Data Preprocessing

Dr. Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida, Gainesville

ranka@cise.ufl.edu

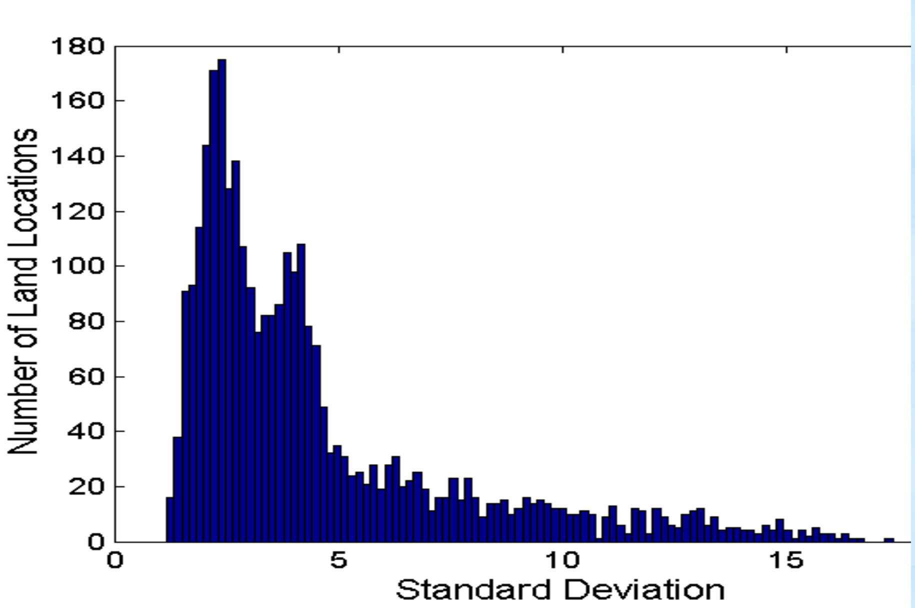
Data Preprocessing

- What preprocessing step can or should we apply to the data to make it more suitable for data mining?
  - Aggregation
  - Sampling
  - Dimensionality Reduction
  - Feature Subset Selection
  - Feature Creation
  - Discretization and Binarization
  - Attribute Transformation

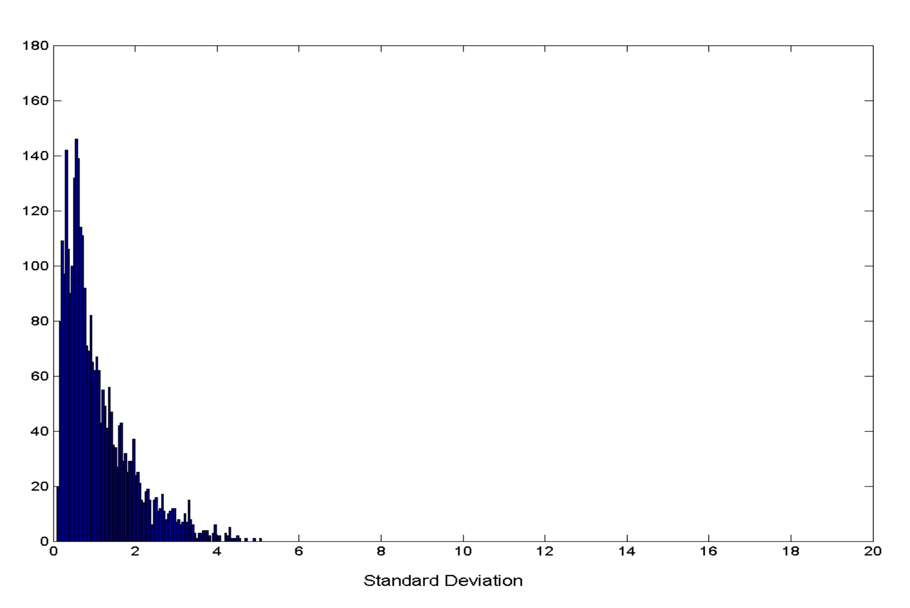
Aggregation

- Aggregation refers to combing two or more attributes (or objects) into a single attribute (or object)
- For example, merging daily sales figures to obtain monthly sales figures
- Why aggregation?
  - Data reduction
    - Allows use of more *expensive* algorithms
  - If done properly, aggregation can act as scope or scale, providing a high level view of data instead of a low level view

- Behavior of group of objects in more stable than that of individual objects
  - The aggregate quantities have less *variability* than the individual objects being aggregated



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

- Sampling is the process of understanding characteristics of data or models based on a subset of the original data. It is used extensively in all aspects of data exploration and mining
- Why sample
  - Obtaining the entire set of "data of interest" is too expensive or time consuming
  - Obtaining the entire set of data may not be necessary (and hence a waste of resources)

Representative Sample

- A sample is representative for a particular operation if it results in *approximately* the same outcome as if the entire data set was used
- A sample that may be representative for one operation, may not be representative for another operation
  - For example, a sample may be representative for histogram along one dimension but may not be good enough for correlation between two dimensions

Sampling Approaches

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement: Once an item is selected, it is removed from the population for obtaining future samples
  - Sampling with replacement: Selected item is *not* removed from the population for obtaining future samples

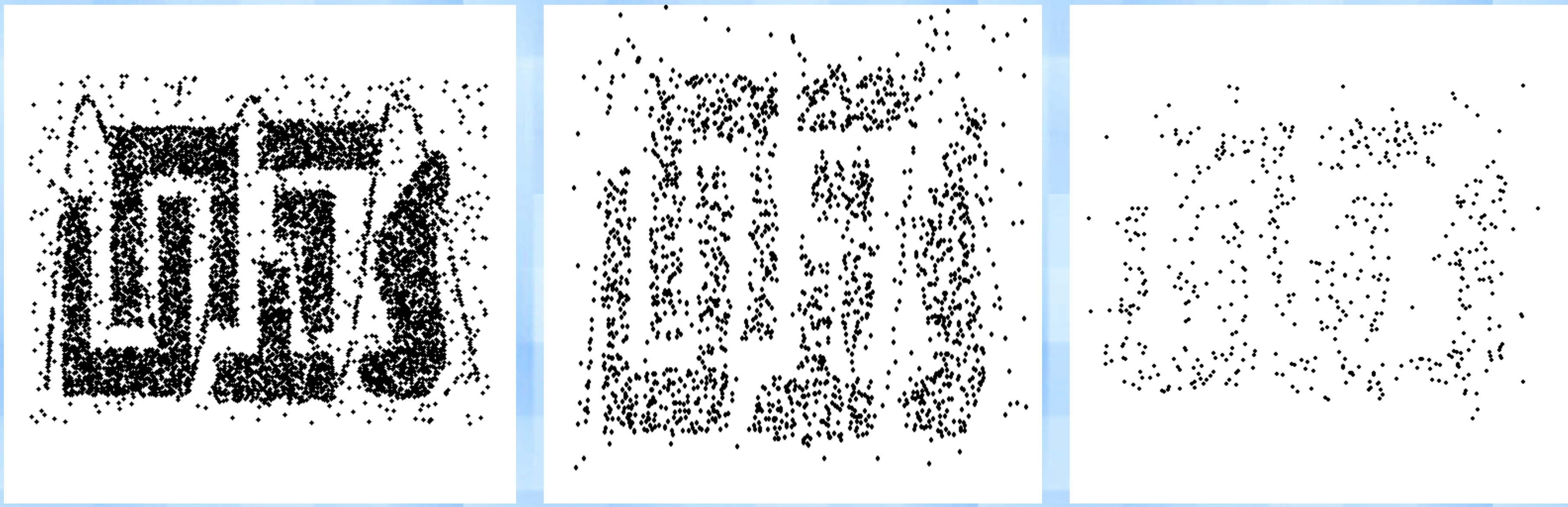
Sampling Approaches

- Stratified Sampling
  - When subpopulations vary considerably, it is advantageous to sample each subpopulation (stratum) independently
  - *Stratification* is the process of grouping members of the population into relatively homogeneous subgroups before sampling
  - The strata should be mutually exclusive : every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive : no population element can be excluded
  - Then random sampling is applied within each stratum. This often improves the representative-ness of the sample by reducing sampling error

Sample Size

- Even if proper sampling technique is known, it is important to choose proper sample size
- Larger sample sizes increase the probability that a sample will be *representative*, but also eliminate much of the advantage of sampling
- With smaller sample size, patterns may be missed or erroneous patters detected

Sample Size / Fraction



8000 points

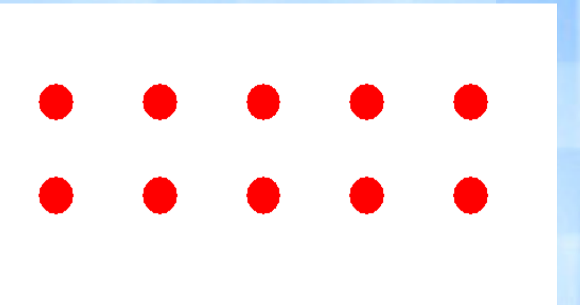
2000 points

500 points

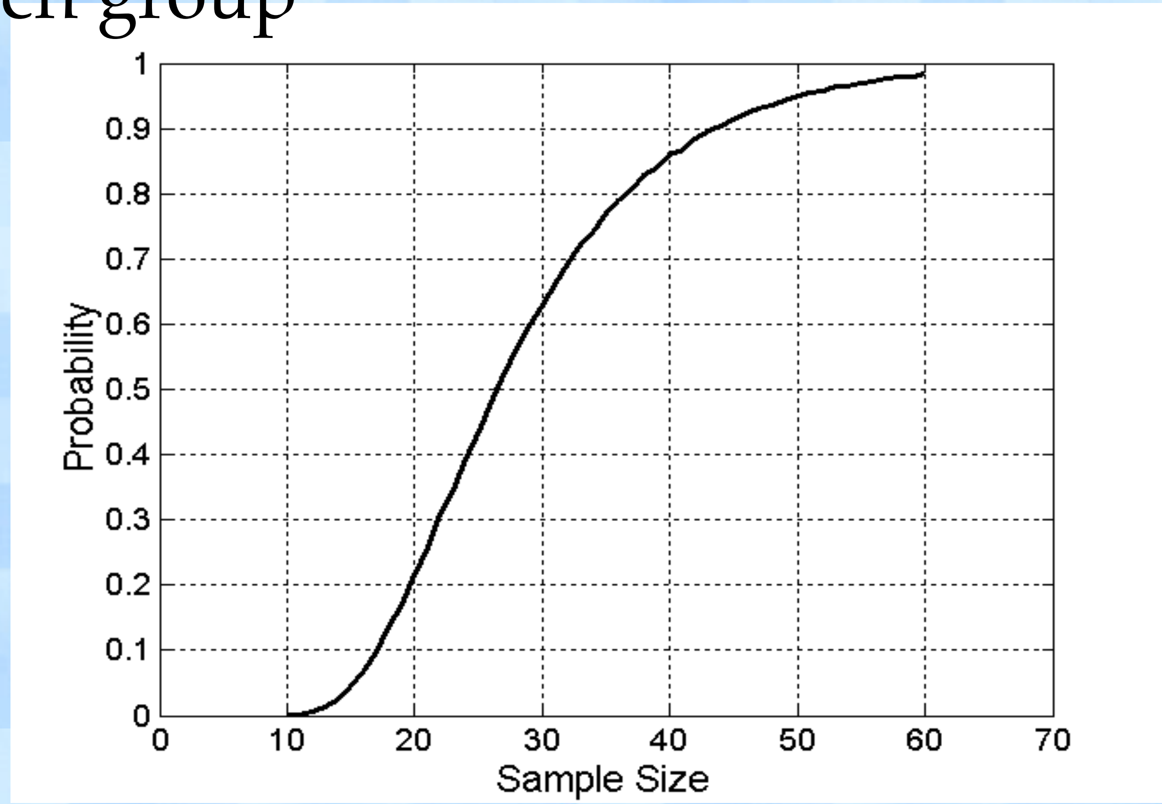
Example of the Loss of Structure with Sampling

Sample Size

- Sample size required to obtain at least one sample from each group



Ten Clusters

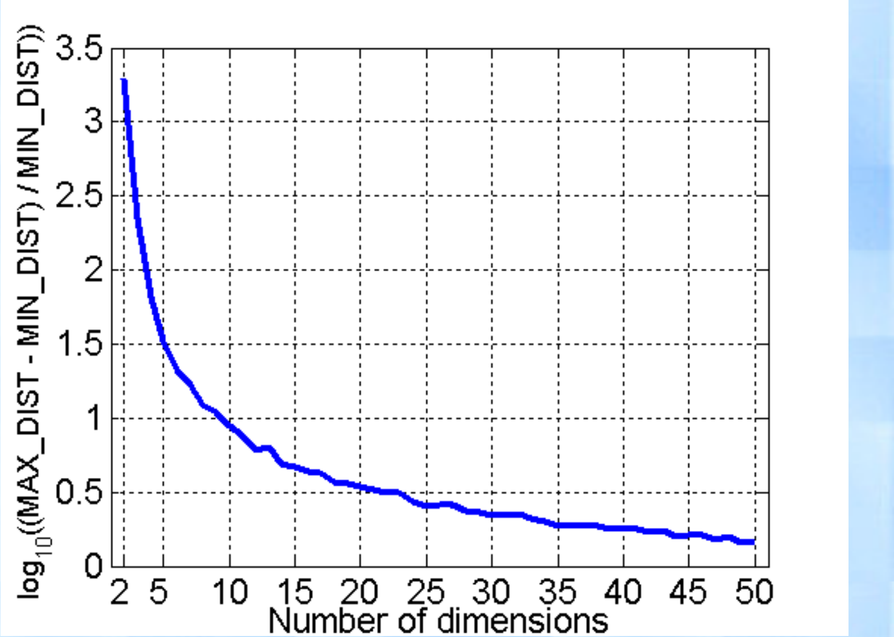


Probability a sample contains points from each of ten groups

Dimensionality Reduction

- Curse of dimensionality: Data analysis becomes significantly harder as the dimensionality of the data increases

Curse of Dimensionality



Decrease in the relative distance between points
As dimensionality increases

Dimensionality Reduction

- Determining dimensions (or combinations of dimensions) that are important for modeling
- Why dimensionality reduction?
  - Many data mining algorithms work better if the dimensionality of data (i.e. number of attributes) is lower
  - Allows the data to be more easily visualized
  - If dimensionality reduction eliminates irrelevant features or reduces noise, then quality of results may improve
  - Can lead to a more understandable model

Dimensionality Reduction

- *Redundant features* duplicate much or all of the information contained in one or more attributes
  - The purchase price of product and the sales tax paid contain the same information
- *Irrelevant features* contain no information that is useful for data mining task at hand
  - Student ID numbers would be irrelevant to the task of predicting their GPA

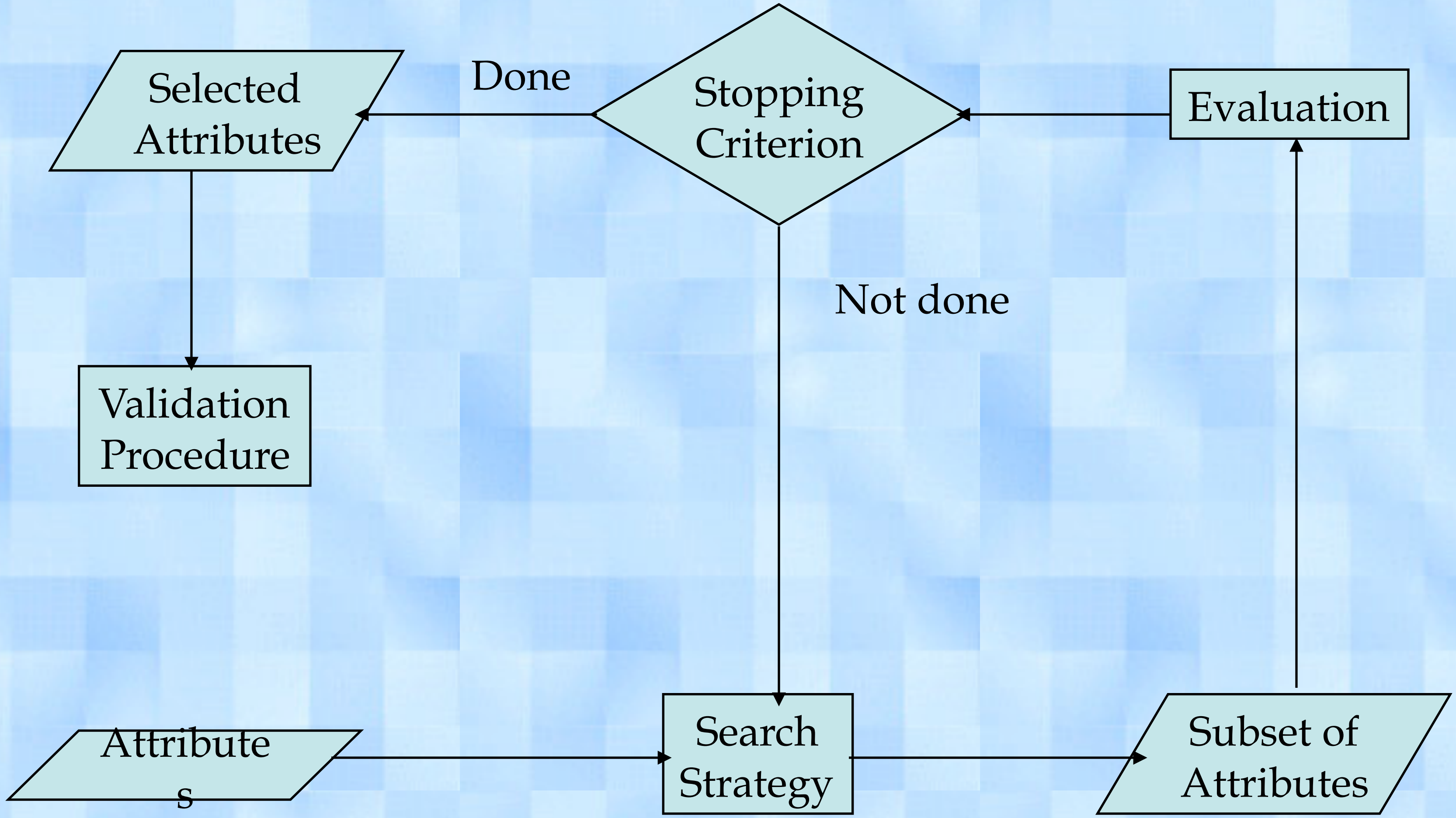
Dimensionality Reduction using Principal Component Analysis

- Some of the most common approaches for dimensionality reduction, particularly for continuous data, use a linear or non-linear projection of data from a high dimensional space to a lower dimensional space
- PCA is a linear algebra technique for continuous attributes that finds new attributes (principal components) that
  - Are linear combinations of original attributes
  - Are orthogonal to each other
  - Capture the maximum amount of variation in data

Dimensionality Reduction by Feature Subset Selection

- There are three standard approaches to feature selection:
  - Embedded approaches: Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches: Features are selected before the data mining algorithm is run
  - Wrapper approaches: Use the target data mining algorithm as a black box to find the best subset of attributes (typically without enumerating all subsets)

Architecture for Feature Subset Selection



Flowchart of a feature subset selection process

Feature Creation

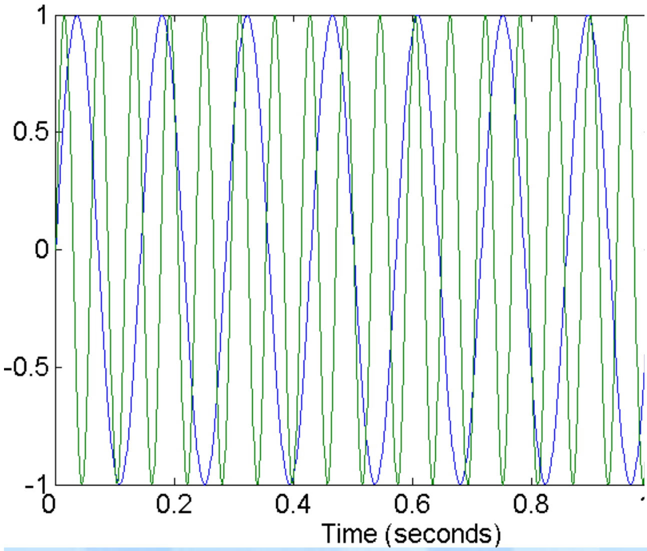
- Sometimes, a small number of *new* attributes can capture the important information in a data set much more efficiently than the original attributes
- Also, the number of new attributes can be often smaller than the number of original attributes. Hence, we get benefits of dimensionality reduction
- Three general methodologies:
  - Feature Extraction
  - Mapping the Data to a New Space
  - Feature Construction

Feature Extraction

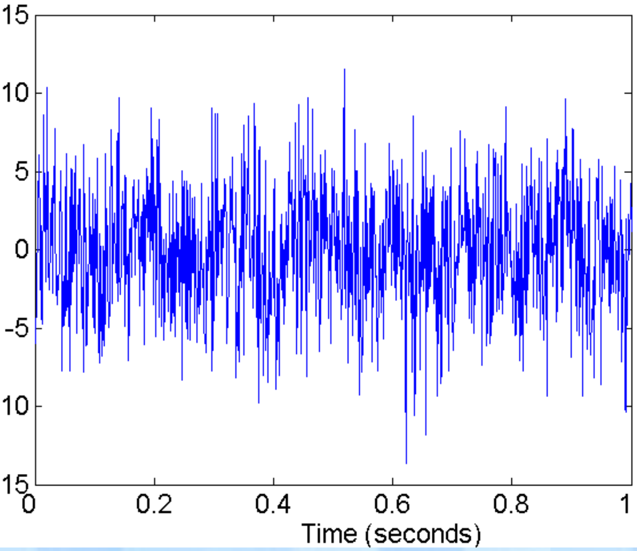
- One approach to dimensionality reduction is feature extraction, which is creation of a new, smaller set of features from the original set of features
- For example, consider a set of photographs, where each photograph is to be classified whether its human face or not
- The raw data is set of pixels, and as such is not suitable for many classification algorithms
- However, if data is processed to provide high-level features like presence or absence of certain types of edges or areas correlated with presence of human faces, then a broader set of classification techniques can be applied to the problem

Mapping the Data to a New Space

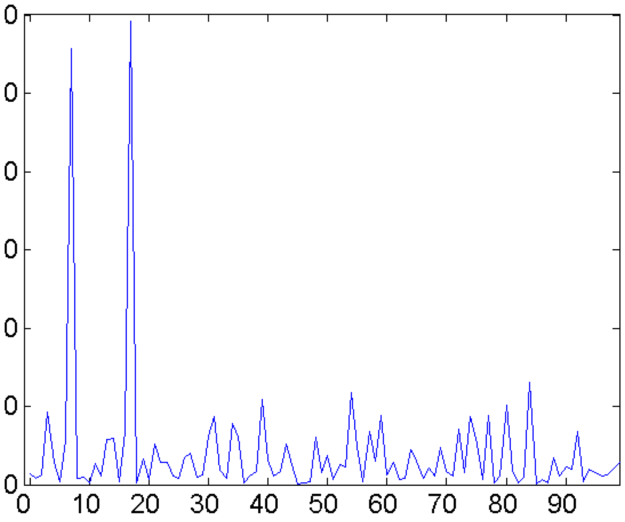
- Sometimes, a totally different view of the data can reveal important and interesting features
- Example: Applying Fourier transformation to data to detect time series patterns



Original Time Series



Time Series with noise



Frequency plot

Feature Construction

- Sometimes features have the necessary information, but not in the form necessary for the data mining algorithm. In this case, one or more new features constructed out of the original features may be useful
- Example, there are two attributes that record volume and mass of a set of objects
- Suppose there exists a classification model based on material of which the objects are constructed
- Then a density feature constructed from the original two features would help classification

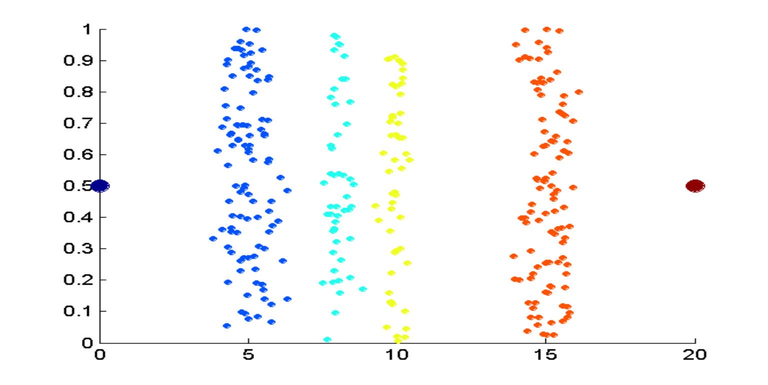
Discretization and Binarization

- Discretization is the process of converting a continuous attribute to a discrete attribute
- A common example is rounding off real numbers to integers
- Some data mining algorithms require that the data be in the form of categorical or binary attributes. Thus, it is often necessary to convert continuous attributes in to categorical attributes and / or binary attributes
- Its pretty straightforward to convert categorical attributes in to discrete or binary attributes

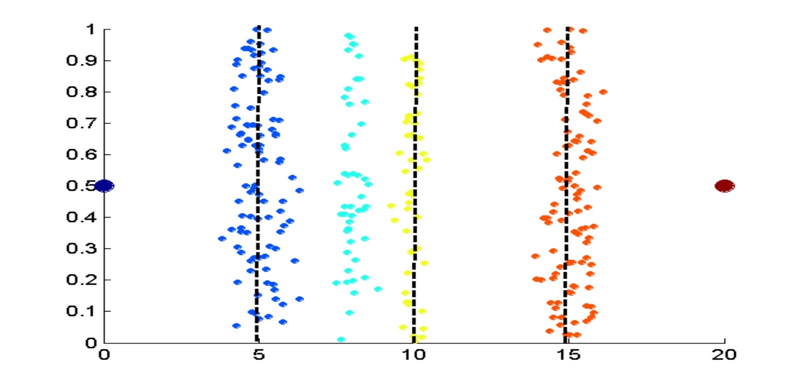
Discretization of Continuous Attributes

- Transformation of continuous attributes to a categorical attributes involves
  - Deciding how many categories to have
  - How to map the values of the continuous attribute to categorical attribute
- A basic distinction between discretization methods for classification is whether class information is used (supervised) or not (unsupervised)

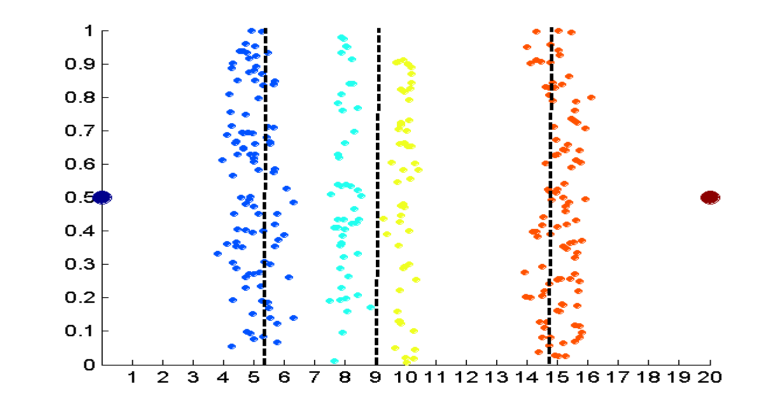
Different Discretization Techniques



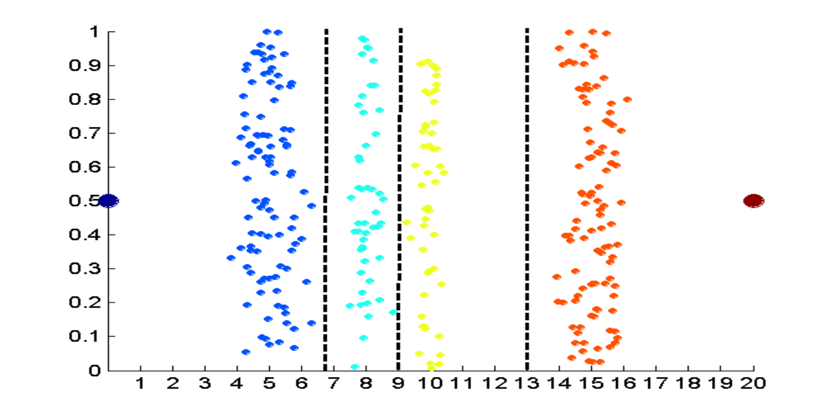
Data



Equal interval width



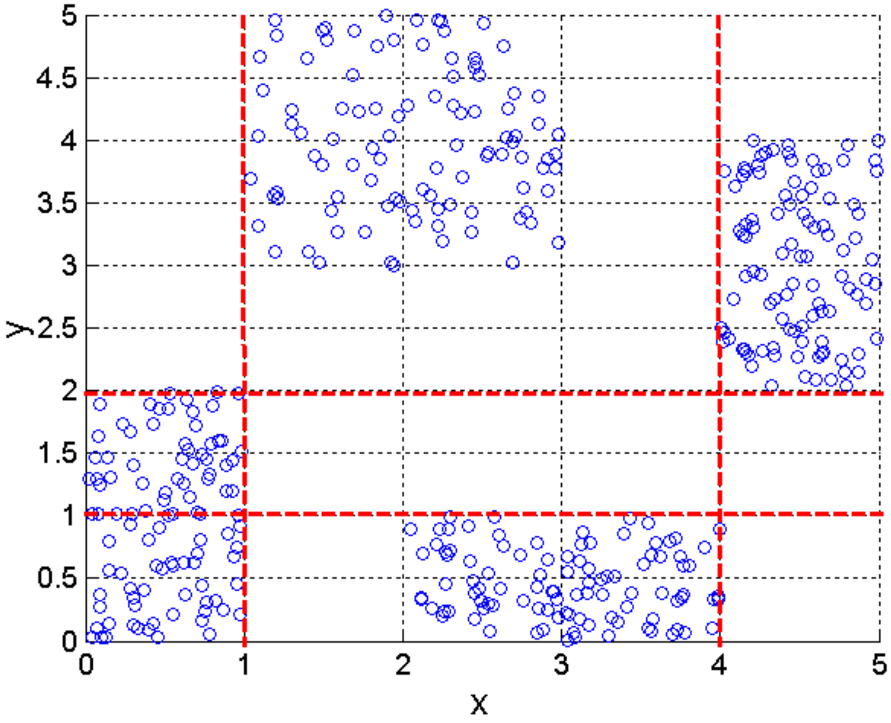
Equal frequency



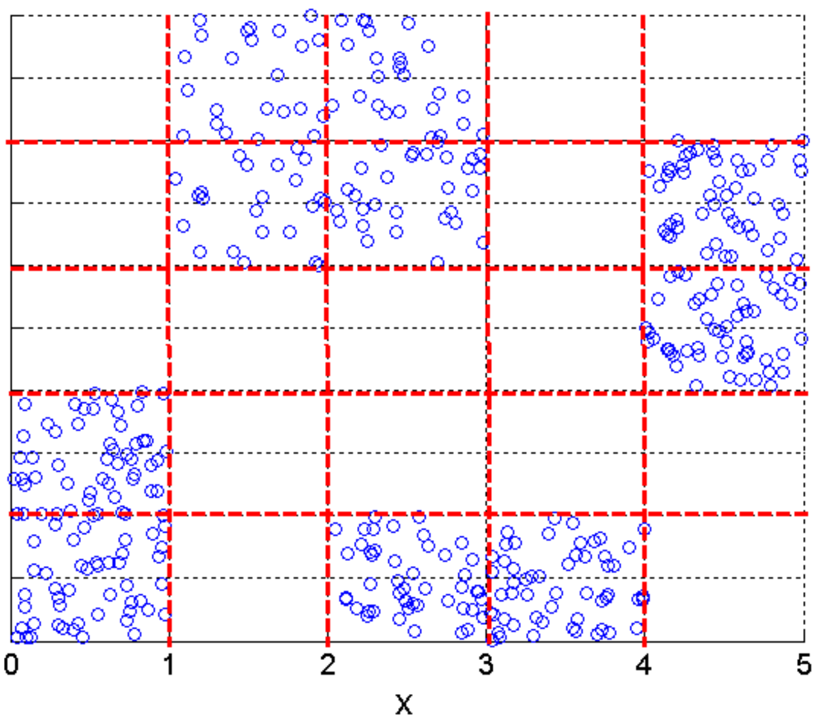
K-means

Different Discretization Techniques

- Entropy based approach



3 categories for both *x* and *y*



5 categories for both *x* and *y*

Attribute Transformation

- An *attribute transformation* refers to a transformation that is applied to all values of an attribute, i.e., for each object, the transformation is applied to the value of the attribute for that object
- There are two important types of attribute transformations
  - Simple function transformations
    - Example: x^k, log x, e^x, sqrt x, etc.
  - Standardization or normalization