"Artificial Intelligence and Machine Learning"
BECE309L

Module - 5
Data Preparation for Machine Learning

Dr. Rabindra Kumar Singh
Associate Professor(Sr.)
School of Computer Science and Engineering
VIT - Chennai

Module - 5 Content :

- Basics of Vectors & Matrices
- Overview: Data Cleaning, Integration, Transformation & Reduction.

Vectors and Matrices :

- **Definition of Vector**: A collection of complex or real numbers, generally put in a column

$$\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} = [v_1 \cdots v_N]^T$$

Transpose

- **Vector Addition**: Add element-by-element

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix} \qquad \mathbf{a} + \mathbf{b} = \begin{bmatrix} a_1 + b_1 \\ \vdots \\ a_N + b_N \end{bmatrix}$$

- **Scalar** : A real or complex number.
- **Multiplying a Vector by a Scalar**

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \qquad \alpha\mathbf{a} = \begin{bmatrix} \alpha a_1 \\ \vdots \\ \alpha a_N \end{bmatrix}$$

- **Vector Space** : A set V of N-dimensional vectors (with a corresponding set of scalars) such that the set of vectors is:
  - "closed" under vector addition
  - "closed" under scalar multiplication
- **Matrix**: Is an array of (real or complex) numbers organized in rows and columns. Here is a 3x4 example:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$

- We'll sometimes view a matrix as being built from its columns; The 3x4 example above could be written as:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \mid \mathbf{a}_2 \mid \mathbf{a}_3 \mid \mathbf{a}_4 \end{bmatrix} \qquad \mathbf{a}_k = \begin{bmatrix} a_{1k} & a_{2k} & a_{3k} \end{bmatrix}^T$$

- We'll take two views of a matrix:
  - "Storage"for a bunch of related numbers (e.g., Cov. Matrix)
  - A transform (or mapping, or operator) acting on a vector(e.g., discrete Fourier transform(DFT), observation matrix, etc.... as we'll see
- **Matrix as Transform**: Our main view of matrices will be as "operators"that transform one vector into another vector.
- Consider the 3x4 example matrix above. We could use that matrix to transform the 4-dimensional vector vinto a 3-dimensional vectoru:

$$\mathbf{u} = \mathbf{A}\mathbf{v} = \left[\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \mathbf{a}_3 \,|\, \mathbf{a}_4\right] \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + v_3\mathbf{a}_3 + v_4\mathbf{a}_4$$

Clearly **u** is built from the columns of matrix **A**; therefore, it must lie in the span of the set of vectors that make up the columns of **A**.

Note that the columns of **A** are 3-dimensional vectors… so is **u.**

Data Pre-processing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called samples , examples, instances, data points, objects, tuples.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - E.g., customer_ID, name, address
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal**: categories, states, or "names of things"
  - Hair_color = auburn, black, blond, brown, grey, red, white
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
  - **Ordinal**
    - Values have a meaningful order (ranking) but magnitude between successive values is not known.
    - Size = small, medium, large, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., temperature in C° or F°, calendar dates
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
    - e.g., temperature in Kelvin, length, counts, monetary quantities

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., 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

Why Data Preprocessing?

Data in the real world is dirty

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“ ”
- noisy: containing errors or outliers
  - e.g., Salary=“-10”
- inconsistent: containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view: Accuracy, Completeness, Consistency, Timeliness, Believability, Value added, Interpretability, Accessibility
- **Broad categories**: Intrinsic, contextual, representational, and accessibility

Major Tasks in Data Preprocessing

- **Data cleaning** -Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration** - Integration of multiple databases, data cubes, or files
- **Data transformation** -Normalization and aggregation
- **Data reduction** - Obtains reduced representation in volume but produces the same or similar analytical results
- **Data discretization** - Part of data reduction but with particular importance, especially for numerical data

Data Pre-processing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction

Mining Data Descriptive Characteristics

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for grouped data):
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula(unimodal): $mean - mode = 3(mean - median)$

Data Pre-processing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction

Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into equal-frequency (equi-depth) bins:
  - **Bin 1**: 4, 8, 9, 15
  - **Bin 2**: 21, 21, 24, 25
  - **Bin 3**: 26, 28, 29, 34
- Smoothing by bin means:
  - **Bin 1**: 9, 9, 9, 9
  - **Bin 2**: 23, 23, 23, 23
  - **Bin 3**: 29, 29, 29, 29
- Smoothing by bin boundaries:
  - **Bin 1**: 4, 4, 4, 15
  - **Bin 2**: 21, 21, 25, 25
  - **Bin 3**: 26, 26, 26, 34

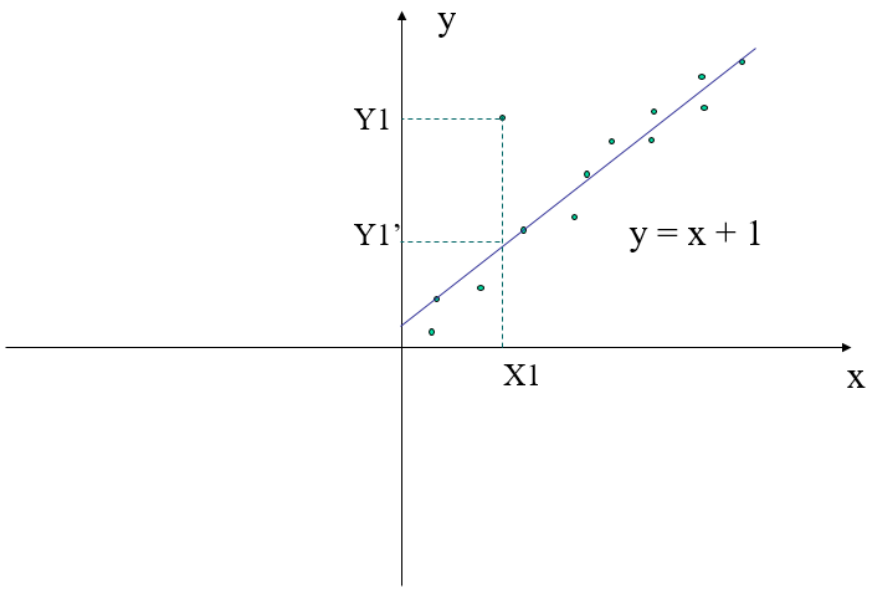
Problem

Problem

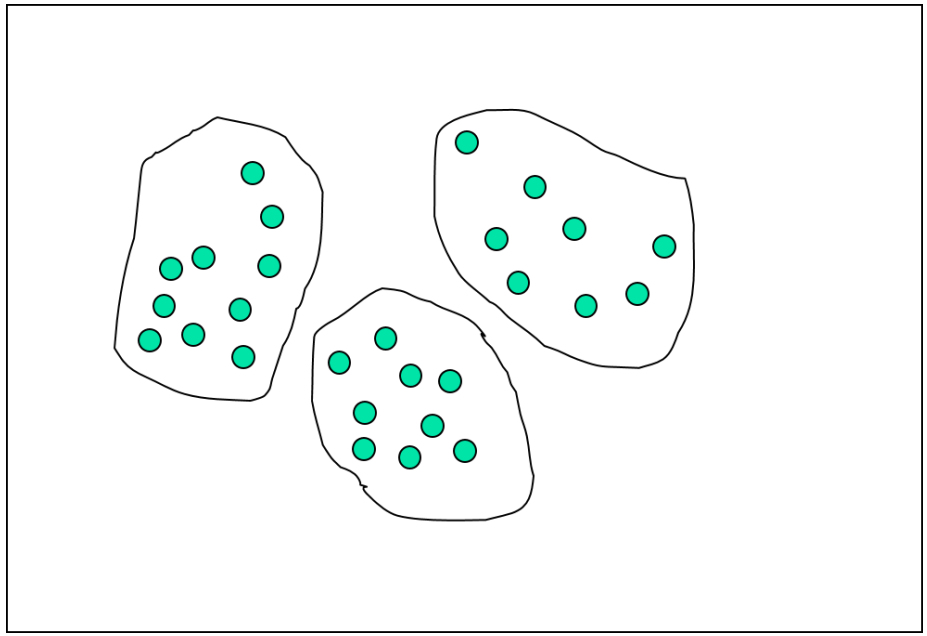
Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45,46, 52, 70.

- Use smoothing by bin means to smooth the above data, using a bin depth of 3.

Regression



Cluster Analysis



Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface

Data Pre-processing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: **e.g., A.cust-id ≡ B.cust - #**
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification:* The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

  $$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - A)(b_i - B)}{N\sigma_A\sigma_B}$$

  where N is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B.
- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated

Correlation Analysis (Categorical Data)

- χ^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\chi 2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})2}{e_{ij}}$$

- O_{ij} is the observed frequency of the joint event(A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j) which is calculated as

$$e_{ij} = \frac{count(A = a_i) Xcount(B = b_j)}{N}$$

- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	male	female	Sum (row)
Fiction	250(90)	200(360)	450
not fiction	50(210)	1000(840)	1050
sum(col.)	300	1200	1500

- $e_{11} = \frac{count(male) X count(fiction)}{N} = \frac{300X450}{1500} = 90$
- χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

- Min-max normalization: to [new_min_A, new_max_A]
  $v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$
  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$
- Z-score normalization (μ : mean, σ : standard deviation):
  $v' = \frac{v - \mu_A}{\sigma_A}$
  - Ex. Let $\mu = 54{,}000$, $\sigma = 16{,}000$. Then, $\frac{73{,}000 - 54{,}000}{16{,}000} = 1.225$
- Normalization by decimal scaling
  $\frac{v'}{10^j}$ Where j is the smallest integer such that Max(|v'|) < 1

Problem

Problem

Use the two methods below to normalize the following group of data:
200; 300; 400; 600; 1000 (a) min-max normalization by setting min = 0 and max = 1
(b) z-score normalization

(a) min-max normalization by setting min = 0 and max = 1
original data: 200 300 400 600 1000
$[0, 1]$ normalized: 0 0.125 0.25 0.5 1
(b) z-score normalization
original data: 200 300 400 600 1000
z-score: -1.06 -0.7 -0.35 0.35 1.78 2.12

Data Pre-processing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction

Data Reduction Strategies

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation:
  - Attribute subset selection
  - Dimensionality reduction — e.g., remove unimportant attributes
  - Data Compression
  - Numerosity reduction — e.g., fit data into models
  - Discretization and concept hierarchy generation

Data Cube Aggregation

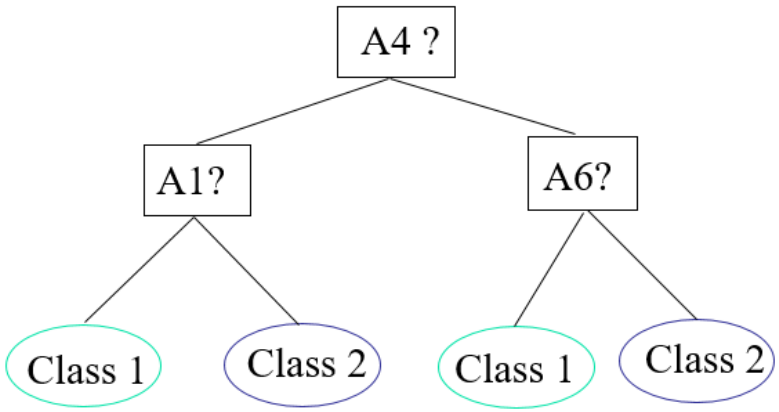
- The lowest level of a data cube (base cuboid)
  - The aggregated data for an individual entity of interest
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features.
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Decision-tree induction

Example of Decision Tree Induction

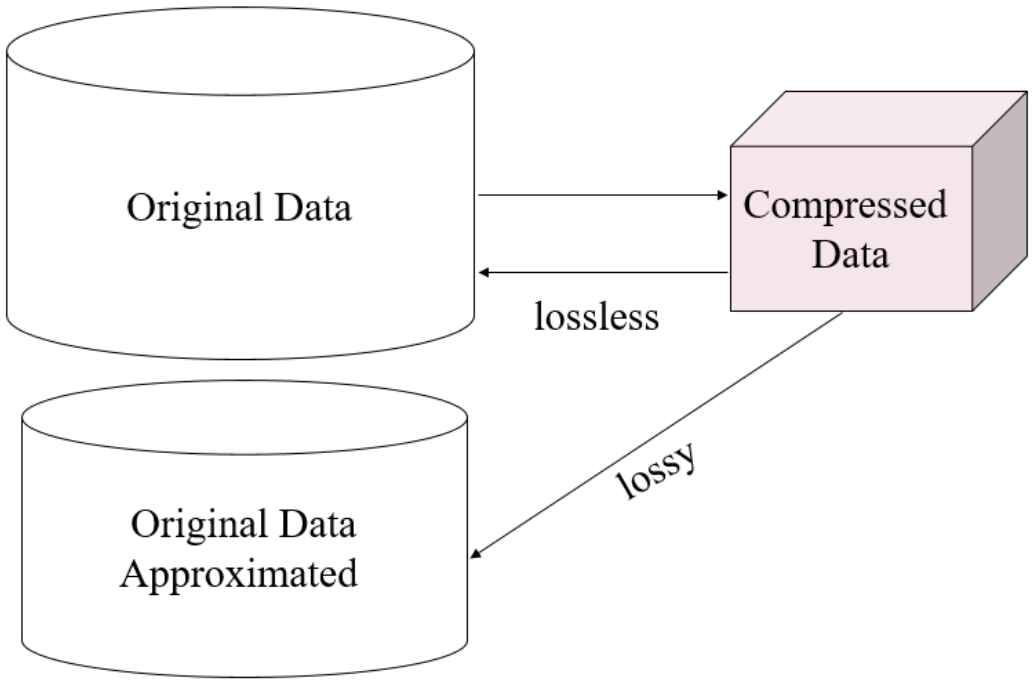
- Initial attribute set: {A1, A2, A3, A4, A5, A6 }



- —-> Reduced attribute set: {A1, A4, A6 }

Data Compression

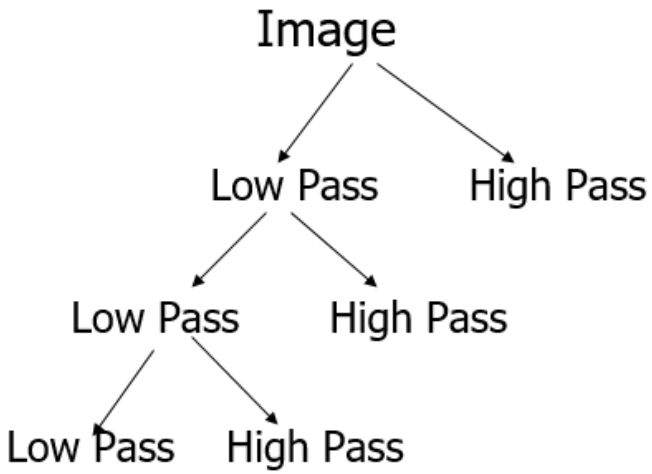
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

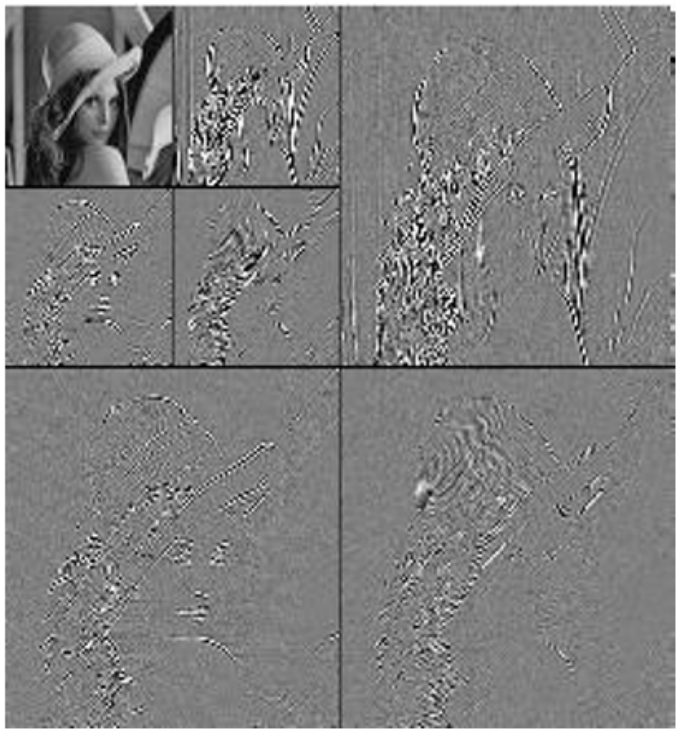


Dimensionality Reduction: Wavelet Transformation

- Discrete wavelet transform (DWT): linear signal processing, multi-resolutional analysis
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
  - Length, L, must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two set of data of length L/2
  - Applies two functions recursively, until reaches the desired length

DWT for Image Compression

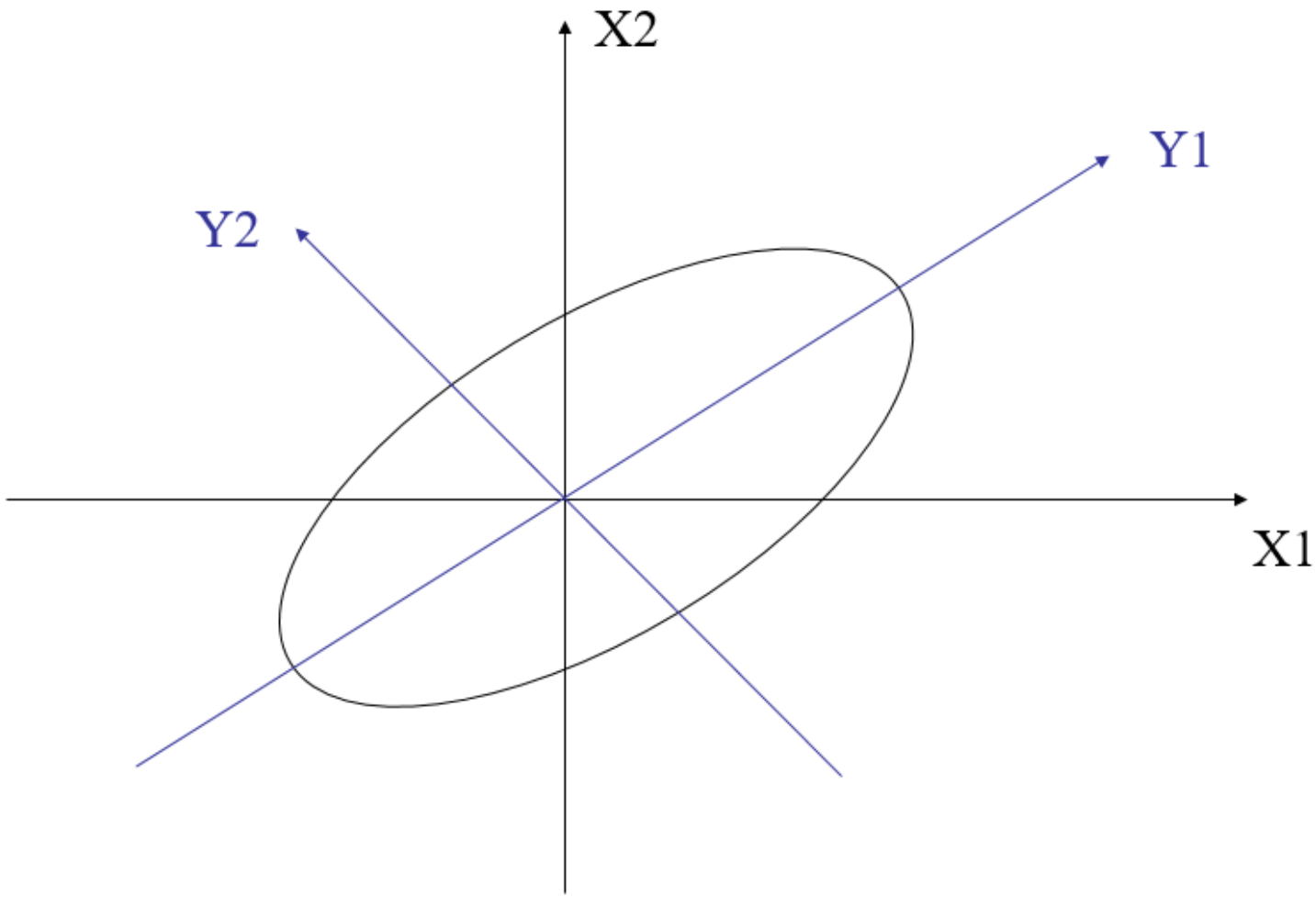




Dimensionality Reduction: Principal Component Analysis (PCA)

- Given N data vectors from n-dimensions, find $k \leq n$ orthogonal vectors (principal components) that can be best used to represent data
- Steps
  - Normalize input data: Each attribute falls within the same range
  - Compute k orthonormal (unit) vectors, i.e., principal components
  - Each input data (vector) is a linear combination of the k principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only
- Used when the number of dimensions is large

Principal Component Analysis (PCA)



Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Example: Log-linear models - obtain value at a point in m-D space as the product on appropriate marginal subspaces
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

Data Reduction Method (1): Regression and Log-Linear Models

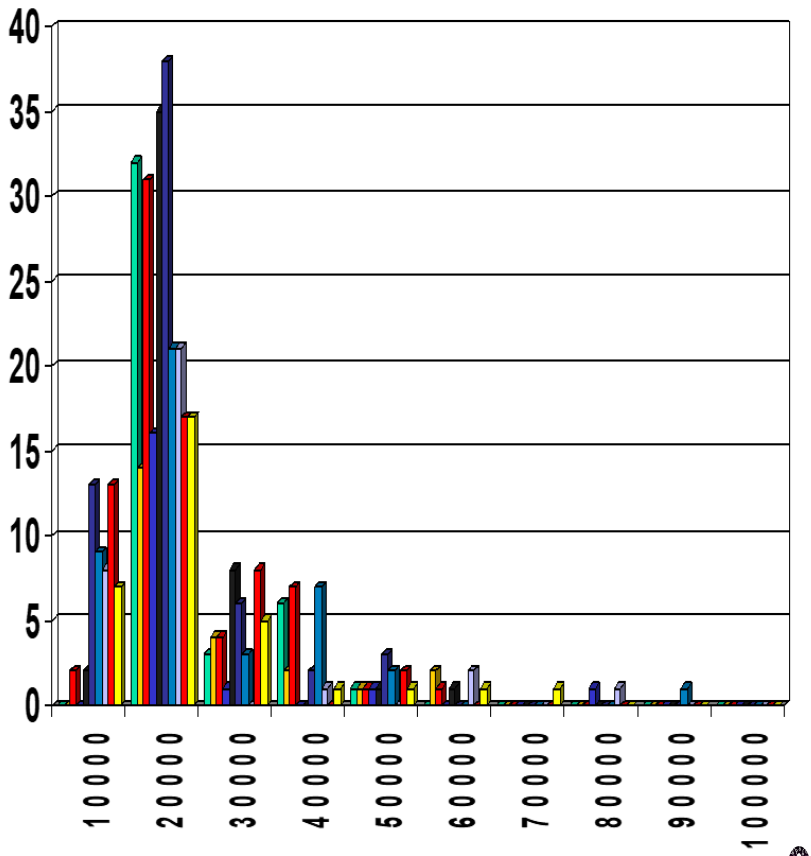
- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability distributions

Regress Analysis and Log-Linear Models

- Linear regression: $Y = wX + b$
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of Y1, Y2, . . . , X1, X2, ...
- Multiple regression: $Y = b0 + b1X1 + b2X2.$
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - The multi-way table of joint probabilities is approximated by a product of lower-order tables
  - Probability: $p(a, b, c, d) = \alpha ab \beta ac \chi ad \delta bcd$

Data Reduction Method (2): Histograms

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range Equal-frequency (or equal-depth)
  - V-optimal: with the least histogram variance (weighted sum of the original values that each bucket represents)
  - MaxDiff: set bucket boundary between each pair for pairs have the $\beta - 1$ largest differences



Example

The following data are a list of prices of commonly sold items.
1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14,15,15, 15,15,15,15,
18,18,18,18,18,18,18,18,20,20,20,20,20,20,20,21,21,21,21,
25,25,25,25,28,28,30,30,30

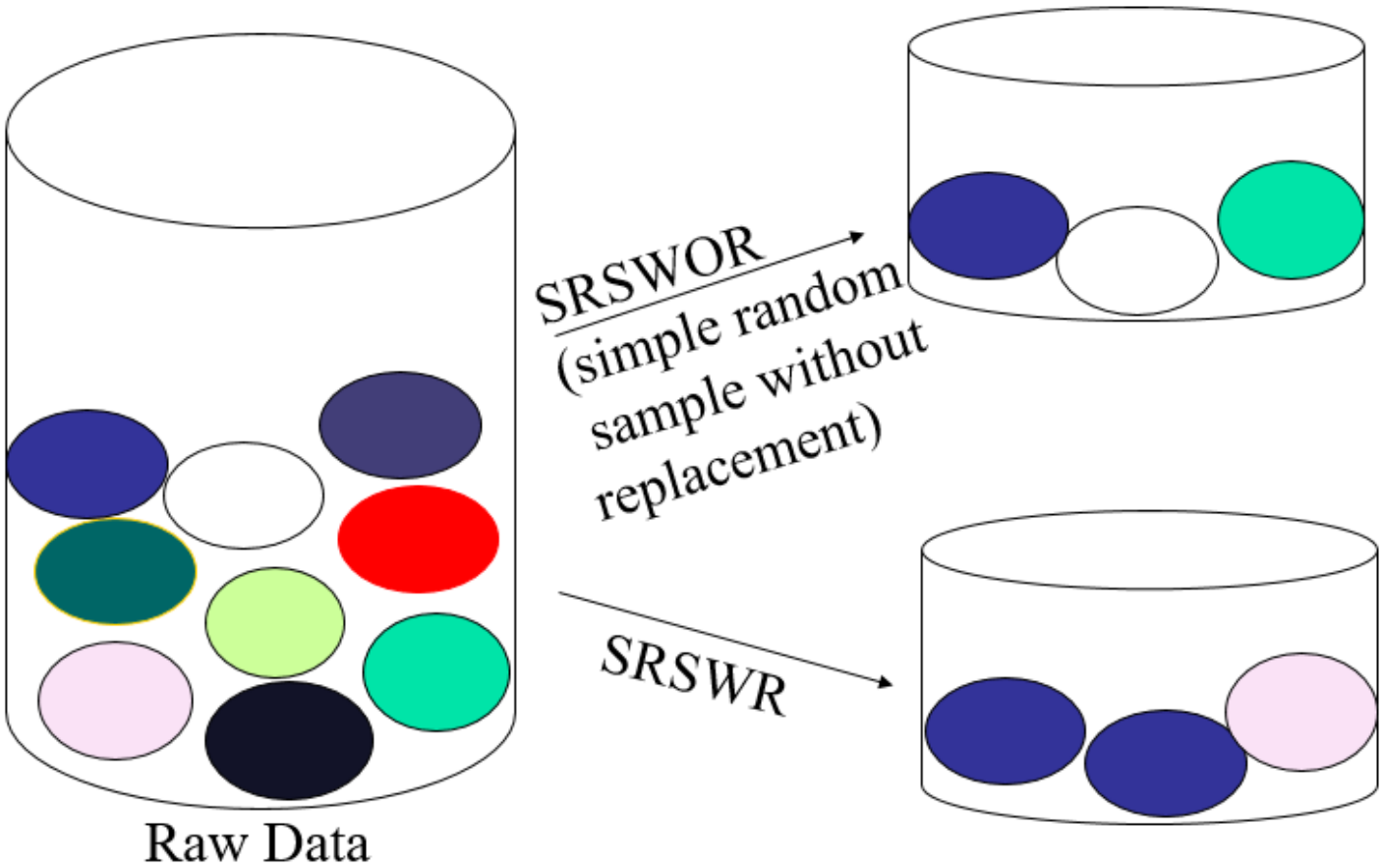
Data Reduction Method (3): Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

Data Reduction Method (4): Sampling

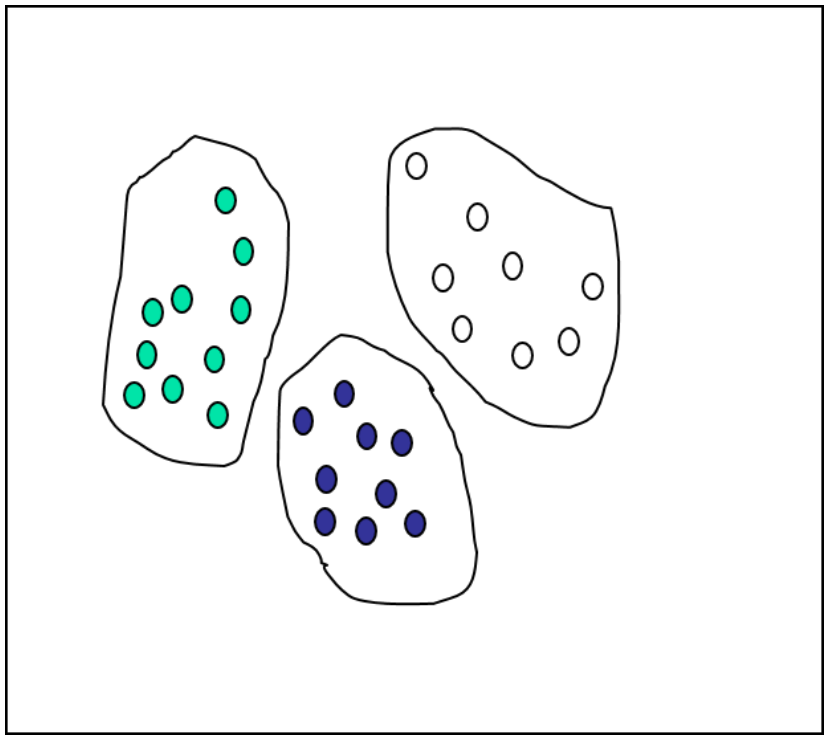
- Sampling: obtaining a small sample s to represent the whole data set N
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Note: Sampling may not reduce database I/Os (page at a time)

Sampling: with or without Replacement

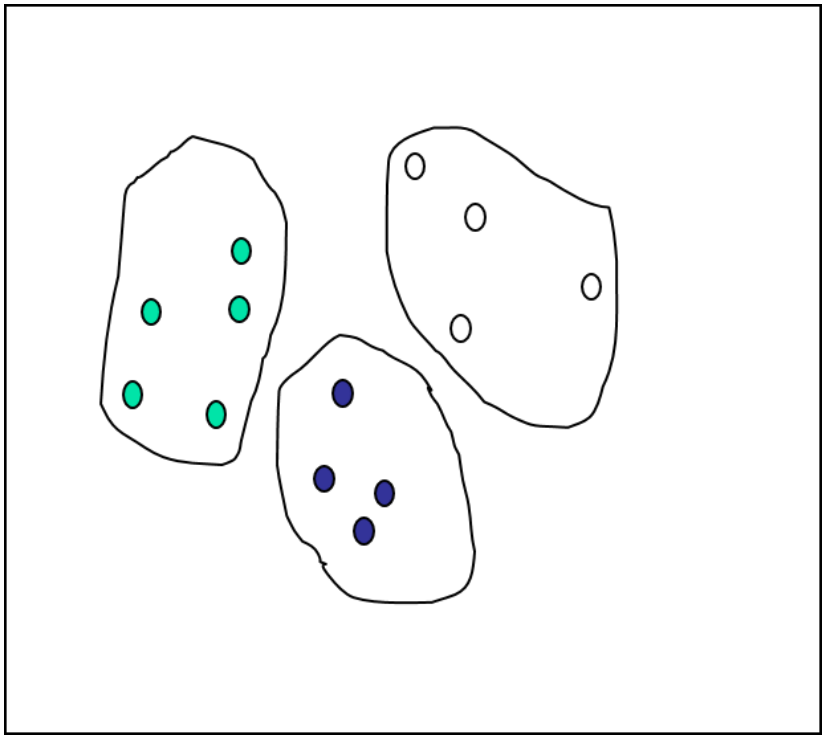


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample




thank
YOU