Statistical Roots of Machine Learning, Deep Learning, Artificial Intelligence, Big Data Analytics and Data Mining

Ali Serhan Koyuncugil,Nermin Ozgulbas

***Abstract*— In this study, statistical roots of main intelligent data analytics are discussed. Furthermore, vital elements of statistical data analysis are given and its extensions on Machine Learning, Deep Learning, Artificial Intelligence, Big Data Analytics and Data Mining are explained in details with method examples.**

***Index Terms*— Artificial intelligence, big data analytics, deep learning, machine learning**

I. Introduction

WHILE computational accelaration getting day by day, data from different sources in different formats increasing exponentially. Therefore, the need for processing, understanding, modelling and valuating this enormous data has been caused new approaches, disciplines and techniques such as Machine Learning, Deep Learning, Artificial Intelligence, Big Data Analytics, Data Mining and more. Actually, it is hard to follow this sequential methods flood. Because, in every new day almost the same things are coming us with new labels. On the other hand, all these brand new data technologies have the same roots which is statistical data analysis or simply statistics. In this study, statistical bases of almost all developing intelligent data analytics are presented in details..

II. SOME VITAL STATISTICAL ELEMENTS

A. Normal Distribution

Normal Distribution or Gauss Distribution is one of the vital elements in statistical theory and one of the vital basis for (statistical) data analysis:

- Most of natural events fit Normal Distribution
- Most of random distributions converge to Normal Distribution
- Most of discrete or continuous variables converge to Normal Distribution under some assumptions

One of the most frequently use of Normal Distribution is Linear Regression. Furthermore, Principal Component Analysis and Discriminant Analysis as an extension of Linear Regression **[1]**.

B. Parametric Methods vs. Non-Parametric Methods

One of the main important issues in statistical data analysis is find a suitable distribution (distribution fitting) for variables. Find a suitable distribution means that variables are suitable for Parametric Methods. Otherwise, non-parametric methods or order statistics (ranks based) methods should use for data analysis. Parametric methods mostly prefer instead of non-parametric ones. Because, parametric methods require (probability) distributions and their accuracy always better than non-parametric ones **[1]**. Usually Goodness of Fit Tests (Kolmogorov-Smirnov, Chi-square etc.) use to identify the suitable probability distribution. Then the suitable parametric methods use due to the probability distribution. In case, the data fails to distribution fitting, then suitable non-parametric methods should use for analysis or modelling.

C. Non-Linearity

Most of the multivariate statistical data analysis which are using as a base for machine learning, deep learning, artificial intelligence and big data analytics methods based on linear methods such as Linear Regression, Principal Component Analysis and Linear Discriminant Analysis. In case lack of linearity conditions the non-linear modelling should use for data analysis. Currently, one of the most popular ways to analyze non-linear data is Neural Networks. Mainly, Neural Networks Method uses activation functions such as Sigmoid Function which is given in Figure 1 **[2]**, **[3]**. Furthermore, Sigmoid Function converges Linear Regression in case of modelling linear data. On the other hand, in case understand the data is linear before apply the NN then it is possible to develop a Linear Regression Model with better solutions than the NN Model. As a result, the most important fact is to apply the suitable method to the suitable data.

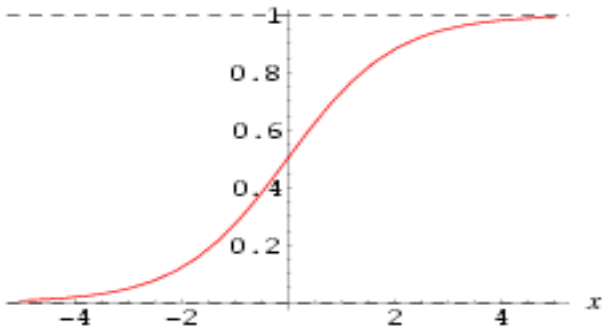

Figure 1. Sigmoid Function

Manuscript received July 13, 2019; revised August 10, 2019.

A. S. Koyuncugil, Department of Insurance and Risk Management, Baskent University, Baglica Campus, Eskisehir Yolu 18. Km. Ankara, Turkey (Phone: + 90 532 665 70 84; e-mail: koyuncugil@ baskent.edu.tr).

N. Ozgulbas, Department of Healthcare Management, Baskent University, Baglica Campus, Eskisehir Yolu 18. Km. Ankara, Turkey (e-mail: ozgulbas@baskent.edu.tr).

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

One of the other ways for modelling non-linear data is piecewise linearity approach. Mainly, piecewise linearity aims to model the non-linear data with linear pieces. Piecewise linear example for Normal Distribution is given in Figure 2 **[4]**.

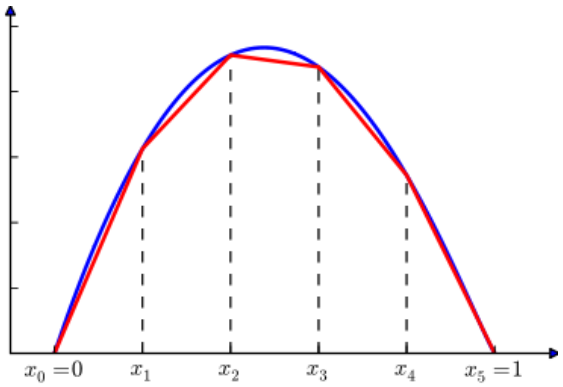

Figure 2. Normal Distribution and its piecewise linearity approach

D. Supervised Methods vs. Unsupervised Methods

One of the most confused concept for data analysis is to use supervised or unsupervised methods. It is possible to make different explanations for this concept but its simply about whether we can determine the output zone before analysis or not. In case the output zone can determine before the analysis then we can call it supervised. Otherwise, unsupervised **[1]**.

We can easily use Cluster Analysis for determine the distinction between supervised and unsupervised methods. In K-means Cluster Analysis we know the data will have K clusters before we begin the analysis. So, the K-means Analysis can easily identify as Supervised. On the other hand, we couldn't know the number of clusters before we begin Hierarchical Cluster Analysis. So, the Hierarchical Cluster Analysis can identify as Unsupervised.

Mainly, unsupervised methods can use to identify/explore the data and then the supervised methods can use to verify and model the data.

E. Predictors

Statistical Learning Theory or in other words 'Learning from Data' aims to predict non-observed data based on past data. In every data analysis we have to find statistically significant predictors for predicting the better non-observed data [3], [5].

III. STATISTICAL ROOTS OF INTELLIGENT ANALYTICS

A. Machine Learning

We may define Machine Learning as programming the machines for solve a problem with a sample data. Programming means developing a predictional model for predicting the non-observed data based on past (sample) data **[1], [3]**. A classification diagram of Machine Learning Techniques are given in Figure 3 **[6]**.

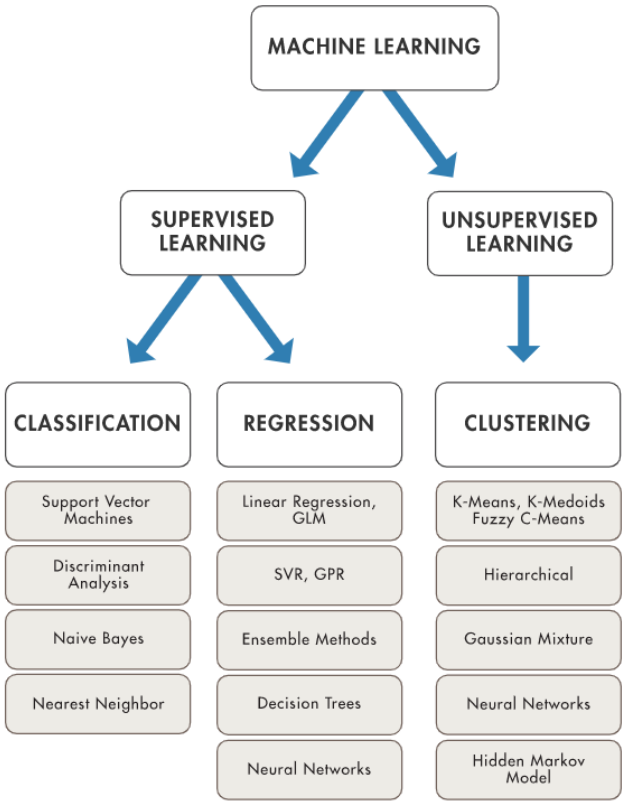

Figure 3. Machine Learning Classification

When we consider the Figure 3, we may easily see that we are talking about the some of the Multivariate Statistical Data Analysis methods. Many machine learning methods come from the modification of the sub-set of Multivariate Statistical Methods.

B. Deep Learning

Nowadays, one the most popular methods is Deep Learning after availability of high performance computational devices such as GPU (or TPU) cards. Mainly, image recognition accuracy makes Deep Learning popular. Mainly, Deep Learning is a sub-set of Machine Learning which is using for big data with a high performance computers. One of the main distinctions between the Machine Learning and Deep Learning is to identification of the interactions between the variables. It is easy to identify the relation between variables in machine learning. On the other hand, many unknown computational units (hidden layers) are using to model the relation between input and output for better prediction **[7].**

However, when we are talking about the Deep Learning, it means that simply we are talking about Neural Networks which is based Non-Linear Statistical Modelling or in general sequential use Machine Learning Techniques which are Multivariate Statistical Data Analysis Techniques.

C. Artificial Intelligence

Artificial Intelligence is simply defines that an artificial presence (computer) which shows intelligence. It means that a computer/machine which makes a decision by itself on an unobserved case **[1]**.

Main difference between Artificial Intelligence and Machine Learning is that AI is developed for decision making and ML is developed for learning new things from

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

data. Therefore, we may define AI is a developed ML System for decision making.

Most popular AI methods are given below **[8]**:

Case Based Reasoning: Using similarity measures which are using in Statistical and Probability Theory.

Rule Based Reasoning: Using if-then rules in different forms. Mainly it uses Conditional Probability Theory.

Artificial Neural Networks: It is a non-linear modelling method which was discuss in Section 2.3.

Genetic Algorithms: A search technique based on sampling which is based chromosomes behaviors.

Fuzzy Systems: (Ordinary) Probability Theory assigns values as {0,1} while Fuzzy Set Theory assigns values [0,1] via Membership Functions.

D. Big Data Analytics

At the beginning of 2000's increasing data assumed as a chance and sequentially databases, data marts and data warehouses were developed for processing and valuating the data for better (strategic) decision making. On the other hand, at the beginning of 2010's huge, flowing data from different sources (such as social media) from different formats (such as image, video, voice) became a curse and a new understanding developed beyond the (structured/ SQL) databases as File Systems (mainly calls NoSQL). As a result, all intelligent analytics modified as big data analytics with some minor changes in Statistical Analysis, ML, DL and AI techniques mainly in data manipulations (replication, duplication etc.) for prevention of loss of data.

E. Data Mining

From Knowledge Discovery Concept since 1976 until now all evolutional statistical analysis, ML, DL, AI methods gathered under one umbrella as data mining for analyzing big data for strategic decision making **[3], [9], [10]**.

IV. CONCLUSION

In this study, statistical bases of all popular intelligent analytical techniques are presented in methodological details. It reveals that adequate multivariate statistical data analysis knowledge is a must for an accurate intelligent analytical model development which are mainly Machine Learning, Deep Learning, Artificial Intelligence, Big Data Analytics and Data Mining.

REFERENCES

[1] A.S. Koyuncugil, 'Lecture notes for data mining course', [Online]. Available: http://www.koyuncugil.org/dosyalar/vmders.pdf.

[2] A. Berson, S. Smith and K. Thearling, Buildind data mining applications for CRM. McGraw Hill, 510, USA, 1999.

[3] A.S. Koyuncugil, 'Fuzzy Data Mining and Its Application to Capital Markets' Unpublished Doctoral Dissertation, Ankara University, Ankara, Turkey, 2006

[4] Wikimedia Commons, [Online]. Available: https://en.wikipedia.org/wiki/Piecewise_linear_function [4]).

[5] T. Hastie, R. Tibshirani J. Friedman, 'The elements of statistical learning; data mining, inference and prediction', Springer Series in Statistics, 533, USA, 2001

[6] Machine learning from MathWork https://www.mathworks.com/discovery/machine-learning.html

[7] Deep learning from MathWork https://www.mathworks.com/discovery/deep-learning.html

[8] S.H. Chen, A.J. Jakeman and J.P. Norton, 'Artificial Intelligence techniques: An introduction to their use for modelling environmental systems', Mathematics and Computers in Simulation 78 (2008) 379–400

[9] **U.** Fayyad, G. Piatetsky-Shapiro and P. Symth, 'From data mining to knowledge discovery in databases', AI Magazine, 17(3); 37-54, 1996

[10] A.S. Koyuncugil and N. Ozgulbas, N (Eds.), 'Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection', IGI Global, USA, 2010

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)