Data Engineering

204426

Data Transformation

Outline

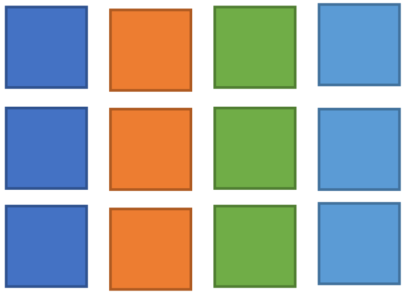
- Definition of Data Transformation
- Data Representation
- Types of Data
- Encoding of Categorical Data
- Normalization and Standardization

Data Transformation

- Process of converting data from one format to another
  - Typically, from the format of a source system into the required format of a destination system.
- Can be simple or complex based on the required changes
- Data transformation can be divided into the following steps:
  1. **Data discovery**: identifying and understanding the data in its source format.
  2. **Data mapping**: defining how individual fields are mapped, modified, joined, filtered, aggregated etc. to produce the final desired output.
  3. **Code generation**: generating executable code that will transform the data based on the desired and defined data mapping rules.
  4. **Code execution**: the generated code is executed against the data to create the desired output.
  5. **Data review:** ensuring the output data meets the transformation requirements.

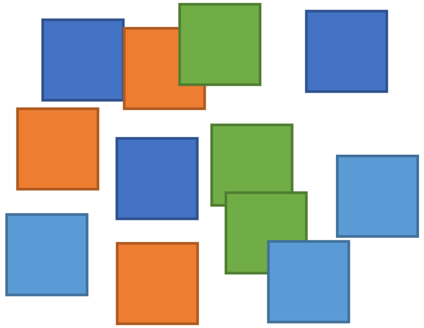
Data Discovery

- What we should identify:
  - Data Structure



Structured Data

- Stands for information that is highly organized, factual, and to-the-point
- Comes in the form of letters and numbers that fit nicely into the rows and columns of tables
- Exists in a format of relational databases



Unstructured Data

- Doesn’t have any pre-defined structure to it
- Comes in all its diversity of forms
- opt for non-relational databases

Data Discovery

- What we should identify:
  - Variables
    - Data type
    - Data format: text, integer, floating point, date-time, currency, etc.
    - Range of data
    - Units of measurement

Types of Data

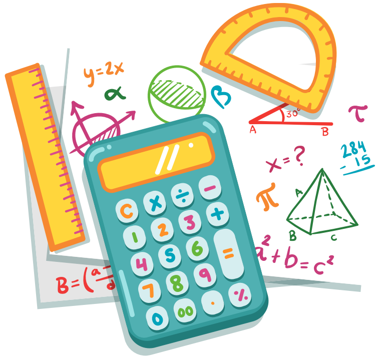
Quantitative Data

- This data can be described using **numbers.**
- **Basic mathematical procedures** are possible on the set.

Qualitative Data

- This data cannot be described using numbers and basic mathematics.
- This data is generally described using natural **categories and language**.

Types of Data



Numeric Attributes - Quantitative

- One that has a real-valued or integer-valued domain.
- Such as age, height, grade, frequency, etc.

Categorical Attributes

- One that has a set-valued domain composed of a set of symbols.
- Such as Gender = {M,F},
  Education = {High School, BS, MS, PhD}, etc.

Discrete

- Take on a finite or countably infinite set
- Such as integer, grade, number of object, etc.

Continuous

- Take on any real value
- Such as height, weight, size, etc.

Types of Data

Categorical Attributes

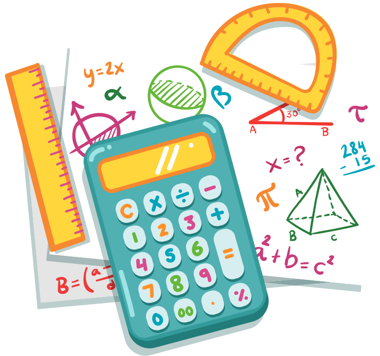
Nominal

- Attribute values in the domain are unordered.
- Can only equality (=) compare.
- Such as gender, type of hair, etc.

Ordinal

- Attribute values are ordered.
- Can both equality (=) and inequality (<, >) compare.
- Such as education, feel (unhappy, OK, happy), etc.

Types of Data



Numeric Attributes

Interval-scaled

- Can compute only differences (addition or subtraction)
- For example, temperature measured in ˚C or ˚F.
  - If it is 20 ˚C on one day and 10 ˚C on previous day
  - We **can** talk about a temperature drop of 10˚C.
  - We **cannot** say that it is twice as cold as the previous day.

Ratio-scaled

- Can compute both differences and ratio between values,
- For example, age.
  - If Jone is 20 years old and Jim is 10 years old.
  - We **can** say that Jone older than Jim with 10 years.
  - We **can** say that Jone is twice as old as Jim.

Types of Data

Summary of data types and scale measures

Provides	Nominal	Ordinal	Interval-scaled	Ratio-scaled
The order of values is known		/	/	/
“Count,” aka “Frequency of Distribution”	/	/	/	/
Mode	/	/	/	/
Median		/	/	/
Mean			/	/
Can quantify the difference between each values			/	/
Can add or subtract values			/	/
Can multiple and divide values				/
Has “true zero”				/

https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

Data Mapping

Translation and Mapping

- Representing codes can be mapped to the relevant values
  - Map faculty code 05 to “Science”
  - Map abbreviate month name to full name
- Converts data from formats used in one system to formats appropriate for a different system.
  - Web data might arrive in the form of hierarchical JSON or XML files, but need to be translated into row and column data.
  - Convert format of date-time data.

Data Mapping

Filtering

- To distill a data source to only what the user needs by eliminating repeated, irrelevant, or overly sensitive data.
- In its most practical form, data filtering simply involves the selection of specific rows, columns, or fields to display from the dataset.
- Example:
  - If the end-user doesn’t need to see the addresses or Social Security numbers of each client in the report, data filtering will scrub them from the report.

Data Mapping

Aggregation and Summarization

- The process where raw data is gathered and expressed in a summary form for statistical analysis.
  - Average
  - Minimum, Maximum
  - Summation, Count

Data Mapping

Aggregation and Summarization

- Example
  - Transforming a time series of customer transactions to hourly or daily sales counts.
  - Companies often collect data on their online customers and website visitors. The aggregate data would include statistics on customer demographic and behavior metrics, such as average age or number of transactions.
- Data aggregation is any process in which data is brought together and conveyed in a summary form. It is typically used before the performance of the statistical analysis.

Data Mapping

Attribute Construction

- New attributes are constructure and added from the given set of attributes.
- Example
  - Constructure and add BMI attribute calculated by using height and weight attributes

Data Mapping

Attribute Construction

- New attributes are constructure and added from the given set of attributes.
- Example
  - Constructure and add BMI attribute calculated by using height and weight attributes

Data Mapping

Discretization

- Converting continuous data attribute values into a finite set of intervals
- Association with each interval some specific data value

Data Mapping

Enrichment

- Data from different sources can be merged to create denormalized, enriched information.
- Long or freeform fields may be split into multiple columns.
- Example:
  - A customer’s transactions can be rolled up into a grand total and added into a customer information table
  - Full name be split into first name, middle name and last name.

Data Mapping

Normalization

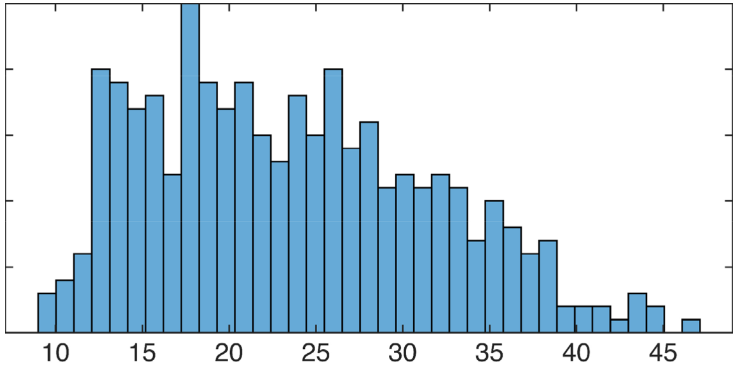
- Converting the source data into another format that allows processing data effectively
- The use of data mining normalization has a number of advantages:
  - The application of data mining algorithms becomes easier
  - The data mining algorithms get more effective and efficient
  - The data is converted into the format that everyone can get their heads around
  - The data can be extracted from databases faster
  - It is possible to analyze the data in a specific manner

Data Mapping

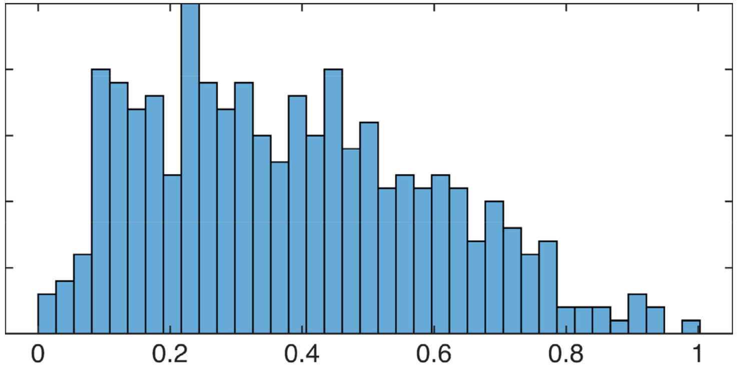
Normalization

- **Min-Max Normalization:** Scale a variable to have a values between 0 and 1

$$x_{min-max} = \frac{x - x_{min}}{x_{max} - x_{min}}$$



Data distribution before normalized



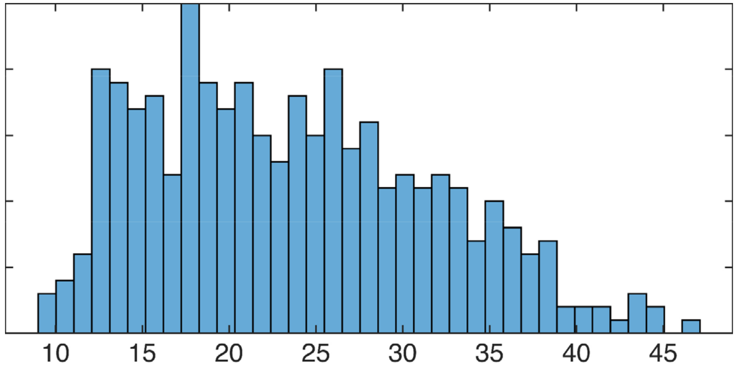
Data distribution after normalized

Data Mapping

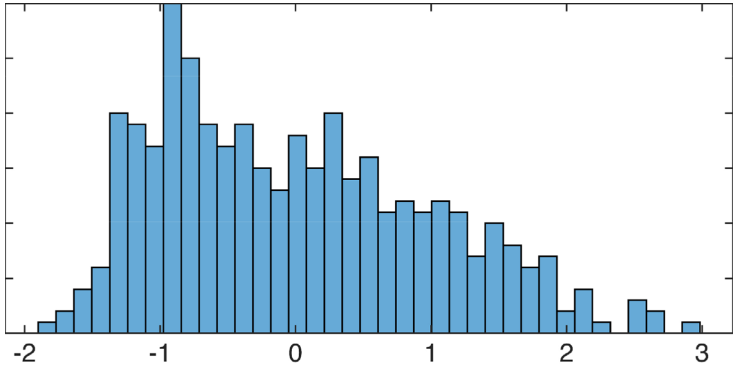
Normalization

- **Z-score Normalization:** Transforms data to have a mean of zero and a standard deviation of 1.

$$x_{z-socre} = \frac{x - \bar{x}}{S.D.}$$



Data distribution before standardized



Data distribution after standardized

Data Mapping

Anonymization and Encryption

- Data containing personally identifiable information, or other information that could compromise privacy or security, should be anonymized before propagation.
- Remove personally identifiable information
- Encryption personal identifier
- Encryption of private data is a requirement in many industries, and systems can perform encryption at multiple levels, from individual database cells to entire records or fields.

Data Mapping

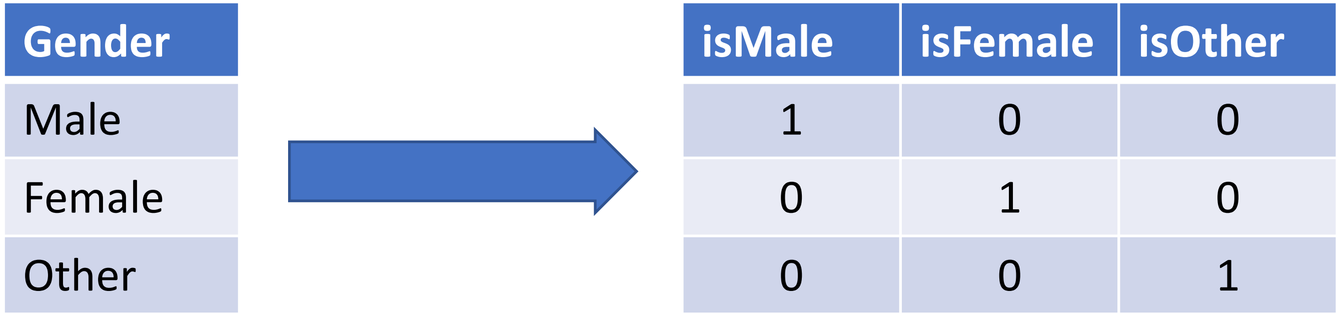
Encoding Categorical Data

- Machine learning algorithms and deep learning neural networks require that input and output variables are numbers.
- Categorical data must be encoded to numbers before we can use it to fit and evaluate a model.

Data Mapping

Encoding Categorical Data

- **One Hot Encoding:** Map each category to a vector that contains 1 and 0
  - 1 - presence of the feature
  - 0 - absence of the feature



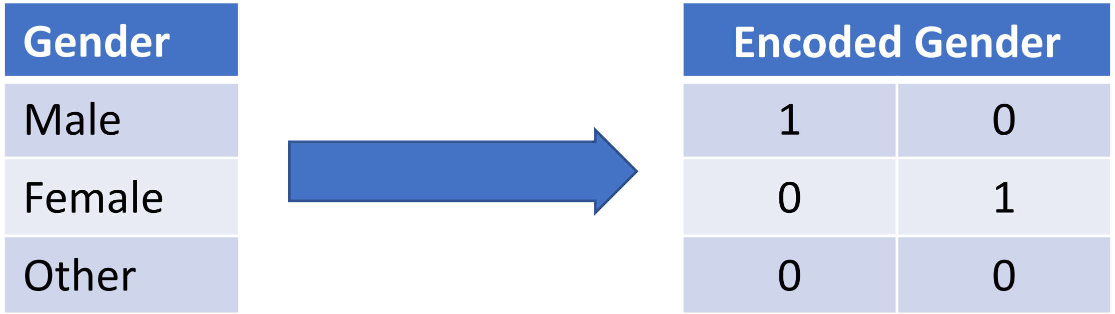
Gender
Male
Female
Other

isMale	isFemale	isOther
1	0	0
0	1	0
0	0	1

Data Mapping

Encoding Categorical Data

- **Dummy Encoding**
  - Similar to one-hot encoding.
  - The dummy encoding is a small improvement over one-hot-encoding.
  - Dummy encoding uses N-1 features to represent N labels.



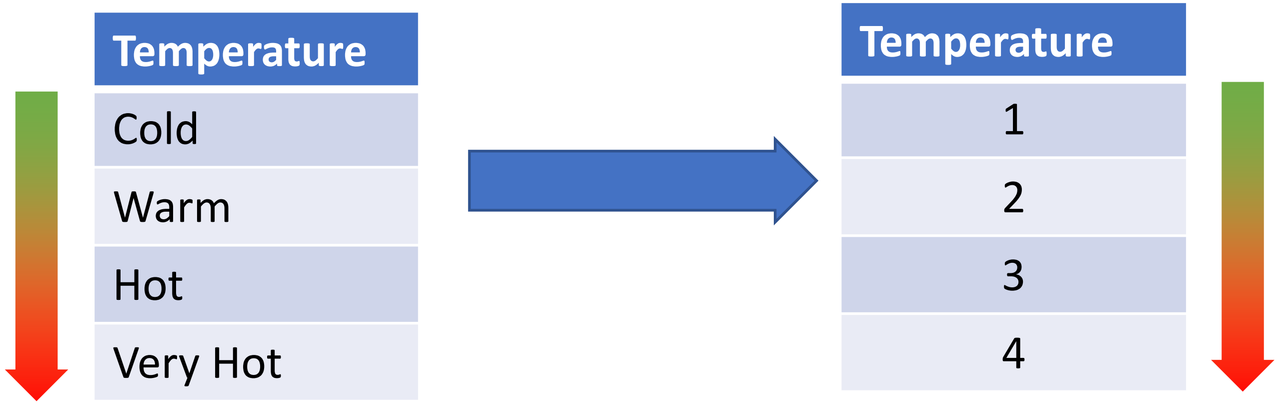
Gender
Male
Female
Other

Encoded Gender	
1	0
0	1
0	0

Data Mapping

Encoding Categorical Data

- **Ordinal Encoding**
  - The encoding of variables retains the ordinal nature of the variable
  - Each category is assigned a value from 1 through the number of possible values by considering the order of values.



Temperature
Cold
Warm
Hot
Very Hot

Temperature
1
2
3
4

References

- Paul Crickard (2020). *Data Engineering with Python*. Birmingham, UK: Packt Publishing Ltd.
- https://www.stitchdata.com/resources/data-transformation/
- https://analyticsindiamag.com/top-8-data-transformation-methods/
- https://www.xplenty.com/blog/data-transformation-explained/