

Chapter 2. Data, Measurements, and Data Preprocessing

- Data Types
- Statics of Data
- Similarity and Distance Measures
- Data Quality, Data Cleaning and Data Integration
- Data Transformation
- Dimensionality Reduction
- Summary

1

1

Types of Data Sets: (1) Record Data

- Relational records
 - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove	12.00	4.00	1.00	240.00	257.00	
Active Outdoors Lyra Glove	10.00	6.00		323.00	339.00	
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lyra Glove		2.00		143.00	145.00	
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

- Transaction data

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

Document	term	freq	tf	df	score	name	ui	oi	tmou	series
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

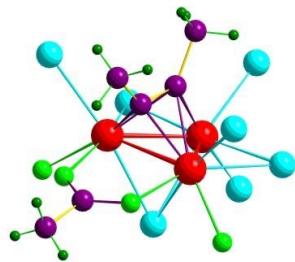
- Document data: Term-frequency vector (matrix) of text documents

2

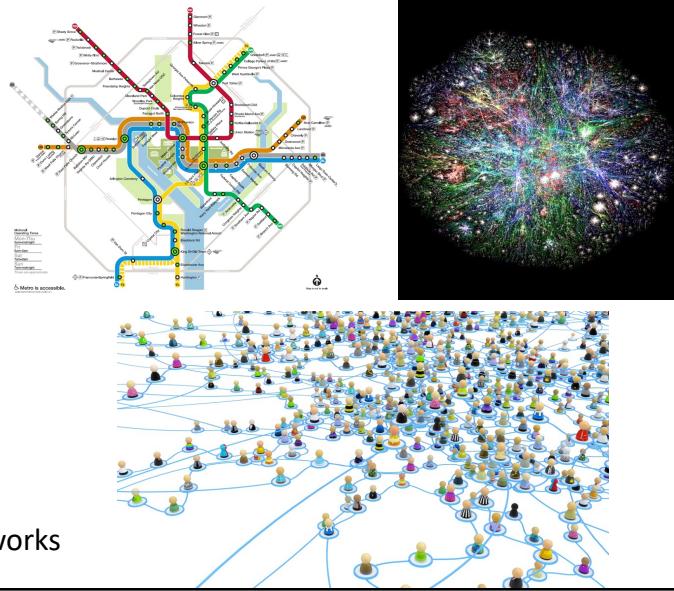
2

Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web



- Molecular Structures
- Social or information networks

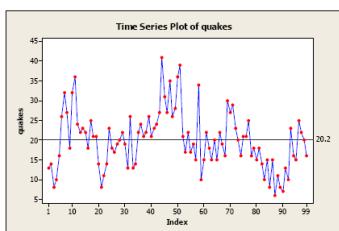


3

3

Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series



- Sequential Data: transaction sequences
- Genetic sequence data

Start

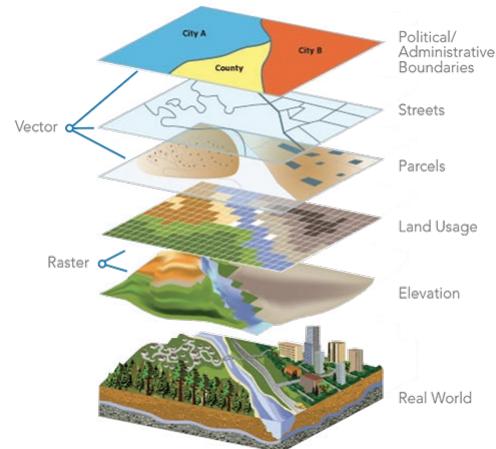
Human	GTTTGAGG	- ATGTCATCA	AATGGCTCCTTCA	TCCCTCATTTACAGACC	TGCCCA
Chimpanzee	GTTTGAGG	- ATGTCATCA	AATGGCTCCTTCA	TCCCTCATTTACAGACC	TGCCCA
Macaque	GTTTGAGG	- ATGTCATCA	AATGGCTCCTTCA	TCCCTCATTTACAGACC	TGCCCA
Human	GACATTCTGCCTAGCAGC	TTGGCTATTATCTG	CTTCTAAACTTAGT	ATTTGGGTG	
Chimpanzee	GACATTCTGCCTAGCAGC	TTGGCTATTATCTG	CTTCTAAACTTAGT	ATTTGGGTG	
Macaque	GACATTCTGCCTAGCAGC	TTGGCTATTATCTG	CTTCTAAACTTAGT	ATTTGGGTG	
Human	GATCTGGAGAC	AA - CTC	GAATAAAAGG	TAATTATTTT	CIAAACAAA
Chimpanzee	GATCTGGAGAC	AA - CTC	GAATAAAAGG	TAATTATTTT	CIAAACAAA
Macaque	GATCTGGAGAC	AA - CTC	GAATAAAAGG	TAATTATTTT	CIAAACAAA
Human	CAGAATAGAT	TTAGCAATTACTTC	TTAGATATTTTAA	TTCTATATTCCTTA	
Chimpanzee	CAGAATAGAT	TTAGCAATTACTTC	TTAGATATTTTAA	TTCTATATTCCTTA	
Macaque	CAGAATAGAT	TTAGCAATTACTTC	TTAGATATTTTAA	TTCTATATTCCTTA	
Human	CCCCTGAGT	GAATGCTGACTTTC	AAAGCAGGTATACA	--- TTATG	
Chimpanzee	CCCCTGAGT	GAATGCTGACTTTC	AAAGCAGGTATACA	--- TTATG	
Macaque	CCCCTGAGT	GAATGCTGACTTTC	AAAGCAGGTATACA	--- TTATG	
Human	UACAGG	TAAGTAAAAAA	ATATTAATTTATTC	H T Y F S K	
Chimpanzee	UACAGG	TAAGTAAAAAA	ATATTAATTTATTC	H T Y F S K	
Macaque	UACAGG	TAAGTAAAAAA	ATATTAATTTATTC	H T Y F S K	

4

4

Types of Data Sets: (4) Spatial, image and multimedia Data

- ❑ Spatial data: maps



- ❑ Image data:
- ❑ Video data:

5

5

Important Characteristics of Structured Data

- ❑ Dimensionality
 - ❑ Curse of dimensionality
- ❑ Sparsity
 - ❑ Only presence counts
- ❑ Resolution
 - ❑ Patterns depend on the scale
- ❑ Distribution
 - ❑ Centrality and dispersion

6

6

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

7

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 (e.g., red, blue)
 - ❑ Binary (e.g., {true, false})
 - ❑ Ordinal (e.g., {freshman, sophomore, junior, senior})
 - ❑ Numeric: quantitative
 - ❑ Interval-scaled: 100°C is interval scales
 - ❑ Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50 °K
 - ❑ Discrete vs. Continuous Attributes

8

8

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

9

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*

10

10

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

11

11

Statics of Data

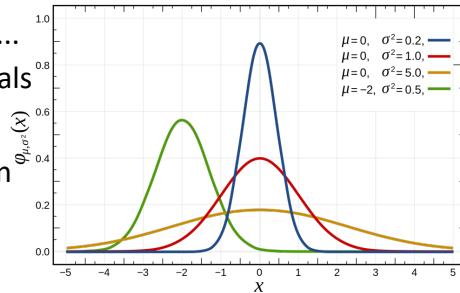
- Measuring the Central Tendency
- Measuring the Dispersion of Data
- Covariance and Correlation Analysis
- Graphic Displays of Basic Statics of Data

12

12

Basic Statistical Descriptions of Data

- Motivation
 - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
 - Median, max, min, quantiles, outliers, variance, ...
- 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



13

13

Measuring the Central Tendency: (1) Mean

- Mean (algebraic measure) (sample vs. population):

Note: n is sample size and N is population size.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- Trimmed mean:

- Chopping extreme values (e.g., Olympics gymnastics score computation)

14

14

Measuring the Central Tendency: (2) Median

□ Median:

□ Middle value if odd number of values, or average of the middle two values otherwise

□ Estimated by interpolation (for *grouped data*):

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Approximate median

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq})_l}{\text{freq}_{\text{median}}} \right) \text{width}$$

Low interval limit Sum before the median interval Interval width ($L_2 - L_1$)

15

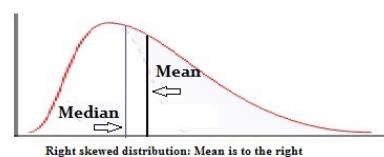
15

Measuring the Central Tendency: (3) Mode

□ Mode: Value that occurs most frequently in the data

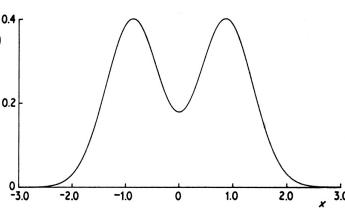
□ Unimodal

□ Empirical formula:

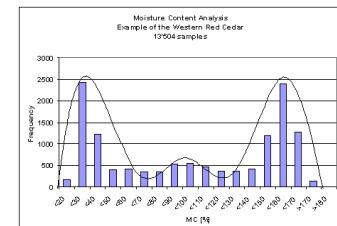
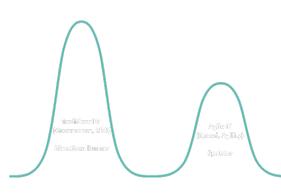
\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})


□ Multi-modal

□ Bimodal



□ Trimodal

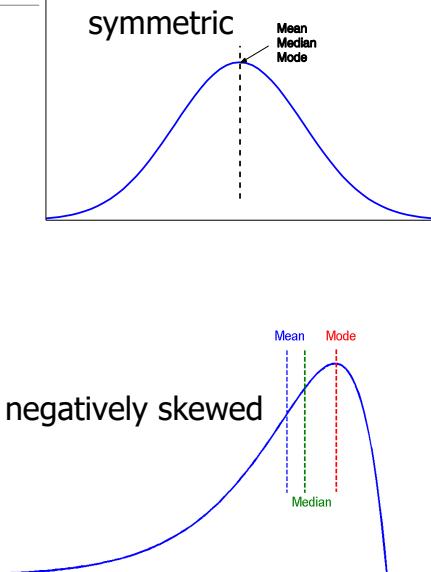
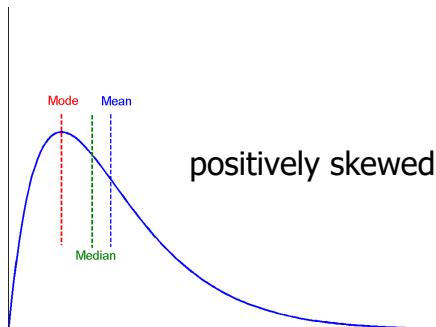


16

16

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

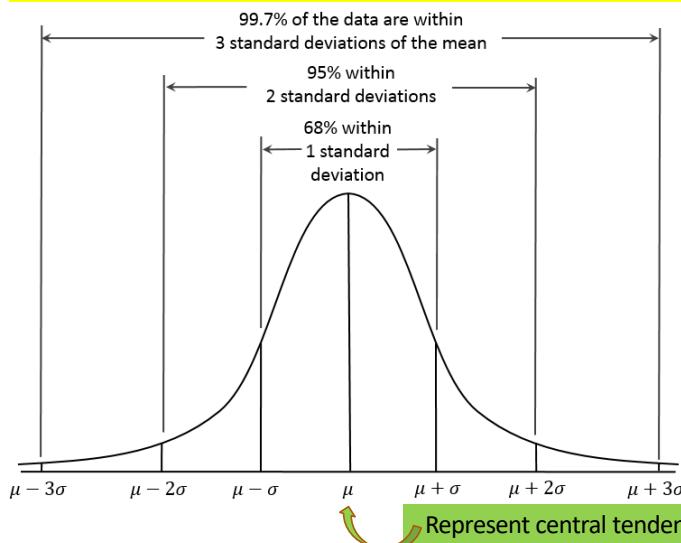


17

17

Properties of Normal Distribution Curve

← ----- Represent data dispersion, spread ----- →



18

18

Measures Data Distribution: Variance and Standard Deviation

- Variance and standard deviation (*sample: s, population: σ*)

- **Variance:** (algebraic, scalable computation)

- Q: Can you compute it incrementally and efficiently?

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \right]$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

Note: The subtle difference of formulae for sample vs. population

- n : the size of the sample
- N : the size of the population

- **Standard deviation s (or σ)** is the square root of variance s^2 (or σ^2)

19

19

Correlation Analysis (for Categorical Data)

- **X² (chi-square) test:**

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

observed
↓
expected

- Null hypothesis: The two distributions are independent
- The cells that contribute the most to the X² value are those whose actual count is very different from the expected count
 - The larger the X² value, the more likely the variables are related
- Note: 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

20

20

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (X1)	200 (X2)	450
Not like science fiction	50 (X3)	1000 (X4)	1050
Sum(col.)	300	1200	1500

- ❑ Null hypothesis: The two distributions are independent
 - ❑ What does that mean?
 - ❑ The ratio between people who play chess vs not play chess is the same for both groups of like science fiction and not like science fiction
 - ❑ $X_1:X_2=X_3:X_4=300:1200$
 - ❑ $X_1:X_3=X_2:X_4=450:1050$
 - ❑ $X_1+X_2=450 \quad X_3+X_4=1050$
 - ❑ $X_1+X_3=300 \quad X_2+X_4=1200$

21

21

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

How to derive 90?
 $450/1500 * 300 = 90$

We can reject the null hypothesis of independence at a confidence level of 0.001

- ❑ 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

22

22

Chi-Square Calculation: An Example

	A	B	C	D	Sum (row)
1					200
0					1000
Sum(col.)	300	300	300	300	1200

- ❑ Degree of freedom
 - ❑ (#categories_in_variable_A -1)(#categories_in_variable_B -1)
 - ❑ number of values that are free to vary

23

23

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

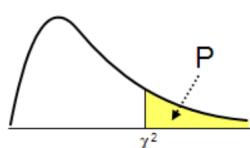
$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- ❑ Degree of freedom =?

We can reject the null hypothesis of independence at a confidence level of 0.001



Values of the Chi-squared distribution



DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458

24

24

Variance for Single Variable (Numerical Data)

- The variance of a random variable X provides a measure of how much the value of X deviates from the mean or expected value of X :

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- where σ^2 is the variance of X , σ is called *standard deviation*
 μ is the mean, and $\mu = E[X]$ is the expected value of X
- That is, variance is the expected value of the square deviation from the mean
- It can also be written as: $\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$

- Sample variance

$$s^2 = \frac{1}{n} \sum_i^n (x_i - \hat{\mu})^2 \quad s^2 = \frac{1}{n-1} \sum_i^n (x_i - \hat{\mu})^2$$

25

25

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1 ; similarly for μ_2

- Sample covariance between X_1 and X_2 : $\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$
- Sample covariance is a generalization of the sample variance:

$$\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i1} - \hat{\mu}_1)$$

- **Positive covariance:** If $\sigma_{12} > 0$
- **Negative covariance:** If $\sigma_{12} < 0$

26

26

Covariance for Two Variables

- **Independence:** If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true
- Some pairs of random variables may have a covariance 0 but are not independent
- Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence
- **Example:**

X_1	1	-1
X_2	0	1

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$$E(X_1) = ?$$

$$E(X_2) = ?$$

$$E(X_1 X_2) = ?$$

27

27

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
- (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
- Covariance formula
- $$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$
- Its computation can be simplified as:
$$\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$$
 - $E(X_1) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$
 - $E(X_2) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$
 - $\sigma_{12} = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

28

28

Correlation between Two Numerical Variables

- **Correlation** between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

- **Sample correlation** for two attributes X_1 and X_2 :

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}}$$

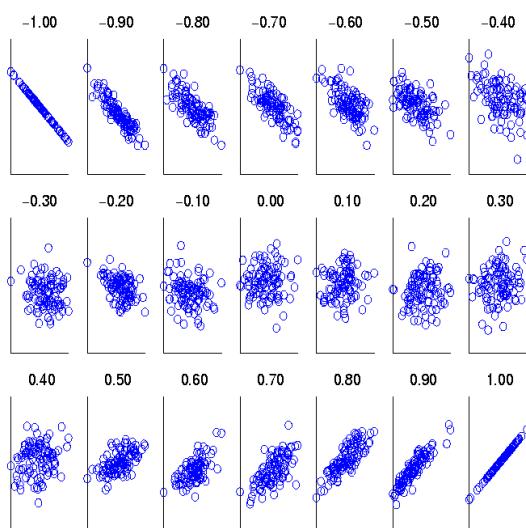
where n is the number of tuples, μ_1 and μ_2 are the respective means of X_1 and X_2 , σ_1 and σ_2 are the respective standard deviation of X_1 and X_2

- If $\rho_{12} > 0$: A and B are positively correlated (X_1 's values increase as X_2 's)
 - The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- If $\rho_{12} < 0$: negatively correlated

29

29

Visualizing Changes of Correlation Coefficient



- Correlation coefficient value range: $[-1, 1]$
- A set of scatter plots shows sets of points and their correlation coefficients changing from -1 to 1

30

30

Covariance Matrix

- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2 X 2 covariance matrix as

$$\begin{aligned}\Sigma &= E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} (X_1 - \mu_1 \quad X_2 - \mu_2) \right] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\end{aligned}$$

- Generalizing it to d dimensions, we have,

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

31

31

Graphic Displays of Basic Statistical Descriptions

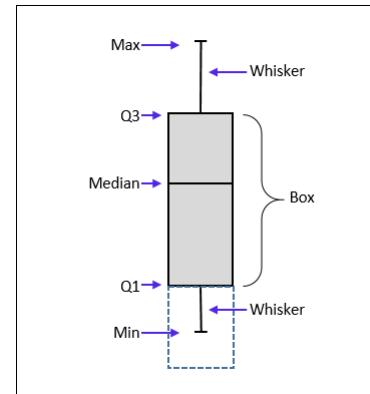
- Boxplot:** graphic display of five-number summary
- Histogram:** x-axis are values, y-axis repres. frequencies
- Quantile plot:** each value x_i is paired with f_i indicating that approximately $100f_i\%$ of data are $\leq x_i$
- Quantile-quantile (q-q) plot:** graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane

32

32

Boxplots

- Measuring the Dispersion of Data
- Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
- Five number summary: min, Q_1 , median, Q_3 , max
- Boxplot: Data is represented with a box
 - Q_1, Q_3 : The ends of the box are at the first and third quartiles, i.e.,
 - Median (Q_2) is marked by a line within the box
 - Whiskers: two lines outside the box extended to Minimum and Maximum

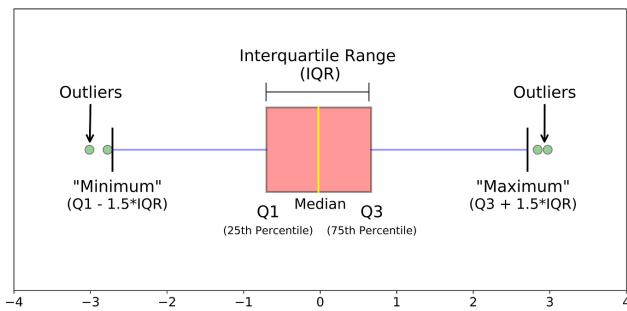


33

33

Boxplots

- Inter-quartile range: $IQR = Q_3 - Q_1$ (the length of the box is IQR)
- Outliers: points beyond a specified outlier threshold, plotted individually
- Outlier: usually, a value higher/lower than $1.5 \times IQR$

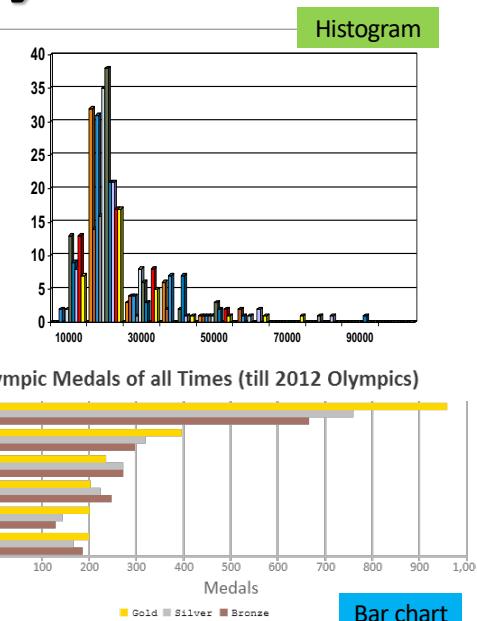


34

34

Histogram Analysis

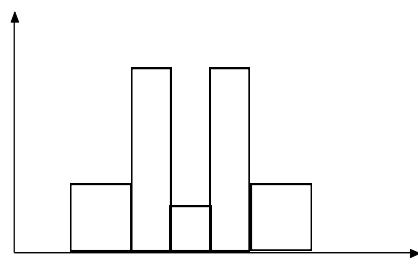
- Histogram: Graph display of tabulated frequencies, shown as bars
- Differences between histograms and bar charts
 - Histograms are used to show distributions of variables while bar charts are used to compare variables
 - Histograms plot binned quantitative data while bar charts plot categorical data
 - Bars can be reordered in bar charts but not in histograms
 - Differs from a bar chart in that it is the area of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width



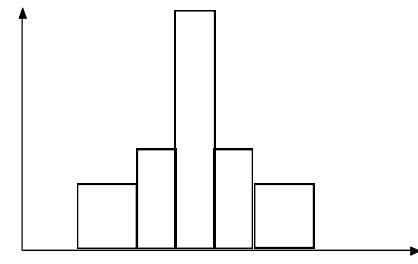
35

35

Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
- The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

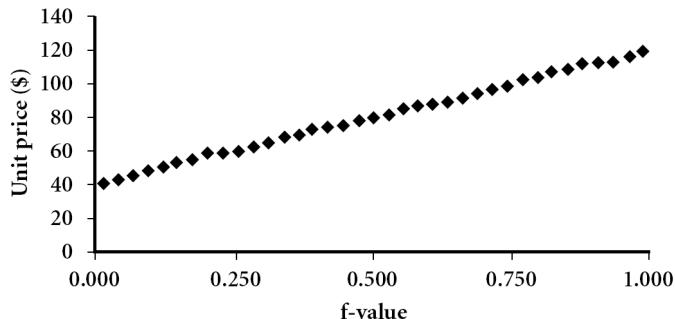


36

36

Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
 - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 $f_i\%$ of the data are below or equal to the value x_i

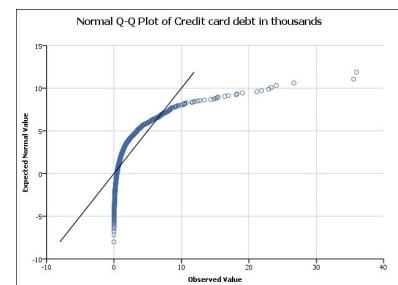
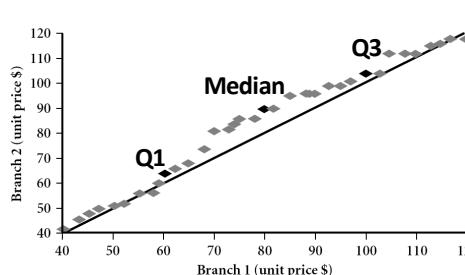


37

37

Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2



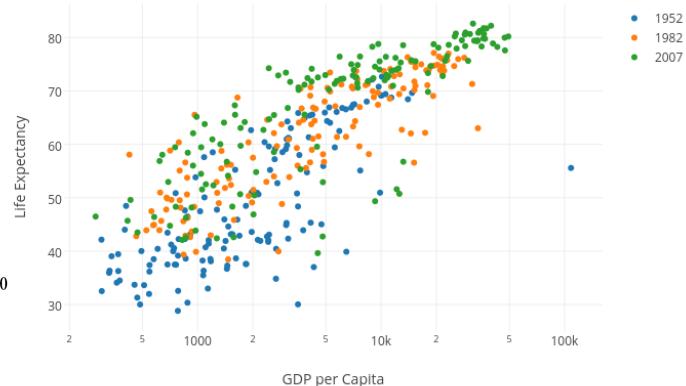
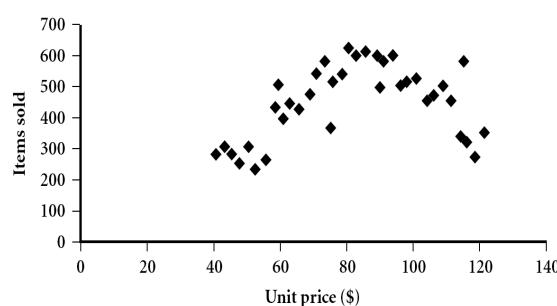
If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line.

38

38

Scatter plot

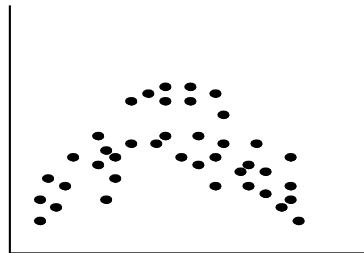
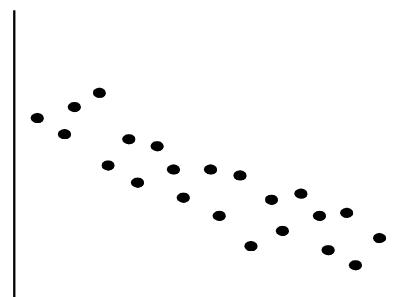
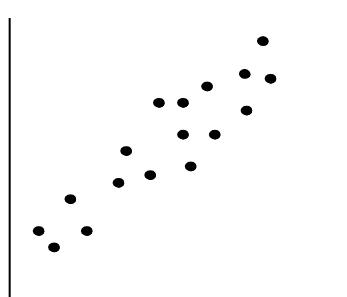
- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



39

39

Positively and Negatively Correlated Data

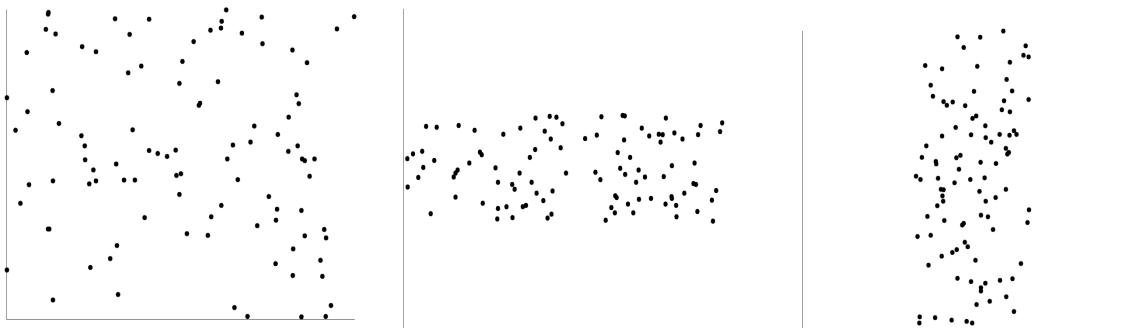


- The left half fragment is positively correlated
- The right half is negative correlated

40

40

Uncorrelated Data



41

41

Similarity and Distance Measures

- Data Matrix versus Dissimilarity Matrix
- Proximity Measures for Nominal Attributes
- Proximity Measures for Binary Attributes
- Dissimilarity of Numeric Data: Minkowski Distance
- Proximity Measures for Ordinal Attributes
- Dissimilarity for Attributes of Mixed Types
- Cosine Similarity
- Measuring Similar Distributions: The Kullback-Leibler Divergence
- Capturing Hidden Semantics in Similarity Measures

42

42

Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function**
 - A real-valued function that quantifies the similarity between two objects
 - Measure how two data objects are alike: The higher value, the more alike
 - Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- **Dissimilarity (or distance) measure**
 - Numerical measure of how different two data objects are
 - In some sense, the inverse of similarity: The lower, the more alike
 - Minimum dissimilarity is often 0 (i.e., completely similar)
 - Range [0, 1] or [0, ∞), depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity

43

43

Data Matrix and Dissimilarity Matrix

- **Data matrix**
 - A data matrix of n data points with l dimensions $\rightarrow D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$
- **Dissimilarity (distance) matrix**
 - n data points, but registers only the distance $d(i, j)$ (typically metric)
 - Usually symmetric, thus a triangular matrix $\rightarrow \begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$
 - **Distance functions** are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
 - Weights can be associated with different variables based on applications and data semantics

44

44

Standardizing Numeric Data

- Z-score:
$$z = \frac{x - \mu}{\sigma}$$
- X: raw score to be standardized, μ : mean of the population, σ : standard deviation
- the distance between the raw score and the population mean in units of the standard deviation
- negative when the raw score is below the mean, "+" when above
- An alternative way: Calculate the mean absolute deviation

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$$

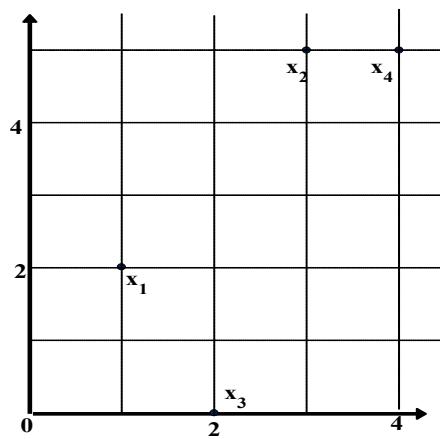
$$\square \text{ standardized measure (z-score): } z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

45

45

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

Dissimilarity Matrix (by Euclidean Distance)

	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$		3.61		
$x3$			2.24	
$x4$				5.39

46

46

Distance on Numeric Data: Minkowski Distance

- ❑ **Minkowski distance:** A popular distance measure

$$d(i, j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{il} - x_{jl}|^p}$$

where $i = (x_{i1}, x_{i2}, \dots, x_{il})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jl})$ are two l -dimensional data objects, and p is the order (the distance so defined is also called L_p norm)

- ❑ Properties

- ❑ $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positivity)
- ❑ $d(i, j) = d(j, i)$ (Symmetry)
- ❑ $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- ❑ A distance that satisfies these properties is a **metric**
- ❑ Note: There are nonmetric dissimilarities, e.g., set differences

47

47

Special Cases of Minkowski Distance

- ❑ $p = 1$: (L_1 norm) **Manhattan (or city block) distance**

- ❑ E.g., the Hamming distance: the number of bits that are different between two binary vectors $d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{il} - x_{jl}|$

- ❑ $p = 2$: (L_2 norm) **Euclidean distance**

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{il} - x_{jl}|^2}$$

- ❑ $p \rightarrow \infty$: (L_{\max} norm, L_{∞} norm) **“supremum” distance**

- ❑ The maximum difference between any component (attribute) of the vectors

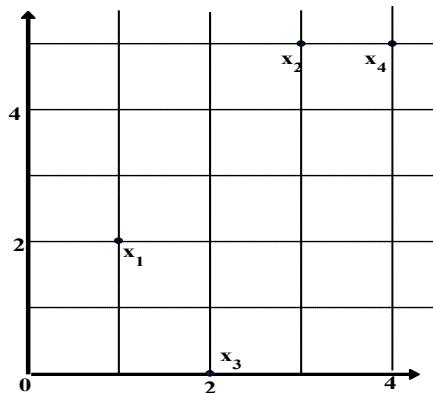
$$d(i, j) = \lim_{p \rightarrow \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{il} - x_{jl}|^p} = \max_{f=1}^l |x_{if} - x_{jf}|$$

48

48

Example: Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L_2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

49

49

Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object j		sum
		1	0	
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
sum	$q+s$	$r+t$	p	

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for symmetric binary variables

$$\text{Distance measure for asymmetric binary variables: } d(i, j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (*similarity* measure for

asymmetric binary variables):

$$\text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q+r+s}$$

(a concept discussed in Pattern Discovery)

50

50

Example: Dissimilarity between Asymmetric Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Mary				
Jack	1	0	Σ_{row}	
	0	1	3	4
Σ_{col}	3	3	6	

- Gender is a symmetric attribute (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- Distance: $d(i, j) = \frac{r + s}{q + r + s}$

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Jim				
Jack	1	0	Σ_{row}	
	0	1	3	4
Σ_{col}	2	4	6	

Mary				
Jim	1	0	Σ_{row}	
	0	2	2	4
Σ_{col}	3	3	6	

51

51

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
 - Example: Color (red, yellow, blue, green), profession, etc.
- Method 1: Simple matching
 - m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
 - Creating a new binary attribute for each of the M nominal states

52

52

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
 - Replace *an ordinal variable value* by its rank: $r_{if} \in \{1, \dots, M_f\}$
 - Map the range of each variable onto $[0, 1]$ by replacing *i*-th object in the *f*-th variable by
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
 - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
 - Then distance: $d(\text{freshman}, \text{senior}) = 1$, $d(\text{junior}, \text{senior}) = 1/3$
 - Compute the dissimilarity using methods for interval-scaled variables

53

53

Attributes of Mixed Type

- A dataset may contain all attribute types
 - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

$$d(i, j) = \frac{\sum_{f=1}^p w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p w_{ij}^{(f)}}$$

- If *f* is numeric: Use the normalized distance
- If *f* is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise
- If *f* is ordinal
 - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
 - Treat z_{if} as interval-scaled

54

54

Cosine Similarity of Two Vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \times \|d_2\|}$$

where \bullet indicates vector dot product, $\|d\|$: the length of vector d

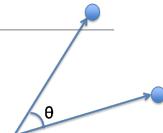
55

55

Example: Calculating Cosine Similarity

- Calculating Cosine Similarity: $\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \times \|d_2\|}$ $sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$
- where \bullet indicates vector dot product, $\|d\|$: the length of vector d
- Ex: Find the **similarity** between documents 1 and 2.
- $d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ $d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$
- First, calculate vector dot product

$$d_1 \bullet d_2 = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$
- Then, calculate $\|d_1\|$ and $\|d_2\|$
- $$\|d_1\| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = \sqrt{6.481}$$
- $$\|d_2\| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{4.12}$$
- Calculate cosine similarity: $\cos(d_1, d_2) = 25 / (6.481 \times 4.12) = 0.94$

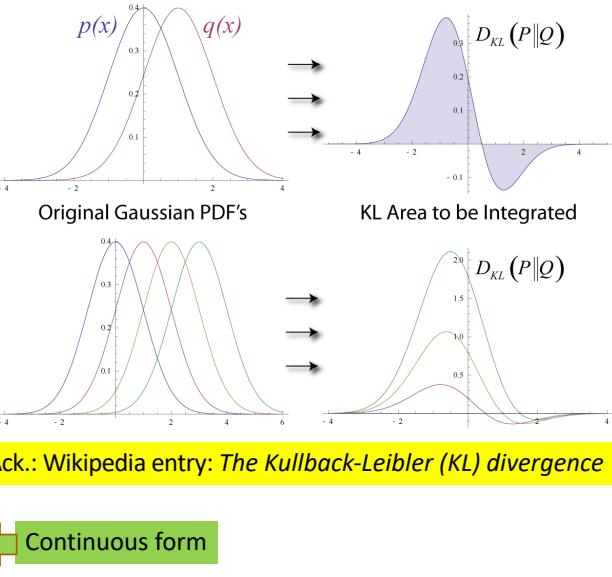


56

56

KL Divergence: Comparing Two Probability Distributions

- The Kullback-Leibler (KL) divergence:
Measure the difference between two probability distributions over the same variable x
- $D_{KL}(p(x) || q(x))$: divergence of $q(x)$ from $p(x)$, measuring the information lost when $q(x)$ is used to approximate $p(x)$



57

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Discrete form

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

Continuous form

More on KL Divergence

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

- Measures the expected number of extra bits required to code samples from $p(x)$ ("true" distribution) when using a code based on $q(x)$, which represents a theory, model, description, or approximation of $p(x)$
- The KL divergence is not a distance measure, not a metric: asymmetric, not satisfy triangular inequality ($D_{KL}(P||Q)$ does not equal $D_{KL}(Q||P)$)
- In applications, P typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while Q typically represents a theory, model, description, or approximation of P.
- In other words, it is the amount of information lost when Q is used to approximate P .
- The KL divergence is sometimes also called the information gain achieved if P is used instead of Q . It is also called the relative entropy of P with respect to Q .

58

58

Subtlety at Computing the KL Divergence

- ❑ Base on the formula, $D_{KL}(P||Q) \geq 0$ and $D_{KL}(P || Q) = 0$ if and only if $P = Q$
- ❑ How about when $p = 0$ or $q = 0$?
$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$
- ❑ $\lim_{p \rightarrow 0} p \log p = 0$
- ❑ when $p \neq 0$ but $q = 0$, $D_{KL}(p || q)$ is defined as ∞ , i.e., if one event e is possible (i.e., $p(e) > 0$), and the other predicts it is absolutely impossible (i.e., $q(e) = 0$), then the two distributions are absolutely different
- ❑ However, in practice, P and Q are derived from frequency distributions, not counting the possibility of unseen events. Thus *smoothing* is needed
- ❑ Example: $P : (a : 3/5, b : 1/5, c : 1/5)$. $Q : (a : 5/9, b : 3/9, d : 1/9)$
 - ❑ need to introduce a small constant ϵ , e.g., $\epsilon = 10^{-3}$
 - ❑ The sample set observed in P , $SP = \{a, b, c\}$, $SQ = \{a, b, d\}$, $SU = \{a, b, c, d\}$
 - ❑ Smoothing, add missing symbols to each distribution, with probability ϵ
 - ❑ $P' : (a : 3/5 - \epsilon/3, b : 1/5 - \epsilon/3, c : 1/5 - \epsilon/3, d : \epsilon)$
 - ❑ $Q' : (a : 5/9 - \epsilon/3, b : 3/9 - \epsilon/3, c : \epsilon, d : 1/9 - \epsilon/3)$
 - ❑ $D_{KL}(P' || Q')$ can then be computed easily

59

Capturing Hidden Semantics in Similarity Measures

- ❑ The above similarity measures cannot capture hidden semantics
 - ❑ Which pairs are more similar: Geometry, algebra, music, politics?
- ❑ The same bags of words may express rather different meanings
 - ❑ “The cat bites a mouse” vs. “The mouse bites a cat”
 - ❑ This is beyond what a vector space model can handle
- ❑ Moreover, objects can be composed of rather complex structures and connections (e.g., graphs and networks)
- ❑ New similarity measures needed to handle complex semantics
 - ❑ Ex. Distributive representation and representation learning

60

60

-
- 17 types of similarity and dissimilarity measures used in data science.
 - <https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681>

 - 9 Distance Measures in Data Science
 - <https://www.maartengrootendorst.com/blog/distances/>

61

61

61

Data Quality, Data Cleaning and Data Integration

- Data Quality Measures
- Data Cleaning
- Data Integration

62

62

What is Data Preprocessing? — Major Tasks

- ❑ **Data cleaning**
 - ❑ Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- ❑ **Data integration**
 - ❑ Integration of multiple databases, data cubes, or files
- ❑ **Data reduction**
 - ❑ Dimensionality reduction
 - ❑ Numerosity reduction
 - ❑ Data compression
- ❑ **Data transformation and data discretization**
 - ❑ Normalization
 - ❑ Concept hierarchy generation

63

63

Why Preprocess the Data? — Data Quality Issues

- ❑ Measures for data quality: A multidimensional view
 - ❑ Accuracy: correct or wrong, accurate or not
 - ❑ Completeness: not recorded, unavailable, ...
 - ❑ Consistency: some modified but some not, dangling, ...
 - ❑ Timeliness: timely update?
 - ❑ Believability: how trustable the data are correct?
 - ❑ Interpretability: how easily the data can be understood?

64

64

Data Cleaning

- ❑ Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
 - ❑ Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - ❑ e.g., *Occupation* = “ ” (missing data)
 - ❑ Noisy: containing noise, errors, or outliers
 - ❑ e.g., *Salary* = “-10” (an error)
 - ❑ Inconsistent: containing discrepancies in codes or names, e.g.,
 - ❑ *Age* = “42”, *Birthday* = “03/07/2010”
 - ❑ Was rating “1, 2, 3”, now rating “A, B, C”
 - ❑ discrepancy between duplicate records
 - ❑ Intentional (e.g., *disguised missing data*)
 - ❑ Jan. 1 as everyone’s birthday?

65

65

Incomplete (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 were not entered due to misunderstanding
 - ❑ Certain data may not be considered important at the time of entry
 - ❑ Did not register history or changes of the data
- ❑ Missing data may need to be inferred

66

66

How to Handle Missing Data?

- ❑ Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % 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**

67

67

Noisy Data

- ❑ **Noise:** random error or variance in a measured variable
- ❑ **Incorrect attribute values** may be due to
 - ❑ Faulty data collection instruments
 - ❑ Data entry problems
 - ❑ Data transmission problems
 - ❑ Technology limitation
 - ❑ Inconsistency in naming convention
- ❑ **Other data problems**
 - ❑ Duplicate records
 - ❑ Incomplete data
 - ❑ Inconsistent data

68

68

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
- ❑ Semi-supervised: Combined computer and human inspection
 - ❑ Detect suspicious values and check by human (e.g., deal with possible outliers)

69

69

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
 - ❑ Integration of the two processes
 - ❑ Iterative and interactive (e.g., Potter's Wheels)

70

70

Data Integration

- ❑ Data integration
 - ❑ Combining data from multiple sources into a coherent store
- ❑ Why data integration?
 - ❑ Help reduce/avoid noise
 - ❑ Get a more complete picture
 - ❑ Improve mining speed and quality
- ❑ Schema integration:
 - ❑ e.g., A.cust-id ≡ B.cust-#
 - ❑ Integrate metadata from different sources
- ❑ Entity identification:
 - ❑ Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

71

71

Handling Noise in Data Integration

- ❑ Detecting data value conflicts
 - ❑ For the same real world entity, attribute values from different sources are different
 - ❑ Possible reasons: no reason, different representations, different scales, e.g., metric vs. British units
- ❑ Resolving conflict information
 - ❑ Take the mean/median/mode/max/min
 - ❑ Take the most recent
 - ❑ Truth finding: consider the source quality
- ❑ Data cleaning + data integration

72

72

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
- ❑ What’s the problem?
 - ❑ $Y = 2X \rightarrow Y = X_1 + X_2 \quad Y = 3X_1 - X_2 \quad Y = -1291X_1 + 1293X_2$
- ❑ Redundant attributes may be detected by correlation analysis and covariance analysis

73

73

Data Transformation

- ❑ Normalization
- ❑ Discretization
- ❑ Data Compression
- ❑ Sampling

74

74

Data Transformation

- ❑ A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- ❑ Methods
 - ❑ Smoothing: Remove noise from data
 - ❑ Attribute/feature construction
 - ❑ New attributes constructed from the given ones
 - ❑ Aggregation: Summarization, data cube construction
 - ❑ Normalization: Scaled to fall within a smaller, specified range
 - ❑ min-max normalization
 - ❑ z-score normalization
 - ❑ normalization by decimal scaling
 - ❑ Discretization: Concept hierarchy climbing

75

75

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 \$73,000 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):

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

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

❑ Ex. Let $\mu = 54,000$, $\sigma = 16,000$. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- ❑ Normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$

76

76

Discretization

- ❑ Three types of attributes
 - ❑ Nominal—values from an unordered set, e.g., color, profession
 - ❑ Ordinal—values from an ordered set, e.g., military or academic rank
 - ❑ Numeric—real numbers, e.g., integer or real numbers
- ❑ Discretization: Divide the range of a continuous attribute into intervals
 - ❑ Interval labels can then be used to replace actual data values
 - ❑ Reduce data size by discretization
 - ❑ Supervised vs. unsupervised
 - ❑ Split (top-down) vs. merge (bottom-up)
 - ❑ Discretization can be performed recursively on an attribute
 - ❑ Prepare for further analysis, e.g., classification

77

77

Data Discretization Methods

- ❑ Binning
 - ❑ Top-down split, unsupervised
- ❑ Histogram analysis
 - ❑ Top-down split, unsupervised
- ❑ Clustering analysis
 - ❑ Unsupervised, top-down split or bottom-up merge
- ❑ Decision-tree analysis
 - ❑ Supervised, top-down split
- ❑ Correlation (e.g., χ^2) analysis
 - ❑ Unsupervised, bottom-up merge
- ❑ Note: All the methods can be applied recursively

78

78

Simple Discretization: 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

79

79

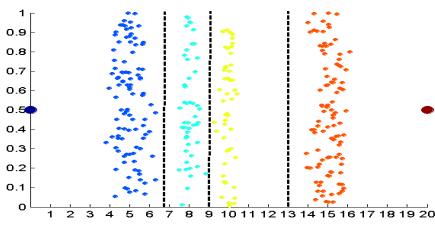
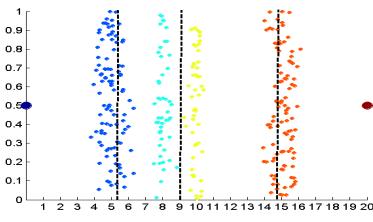
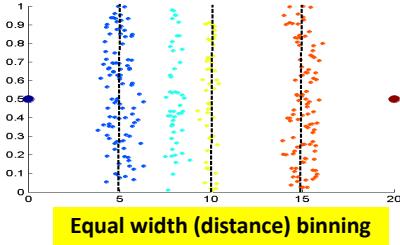
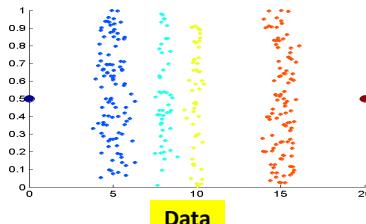
Example: Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 21, 24, 25, 26, 28, 29, 34
- * Partition into equal-frequency (**equal-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

80

80

Discretization Without Supervision: Binning vs. Clustering



81

81

Discretization by Classification & Correlation Analysis

- ❑ Classification (e.g., decision tree analysis)
 - ❑ Supervised: Given class labels, e.g., cancerous vs. benign
 - ❑ Using *entropy* to determine split point (discretization point)
 - ❑ Top-down, recursive split
 - ❑ Details to be covered in Chapter “Classification”
- ❑ Correlation analysis (e.g., Chi-merge: χ^2 -based discretization)
 - ❑ Supervised: use class information
 - ❑ Bottom-up merge: Find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
 - ❑ Merge performed recursively, until a predefined stopping condition

82

82

Concept Hierarchy Generation

- ❑ Concept hierarchy organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- ❑ Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- ❑ Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- ❑ Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- ❑ Concept hierarchy can be automatically formed for both numeric and nominal data—For numeric data, use discretization methods shown

83

83

Concept Hierarchy Generation for Nominal Data

- ❑ Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
 - ❑ *street < city < state < country*
- ❑ Specification of a hierarchy for a set of values by explicit data grouping
 - ❑ {Urbana, Champaign, Chicago} < Illinois
- ❑ Specification of only a partial set of attributes
 - ❑ E.g., only *street < city*, not others
- ❑ Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
 - ❑ E.g., for a set of attributes: {*street, city, state, country*}

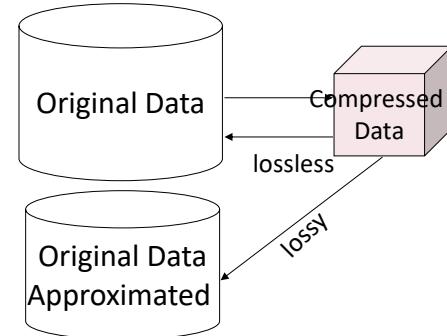
84

84

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
- ❑ Time sequence is not audio
 - ❑ Typically short and vary slowly with time
- ❑ Data reduction and dimensionality reduction may also be considered as forms of data compression

85



Lossy vs. lossless compression

85

Data Cube Aggregation

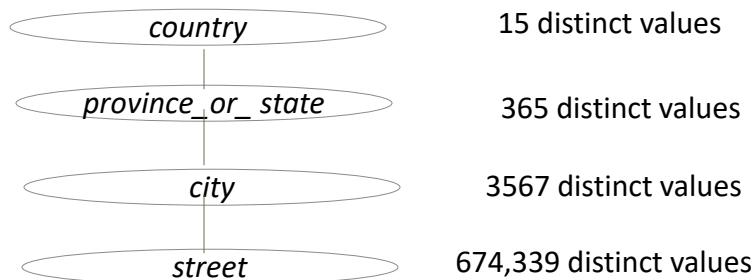
- ❑ The lowest level of a data cube (base cuboid)
 - ❑ The aggregated data for an **individual entity of interest**
 - ❑ E.g., a customer in a phone calling data warehouse
- ❑ Multiple levels of aggregation in data cubes
 - ❑ Further reduce the size of data to deal with
- ❑ Reference appropriate levels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

86



Automatic Concept Hierarchy Generation

- ❑ Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
- ❑ The attribute with the most distinct values is placed at the lowest level of the hierarchy
- ❑ Exceptions, e.g., weekday, month, quarter, year



87

87

Sampling

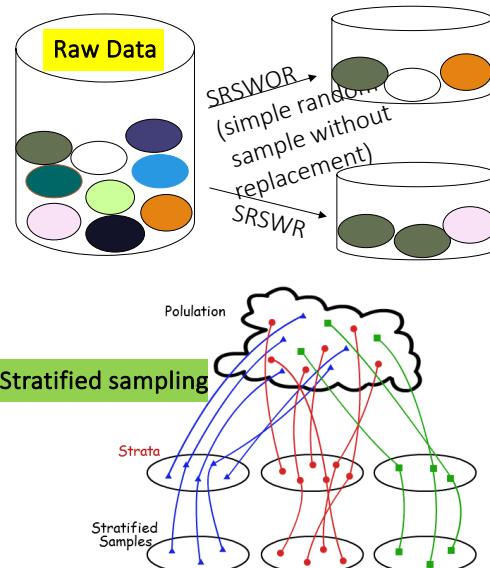
- ❑ Sampling: obtaining a small sample s to represent the whole data set N
- ❑ Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- ❑ Key principle: Choose a **representative** subset of the data
 - ❑ Simple random sampling may have very poor performance in the presence of skew
 - ❑ Develop adaptive sampling methods, e.g., stratified sampling:
- ❑ Note: Sampling may not reduce database I/Os (page at a time)

88

88

Types of Sampling

- Simple random sampling:** equal probability of selecting any particular item
- Sampling without replacement**
 - Once an object is selected, it is removed from the population
- Sampling with replacement**
 - A selected object is not removed from the population
- Stratified sampling**
 - Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)



89

89

Data Reduction

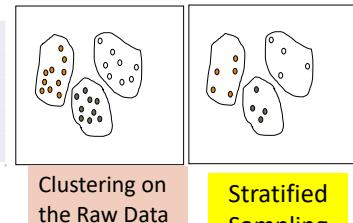
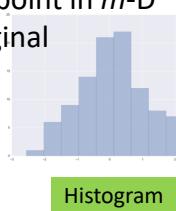
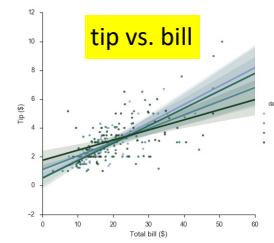
- Data reduction:**
 - Obtain a reduced representation of the data set
 - much smaller in volume but yet produces *almost* the same analytical results
- Why data reduction?—A database/data warehouse may store terabytes of data
 - Complex analysis may take a very long time to run on the complete data set
- Methods for data reduction** (also *data size reduction* or *numerosity reduction*)
 - Regression and Log-Linear Models
 - Histograms, clustering, sampling
 - Data cube aggregation
 - Data compression

90

90

Data Reduction: Parametric vs. Non-Parametric Methods

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- Parametric methods** (e.g., regression)
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
 - Ex.: 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, ...

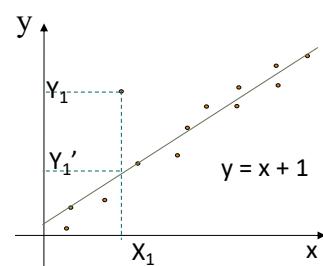


91

91

Parametric Data Reduction: Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a **dependent variable** (also called **response variable** or **measurement**) and of one or more **independent variables** (also known as **explanatory variables** or **predictors**)
- The parameters are estimated so as to give a "best fit" of the data
- Most commonly the best fit is evaluated by using the **least squares method**, but other criteria have also been used



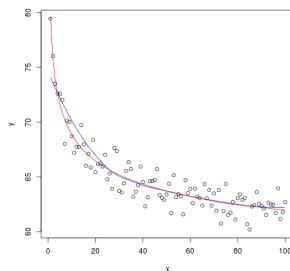
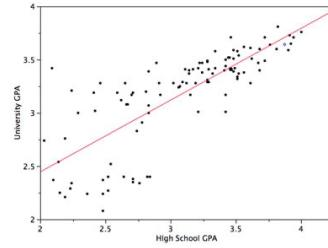
- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

92

92

Linear and Multiple Regression

- ❑ Linear regression: $Y = w X + b$
 - ❑ Data modeled to fit a straight line
 - ❑ Often uses the least-square method to fit the line
 - ❑ 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 $Y_1, Y_2, \dots, X_1, X_2, \dots$
- ❑ Nonlinear regression:
 - ❑ Data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables
 - ❑ The data are fitted by a method of successive approximations

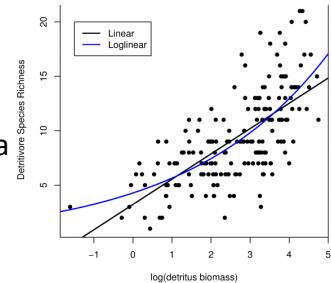
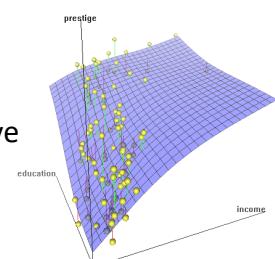


93

93

Multiple Regression and Log-Linear Models

- ❑ Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
 - ❑ Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
 - ❑ Many nonlinear functions can be transformed into the above
- ❑ Log-linear model:
 - ❑ A math model that takes the form of a function whose logarithm is a linear combination of the parameters of the model, which makes it possible to apply (possibly multivariate) linear regression
 - ❑ Estimate the probability of each point (tuple) in a multi-dimen. space for a set of discretized attributes, based on a smaller subset of dimensional combinations
 - ❑ Useful for dimensionality reduction and data smoothing

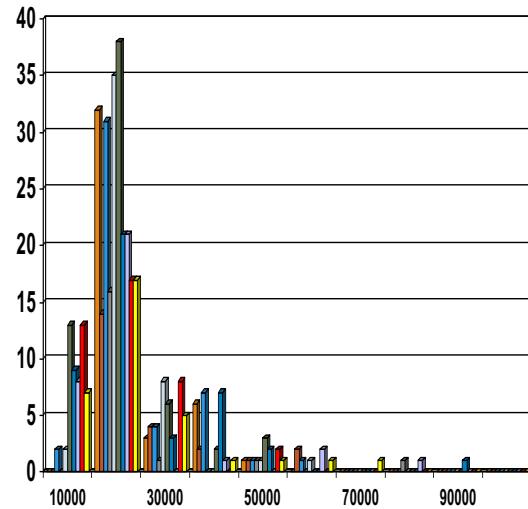


94

94

Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
 - Equal-width: equal bucket range
 - Equal-frequency (or equal-depth)

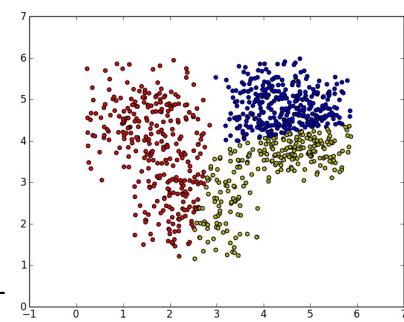


95

95

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
- Cluster analysis will be studied in depth in Chapter 10



96

96

Dimensionality Reduction

- ❑ What Is Dimensionality Reduction?
- ❑ Dimensionality Reduction Methods
 - ❑ Principal Component Analysis
 - ❑ Attribute Subset Selection
 - ❑ Nonlinear Dimensionality Reduction Methods

97

97

What Is Dimensionality Reduction?

- ❑ Curse of dimensionality
 - ❑ When dimensionality increases, data becomes increasingly sparse
 - ❑ Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
 - ❑ The possible combinations of subspaces will grow exponentially
- ❑ Dimensionality reduction
 - ❑ Reducing the number of random variables under consideration, via obtaining a set of principal variables
- ❑ Advantages of dimensionality reduction
 - ❑ Avoid the curse of dimensionality
 - ❑ Help eliminate irrelevant features and reduce noise
 - ❑ Reduce time and space required in data mining
 - ❑ Allow easier visualization

98

98

Dimensionality Reduction Methods

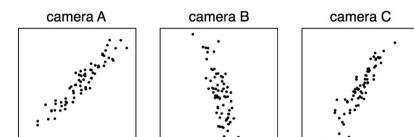
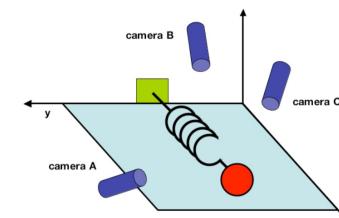
- Dimensionality reduction methodologies
 - **Feature selection:** Find a subset of the original variables (or features, attributes)
 - **Feature extraction:** Transform the data in the high-dimensional space to a space of fewer dimensions
- Some typical dimensionality reduction methods
 - Principal Component Analysis
 - Attribute Subset Selection
 - Nonlinear Dimensionality Reduction

99

99

Principal Component Analysis (PCA)

- PCA: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called *principal components*
- The original data are projected onto a much smaller space, resulting in dimensionality reduction
- Method: Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



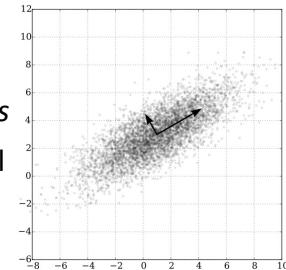
Ball travels in a straight line. Data from three cameras contain much redundancy

100

100

Principal Component Analysis (Method)

- ❑ Given N data vectors from n -dimensions, find $k \leq n$ orthogonal vectors (*principal components*) best used to represent data
 - ❑ 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, to reconstruct a good approximation of the original data)
 - ❑ Works for numeric data only



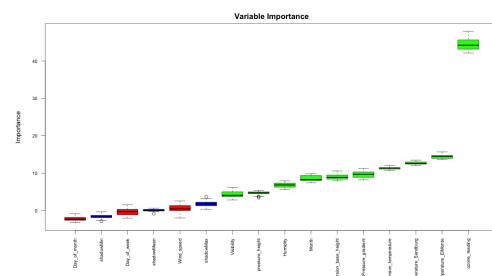
Ack. Wikipedia: Principal Component Analysis

101

101

Attribute Subset Selection

- ❑ Another way to reduce dimensionality of data
 - ❑ Redundant attributes
 - ❑ Duplicate much or all of the information contained in one or more other attributes
 - ❑ E.g., purchase price of a product and the amount of sales tax paid
 - ❑ Irrelevant attributes
 - ❑ Contain no information that is useful for the data mining task at hand
 - ❑ Ex. A student's ID is often irrelevant to the task of predicting his/her GPA



102

102

Heuristic Search in Attribute Selection

- ❑ There are 2^d possible attribute combinations of d attributes
- ❑ Typical heuristic attribute selection methods:
 - ❑ Best single attribute under the attribute independence assumption: choose by significance tests
 - ❑ Best step-wise feature selection:
 - ❑ The best single-attribute is picked first
 - ❑ Then next best attribute condition to the first, ...
 - ❑ Step-wise attribute elimination:
 - ❑ Repeatedly eliminate the worst attribute
 - ❑ Best combined attribute selection and elimination
 - ❑ Optimal branch and bound:
 - ❑ Use attribute elimination and backtracking

103

103

Attribute Creation (Feature Generation)

- ❑ Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- ❑ Three general methodologies
 - ❑ Attribute extraction
 - ❑ Domain-specific
 - ❑ Mapping data to new space (see data reduction)
 - ❑ E.g., Fourier transformation, wavelet transformation, manifold approaches (not covered)
 - ❑ Attribute construction
 - ❑ Combining features (see discriminative frequent patterns in Chapter on “Advanced Classification”)
 - ❑ Data discretization

104

104

Summary

- Data types and attribute types
 - Nominal, binary, ordinal, numerical, discrete vs. continuous attributes
- Statistics of data
 - Central tendency, dispersion, covariance and correlation, graphical displays
- Measure data similarity and correlation
 - Proximity measures for nominal, binary, numerical, ordinal and mixed types
 - Cosine similarity, KL divergence
- Data quality measures, data cleaning, and data integration
- Data transformation: normalization, discretization, data compression and sampling
- Dimensionality reduction methodologies
 - Principal Component Analysis (PCA), attribute subset selection, and nonlinear dimensionality reduction

110

110