DATA ANALYTICS

Unit 1:Data Integration

Bharathi.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:Data Integration

Bharathi R

Department of Computer Science and Engineering

Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
  - Data Integration must be careful to avoid
    1. Redundancies
    2. Inconsistencies

Data Integration

- **Challenges in Data integration**:

1. Semantic heterogeneity
2. Structure of data

domain — by different set of people

Data Integration : Solving the challenges

1. How can we match schema and objects from different sources
   - This is called "Entity identification problem"
2. Are any attributes are correlated?
   - Redundancy and correlation analysis
3. Avoid tuple duplication
4. Detection and resolution of data value conflicts

1) Correlation coefficient "pearson" & 2) covariance
Quantitative → Numerical
Qualitative
→ Nominal
↘ Chi square test for independence
H_0
H_1

Data Integration

1. Entity identification problem:
   - Identify real world entities from multiple data sources,

     e.g., Bill Clinton = William Clinton

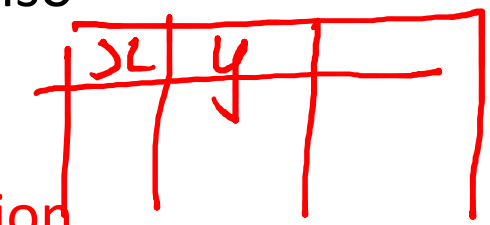
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate "metadata" from different sources

2. Handling Redundancy in Data Integration

- Redundancy is another important issue in data integration. An attribute (such as annual revenue, for instance) may be redundant if it can be "derived" from another attribute or set of attributes.

- Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

- Some redundancies can be detected by correlation analysis.



30 xx.1.2010

Nominal
Numerical

Correlation Analysis (Nominal Data)

- **X^2 (chi-square) test**

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

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count

Chi-Square Calculation: An Example

Suppose that a group of 1500 people was surveyed. The gender of each person was noted. Each person was polled as to whether his or her preferred type of reading material was fiction or nonfiction. Thus, we have two attributes, gender and preferred reading.

The observed frequency (or count) of each possible joint event is summarized in the contingency table shown in Table below, where the numbers in parentheses are the expected frequencies.

Chi-Square Calculation: An Example

χ^2 & p value
χ^2 table.

2 attributes
1) gender < M / F
2) Reading material < Fic / X Scie Fic

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

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

	M	F	Sum
Fic	250	200	450
X Fic	50	1000	1050
Tot:	300	1200	1500

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

$\sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$

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

Observed values

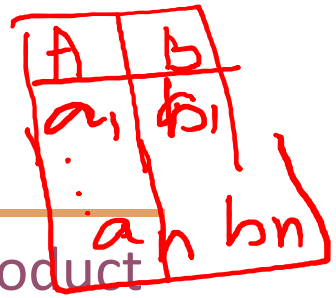
Expected value

$e_{ij} = \frac{(\text{row total})(\text{col tot})}{\text{total sample size}}$

- It shows that gender and preferred reading are correlated in the group

H_0 : Gender & Reading are indep
H_1 : " & " are dependent

Correlation Analysis (Numeric Data)



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

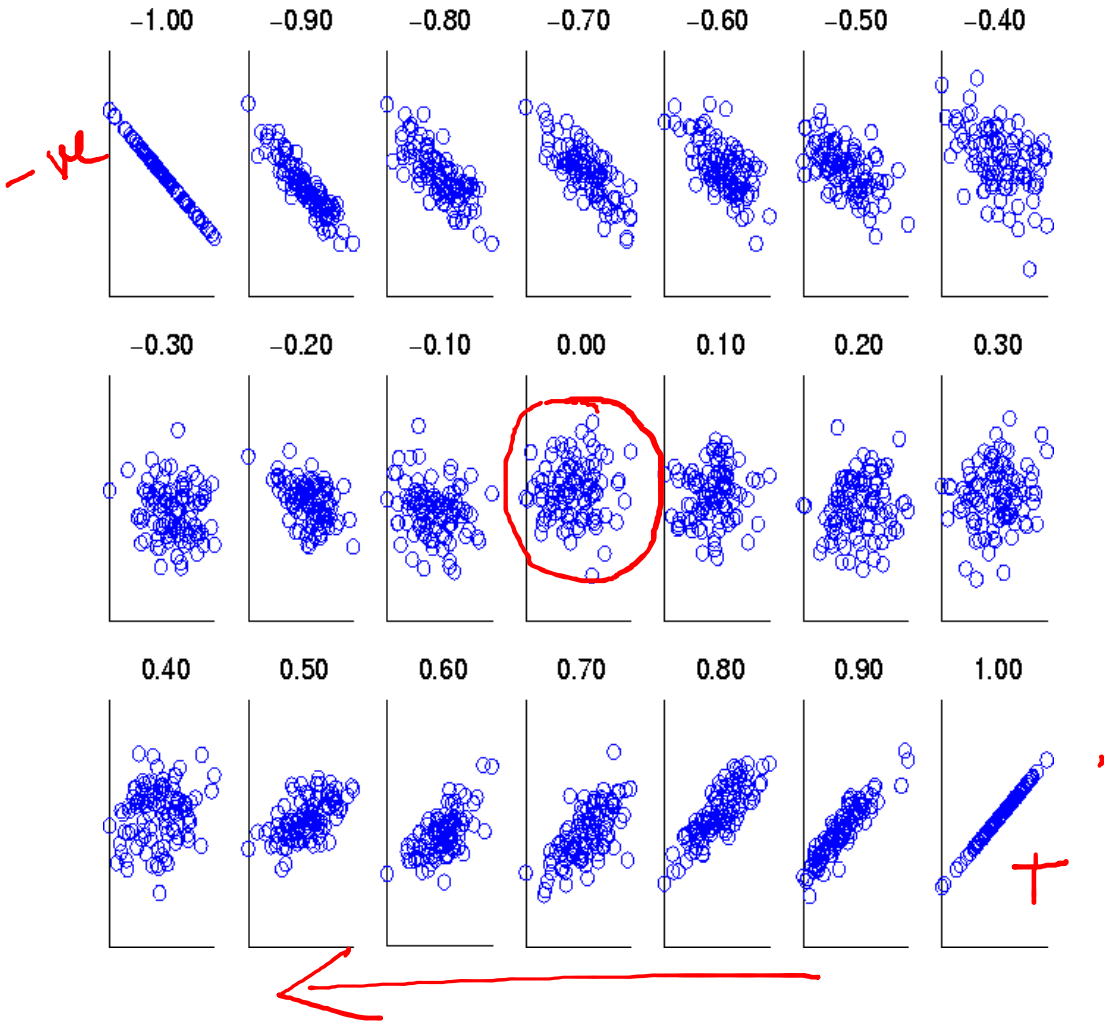
$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

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

- 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

DATA ANALYTICS

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Covariance (Numeric Data)

- Covariance is similar to correlation

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

Correlation coefficient:

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

where n is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

Covariance (Numeric Data)

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 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

Co-Variance: An Example

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

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

Expected value = mean

William Navidi

Co-Variance: An Example

Suppose two stocks A and B 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?

A	B
2	5
3	8
5	10
4	11
6	14

$E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$

$E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$

$Cov(A,B) = (2\times5+3\times8+5\times10+4\times11+6\times14)/5 - 4 \times 9.6 = 4$

Thus, A and B rise together since Cov(A, B) > 0.

$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$

$=$

2. 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 revenue2.
- 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

3. Tuple Duplication

- In addition to detecting redundancies between attributes, duplication should also be detected at the tuple level (e.g., where there are two or more identical tuples for a given unique data entry case).

- Inconsistencies often arise between various duplicates, due to inaccurate data entry or updating some but not all data occurrences.

- For example, if a purchase order database contains attributes for the purchaser's name and address instead of a key to this information in a purchaser database, discrepancies can occur, such as the same purchaser's name appearing with different addresses within the purchase order database.

4. Detecting and resolving data value conflicts

- For the same real world entity, attribute values from different sources are different
- Possible reasons: different representations, different scales, e.g., metric vs. British units
- For a hotel chain, the price of rooms in different cities may involve not only different currencies but also different services (e.g., free breakfast) and taxes.
- One university may adopt a quarter system, offer three courses on database systems, and assign grades from A+ to F, whereas another may adopt a semester system, offer two courses on databases, and assign grades from 1 to 10.

Exercise

A food court manager wants to know if there is a relationship between gender (male or female) and the preferred condiment on a burgers. The following table summarizes the results. Test the hypothesis with a significance level of 10%.

		Condiment			
		Ketchup	Mustard	Relish	Total
Gender	Male	15	23	10	48
	Female	25	19	8	52
	Total	40	42	18	100

observed.

significance level.

p_value

plausible

Chi square test

H_0 : independent

H_1 : dependent

$\chi^2 = \sum \frac{(Ob.-E)^2}{E}$ → find Expected value

df.

Exercise

The average share prices of two companies (X and Y)over the past 12 months are shown in Table . Calculate the Pearson correlation coefficient.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

DATA ANALYTICS

Exercise

- Explain how redundancy is handled in data integration.
- Compare and contrast Correlation and Covariance.

Chi-square table

	p value											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
	25%	20%	15%	10%	5%	2.5%	2%	1%	0.05%	0.025%	0.01%	0.005%
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.6	11.98	13.82	15.2
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.59	6.74	7.78	9.49	11.14	11.67	13.23	14.86	16.42	18.47	20
5	6.63	7.29	8.12	9.24	11.07	12.83	13.33	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.53	14.45	15.03	16.81	13.55	20.25	22.46	24.1
7	9.04	5.8	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.63	21.67	23.59	25.46	27.83	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42

$$df = (r-1)(c-1)$$
$$= (2-1)(2-1) = 1$$

References

Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.

Chapter 3: Data Preprocessing

Ref Book:

THANK YOU

Bharathi. R

Department of Computer Science

rbharathi@pes.edu