University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2020
Lecture 05 (Sep 8)

COMPUTING AND SOFTWARE

Career & Internship Fair

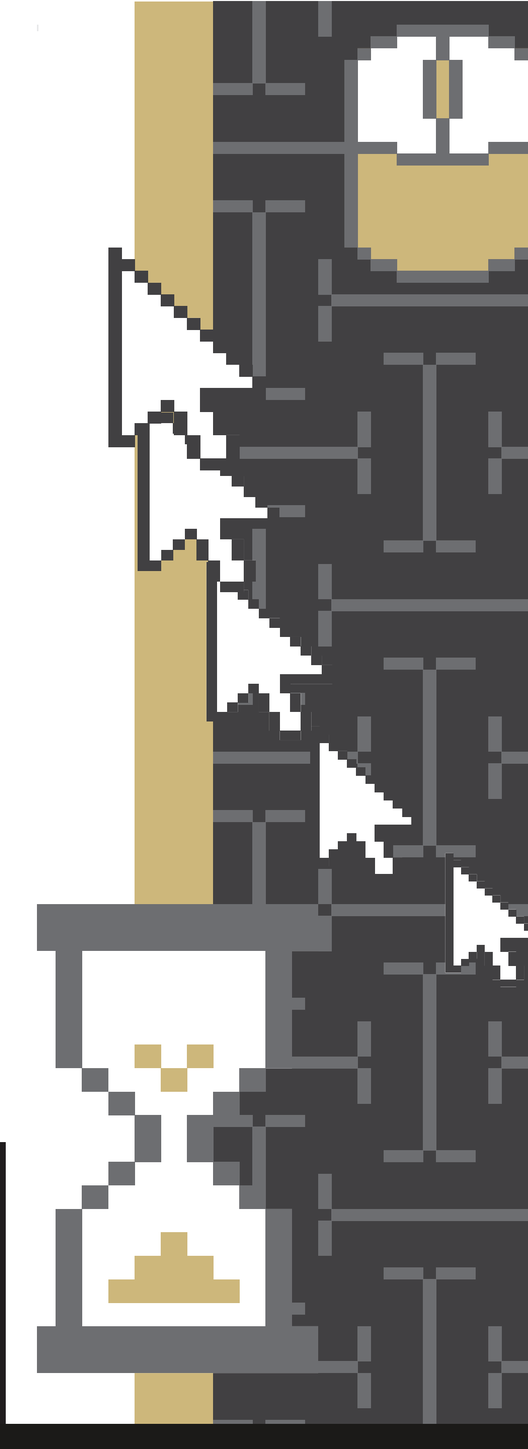
Tuesday, Sept. 15

11 am - 4 pm

VIRTUAL: On Handshake

COMPLETE YOUR PROFILE THEN SIGN UP!

Employers from a variety of companies will be recruiting for computer science, software development or engineering and all levels. Speak with engineers one-on-one and learn about a broad spectrum of work opportunities.



Career Services
UNIVERSITY OF COLORADO **BOULDER**

colorado.edu/career

Reminder

- Homework 1
  - posted in Canvas, due at 9:30am, Th, Sep 10
  - HW1 for CSCI 4502 vs. CSCI 5502
  - Jupyter Notebook for Q3
  - SUBMIT your attempt in Canvas before deadline
  - check syllabus for office hours

Review

- Chapter 2: Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Review: Chap 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization

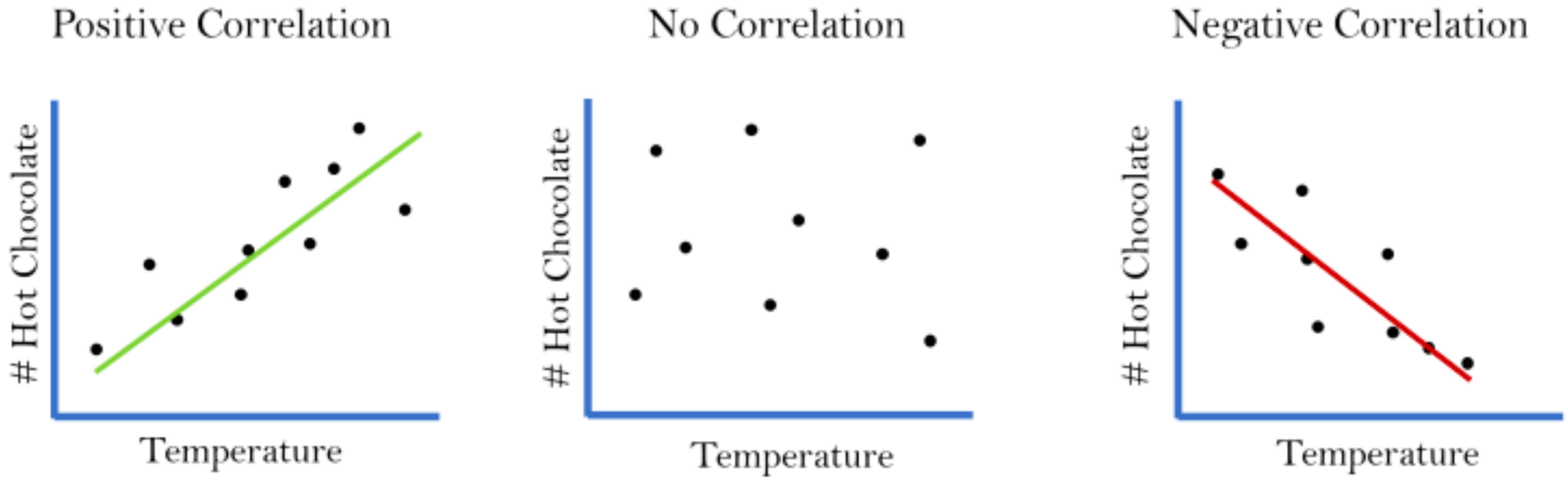
Data Integration

- Combines data from multiple sources
- Entity identification
  - schema integration, object matching
  - e.g., student_id vs. student_number
- Redundant data
  - different naming, derived data
  - may be detected by correlation analysis

Correlation Analysis (1)

- Correlation coefficient (numerical data)

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



https://elisemmyersblog.files.wordpress.com/2017/08/correlations.png?w=736

Correlation Analysis (2)

- X^2 (chi-square) test (categorical data)

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

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

Chi-Square Test: An Example

	play chess	not play chess	total
like fiction	250	200	450
not like fiction	50	1000	1050
total	300	1200	1500

$$e_{11} = \frac{\#(like\ fiction) \times \#(play\ chess)}{N} = \frac{300 \times 450}{1500} = 90$$

Chi-Square Test: An Example

	play chess	not play chess	total
like fiction	250 (90)	200 (360)	450
not like fiction	50 (210)	1000 (840)	1050
total	300	1200	1500

$$e_{11} = \frac{\#(like\ fiction) \times \#(play\ chess)}{N} = \frac{300 \times 450}{1500} = 90$$

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

Correlation Analysis (3)

- Does correlation imply causality?
  - sleeping with one's shoes on is strongly correlated with waking up with a headache
  - the more fireman fighting a damage, the more damage there is going to be
  - as ice cream sales increases, the rate of drowning deaths increases sharply
  - correlation does not imply causality!

http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Chapter 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization

Data Reduction

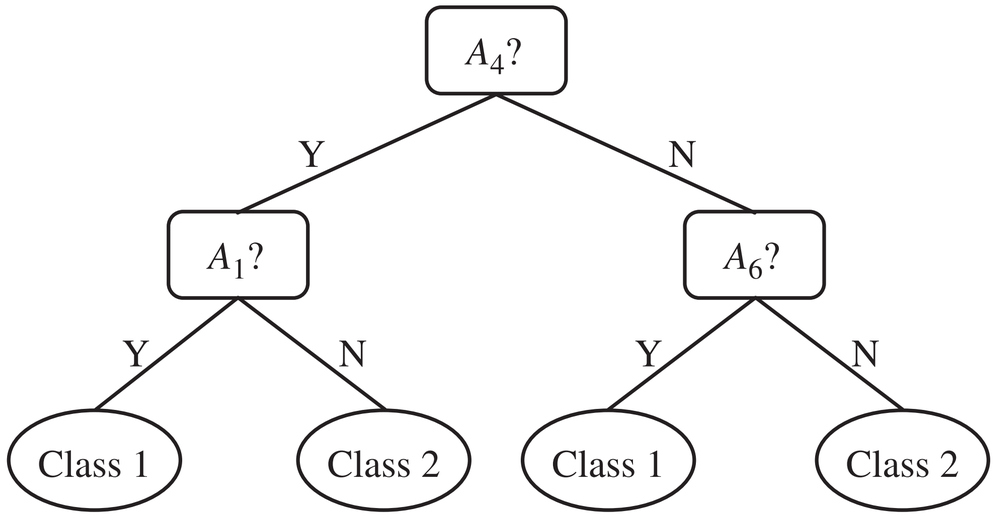
- Why data reduction?
  - massive data sets
  - mining takes a long time
- Goal of data reduction
  - data set is much smaller in volume
  - produces (almost) the same mining results

Data Reduction Strategies

- Dimensionality reduction
  - attribute subset selection
  - Wavelet transform
  - principle component analysis (PCA)
- Numerosity reduction
  - regression, log-linear models
  - data cube aggregation
  - histograms, clustering, sampling

Attribute Subset Selection

- Remove irrelevant or redundant attributes

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Dimensionality Reduction

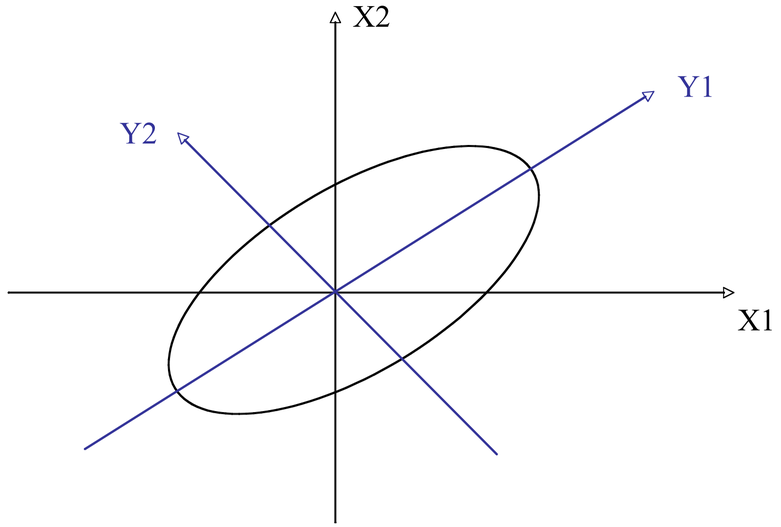
- Discrete wavelet transform (DWT)
  - linear signal processing, multi-resolution
  - store a small fraction of the strongest wavelet coefficients

20 coeffs 100 coeffs 400 coeffs 16,000 coeffs

http://grail.cs.washington.edu/projects/query/

Dimensionality Reduction

- Principal component analysis (PCA)
  - given N data vectors of n dimensions
  - find k <= n orthogonal vectors (principal components) that can
  - best represent the data
  - for numerical data only
  - used when n is large

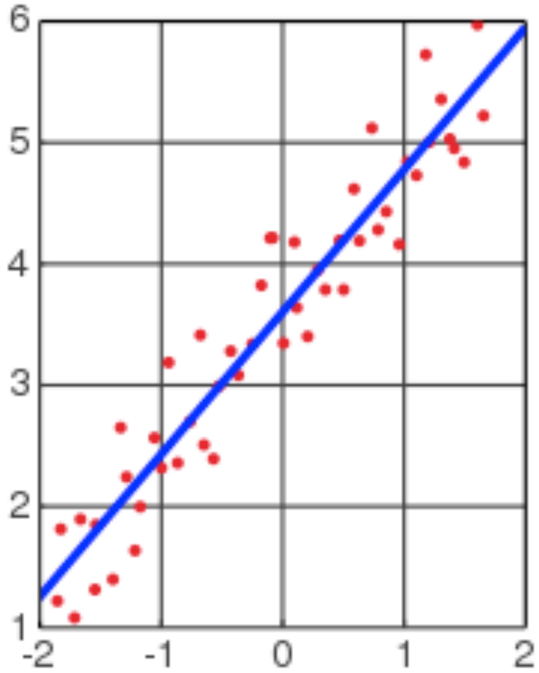


Numerosity Reduction

- Use alternative, smaller data representations
- Parametric methods
  - assume the data fits some model
  - estimate model parameters
  - store the parameters, discard the data
- Non-parametric methods
  - do not assume models
  - e.g., histograms, clustering, sampling

Regression & Log-Linear Models

- Linear regression
  - Y = w X + b
- Multiple regression
  - Y = b0 + b1 X1 + b2 X2
- Log-linear models
  - approximate multi-dimensional probability distributions with lower-dimensional distributions



http://en.wikipedia.org/wiki/Linear_regression

Data Cube Aggregation

- E.g., quarterly sales => annual sales
- Multiple levels of aggregation may be possible
- Use the smallest representation which is enough for the task

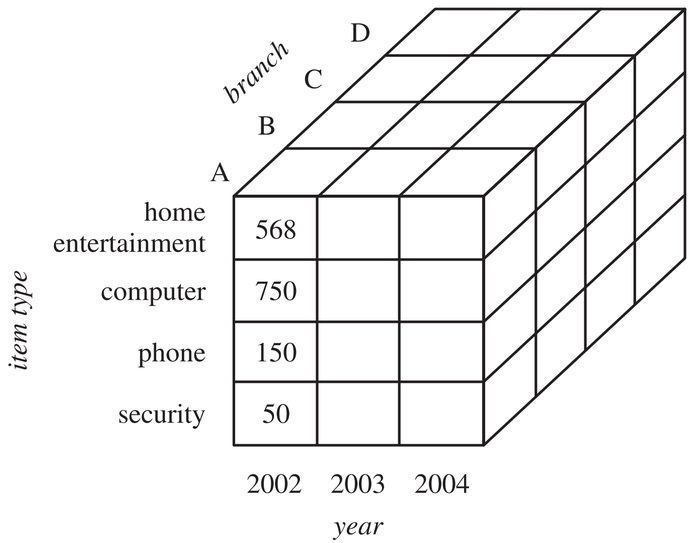
Year 2004

Year 2003

Year 2002

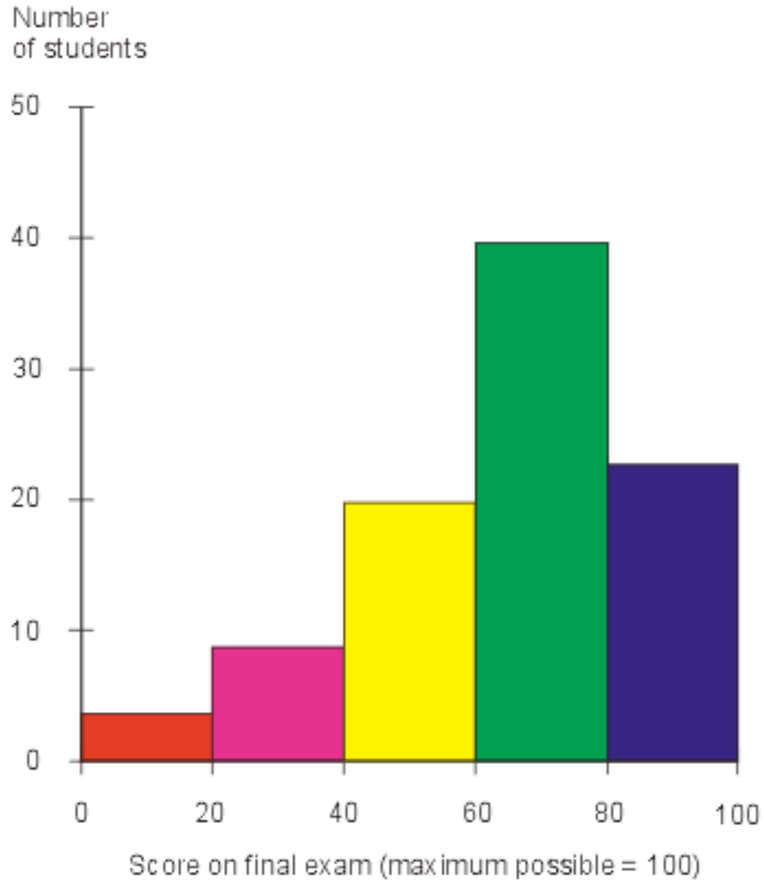
Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000



Histograms

- Divide data into buckets and store average (or sum) for each bucket
- Partitioning rules
  - equal-width
  - equal-frequency
  - v-optimal
  - max-diff



http://media.techtarget.com/digitalguide/images/Misc/iw_histogram.gif

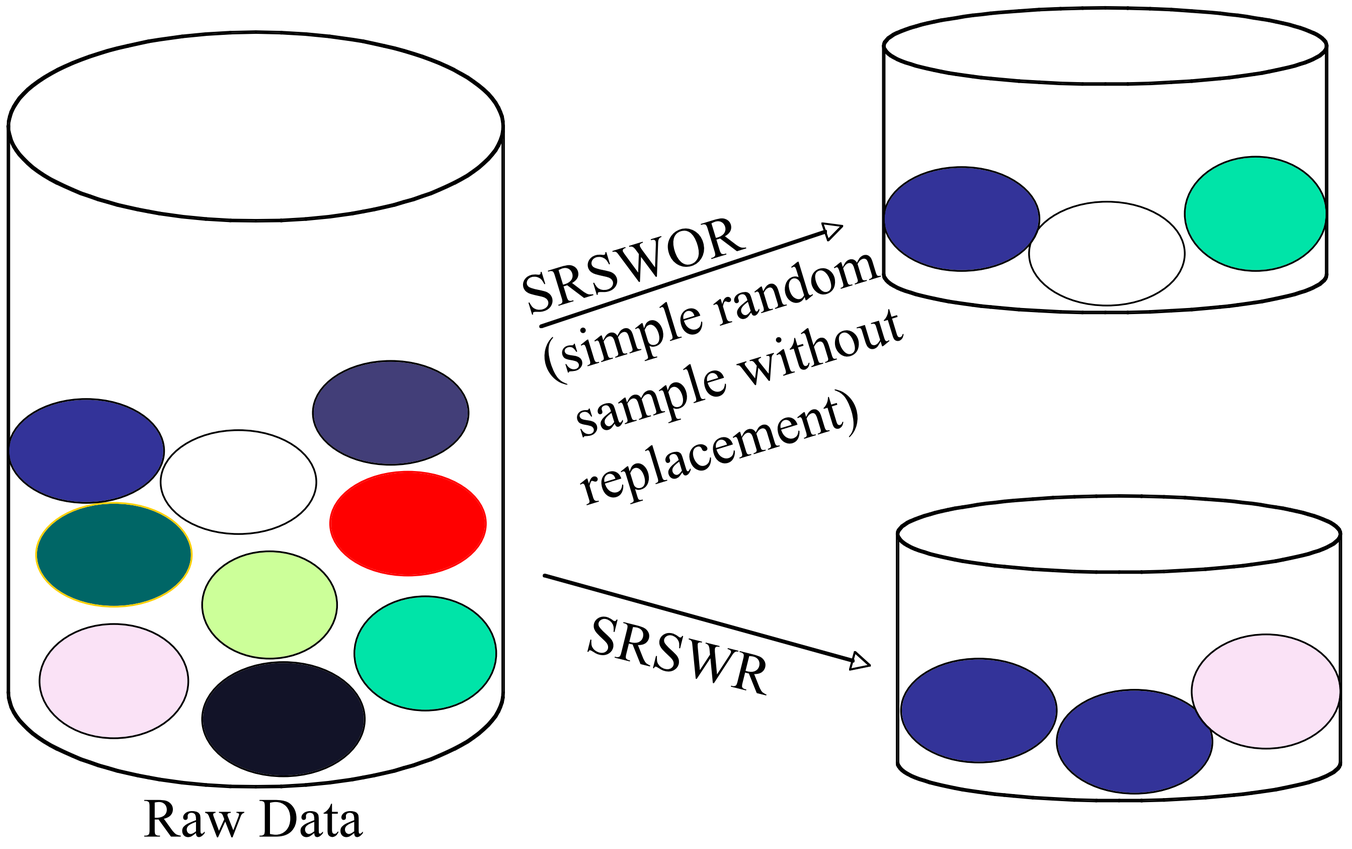
Clustering

- Partition data into clusters based on similarity
- Store cluster representations only
  - e.g., centroid and diameter
- Can have hierarchical clustering
- Many choices of clustering definitions and clustering algorithms

Sampling

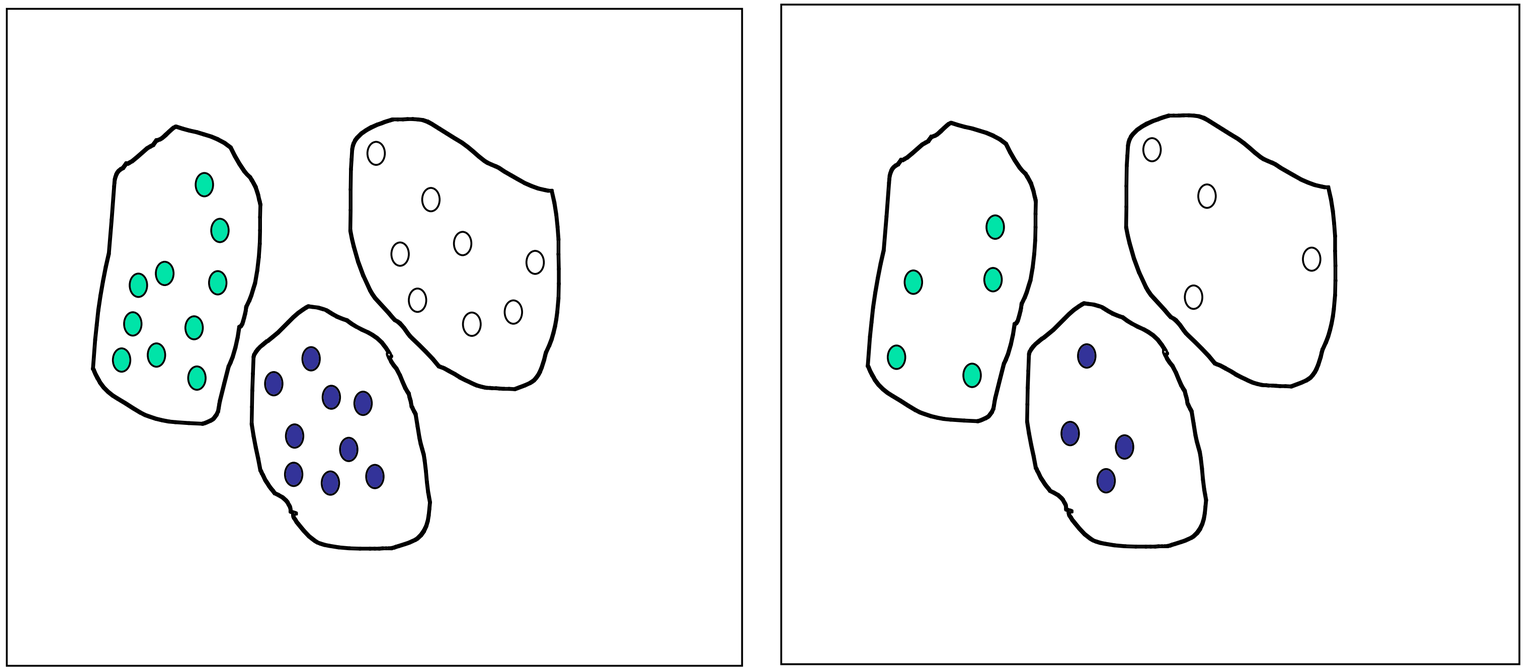
- Use a small sample to represent whole data
- Choose a representative subset of the data
  - simple random sampling may have very poor performance in the presence of skew
- Simple random sample without replacement
- Simple random sample with replacement
- Cluster sample
- Stratified sample

Sample w/ or w/o Replacement



Cluster or Stratified Sampling

- Approximate the percentage of each class



raw data

cluster/stratified sample

Chapter 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
  - e.g., daily sales => monthly, annual sales
- Generalization: concept hierarchy climbing
  - e.g., street => city => state
- Normalization: scale to fall within a range
- Attribute/feature construction: new attributes constructed from existing ones

Normalization (1)

- Min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - e.g., income range [$12,000, $98000] normalize to [0.0, 1.0], then $73,600 is mapped to

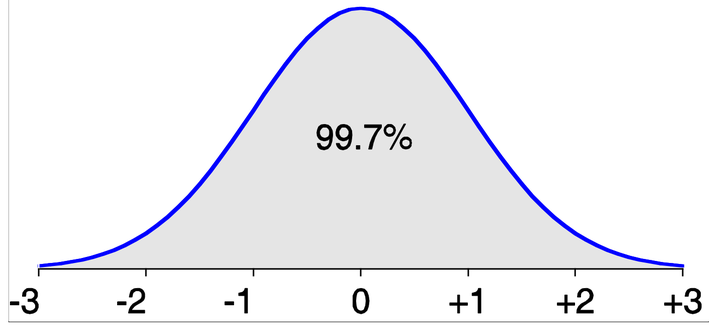
$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- Z-score normalization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - e.g., mean = 54,000
  - stdev = 16,000
  - then

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$



Normalization (2)

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

- where j is the smallest integer s.t. Max(|v'|)<1

- e.g., range [-986, 917]
  - j = 3, divide by 1000
  - -986 => -0.986
  - 917 => 0.917

Discretization

- Three types of attributes
  - nominal: unordered set (e.g., profession)
  - ordinal: ordered set (e.g., military rank)
  - continuous: e.g., integer or real numbers
- Discretization
  - divide continuous range into intervals
  - interval labels used to replace data values
  - supervised vs. unsupervised, split vs. merge

Discretization Methods

- Binning: split, unsupervised
- Histogram analysis: split, unsupervised
- Clustering analysis: split/ merge, unsupervised
- Entropy-based discretization: split, supervised
- Interval merging by X^2 analysis:
  - merge, supervised
- Intuitive partitioning:
  - split, unsupervised

Entropy-Based Discretization

- Partition D into D_1 and D_2 at boundary A

$$Info_A(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2)$$

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

- Pick boundary A with minimum $Info_A(D)$
  - "purer" distribution has lower entropy
- Apply recursively to each partition
- Top-down split, supervised (uses class info)

Interval Merge by X^2 Analysis

- Bottom-up merge, supervised
- Merge the best neighboring intervals
  - intervals w/ most similar class distributions
- ChiMerge
  - merge adjacent intervals w/ min X^2 value
    - i.e., class is independent of interval
  - stopping criterion
    - significance, #intervals, inconsistency, ...

Concept Hierarchy Generation

- Categorical data
- Partial/total ordering of attributes
  - street < city < state < country
- Automatic concept hierarchy generation
  - fewer distinct values => higher level
  - e.g., street, city, state, country
  - exceptions
    - e.g., weekday, month, quarter, year

Chapter 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization