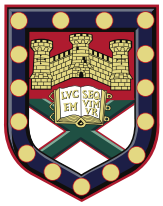

University of Exeter

Can statistics help us to understand deep learning?

Johannes Smit

May 2019

Abstract

All theses/dissertations must include an abstract of approximately 300 words bound in with each copy and placed so as to follow the title page.

Acknowledgements

Hi mum.

Contents

Chapter 1

Introduction

1.1 What is machine learning?

To-do: *Machine learning is very complex to human understanding Machine learning is widely used in many high stakes scenarios — give examples It is important to be able to look inside the 'black box' of machine learning — explain how this would be useful in each of the given examples We can use statistical methods to try and recover the information in a ML algorithm — give some idea how*

Chapter 2

Deep Learning

2.1 Machine Learning

2.1.1 Neural Networks

The Neuron

2.1.2 Backpropagation

> **To-do:** *Section on keras/tensorflow?*

2.2 Example

To demonstrate the possibility of using statistical methods to understand the process of deep learning, we use a simple function $f(x) = x + 5\sin(x)$. 256 evenly spread datapoints were taken from this function, and noise following $\epsilon \sim \mathcal{N}(0, 0.1)$ was applied.

> **To-do:** *Possibly also use a more complex example?*

2.2.1 Training the neural network

This function was learnt by a neural network with 8 layers, each with 10 neurons and the $\tanh(\cdot)$ activation function, except for the final layer which used a linear activation function. The result of this learning is seen in Figure 2.1.

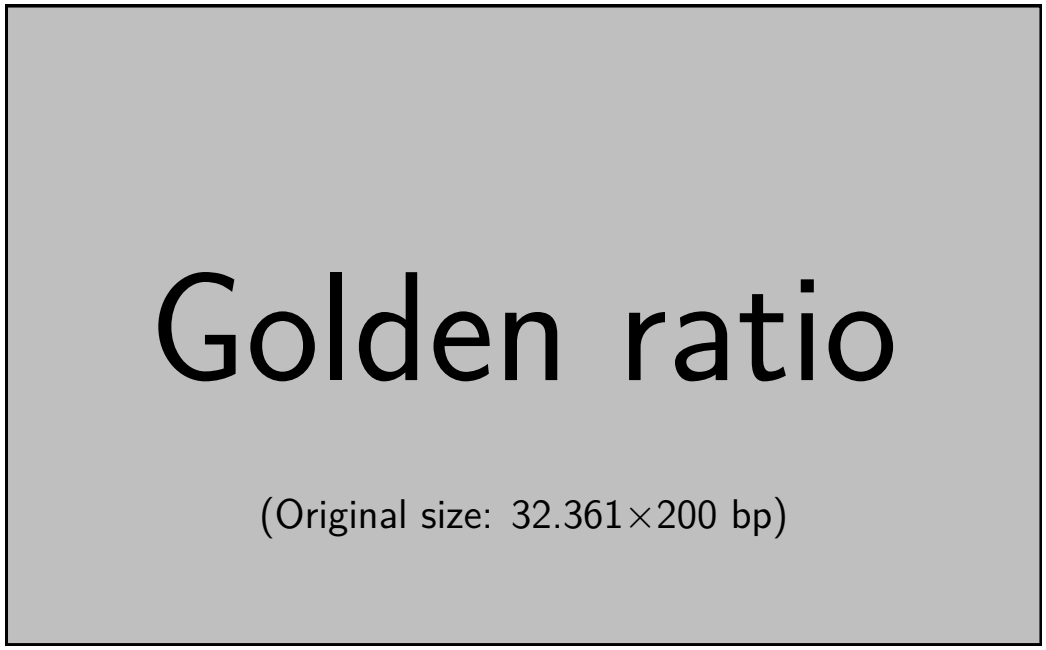

Figure 2.1: The output of the neural network.

To-do: *Information on choosing the right size NN*

2.2.2 Ordering of the datapoints

Due to the form of stochastic gradient descent used to train the model, if the order of the datapoints is not randomised, then the optimisation algorithm can more easily get stuck in a local minimum. An example of this is seen in Figure 2.2, where the same neural network has been trained on the example data.

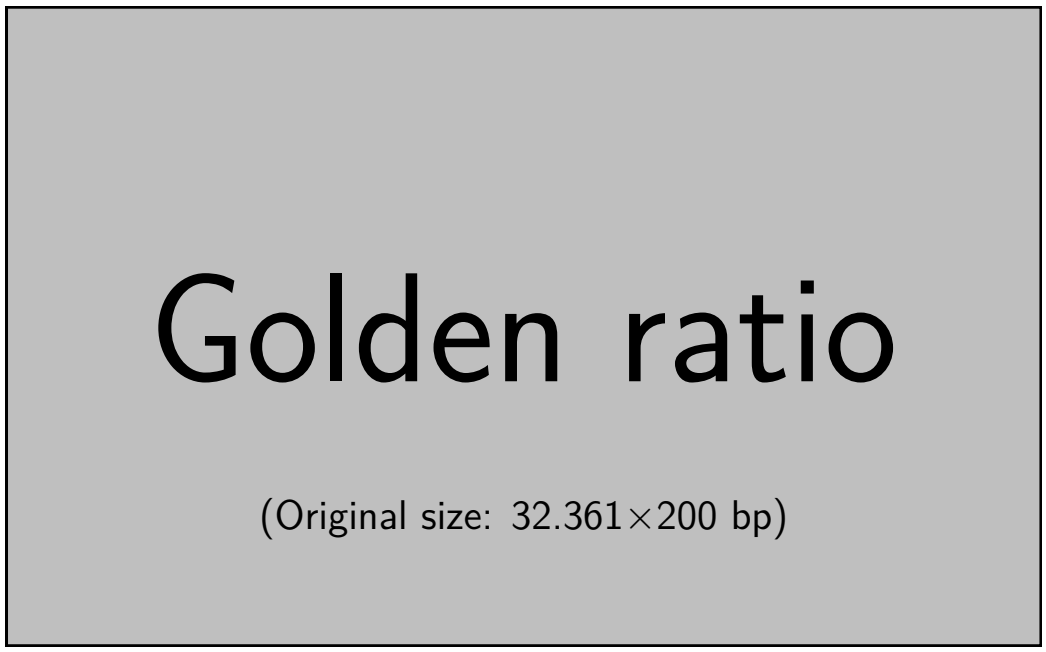

Figure 2.2: The output of the same neural network trained on the same data but ordered differently.

Chapter 3

Opening the Black Box

3.1 Previous papers on this subject

3.2 Regression

One method of opening the black box of machine learning is to use regression with many different predictors.

3.2.1 Linear regression

In this case, x, x^2, $\sin(x)$, $\sin(2x)$, $\sin(x/2)$, $\cos(x)$, $\cos(2x)$ and $\cos(x/2)$ were used. This produced a complex model with many coefficients close to zero.

To-do: *table of linear regression fit*

3.2.2 Stepwise regression

It is then possible to use stepwise regression to reduce the number of parameters in this linear model.

To-do: *table of stepwise regression*

While the relevant estimators x and $\sin(x)$ were identified and their coefficients fairly accurately estimated, a few other variables were also identified as significant. This method is only likely to work with a perfect or near perfect fit with no noise, which is unrealistic for real applications.

3.2.3 LASSO

Alternatively, we can use the Least Absolute Shrinkage and Selection Operator(LASSO) method to select the significant variables from the full model. The LASSO method constrains the sum of the absolute values of the model parameters, regularising the least influential parameters to zero. We vary the hyperparameter λ to change how regularised the coefficients are, as seen in Figure 3.1. We then use k-fold cross validation to find the optimal hyperparameter λ.

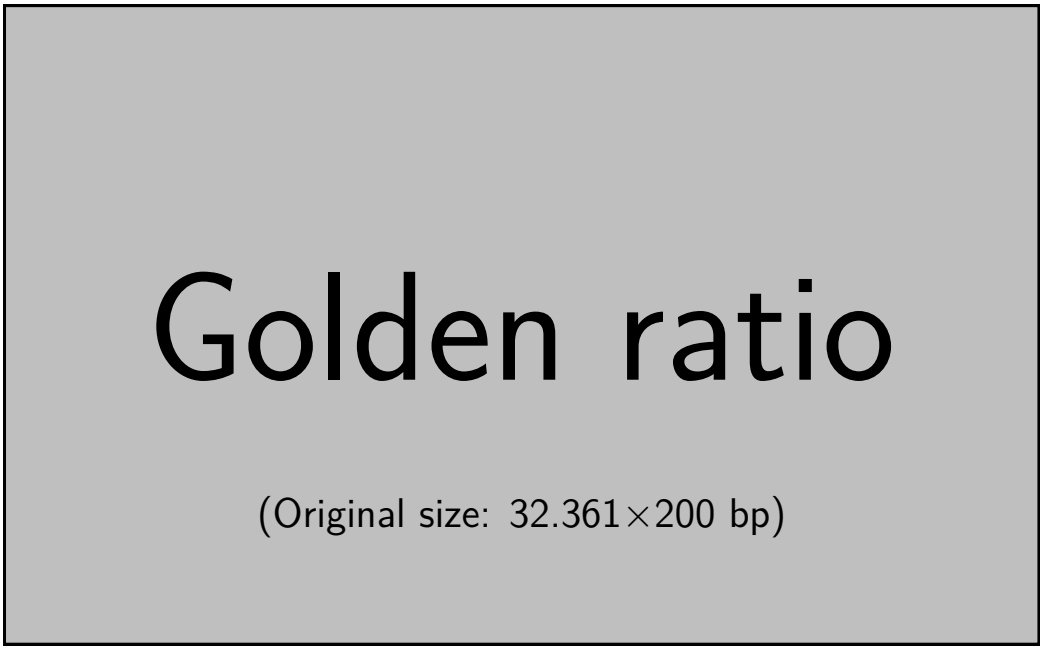

Figure 3.1: Changing the hyperparameter λ.

To-do: *table of LASSO coefs*

3.3 Gaussian process regression

We can use Gaussian Processes (GPs) to try and model the output of the ANN. GPs are an extension of the multivariate normal distribution to an infinite dimensional process with a mean function and covariance function instead of a mean vector and covariance matrix. Because GPs are very flexible, applying a GP to the output of the ANN is likely to result in a close fit.

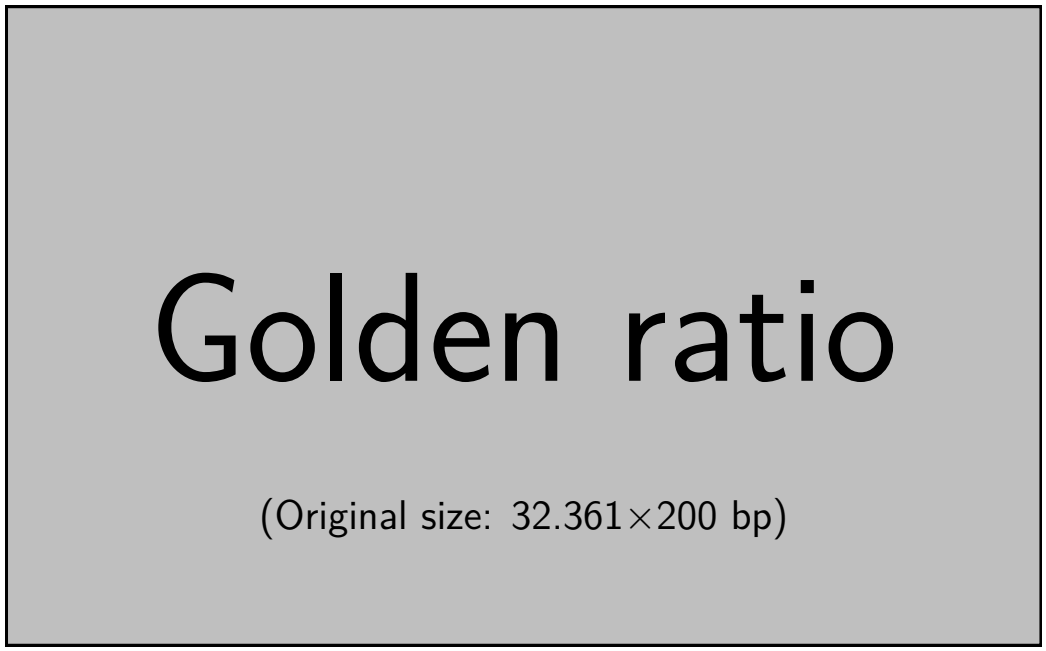

Figure 3.2: The fit of the GP.

Chapter 4

Conclusion

4.1 How well have each of the attempts worked?

4.2 What could be improved?

4.3 How useful would more research on this topic be?

4.4 What should future research on this topic focus on?

Appendix A

Code

Appendix here.