Unit-2

Data Pre-processing

Data Preprocessing

- Data preprocessing is the process of **transforming raw data into an understandable** format.
- The **quality of the data** should be checked before applying machine learning or data mining algorithms.


Data Integration
Data Cleaning
Data Preprocessing
Data Transformation
Data Reduction or Dimension Reduction
V7 Labs

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	0
1	Spain	27.0	48000.0	1
2	Germany	30.0	54000.0	0
3	Spain	38.0	61000.0	0
4	Germany	40.0	NaN	1
5	France	35.0	58000.0	1
6	Spain	NaN	52000.0	0
7	France	48.0	79000.0	1
8	Germany	50.0	83000.0	0
9	France	37.0	67000.0	1

Data cleaning

- Missing Values
  - Ignore/Fill Missing values
- Noisy Data
  - Binning (Mean/ Median/ Boundary Values)
  - Regression
  - Clustering

Data integration

- The process of combining multiple sources into a single dataset. The Data integration process is one of the main components in data management.

Data Integration

- **Schema integration**: Integrates metadata(a set of data that describes other data) from different sources.
- **Entity identification problem:** Identifying entities from multiple databases.
- **Detecting and resolving data value concepts**: The data taken from different databases while merging may differ from another database. For example, the date format may differ like “MM/DD/YYYY” or “DD/MM/YYYY”.

Data Transformation

- The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements.

Data Transformation

- **Normalization:**
  It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
- **Attribute Selection:**
  In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

Data Transformation

- **Discretization**: The continuous data here is split into intervals. interval like (3 pm-5 pm, 6 pm-8 pm).
- **Concept Hierarchy Generation:**
  Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Data Reduction

- This process helps in the reduction of the volume of the data which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space.

Data Reduction

- **Attribute Subset Selection:**
  The highly relevant attributes should be used, rest all can be discarded.
- **Numerosity Reduction:**
  This enable to store the model of data instead of whole data, for example: Regression Models.
- **Dimensionality Reduction:**
  This reduce the size of data by encoding mechanisms.It can be lossy or lossless.

Data Preprocessing: Best practices

- The first step in Data Preprocessing is to understand your data. Just looking at your dataset can give you an intuition of what things you need to focus on.
- Use statistical methods or pre-built libraries that help you visualize the dataset and give a clear image of how your data looks in terms of class distribution.
- Summarize your data in terms of the number of duplicates, missing values, and outliers present in the data.

Data Preprocessing: Best practices

- Drop the fields you think have no use for the modeling or are closely related to other attributes. Dimensionality reduction is one of the very important aspects of Data Preprocessing.
- Do some feature engineering and figure out which attributes contribute most towards model training.

Splitting dataset into Training and Testing set.

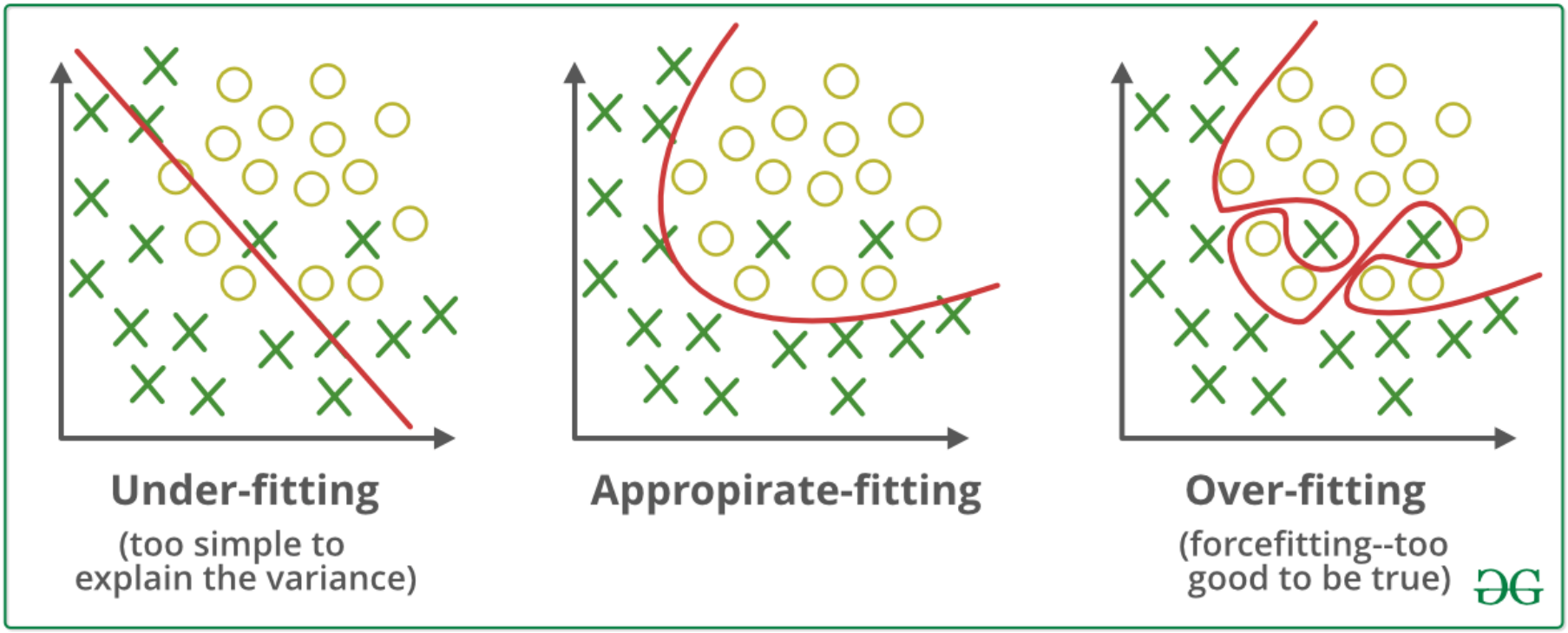
- Training Data
  - The observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables.
- Test Data
  - The test set is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set.

Basics

- **Bias:** Assumptions made by a model to make a function easier to learn.
- **Variance:** If you train your data on training data and obtain a very low error, upon changing the data and then training the same previous model you experience a high error, this is variance.

Under & Over Fitting

- **Underfitting:**
  - A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.
- **Overfitting:**
  - A statistical model is said to be overfitted when we train it with a lot of data *(just like fitting ourselves in oversized pants!)*.


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

References

- https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/
- https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/
- https://www.v7labs.com/blog/data-preprocessing-guide