INTRODUCTION TO DATA ANALYTICS

Class #6

Data Pre-processing

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology
IIIT Sri City

QUOTE OF THE DAY..

Nothing great was ever achieved without enthusiasm.

- RALPH WALDO EMERSON, American philosopher

TODAY'S DISCUSSION…

- Data Cleaning

- Data Integration

- Data Transformation

DATA CLEANING

DATA CLEANING

Real data are raw in the sense that they are incomplete, noisy and inconsistent. Data cleaning is the operation of finding and removing false or corrupt records from a database, and refers to identifying incorrect, irrelevant, incomplete, inaccurate, or parts of the data and then modifying, replacing, erasing false & misleading data.

Data cleaning includes the activity to

1) Fill missing values

2) Smooth out noise

3) Correct inconsistencies in the data

FILLING MISSING VALUES

It is usual for an object to be missing one or more attribute values. In some cases, the information was not collected. In other cases, some attributes are not applicable to all objects.

There are several strategies to deal with missing values.

1) Eliminate data objects or attributes
2) Estimate missing values

FILLING MISSING VALUES

Eliminate data objects or attributes:

This is the simple and effective strategy to eliminate objects or attributes with missing values.

- This method is not very effective, unless the objects contains several attributes with missing values.
- However, even a partially specified data object contains some information, and if many objects have missing values, then a reliable analysis is not ensured or impossible.
- It is especially poor when the percentage of missing values per attribute varies considerably.

FILLING MISSING VALUES

Estimating missing values:

- Use the attribute mean to fill in the missing values.
- Use a global constant (default value) to fill in the missing values.
- Use the attribute mean for all samples belonging to the same class.
- Use the most probable values (with regression, Bayesian formalism, decision tree induction, etc.)

✔ All these strategies are not foolproof strategies.

✔ Filled-in value may bias the data analysis, if they appear in large number or deviate from actual value.

✔ The last strategy is the best, however it is computationally expensive.

SMOOTHING NOISY DATA

Noise is a random error or variance in a measured variable. Following are the three noisy data smoothing techniques.

a) Binning
b) Regression
c) Clustering

SMOOTHING NOISY DATA

Binning: Binning methods smooth a sorted data value by consulting its neighborhood, i.e., values around it.

Example 1: Partitioning into equal frequency bins.

- First sort the data and then partition into equal frequency bins of a chosen size.
- Each value in the bin is replaced by the mean value of the bin.

Suppose given data values are 34, 28, 21, 24, 4, 8, 15, 21, 25

1. **Sort:** 4, 8, 15, 21, 21, 24, 25, 28, 34

2. **Bin** them into bins of size 3 (say).

   Bin 1 : 4, 8, 15
   Bin 2 : 21, 21, 24
   Bin 3 : 25, 28, 34

3. **Replace** each value in the bin by the **mean value** of the bin.

   Bin 1 : 9, 9, 9
   Bin 2 : 22, 22, 22
   Bin 3 : 29, 29, 29

Note: Smoothing by bin medians (or mode) can be employed.

SMOOTHING NOISY DATA

Example 2: Smoothing by bin boundaries

- Identify the minimum and maximum values in each bin as the bin boundaries.
- Each bin value is replaced by the closest boundary value.

Suppose given data values are 34, 28, 21, 24, 4, 8, 15, 21, 25

1. **Sort:** 4, 8, 15, 21, 21, 24, 25, 28, 34
2. **Bin** them into bins of size 3 (say).

   Bin 1 : 4, 8, 15
   Bin 2 : 21, 21, 24
   Bin 3 : 25, 28, 34

3. **Replace** each value in the bin by the **closest boundary value**.

   Bin 1 : 4, 4, 15
   Bin 2 : 21, 21, 24
   Bin 3 : 25, 25, 34

Note:

- **In general, the larger the width, the greater the effect of the smoothing.**
- **Binning is also used as a discretization technique.**

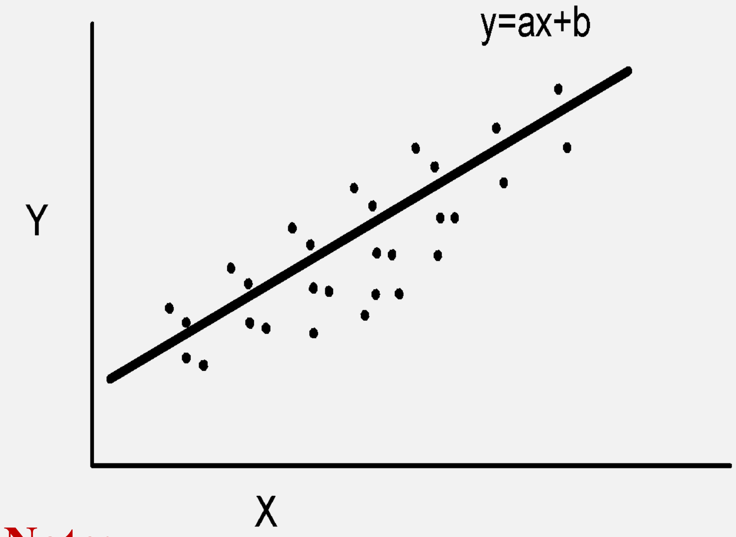
SMOOTHING NOISY DATA

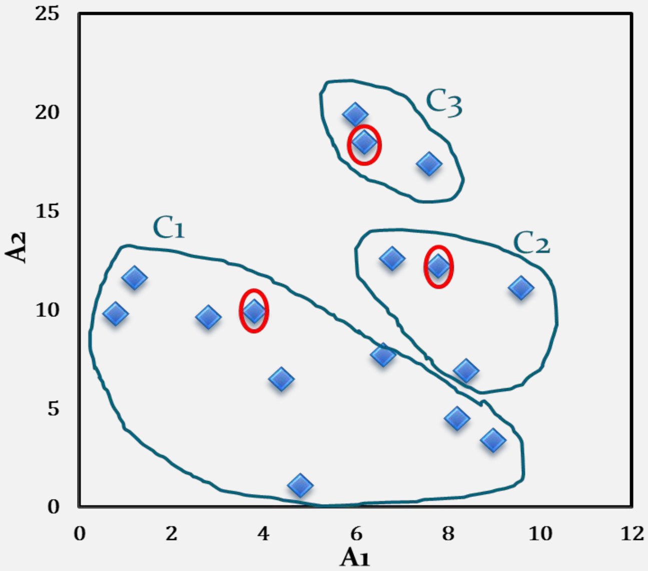
Regression:

Data can be smoothed by fitting the data to a function (the function can be derived with regression technique).

Clustering:

Similar values are organized into groups, or clusters. There are data, which does not belong to any cluster are called outliers (i.e., noisy data) and these can be adjusted to put them into their nearest clusters.





Note:

- **Regression and clustering will be discussed in detail in later lectures.**

CORRECT INCONSISTENCIES IN THE DATA

Outlier detection using Quartiles:

This technique tags two types of outliers: mild and extreme outliers.

Mild outliers: which are just outside the range

$\{Q_1 - 1.5 \times IQR, \dots, Q_3 + 1.5 \times IQR\}$

Extreme outliers: which are beyond the range

$\{Q_1 - 3 \times IQR, \dots, Q_3 + 3 \times IQR\}$

Where $IQR = Q_3 - Q_1$ and Q_1, Q_3 being thee first and third quartiles.

DATA INTEGRATION

DATA INTEGRATION

Data integration is merging data from various data sources. The major issue to be considered while integrating data is **redundancy**.

Correlation analysis to detect redundancy:

Some redundancy can be detected by correlation analysis. In other words. Given two attributes, such analysis can measure how strongly one attribute implies the other.

PEARSON’S CORRELATION ANALYSIS

KARL PEARSON’S CORRELATION COEFFICIENT

- This method is applicable to find correlation coefficient between two **numerical** attributes.

Definition: **Karl Pearson’s correlation coefficient**

Let us consider two attributes are X and Y.

The Karl Pearson’s coefficient of correlation is denoted by r^* and is defined as

$$r^* = \frac{\sum_{1=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N.\sigma_X.\sigma_Y}$$

where X_i = i – th value of X – variable

$\bar{X}$ = mean of X

Y_i = i – th value of Y – variable

$\bar{Y}$ = mean of Y

N = number of pairs of observation of X and Y

σ_X = standard deviations of X

σ_Y = standard deviation of Y

KARL PEARSON'S CORRELATION COEFFICIENT

- This is also called **Pearson's Product Moment Correlation** or **Galton coefficient.**
- **Note that:** $-1 \leq r^{*} \leq +1$

If $\boldsymbol{r^{*} > 0}$: The attributes (say, A and B) are positively correlated. This implies that the values of A increases as the value of B increases. Further, the higher the value, the stronger the connection. Thus, a higher value of r^{*} means either A or B can be removed due to redundancy.

If $\boldsymbol{r^{*} = 0}$: A and B are independent and there is no correlation between them.

If $\boldsymbol{r^{*} < 0}$: A and B are called negatively correlated. This implies that the values of one attribute increases as the value of other attribute decreases. In this case, also neither A nor B are redundant and hence none should be removed.

Note: If A and B are correlated, then this does not mean that A causes B. For Eg.: Say A = number of hospitals and B = number of car theft in a given city. Here, A is not a cause of B or vice-versa. Rather A, B may be linked to another attribute c.

X2-CORRELATION ANALYSIS

CHI-SQUARED TEST OF CORRELATION

- This method is also alternatively termed as Pearson's χ^2–test or simply χ^2 -test
- This method is applicable to categorical (discrete) data only.

  - Suppose, two attributes *A* and *B* with categorical values

$$A = a_1, a_2, a_3,\ldots.., a_m \quad \text{and}$$

$$B = b_1, b_2, b_3,\ldots.., b_n$$

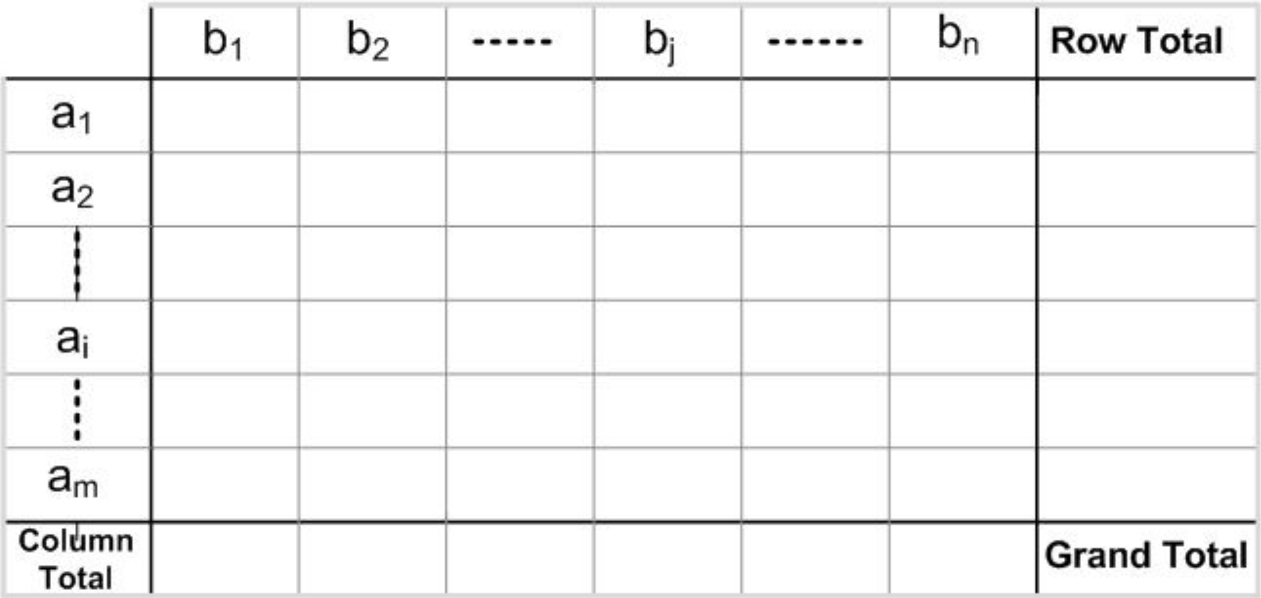
having *m* and *n* distinct values.

Between whom we are to find the correlation relationship.

X^2 –TEST METHODOLOGY

Contingency Table

Given a data set, it is customary to draw a contingency table, whose structure is given below.

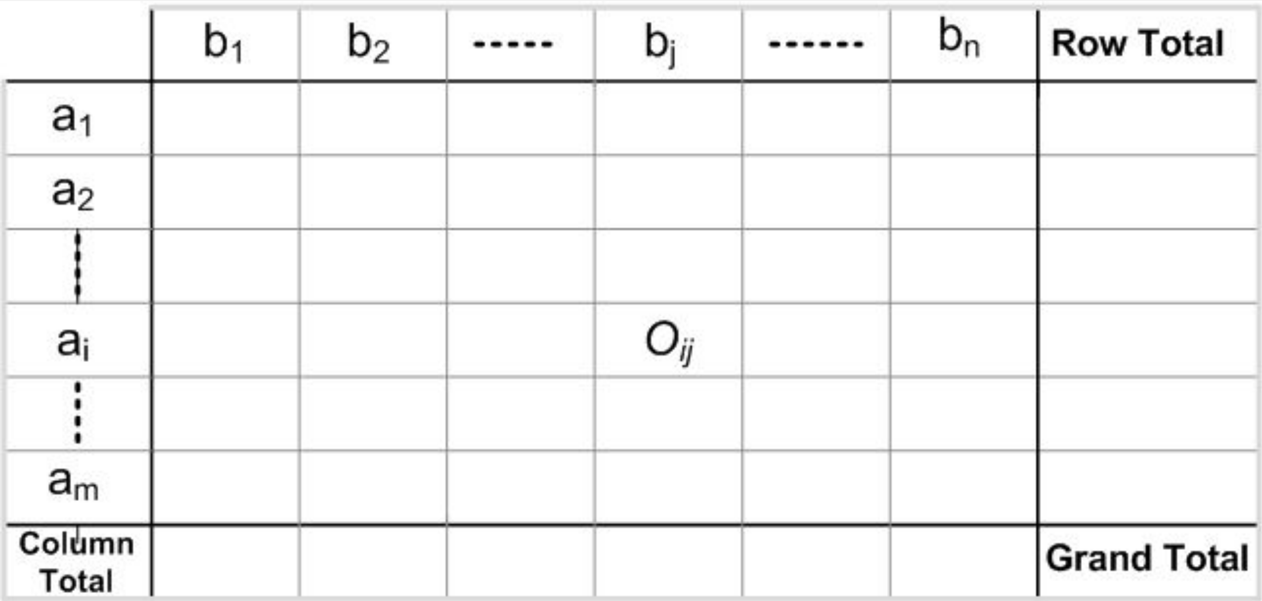


	b_1	b_2	-----	b_j	------	b_n	Row Total
a_1							
a_2							
⋮							
a_i							
⋮							
a_m							
Column Total							Grand Total

X^2 –TEST METHODOLOGY

Entry into Contingency Table: Observed Frequency

In contingency table, an entry O_{ij} denotes the event that attribute A takes on value a_i and attribute B takes on value b_j (i.e., $A = a_i$, $B = b_j$).



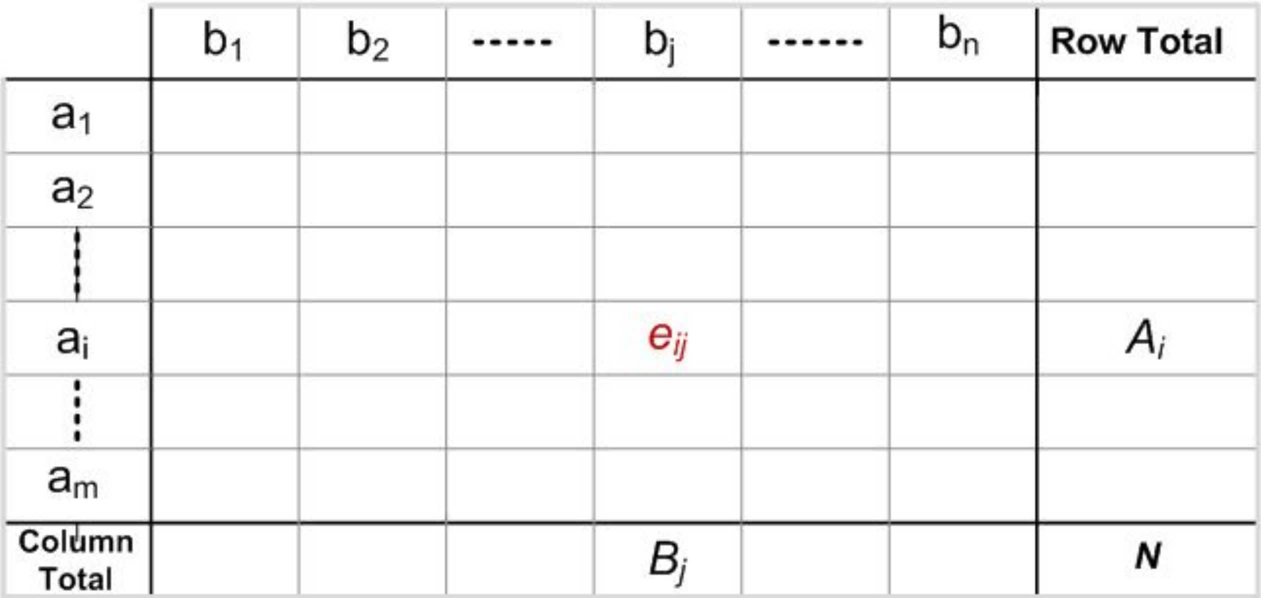
	b_1	b_2	-----	b_j	------	b_n	Row Total
a_1							
a_2							
⋮							
a_i				O_{ij}			
⋮							
a_m							
Column Total							Grand Total

X^2 –TEST METHODOLOGY

Entry into Contingency Table: Expected Frequency

In contingency table, an entry e_{ij} denotes the expected frequency, which can be calculated as

$$e_{ij} = \frac{Count(A = a_i)\times Count(B = b_j)}{Grand\ Total} = \frac{A_i \times B_j}{N}$$



	b_1	b_2	-----	b_j	------	b_n	**Row Total**
a_1							
a_2							
⋮							
a_i				e_{ij}			A_i
⋮							
a_m							
Column Total				B_j			N

X^2 – TEST

Definition: χ^2-**Value**

The $\boldsymbol{\chi}^2$ value (also known as the Pearson's $\boldsymbol{\chi}^2$ test) can be computes as

$$\boldsymbol{\chi}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

where o_{ij} is the **o**bserved frequency

e_{ij} is the **e**xpected frequency

X^2 – TEST

- The cell that contribute the most to the χ^2 value are those whose actual count is very different from the expected.

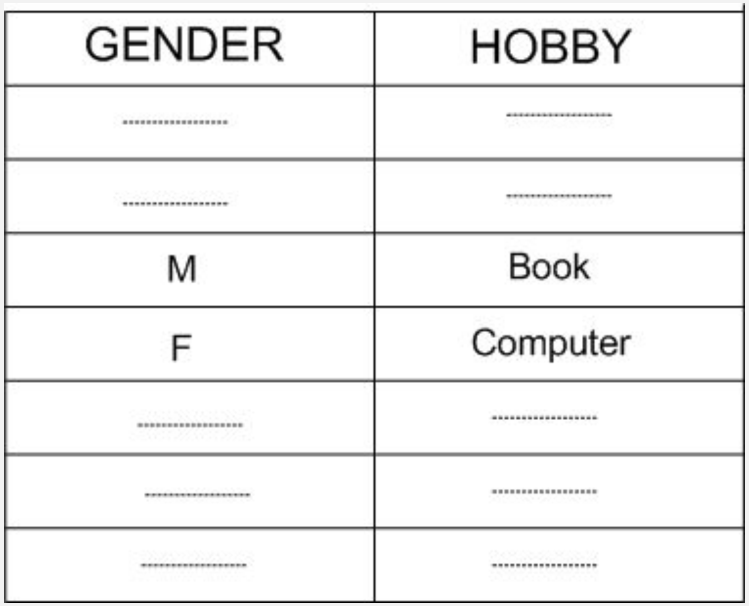
- The χ^2 statistics tests the hypothesis that A and B are independent. The test is based on a significance level, with $(n\text{-}1) \times(m\text{-}1)$ degrees of freedom., with a contingency table of size $n \times m$

- If the hypothesis can be rejected, then we say that A and B are statistically related or associated. (We will see this in detail later).

X^2 – TEST

Example 6.1: Survey on Gender versus Hobby.

- Suppose, a survey was conducted among a population of size 1500. In this survey, gender of each person and their hobby as either “book” or “computer” was noted. The survey result obtained in a table like the following.



GENDER	HOBBY
....	....
....	....
M	Book
F	Computer
....	....
....	....
....	....

- We have to find if there is any association between Gender and Hobby of a people, that is, we are to test whether “gender” and “hobby” are correlated.

X^2 – TEST

Example: Survey on Gender versus Hobby.

- From the survey table, the observed frequency are counted and entered into the contingency table, which is shown below.

		GENDER		
		Male	Female	Total
HOBBY	Book			
	Computer			
	Total			

X^2 – TEST

Example: Survey on Gender versus Hobby.

- From the survey table, the expected frequency are counted and entered into the contingency table, which is shown below.

		GENDER		
		Male	Female	Total
HOBBY	Book			
	Computer			
	Total			

X^2 – TEST

- Using equation for χ^2 computation, we get

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

$$= 507.93$$

- This value needs to be compared with the tabulated value of χ^2 (available in any standard book on statistics) with 1 degree of freedom (for a table of $m \times n$, the degrees of freedom is $(m-1)\times(n-1)$; here $m = 2$, $n = 2$).

- For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.01 significance level is 10.828. Since our computed value is above this, we reject the hypothesis that "Gender" and "Hobby" are independent and hence, conclude that the two attributes are *strongly correlated* for the given group of people.

X^2 – TEST

Example 6.2: Hypothesis on “accident proneness” versus “driver’s handedness”.

- Consider the following contingency table on car accidents among left and right-handed drivers’ of sample size 175.
- Hypothesis is that *“fatality of accidents is independent of driver’s handedness”*

		HANDEDNESS		
		Left-Handed	Right-Handed	Total
FATALITY	Non-Fatal			
	Fatal			
	Total			

- Find the correlation between Fatality and Handedness and test the significance of the correlation with significance level 0.1%.