Data Science

Unit - 1

07/09/2023

Introduction to Python :-

Guido Van Rossum (developer of python)

Latest version - Python 3.10.6
(3 → version no., 10 → updates, 6 → updates)

features of python -

- Simple and Straight forward Syntax
- Case Sensitive
- Multi-paradigm programming language [procedural-oriented, functional-oriented, Object-oriented]
- Dynamically Typed.
- Emphasis on code readability
- Automatic memory management.
- ★ Large library support.
- Huge community
- Platform independent.

Web frameworks - Django, Flask etc.

Libraries like numpy, Scipy, matplotlib are used in Scientific computations.

// First program -

```
print("Enter two numbers")
a = int(input())
b = int(input())
c = a+b
print("Sum is", c)
```

Comments -

- Single line comment - # here is python class.
- multi line comment - """ ... """ threetimes double quotes.

Dynamic Type- Not only the value of a variable may change during program execution but the type as well.

eg-
```
x = 5       # type of x is int
x = 5.7     # type of x is float
x = True    # type of x is bool.
x = "Anushka"  # type of x is str.
x = 3+4j    # type of x is complex.
```
for finding the datatype of the variable.
⇒ type(x)

data type is always a class in Python.

Numbers	Boolean	String
int	bool	str
float		
complex		

double is not there in Python.
char is also not there in Python.

1. Instance object
2. Function object
3. Class object

★ **Import-**

module~ .py file is a module
module contains python code with three kinds of reusable elements.

① variables.
② functions.
③ classes.

eg- import A1
print(A1.x)

★ kwlist contains all keywords list. [35 keywords]

eg- import keyword
print("There are total", len(keyword.kwlist), "keywords in Python")
print(keyword.kwlist)

Operators- There is no ++ (increment) or --(decrement) operator in python.

→ Arithmetic, Relational, Logical, Bitwise, Assignment
→ Identity operator ~ is, is not
→ membership operator ~ in, not in

Arithmetic~ ** is for calculating power.

→ **floor division-** [//] double slash

eg- 3//4 = 0 both int result int
eg- 3/4 = 0.75 one of them is float result float.

→ 6>5>4
In python, it check for each operator ie, 6>5 or 5>4
both true then return true.

⇒ Non-empty string ~ true | Non-zero number - true
empty string - false "" | zero - false

through id we print "address".

eg-
```
x = 5
y = 5
id(x)
436522256
```

Unit-2

Model Designing

Data Preprocessing – Data preprocessing refers to the cleaning, transforming and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific task.

Data Quality ⋆ preprocessing is the process of converting raw data into a format that is understandable and usable.

It ensures that data quality is consistent before applying any Machine learning.

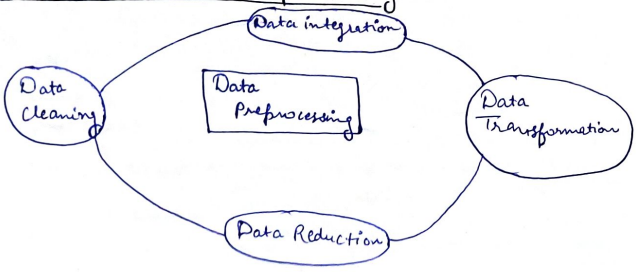
Need of Data Preprocessing / Data Preprocessing Importance :–

The main objective of that Data preprocessing is to ensure and check the quality of data.

It checks the following things for the input data :–

1) Accuracy
2) Completeness
3) Consistent
4) Timeliness
5) Trustable
6) Interpretability

Major tasks in Data Preprocessing –



- **Data Quality** – Data quality is how we describe the state of any given dataset. It measures objective elements such as completeness, accuracy and consistency.

Accuracy – check for correctness, accurate or not.

Completeness – not recorded, unavailable.

Consistency – some modified but some not, dangling.

Timeliness – timely update

Believability – how trustable the data is correct.

Interpretability – how easily the data can be understood.

⋆ Data Cleaning :–

- Fill in missing values, smooth noisy data, identify or remove outliers and resolve inconsistencies and duplicates.

Data cleaning uses methods to handle incorrect, incomplete, inconsistent or missing values, identifying or correcting errors.

Various techniques can be used for data cleaning such as Removal and transformation.

⋆ Data Integration :–

Data integration can be defined as combining data from multiple sources and integration of multiple databases, files and then create a unified dataset.

Data Integration, it requires handling data with different formats, structures and semantics.

☆ Data Reduction –

Data Reduction is used to reduce the volume or size of the input data. Its main effective is to reduce storage and analysis costs and improve storage efficiency. Data compression is used in Data Reduction. It can be achieved through feature selection and feature extraction.

☆ Data Transformation – It is a process of converting data into a format that helps in building efficient ML models and also in analysis.

Techniques used in data transformation includes normalization, standardization & discretization.

☆ Data Discretization – Data discretization is a process of converting numerical or continuous variables into a set of intervals / bins. This makes data easier to analyze. It requires categorical data.

It can be achieved through techniques such as equal width, binning, clustering.

☆ Data Normalization – Normalization is used to scale the data to a common range, such as b/w 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common approaches to performing normalization are Min-Max normalization, Data Standardization or Data Scaling etc.

☆ Data Standardization – It is used to transform the data to have zero mean and unit variance.

Data cleaning – To handle irrelevant & missing parts.

It contains total 6 steps –

- Handling Missing Values
- Removes duplicate
- Fix errors
- Removing irrelevant data
- Handling outliers
- Convert data type.

a) Fill missing values.
b) Smooth out noise and outliers.
c) Remove inconsistencies

1. Handle missing Data :- During cleaning, handling missing values is one of most common task. It may contain missing values which need a fix before analysis of data. We can handle By :-

- Either removing the records that have missing values or
- Filling the missing values using some statistical technique or by gathering data understanding.

a) Ignore the tuples – This approach is suitable only when the data set we have is quite larger and multiple values are missing within a tuple.

b) Fill the missing values – To fill missing values, we can done it manually by attribute mean or the most probable value.

2. Noisy Data :-

* Noise is a random error or variance in a measured variable.

Noisy data is a meaningless data that cannot be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc.

It can be due to "Incorrect attribute values.

- faulty data collection
- data entry problems
- data transmission problems
- technology limitations
- inconsistency in naming convention.

OR

- duplicate records, incomplete or inconsistent data.

Outliers - Cleaning data from outliers is a process of identifying & handling extreme values that deviate significantly from the majority of the data.

How to handle Noisy data - 'Smoothning'

1. **Binning Method** - This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size (or equal-frequency bins) and then various methods are performed to complete the task. Each segments is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
   - Smoothing by BIN means, Bin median, Bin boundaries.
2. **Regression** - Smooth by fitting the data into regression functions. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables.
3. **Clustering** - (detect and remove outliers)
   This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters, and then remove the outliers.

→ Binning is the process of changing numerical variables into categorical counterparts. The no. of categorical counterparts depends on the no. of bins specified by the user.

- Data cleaning as a process -

- **Data discrepancy detection** -

1) Data discrepancy detection use meta data (eg. domain, range, dependency, distribution)
2) Check field overloading.
3) Check uniqueness rule, consecutive rule and null rule
4) Use commercial tools
   - Data scrubbing - Scrubbing is also known as data cleaning. The data cleaning process detects and removes errors and anomalies and improves data quality. Use simple domain knowledge (e.g, postal code, spell-check) to detect errors and make corrections.
   - Data auditing - A data-audit is a step-by-step process that examines every step of the data science process. By analyzing data to discover rules and relationship to detect violators (eg. correlation and clustering to find outliers)

- **Data migration & Integration** -

Data migration is the process of moving data from one location to another, one format to another, or one application to another.

Data Integration refers to the process of combining data from multiple sources into a single unified view.

- Data migration tools allow transformations to be specified.
- ETL (Extraction / Transformation / loading) tools: allow users to specify transformations through a graphical user interface.

Data Integration

Combines data from multiple sources into a coherent store

- Schema Integration :- eg. A.cust-id ≡ B.cust-#
  Integrate meta data from different sources.
- Detecting and resolving data value conflicts ⇒ unwanted Redundancy attributes eg. age

✱ Handling Redundancy in Data Integration –

Redundant data occur often when integration of multiple databases –

- Object Identification – The same attribute or object may have different names in different databases.
- Derivable data – One attribute may be "derived" attribute in another table.

Redundant attributes may be able to be detected by correlation analysis and covariance analysis.

Redundancy Detection –

- An attribute may be redundant if it can be derived or obtained from another attribute or set of attributes.
- Inconsistencies in attributes can also cause redundancies in the resulting data set.
- Some ~~redundancy~~ can be detected by correlation analysis.

Note – It should be kept in mind that Data processing techniques are not necessarily applied in strict order.
There may be no use of some techniques in some cases.

① Smoothning by Bin means

eg- Istly sort the data then make bins

bin 1 - 4, 8, 15	9, 9, 9
bin 2 - 21, 21, 24	22, 22, 22
bin 3 - 25, 28, 34	29, 29, 29

② Smoothing by Bin Boundaries

Bin 1 - 4, 4, 15 — mean by which boundary of then replace

Bin 2 - 21, 21, 24

Bin 3 - 25, 25, 34

Correlation Analysis (Nominal Data)

* χ^2 (chi-square) test :- It checks for NULL hypothesis (→ both attributes are independent of each other)

$$\chi^2 = \sum \frac{(\text{Observed} - \text{expected})^2}{\text{expected}}$$

- The larger the χ^2 value, the more likely the variables are related.
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count.
- Correlation does not imply causality.

Correlation Analysis (Numeric Data)

Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{AB} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples.
$\bar{A}$ and $\bar{B}$ are respective means of A and B,
σ_A and σ_B are the respective standard deviation of A & B
$\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's), The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated.

Correlation (viewed as linear Relationship)

- Correlation measures the linear relationship between objects.

For compute correlation, we standardize data objects A and B and then take their dot product.

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A' \cdot B'$$

Covariance - Covariance is similar to correlation.

$$\text{Cov}(A, B) = E((A - \bar{A})(B - \bar{B})) = \sum_{i=1}^{n} \frac{(a_i - \bar{A})(b_i - \bar{B})}{n}$$

$$r_{A,B} = \frac{\text{Cov}(A, B)}{\sigma_A \sigma_B}$$

where n is the number of tuples,
$\bar{A}$ and $\bar{B}$ are the respective mean or expected values of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- Positive covariance - $\text{Cov}_{A,B} > 0$ larger than their expected values.
- Negative covariance - $\text{Cov}_{A,B} < 0$ smaller than the expected values.
- Independence - $\text{Cov}_{A,B} = 0$ but converse is not true.

Data Reduction

Data Reduction Strategies -

Why data reduction? - A database / data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Strategies -

- **Dimensionality reduction** :- It removes unimportant attributes. This technique involves reducing the number of features in the dataset, either by removing features that are not relevant or by combining multiple features into a single feature.

a) **Wavelet transforms** - Wavelet transform is a signal processing technique that transforms 'linear signals'. It is a key tool for signal analysis.

b) **Principal Components Analysis** (PCA) :- PCA is a statistical technique that reduces the dimensionality of large datasets with many variables. It does this by transforming the original variables into a smaller set of new variables that are linearly uncorrelated and capture most of the information in the original data. PCA can help with analyzing, visualizing and summarizing data, as well as identifying patterns and correlations.

→ It converts a set of correlated variables (larger) to a set of smaller uncorrelated variables (called principal components) if reduce the overfitting problem.

→ PCA can be used for variety of purposes, including data visualization, feature selection and data compression.

Feature Selection - PCA can be used to identify the most important variables in a dataset (in feature selection).

Data Compression - In Data compression, PCA can be used to reduced the size of a dataset without losing important information.

c) Feature subset selection - Feature selection is the most critical pre-processing activity in any machine learning process. It intends to select a subset of attributes or features that makes the most meaningful contribution to a ML activity. To remove irrelevant or redundant features from the dataset.

d) Feature Creation - Feature Creation is the process of generating new features based on domain knowledge or by observing patterns in data.

- Numerosity Reduction :- In this reduction technique, the actual data is replaced with mathematical models or smaller representations of the data instead of actual data, it is important to only store model parameter.

  Or non-parametric methods such as clustering, histogram and sampling.

  a) Regression & log-linear models
  b) Histograms, clustering, sampling
  c) Data cube aggregation

Histograms, Clustering, Sampling - Histogram is used to partition the value for the attribute X into disjoint ranges called Brackets.

Clustering means grouping similar data together.

Sampling - It is a preprocessing step in which a subset of the dataset is selected and subjected to various data science methods.

Data Compression - It is a process of reducing the amount of data needed for the storage or transmission of a given piece of information.

Data Reduction

1. Dimensionality Reduction - Remove redundant attributes.

a) Curse of dimensionality -
- When dimensionality increases, data becomes increasingly sparse.
- Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
- The possible combinations of subspaces will grow exponentially.

b) Dimensionality Reduction -
- Avoid the curse of dimensionality.
- Help eliminate irrelevant features and reduce noise.
- Reduce time and space required in data mining.
- Allow easier visualization.

c) Dimensionality reduction Techniques -
- Wavelet transforms
- Principal Component Analysis
- Supervised and non-linear techniques (feature selection)

Wavelet Transform

Wavelet Transform decomposes a signal into different frequency sub-bands.

- Applicable to n-dimensional signals.
- Data are transformed to preserve relative distance between objects at different levels of resolutions.
- Used for image compression and digital signal processing

☆ Discrete Wavelet Transformation - DWT for linear signal processing and multi-resolution analysis.

- It is similar to discrete Fourier transform (DFT), but better than lossy compression, localized in space.

Wavelet Decomposition - A math tool for space-efficient hierarchical decomposition of functions.

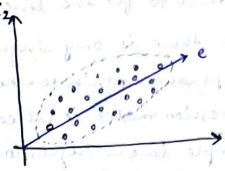
- Many small detail coefficients can be replaced by 0's and only the significant coefficients are retained.

Why wavelet Transforms :-

- Effective removal of outliers
- multi-resolution.
- Efficient. complexity $O(N)$
- Only applicable to low dimensional data.

Principal Component Analysis (PCA)

- In PCA, we have to find a projection that captures the largest amount of variation in data.
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.



Steps -

- Normalize input data : Each attribute falls within the same range
- Compute principal components.
- Each input data (vector) is a linear combination of the k principal component vectors.
- Sorted in order of decreasing "significance" or strength.
- By eliminating weak components, size of data will be reduced.

2. Numerosity reduction - Reduce data volume by choosing alternative, smaller forms of data representation. It is beneficial in situations where the dataset is too large to be processed efficiently or where the dataset contains a large amount of redundant or irrelevant data points.

- Parametric method -

Regression :- Data can be made smooth by fitting it to a regression function. The regression used may be linear or multiple. A statistical method used to estimate the relationship between a dependent variable & one or more independent variables.

- Non-parametric method -

Histograms, Clustering, Sampling etc.

Linear Regression

- Data modeled to fit a straight line.
- Often uses the least-square method to fit the line.
- When there is only single independent attribute, such regression model is called Simple linear regression

Multiple Regression

- Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- When there are multiple independent attributes, such regression models are called multiple linear regression.

Log-linear model - It can be used to estimate the probability of each data point in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.

Regression Analysis

Regression analysis have 3 types —
Linear, Multiple, Non-linear Regression

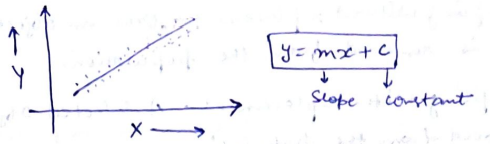
Linear Regression – $Y = wX + b$

Two regression coefficient, w and b specify the line and are to be estimated by using the data.

→ Simple linear regression is a model that assesses the relationship between a dependent variable & an independent variable. The simple linear regression is also expressed as–

$Y = a + bX + \epsilon$, dependent variable is continuous in nature.

where, Y = dependent variable
X = independent variable
a = Intercept
b = Slope
ϵ = residual (error)



Histogram Analysis – Histogram is used to partition the value for the attribute X, into disjoint ranges called Brackets. There are some partitioning rules —

- Equal frequency partitioning
- Equal width partitioning

Clustering – Clustering refers grouping similar data together into clusters. Clustering is used to reduce the size of database by replacing similar datapoints.

Sampling

This technique involves selecting a subset of the data to work with, rather than using the entire dataset. This can be useful for reducing the size of a dataset, while preserving the important information. It can be done using techniques such as random sampling, stratified sampling & systematic sampling.

Types of Sampling

1) Simple random sampling – There is an equal probability of selecting any particular item.

2) Sampling without replacement - Once an object is selected, it is removed from the population.

3) Sampling with replacement - A selected object is not removed from the population.

4) Stratified Sampling - In this type of sampling, we divide the population into subgroups based on different traits like gender, category etc. and then we select the sample(s) from these subgroups :

Data Cube Aggregation

This technique is used to aggregate data in a simpler form, also it summarizes the data.

Data cube store multi-dimensional aggregated information. Aggregation operations are applied to the data in the construction of a data cube.

3) **Data Compression :-** This involves compressing the dataset while preserving the important information. It is used to reduce the size of dataset for storage & transmission process.

Compression is of 2 types -

* lossless compression
* Lossy compression.

Data Transformation & Data Discretization

★ **Data Transformation** - It refers to process of converting raw data into a format that is suitable for analysis & modeling and that is free of errors.

Methods -

- Smoothing - remove noise from data
- attribute/feature construction - new attributes constructed from the given ones.
- Aggregation - summarization, data cube construction
- Normalization :
  * min-max normalization
  * z-score normalization
  * normalization by decimal scaling.
- Discretization - concept hierarchy climbing.

Normalization

It refers to scaling the data to a common range of values, such as between 0 and 1, to facilitate comparison & analysis. Techniques that are used for normalization are :

1. **Min-Max Normalization :-** to $[new\_min_A, new\_max_A]$

- This transforms the original data linearly.
- Let min_A is minima and max_A is maxima of an attribute
- v is the value you want to plot in new range. (old value)
- v' is the new value you get after normalizing the old value

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

2. Z-score normalization :- (μ = mean, σ = Standard deviation)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

v = old value
v' = new value.

- In this normalization, (zero-mean normalization) the values of an attribute (A), are normalised based on the mean of A and its standard deviation.
- A value, v, of attribute A is normalized to v' by computing

3. Decimal Scaling :- $v' = \frac{v}{10^j}$, where j is the smallest integer such that $\max(|v'|) < 1$

- It involves data transformation by dragging the decimal points of values.
- A value, v, of attribute A is normalized to v' by computing
- For normalizing the values we divide the numbers by 100 (i.e. $j = 2$) so that values come out to be as; 0.98, 0.97 and so on.

Discretization

attributes

Converting continuous data into discrete categories or bins. With minimum data loss.

Three types of attributes -
- Nominal - value from unordered set
- Ordinal - value from ordered set
- Numeric - real numbers, integers.

→ Reduce data size by discretization.
→ Supervised & Unsupervised.
→ Discretization can be performed recursively on an attribute.

Data discretization Methods -

All the methods can be applied recursively.

1. Binning - (Top down split, unsupervised)
It refers a data smoothing technique that helps to group a huge no. of continuous values into smaller value

2. Histogram Analysis - Histogram refers to a plot used to represent the underlying frequency dist^n of continuous data set. Histogram assigns the data inspection for data distribution.

3. Clustering analysis - A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

4. Decision-tree analysis - applies to classification & prediction
Data discretization refers to a decision-tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure.
In a numeric attribute discretization, first you need to select the attribute that has the least entropy, and then run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using same splitting criteria. Generate rules for simple interpretation & understanding.

5. Correlation (χ^2) analysis - Discretizing data by linear regression technique, you can get the best neighbouring interval & then the large intervals are combined to develop a larger overlap to form final 20 overlapping intervals.

Simple Discretization : Binning

- Equal-Width (distance) partitioning
- Equal-depth (frequency) partitioning

→ Equal-Width binning divides the range into N intervals of equal size width of intervals.

→ Equal-depth binning divides the range into N intervals, each containing approximately same number of records.

Example of Binning methods for data Smoothing

eg- 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.

- Partition into equal-frequency (equi-depth) bins :-

  Bin 1 : 4, 8, 9, 15

  Bin 2 : 21, 21, 24, 25

  Bin 3 : 26, 28, 29, 34

- Smoothing by Bin means :

  Bin 1 : 9, 9, 9, 9

  Bin 2 : 23, 23, 23, 23

  Bin 3 : 29, 29, 29, 29

- Smoothing by bin boundaries :

  Bin 1 : 4, 4, 4, 15

  Bin 2 : 21, 21, 25, 25

  Bin 3 : 26, 26, 26, 34

- Concept Hierarchy Generation -

Concept hierarchies can be used to recursively reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts.

- Concept hierarchy organizes concepts (attribute values) hierarchically and is usually associated with each dimension in a data warehouse.
- Concept hierarchy facilitate drilling and rolling in data warehouses to view data in multiple granularity.
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers.
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, discretization is used.

Automatic Concept Hierarchy Generation

Some hierarchies can be automatically generated based on the analysis of the no. of distinct values per attribute in the data set.

- The attribute with the most distinct values is placed at the lowest level of the hierarchy.
- Exceptions eg- weekday, month, quarter, year.

country
|
province or State
|
city
|
Street

Train/Test :- It is a method to measure the accuracy of the model.

It is called train/test because in this we split the data set into two sets : a training set & a testing set.

Training Set

Train the model means create the model.

A training set is a portion of a data set used to fit (train) a model for prediction or classification of values that are known in the training set, but unknown in other (future) data.

The training set is used in conjuction with validation and/or test sets that are used to evaluate different models.

- The training data varies depending on whether we are using Supervised learning or unsupervised learning Algorithm.
- This is the actual dataset from which a model trains ie. the model sees and learns from this data to predict the outcome or to make the right decisions.

Testing Set

Test the model means, test the accuracy of model.

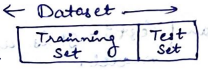
This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset.

The test dataset is another subset of original data which is independent of the training dataset.

- This dataset is independent of training set but has a somewhat similar type of probability distribution of classes & is used as a benchmark to evaluate the model, used only after the training of the model is complete.

Need of splitting dataset into Train & Test Set :-

It is a part of data preprocessing, by this we can improve the performance of our model & give better predictability. For splitting dataset, we can use the train_test_split function of scikit-learn.



If we train & test the model with two different datasets then it will decrease the performance of model.

Hence it is important to split a dataset into two parts train & test set.

Difference b/w Training data vs. Testing data—

- The main difference between training data and testing data is that training data is the subset of original data that is used to train the ML model, whereas testing data is used to check the accuracy of model.
- The training dataset is generally larger in size compared to the testing dataset.
- Training data is well known to the model as it is used to train the model, whereas testing data is like new data to the model.

Normalization

Normalization is scaling the data to be analyzed to a specific range such as [0.0, 1.0] to provide better results. Normalization contribute toward the success of data extract process.

Need- Normalization is required when we are dealing with attributes on a different scale.

—Model Selection—

It is a process of deciding which techniques are best suited to solve the given problem.

★ Training dataset –

- Training data is the biggest (in-size) subset of original dataset, which is used to train or fit the ML model.
- First, the training data is fed to ML algo, which lets them learn how to make predictions for the given task.
- Type of training data we provide to the model is highly responsible for model's accuracy and prediction ability.
- Better the quality of training data, the better will be the performance of the model.
- Training data is approximately more than or equal to 60% of total data for ML model.

★ Test dataset –

- This dataset evaluates the performance of the model and ensures that the model can generalize well with new or unseen data.
- It is another subset of original data which is independent of training dataset.
- Used as a benchmark for model evaluation, once the model training is completed.
- Usually, test data is 20-25% of total original dataset.
- The test data should –
  ① represent a part of original data.
  ② large enough to give meaningful predictions.

Overfitting –

- A model is said to be overfitted when it performs quite well with training dataset but does not generalize well with the new or unseen (test) dataset.
- Occurs when model tries to cover all datapoints and start catching noise present in data.
- accuracy and efficiency of model degrade.
- complex model has high chance of overfitting.
- ways to avoid overfitting are –
  a) cross validation method, early stopping the training or by regularization etc.

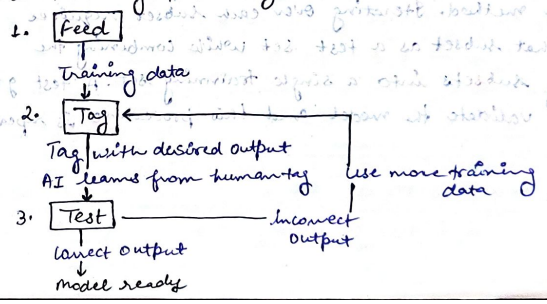
Underfitting –

Model is said to be under-fitted when it is not able to capture the underlying trend of the data.

- occurs when model shows poor performance even with training set.
- generally occurs if model is not perfectly suitable for the problem that we are trying to solve.
- ways to avoid underfitting are –
  increasing training time and increasing no. of features in training set.

Note – Training dataset is larger than test dataset.
General splitting ratios: 80:20, 70:30 or 90:10

How do training & testing data work in Machine Learning



Model Selection

Model Selection is the process of deciding which learning technique to use to model our data.

for example - while attempting to solve a classification issue, we may consider using Logistic Regression, Support Vector Machines, trees and other methods. It is also necessary to make choices on the degree of linear regression techniques while solving a regression problem.

- In cross-Validation - Cross-validation is often used as a benchmark for measuring how well a model generalizes to new data. It plays a role in two crucial steps of data analysis: model selection & evaluation.

Model complexity Selection is the process of deciding what kind of model to use.

- Validation Set - A portion of a data set used in data mining to assess the performance of prediction or classification models that have been fit on a separate portion of the same data set (the training set).
- K-Fold Cross Validation - The dataset is shuffled and then divided into K groups at random to implement the cross-validation method. Iterating over each subset requires treating that subset as a test set while combining the remaining subsets into a single training set. A test group is used to validate the model and this procedure is repeated K-times.

Stratified Cross-Validation :- The splitting of data into folds may be governed by criteria such as ensuring that each fold has the same proportion of observations with a given categorical value, such as the class outcome value. This is called Stratified cross-validation.

Bootstrap method - The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement.

- The bootstrap method involves iteratively resampling a dataset with replacement.
- That when using the bootstrap you must choose the size of the sample and the number of repeats.
- The scikit-learn provides a function that you can use to resample a dataset for the bootstrap method.

Model Evaluation

Model evaluation is the process of figuring out how well the model performs at guessing something.

The evaluation is usually handled with a test dataset. Model evaluation metrices are used to assess goodness of fit between model and data, to compare different models in this context.

Model Evaluation techniques -

1. Accuracy - Accuracy is a simple and common measure for whether or not predictions were correct compared to known outcomes. It is just a fraction of test instances that were labeled correctly. Measure the performance of the model.

Measures of accuracy reveal that the model makes 990 out of 1000 correct predictions; giving it an accuracy of $(990/1000) * 100 = 99\%$. It is the ratio of total correct instances to the total instances. Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

2. Confusion Matrix – A confusion matrix is a way to compare known outcomes and guessed outcomes.

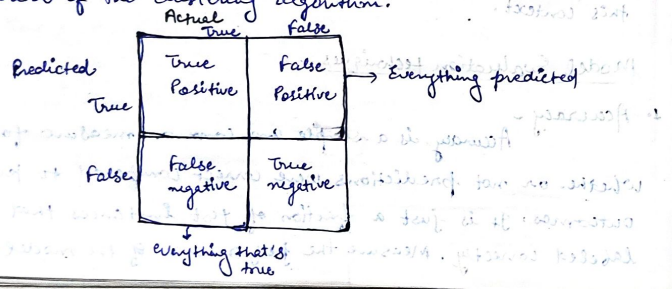
If it's a binary classifier where there are two possible guesses (True/False, Yes/No) etc. then the matrix ends up looking like a 2x2 grid. One axis represents the true outcome and another axis represents the guessed or predicted outcome.

The combinations include:

- True Positive – the things that are true, that were predicted true.
- True Negative – false, predicted false.
- False Positive – actually false, but predicted true.
- False Negative – actually true, predicted false.

→ Each Box of the confusion matrix gets a count, percentage or proportion reflecting the outcomes of the different predictions from the model.

→ The higher the concentration of observations in the diagonal of the confusion matrix, the higher the accuracy / predictive power of the clustering algorithm.



3. Precision – Precision is the ratio of the true positives over the sum of the true positive + false positive.

It's a way of assessing how reliable the true guesses are.

$$\text{Precision} = \frac{TP}{TP+FP}$$

4. Recall – Recall is the ratio of the true positive divided by the sum of the true positive + the false positive or - everything that is true. The percentage of positive instances that were properly detected, or recall.

$$\text{Recall} = \frac{TP}{TP+FN}$$

5. F1-Score – F_1-score is used to evaluate the overall performance of a classification model. It is harmonic mean of precision + recall.

$$F_1\text{-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC- ROC curve –

ROC – Receiver Operating Characteristics

AUC – Area Under Curve.

ROC- curve – ROC curve is the graphical representation of the effictiveness of the binary classification model. It plots the true positive rate (TPR) vs the false positive rate (FPR) at different classification thresholds.

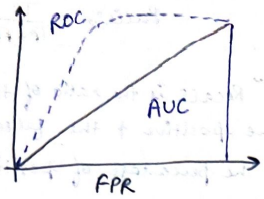
- ROC curve is almost independent of the response rate.

AUC- Curve – AUC curve represents the area under the ROC curve. It measures the overall performance of the binary classification model. As both TPR + FPR ranges b/w 0 to 1. So area will always lie between 0 and 1 & A greater value of AUC denotes

better model performance.

Main goal is to maximize this area in order to have the highest TPR and lowest FPR at the given threshold.

Threshold - It is a particular value beyond which you say a point belongs to a particular class.



TPR and FPR -

TPR - True positive rate

FPR - False positive rate

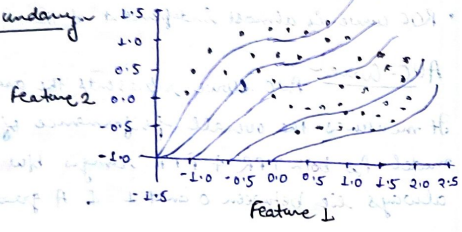
Non-linear decision Boundaries - A Non-linear decision boundary is a curved-line that separates the data into two or more classes. Non-linear decision boundaries are used when the classification problem is not linearly separable.

Non-linear decision boundaries can take different forms such as parabolas, circle, ellipses etc.

Decision Boundary - It refers to a line or curve (or hyperplane) that divides the data into two or more categories based on their features.

→ The objective of decision boundary is to make accurate predictions on unseen data by identifying the correct class for a given input.

Non-linear decision boundary -



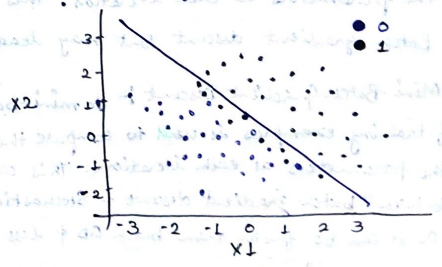
How decision Boundaries are generated -

- Support Vector Machine (SVM)
- Decision trees

Support Vector Machine (SVM) - SVM is supervised learning algorithm that is used to find the best decision boundary that separates the data into two or more classes. SVM uses a technique called kernel trick to transform the data into a higher-dimensional space where a linear decision boundary can be found.

Decision Trees - Decision trees are a type of Supervised learning algorithm that is used to generate a decision boundary by recursively partitioning the feature space into smaller subspaces. The decision boundary is represented by a tree-like structure where each node represents a decision based on a feature value and each leaf represents a class label.

Linear decision Boundary - A linear decision boundary is a straight line that separates the data into two classes. It is the simplest form of decision boundary and is used when the classification problem is linearly separable. Linear decision boundary can be expressed in the form of a linear equation $y = mx + b$, where m is the slope of line & b is y-intercept.



Gradient Descent - It is used to optimize the weight & biases based on the cost function. Cost function evaluates the difference between the actual and predicted outputs. It works by iteratively adjusting the weights or parameters of the model in the direction of the negative gradient of the cost function until the minimum of the cost function is reached.

Gradient Descent Strategies -

- Batch Gradient Descent :- To update the model parameter values like weight & bias, the entire training dataset is used to compute the gradient & update the parameters at each iteration. This can be slow for large datasets but may lead to a more accurate model.

  It is effective for convex or relatively smooth error manifolds because it moves directly toward an optimal solution by taking a large step in the direction of the negative gradient of cost function.

- Stochastic Gradient Descent [SGD] :- In SGD, only one training example is used to compute the gradient & update the parameters at each iteration. This can be faster than Batch gradient descent but may lead to more noise in updates

- Mini-Batch Gradient Descent :- In mini-batch GD a small batch of training examples is used to compute the gradient & update the parameters at each iterations. This can be a good compromise between batch gradient descent & Stochastic gradient descent, as it can be faster than batch GD & less noisy than SGD.