Basic Machine Learning: Data Preprocessing

Oleh: Muhammad Angga Muttaqien | Founder & Mentor

Goal

Goal

To understand how important data preprocessing is in term of building machine learning model

- Understand & Prepare the datasets for machine learning modeling process
- Important task technique that transforms raw data into a more understandable, useful and efficient format

Outline

Outline

- Data Handling

- Data Transformation
  - ❑ Categorical data transformation
  - ❑ Numerical data transformation

- Feature Analysis
  - ❑ Dimensional reduction & TDA
  - ❑ Feature analysis
  - ❑ Feature engineering

Content

Data Handling

Our real world data is generally:

Missing or incomplete: Certain attributes or values or both are missing or only aggregate data is available

Noisy: Data consists of errors, outliers, inconsistency

Categorical or Continuous: Data represents categorical or continuous values that are not standardized

Data Handling

How to deal with such data?

- Replace with default value
- Replace with mean/median/mode values
- Drop data

Which one to use?

Data Transformation

To transform data into form that can be learned easier by computer

To simplify data by decreasing its value scale

All based on whether the data type is categorical or numerical

Data Transformation

To transform data into form that can be learned easier by computer

To simplify data by decreasing its value scale

To standardize data type (categorical or numerical)

Data Transformation (Categorical)

The idea is to make all features that contain categorical data for having numerical information. Generally there are 2 method that can be used:

1. Label Encoder
2. One Hot Encoder

	Economy Level	Gender	Occupation
0	Medium	Male	Programmer
1	High	Female	Auditor
2	Medium	Female	Manager
3	Low	Male	Teacher
4	Medium	Male	Marketing

Machine cannot understand such string value in 'Economy Level', 'Gender', 'Occupation' so somehow we need to change them into **numerical value**

Data Transformation (Categorical)

Label Encoder

Data Transformation (Categorical)

One Hot Encoder
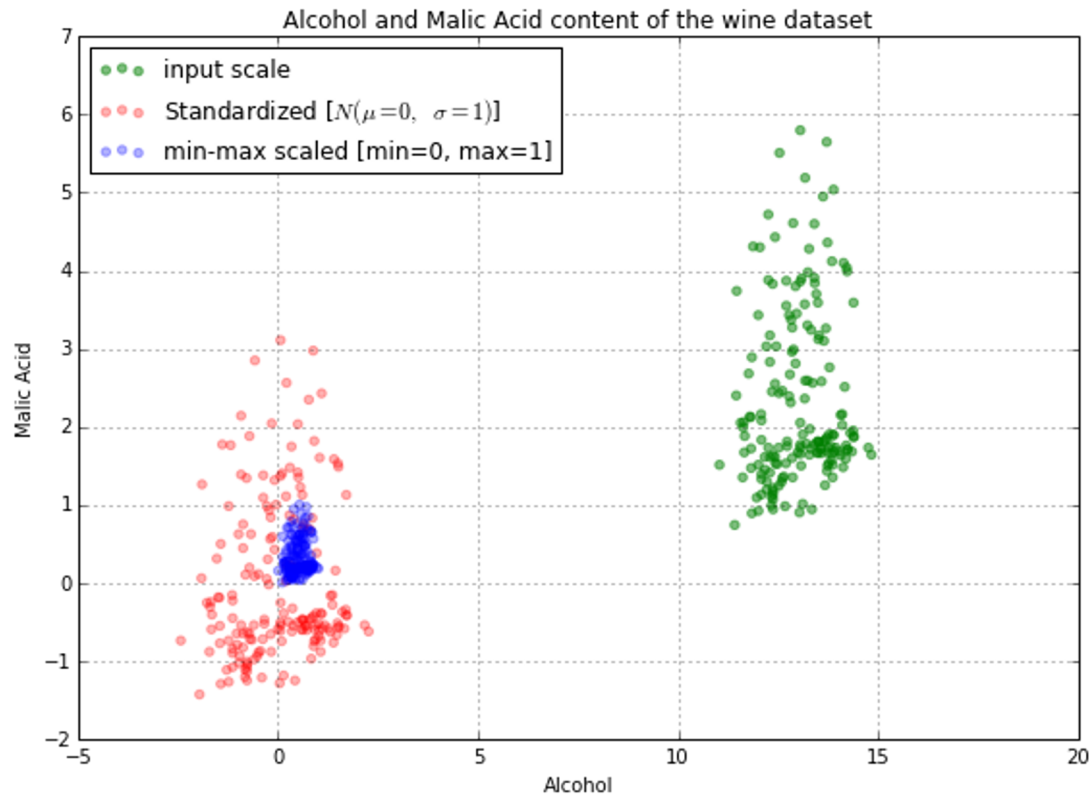
Data Transformation (Numerical)

The idea is to make all feature that contain numerical data for having same scale. Generally there are 2 method that can be used:

1. Standard Scaler
2. Min-Max Scaler

	Age	Gained Calory per Day	Gross Income
0	25	500	5000000
1	27	300	7000000
2	20	700	2500000
3	30	500	5000000
4	22	1000	1000000

- 'Age' feature has tens scale
- 'Gained Calory per Day' feature has hundreds scale
- 'Gross Income' feature has millions scale

Data Transformation (Numerical)

Alcohol and Malic Acid content of the wine dataset

Legend:
- input scale (green)
- Standardized [$N(\mu=0, \sigma=1)$] (red)
- min-max scaled [min=0, max=1] (blue)

❏ Standard Scaler transform data distribution into **normal distribution**

❏ Min Max Scaler transform data range **into 0 ~ 1**

❏ Rule of thumb:

- Use Min Max Scaler as the default if you are transforming a feature

- Use Standard Scaler if you need a relatively normal distribution

Feature Analysis

Feature Selection	Feature Engineering	Feature Extraction
- Based on domain knowledge - Exclude unnecessary features - Heatmap and many other visualization techniques support	- Based on domain knowledge - Generating new feature from other related datasets	- Keep all information from all feature before extracting the information - Popular techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Auto Encoder (Neural Network Approach)

Thanks!