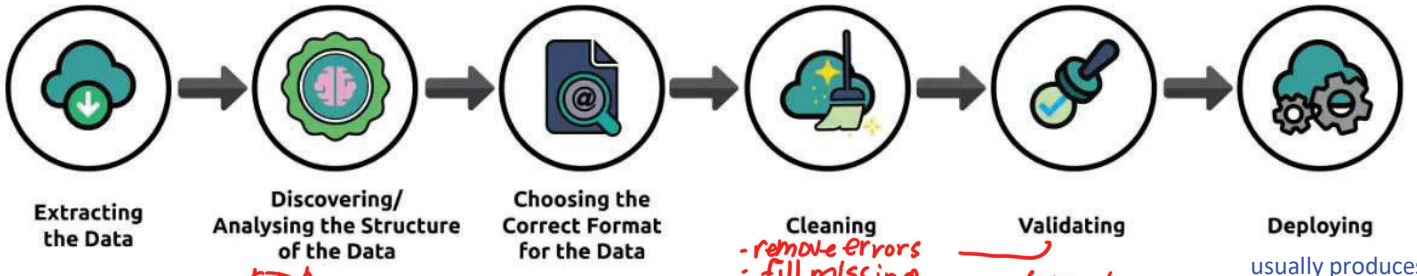
Lecture 2: Data Preparation & Feature Engineering

Assoc Prof Tham Chen Khong
Dept of Electrical & Computer Engineering (ECE)
NUS
E-mail: eletck@nus.edu.sg

Data Wrangling

- Data wrangling is the act of mapping raw data into another format more suitable for another purpose.



From sensors, database, data sources, screen scraping etc.

EDA
Exploratory Data Analysis

also: SQL/no-SQL/graph database

Cleaning
- remove errors
- fill missing data

Validating
correct, makes sense

Deploying
usually produces CSV or JSON file
→ build ML models

Example of a data wrangling tool in Python is Pandas

Feature Engineering

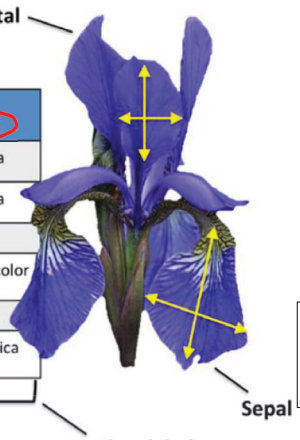
→ relevant inputs to train a ML model

- Feature engineering involves processing the raw data, selecting relevant features and possibly creating new features from raw data to be used for ML
  - analyze the raw data in order to extract a new or more valuable set of features

Example: Iris dataset
https://archive.ics.uci.edu/ml/datasets/iris

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

Refer to scikit-learn
Feature Extraction (section 6.2) and
Preprocessing Data (section 6.3)

1. Data Cleaning - Imputation

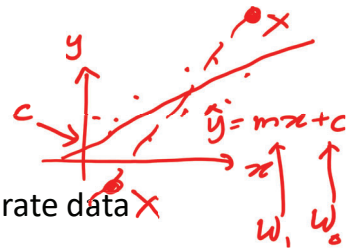
- When it comes to preparing data for machine learning, missing values are one of the most typical issues.
- Human errors, data flow interruptions, privacy concerns, and other factors could all contribute to missing values.
- Missing values have an impact on the performance of machine learning models.

- The main goal of imputation is to handle these missing values. There are two types of imputation:

1. **Numerical Imputation:**
   - Strategies range from simple (e.g., replacing missing values with the mean of the column) to sophisticated (e.g., using a model to handle such data).
   - For a baseline imputation approach using the mean, median, or most frequent value, the `SimpleImputer` class in scikit-learn can be used (section 6.4).
   - Alternatively, to fill all missing values (i.e. `NaN`) in a Pandas dataframe with certain value, use the Pandas `fillna()` function.
2. **Categorical Imputation:**
   - When dealing with categorical columns, one can replace missing values with the most frequent value in the column.
   - If the values in the column are evenly distributed and there is no dominating value, imputing a category like "`Other`" may be a better choice.

extreme values (big/small)

2. Outlier Removal

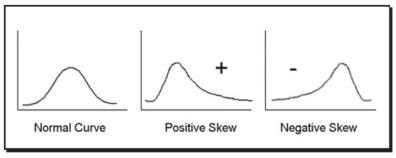


- Outlier handling is a technique for removing outliers from a dataset.
- This method can be used on a variety of scales to produce a more accurate data representation. This has an impact on the model's performance.
- Depending on the model, the effect could be large or minimal; for example, linear regression is particularly susceptible to outliers. This procedure should be completed prior to model training.
- Various methods of handling outliers include:
- **Removal**: Outlier-containing entries are deleted from the distribution. However, if there are outliers across numerous variables, this strategy may result in a big chunk of the data being missed.
- **Replacing values**: Alternatively, the outliers could be handled as missing values and replaced with suitable imputation.
- **Capping**: Using an arbitrary value or a value from a variable distribution to replace the maximum and minimum values.
- **Discretization**: Discretization is the process of converting continuous variables, models, and functions into discrete ones. This is accomplished by constructing a series of continuous intervals (or bins) that span the range of the desired variable/model/function.

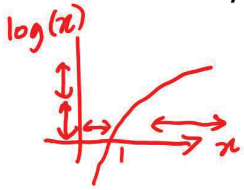
min e.g. 17 max e.g. 22

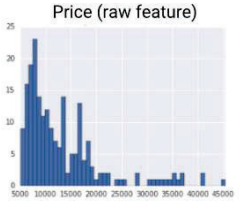
−10 → 1: 0–5; 2: 6–10; 3: 11–15; 20

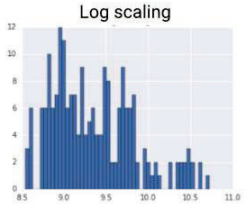
3. Transform



- Log transform is a commonly used technique to turn a skewed distribution into a normal or less-skewed distribution.
- Take the log of the values in a column and utilise those values as the column. The effect is that the smaller values range is expanded and the larger values range is reduced.
- The data may become closer to being normally distributed.







- Pandas example:
  - `df[log_price] = np.log(df['Price'])`
- Others: square root transform etc.

4. Categorical Data: One Hot Encoding

- A one hot encoding is a representation of categorical variables as binary vectors.
- This first requires that the categorical values be mapped to integer values.
- Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1, e.g.

$$\text{red} = [1, 0, 0]$$
$$\text{yellow} = [0, 1, 0]$$
$$\text{green} = [0, 0, 1]$$

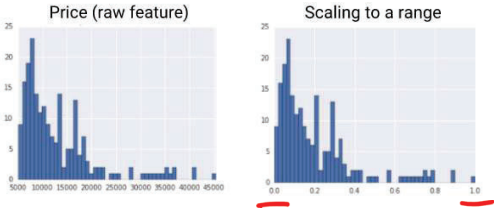
- *note*: there are other ways to encode categorical data, e.g. label encoding, {+1,-1,…} etc.

do this only for ordinal categories, otherwise do target encoding

5. Scaling and Normalization - 1

- Feature scaling is one of the most pervasive and difficult problems in machine learning, yet it is one of the most important things to get right. In order to train a predictive model, we need data with a known set of features that needs to be scaled up or down as appropriate.
- After a scaling operation, the continuous features become similar in terms of range. Although this step is not required for many algorithms, it is still a good idea to do so. Note that distance-based algorithms like k-nearest neighbours and K-means, on the other hand, require scaled continuous features as model input.
- There are two common ways of doing scaling :
- 1. **Normalization**: All values are scaled in a specified range between 0 and 1 via normalisation (or **min-max normalisation**). This modification has no influence on the feature's distribution, however it does exacerbate the effects of outliers due to lower standard deviations. As a result, it is advised that outliers be dealt with prior to normalisation.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad x_i = \frac{x_i^{raw} - x^{min}}{x^{max} - x^{min}}, \quad i = 1, 2, ..., M$$

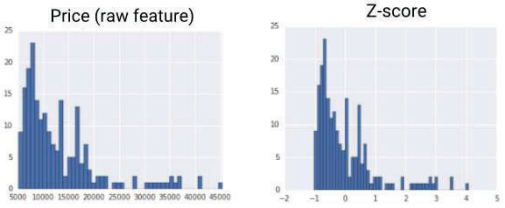


- scikit-learn `sklearn.preprocessing.MinMaxScaler`

5. Scaling and Normalization - 2

- 2. **Standardization**: Standardization (also known as **z-score normalisation**) is the process of scaling values while accounting for standard deviation. If the standard deviation of features differs, the range of those features will likewise differ. The effect of outliers in the characteristics is reduced as a result. To arrive at a distribution with a 0 mean and variance 1, all the data points are subtracted by their mean and the result divided by the distribution's standard deviation.

$$z = \frac{x - \mu}{\sigma} \qquad x_i = \frac{x_i^{raw} - E[X]}{\sigma(X)}, \quad i = 1, 2, ..., M$$

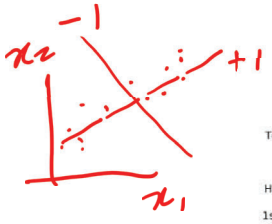


- scikit-learn `sklearn.preprocessing.StandardScaler`

Other Aspects of Feature Engineering

- Feature Extraction (advanced topic)
  - transforming arbitrary data, such as text or images, into numerical features usable for machine learning (e.g. irises dataset)
  - refer to scikit-learn section 6.2
- Feature Selection
  - remove features that are not useful
    - e.g. examine **correlation of features to target value *y*** (use training data) label
      - use `seaborn.heatmap()` to visualize and pick features
    - refer to scikit-learn section 1.13
- Dimensionality Reduction
  - Principal Component Analysis (PCA) (refer to `sklearn.decomposition.PCA`)
    - linearly transform data into a new coordinate system where the much of the variation in the data can be visible with fewer dimensions (covered in EE4300)
- Feature Learning (advanced topic)
  - e.g. word embedding used in transformers, LLM

seaborn.heatmap() of correlation



Target value

Stronger correlation on both ends of the spectrum pops out in darker, weaker correlation in lighter shades.

Thank You

Questions?

Assoc Prof Tham Chen Khong

E-mail: eletck@nus.edu.sg