Introduction to Data Science - An Overview

Thangakumar Jeyaprakash^{#1}, Padmaveni K^{#2}

^{#1#2}Associate Professor and Department of Computer Science and Engineering
^{#1#2}Hindustan Institute of Technology and Science, India

Abstract Data science plays a vital role in the research field of computer science and engineering which involves collection of data, transformation, processing, describing, and modelling. In this article, fundamental theory of Data Science, Machine learning and Deep Learning with the scope and opportunities has been discussed. This helps the researchers to get a clarity on data science and its importance.

Keywords— Data science, Machine learning, Deep Learning

I. INTRODUCTION

Data science is the science of collecting, storing, processing, describing, and modelling of data

Fig 1. Five tasks of Data science

II. DATA COLLECTION:

Data scientist must be work on the following questions to collect the data

- What to collect?
- How to collect?
- Where to collect?
- Label data or not?

Suppose a data scientist must collect data [1] on shopping mart, he needs to identify the items need to be purchased together to do some statistical modelling [2]. For example, a mobile phone may be purchased along with SD card, charger etc. a laptop may be purchased together with laptop bag, and wireless mouse. Such data helps the data scientist to do some statistics like which items do customer purchase together? Or which items moving faster etc. Data scientist should have the knowledge of structured query language or python packages to do analytics.

III. DATA STORAGE

The collected data would be a transactional or operational data [3] such as an insurance schemes, stock inventory, customer records, employee records, student records, patients records, purchase orders, stock market, school children record etc.

Fig 2: Data records

These data can be divided into structured data and unstructured data [4]. Data in a structured format or readily available for analytics collected from multiple relational database management system (RDBMS) are known as structured data which can be integrated into a common repository to support data analytics. The unstructured data referred as text, image, voice data, data from social blogs to do some sentimental analysis. The data which are unstructured needs to be cleaned or pre-processed to support analytics. A large collection of data storage with both structured and unstructured data is known as data lake referred to as a Big Data. It consists of uncured, useful or the data is not useful, and the data from everywhere.

IV. DATA PROCESSING

The data processing involves data wrangling, data cleaning and data scaling

Figure 3: Data Processing

A. Data wrangling or Data Munging

Data wrangling involves the extraction, transformation and load known as ETL process. The extraction of data from one database to another will be transformed into another form to support analytics which will be loaded to the data warehouse is known as Data wrangling or Munging. Some of the ETL [5] tools are IBM infosphere, informatica, Oracle data integrator, Alooma, Fivetran, Snaplogic etc.

B. Data Cleaning

When extracting data from multiple databases, some values may be missed in columns. To fill such missing values, to standardize the keyword tags, to correct the spelling errors or to identify and remove unwanted data, data cleaning [6] is required.

C. Data Scaling

The normalizing and standardizing of data are known as Data scaling. For example, converting all kilometres to miles, rupees to dollars.

D. Describe:

After processing the data, it can be described in a visualized form using a different type of plots such as box plot, scattered plot, or line chart. For example, a car show room want to visualize the car sales in terms of colours such as red, blue and green, then the data can be described with the following example.

TABLE 1:

Car sales with respect to colour

Colour	2018	2019
Red	9	7
Blue	7	5
Green	5	3

In figure 4, the data has been described visually. It concludes, red has the maximum sales every year compared to other car colours while comparing the car sales for the 2018 and 2019 with respect to red, blue and green. With this sales analysis, the data scientist may generate a report to increase the production of red coloured vehicles to attract more customers and to increase sales.

Fig 4: Number of cars sold vs car colour

V. DATA MODELLING

The data can be modelled after collecting, storing, processing, and describing the data. The data can be modelled in two ways.

- A. Statistical Modelling
- B. Algorithm Modelling

A. Statistical Modelling:

We assume very simple models which allows robust statistical analysis which gives statistical guarantees.

$$y = f(x)$$

Where y is outcome, f is function and x is input. The outcome will be depending upon the value of x . x may be anything. If you want to make a coffee, the taste of the coffee is depending upon the quantity of coffee power, sugar and milk. Here, the outcome y is coffee, and x are coffee power, sugar and milk.

$$y = f(x_1, x_2, x_3)$$

Where x_1 , x_2 , x_3 are coffee power, sugar and milk respectively. These simple models can be derived using statistical modelling. The statistical modelling involves linear regression, logistic regression and linear discriminant analysis.

B. Algorithm modelling:

The complex problems can be derived using algorithm modelling where we can train a model using the tasks of data sciences. More multi-dimensional data [7] will be used here for statistical analysis. It mainly

focuses on prediction and optimization with more solutions for Machine learning [8] and Deep learning algorithms. It involves the statistical modelling methods such as linear regression, logistic regression, linear discriminant analysis and the algorithm-based methods such as Decision[9][10 trees, k-Nearest neighbour, Support vector machine, Naïve Bayes methods and Multi layered neural networks.

Fig 5: Algorithm Modelling

VI. CONCLUSIONS

This paper focuses mainly the fundamental theory of data sciences in a detailed approach which helps researchers to get clarity in collecting data, data storage, data processing, describing, data modelling. Machine learning and deep learning algorithms

REFERENCES

- [1] "Data Science and Analytics", Department of Industrial and Systems Engineering Lehigh University Technical Report 15T-009
- [2] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: Proceedings of ICDT. pp. 398–416. Springer (1999)
- [3] Aden, C, Kleppin, L, Schmidt, G and Schröder, W. 2009. Consolidation, visualisation and analysis of forest condition relevant data in the WebGISWaldIS. In: Strobl, J, Blaschke, Th and Griesebner, G (eds.), Angewandte Geoinformatik 2009, 506–515
- [4] Chavan, V and Penev, L. 2011. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics, 12(Supl. 15): S2. DOI: <https://doi.org/10.1186/1471-2105-12-S15-S2>
- [5] Tseng, V.S., Wu, C.W., Fournier-Viger, P., Philip, S.Y.: Efficient algorithms for mining the concise and lossless representation of high utility itemsets. IEEE Trans. Knowl. Data Eng. 27(3), 726–739 (2015)
- [6] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmman, T., Sun, S., Zhang, W.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 601–610. ACM (2014)
- [7] Gramener: Gramener—a data science company. <https://gramener.com/> (2018). Accessed Dec 2018
- [8] Polyzou, A., Karypis, G.: Grade prediction with models specific to students and courses. Int. J. Data Sci. Anal. 2(3–4), 159–171 (2016)
- [9] Lam Ky Nhan, Phuong Hoang Yen "The Effects of Using Infographics-based Learning on EFL Learners' Grammar Retention" International Journal of Science and Management Studies (IJSMS) V4.I4 (2021): 255-265
- [10] Nadia Ulfa Agustin Hilman, Maya Ariyanti, AstriGhina "Social Media Marketing Effect towards Purchase Decision at the Embroidery MSMEs in Tasikmalaya" International Journal of Science and Management Studies (IJSMS) V4.I4 (2021): 202-209.