Data preprocessing

T.R.Lekhaa
AP-IT
SNSCE

Data preprocessing

Why data preprocessing?

- Real world data can be incomplete, noisy and inconsistent form.

- These data needs to be preprocessed in order to help improve the quality of the data, and quality of the mining results.

Several data preprocessing techniques

- Data cleaning
  - – Applied to remove noise, inconsistent data
- Data integration
  - – Merges data from multiple sources
- Data reduction
  - – Reduce data size by aggregating, eliminating redundant features
- Data transformations
  - – Normalization – data are scaled to fall within a smaller range(0.0 to 1.0) -> improves accuracy & efficiency

Why preprocess the data?

- Factors comprising data quality
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - interpretability

Data preprocessing techniques/ major tasks in data preprocessing

- **Data cleaning**
  - Fill in missing values, smoothing noisy data, identifying or removing outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**

  - reduce the data size by aggregating, eliminating, or clustering etc
  - Strategy: Dimensionality Reduction :
    - Data encoding schemes are applied to obtain reduced or compressed data
    - Data Compression technique: Eg. Wavelet transforms, PCA (Principal Component Analysis)
    - Attribute subset selection: Eg. Removing irrelevant attributes
    - Attribute construction: Eg. Small set of more useful attributes derived from original set
  - Strategy: Numerosity Reduction
    - Data replaced by alternative, smaller representation using parametric models (eg. Regression or log-linear models) or nonparametric models (e.g histograms, clusters, sampling or data aggregation)
- **Data transformation**
  - The data are transformed or consolidated into forms appropriate for mining.
  - Normalization, Data discretization, concept hierarchy

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"

How to Handle Missing Data?

- Ignore the tuple
  - (If Class label miss), not effective
- Fill in the missing value manually
  - Time consuming & not feasible
- Fill in it automatically with
  - a global constant (unknown or infinity symbol), simple but mining program consider unknown as one class
  - Use a measure of central tendency for the attribute mean (eg mean or median)
    - Mean – symmetric data
    - Median – skewed data (positive skewed – values < median, negative skewed – values > median)
  - the attribute mean for all samples belonging to the same class
    - eg if classified customer according to credit risk, may replace the missing value with mean value for customers in same credit risk category
    - If data distribution skewed – median value is better choice
  - the most probable value
    - Determined based on regression, inference-based tools using Bayesian or decision tree induction

Noisy Data

- What is noise?
- Random error or variance in a measured variable

How to Handle Noisy Data?

- Binning
  - Smooth a sorted data value by consulting its neighborhood i.e the values around it.
  - first sort data and partition into (equal-frequency) bins or buckets
  - 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
  - Linear Regression –finding best line to fit two attributes or variables so that one attribute used to predict other
  - Multiple Linear Regression – more than two attributes involved
- Outlier Analysis – detected by Clustering
  - detect and remove outliers

- Data smoothing methods are also used for
  - Data discretization (transformation) & data reduction

Binning - Example

- ❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equal-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

Data Cleaning as a Process

- Data discrepancy detection

  - first step in data cleaning

Caused by several factors: poorly designed data entry forms , human errors in data entry, deliberate errors (don’t like to give infor.) data decay(outdated addresses), errors in instrumentation devices

- – Uses the knowledge of metadata
- – Check unique rule, consecutive rule and null rule
- – Use commercial tools
  - Data scrubbing: use simple domain knowledge (e.g., spell-check) to detect errors and make corrections
  - Data auditing: by analyzing data to discover rules and relationship (e.g., correlation and clustering to find outliers)

- Data transformation
  - – Data migration tools: allow transformations to be specified eg. Replace the string “roll no” by “serial no”
  - – ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface

Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
  - Integrate metadata from different sources
- Entity identification problem:
  - Issue in integration: **Schema Integration and Object matching**
  - How equivalent real-world entities from multiple data sources be matched? – entity Identification Problem
  - Same entity can be represented in different forms, e.g., customer-id == cust-number
  - Metadata can be used to avoid errors in schema integration
- Redundancy and correlation analysis – is another issue. It can be detected by correlation analysis
- Tuple Duplication
- 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(imperial units)

Redundancy and Correlation Analysis

- Correlation analysis
- Given two attributes, correlation analysis measure how strongly one attribute implies on another.
- Nominal Data - X^2 (chi-square) test
- Numeric attributes – correlation coefficient & covariance

X^2 correlation for Nominal Data

- Correlation between two attributes A and B discovered by chi-square test
- A has c distinct values namely, $a_{1,}a_{2,}.....\ a_c$
- B has r distinct values namely $b_{1,}\ b_{2,}...\ b_r$
- Data tuples between A and B shown as contingency table, with c values of A making up columns and r values of B making up rows.

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

The larger the X^2 value, the more likely the variables are related

Contd..,

Observed frequency – actual count

Expected frequency can be calculated as

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

N – number of data tuples

$count(A = a_i)$ - no. of tuples having value a_i

$count(B = b_j)$ - no. of tuples having value b_j

Example – chi-square calculation

- Problem:
  - 1500 people
  - Gender of each person noted
  - Preferred type of reading material – fiction or nonfiction
- Two attributes: gender & preferred_reading
- Observed frequency from contingency table:

	Male	Female	Total
Fiction	250(90)	200(360)	450
Non-fiction	50(210)	1000(840)	1050
Total	300	1200	1500

- Expected frequency of cell(male, fiction) is as follows

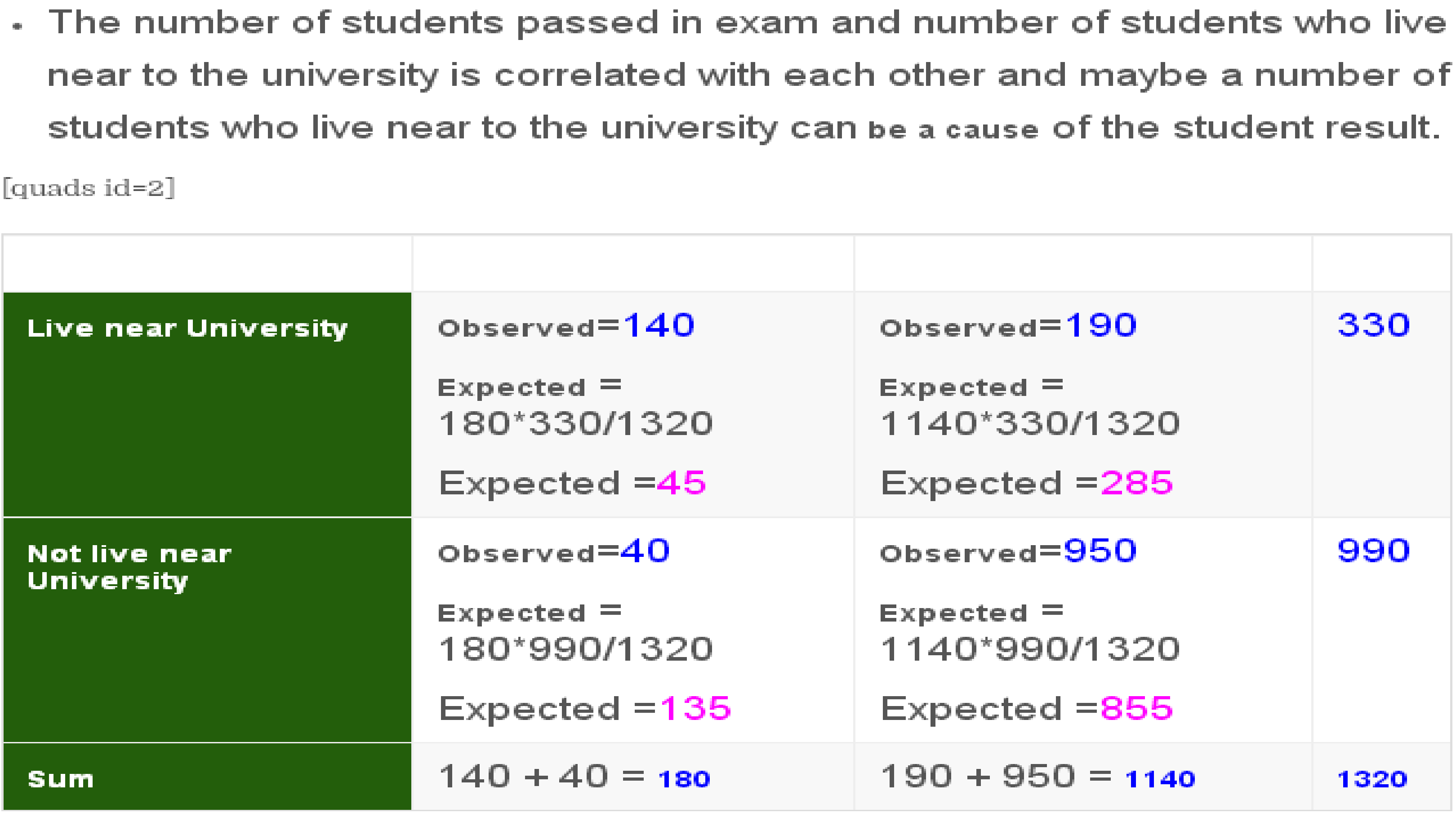
e_{11} = [count(male) * count(fiction)] / n = (300*450)/1500 = 90

- Chi-square calculation:

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

- 2*2 table, DF(Degree of Freedom) = (2-1)*(2-1) = 1
- For DF =1, value needed to reject hypothesis at 0.001level is 10.828
- Hence calculated value is (507.93) > 10.828, can reject gender and χ^2 preferred_reading attributes are independent and conclude that they are strongly correlated

Example



- The number of students passed in exam and number of students who live near to the university is correlated with each other and maybe a number of students who live near to the university can be a cause of the student result.

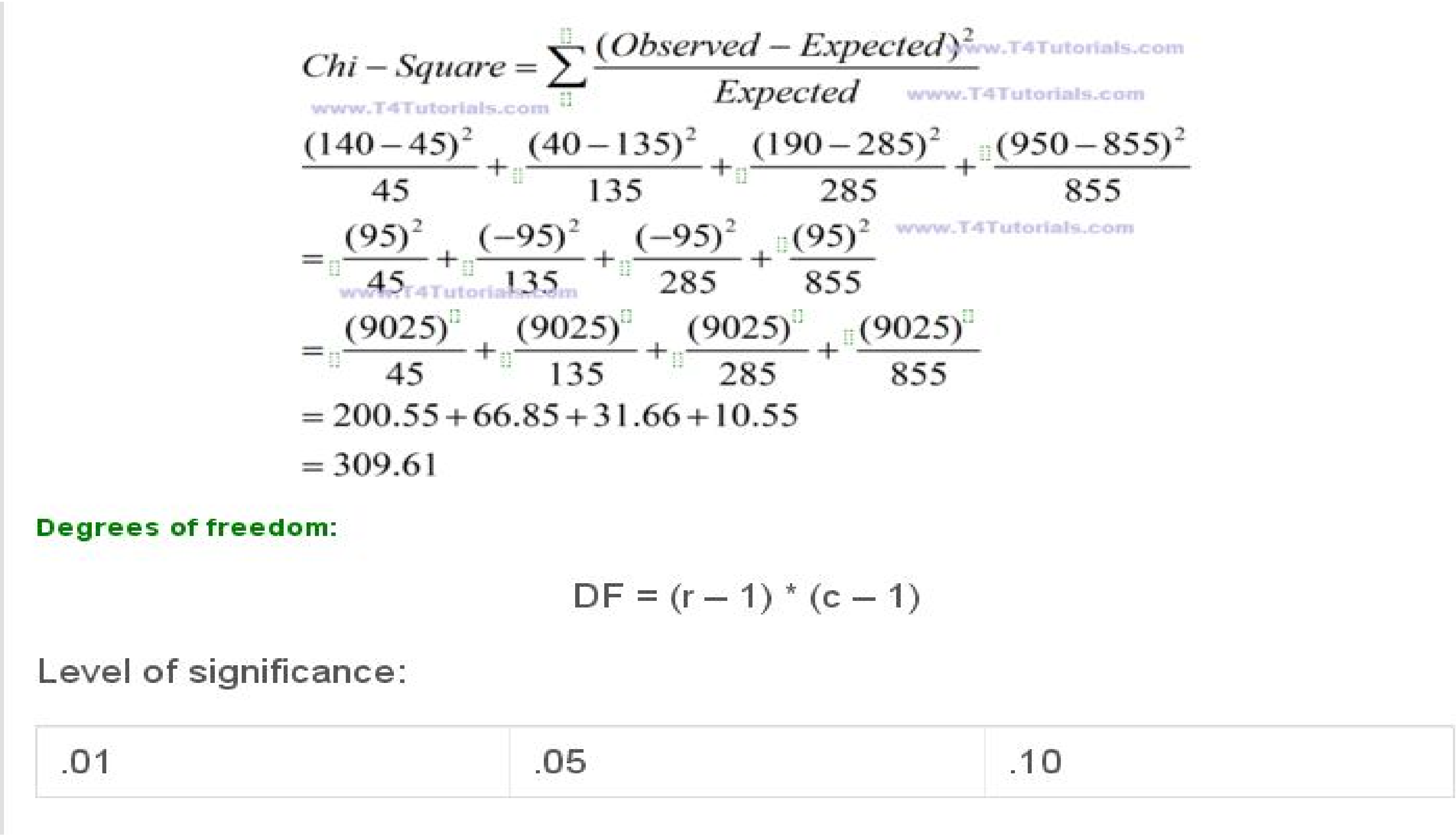
[quads id=2]

Live near University	Observed=140
	Expected = 180*330/1320
	Expected =45
		Observed=190
		Expected = 1140*330/1320
		Expected =285
			330

Not live near University	Observed=40
	Expected = 180*990/1320
	Expected =135
		Observed=950
		Expected = 1140*990/1320
		Expected =855
			990

Sum	140 + 40 = 180	190 + 950 = 1140	1320

Solution



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

$$\frac{(140-45)^2}{45} + \frac{(40-135)^2}{135} + \frac{(190-285)^2}{285} + \frac{(950-855)^2}{855}$$

$$= \frac{(95)^2}{45} + \frac{(-95)^2}{135} + \frac{(-95)^2}{285} + \frac{(95)^2}{855}$$

$$= \frac{(9025)^{\square}}{45} + \frac{(9025)^{\square}}{135} + \frac{(9025)^{\square}}{285} + \frac{(9025)^{\square}}{855}$$

$$= 200.55 + 66.85 + 31.66 + 10.55$$

$$= 309.61$$

Degrees of freedom:

DF = (r – 1) * (c – 1)

Level of significance:

.01	.05	.10

Data Integration

- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data reduction

- **Data reduction**: 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

Data reduction strategies

- Data cube aggregation – where aggregation operations are applied to the data for construction of a data cube.
- Attribute subset selection – reduces the data set size by removing irrelevant or redundant attributes. Goal – to find a minimum set of attributes.
- Dimensionality reduction, e.g., remove unimportant attributes
  - Wavelet transforms
  - Principal Components Analysis (PCA)
  - Feature subset selection
- Numerosity reduction (some simply call it: Data Reduction)
  - Regression and Log-Linear Models
  - Histograms, clustering, sampling
  - Data cube aggregation
- Data compression

Data reduction

Attribute subset selection:

- stepwise forward selection
- Stepwise backward elimination
- Combination of forward selection and backward elimination

Data reduction

- Dimensionality reduction
  - Data encoding or transformations are applied so as to obtain reduced or compressed representation of the original data.
  - Lossless- if the original data can be reconstructed from the compressed data without any loss of information
  - Lossy- if the original data can be reconstructed from the compressed data with loss of information.

Data Transformation

- The data are transformed or consolidated into forms appropriate for mining.
- Methods
  - Smoothing: Remove noise from data
  - 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 [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  – Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 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 μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

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

Where *j* is the smallest integer such that Max(|ν'|) < 1