Detection of human alertness using supervised learning

Anders Hørsted

Kongens Lyngby 2011
IMM

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

This is the summary/abstract

Resumé

På dansk

Preface

This thesis was prepared at Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with different aspects of mathematical modeling of systems using data and partial knowledge about the structure of the systems. The main focus is on extensions of non-parametric methods, but also stochastic differential equations and neural networks are considered.

The thesis consists of a summary report and a collection of ten research papers written during the period 1996–1999, and elsewhere published.

Lyngby, December 1999

Henrik Aalborg Nielsen

Acknowledgements

I thank my...

Contents

1 Introduction **1**
- 1.1 The competition . . . 2
  - 1.1.1 Problemformulering . . . 2

2 The Ford Challenge **3**
- 2.1 Other online machine learning competitions . . . 3
  - 2.1.1 Netflix Prize . . . 4
  - 2.1.2 KDD Cup . . . 4
  - 2.1.3 And many others . . . 5
- 2.2 Kaggle.com and The Ford Challenge . . . 5
  - 2.2.1 The data set . . . 7

3 Data exploration **9**
- 3.1 A note about Test data, training data . . . 9
- 3.2 Calculating common statistics . . . 9
- 3.3 Determining the datatype of features . . . 10
  - 3.3.1 Unique values . . . 10
  - 3.3.2 Plotting some features . . . 10
- 3.4 Finding possible discriminating features . . . 10
  - 3.4.1 Testing binary features . . . 10
  - 3.4.2 Scatterplots . . . 10
- 3.5 Making a Principal Component Analysis . . . 10
  - 3.5.1 The math behind PCA . . . 10
- 3.6 Conclusions . . . 10

4 Recreating winning approach **11**

5 Improving the winning approach **13**

6 Other classification methods **15**

7 Theory of classification **17**
7.1 AUC 17

8 Workflow and tools **19**

9 Discussion **21**

10 Conclusion **23**

A Appendices **25**
A.1 Source code for calculating common statistics 26
A.2 Results of calculating common statistics 27

Chapter 1

Introduction

Igennem de sidste 30 år er computerens ydeevne vokset markant. Den forbedrede ydeevne har givet mulighed for at udføre statistisk dataanalyse, i et omfang der ikke tidligere har været muligt. En af de ting der er blevet mulighed for, er at programmere såkaldte "classifiers". Et godt eksempel på en classifier, er de programmer bankerne bruger til at opdage evt. snyd med kreditkort. Baseret på alle tidligere korttransaktioner, og viden om hvilke der var "falske" transaktioner, kan bankerne programmere en classifier, der med høj præcision kan forudsige om en ny transaktion er snyd eller ej.
Denne opgave tager udgangspunkt i en konkurrence på hjemmesiden kaggle.com. Konkurrencen er udbudt af Ford Motors og handler om at udvikle en classifier, der kan forudsige om en bilist er ved at blive ukoncentreret, mens han/hun kører bil. Til at udvikle denne classifier har Ford foretaget en række målinger på bilister, mens de kørte bil, og til hver måling er det blevet noteret om bilisten var opmærksom eller ej. Med udgangspunkt i disse målinger udarbejdes – ved brug af de mest gængse metoder – en række classifiers, og deres evne til at klassificere ny data sammenlignes.

1.1 The competition

1.1.1 Problemformulering

I denne opgave vil jeg...

- ... bruge de mest gængse klassifikationsmodeller (nearest neighbour, logistic regression, neural networks og SVM) til at lave en classifier der (forhåbentlig) kan forudsige om en bilist er ved at falde i søvn.
- ... undersøge hvor stor indflydelse den indledende databehandling (feature selection og outlier removal) har på det endelige resultat.
- ... undersøge om classifieren kan forbedres ved at implementere en Hidden Markov Model, der tager hensyn til det temporale aspekt af data.
- ... implementere en ensemble classifier, der kombinerer resultatet af flere classifiers i én classifier.

Chapter 2

The Ford Challenge

In this chapter, the ford competition that is the basis for this report, is introduced. Before introducing the ford challenge, a short overview of other online machine learning competitions is given. As part of introducing the ford competition, the kaggle.com website that hosted the ford competition is presented, and the data set used in the ford competition is described in detail. But as metioned the chapter starts with a short overview of other online machine learning competitions.

2.1 Other online machine learning competitions

To get a little perspective on the ford challenge, this section gives a short review of some past and present online machine learning competitions.

2.1.1 Netflix Prize

One of the most talked about competitions, may very well be the Netflix Prize. The Netflix competition was launched on October 2, 2006 and the aim of the competition was to predict how users would grade new movies, based on a large dataset of previous grades. Why did the Netflix Prize gather a lot of attention? First of all the grand prize was $1M, which is a lot of money. And secondly the dataset was huge, consisting of 100,480,507 ratings given by 480,189 users, leaving room for a lot of interesting new techniques to be tested.

When the Netflix Prize was awarded on September 18, 2009, 5169 different teams had submitted at least one entry to the competition. The winning team consisted of three different teams, that at one point decided to team-up and compete as a join-team.

Netflix originally wished to follow up the Netflix Prize, with another competition but decided to dismiss the idea, due to a lawsuit regarding privacy concerns related to the first Netflix Prize. [1]

2.1.2 KDD Cup

Although the Netflix Prize gather a lot attention, it wasn't the first online machine learning competition. An example of an earlier competition, is the KDD Cup. The KDD Cup is a competition that is held every year, and that started back in 1997. The subject changes every year, and can be anything from mining purchase data from an online store, to computer aided detection of breast cancer (the 2000 and 2008 competitions respectively).

This year the KDD Cup is held in coorporation with Yahoo! Labs and the task is to predict user ratings of musical items (both tracks, albums, artists and genres). One note worthy detail about the 2011 KDD Cup is the huge data set, containing over 300 million ratings of more than 600,000 distinct items. [2]

[1] The section about The Netflix Prize is based on Wikipedia (2011) and Netflix (2011)

[2] Read more about the KDD Cup history at ACM (2011). For more info about the 2011

2.1.3 And many others

The Netflix Prize and the KDD Cup are just two examples of online machine learning competitions. Many others exists, such as

- *The Hearst Challenge 2011* - Every year The Hearst Corporation hosts a machine learning competition. This year the task is to data mine the history of 1.8 million emails sent to subscribers of Hearst's publications, and then predict who will open emails in the future. Read more at `http://www.hearstchallenge.com`
- *The Reclab Prize* - RichRelevance is a company that specializes in online product recommendation. They offer a $1M prize for the team that first improves their product recommendation algorithm by 10%. Read more at `http://www.overstockreclabprize.com`
- *The Heritage Health Prize* -

Since almost all the machine learning competition websites, needs the same functionality, websites that specializes in hosting machine learning competitions has appeared. One example of a website that provides a hosting platform for machine learning competitions is the kaggle.com website, who hosts The Ford Challenge.

2.2 Kaggle.com and The Ford Challenge

As already mentioned, kaggle.com hosts machine learning competitions for universities and corporations. The first competition hosted by kaggle.com was started in April 2010, and since then a total of 18 competitions have been held. Every competition at kaggle.com has some background information, links to the data sets, a submission system and a forum, as seen in figure 2.1.

competition see Labs (2011)

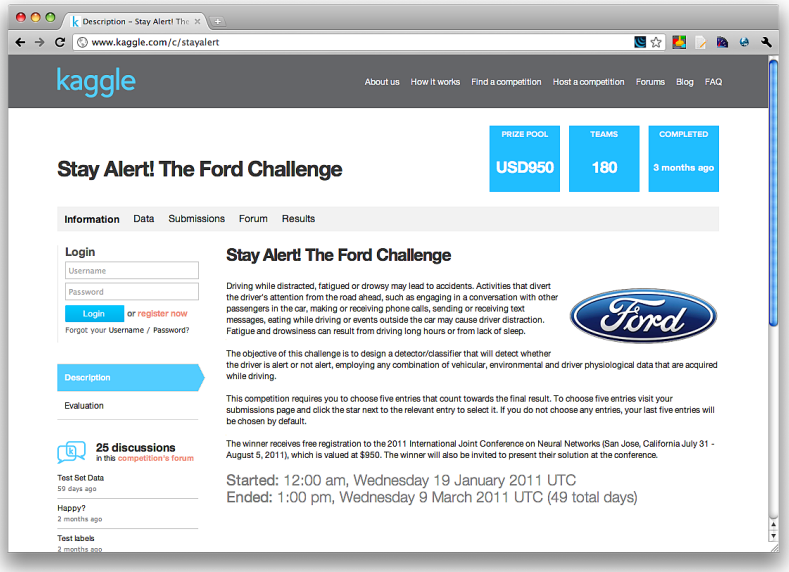

Figure 2.1: The Ford Challenge frontpage at the kaggle.com website

The Ford Challenge began on January 19, 2011. The task was to create a classifier that is able to detect when a driver is about to get distracted while driving. The dataset that Ford made available for the competition consisted of measurements of 30 different features, measured on drivers along with a binary feature (IsAlert) that was 1 if the driver was alert and 0 otherwise. The 30 features was a mix of environmental, driver physiological and vehicular features[3]. Based on this dataset a classifier should predict the IsAlert feature of a distinct test dataset held by Ford.

One detail that made the competition slightly different than many other competitions, was that Ford would not disclose any information about what the different features represented[4]. The official reason was that (Abou-Nasr, 2011b)

> "We like to encourage the participants to pursue classification without preconceived notions based on prior knowledge of the

[3]The dataset is explained in much more detail in the section 2.2.1

[4]see forum replies from the Ford spokesperson (Abou-Nasr, 2011a,c)

> subject, focusing on variables which lead them (based on their own experiments) to better classification."

Doubts about the true motive behind the lack of details about the features, was expressed by, what later turned out to be, the winner of the competition (Inference, 2011)

The performance of the classifiers was measured by calculating the AUC (see section 7.1) of the classifiers, on the test set. A limit of two submissions per contestant was set as a way to counteract the possibility of someone reverse-engineering the IsAlert-feature of the test dataset. This could of cause be circumvented by one person registering more than once. And this was exactly what happened. On March 9, 2011 when the competition ended, two users had achieved exactly the same AUC (six significant digits) and other contestants immediately questioned the probability of two unrelated users getting exactly the same AUC (Pardos, 2011). One of the two leaders (Rosanne/Shen) quickly admitted that he and his friend had indeed used two accounts to get 4 submission per day (Rosanne, 2011), and after some discussion in the forum, the leaders were disqualified. The end result was that the user Inference in third position was declared the official winner of the competition.

Two weeks after the competition ended, Inference described the technique used to win the competition. This will be described in detail in chapter 4. But before that chapter, the data set will be described in details in the next section.

2.2.1 The data set

The dataset used to create a classfier was released on the competition website, the day the competition started. The dataset consisted of a number of trials and each trial was approximately 2 minutes of sequential data recorded every 100ms during a driving session on the road or in a driving simulator (Kaggle.com, 2011). As the interval between two *rows* was 100ms, each trial consists of approximately $2\,\text{minutes} \cdot 60\,\frac{\text{secs}}{\text{minute}} \cdot 10\,\frac{\text{rows}}{\text{sec}} = 1200$ rows.

TrialID	ObsNum	IsAlert	P1	...	P8	E1	...	E11	V1	...	V11
0	0	0	12.2	...	1.2	4.3	...	33	12	...	7.34
⋮											⋮
0	1200	1	11.1	...	10.7	1.3	...	21	8	...	8.82
⋮											⋮
510	1198	0	11.1	...	10.7	1.3	...	21	8	...	8.82

Table 2.1: Structure of the data set

Each row has a total of 33 data columns structured as shown in table 2.1. Some details are worth noticing:

- The TrialID starts at 0 and the trials from 469 to 479 (both inclusive) are missing. The last trial has TrialID=510. This gives a total of exactly 500 trials.
- The ObsNum also starts at 0 there are not exactly 1200 observation for every trial.
- The total number of rows is 604,229
- The row number is not part of the data set, so a row is uniquely identified by the pair (TrialID, ObsNum).

As mentioned before no additional information about what the different features represent or what datatype (discrete, continous) they are, was disclosed by Ford. The only way to get these informations is by doing a thorough data exploration of the data set and that is what the next chapter is about.

CHAPTER 3

Data exploration

Here I describe the various data exploration techniques I have used. Lots of nice graphs. PCA. Boxplots. Feature plots. Scatter Plots. Unique Values. Is it discrete, binary, continuous.

3.1 A note about Test data, training data

3.2 Calculating common statistics

To start out the data exploration four common statistics, namely the mean, min, max and standard deviation, of every feature across the whole dataset was calculated. The source code for the calculations can be found in appendix A.1 and all results in appendix A.2

Mean, min, max, std for whole data set. Conclusions: P8, V7, V9 are zero throughout the data set, V5, E9 could be binary. Drivers only alert a little over half the time...

3.3 Determining the datatype of features

3.3.1 Unique values

3.3.2 Plotting some features

3.4 Finding possible discriminating features

3.4.1 Testing binary features

3.4.2 Scatterplots

3.5 Making a Principal Component Analysis

3.5.1 The math behind PCA

3.6 Conclusions

Chapter 4

Recreating winning approach

Here I describe how I have tried to recreate the winning approach. How I measure performance. The problem that I do not have access to the test data set used by Inference. My results. The scikits.learn library that I have used.

Chapter 5

Improving the winning approach

Here I try to improve the winning approach, by doing my own feature selection. Forward selection. Lasso. Cross validating with Lasso. Using window instead of running.

Chapter 6

Other classification methods

Here I describe some alternatives to the logistic regression used by Inference. Hoping to get a result or two from SVM or Neural Network.

Chapter 7

Theory of classification

Here I describe some general results about classification. Bayes error rate. No free lunch theorem. Ugly duckling theorem.

7.1 AUC

Chapter 8

Workflow and tools

Chapter 9

Discussion

Bla, bla, bla.

Chapter 10

Conclusion

More bla, bla, bla.

Chapter A

Appendices

A little introduction and then a new page

A.1 Source code for calculating common statistics

```
from json import dump

from src.data_interface import trd, L
from src.utils import get_path


sess_root = get_path(__file__) + '/../scr'

summary_statistics = ['min', 'max', 'mean', 'std']
features_to_calculate = L[2:]
calculations = {}

for feature_name in features_to_calculate:
    calculations[feature_name] = {}
    for statistic in summary_statistics:
        stat_function = getattr(trd.get_feature(feature_name), statistic)
        calculations[feature_name][statistic] = stat_function()

f = open(sess_root + '/summary_statistics.json', 'w')
dump(calculations, f, indent=4)
f.close()
```

A.2 Results of calculating common statistics

Feature	Mean	Min	Max	Std.Dev
E11	1.37	0.00	52.40	5.37
E10	63.38	0.00	127.00	19.23
IsAlert	0.58	0.00	1.00	0.49
V1	75.58	0.00	129.70	44.94
V2	-0.04	-4.79	3.99	0.42
V3	569.78	240.00	1023.00	299.02
E9	0.87	0.00	1.00	0.33
E8	1.39	0.00	9.00	1.62
V6	1699.34	0.00	4892.00	626.27
V7	0.00	0.00	0.00	0.00
E5	0.02	0.01	0.02	0.00
E4	-3.99	-250.00	260.00	34.80
E7	1.81	0.00	25.00	2.95
E6	358.50	260.00	513.00	27.84
E1	10.54	0.00	243.99	14.00
E3	0.30	0.00	4.00	1.03
E2	105.05	0.00	360.00	128.33
P2	12.01	-45.63	71.17	3.74
P3	1028.00	504.00	2512.00	309.70
P1	35.45	-22.48	101.35	7.36
P6	843.73	128.00	228812.00	2795.32
P7	78.40	0.26	468.75	18.79
P4	63.99	23.89	119.05	19.76
P5	0.18	0.04	27.20	0.39
P8	0.00	0.00	0.00	0.00
V4	21.24	0.00	484.49	67.21
V8	12.46	0.00	82.10	11.54
V9	0.00	0.00	0.00	0.00
V5	0.18	0.00	1.00	0.38
V10	3.28	1.00	7.00	1.27
V11	11.59	1.68	262.45	8.43

Bibliography

Mahmoud Abou-Nasr. Kaggle forum: Discrete or continuous values - Reply 2, 2011a. URL http://www.kaggle.com/c/stayalert/forums/t/266/discrete-or-continuous-values/1635#post1635.

Mahmoud Abou-Nasr. Kaggle forum: Were units changed? - Reply 2, 06 2011b. URL http://www.kaggle.com/c/stayalert/forums/t/268/were-units-changed/1641#post1641.

Mahmoud Abou-Nasr. Kaggle forum: About the parameter - Reply 3, 2011c. URL http://www.kaggle.com/c/stayalert/forums/t/317/about-the-parameter/1861#post1861.

ACM. KDD Cup Center, 06 2011. URL http://www.sigkdd.org/kddcup/index.php.

Inference. Kaggle forum: Ford - Reply 3, 06 2011. URL http://www.kaggle.com/c/stayalert/forums/t/295/ford/1743#post1743.

Kaggle.com. The Ford Challenge Data Files, 2011. URL http://www.kaggle.com/c/stayalert/Data.

Yahoo! Labs. KDD Cup 2011, 06 2011. URL http://kddcup.yahoo.com/.

Netflix. Netflix Prize Leaderboard, 06 2011. URL http://www.netflixprize.com/leaderboard.

Zach Pardos. Kaggle forum: Top two teams with same AUC - Reply 1, 06 2011. URL http://www.kaggle.com/c/stayalert/forums/t/327/top-two-teams-with-same-auc/1937#post1937.

Rosanne. Kaggle forum: Top two teams with same AUC - Reply 4, 06 2011. URL http://www.kaggle.com/c/stayalert/forums/t/327/top-two-teams-with-same-auc/1957#post1957.

Wikipedia. Netflix Prize, 06 2011. URL http://en.wikipedia.org/wiki/Netflix_Prize.