Machine Learning Concepts and its Applications

K.UlagaPriya, K.Kalaivani, A.Saritha

Department of Computer Science and Engineering

Vels Institute of Science Technology and Advanced Studies, Chennai, India.

Abstract

Machine Learning is all about learning the data using various algorithms. Huge amount of data is available everywhere and it is essential to analyse the existing data and provide useful information. By using Machine Learning algorithm, user can predict the future using the existing data or the user can provide an Inference about the existing data based on the relationship among the data. Various applications like Financial Marketing analysis, Natural Language Processing, Telecommunication, Fraud Detection etc, uses Machine learning Techniques. Basically learning can be done using two approaches such as parametric and non parametric approach. In parametric approach, some assumptions are made before the algorithm is applied. In non parametric no make such explicit assumptions are made.

1. Introduction

Major classification of Machine Learning techniques are Supervised Learning and Unsupervised Learning. In Supervised Learning each set of Input variables is associated with one output variable. In UnSupervised Learning there is no output variable, and hence the pattern of data is analysed. There are several techniques for Supervised Learning and they are as follows:

- Linear Regression
- Logical Regression
- Linear Discriminant Analysis
- K-Nearest Neighbor
- Ridge Regression
- Lasso Regression
- Support Vector Machine
- Decision Trees
- Random Forest

Following are the unsupervised Technique

- ▪ K means Clustering
- ▪ Hierarchical Clustering
- ▪ Probablistic Clustering

2. Regression Vs Classification Problems

Input variables can be classified as quantitative or qualitative. Quantitative variables can be numerical values whereas qualitative variables are categorical variables. Problems which are related to Quantitative is regression problem. Problems which are related to Qualititative are termed as classification The selection of learning methods is based on whether the output variable is Quantitative or Qualitative i.e Linear regression is used when it is quantitative and Logistic regression is used when it is qualitative.

3. Overfitting Vs Underfitting

Before applying a machine learning technique, the dataset needs to be spitted as training data and test data. Usually 70:30 split is preferred. For example student dataset, 70% of data can be considered as training set and the remaining 30% of dataset can be treated as test data set. After a model or algorithm has been processed by using the training set, you test the model by making predictions against the test set. A common danger in machine learning is overfitting — producing a model that performs well on the data you train it on but performs poorly to any new data or test data. The other side of this is underfitting , producing a model that doesnt perform well even on the training data. The remedy is to move on and try alternate machine learning algorithms

4. Bias Vs Variance Tradeoff

Bias is the difference between the actual value and the predicted value. Usually in Parametric methods bias will be high. Generally Linear Regression, Linear Discriminant analysis, Logistic regression has high bias since it has high assumptions. K-Nearest Neighbor, Support Vector Machine has low bias since the assumptions are low.
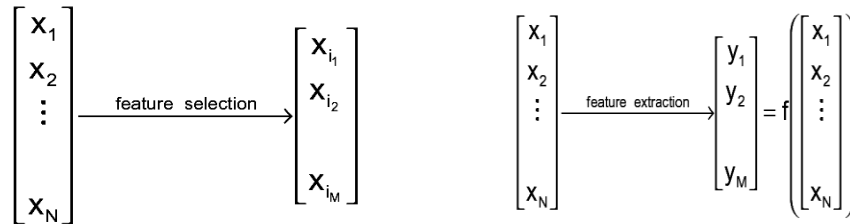
Variance is the amount of change respective to new data.When the model is build with training set ,and when new data is introduced the variance should be less. Generally Linear Regression, Linear Discriminant analysis, Logistic regression has low variance because when new data point is introduced the variance will be low .Whereas when new data point is introduced in models like K-Nearest Neighbor, Support Vector Machine has high variance.

5. Feature Selection Vs Feature Extraction

The two general approaches for dimensionality reduction are :

Feature Extraction is nothing but transforming the existing Features into a lower dimension space

Feature Selection is the selecting a subset of existing features without any change.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature selection}} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_M} \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \right)$$

6. Machine Learning a Comparative Study

Artificial Intelligence VS Machine Learning Vs Deep Leaning : Artificial Intelligence is the science of making machines to think as like human beings. It is a task of using computers to understand human intelligence. It is a t Machine Learning is a subset of Artificial Intelligence. It is the study of predicting the future based on existing data and providing inference about the data using various algorithms. Deep learning is one of many approaches to machine learning.
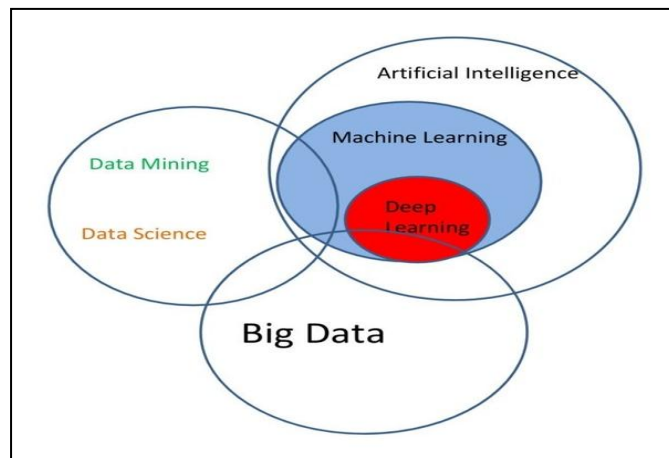
Data Mining vs Data Analytics vs Machine Learning :

One key difference between machine learning and data mining is how they are used and applied in our everyday lives. Datamining is collecting huge amount of information and using that information some insights can be done. Data analytics is part of data science where data analysis are done using statistical methods.

7. Machine learning in various fields

Sehla Loussaief ; Afef Abdelkrim in 2016 [1] described Machine learning framework for image classification. There are several feature extraction is available and this paper evaluates the use of the classical SURF technique against global color feature extraction. The main idea of this paper is to identify the best machine learning techniques to recognize the stop sign images. Berina Alic, Lejla Gurbeta , Almir Badnjevic in 2017 [2]demonstrated   machine learning techniques for classification of diabetes and cardiovascular diseases. This paper presents the overview of machine learning techniques in classification of diabetes and cardiovascular diseases (CVD) using Artificial Neural Networks (ANNs) and Bayesian Networks (BNs).It was identified that Neural networks gives better accuracy during classification. M. Krendzelak ; F. Jakab in 2016 [3]published Text categorization with machine learning and hierarchical structures. This paper aims to provide the overall view of using different machine learning techniques for text categorization

Sunil Kr.Jha ZulfiqarAhmad 2018[4] analysed about Soil microbial dynamics prediction using machine learning regression methods. This used Fuzzy based system for prediction. The performance measures is identified and published the results. LukasLingitza, ViolaGallinaa has done a Lead time prediction using machine learning algorithms. Nasser H.SweilamaA.A.TharwatbN.K.Abdel Moniemc[5] did a comparative study on Support vector machine for diagnosis cancer disease. Using a non linear classifier the diagnosis and prediction of cancer is done using support vector machine. Stroke prediction, Stock Forecasting, Cost Prediction Modelling[Jiang Tiejun ; Zhang Huaiqiang] using SVM uses kernel functions for prediction which proposes high accuracy when compared to other models. Performance tuning methods are applied for better prediction.

8. Conclusion

Thus the various concepts related to machine learning is discussed and the basic differences between the terminologies in the world of data analytics. The comparative study exhibits the effective way of choosing the appropriate technique for a particular application places a vital role in determining the accuracy. The usage of machine learning is very wide and it is one of the emerging trends in Information Technology.

9. References

[1]Sehla Loussaief ; Afef Abdelkrim Machine learning framework for image classification Published in 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT).IEEE

[2]Berina Alic, Lejla Gurbeta , Almir Badnjevic in 2017 [2]demonstrated Machine learning techniques for classification of diabetes and cardiovascular diseases Published in 2017 6th Mediterranean Conference on Embedded Computing (MECO), IEEE

[3] Krendzelak ; F. Jakab Text categorization with machine learning and hierarchical structures Published in 2015 13th International Conference on Emerging eLearning Technologies and Applications (ICETA),IEEE

[4]Sunil Kr.Jha ZulfiqarAhmad Soil microbial dynamics prediction using machine learning regression methods, ,Elsevier

[5]LukasLingitza, ViolaGallinaa Lead time prediction using machine learning algorithms, 51st CIRP Conference on Manufacturing Systems,Elsevier

[6]NasserH.SweilamaA.A.TharwatbN.K.Abdel Moniem A comparative study on Support vector machine for diagnosis cancer disease Elsevier.

[7] Jiang Tiejun ; Zhang Huaiqiang Study on Cost Prediction Modeling with SVM , 2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, IEEE