Chapter 3. Data Processing: Data Cleaning

Meng Jiang

Data Science

Why? Data Quality Issues

- Measures for data quality: A multidimensional view
  - _____
  - _____
  - Completeness: not recorded, unavailable, …
  - _____
  - _____
  - _____

Why? Data Quality Issues

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Believability: how trustable the data are correct?
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, …
  - Timeliness: timely update?
  - Interpretability: how easily the data can be understood?

Data Preprocessing

- **Data cleaning**
- Data integration
- Data reduction
- Dimensionality reduction

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = " " (missing data)
  - _____
  - _____
  - _____

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and 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"
    - Was rating "1, 2, 3", now rating "A, B, C"
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data were not entered due to misunderstanding
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple
- Fill in the missing value manually: tedious + infeasible?

How to Handle Missing Data?

- Ignore the tuple
- Fill in the missing value manually: tedious + infeasible?
- Fill in it *automatically* with
  - _____
  - _____
  - _____
  - _____

How to Handle Missing Data?

- Ignore the tuple
- Fill in the missing value manually: tedious + infeasible?
- Fill in it *automatically* with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention
- Other data problems
  - Duplicate records
  - Incomplete data
  - Inconsistent data

How to Handle Noisy Data?

- Binning
  - First sort data and partition into (equal-frequency) bins
  - 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
- Clustering
  - Detect and remove outliers
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

Data Preprocessing

- Data cleaning
- **Data integration**
- Data reduction
- Dimensionality reduction

Data Integration

- Data integration
  - Combining data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- 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

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 revenue
- Redundant attributes may be able to be detected by *correlation analysis (often for categorical attributes)* and *covariance analysis (often for numerical attributes)*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis

	Play chess	Not play chess	Sum (row)
Like science fiction			450
Not like science fiction			1050
Sum(col.)	300	1200	1500

Correlation Analysis

	Play chess	Not play chess	Sum (row)
Like science fiction	90	360	450
Not like science fiction	210	840	1050
Sum(col.)	300	1200	1500

How to derive "90"?

450/1500 * 300 = 90

Correlation Analysis

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

Correlation Analysis

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

observed

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

expected

- **Null hypothesis:** The two distributions are independent
- The cells that contribute the most to the X^2 value are those whose actual count is different from the expected count
  - The larger the X^2 value, the more the null hypothesis of independence is rejected, and the more likely the variables are related

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

Example: Chi-Square Calculation

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

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

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

We can reject the null hypothesis of independence at a confidence level of 0.001.

- It shows that like_science_fiction and play_chess are correlated.

Example: Chi-Square Calculation

Degrees of freedom (df)	χ^2 value[19]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Correlation Analysis

- Note: Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population
- Causal analysis

Effective Promotional Strategies Selection in Social Media: A Data-Driven Approach by K. Kuang, M. Jiang, P. Cui, J. Sun, S. Yang. IEEE Transactions on Big Data (TBD), 2017.

Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing by K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2017.

Variance for Single Variable (Numerical Data)

- The variance of a random variable *X* provides a measure of how much the value of *X* deviates from the mean or expected value of *X:*

$$\sigma^2 = \mathrm{var}(X) = E[(\mathrm{X} - \mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

  - where σ^2 is the variance of X, σ is called *standard deviation*

    μ is the mean, and **$\mu = E[X]$** is the expected value of X
  - That is, variance is the expected value of the square deviation from the mean
  - It can also be written as:

$$\sigma^2 = \mathrm{var}(X) = E[(\mathrm{X} - \mu)^2] = E[\mathrm{X}^2] - \mu^2 = E[\mathrm{X}^2] - [E(x)]^2$$

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

  where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1; similarly for μ_2

- **Positive covariance:** If $\sigma_{12} > 0$
- **Negative covariance**: If $\sigma_{12} < 0$
- **Independence**: If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true
  - Some pairs of random variables may have a covariance 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: Calculation of Covariance

- Suppose two stocks X_1 and X_2 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?

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 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?
- Covariance formula

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

- Its computation can be simplified as: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
  - $E(X_1) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$
  - $\sigma_{12} = (2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14) / 5 - 4 \times 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

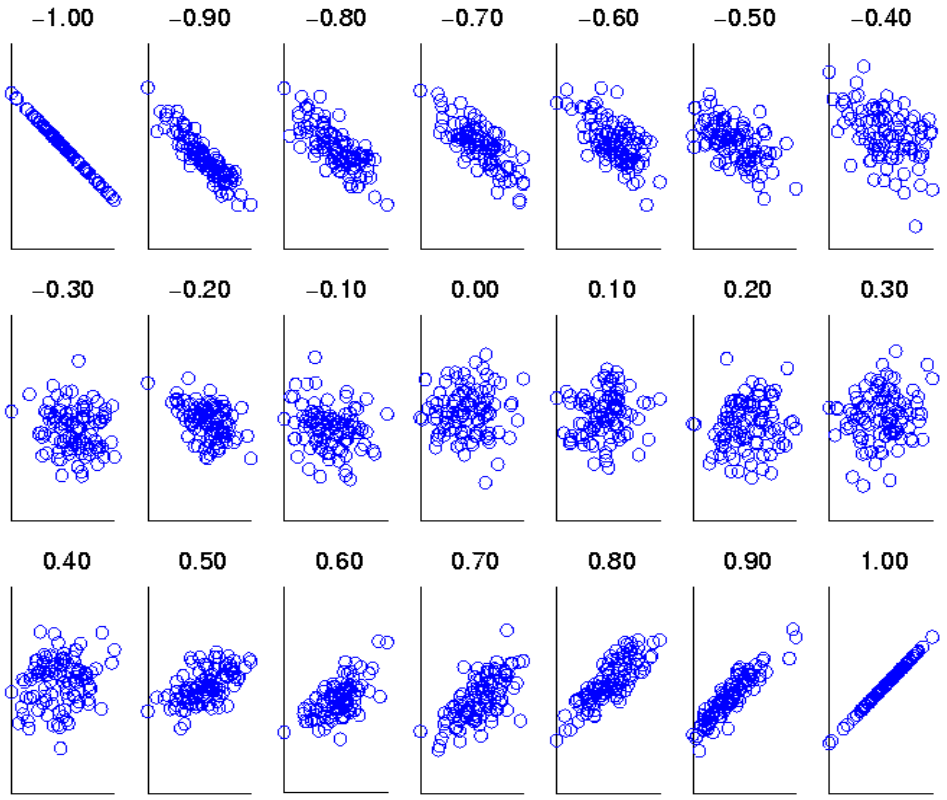
- **Correlation** between two variables X1 and X2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

- If ρ12 > 0: A and B are positively correlated (X1's values increase as X2's)
  - The higher, the stronger correlation
- If ρ12 = 0: independent (under the same assumption as discussed in co-variance)
- If ρ12 < 0: negatively correlated

Visualizing Changes of Correlation Coefficient

- Correlation coefficient value range: [–1, 1] Can you prove the range?
- A set of scatter plots shows sets of points and their correlation coefficients changing from –1 to 1



Covariance Matrix

- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2 * 2 covariance matrix as

$$\Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = E[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix}(X_1 - \mu_1 \quad X_2 - \mu_2)]$$

$$= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

- Generalizing it to *d* dimensions, we have,

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \mathbf{\Sigma} = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Discussion

- Can you use Chi-Square or p-value (doing correlation analysis) to select meta paths (as features) for relationship prediction?

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4*
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995