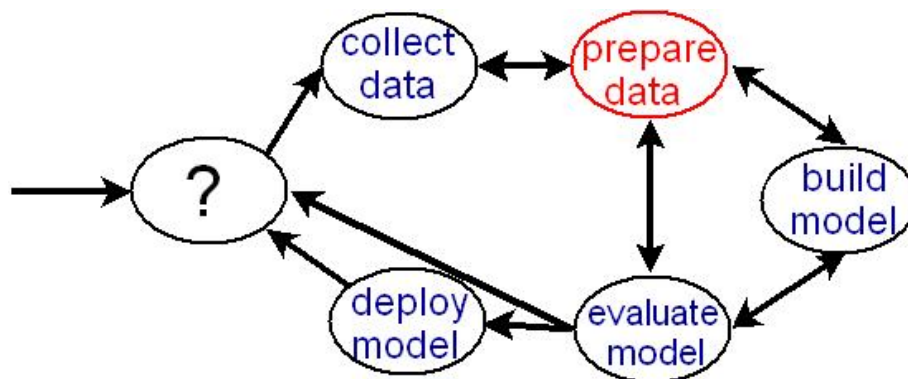


Data Mining

COIS 4400H/AMOD 5400H

Preprocessing (Chapter 2)

Data Preparation



Data Preprocessing

Aggregation

Sampling

Feature subset selection

Discretization and binarization

Dimensionality reduction

Attribute creation and transformation

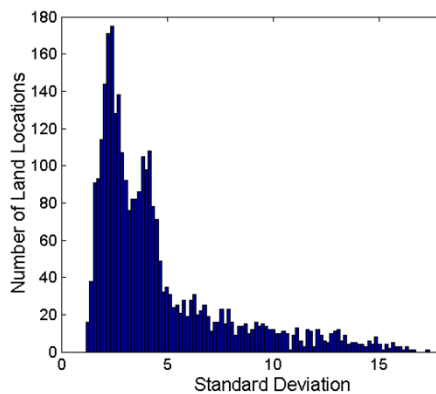
Data quality issues

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

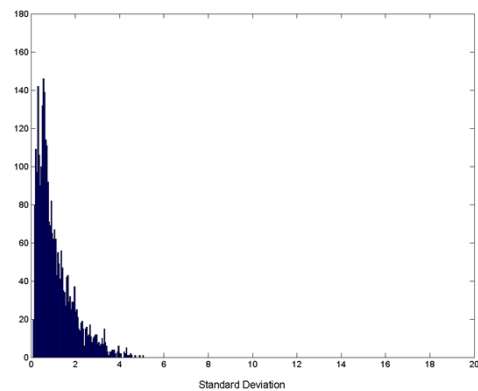
Data reduction: Reduce the number of attributes or objects

Change of scale: Cities aggregated into regions, states, countries, etc

More “stable” data : Aggregated data tends to have less variability



**Standard Deviation of Average
Monthly Precipitation**



**Standard Deviation of Average
Yearly Precipitation**

Sampling Techniques

Simple Random Sampling

each object selected with equal probability

Sampling without replacement

remove object from sample if selected

Sampling with replacement

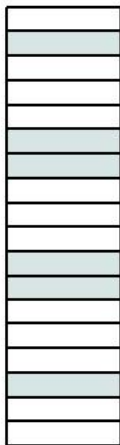
do not remove object from sample if selected

Stratified sampling

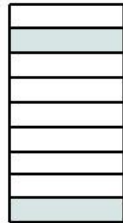
split the data into several partitions
random samples from each partition

Data preparation: sampling

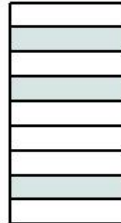
original data



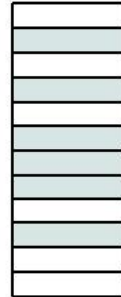
random sample



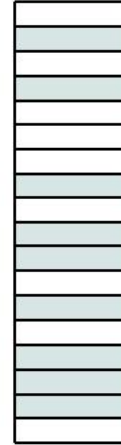
stratified sample



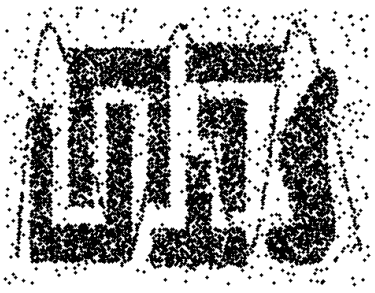
undersampling



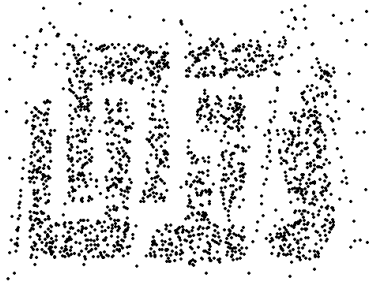
oversampling



Sample Size



8000 points



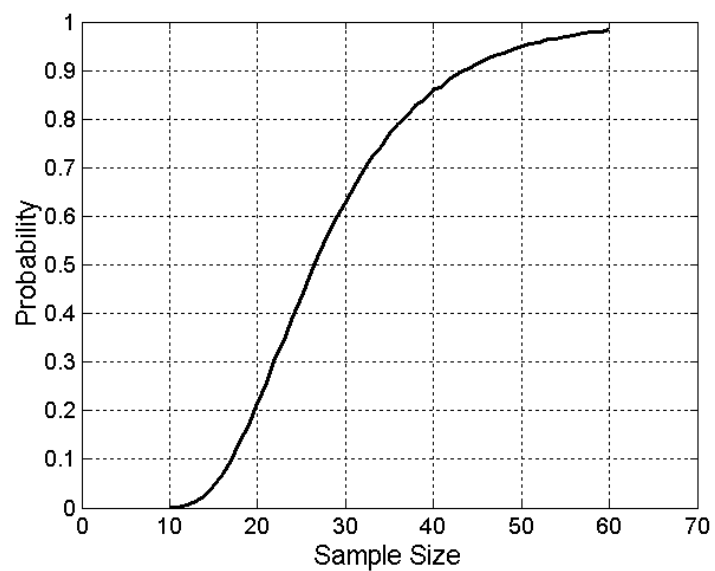
2000 Points



500 Points

Sample Size

What sample size is necessary to get at least one object from each of 10 groups?



Feature Subset Selection

remove redundant features

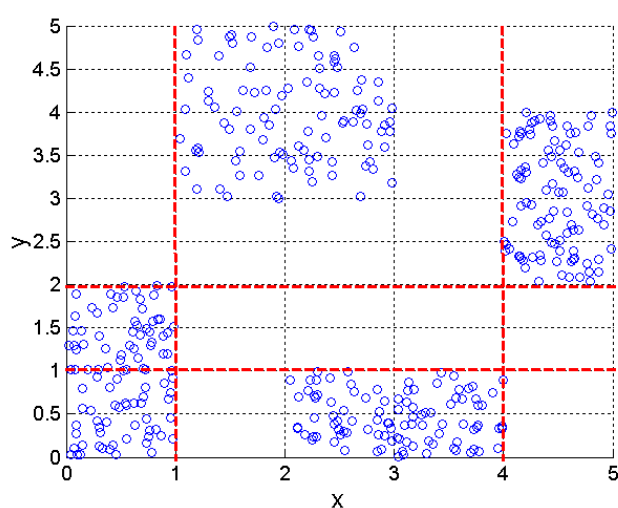
- duplicate much or all of the information contained in one or more other attributes
- Example: purchase price of a product and the amount of sales tax paid

remove irrelevant features

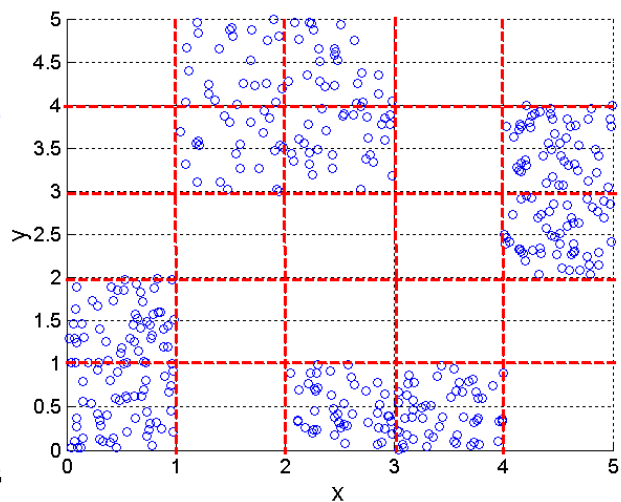
- contain no information that is useful for the data mining task at hand
- Example: students' ID is often irrelevant to the task of predicting students' GPA

Discretization Using Class Labels

(so we're assuming we know what the groups are)

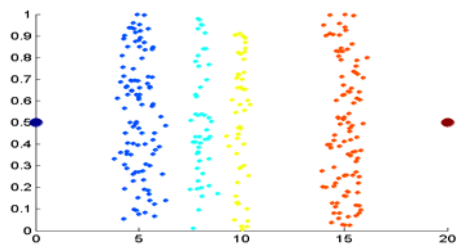


3 categories for both x and y

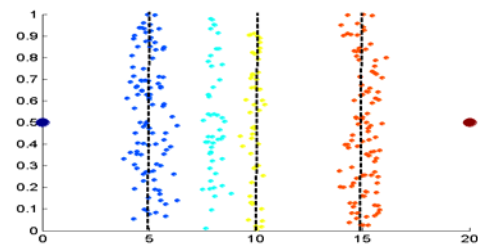


5 categories for both x and y

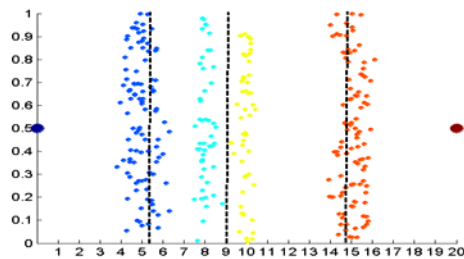
Discretization Without Using Class Labels



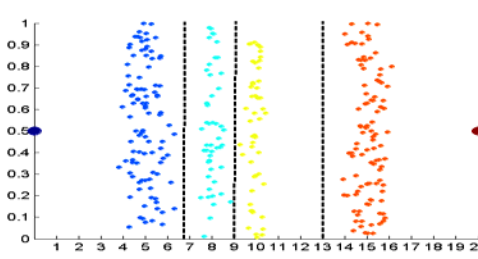
Data



Equal interval width



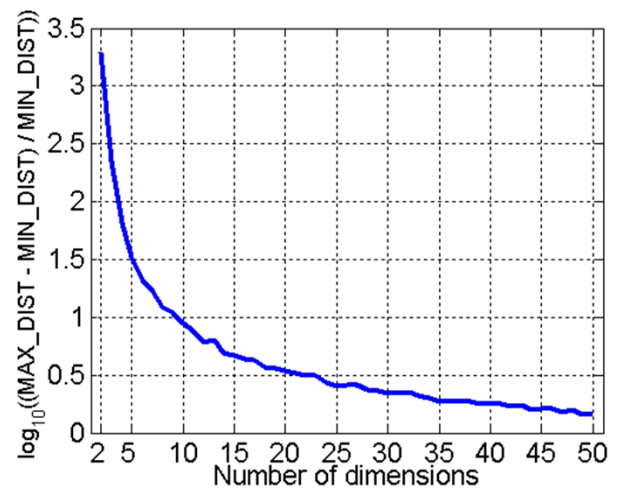
Equal frequency



K-means

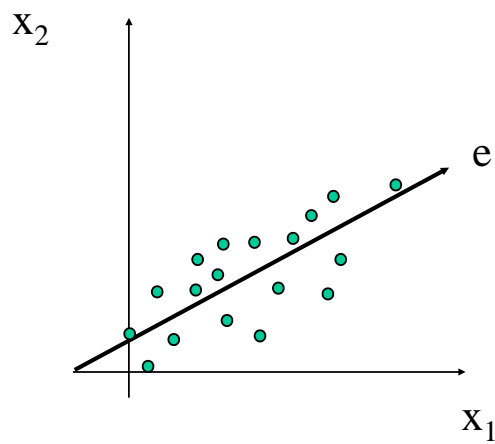
Curse of Dimensionality

When dimensionality increases,
data becomes increasingly sparse
in the space that it occupies



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

Dimensionality Reduction: PCA



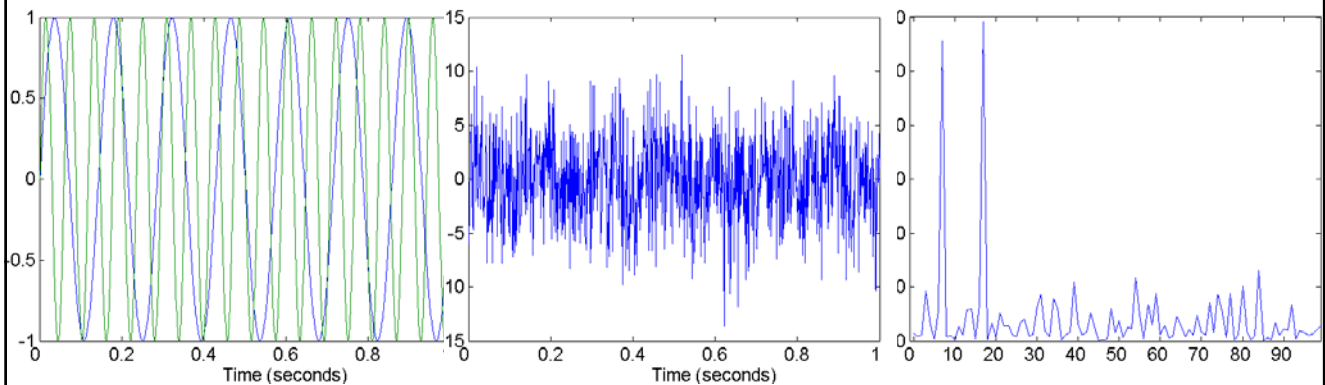
Dimensionality Reduction: PCA

Dimensions = 206



Feature Creation: Mapping Data to a New Space

one possibility: Fourier transform



Two Sine Waves

Two Sine Waves + Noise

Frequency

Attribute Transformation

- categorical to numeric
- numeric to categorical: see binning and discretization
- Principal Component Analysis
- Normalization and Standardization

Transformation: categorical to numeric

map to circle, sphere or hypersphere

- may work if categories are ordinal (i.e. days of the week)
- usually produces poor results otherwise

map to generalized tetrahedron:

- to uniquely represent k possible attribute values, we need k new attributes.
- works for both ordinal and nominal data

Data normalization

min-max normalization

z-score normalization (standardization)

normalization by decimal scaling