Topic Models

Anders Hørsted - s082382
Vejledere: Morten Mørup og Ole Winther

Kongens Lyngby, 1. september 2012
IMM - Fagprojekt

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Summary

In this project, the logistic regression model used to win the online machine learning competition “Stay Alert! The Ford Challenge” is recreated, using the Python machine learning library scikits.learn. Performance similar to the performance of the winning model is achieved by the new model.

By doing a forward feature selection, a model is found that uses the same number of features but consistently performs better than the winning model.

Finally it is tried to further enhance the performance by training a neural network model in the Python library PyBrain. No significant improvements are achieved by the neural network, compared to the logistic regression, but this is in part attributed to lack of computational resources.

Resumé

I dette projekt bliver den logistiske regressions model, der blev brugt til at vinde machine learning konkurrencen “Stay Alert! The Ford Challenge”, genskabt ved brug af Python-biblioteket scikits.learn. Resultater magen til resultaterne for den vindende model opnås af den nye model

Ved at lave en *forward feature selection*, laves en ny model, der bruger det samme antal features, men som konsekvent giver bedre resultater end vindermodellen.

Til sidst forsøges det at forbedre resultaterne yderligere, ved at træne et neuralt netværk i Python biblioteket PyBrain. Ingen markante forbedringer opnås af det neurale netværk, sammenlignet med den logistiske regression, men dette tilskrives til dels mangel på computerkraft.

Preface

This project is the result of attending the course "01666 Project work - Bachelor of Mathematics and Technology" at the Technical University of Denmark. The course is mandatory for all Bachelors at Mathematics and Technology and counts for 10 ECTS points.

The subject of the project is held within the discipline called machine learning that, depending on how it is defined, can be seen as part of artificial intelligence or applied statistics.

The target group of the report is any student that have completed the course "01005 Mathematics 1" at DTU or a similar course at another university.

All files used in this project are available at `https://github.com/alphabits/ford_challenge/tree/master/`.

Anders Hørsted
Christianshavn, June 2011

Acknowledgements

I would like to thank my two supervisors Morten Mørup and Ole Winther, for the great help they have given me during this project. At no time have I had trouble getting help when I had questions, even when I showed up unannounced. It has been a great pleasure to have Morten and Ole as supervisors.

Contents

Chapter 1

Introduction

This decade could very well be the most exciting decade for anyone interested in data analysis. The ease with which data can be collected, combined with the very cheap storage prices makes it likely that before the end of this decade, most details about a person's life will in some way be collected digitally. The increasing amount of data calls for a continous refinement and advancement of both the computational and statistical techniques used to analyze the data. As it is exactly this borderland between statistics and computer science, that is the focus area of machine learning, it seems probable that much of the advancement in the near future will happen within the machine learning community. This of cause, makes machine learning an interesting study area and it wasn't a difficult decision to choose machine learning as the general topic for this project.

Although the general topic was chosen, the question still remained about what should be the specific subject of the project. The idea soon emerged that the project could be based on one of the many online machine learning competitions, that have appeared within the last couple of years. The idea behind these online competitions, is that some private or public organization, releases a dataset along with a specific task to be solved using

the dataset. One such competition, called The Ford Challenge, was announced in the beginning of January and as it was an exciting competion it was chosen as the subject for this project.

In The Ford Challenge a model must be build, that can predict when a driver is about to become inattentive. By using data collected from various drivers it should be possible to let a computer detect patterns that are related to an inattentive driver. The task is indeed exciting but as the competition already ended in the beginning of March, active participation in the competition couldn't be the main focus of this project. Instead it was chosen to try and recreate the model of the Ford Challenge winner. When a performance similar to that of the winner was achieved, various techniques to improve on the winning model could be tried. Before any model could be build though, some thorough data exploration had to be done, as almost no information about the dataset was revealed by Ford.

Apart from the goals specifically related to The Ford Competition, it was also strongly desired to set up a working environment for data analysis, based solely on open source tools. When these tools were found some time should be used to set up a flexible and simple working environment.

The problem statement of this project is:

- Try to decide the datatype for all features in the dataset, by doing a thorough data exploration
- Try to recreate the winning model of The Ford Competition.
- Improve on the winning model by using other models than the winner.
- Set up a working environment for data analysis, and reflect on possible ways to ease future data analysis projects.

Much time have went into solving the above problem statement, but much time was also spent working on this report. The report starts out by presenting the Ford Challenge, along with descriptions of some other machine learning competitions. After introducing the Ford Challenge, the dataset of the Ford Challenge is introduced in details. Then the data is thoroughly

explored and some time is spent trying to decide the datatype of the various features. After the data exploration, a chapter about the theory of binary classification follows. The theory is kept at a basic level and should only be used as a quick summary. After the theory comes the modelling chapters. Results obtained from the various experiments are presented and a basic confidence interval is calculated. All results are summarized in the last chapter of the modelling section. Following the modelling section is a chapter about project management and the software used for this project. Normally this chapter would have been an appendix, but since setting up a working environment is part of the problem statement a chapter is devoted to this subject. Finally the report closes with a discussion, conclusion and appendices.

Bibliography

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science + Business Media, LLC, 1st edition.