

Data Mining

Pre-Processing

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview 
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Learning Outcomes

- Apply data preprocessing techniques for cleaning, integration, reduction and transformation of data
- Use correlation analysis to identify relationship between attributes


Data Quality: Why Preprocess the Data?

- 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?

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 reduction**
 - Dimensionality reduction
 - Data compression
- **Data transformation and data discretization**
 - Normalization
 - Concept hierarchy generation

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning 
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, 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?

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 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 (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”
 - 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

Exercise 1

- Given the following data

Age	Car Accidents	Class (Gender)
21	3	F
?	4	F
20	3	F
22	?	M
19	5	M
18	?	M
?	7	M

- Find the missing values using
 - Mean
 - Class Average
 - Imputation/Nearest-neighbor

Exercise 1

- Find the missing values using

- Mean

- Age:
 - Car Accidents:

- Class Average

- Age
 - Female:
 - Male:
 - Car Accidents
 - Female:
 - Male:

Age	Car Accidents	Class (Gender)
21	3	F
?	4	F
20	3	F
22	?	M
19	5	M
18	?	M
?	7	M

Age	Car Accidents	Class (Gender)
21	3	F
	4	F
20	3	F
22		M
19	5	M
18		M
	7	M

Age	Car Accidents	Class (Gender)
21	3	F
	4	F
20	3	F
22		M
19	5	M
18		M
	7	M

Exercise 1 - Sol

Age	Car Accidents	Class (Gender)
21	3	F
?	4	F
20	3	F
22	?	M
19	5	M
18	?	M
?	7	M

- Find the missing values using
 - Mean
 - Age: $(21+20+22+19+18)/5=20$
 - Car Accidents: 4.4

Age	Car Accidents	Class (Gender)
21	3	F
20	4	F
20	3	F
22	4.4	M
19	5	M
18	4.4	M
20	7	M

Class Average

- Age
 - Female: 20.5
 - Male: 19.7
- Car Accidents
 - Female: 3.33
 - Male: 6

Age	Car Accidents	Class (Gender)
21	3	F
20.5	4	F
20	3	F
22	6	M
19	5	M
18	6	M
19.7	7	M

Exercise 1

- Find the missing values using
 - Imputation/Nearest-neighbor

	Age	Car Accidents	Class (Gender)
A	21	3	F
B	?	4	F
C	20	3	F
D	22	?	M
E	19	5	M
F	18	?	M
G	?	7	M

	Age	Car Accidents	Class (Gender)
A	21	3	F
B		4	F
C	20	3	F
D	22		M
E	19	5	M
F	18		M
G		7	M

Exercise 1 - Sol

- Find the missing values using
 - Imputation/Nearest-neighbor

	Age	Car Accidents	Class (Gender)
A	21	3	F
B	?	4	F
C	20	3	F
D	22	?	M
E	19	5	M
F	18	?	M
G	?	7	M

A, C, D, and F are closest to B. We can either take the average $(21+20+22+18)/4$ or pick the first one (21)

	Age	Car Accidents	Class (Gender)
A	21	3	F
B	21	4	F
C	20	3	F
D	22	7	M
E	19	5	M
F	18	7	M
G	22	7	M

D is closest to G, so we use its value

Distance Matrix

	A	B	C	D	E	F	G
A							
B	1						
C	1	1					
D	2	1	3				
E	5	2	4	1			
F	4	1	3	2	1		
G	5	4	5	0	2	0	

One approach could be to ignore nearest neighbors who have missing values

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** which require 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)

Binning Example

- Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34
- Partition into (equal-frequency) bins:
 - Bin 1: 4, 8, 15
 - Bin 2: 21, 21, 24
 - Bin 3: 25, 28, 34
 - **Smoothing by bin means:**
 - Bin 1: 9, 9, 9
 - Bin 2: 22, 22, 22
 - Bin 3: 29, 29, 29
 - **Smoothing by bin boundaries:**
 - Bin 1: 4, 4, 15
 - Bin 2: 21, 21, 24
 - Bin 3: 25, 25, 34

Data Cleaning as a Process

■ Data discrepancy detection


- Use metadata (e.g., domain, range, dependency, central tendency and distribution)
- Check field overloading
 - inconsistent use of codes and inconsistent data representations
 - E.g. "2010/12/25" and "25/12/2010"
- Check uniqueness rule: A unique attribute should have unique val.
- Consecutive rule: No missing values between min and max values
- Null rule (entries like "don't know" or "?" should be transformed to blank)
- Use commercial tools for discrepancy detection
 - 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 Cleaning as a Process

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

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration 
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Integration

- **Data integration:**
 - Combines data from multiple sources into a coherent store
- Schema integration: e.g., $A.cust-id \equiv B.cust-\#$
 - Integrate metadata from different sources
- **Entity identification problem:**
 - Identify real world entities from multiple data sources, e.g., M Hamza = Muhammad Hamza
- Detecting and resolving data value conflicts
 - For the same real world entity, attribute values from different sources are different. (e.g., Data codes for pay type in one database may be "H" and "S", and 1 and 2 in another)
 - Possible reasons: different representations, (e.g. discount on orders vs lineitems) different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integrating 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* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

- **X² (chi-square) test**

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

- The larger the X² value, the more likely the variables are related
- The cells that contribute the most to the X² value are those whose actual count is very different from the expected count
- 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

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

$\frac{300 \times 450}{1500} = 90$

	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

- χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

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

- For this 2×2 table, the degrees of freedom are $(2 - 1)(2 - 1) = 1$. For 1 degree of freedom, at the 0.05 significance level table value is $3.841 < 507$. So, correlated

Chi-Square Table

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	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	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	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
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Exercise 2

- There are 3000 people who have iPhone, out of these, 2000 have installed WhatsApp.
- There are 2000 people who don't have iPhone, out of these, 1750 have installed WhatsApp
- Is there a correlation between having iPhone and using WhatsApp?

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

Exercise 2

- There are 3000 people who have iPhone, out of these, 2000 have installed WhatsApp.
- There are 2000 people who don't have iPhone, out of these, 1750 have installed WhatsApp

	iPhone	No iPhone	Total
WhatsApp	2000 ()	1750 ()	3750
No WhatsApp	1000 ()	250 ()	1250
Total	3000	2000	5000

Expected value WhatsApp & iPhone =

Expected value WhatsApp & No iPhone =

Expected value No WhatsApp & iPhone =

Expected value No WhatsApp & No iPhone =

Correlation =

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

Exercise 2 - Sol

- There are 3000 people who have iPhone, out of these, 2000 have installed WhatsApp.
- There are 2000 people who don't have iPhone, out of these, 1750 have installed WhatsApp

	iPhone	No iPhone	Total
WhatsApp	2000 (2250)	1750 (1500)	3750
No WhatsApp	1000 (750)	250 (500)	1250
Total	3000	2000	5000

Expected value WhatsApp & iPhone = 27.7

Expected value WhatsApp & No iPhone = 83

Expected value No WhatsApp & iPhone = 41

Expected value No WhatsApp & No iPhone = 125

Correlation = 276

Chi-Square Table

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	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	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	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
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Exercise 3

- Find the correlation between color and gender
- Find the correlation between color and food

ID	Gender	Fav. Color	Fav. Food
1	M	Blue	Cake
2	M	Blue	Pasta
3	F	Red	Pasta
4	M	Red	Cake
5	F	Red	Cake
6	F	Pink	Pasta
7	F	Red	Cake
8	M	Blue	Pasta
9	F	Red	Pasta
10	M	Blue	Cake

Exercise 3

- Color and gender

	Blue	Red	Pink	Total
Male				
Female				
Total				

- Color and food

	Blue	Red	Pink	Total
Cake				
Pasta				
Total				

ID	Gender	Fav. Color	Fav. Food
1	M	Blue	Cake
2	M	Blue	Pasta
3	F	Red	Pasta
4	M	Red	Cake
5	F	Red	Cake
6	F	Pink	Pasta
7	F	Red	Cake
8	M	Blue	Pasta
9	F	Red	Pasta
10	M	Blue	Cake

Exercise 3

- Color and gender

	Blue	Red	Pink	Total
Male	4 ()	1 ()	0 ()	5
Female	0 ()	4 ()	1 ()	5
Total	4	5	1	10

- Color and food

	Blue	Red	Pink	Total
Cake	2 ()	3 ()	0 ()	5
Pasta	2 ()	2 ()	1 ()	5
Total	4	5	1	10

ID	Gender	Fav. Color	Fav. Food
1	M	Blue	Cake
2	M	Blue	Pasta
3	F	Red	Pasta
4	M	Red	Cake
5	F	Red	Cake
6	F	Pink	Pasta
7	F	Red	Cake
8	M	Blue	Pasta
9	F	Red	Pasta
10	M	Blue	Cake

Chi-Square Table

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	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	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	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
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Exercise 3 - Sol

- Color and gender

	Blue	Red	Pink	Total
Male	4 (2)	1 (2.5)	0 (0.5)	5
Female	0 (2)	4 (2.5)	1 (0.5)	5
Total	4	5	1	10

Correlation (color,gender) = $(4-2)^2/2 + (1-2.5)^2/2.5 + (0-0.5)^2/0.5 + (0-2)^2/2 + (4-2.5)^2/2.5 + (1-0.5)^2/0.5 = 6.8 \rightarrow$ Correlated

- Color and food

	Blue	Red	Pink	Total
Cake	2 (2)	3 (2.5)	0 (0.5)	5
Pasta	2 (2)	2 (2.5)	1 (0.5)	5
Total	4	5	1	10

Correlation (color,food) = $(2-2)^2/2 + (3-2.5)^2/2.5 + (0-0.5)^2/0.5 + (2-2)^2/2 + (2-2.5)^2/2.5 + (1-0.5)^2/0.5 = 1.2 \rightarrow$ Not Correlated

ID	Gender	Fav. Color	Fav. Food
1	M	Blue	Cake
2	M	Blue	Pasta
3	F	Red	Pasta
4	M	Red	Cake
5	F	Red	Cake
6	F	Pink	Pasta
7	F	Red	Cake
8	M	Blue	Pasta
9	F	Red	Pasta
10	M	Blue	Cake

Correlation Analysis (Numeric Data)

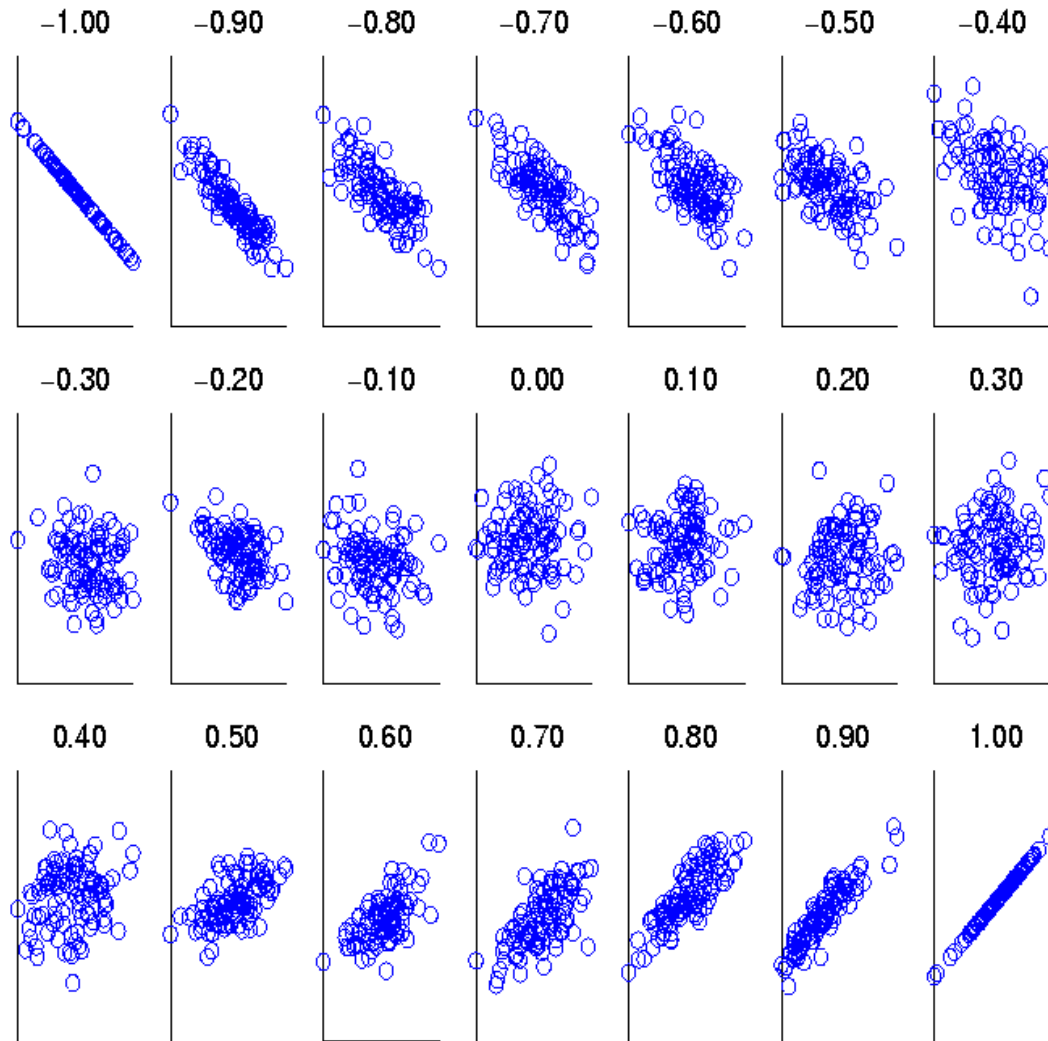
- Correlation coefficient (also called **Pearson's product moment coefficient**)

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

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 , and $\sum(a_i b_i)$ is the sum of the AB cross-product.

- $-1 \leq r_{A,B} \leq +1$.
- 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; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1.

Exercise 4

Given the following data

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

	Age	Car Accidents
	23	3
	21	4
	20	3
	22	2
	19	5
	18	6
	17	7
Average	20.0	4.3
StdDev	2.16	1.7

Find the correlation between Age and Car Accidents using Pearson correlation

Exercise 4

Find the correlation between Age and Car Accidents using Pearson correlation

	Age	Car Accidents
	23	3
	21	4
	20	3
	22	2
	19	5
	18	6
	17	7
Average	20.0	4.3
StdDev	2.16	1.8

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

Exercise 4 - Sol

Find the correlation between Age and Car Accidents using Pearson correlation


Age	Car Accidents
23-20= 3	3-4.3 = -1.3
21-20= 1	4-4.3 = -0.3
20-20= 0	3-4.3 = -1.3
22-20= 2	2-4.3 = -2.3
19-20= -1	5-4.3 = 0.7
18-20= -2	6-4.3 = 1.7
17-20 = -3	7-4.3 = 2.7
2.16	1.8

$$\text{Correl}(\text{age}, \text{acc}) = [(3 \cdot -1.3 + 1 \cdot -0.3 + 0 \cdot -1.3 + 2 \cdot -2.3 - 1 \cdot 0.7 - 2 \cdot 1.7 - 3 \cdot 2.7)] - [7 \cdot 20 \cdot 4.3] / ((7-1) \cdot 2.16 \cdot 1.8) = -0.9$$

-> -0.9 shows that there is a strong negative correlation between the attributes age and # of car accidents

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

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction 
- Data Transformation and Data Discretization
- Summary

Data Reduction Strategies

- **Data reduction:** Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
 - **Dimensionality reduction**, e.g., remove unimportant attributes
 - Principal Components Analysis (PCA)
 - Feature subset selection
 - **Sampling**

Data Reduction 1: 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

- **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

- **Dimensionality reduction techniques**

- Principal Component Analysis (Not covered)
- Supervised and nonlinear techniques (e.g., feature selection)

Attribute Subset Selection

- One 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
 - E.g., students' ID is often irrelevant to the task of predicting students' GPA

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

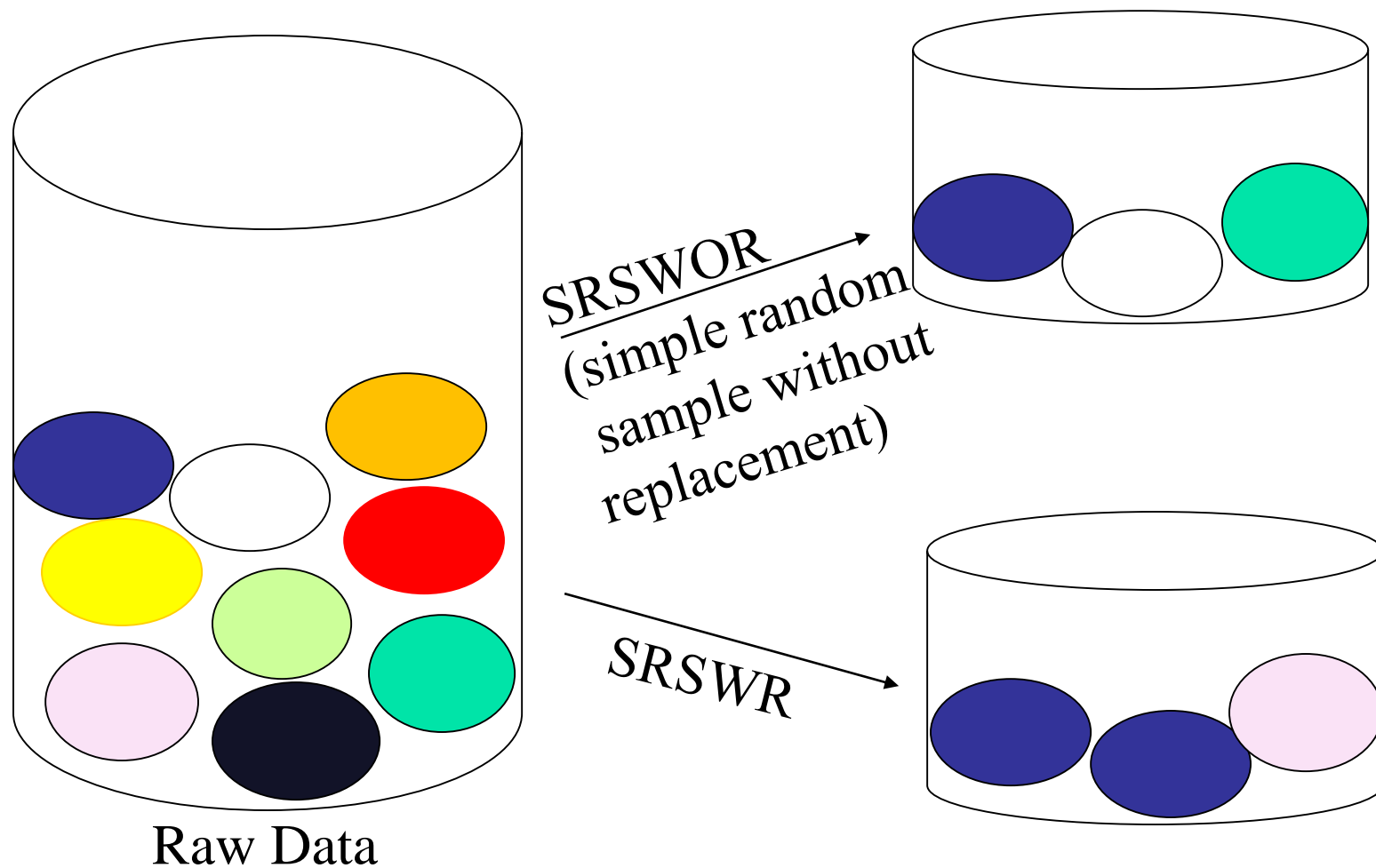
Sampling

- 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)

Types of Sampling

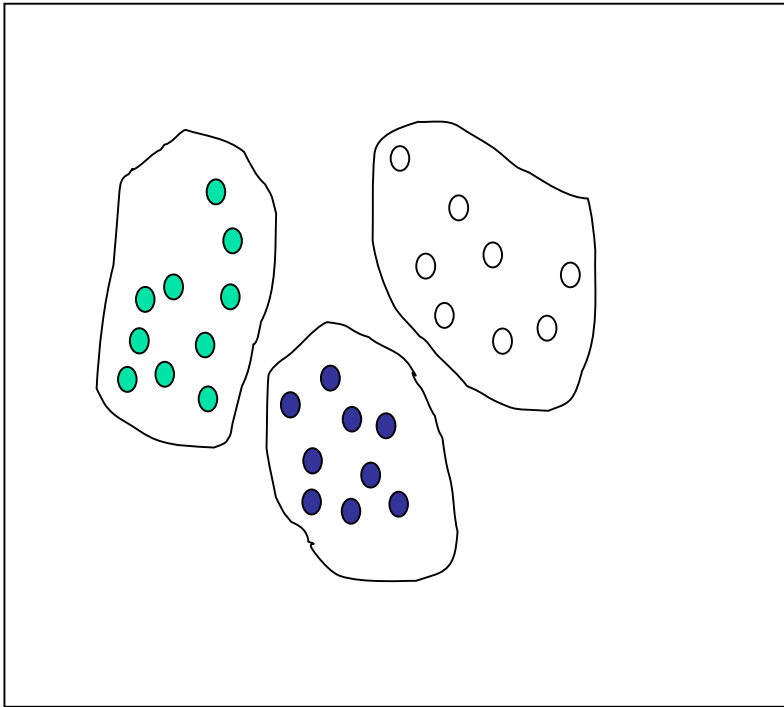
- **Simple random sampling**
 - There is an equal probability of selecting any particular item
- **Sampling without replacement**
 - Once an object is selected, it is removed from the population
- **Sampling with replacement**
 - A selected object is not removed from the population
- **Stratified sampling:**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

Sampling: With or without Replacement

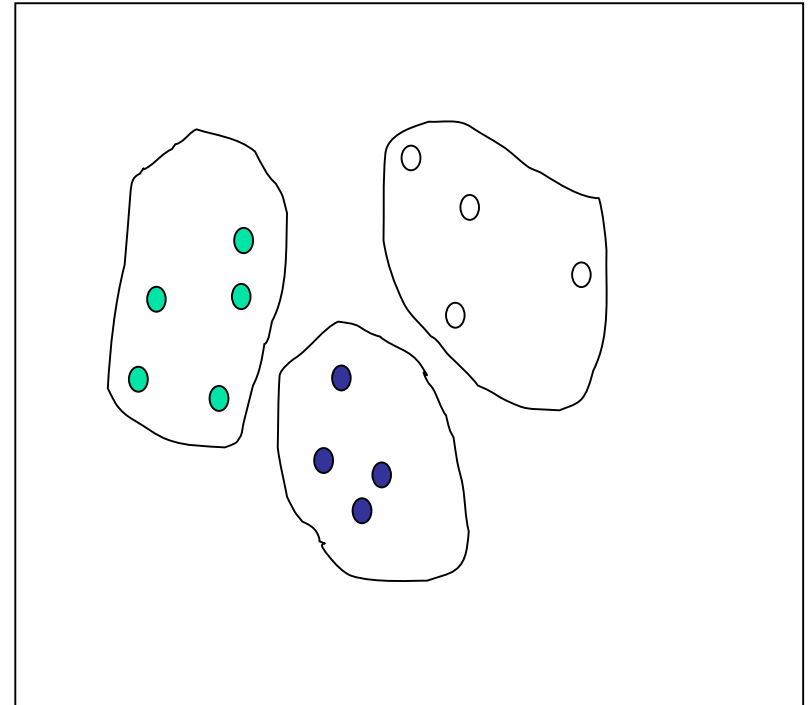


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary



Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that 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

Normalization

- **Min-max normalization:** to $[\text{new_min}_A, \text{new_max}_A]$

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

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

- **Normalization by decimal scaling**

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

Exercise 5

- Given the following data, normalize the attributes Age, Salary, and Grade using

- Min-Max

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- Z-score

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

- Decimal Scaling

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

	Age	Salary (1000s)	Grade
Ali	20	35	1
Bilal	25	20	0.5
Ehsan	20	30	1
Faris	20	25	0
Average	21.3	27.5	0.6
St. Dev.	2.2	5.6	0.4

Exercise 5 – Min-Max

- Given the following data, normalize the attributes Age, Salary, and Grade using
- Min-Max [0,1]
 - minAge = 20
 - maxAge = 25
- Z-score
- Decimal Scaling

	Age	Age'	Salary (1000s)	Salary'	Grade	Grade'
Ali	20		35		1	
Bilal	25		20		0.5	
Ehsan	20		30		1	
Faris	20		25		0	

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

Exercise 5 – Min-Max (Sol)

$$(30-20)/(35-20) = 10/15 = 0.67$$

$$((20-20)/(25-20))*(1-0) + 0 = 0$$

- Given the following data, normalize the attributes Age, Salary, and Grade using
- Min-Max [0,1]
 - minAge = 20
 - maxAge = 25
- Z-score
- Decimal Scaling

	Age	Age'	Salary (1000s)	Salary'	Grade	Grade'
Ali	20	0	35	1	1	1
Bilal	25	1	20	0	0.5	0.5
Ehsan	20	0	30	0.67	1	1
Faris	20	0	25	0.33	0	0

$$(25-20)/(35-20) = 5/15 = 0.33$$

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

Exercise 5 – Z-Score

- Given the following data, normalize the attributes Age, Salary, and Grade using
- Z-score

	Age	Age'	Salary (1000s)	Salary'	Grade	Grade'
<i>Ali</i>	20		35		1	
<i>Bilal</i>	25		20		0.5	
<i>Ehsan</i>	20		30		1	
<i>Faris</i>	20		25		0	
<i>Average</i>	21.3		27.5		0.6	
<i>St. Dev.</i>	2.2		5.6		0.4	

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

Exercise 5 – Z-Score

$$(30-27.5)/5.6 = 0.4$$

- Given the following data, normalize the attributes Age, Salary, and Grade using
- Z-score

	Age	Age'	Salary (1000s)	Salary'	Grade	Grade'
Ali	20	-0.6	35	1.3	1	1.0
Bilal	25	1.7	20	-1.3	0.5	-0.3
Ehsan	20	-0.6	30	0.4	1	1.0
Faris	20	-0.6	25	-0.4	0	-1.5
Average	21.3		27.5		0.6	
St. Dev.	2.2		5.6		0.4	

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

Exercise 5 – Decimal Scaling

- Given the following data, normalize the attributes Age, Salary, and Grade using
- Decimal Scaling

	Age	Age'	Salary (1000s)	Salary'	Grade	Grade'
<i>Ali</i>	20		35		1	
<i>Bilal</i>	25		20		0.5	
<i>Ehsan</i>	20		30		1	
<i>Faris</i>	20		25		0	

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

Exercise 5 – Decimal Scaling

$$30/100 = 0.3$$

- Given the following data, normalize the attributes Age, Salary, and Grade using
- Decimal Scaling

	Age	Age'	Salary (1000s)	Salary'	Grade	Grade'
<i>Ali</i>	20	0.2	35	0.4	1	1.0
<i>Bilal</i>	25	0.3	20	0.2	0.5	0.5
<i>Ehsan</i>	20	0.2	30	0.3	1	1.0
<i>Faris</i>	20	0.2	25	0.3	0	0.0

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

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

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

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

Exercise 6

- Given the following data
- Draw the bins for Age using
 - Equi-width (2 bins)
 - Equi-width (3 bins)
 - Equi-depth (3 bins)

Age	Car Accidents
23	3
21	4
20	3
22	2
19	5
18	6
17	7
29	1
18	5
18	6
23	2
20	4

Exercise 6

$$W = (B - A) / N$$

- Draw the bins for Age using
 - Equi-width (2 bins) [A=17, B=29, N=2]
 - $W =$
 - Bins:
 - Equi-width (3 bins)
 - $W =$
 - Bins:
 - Equi-depth (3 bins)
 - Bins:

Equi-width (2 bins)		Equi-width (3 bins)		Equi-depth (3 bins)	
Age	Bin	Age	Bin	Age	Bin
23		23		23	
21		21		21	
20		20		20	
22		22		22	
19		19		19	
18		18		18	
17		17		17	
29		29		29	
18		18		18	
18		18		18	
23		23		23	
20		20		20	

$$17+6=23 \quad 23+6=29$$

Exercise 6

$$W = (B - A) / N$$

- Draw the bins for Age using
 - Equi-width (2 bins) [A=17, B=29, N=2]
 - $W = (29-17)/2 = 6$
 - Bins: 17 to 23, >23 to 29
 - Equi-width (3 bins)
 - $W = (29-17)/3 = 4$
 - Bins: 17 to 21 , >21 to 25 , >25 to 29
 - Equi-depth (3 bins)
 - Sort the numbers:
17, 18, 18, 19, 19, 20, 21, 22, 23, 23, 29
 - Divide into three groups
17, 18, 18, 18 | 19, 20, 20, 21 | 22, 23, 23, 29
 - Bins: 17-18, 19-21, 22-29

Equi-width (2 bins)		Equi-width (3 bins)		Equi-depth (3 bins)	
Age	Bin	Age	Bin	Age	Bin
23	17-23	23	>21 to 25	23	22-29
21	17-23	21	17 to 21	21	19-21
20	17-23	20	17 to 21	20	19-21
22	17-23	22	>21 to 25	22	22-29
19	17-23	19	17 to 21	19	19-21
18	17-23	18	17 to 21	18	17-18
17	17-23	17	17 to 21	17	17-18
29	>23 to 29	29	>25 to 29	29	22-29
18	17-23	18	17 to 21	18	17-18
18	17-23	18	17 to 21	18	17-18
23	17-23	23	>21 to 25	23	22-29
20	17-23	20	17 to 21	20	19-21

Assignment 2

- Given the following data

Age	Car Accidents
23	3
21	4
20	3
22	2
19	5
18	6
17	7

- Draw the bins for Age using
 - Equi-width (2 bins)
 - Equi-width (3 bins)
 - Equi-depth (3 items)

Concept Hierarchy Generation

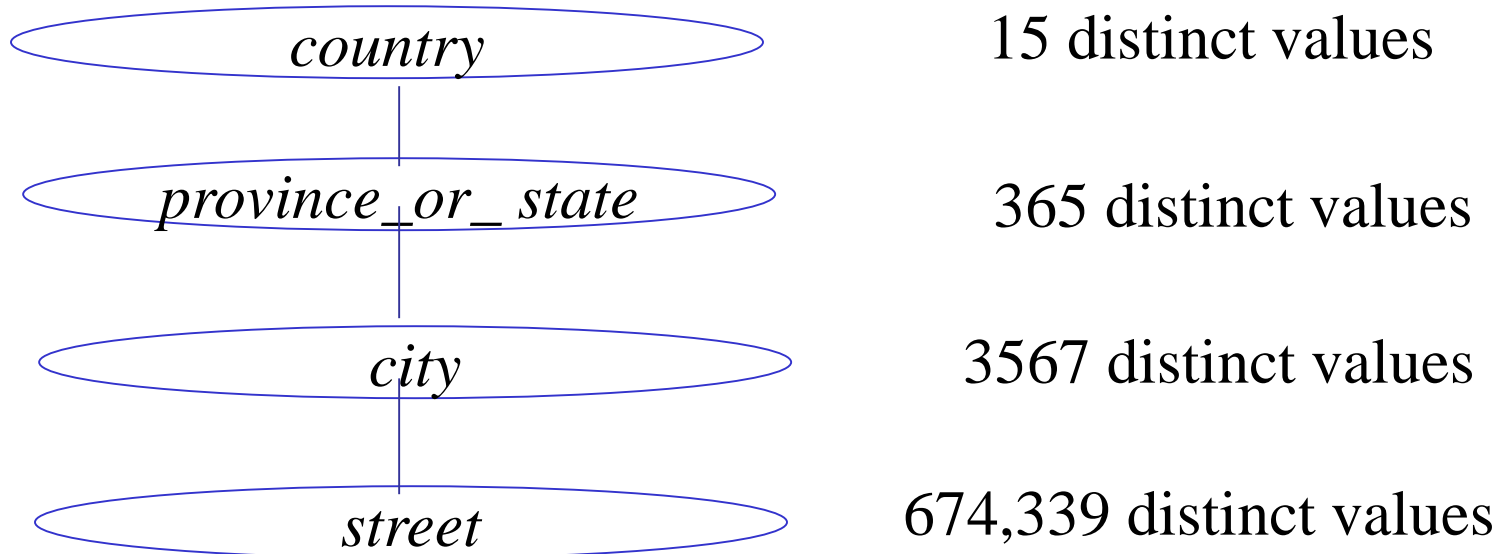
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively 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.

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 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*}

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



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary 

Summary

- **Data quality:** accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning:** e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
 - Entity identification problem
 - Remove redundancies
 - Detect inconsistencies
- **Data reduction**
 - Dimensionality reduction
 - Data compression
- **Data transformation and data discretization**
 - Normalization
 - Concept hierarchy generation

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Comm. of ACM*, 42:73-78, 1999
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), Dec. 1997
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol.23, No.4
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, VLDB'2001
- T. Redman. *Data Quality: Management and Technology*. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995