Name: Abhishek Srivastava

Registration Number: 19BCE10071

ACTIVITY- 5

Investigate the Handling of Redundancy in Data Integration.

Explore the usage of correlation analysis and covariance analysis towards eliminating the redundant attributes along with relevant computations and scenario analysis.

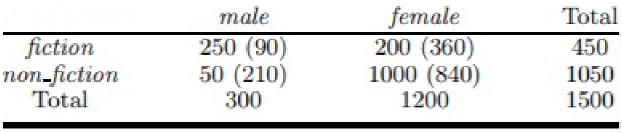
Data Redundancy- Redundancy is another important issue in data integration. An attribute (such as annual revenue, for instance) may be redundant if it can be "derived" from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

- Some redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data. For nominal data, we use the χ 2 (chisquare) test.
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality
- For numeric attributes, we can use the correlation coefficient and covariance, both of which access how one attribute's values vary with those of another.

Correlation Analysis towards eliminating the redundant attributes-

χ 2 Correlation Test for Nominal Data

Example - Suppose that a group of 1,500 people was surveyed. The gender of each person was noted. Each person was polled as to whether their 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.



	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

The expected frequencies are calculated based on the data distribution for both attributes using Equation

eij = count (A = ai) × count (B = bj)/n

Using Equation this, we can verify the expected frequencies for each cell. For example, the expected frequency for the cell (male, fiction) is

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90,$$

Name: Abhishek Srivastava

Registration Number: 19BCE10071

and so on. Notice that in any row, the sum of the expected frequencies must equal the total observed frequency for that row, and the sum of the expected frequencies in any column must also equal the total observed frequency for that column. Using Equation

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

for χ 2 computation, we get

$$\begin{aligned} \chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93. \end{aligned}$$

For this 2 × 2 table, the degrees of freedom are (2 − 1)(2 − 1) = 1. For 1 degree of freedom, the χ 2 value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the χ 2 distribution, typically available from any textbook on statistics). Since our computed value is above this, we can reject the hypothesis that gender and preferred reading are independent and conclude that the two attributes are (strongly) correlated for the given group of people.

Correlation Coefficient for Numeric Data

For numeric attributes, we can evaluate the correlation between two attributes, A and B, by computing the correlation coefficient (also known as Pearson's product moment coefficient, named after its inventer, Karl Pearson). This is

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

where n is the number of tuples, and are the respective means of A and B, σA and σB are the respective standard deviation of A and B, and Σ(aibi) is the sum of the AB cross-product.

If rA,B > 0, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.

rA,B = 0: independent; rAB < 0: negatively correlated

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

Name: Abhishek Srivastava

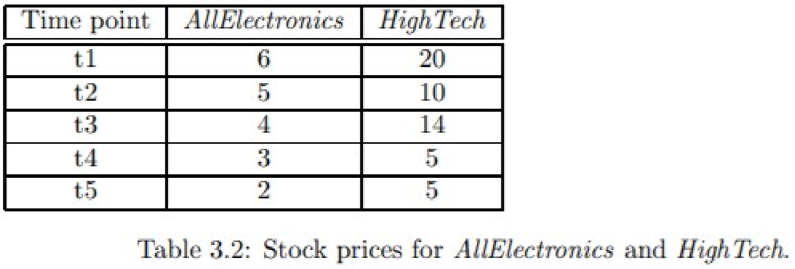
Registration Number: 19BCE10071

where n is the number of tuples, and are the respective mean or expected values of A and B, σA and σB are the respective standard deviation of A and B.

- Positive covariance: If CovA,B > 0, then A and B both tend to be larger than their expected values.
- Negative covariance: If CovA,B < 0 then if A is larger than its expected value, B is likely to be smaller than its expected value.
- Independence: CovA,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.

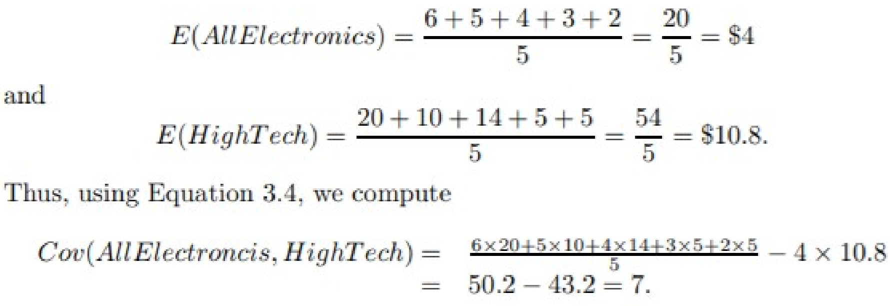
Example of Covariance analysis of numeric attributes

Consider Table below, which presents a simplified example of stock prices observed at five time points for AllElectronics and HighTech, some high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?



Time point	*AllElectronics*	*HighTech*
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Table 3.2: Stock prices for *AllElectronics* and *HighTech*.



$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

and

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.8.$$

Thus, using Equation 3.4, we compute

$$\begin{aligned} Cov(AllElectroncis, HighTech) &= \frac{6\times20+5\times10+4\times14+3\times5+2\times5}{5} - 4\times10.8 \\ &= 50.2 - 43.2 = 7. \end{aligned}$$

Therefore, given the positive covariance we can say that stock prices for both companies rise together.