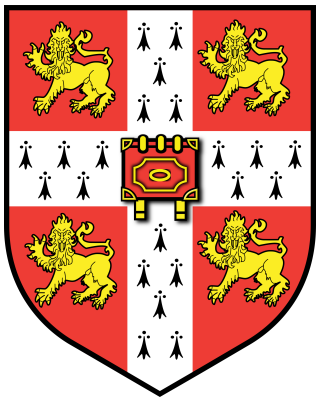
Fairness and the Bias-Variance Trade-Off

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 28, 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:

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 popular fairness criteria alongside bias errors and variance errors in the context of recidivism data.

Acknowledgements?

Keywords *fairness, fair machine learning, bias-variance trade-off*

Contents

List of Figures

List of Tables

1. Introduction

1.1 Overview

The bias-variance trade-off is a central consideration in supervised machine learning. 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, missing nuances in the data, whereas variance errors arise due to limits on available training data, modelling the training data too closely. The bias-variance trade-off has been widely explored in relation to automated decision making [**?**, 1–3], including how predictions vary depending on the trade-off between these error types. Little explored however is the relationship between bias errors, variance errors, and fairness. This work aims to delve into the implications of these error types in 'fair' machine learning (fair-ML): a growing critical area of research.

The use of machine learning is becoming ever more prevalent in society, increasingly influencing life-changing decisions in domains such as criminal justice, employment, and education. Historically, many of these domains were infiltrated with cognitive biases and discriminatory decisions. The use of machine learning in these areas runs the risk of perpetuating this discrimination if not properly understood. This concern has sparked an important new wave of research on fair-ML [4–7], which aims to ensure that these algo-

rithms are as non-discriminatory as possible. There are two major categories of approaches to addressing this: pre-processing - ensuring the data used to train the model is free from biases; and post-processing - ensuring the model does not introduce further discrimination or perpetuate biases in the data. ***maybe this (italicised) bit doesn't go in intro?*** *Minimising discrimination in data collection and the pre-processing phase is undeniably important, but very difficult to approach, given that many unconscious cognitive biases run deep within society and may present themselves in data in an inconspicuous manner. Although there is growing interdisciplinary research exploring the impact of societal structures on datasets [8], and the importance of context-aware fair-ML has been raised [9], addressing these inherent biases is truly only possible through societal reform.* Therefore, like much of the fair-ML research, we will be exploring fairness in relation to the latter approach - post-processing - since this is the area in which model choice and optimisation has the greatest opportunity to impact fairness (or lack thereof).

The fair-ML community is exploding with definitions of what it means for a model to be non-discriminatory. Some of the most popular definitions of fairness include *demographic parity, statistical parity, counterfactual fairness, equalised odds*, and *calibration*, some of which are discussed further in section 3. In our work, we choose to use equalised odds as our fairness criteria; a popular and reasonable metric for analysing fairness in binary classification. We use this equalised odds criteria to minimise the discrimination of our models with respect to protected characteristics, before exploring the bias-variance trade off for said 'fair' models - evaluating the difference in the behaviour of models with a proportionally high bias error or proportionally high variance error. Henceforth, we introduce the idea that 'fair' models with high bias errors may introduce new categories of discrimination, and therefore why it may be valuable to favour variance errors in fair machine learning models.

This exploration will be done in the criminal justice domain, a life-changing area in which the use of machine learning has been criticised (predominantly due to concerns of unfair discrimination [10]). In particular, we will focus on recidivism data, which has been used for predicting the likelihood of criminals to re-offend in systems such as COMPAS. These systems have been used in pre-trial, parole, and sentence decisions; it is therefore imperative that the models used are optimised for non-discrimination. We will explore the bias, variance, and fairness relationship by: first designing simple classifiers to predict a binary recidivism classification; then correcting these models for fairness using protected characteristics; and finally interrogating predictions made in relation to the bias and variance errors of the models, to determine the varying treatment of individuals with particular characteristics in models exhibiting different error types, finding potential new categories of discrimination.

1.2 Report Outline

outline of chapters and what they include

2. Background

We will now introduce the relevant background and definitions used in our exploration, and refer to these throughout the remainder of the report.

2.0.1 Bias and Variance

Bias-variance decomposition for zero-one loss

First explored for squared loss in regression models by Efron [11], the bias-variance decomposition allows a model's error to be decomposed into the statistical bias and variance. Tibshirani introduced the bias and variance decomposition for classification rules [1], providing a general decomposition for any prediction error measure and using bootstrapping (see 2.0.2) to derive a sample-based estimate of the prediction error decomposition. This prediction error decomposition is also explored by Dietterich, focusing on zero-one loss (misclassification error) in [12], as applied to a bootstrap classifier. The zero-one loss is given as a probability of misclassification, $P_S(x)$ for each example x. Given a training set S of size n, a classifier C which outputs a prediction class Y_x, and the true class of x T_x, the zero-one loss is:

$$P_S(x) = \begin{cases} 1 & \text{if } Y_x \neq T_x \\ 0 & \text{if } Y_x = T_x \end{cases}$$

So the probability is 1 if C misclassifies x and 0 otherwise.

We can then draw a set of i bootstrap training sets consisting of training sets S_i, each of size n. We can apply C to each x to construct a set of predictions for each x, Y_x. From this, we can find the **average** misclassification probability for each x, $\hat{P}(C, n, x)$ as:

$$\hat{P}(C, n, x) = \frac{1}{i} \sum_{i=1}^{i=i} P_{Si}(x)$$

In other words, this is the expected error rate of C for any x:

$$Error(C, n, x) = \hat{P}(C, n, x)$$

In this paper, since we are using the zero-one loss as our error function, we can define this as:

$$L_{ZO}(C, n, x) = Error(C, n, x)$$

Therefore:

$$L_{ZO}(C, n, x) = \hat{P}(C, n, x)$$

Further, we can define the bias and variance as follows. For any point x, if $L_{ZO}(C, n, x) > 0.5$, on average, x is misclassified. Hence, x is a systematic error, or error due to *bias*. Therefore, the bias of a classifier C, trained on training sets of size n can be defined as:

$$Bias(C, n, x) = \begin{cases} 0 & \text{if } \hat{P}(C, n, x) \leq 0.5 \\ 1 & \text{if } \hat{P}(C, n, x) > 0.5 \end{cases}$$

Then, the variance of C at x can be defined as the difference between the error rate and the bias, so variance is the error rate relative to the bias:

$$Variance(C, n, x) = \begin{cases} \hat{P}(C, n, x) & \text{if } \hat{P}(C, n, x) \leq 0.5 \\ 1 - \hat{P}(C, n, x) & \text{if } \hat{P}(C, n, x) > 0.5 \end{cases}$$

These definitions allow the bias and variance for any C to be measured at any x, for example, by using bootstrapping. Finally, the bias and variance

for a classifier can be calculated by simply averaging the bias and variance values over all x (as according to the decomposition by Domingos [2]).

[1, 2, 12]

Bias-Variance Tradeoff

say something about what the tradeoff is?

2.0.2 Bootstrapping

Bootstrapping

Bootstrapping [13] is a sampling with replacement procedure introduced by Efron and Tibshirani. It works in the following way: let S be a training dataset of size n. A bootstrap replicate training set, S_i, can be created by randomly drawing n samples from S with replacement. Therefore, at each turn, each example x_i will have a probability of $1/n$ of being selected. This results in some examples appearing multiple times in the new bootstrap training set S_i, and some not at all. This procedure is repeated many times in order to create a set of bootstrap training sets Z. Then, a classification for each example x_i in the testing set can be found by training a classifier C on each S_i.

Estimating bias and variance with bootstrapping

Let Y_x be the set of classifications found by C for all Z for each x. Y_x can then be used to estimate the bias and variance errors for each x_i, as above in **??**. The bias and variance for all x_i can then be averaged in order to determine the overall bias and variance errors of the model.

put eqn here?

Bagging

In order to determine an overall classification for each x, a voting procedure is carried out: the class with the highest probability of prediction across all Z is chosen as the final classification for that x_i. In the binary case under zero-one loss, this is simply the modal classification. Breiman defined this procedure as *Bagging* [14] (bootstrap aggregating); this procedure also reduces the zero-one loss of the models (hence reducing bias and variance).

2.0.3 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?

2.0.4 Equalised odds

Equalised odds [15], introduced by Hardt et al., is a popular fairness criteria that ensures non-discrimination against a specific protected characteristic in the following way: let $\hat{Y}$ be a classifier, A be a protected characteristic and Y be a classification. $\hat{Y}$ satisfies equalised odds with respect to A if $\hat{Y}$ and A are independent conditional on Y, ensuring that the classifier will predict Y at the same rate regardless of the value of A. This allows $\hat{Y}$ to depend on A but only through the target Y, encouraging the use of features that relate to Y directly rather than through A. In the binary case:

$$P\{\hat{Y} = 1|A = 0, Y = y\} = P\{\hat{Y} = 1|A = 1, Y = y\}, y \in 0, 1$$

For $y = 1$, this constraint requires $\hat{Y}$ to have equal *true positive rates (TPR)* across the two demographics $A = 0$ and $A = 1$. For $y = 0$, it requires equal *false positive rates (FPR)* across the two demographics $A = 0$ and $A = 1$. Therefore, equal accuracy is enforced across all demographics. This definition of fairness however does not allow for many different protected characteristics to be simultaneously enforced. This fact, and the nature of real-world datasets, means that it may be impossible to completely satisfy

the criteria perfectly. Instead of ensuring exact equality of TPR and FPR across demographics, the level of discrimination with respect to equalised odds can be measured.

elaborate on the above - define the discrimination measure - minimise diff in FPR and TPR for demographics. by minimising zero-one loss? - need to find good/optimal predictor - minimising loss

2.0.5 Recidivism

discussion of relevance and controversy in compas etc.

3. Related Work

3.0.1 Fairness

discussion including definitions, strengths, weaknesses of each.

Equalised Odds

Why chosen equalised odds.

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: - Minimsise loss (zero-one misclassification loss) (loss function) - (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). - Fit the model on the training data (which is one bootstrap data sample as defined above) - Perform classification for each bootstrap sample separately, and store these in a DataFrame, to be passed into the bias/variance calculations.

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

The classification process then uses a bootstrapping procedure with the chosen model, to generate predictions of recidivism classifications (1 = will not reoffend (positive case); 0 = will reoffend (negative case)).

Bootstrapping [16] is a sampling with replacement procedure. Here, the sample size is the same as the size of the (training) dataset. The bootstrapping procedure is run many times to generate different training datasets, which will then be used for classification. In turn, the classification results will be used to calculate and study the bias and variance errors.

5. Evaluation

6. Discussion and Further Work

7. Summary and Conclusions

Bibliography

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

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

[3] 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.

[4] 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.

[5] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

[6] Luca Oneto and Silvia Chiappa. Fairness in machine learning. In *Recent Trends in Learning From Data*, pages 155–196. Springer, 2020.

[7] Elena Beretta, Antonio Santangelo, Bruno Lepri, Antonio Vetrò, and Juan Carlos De Martin. The invisible power of fairness. how machine learning shapes democracy. In *Canadian Conference on Artificial Intelligence*, pages 238–250. Springer, 2019.

[8] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 167–176, 2020.

[9] Michelle Seng Ah Lee. Context-conscious fairness in using machine learning to make decisions. *AI Matters*, 5(2):23–29, 2019.

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

[11] Bradley Efron. Regression and anova with zero-one data: Measures of residual variation. *Journal of the American Statistical Association*, 73(361):113–121, 1978.

[12] Thomas G Dietterich and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Technical report, Department of Computer Science, Oregon State University, 1995.

[13] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[14] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[16] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.