

Because the standard deviation is the natural unit of measurement for Normal distributions, we can restate observations in terms of how many standard deviations above or below the mean they are.

Observations expressed in standard deviations above or below the mean of a distribution are called **standard scores.**

Standard Score

The standard score for any observation is

$$\text{standard score} = \frac{\text{observation - mean}}{\text{standard deviation}}$$

Standard Scores

EXAMPLE 3 ACT versus SAT scores

Jennie scored 600 on the SAT Mathematics exam. Her friend Gerald took the American College Testing (ACT) test and scored 21 on the math part. ACT scores are Normally distributed with mean 18 and standard deviation 6. Assuming that both tests measure the same kind of ability, who has the higher score?

Jennie's standard score is

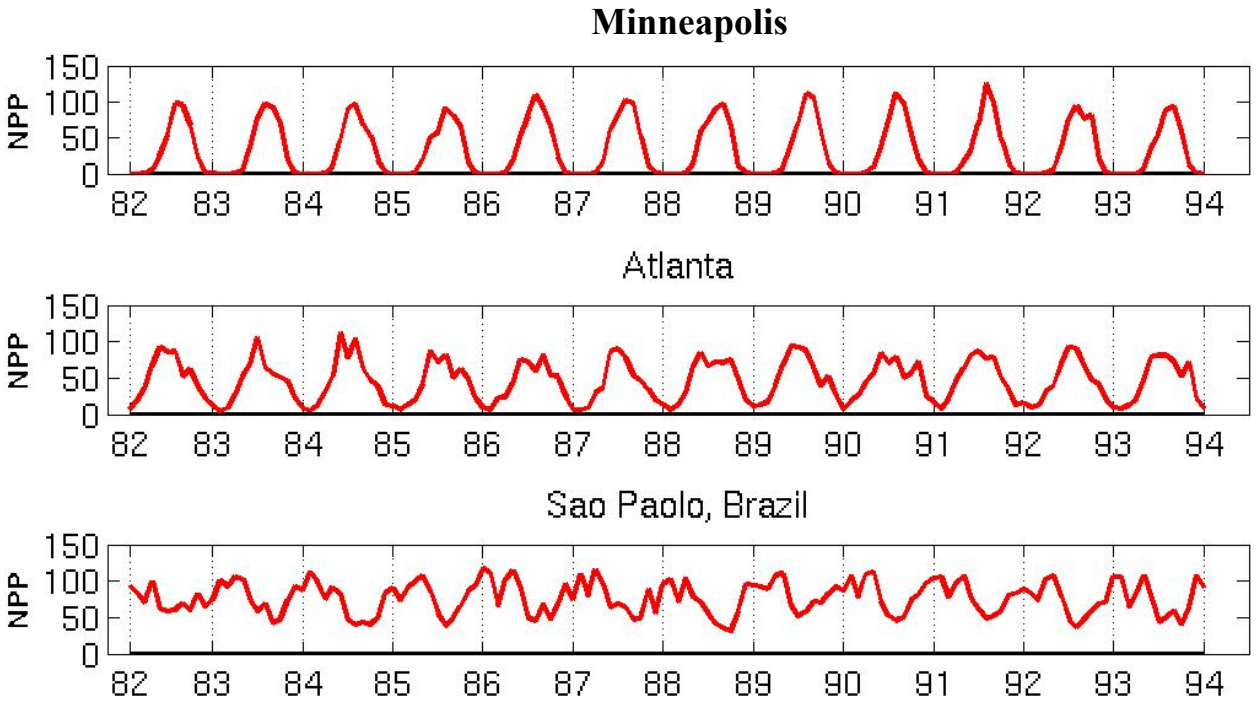
$$\frac{600 - 500}{100} = \frac{100}{100} = 1.0$$

Compare this with Gerald's standard score, which is

$$\frac{21 - 18}{6} = \frac{3}{6} = 0.5$$

Because Jennie's score is 1 standard deviation above the mean and Gerald's is only 0.5 standard deviation above the mean, Jennie's performance is better.

Example: Sample Time Series of Plant Growth

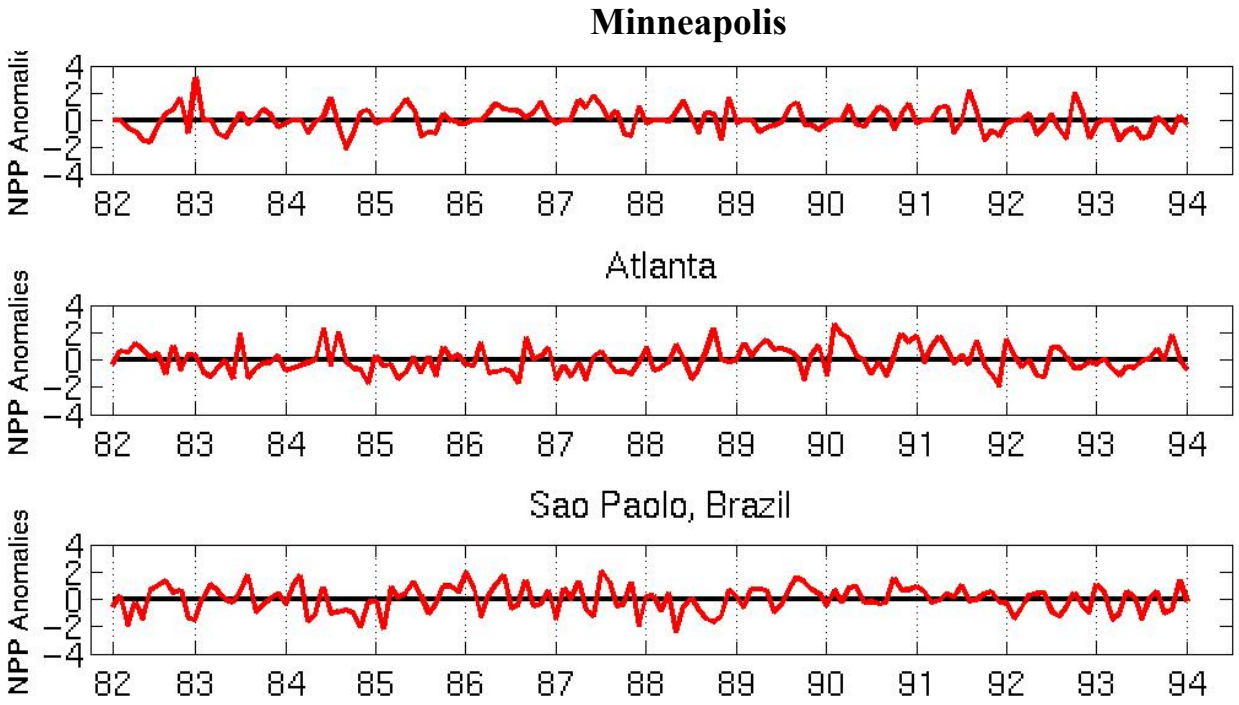


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

Seasonality Accounts for Much Correlation



Normalized using monthly Z Score:

Subtract off monthly mean and divide by monthly standard deviation

Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000

Outline of Second Lecture for Chapter 2

- Basics of Similarity and Dissimilarity Measures
- Distances and Their Properties
- Similarities and Their Properties
- Density

Similarity and Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

p and *q* are the corresponding attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

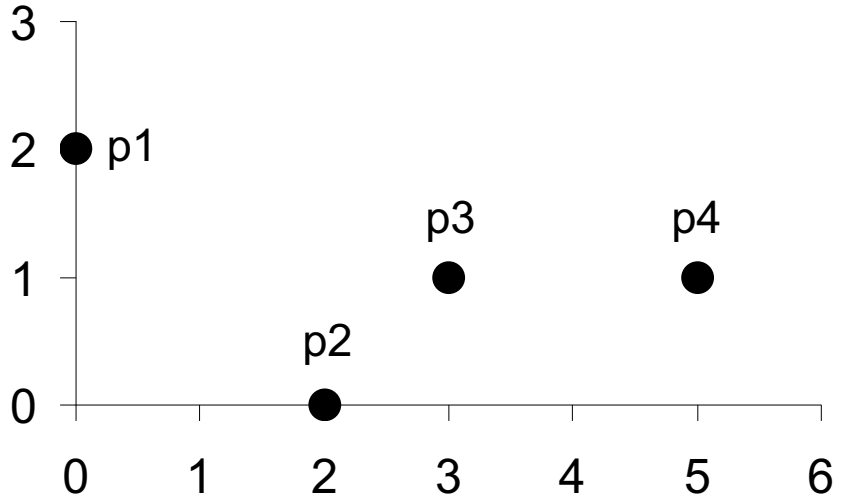
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

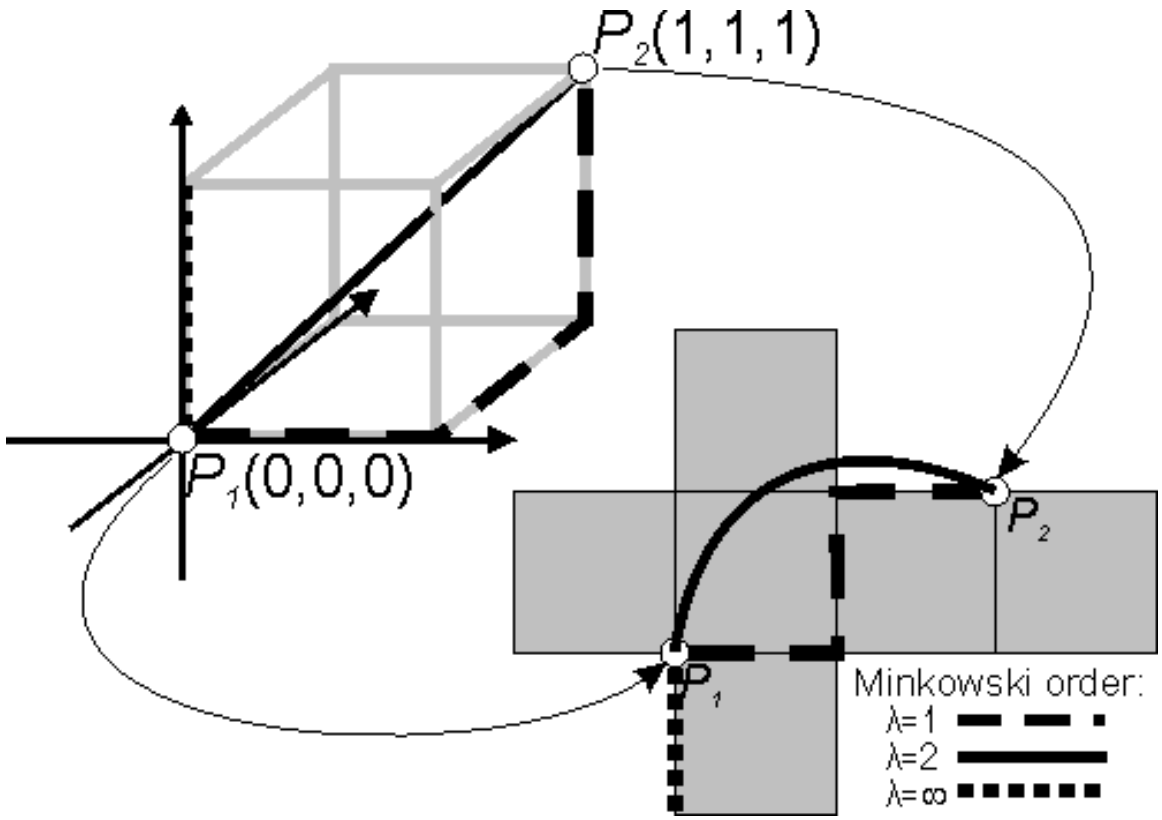
- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Minkowski Distance

- Minkowski Distance: $dist = (\sum_{k=1}^{n} | p_k - q_k |^r)^{\frac{1}{r}}$



Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

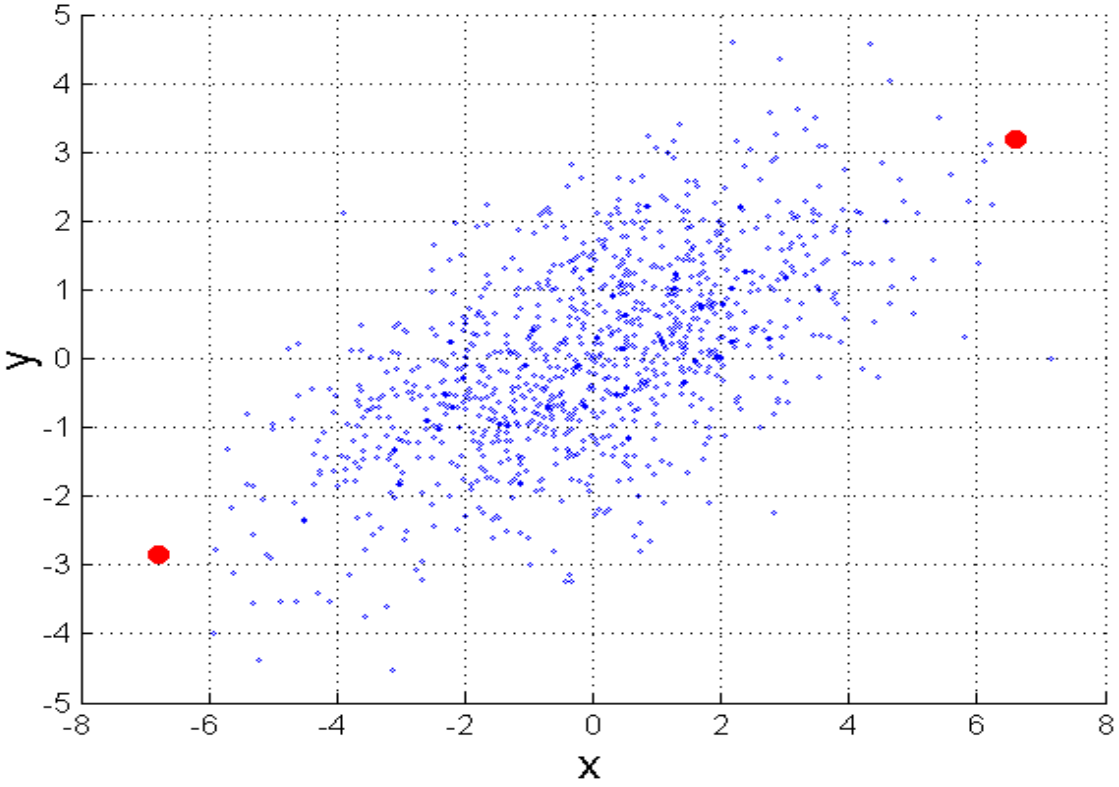
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Mahalanobis Distance

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$



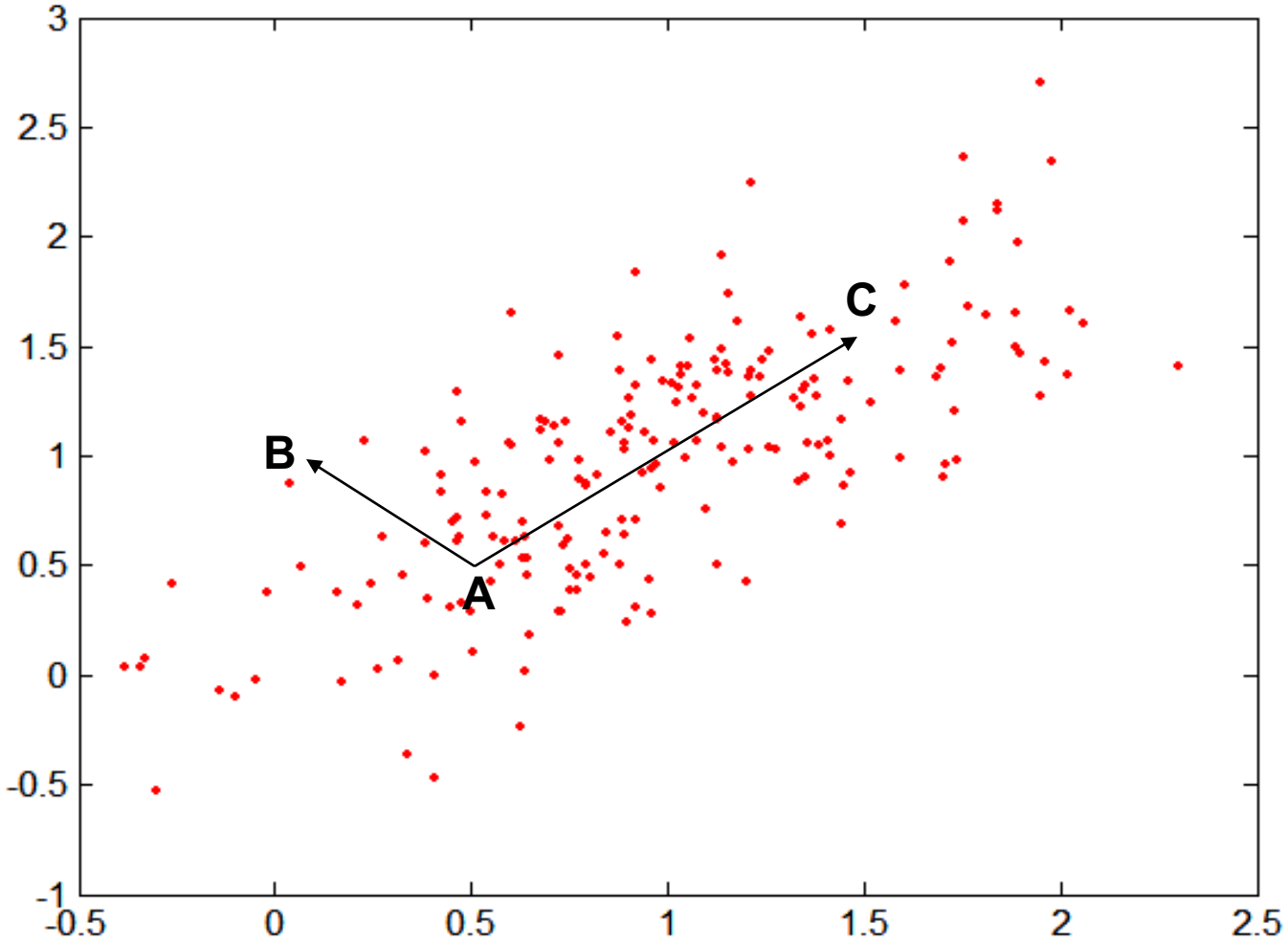
Σ is the covariance matrix of the input data *X*

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

Determining similarity of an unknown Sample set to a known one. It takes Into account the correlations of the Data set and is scale-invariant.

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

  1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
  2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
  3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

  where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes

- Compute similarities using the following quantities

  F01 = the number of attributes where *p* was 0 and *q* was 1

  F10 = the number of attributes where *p* was 1 and *q* was 0

  F00 = the number of attributes where *p* was 0 and *q* was 0

  F11 = the number of attributes where *p* was 1 and *q* was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  = (F11 + F00) / (F01 + F10 + F11 + F00)

  J = number of 11 matches / number of non-zero attributes

  = (F11) / (F01 + F10 + F11)

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

F01 = 2 (the number of attributes where *p* was 0 and *q* was 1)

F10 = 1 (the number of attributes where *p* was 1 and *q* was 0)

F00 = 7 (the number of attributes where *p* was 0 and *q* was 0)

F11 = 0 (the number of attributes where *p* was 1 and *q* was 1)

SMC = (F11 + F00) / (F01 + F10 + F11 + F00)

= (0+7) / (2+1+0+7) = 0.7

$J = (F_{11}) / (F_{01} + F_{10} + F_{11}) = 0 / (2 + 1 + 0) = 0$

Cosine Similarity

more like Jaccard coefficient

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

where $\bullet$ indicates vector dot product and $\| d \|$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$\|d_1\|$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481

$\|d_2\|$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

cos(d_1, d_2) = .3150

Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

prove this = Jaccard coefficient in binary distribution

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Correlation

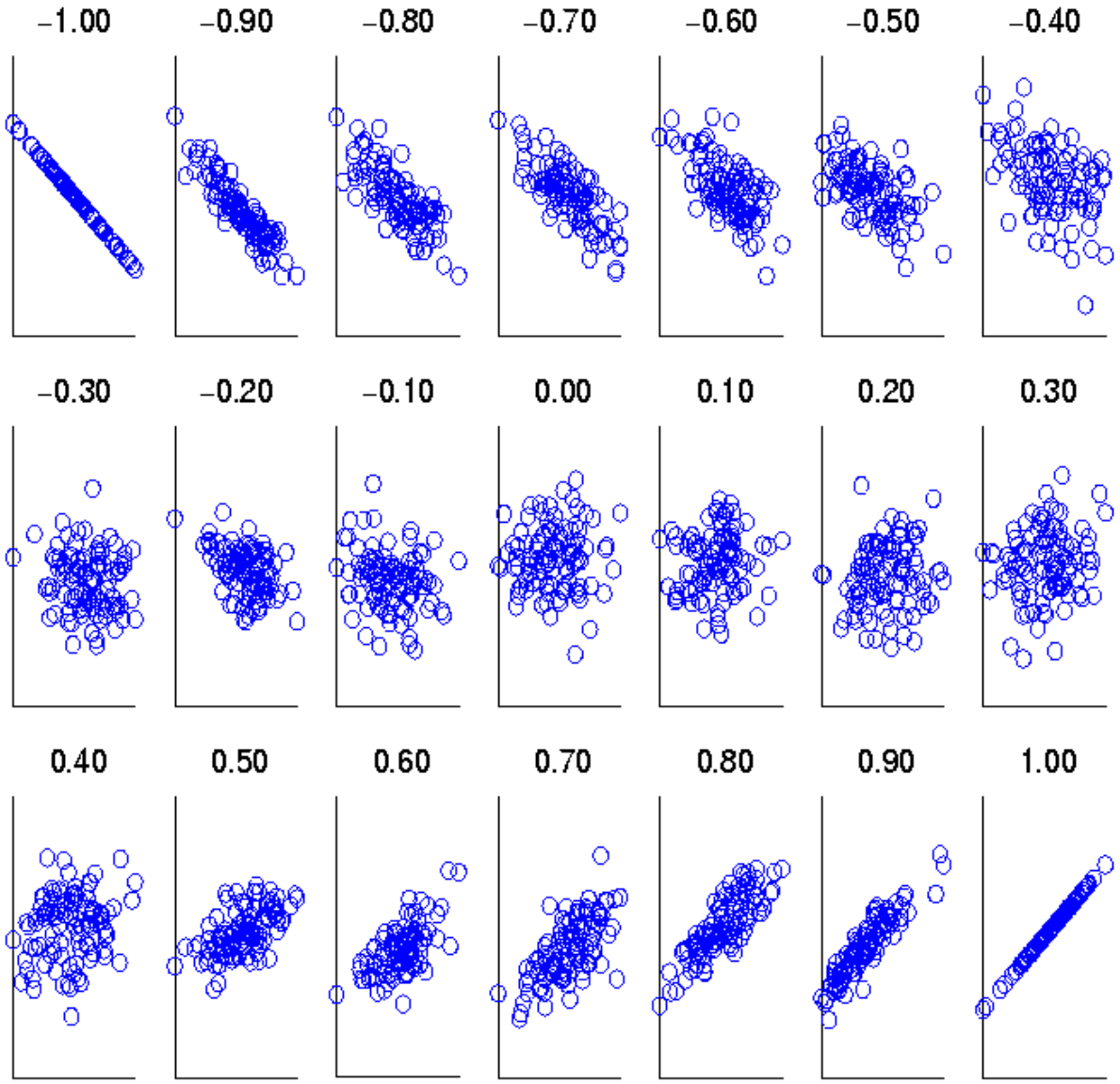
- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - \boldsymbol{mean}(p)) / \boldsymbol{std}(p)$$

$$q'_k = (q_k - \boldsymbol{mean}(q)) / \boldsymbol{std}(q)$$

$$correlation(p,q) = p' \bullet q' / (n-1)$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Drawback of Correlation

- X = (-3, -2, -1, 0, 1, 2, 3)
- Y = (9, 4, 1, 0, 1, 4, 9) $Y = X^2$

- Mean(X) = 0, Mean(Y) = 4

- Correlation

  = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5)

  = 0

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1: For the kth attribute, compute a similarity, $s_k(\mathrm{x}, \mathrm{y})$, in the range [0, 1].

2: Define an indicator variable, δ_k, for the kth attribute as follows:

δ_k = 0 if the kth attribute is an asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the kth attribute

δ_k = 1 otherwise

3. Compute $\mathrm{similarity}(\mathbf{x}, \mathbf{y}) = \dfrac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

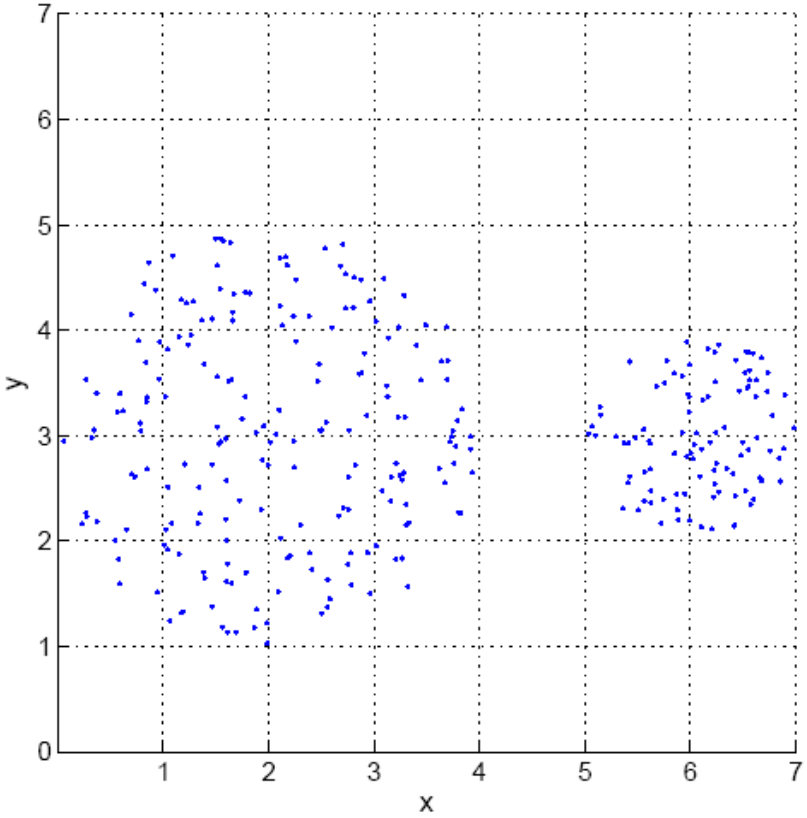
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

Density

- Measures the degree to which data objects are close to each other in a specified area
- The notion of density is closely related to that of proximity
- Concept of density is typically used for clustering and anomaly detection
- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
    - Estimate what the distribution of the data looks like
  - Graph-based density
    - Connectivity

Euclidean Density: Grid-based Approach

- Simplest approach is to divide region into a number of rectangular cells of equal volume and



Grid-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Counts for each cell.

Euclidean Density: Center-Based

- Euclidean density is the number of points within a specified radius of the point

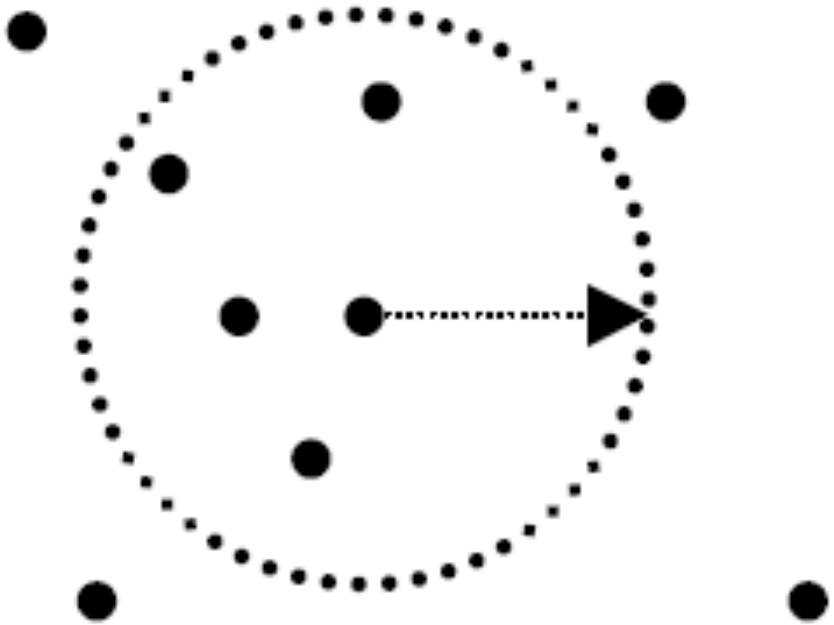


Illustration of center-based density.

Distance in High-Dimensional Space

- Theorem of Instability
  - The Euclidean distance from a point to its closest neighbor approaches the distance to its farthest neighbor, as dimension approaches infinity.
  - That is, for any ε >0, the maximum and minimum interpoint distance are arbitrarly close with Prob P=1

$$\lim_{d\to\infty} P\left(\frac{d_{max}}{d_{min}} - 1 \leq \epsilon\right) = 1.$$

  - Bellman's curve "empty space phenomenon", - the exponentially increase in hypervolume as dimensions are added. If the number of samples remains fixed, the density of samples must exponentially decrease.