Feature Engineering:
Data Preprocessing for Machine Learning

Bachelorarbeit

Zur Erlangung des akademischen Grades „Bachelor of Science (B. Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Siegmann Vorname: Colin M.

[redacted] [redacted] [redacted] [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, den 11.08.2017

Contents

1 Introduction

Machine Learning is an important subject which almost everyone comes in contact with nowadays. Many might have heard about it, some may have not but Machine Learning is embedded in way more fields than most people may know. A lot of research has been done in the field of Machine Learning algorithms, yet the field of Feature Engineering is a subject limited within studies.[1] This, combined with the fact that Feature Engineering incorporates domain knowledge into Machine Learning makes this research field interesting and important. Machine Learning and Feature Engineering provide many possibilities and help companies move forward in many ways. However, it would be exaggerated to consider Machine Learning as replacement for human labor. Some jobs may be made obsolete when Machine Learning is widely applied, therefore, new jobs and even new occupational areas emerge. Especially experts with a lot of experience and domain knowledge are crucial for successful Feature Engineering and thus successful Machine Learning.[2] The purpose of this paper is to qualitatively explain what Machine Learning and Feature Engineering are and what role Feature Engineering plays within Machine Learning. Many scientific papers explain Feature Engineering and Machine Learning rather mathematically. As many people who are affected by Machine Learning are not mathematicians, this paper intents to explain these constructs understandably also for non-mathematicians. The illustration of problems, theories and application examples is made in order to provide a good overview of this very large thematical complex. The vast majority of this work concentrates especially on classification.

Firstly, the conceptualities and boundaries of Machine Learning as well as important concepts will be defined. The role of Machine Learning in the modern world will be analyzed with the help of Gartner´s Hype Cycle. Afterwards different types and tasks of Machine Learning will be differentiated. Some problems as well as performance measurements are revealed. This information allows to explain what Feature Engineering is, how it works and why it is done. Then theoretical approaches to Machine Learning and Feature Engineering are looked at. Afterwards some application fields for Feature Engineering in Machine Learning are presented and practical examples on how Feature Engineering is done, are given. Finally, the revealed information and results are discussed in order to evaluate the role of Feature Engineering for Machine Learning. This discussion leads to limitations and implications for science and practice. At last a conclusion regarding Feature Engineering for Machine Learning is presented.

[1] Cf. Rawat & Khemchandani, 2017, p. 169
[2] Cf. Domingos 2012, pp. 82-83

7 Conclusion

This paper had the intention to qualitatively explain the constructs of Feature Engineering and Machine learning and illustrate which role Feature Engineering has for Machine Learning. This was done as described in the following: Firstly, the relevant conceptualities were explained and defined. The types of Machine Learning, the general interest for it, typical problems and performance measurements were illustrated in order to explain the concept of Feature Engineering. Some theoretical approaches were explained in order to allow a scientific evaluation based on these theories. Afterwards, different application fields for Feature Engineering in Machine Learning classification were explained and practical examples for Feature Engineering were given. These insights allowed to qualitatively discuss the construct of Feature Engineering within Machine Learning. Finally, implications for economy and science were provided and the limitations of this paper were revealed.

It was shown that data pre-processing for Machine Learning is a large and complex field. Machine Learning is widely applied and will be adapted even more in the near future. Many different types of problems can be solved and tasks can be accomplished by Machine Learning. Even though a consensus on the importance of Feature Engineering exists in order to objectively show its benefits, metrics and performance measurements have to be utilized. Many theoretical approaches for Machine Learning exist but specifically Feature Engineering is a rather practical field with many hit-and-miss situations and domain knowledge. Nevertheless, the advantage of Machine Learning is shown in fields as such as text-, audio- and image-classification. However, not all fields are perfectly suitable for Machine Learning. The degree to which Machine Learning and Feature Engineering should be used lies in the decision of executives. They have to analyze the cost-benefit ratio in order to allow for a good decision. Supervised Learning is still state-of-the-art in classification but Unsupervised Learning and automated Feature Engineering will slowly be adapted in the future. Machine Learning also leads to difficult questions regarding liability and ethics. Not all of these questions are able to be answered momentarily. Finally, Machine Learning can provide great contributions for companies and enable new possibilities but it cannot take on responsibilities, it cannot make final decisions and its ethics remain complicated. These issues must still be solved by humans.

"Machine Learning is not magic; it can´t get something from nothing. What it does is get more from less." [99]

– Pedro Domingos

[99] Cf. Domingos 2012, p. 80