Data Management

CT051-3-M

Topic 4 – Data Preprocessing – PART 2

Topic & Structure of Lesson

- Major tasks in data preprocessing
  - Data Integration
  - Data Transformation

Data Preprocessing

- Why preprocess the data?
- Data integration
- Data transformation

Data Integration

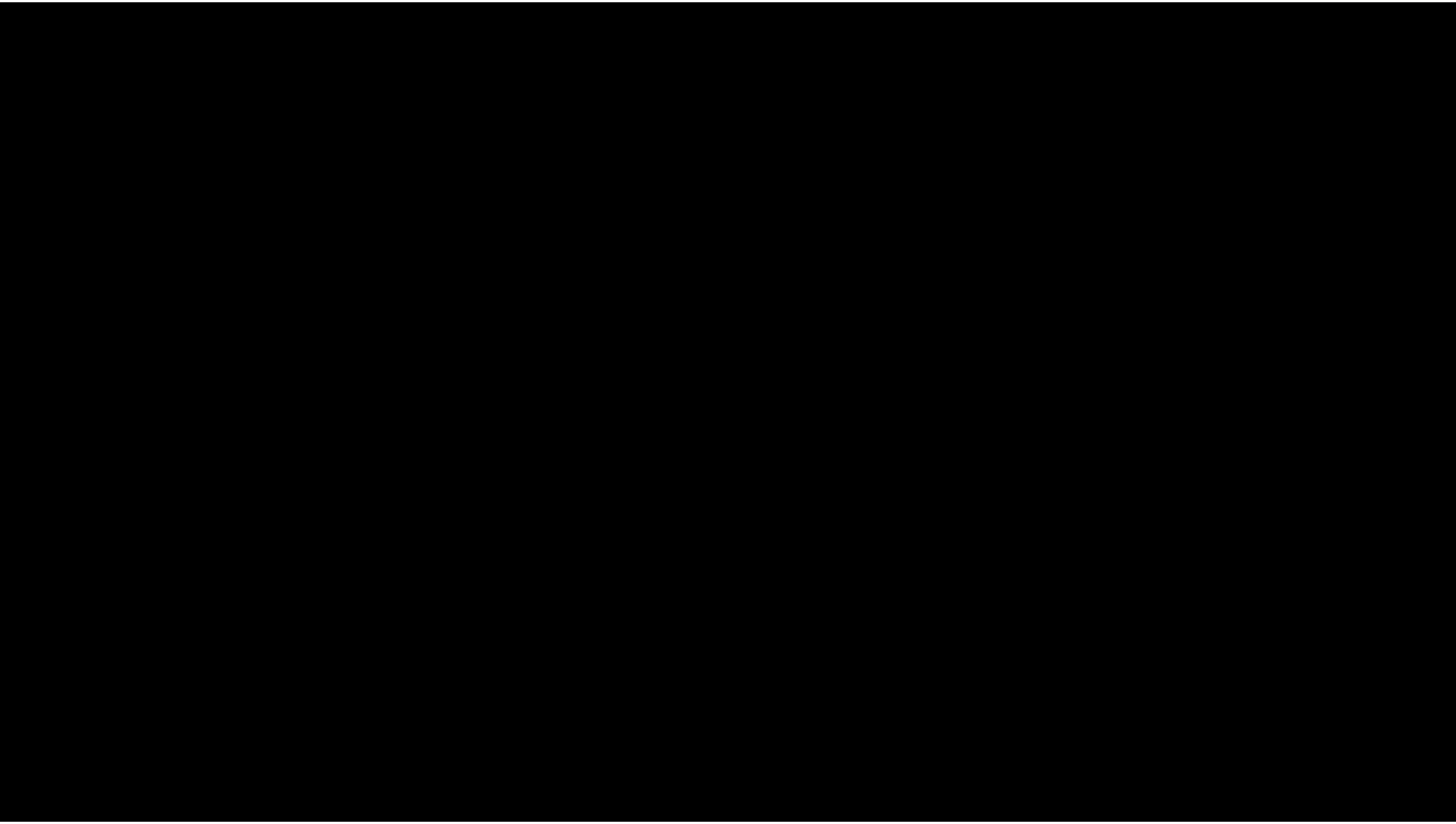
- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value 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 Redundant Data in Data Integration

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a “derived” attribute in another table, e.g., annual revenue

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

Data Integration: A Strategy for the Future



Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation

Data Transformation

One of the important factor to decide difference between powerful and useless model.

Two Objective:

1. Generate new variable - “Analytical/Modeling/Secondary Variables”
2. Fix any potential problems such as missing value and skewed variable distribution.

Examples of Analytical Variables

- Average transaction value over a certain period (month, quarter, year)
- Average number of transactions over a certain period (month, week, etc.)
- Ratio of purchases from a specific product or product category to total purchases
- Ratio of total purchases in a time period to available credit in the same period
- Ratio of each purchase to the cost of shipment (specially when shipping costs are paid by the customer)
- Ratio of number/value of purchased products to size of household

Examples of Analytical Variables Contd…

- Ratio of price of most expensive product to least expensive product
- Ratio of largest to smallest purchase in a time period
- Ratio of total household debt to total household annual income
- Ratio of largest income in a household to total household income
- Ratio of available credit to household income
- Average annual income per household member
- Ratio of debt to available credit
- Average percentage of used portion of credit over a certain period

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing, values for numerical attributes, like *age, may be mapped to higher-level* concepts, like *youth, middle-aged, and senior.*
- Feature Scaling
  - Normalization (Rescaling): scaled to fall within a small, specified range [0…1]
  - Mean Normalization
  - Standardization

Data Transformation: Normalization

- min-max normalization

$$\text{Normalized value} = (x - x_{\min})/(x_{\max} - x_{\min}),$$

where, x, $x_{\min}$, and $x_{\max}$ are income, minimum income, and maximum income, respectively.

Seven income values.

15,000	56,000	22,000	26,000	34,000	44,000	46,000

Normalized income values.

0	1.00	0.17	0.27	0.46	0.71	0.76

Data Transformation: Mean Normalization

Mean normalization:

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$

Where, x is an original value, x' is the normalized value.

Data Transformation: Standardization

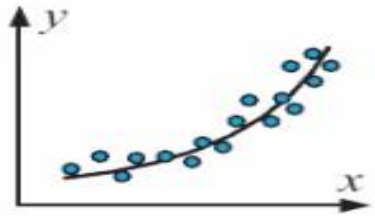
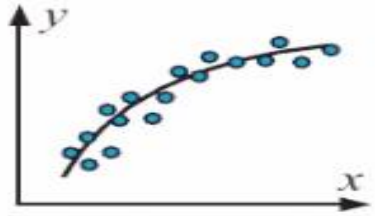
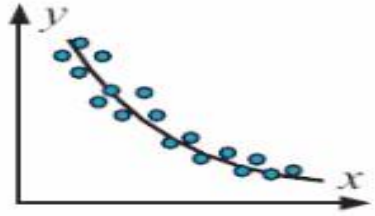
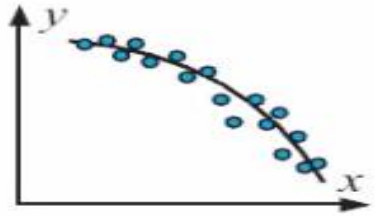
- Standardization (or Z-score normalization or Mean Removal)
- The features will be rescaled so that they'll have the properties of a standard normal distribution with

μ=0 and σ=1

where μ is the mean (average) and σ is the standard deviation from the mean

$$z = \frac{x - \mu}{\sigma}$$

Data Transformation: Using Square, Log or Reciprocal

Curve	Transformations to test
	
	
	
	

Changing the variable distribution

- Changing the distribution of a variable, independent or dependent, can result in significant change in model performance.
- Usually performed to uncover the relationships that were masked by the variable distribution.

Changing the variable distribution

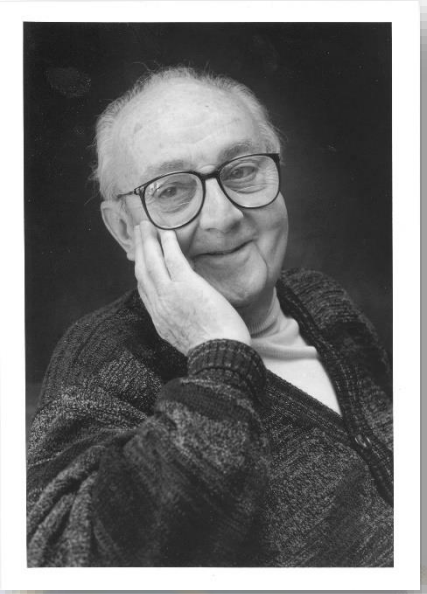
- Three methods for transformation (Continuous variable)
  - Rank Transformation
  - Box-Cox Transformation
  - Histogram

Rank Transformation

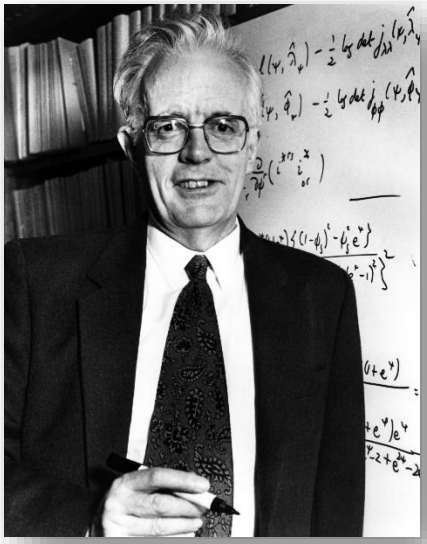
- Simplest method – Valid ONLY for continuous variable
- It just replace the values with Rank

Box-Cox Transformation

George Box



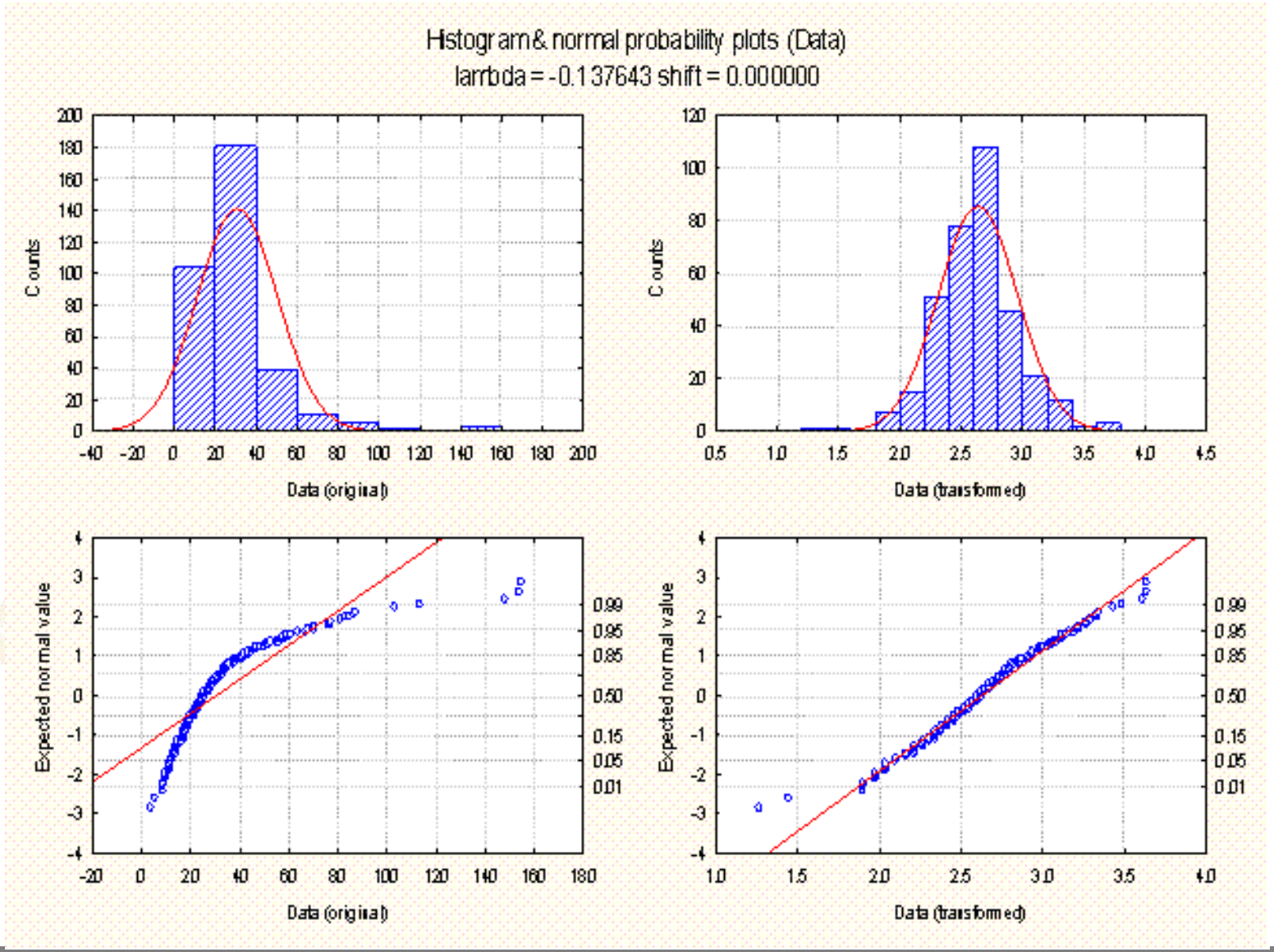
David Cox



- Attempts to transform a continuous variable into an *almost* normal distribution.
- The Box-Cox transformation of the variable x is also indexed by λ, and is defined as

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda}.$$

Box Cox Transformation - Example



Further Reading

- http://onlinestatbook.com/2/transformations/log.html

Question & Answer Session

A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Q & A