Fairness and the Bias-Variance Trade Off

Claire I. Coffey
Clare Hall

A dissertation submitted to the University of Cambridge in partial fulfilment of the requirements for the degree of Master of Philosophy in Advanced Computer Science

University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom

Email: cic31@cam.ac.uk

May 25, 2020

Declaration

I, Claire I. Coffey of Clare Hall, being a candidate for the M.Phil in Advanced Computer Science, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

Total word count: n/a

Signed:

Date:

This dissertation is copyright ©2020 Claire I. Coffey.
All trademarks used in this dissertation are hereby acknowledged.

Abstract

In machine learning generalisation, bias errors arise due to over-simplification and variance errors arise due to limits on available training data. This project considers the implications of these error types in fairness. Are models that exhibit bias errors prone to introducing new categories of unfair discrimination? The project will consider different fairness criteria alongside bias errors and variance errors in the context of recidivism data.

Acknowledgements?

Keywords?

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.0.1 Overview

Bias-Variance Trade Off

Fairness 2 ways in which fairness can be explored: pre and post processing i.e. in data or in model. we focus on latter. Counterfactual fairness, Equalised odds, intersectional fairness (multicalibration), context-aware fairness (positionality?).

Recidivism

Novelty

Description of report

The bias-variance trade-off is a well documented problem in statistics and machine learning [1–5]. It refers to the simultaneous minimisation of two conflicting error types which prevent a model from generalising well: the bias error and the variance error. Bias errors arise due to over-simplification of circumstances, whereas variance errors arise due to limits on available

training data. This project considers the implications of these error types in fairness.

The growth in ubiquity of machine learning algorithms in society has inspired in a new wave of research on fairness [6–8]. This research is essential in order to minimise the potential discrimination of these algorithms. A recent application of machine learning in society is in recidivism, predicting the likelihood that a defendant will re-offend. These algorithms are commonplace and life-changing, as they are used to help determine pre-trial, parole and sentence decisions. It is therefore critical these algorithms are as fair as possible. However, fairness is difficult to define. This has been analysed with respect to recidivism in [9], following the dissection of the widely used recidivism software, Northpointe's COMPAS, which was argued to be unfairly discriminating against black defendants [10, 11]. This finding was disputed by Northpointe [12] due to "differing interpretations of fairness."

In this project I plan to explore how bias errors and variance errors impact fairness, specifically applied to recidivism data. I will look at these two errors both separately and in context of the bias-variance trade-off. Ultimately, the goal of this project is answer the question: *are models that exhibit bias errors prone to introducing new categories of unfair discrimination?* The motivation for this question is the idea that many algorithms favour consistent decision making, which tends to introduce the over-simplification of circumstances, introducing bias errors. This in turn suggests that nuances in the environment may be missed, and certain groups may therefore be discriminated against.

As surveyed in [7] and mentioned above, fairness can be defined in multiple ways, which leads to differing algorithmic approaches. Therefore, I plan to review a variety of prominent definitions and criteria and explore how the bias errors and variance errors in relation to these criteria may differ. Furthermore, I plan to look at the bias-variance trade-off for a variety of supervised learning models. I will utilise existing implementations of these from Python

machine learning libraries, which can be easily manipulated, for example Scikit-learn [13]. I will apply these to the recidivism data used in [11], which is publicly accessible online through the US government, but also through the authors of previous analysis of the data, as in [14]. I will investigate potential patterns and whether a high bias error may behave differently for different models in terms of introducing new categories of discrimination.

What Bias/variance might mean in this context - i.e. varied decision making, what happens if bias or var error is too high - neil's blog post

1.0.2 Report Outline

outline of chapters and what they include

Chapter 2

Background

2.0.1 Bias-Variance Trade Off

Reference Bias/variance decomposition for classification? - talk about noise? How applies to diff models? Bootstrapping.

2.0.2 Fairness

Fairness philosophically? Fairness in machine learning. Multi objective optimisation in fairness to optimise these? how problem can be formulated in this way - as weighted average?

2.0.3 Recidivism

discussion of relevance and controversy in compas etc.

Chapter 3

Related Work

3.0.1 Fairness

discussion including definitions, strengths, weaknesses of each.

Equalised Odds

Why chosen equalised odds.

Chapter 4

Design and Implementation

4.0.1 Data

Preprocessing and Cleaning

We import the data into a pandas DataFrame. We begin by cleaning the data, so the crime descriptions are simplified, removing duplicate categories. For example, we merge descriptions such as 'possession of cocaine' and 'possess cocaine', or 'burglary/weapon' and 'burglary and weapon', by removing prepositions, and replacing abreviations and similies.

Then, the categorical data is split into different fields for each category, and encoded as 0 or 1. For example, an individual with characteristic "sex: male" would be encoded as "male: 1, female: 0". The sex category is then removed.

We then consider which fields to use for prediction. This includes the removal of any fields/columns which contain many NaN values, since these cannot be handled by the classifiers. We choose to remove the columns with many NaNs rather than using an alternative approach such as replacing them with the average so as not to introduce other types of bias. We also then remove rows/individuals containing any further NaN values so there is no longer any NaN values present in the data.

We then normalise all of the data in the dataframe, so that when fed into

the classifier, the predicitons are not skewed (and potentially different forms of bias introduced). We do this by using the StandardScaler in the sklearn preprocessing library, and we normalise the data to have a variance of 1.

Finally, we define the number of testing/training samples desired and split the data into these two sets appropriately.

Data Bias

discuss the existence of this, i.e. systematic bias in criminal justice system and society. although not main focus of our thing. refer to papers about this i.e. context aware fairness etc.

4.0.2 Models / Classification

RBF SVM with diff parameters - why ? -Procedure for testing diff parameters -Model optimisation. we want 2 models to compare which satisfy the following: - (nearly the) same overall error - 1 high bias, 1 high variance - satisfies equalised odds: calculate TP and FP for each demographic/protected characteristic. and minimise the difference between these different demographics (equally predictive across protected characteristics).

4.0.3 Fairness Correction

Equalised odds. Optimising for different demographics by minimising diff between TP and FP across diff demographics.

4.0.4 Bootstrapping

Chapter 5

Evaluation

Chapter 6

Discussion and Further Work

Chapter 7

Summary and Conclusions

Bibliography

[1] Robert Tibshirani. *Bias, variance and prediction error for classification rules*. Citeseer, 1996.

[2] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[3] Pedro Domingos. A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238, 2000.

[4] Alexandros Agapitos, Anthony Brabazon, and Michael O'Neill. Controlling overfitting in symbolic regression based on a bias/variance error decomposition. In *International Conference on Parallel Problem Solving from Nature*, pages 438–447. Springer, 2012.

[5] Giorgio Valentini and Thomas G Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *Journal of Machine Learning Research*, 5(Jul):725–775, 2004.

[6] Niels Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services*, 36(1):15–30, 2018.

[7] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[8] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[9] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[10] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[11] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

[12] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[14] Broward Country Jail Data. Propublica compas analysis. https://github.com/propublica/compas-analysis, 2016.