Data Mining (CS 451)

JAN-MAY’ 18

PRAGYA VERMA

RECAP

- Types of Data
- Importance of Data Pre-processing
- Data Cleaning

CONTENT

- Data Integration
  - Correlation Test for Nominal Data
    - Practice Question

Data Integration

- Data Mining often requires data integration – the merging of data from multiple data stores
- We need to be careful and ensure that we reduce and avoid redundancies and inconsistencies in the resulting data set.
- This can improve the accuracy and speed of the subsequent data mining process
- The semantic heterogeneity and structure of data pose great challenges in data integration

Data Integration (Contd.)

- How can we match schema and objects from different sources?
- For example, how can the data analyst be sure that *customer_id* in one database is same as *cust_no* in another?
- This problem is known **Entity Identification Problem**.
- In order to deal with it, you need to check if any attributes are correlated.

Redundancy and Correlation Analysis

- An attribute (such as annual revenue, age) may be redundant if it can be “derived” from another attribute or set of attributes
- Some redundancies can be detected by **correlation analysis**.
- Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.

Correlation Test for Nominal Data

- Let A and B be two attributes
- Suppose A has c distinct values, namely a_1, a_2, … , a_c. B has r distinct values, namely, b_1, b_2, …, b_r.
- We can represent this in the form of a **contingency table**, with c values of A making up the columns and the r values of B making up the rows.

Correlation Test for Nominal Data (Contd.)

- Let (A_i, B_j) denote the joint event that attribute A takes on value a_i and attribute B takes on value b_j.

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

- Where, o_{ij} = observed frequency (actual count) of the joint event (A_i, B_j); e_{ij} = expected frequency of (A_i, B_j) which can be comp

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

Correlation Test for Nominal Data (Contd.)

- Count ($A = a_i$) is the number of tuples having value a_i for A, and count ($B = b_j$) is the number of tuples having value b_j for B.
- Alpha level of significance denotes probability of rejecting the null hypothesis when it is true.
- Degree of Freedom = (no. of rows – 1) x (no. of columns – 1)
- Check the Chi-squared distribution table for the given values of degree of freedom and level of significance.
- If χ^2 is above the value in the table we can say that the attributes are strongly correlated.

Practice Question 1

Is gender independent of education level? A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Level of significance = 5 %

Next Class

- Correlation Test for Numerical Data