SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

COURSE NAME: 19TS622-MACHINE LEARNING

III YEAR /VI SEMESTER

Unit 1- INTRODUCTION

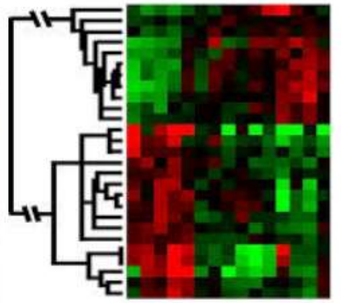
Topic 2: Machine Learning Workflow

Brain Storming

When Do We Use Machine Learning?

- ML is used when:
- Human expertise does not exist (navigating on Mars)
- Humans can’t explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)

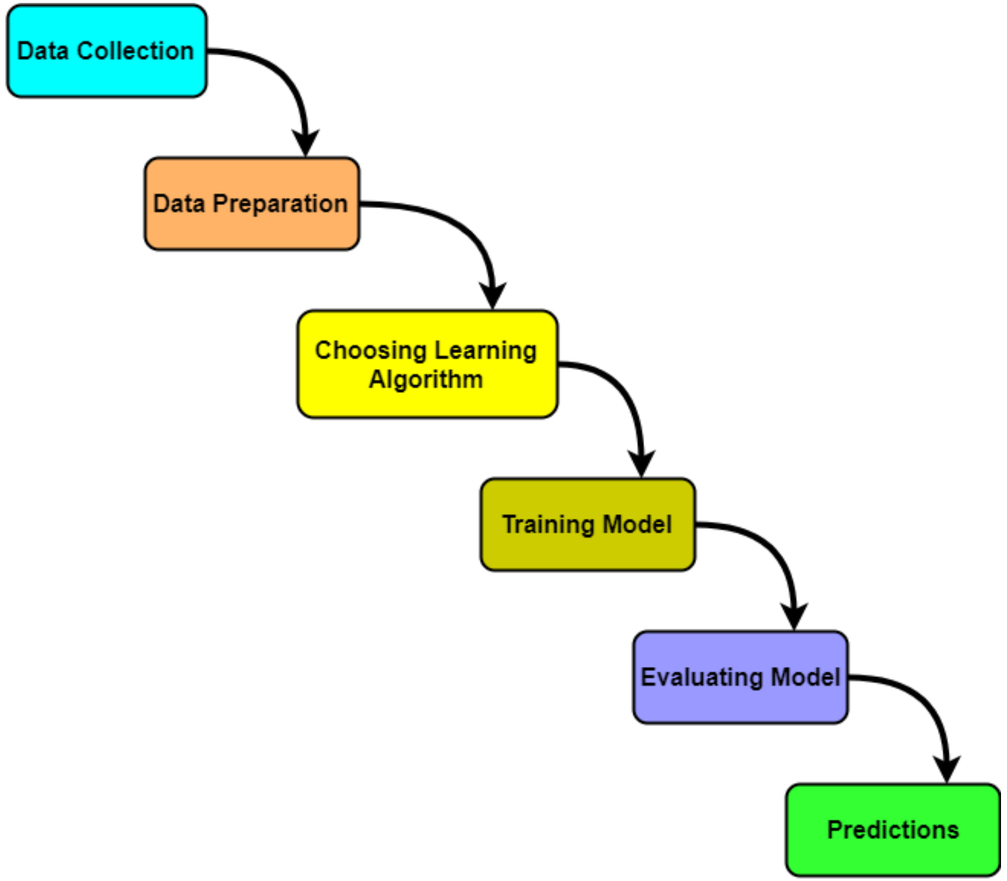


Machine Learning Workflow

- Machine learning workflow refers to the series of stages or steps involved in the process of building a successful machine learning system.

- The various stages involved in the machine learning workflow are-



Machine Learning Workflow

Machine Learning Workflow

- Data Collection
- Data Preparation
- Choosing Learning Algorithm
- Training Model
- Evaluating Model
- Predictions

Data Collection

- Data is collected from different sources.
- The type of data collected depends upon the type of desired project.
- Data may be collected from various sources such as files, databases etc.
- The quality and quantity of gathered data directly affects the accuracy of the desired system....

Data Preparation

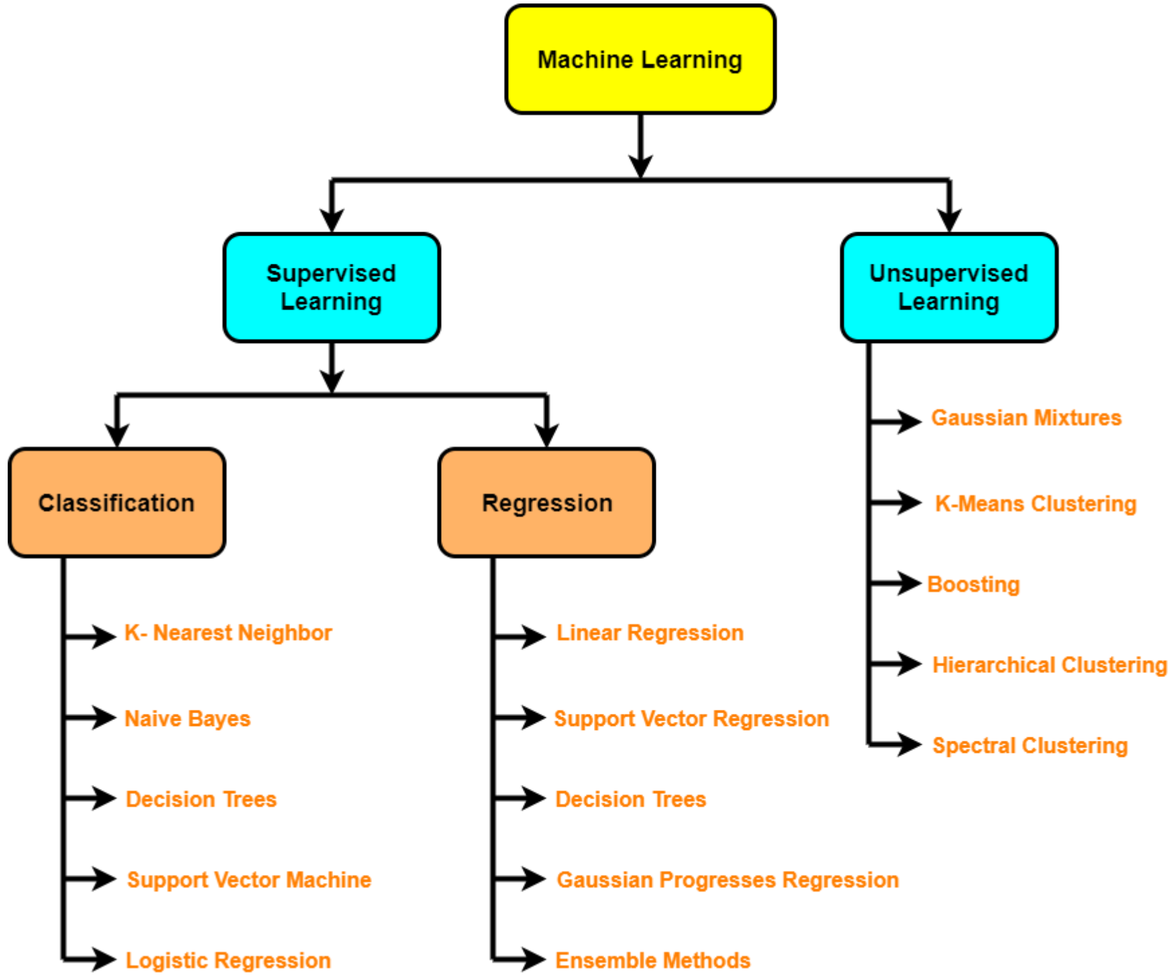
- Data preparation is done to clean the raw data.
- Data collected from the real world is transformed to a clean dataset.
- Raw data may contain missing values, inconsistent values, duplicate instances etc.
- So, raw data cannot be directly used for building a model.

Different methods of cleaning the dataset are-

- Ignoring the missing values
- Removing instances having missing values from the dataset.
- Estimating the missing values of instances using mean, median or mode.
- Removing duplicate instances from the dataset.
- Normalizing the data in the dataset.

Choosing Learning Algorithm

- The best performing learning algorithm is researched.
- It depends upon the type of problem that needs to solved and the type of data we have.
- If the problem is to classify and the data is labeled, classification algorithms are used.
- If the problem is to perform a regression task and the data is labeled, regression algorithms are used.
- If the problem is to create clusters and the data is unlabeled, clustering algorithms are used.

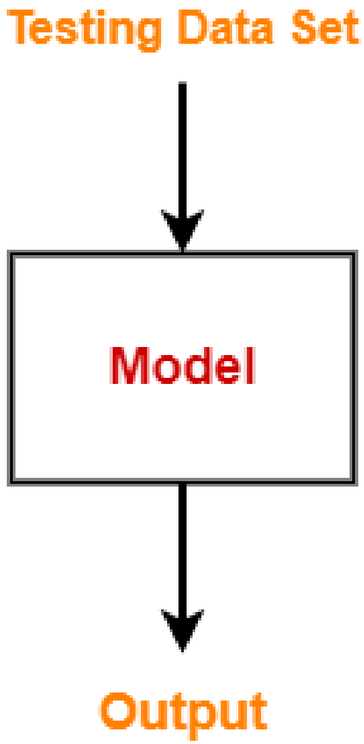


Training Model

- The model is trained to improve its ability.
- The dataset is divided into training dataset and testing dataset.
- The training and testing split is order of 80/20 or 70/30.
- It also depends upon the size of the dataset.
- Training dataset is used for training purpose.
- Testing dataset is used for the testing purpose.
- Training dataset is fed to the learning algorithm.
- The learning algorithm finds a mapping between the input and the output and generates the model.

Evaluating Model

- The model is evaluated to test if the model is any good.
- The model is evaluated using the kept-aside testing dataset.
- It allows to test the model against data that has never been used before for training.
- Metrics such as accuracy, precision, recall etc are used to test the performance.
- If the model does not perform well, the model is re-built using different hyper parameters.
- The accuracy may be further improved by tuning the hyper parameters.



Predictions

- The built system is finally used to do something useful in the real world.
- Here, the true value of machine learning is realized.

Assessment 1

Class C of a “family car”

Prediction: Is car *x* a family car?

Knowledge extraction: What do people expect from a family car?

Output:

Positive (+) and negative (–) examples

Input representation:

x_1: price, x_2 : engine power

References

TEXT BOOKS

1. Alpaydin Ethem, “Introduction to Machine Learning”, MIT Press, Second Edition, 2010

REFERENCES

1. AurélienGéron, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition, o'reilly, (2017)
2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning: with Applications in R”, Springer; First Edition 2013.
3. P. Flach, —Machine Learning: The art and science of algorithms that make sense of data,Cambridge University Press, 2012.
4. Tom M. Mitchell, “Machine Learning”, McGraw-Hill Education (India) Private Limited, 2013.

Thank You