Raw Data to Feature Space*

Christina Fleurisma[1]

***Abstract*— The main focus of this paper is on challenges faced during the extraction from given raw data into feature space along with appropriate subspace in order to develop and train machine learning models.**

I. INTRODUCTION

Data science has become one of the most prominent field in Computer Science. It works by collecting, preparing, analyzing, visualizing, and preserving data. Data Science, especially Big Data and machine learning, there are four essential objectives which goes from the understanding of data, the understanding of machine learning, the understanding of systems, and the understanding of scalability and complexity. In addition, the learning and understanding of programming languages like Python, R, and, Scala and the most recent big data systems such as Hadoop and Spark is crucial for data analysts as well as data scientists.

II. SETTING PROGRAMMING ENVIRONMENT

Machine learning requires specific coding environment in order to collect, visualize, and transform data. For this assignment, anaconda was highly recommended and python was a suggested programming language. In order to download anaconda3, the used of some specific command lines were necessary.

cd Downloads/
bash script.sh
source activate ComputerVision
conda env remove -n ComputerVision

A.

III. PROCEDURE FOR IMAGE SELECTION

A.

In this assignment, one of the challenges was the choice of the data. Choosing which pictures, in this case would determine the learning efficiency of the machine. In this case, the pictures of cherry, mango, and pineapple were used taking into account multiple features such as size, texture, shape, and color. While cherry is small with a glossy texture, round shaped, and dark plain red; mango is medium size with a soft but non glossy texture, oval shape and a fading red to yellow and greenish color; pineapple on the other hand is bigger, has a rough, lumpy texture, a more cylinder shaped bottom with a green hat/leaves.

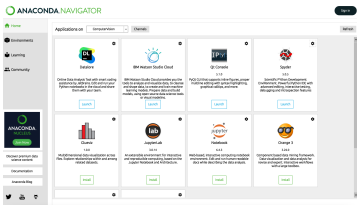

Fig. 1. a snapshot Anaconda3 enviroment

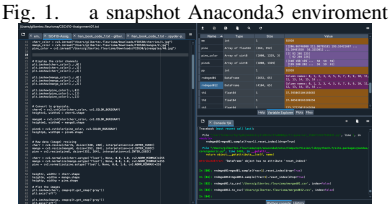

Fig. 2. a snapshot of Spyder IDE

Fig. 3. Cherry photo.

Fig. 4. Mango photo

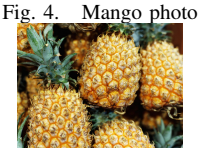

Fig. 5. Pineapple photo

The dataset is accurate, balanced, and complete. It could be considered as big data with no that need to be scaled and normalized.

IV. CONCLUSION

The goal of this assignment were to extract feature from a set of given data, to construct a feature space or spaces, and finally to develop and train machine

*This work was not supported by any organization

[1]H. Kwakernaak is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands h.kwakernaak at papercept.net

[2]P. Misra is with the Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA p.misra at ieee.org

learning models. The challenges faced during this assignment such as creating the programming environment, selecting the data, scaling and normalizing the data were met successfully. Throughout this work, it was shown how computers can process and analyze data.