Mitigating Bias in Machine Learning

Dr Leena Murgai

June 30, 2021

Contents

List of Figures

List of Tables

Notation and conventions

Mathematical notation

- $\mathbb{P}(A)$ denotes probability of event A
- $\mathbb{E}$ denotes expectation
- $\forall$ means for all
- $|$ means such that
- $\in$ means a member of
- $\Rightarrow$ means implies
- $\Leftrightarrow$ means if and only if
- square brackets are inclusive, round brackets are not, for example,

$$x \in [a, b) \quad \Leftrightarrow \quad a \leq x < b$$

Typographical conventions

In this book we mostly follow mathematical typographical conventions for variables. All variables are in italic. We use:

- lowercase letters for scalar variables, e.g. a,
- uppercase letters for random variables, e.g. X.
- lowercase bold typeface letters for (column) vectors, e.g. $\boldsymbol{y}$,
- uppercase bold typeface letters for matrices and vectors of random variables, e.g. $\boldsymbol{X}$.

Notice we have overloaded our notation slightly for matrices and random variables. If it is not clear from the context which we mean, it will be stated explicitly.

We may also use tensor notation where convenient. For the elements of a matrix $\boldsymbol{X}$ we use x_{ij} where i refers to the row and j to the column. Similarly for the elements of a vector $\boldsymbol{y}$, we use y_i where i refers to the row. For the transpose we use T in the superscript so that $\boldsymbol{y}^T$ would be a row vector. We also use $\boldsymbol{x}_i$ to denote the ith row of the matrix $\boldsymbol{X}$.

Data

We will use $\boldsymbol{X}$ to denote the (non-sensitive) feature matrix and $\boldsymbol{Z}$ to denote the sensitive feature matrix (features like gender and race for example). We use $\boldsymbol{y}$ to denote the target variable vector (a column vector with n elements, each corresponding to an sample) and $\hat{\boldsymbol{y}}(\boldsymbol{X}, \boldsymbol{Z})$ to denote the predicted target variable output by our model, which is a function of the features. We shall use $\mathcal{X}$ and $\mathcal{Z}$ to denote the set of possible values our feature vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ can take respectively and $\mathcal{Y}$ to denote the set of all possible values our

outcome y can take. When more appropriate we will use set notation to denote our set of data points which we write as $(\boldsymbol{X}, \boldsymbol{Z}, Y) = \{\boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}_{i=1}^n$

We shall use the same notation for our target and model output for both discrete (classification) and continuous (regression) variables. In the case where the target variable is discrete and derived from a continuous classifier (that is, one where we find the classification by applying a threshold to an underlying score), we denote the underlying score as $\boldsymbol{p}(\boldsymbol{X}, \boldsymbol{Z})$ (if y is a binary variable) in which case we can write,

$$\hat{y}_i(\boldsymbol{X}, \boldsymbol{Z}) = H\left(p_i(\boldsymbol{X}, \boldsymbol{Z}) - \tau\right) \quad \forall\, i$$

where $H(x)$ is the Heaviside step function:

$$H(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{otherwise} \end{cases}$$

and τ is the threshold.

Note that if there was a single sensitive feature (rather than multiple) we would use $\boldsymbol{z}$ (rather than $\boldsymbol{Z}$) to denote it (since it would be a vector) and if the target was multi-class rather binary we would use $\boldsymbol{P}$ (rather than $\boldsymbol{p}$) for the score since we would need a score for each class. If we have n examples, m_x non-sensitive features, and m_z sensitive features then, $\boldsymbol{X}$ is an $n \times m_x$ matrix, $\boldsymbol{Z}$ is an $n \times m_z$ and $\boldsymbol{y}$ and $\boldsymbol{p}$ are vectors with n elements. If $\boldsymbol{y}$ was a multi-class target variable with c possible classes, then $\boldsymbol{P}$ would be an $n \times c$ matrix.

For binary sensitive and target variables $\boldsymbol{z}$ and $\boldsymbol{y}$, we will set the advantaged group and advantageous target class to have the value one, the disadvantaged group and disadvantageous target class will then take the value zero.

Random variables

Following the typographical conventions described above, we use $\boldsymbol{X}$, $\boldsymbol{Z}$ (or Z for a single sensitive feature), Y, $\hat{Y}$ and P (or $\boldsymbol{P}$ for multi-class), to denote the the random variables corresponding to our non-sensitive features, sensitive features, target variable, model predicted target variable and model probability function respectively.

Special values

We will occasionally use $+$ or $-$ in the subscript of a binary variable to respectively denote the advantaged or disadvantaged outcome (or class). For example,

- $Y = y_+$ is the advantageous outcome
- $Y = y_-$ is the disadvantageous outcome
- $Z = z_+$ is the advantaged (privileged) class
- $Z = z_-$ is the disadvantaged (unprivileged) class

For brevity and readability, we shall (on occasion) omit the random variable in the event descriptor of a probability term (if it is obvious which random variable we are referring to). For example, for a binary target variable we might write,

$$\mathbb{P}(Y = y_+) = \mathbb{P}(y_+).$$

Probability density functions

As a shorthand we will use f_X to denote the probability density function for the random variable X. Note then that for a discrete random variable X, we can write,

$$\mathbb{P}(X = x) = f_X(x),$$

while for a continuous random variable we have

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Expectations

We denote the expectation as,

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) = \int_{x \in \mathcal{X}} g(x) f_X(x)\,\mathrm{d}x$$

where we take the expectation of a multi variate function, we will use a subscript to indicate the variable the expectation is taken over, e.g. $\mathbb{E}_X[g(X, Y)]$.

Naming conventions

- n for number of examples or data points
- d for differences
- r for rates

Chapter 1

Motivation and context

This chapter at a glance

- Problems with machine learning in social and political domains
- Contrasting socio-political theories of fairness in decision systems
- The history, application and interpretation of anti-discrimination law
- Association paradoxes and the importance of domain knowledge
- The different types of harm caused by biased systems

Welcome and congratulations on taking this first step towards becoming a better machine learning practitioner and citizen of the world. If you've made it here chances are you've already worked with models and have some awareness of the problem of unfair machine learning algorithms. You might be a student with a foundational course in machine learning under your belt, or a Data Scientist or Machine Learning Engineer, concerned about the impact your models might have on the world. In this book we are going to learn a whole host of techniques for measuring and mitigating bias and unfairness from machine learning models. We will implement them in Python using Jupyter Notebook and analyse their behaviour. In this book I will assume you have some experience with the Python data science stack (Pandas, NumPy, SciPy, Matplotlib and scikit-learn). At the end of this book you will be able to make educated choices about which techniques and methods to use for measuring and mitigating bias in your own models. We will work with IBMs AI Fairness 360 (AIF360) library, the most comprehensive open source library available for measuring and mitigating bias in machine learning models.

Before we embark on this journey it's important to look at the bigger picture. Bias and fairness in machine learning is a complex topic. It is not simply a matter of evaluating some formulas, implementing an algorithm, optimising a value or writing some code. These are critical components, and we'll spend much of this book on such challenges, but the problem of unfair bias in machine learning models is very much an interdisciplinary one. The choices we make in developing machine learning solutions can raise questions that are ethical, social, political, legal and philosophical in nature. Given this, we would be remiss not to start with the broader context to motivate our learning and better understand the multifaceted nature of the problem.

As we'll see in this chapter, building fair models is hard for a whole host of reasons. Fairness is not free, it requires resources. In industry where models are developed and deployed for profit, and companies compete for market share, there are strong incentives to minimise cost and decrease the time to deployment; a strategy not particularly conducive to taking a careful, considered approach. Data can be misleading - we'll see that identifying bias from static data is not just a matter of crunching numbers. Identifying causal

relationships requires domain knowledge - an understanding of the data and metrics. Fairness is a rather subjective notion. Correspondingly there is no single fairness metric one can calculate to know if any given model is fair, or single solution to making a system fairer. How does one decide which metric is the right one and whose responsibility is it to decide? Context is everything and this is one of the two main purposes of this chapter - to provide context. The other is motivation. Why is it important to consider the fairness of our models? What happens if we don't? What's the harm? What does the law say and how does it ensure models are fair? These are some of the questions we'll focus on in this chapter.

We'll start with a brief recap of the different types of machine learning disciplines. We'll discuss machine learning in the context of modelling and advocate for a healthy level of scepticism towards models and their ability to 'learn' and predict the future. We will discuss some theories of fairness in sociopolitical systems that can be related to model training and criteria. We will look at anti-discrimination laws in the US - discuss briefly the history and application of them, and the tensions that exist in their interpretation. We will discuss technical difficulties in identifying bias in data through illustrative examples of Simpson's paradox. Finally we will learn about the different types of harm that are caused by biased systems - many of which are difficult if not impossible to quantify. By the end of this chapter we will have a stronger more grounded sense of why fairness of machine learning systems is important.

1.1 Machine Learning

Machine learning can be described as the study of computer algorithms that improve (or learn) with experience. It can be broadly subdivided into the fields of supervised, unsupervised and reinforcement learning.

Supervised learning For supervised learning problems, the examples come in the form of labelled training data. Given a set of features X and labels (or targets) Y, we want to learn a mapping f, such that $Y = f(X)$, where f generalizes to previously unseen data.

Unsupervised learning For unsupervised learning problems there are no labels Y, only features X. Instead we are interested in looking for patterns and structure in the data. For example, we might want to subdivide the data into clusters of points with similar (previously unknown) characteristics or we might want to reduce the dimensionality of the data (to be able to visualize it or simply to make a supervised learning algorithm more efficient). In other words, we are looking for a new feature Y and the mapping f from X to Y.

Reinforcement learning Reinforcement learning is concerned with the problem of optimally navigating a state space to reach a goal state. The problem is framed as an agent that takes actions, which result in rewards (or penalties). The task is then to maximize the cumulative reward. As with unsupervised learning, the agent is not given a set of examples of optimal actions in various states, but rather must learn them through trial and error. A key aspect of reinforcement learning is the existence of a trade-off between exploration (searching unexplored territory in the hope of finding a better choice) and exploitation (exploiting what has been learned so far).

In our discussions we will focus on the first two categories (essentially algorithms that capture and or exploit patterns in data) primarily because these are the fields in which problems related to bias in machine learning are most pertinent (automation and prediction). As one would expect then, these are also the areas in which many of the technical developments in measuring and mitigating bias have been concentrated.

1.1.1 Machines that learn

The idea that the kinds of technologies described above are learning is an interesting one. The analogy is clear, learning from example is certainly one way to learn, though not without its issues. In less modern disciplines one would simply think of 'training' as solving an equation, optimising the parameters of a model

to best fit the data or finding the most probable parameters of the model (assuming the parametric form is the correct one and that the errors are normally distributed). So where does the terminology come from? The term "machine learning" was coined by Arthur Samuel in the 1950's when, at IBM, he developed an algorithm capable of playing draughts (checkers). By the mid 70's his algorithm was competitive at amateur level. Though it was not called reinforcement learning at the time, the algorithm was one of the earliest implementations of such ideas. Samuel used the term 'rote learning' to describe a memorization technique he implemented where the machine remembered all the states it had visited and the corresponding reward function, in order to extend the search tree.

Today, the term 'artificial intelligence' is used almost interchangeably with 'machine learning'. What's troublesome about words like 'learning' and 'intelligence' to describe these technologies is that these are amazing promises. While for some, they inspire excitement and wonder about where the technology could go, for many more, it shrouds them in mystery, is intimidating and prevents them from being able to challenge decisions made by them or justifies having that right taken away. As a data scientist faced with difficult and subjective decisions around the fairness of the technology, it might seem like not intervening is the neutral choice - to let the machine 'learn' from the data. But as we'll see, not intervening is not the neutral choice. People should be able to challenge the decisions made by 'intelligent' machines, certainly no less than they could if they were made by a person.

1.1.2 Models

Underlying every machine learning algorithm is a model (often several of them) and these have been around for millennia. Based on the discovery of palaeolithic tally sticks (animal bones carved with notches) it's believed that humans have kept numerical records for over 40,000 years. The earliest mathematical models (from around 4,000 BC) were geometric and used to advance the fields of astronomy and architecture. By 2,000 BC, mathematical models were being used in an algorithmic manner to solve specific problems by at least three civilizations (Babylon, Egypt and India).

So what exactly is a model and what's it for? A model is a simplified representation of the real world. The purpose of simplification is to reduce a complex physical system to a size that can be studied. Once we have a model which represents a theoretical understanding of the world (under a series of simplifying assumptions) we can test it by measuring and comparing the results to reality. Based on the results we can assess how accurate our understanding of the world was and update our model accordingly. The idea is to iteratively improve our understanding by figuring out where the model fails.

Historically, or at least in academia, models have been used in the pursuit of knowledge; as a mechanism to understand the world around us and explain why things behave as they do; to prove that the earth could not be flat, explain why the stars move and shift in brightness as they do or, (somewhat) more recently in the case of my PhD, explain why supersonic flows behave uncharacteristically when a shock wave encounters a vortex. As the use of models has been adopted by industry, increasingly their purpose has been geared towards prediction and automation, as a way to monetize that understanding. But the pursuit of profit inevitably creates conflicts of interests. If your goal is to learn more, finding out where your theory is wrong and fixing it is a core part of the game. In business, where the goal is to maximise profit and minimise cost, it need not be.

I recall a joke I heard at school describing how one could tell which field of science an experiment belonged to. If it changes colour, it's biology; if it explodes, it's chemistry and if it doesn't work, it's physics! Models of real world phenomena fail. They are, by their very nature, a reductive representation of an infinitely more complex real world system. They are, by construction, unable to adequately represent outliers or in some cases even just minorities. They capture the dense part of the distribution, and in turn better predict events falling in the same region. One might describe such models as simply interpolating historic data. If you have enough data, any appropriately complex model will work well[1]. It's when you use a model to predict behaviour in the tails of the distribution it is trained on (where one has less data and so uncertainty is greater) or to extrapolate (where one has no data at all) that the model becomes really important.

[1]The caveat of course is, that the more complex the phenomena you are trying to predict, the more data you need, and that growth is exponential (also known as, the curse of dimensionality).

No model is infallible. It's important to really internalise this fact. Interpreted with caution and diligence they can tell us a lot about the way things behave. But treating them as some kind of all knowing oracle that we can set up and walk away from is a mistake. Obtaining adequately rich and relevant data is a major limitation of machine learning models and yet they are increasingly being applied to problems where that kind of data simply doesn't exist. Unprecedented events happen, and individuals a model hasn't seen before will come along. We see examples in the news everyday. For most models, history really doesn't go back that far and the data they are trained and tested on is always just a sample. As I sit here writing this now, we are amidst a global pandemic. The novel Coronavirus or COVID-19 is described as a one in one hundred year event (which, as far as a machine learning model is concerned, is unprecedented). The price of oil became negative for the first time in history. AI inventory trackers, recommendation algorithms, fraud detection, marketing systems and more worldwide have been thrown off by the ways in which people now browse, buy and binge products. Many companies have been forced to step in and introduce manual corrections to their algorithms. All the while many of us have been trying to figure out why it's so hard to find toilet paper!

1.1.3 The new electricity

In 2017, Stanford Researcher Andrew Ng famously dubbed AI as the new electricity, explaining that it would transform every industry in the coming years just as electricity had 100 years ago. Recent years have seen some truly remarkable results that have sparked both excitement and fear over what lies ahead. Progress in the field of deep learning combined with increased availability and decreased cost of computational resources has led to ground breaking results. Machine learning algorithms have been able to surpass human performance on a number of tasks, from disease detection from medical images to causing the 18 time world champion Go grandmaster Lee Se-Dol to retire.

Automation seemingly offers a path to the promised land, making our lives easier, improving the efficiency and efficacy of the many industries we transact with day to day, but there are also growing and legitimate concerns that not everyone will benefit from these advances. Machine learning is already being used to automate decisions in just about every aspect of modern life; deciding which adverts to show to whom, deciding which transactions might be fraud when we shop, deciding who is able to access to financial services such as loans and credit cards, determining our treatment when sick, filtering candidates for education and employment opportunities, in determining which neighbourhoods to police and even in the criminal justice system to decide what level bail should be set at or the length of a given sentence. At almost every major life event, going to university, getting a job, buying a house, getting sick, decisions are being made by machines. By construction, these models encode existing societal biases. They not only proliferate them but are capable of amplifying them and are easily deployed at scale. Understanding the shortcomings of these models and ensuring such technologies are deployed responsibly are essential if we are to safeguard social progress.

1.2 Discrimination, bias, fairness and ethics

In 2016, Cathy O'Neil published a book in which she gave numerous accounts of models in the wild which she described as Weapons of Math Destruction. These models were deployed at scale, opaque and harmful. In one example she describes a model purported to measure teacher performance, the scores of which used to determine which teachers got a bonus and which teachers got fired. The (politically motivated) policy acted as an incentive for teachers to cheat their students' scores. Teachers that unwittingly took up where a cheating teacher left off (and did not cheat themselves), the fall in performance was falsely attributed to them. For the teachers that were fired, there was no explanation for their scores despite evidence showing the model didn't work - the fairness of the algorithm was apparently beyond challenge. In another example the book talks about just in time scheduling algorithms being used (by large retailers like Starbucks, McDonald's and Walmart) to minimise staffing the costs. By taking into account everything from weather patterns to sports events, the algorithms predict footfall and thus staffing needs. The cost saving comes at the expense of their employees. In some cases, employees are given just enough hours to fall short of qualifying for costly health insurance. Employees are subjected to haphazard schedules with little notice that prevent them from

being able to prioritise anything other than work and eliminating the possibility of any opportunity that might enable them to advance beyond the low-wage work pool. Another example talks about models which calculate recidivism risk being used to determine sentencing. Scores are based on the answers given in a questionnaire that include questions about their upbringing, neighbourhood, family and friends, a shocking practice the legality of which is bewildering.

Everyday it seems that new problems with algorithms are exposed, from democracy threatening social media to environmentally unsustainable large scale language models. The problems can seem overwhelming. There are so many different failures at play that it's hard to know where to start. It's not just the models, it's the testing and oversight of them, it's what they are used for, it's the power dynamics between the creators of the technology and the people exposed to it, it's the context and the history behind why they play out the way they do, it's conflicting interests, it's about justice, fairness, social mobility, accountability, transparency. It's about ethics. As data scientist we are not in control of everything but we are certainly not powerless. There are many issues with the way that machine learning models are built and deployed today and data scientists have an important role to play in fixing them. In this book, we will focus on those aspects which are within the control of a data science team, in particular, the modelling and deployment of machine learning systems. In this book, we will assume data is obtained externally. Related data specific topics of security, privacy, transparency, control and consent (though fundamental to the fairness of the machine learning systems), are not within the scope of this book.

Perhaps the biggest challenge in developing models ethically (after the conflicting interests introduced by trading models for currency) is that the questions we must answer are legal, philosophical, social and political in nature, questions to which there is often no definitive answer but rather competing viewpoints. What is right and wrong, fair or unfair, harmful or helpful are in general subjective and there are trade-offs between different values. Correspondingly, (as we'll see in later chapters) there is no single definition of fairness, but rather many. Furthermore, not all definitions of fairness can be satisfied simultaneously. There are trade-offs to be made, between fairness and accuracy, group fairness and individual fairness. So given these subjective choices, how does one decide what to do and whose responsibility is it to choose really? Which fairness metric is the right one? What trade-off is the right one to make? This book will not to provide the answers to these questions - there is unfortunately no silver bullet. Together we will work towards asking the right questions. Building machine learning systems ethically is not about finding the perfect answer every time but rather expanding our perspectives on the technology we develop. It's about looking for the cracks before deploying systems, preventing the foreseeable failures and doing the best we can on the ones we didn't see coming. It's worth recognising that developing technology ethically is not simply an act of benevolence but an essential part of a sustainable business strategy that can withstand the kinds of cultural shifts that happen over a generation.

Let's discuss another example. In 2016, analysis published by Bloomberg uncovered racial disparities in eligibility for Amazon's same day delivery services for Prime customers[2][13]. The study used census data to identify Black and White residents and plot the data points on city maps which simultaneously showed the areas that qualified for the Prime customer same day delivery. The disparities are glaring at a glance. In six major cities, New York, Boston, Atlanta, Chicago, Dallas, and Washington, DC where the service did not have broad coverage, it was mainly Black neighbourhoods that were ineligible. In the latter four cities, Black residents were about half as likely to live in neighbourhoods eligible for Amazon same-day delivery as White residents.

At the time Amazon's process in determining which ZIP codes to serve was reportedly a cost benefit calculation that did not explicitly take race into account but the resemblance of these maps to redlining maps from the 1930's is hard to not see. Redlining was the (now illegal) practice of declining (or raising prices for) financial products to people based on the neighbourhood where they lived. Because neighbourhoods were racially segregated (a legacy that lives on today), public and private institutions were able to systematically exclude minority populations from the housing market and deny loans for house improvements without explicitly taking race into account. Between 1934 and 1962, the Federal Housing Administration distributed

[2]To be clear, the same day delivery was free for eligible Amazon Prime customers on sales exceeding $35. Amazon Prime members pay a fixed annual subscription fee, thus the disparity is in the level of service provided for Prime customers who are eligible verses those that are not.

$120 billion in loans. Thanks to redlining, 98% of these went to White families.

Amazon is a private enterprise, and it is legally entitled to make decisions about where to offer services based on how profitable it is. Some might argue they have a right to be able to make those decisions. Amazon is not responsible for the injustices that created racial disparities in wealth, but the reality is that such disparities in access to goods and services contribute to it. Consider also that same-day delivery for many people would be described as a luxury, but that this must also be considered in context. The cities affected have a long histories of racial segregation and economic inequality resulting from systemic racism now deemed illegal. They are neighbourhoods which to this day are underserved by brick and mortar retailers, where residents are forced to travel further and pay more for household essentials. Now we are in the midst of a pandemic, where once delivery of household goods used to be a luxury, with so many forced to quarantine, suddenly it's become far more valuable. What we consider to be a necessity changes over time, it depends on where one lives, their circumstances and more. Finally, consider the scale of Amazon's operations, in 2016 one third of retail e-commerce spending in the US was with Amazon (that number has since risen to almost 50%).

Rational Prejudice

What do you think? Is it fair for Amazon to provide different service levels based on where customers live in this way?

1.2.1 Fairness as justice

Fairness is subjective. While one might consider it to be equality, another might define it as deservedness. Our beliefs about what fairness means, form the basis of our view of how the world should be, and lie at the very core of who we are. They are shaped by our experiences, (upbringing, class, education, occupation and more). They determine our stance on everything - politics, economics and society. Cathy O'Neil describes algorithms as "opinions embedded in code". Developing machine learning models is not an objective scientific process, it involves making a series subjective choices. One of the most fundamental way in which we impose our opinion is in deciding how we measure success. Success for the same algorithm will look rather different depending on which perspective you look at it from.

Utilitarianism

Let's consider the standard, approach to training a model which is essentially to minimise the aggregate error on the training data. Our objective is to maximise utility. We want to essentially automate past decisions. Replicating the past is a rather narrow objective - efficiency. Get a machine to do what a human did. Increase productivity and reduce cost. The decision process is loosely justified in a utilitarian[3] sense, in that the decision process which maximises accuracy (probability of error) for the greatest number of people (everyone in the training data) is justified. A major flaw with utilitarianism is that it is impossible in practice to account for the unforeseeable longer-term impacts of an action that would impact the happiness or well-being of a population. It's easy to come up with examples of actions which seemed optimal with the information one had at one point but one might not have taken in hindsight.

A political conception of justice

In his theory Justice As Fairness[18], John Rawls takes a different approach. He describes an idealised democratic framework, based on liberal principles and explains how unified laws can be applied (in a free society made up of people with disparate world views) to create a stable sociopolitical system. One in which citizens would not only freely co-operate, but further advocate for it. This was achieved through his political conception of justice which would:

[3] According to a utilitarian doctrine, an action is justified if it maximises the happiness or well-being for the greatest number of people.

1. grant all citizens a set of basic rights and liberties
2. give special priority to the aforementioned rights and liberties over demands to further the general good, e.g. increasing the national wealth
3. assure all citizens sufficient means to make use of their freedoms.

The distinction between liberties and rights can be blurred, indeed they are commonly used interchangeably, but they have distinct meanings. Liberties protect against government actions while governments should proactively protect the rights of it's citizens. The special priority given to the basic rights and liberties in the political conception of justice contrasts with a utilitarian doctrine. The Parallels can be drawn here in machine learning where there is a trade-off between fairness and the utility of our algorithm. Maximising aggregate accuracy does not take into consideration how the errors of the system are distributed. Translating Rawls' political conception of justice might require some minimum accuracy level (maximum probability of error) to be set for all members of the population, even if this would result in a less accurate algorithm on aggregate.

Principles of Justice as Fairness

1. **Liberty principle:** Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which is compatible with the same scheme of liberties for all;
2. **Equality principle:** Social and economic inequalities are to satisfy two conditions:
   (a) **Fair equality of opportunity:** The offices and positions to which they are attached are open to all, under conditions of fair equality of opportunity;
   (b) **Difference principle** They must be of the greatest benefit to the least-advantaged members of society.

The principles of Justice as Fairness are ordered by priority so that fulfilment of the liberty principle takes precedence over the equality principles and fair equality of opportunity takes precedence over the difference principle.

The first principle grants basic rights and liberties to all citizens which are prioritised above all else and cannot be traded for other societal benefits. It's worth spending a moment thinking about what those rights and liberties look like. They are the the basic needs that are important for people to be free, to have choices and the means to pursue their aspirations. Today many of what Rawls considered to be basic rights and liberties are allocated algorithmically; education, employment, housing, healthcare, consistent treatment under the law to name a few.

The second principle requires positions to be allocated meritocratically, with all similarly talented (with respect to the skills and competencies required for the position) individuals having the same chance of attaining such positions i.e. that allocation of such positions should be independent of social class or background. Later we'll see how equality of opportunity might be loosely translated (under some assumptions) to metrics measuring fairness of a classifier across different subgroups of the population.

The third principle acts to prevent redistribution of social and economic currency from the rich to the poor by requiring that inequalities are of maximal benefit to the least advantaged in a society, also described as the maximin principle. In this principle, Rawls does not take the simplistic view that inequality and fairness are mutually exclusive but rather concisely articulates when the existence of inequality becomes unfair.

1.2.2 A brief history of US anti-discrimination laws

It's easy (when you're crunching numbers, interpreting changes to some metric that boils some notion of fairness down to a single number) to forget that anti-discrimination laws were born out of long-standing, vast and systemic discrimination against historically oppressed and disadvantaged classes. It is not possible, in this book, to adequately cover the history behind anti-discrimination laws in the US (let alone the world);

provide a picture of the barriers people faced; how they have contributed to disparities in all measures of prosperity (health, wealth, housing, crime, incarceration) or look at how those disparities have evolved over time (before and after said laws were passed, until today). That said this book would be incomplete without some discussion or a reminder of the context.

Anti-discrimination laws in the US rest on the 14th amendment to the constitution which grants citizens "equal protections of the law". Class action law suit Brown v Board (of Education of Topeka, Kansas) was a landmark case which in 1954 (nine decades after the abolition of slavery), legally ended racial segregation in the US. Justices ruled unanimously that racial segregation of children in public schools was unconstitutional, establishing the precedent that "separate-but-equal" was, in fact, not equal at all. Though Brown v Board did not end segregation in practice, resistance to it in the south fuelled the civil rights movement. In the years that followed the NAACP (National Association for the Advancement of Coloured People) challenged segregation laws. In 1955, Rosa parks refusing to give up her seat on a bus in Montgomery (Alabama) led to sit ins and boycotts, many of them led by Martin Luther King Jr. The resulting Civil rights act of 1964 eventually brought an end to "Jim Crow" laws which barred Blacks from sharing buses, schools and other public facilities with Whites.

After the violent attack by Alabama state troopers on participants of a peaceful march from Selma to Montgomery was televised, The Voting Rights Act of 1965 was passed. It overcame many barriers (including literacy tests), at state and local level, used to prevent Black people from voting. Before this incidents of voting officials asking Black voters to "recite the entire Constitution or explain the most complex provisions of state laws"[14] in the south were common place.

In the years following the second world war, there were many attempts to pass an Equal Pay Act. Initial efforts were led by unions who feared men's salaries would be undercut by women who were paid less for doing their jobs during the war. By 1960, women made up 37% of the work force but earned on average 59 cents for each dollar earned by men. The Equal Pay Act was eventually passed in 1963 in a bill which endorsed "equal pay for equal work". Laws for gender equality were strengthened the following year by the Civil Rights Act of 1964.

Throughout the 1800's the American federal government displaced Native American communities to facilitate White settlement. In 1830 the Indian Removal Act was passed in order to relocate hundreds of thousands of Native Americans. Over the following two decades, thousands of those forced to march hundreds of miles west on the perilous "Trail of Tears" died. By the middle on the century, the term "manifest destiny" was popularlised to describe the belief that White settlement in North America was ordained by God. In 1887, the Dawes Act laid the groundwork for the seizing and redistribution of reservation lands from Native to White Americans. Between 1945 and 1968 the federal government terminated recognition of more than 100 tribal nations placing them under state jurisdiction. Once again Native Americans were relocated, this time from reservations to urban centres.

In addition to displacing people of colour, the federal government also enacted policies that reduced barriers to home ownership almost exclusively for White citizens - subsidizing the development of prosperous suburbs, guaranteeing mortgages and enabling access to job opportunities by building highway systems for White commuters, often through communities of colour. Even government initiatives aimed at helping veterans of World War II to obtain home loans accommodated Jim Crow laws allowing exclusion of Black people. In the wake of the Vietnam war, just days after the assassination of Martin Luther King J, the Fair Housing Act of 1968 was passed, prohibiting discrimination concerning the sale, rental and financing of housing based on race, religion, national origin or sex.

The Civil Rights Act of 1965 acted as a catalyst for many other civil rights movements, including those protecting people with disabilities. The Rehabilitation Act (1973) removed architectural, structural and transportation barriers and set up affirmative action programs. The Individuals with Disabilities Education Act (IDEA 1975) required free, appropriate public education in the least restrictive environment possible for children with disabilities. The Air Carrier Access Act (1988) which prohibited discrimination on the basis of disability in air travel and ensured equal access to air transportation services. The Fair Housing Amendments Act (1988) prohibited discrimination in housing against people with disabilities.

Title IX of the education amendments of 1972 prohibits federally funded educational institutions from discriminating against students or employees based on sex. The law ensured that schools (elementary to

university level) that were recipients of federal funding (nearly all schools) provided fair and equal treatment of the sexes in all areas, including athletics. Before this few opportunities existed for female athletes. The National Collegiate Athletic Association (NCAA) offered no athletic scholarships for women and held no championships for women's teams. Since then the number of female college athletes has grown five fold. The amendment is credited with decreasing dropout rates and increasing the numbers of women gaining college degrees.

The Equal Credit Opportunity Act was passed in 1974 when discrimination against women applying for credit in the US was rife. It was common practice for mortgage lenders to discount incomes of women that were of 'child bearing' age or simply deny credit to them. Two years later the law was amended to prohibit lending discrimination based on race, color, religion, national origin, age, the receipt of public assistance income, or exercising one's rights under consumer protection laws.

In 1978, congress passed the Pregnancy Discrimination Act in response to two Supreme Court cases that ruled that excluding pregnancy related disabilities from disability benefit coverage was not gender based discrimination, and did not violate the equal protection clause.

Table 1.1 shows a (far from exhaustive) summary of regulated domains with corresponding US legislation. Note that legislation in these domains extend to marketing and advertising not just the final decision. Table 1.2 provides a list of protected characteristics under US federal law with corresponding legislation (again not exhaustive).

Table 1.1: Regulated domains in the private sector under US federal law.

Domain	Legislation
Finance	Equal Credit Opportunity Act
Education	Civil Rights Act (1964) Education Amendment (1972) IDEA (1975)
Employment	Equal Pay Act(1963) Civil Rights Act (1964)
Housing	Fair Housing Act (1968) Fair Housing Amendments Act (1988)
Transport	Urban Mass Transit Act (1970) Rehabilitation Act (1973) Air Carrier Access Act (1988)
Public accommodation[a]	Civil Rights Act (1964)

[a]Prevents refusal of customers.

1.2.3 Application of the law

The descriptions above might might lead you to believe the the problem of discrimination (at least for those who have the means to take legal action) is largely solved by the law. It's worth taking a look at how anti-discrimination laws in the US are applied. For illustrative purposes, we look at Title VII of the Civil rights act of 1964 in the context of employment discrimination. For a more detailed discussion see Barocas & Selbt[9].

Legal liability for discrimination against protected classes can be established as disparate treatment and/or disparate impact. Disparate treatment (also described as direct discrimination in Europe) refers to both differing treatment of individuals based on protected characteristics, and intent to discriminate. Disparate impact (also described as indirect discrimination in Europe) does not consider intent but addresses policies and practices that disproportionately impact protected classes.

Table 1.2: Protected characteristics under US Federal Law.

Protected Characteristic	Legislation
Race	Civil Rights Act (1964)
Sex	Equal Pay Act (1963) Civil Rights Act (1964) Pregnancy Discrimination Act (1978)
Religion	Civil Rights Act (1964)
National Origin	Civil Rights Act (1964)
Citizenship	Immigration Reform & Control Act
Age	Age Discrimination in Employment Act (1967)
Familial status	Civil Rights Act (1968)
Disability status	Rehabilitation Act of 1973 American with Disabilities Act of 1990
Veteran status	Veterans' Readjustment Assistance Act 1974 Uniformed Services Employment & Reemployment Rights Act
Genetic Information	Civil Rights Act(1964)

Disparate treatment

Disparate treatment effectively prohibits rational prejudice (backed by data showing the protected feature to be correlated) as well as denial of opportunities based on protected characteristics. It effectively prevents the use of protected characteristics as features in machine learning algorithms. It's noteworthy that in the case of disparate treatment, the actual impact of using the protected features on the outcome is irrelevant; so even if a company could show that the target variable produced by their model had zero correlation with the protected characteristic, the company would still be liable for disparate treatment. This fact is somewhat bizarre given that not using the protected feature in the algorithm provides no guarantee that the algorithm is not biased in relation to it. Indeed an organisation could very well use their data to predict the protected characteristic.

In an effort to avoid disparate treatment liability, many organisations do not even collect data relating to protected characteristics, leaving them unable to accurately measure, let alone address, bias in their algorithms, even if they might want to[4]. In summary, disparate treatment as applied today does not resolve the problem of unconscious discrimination against disadvantaged classes through their use of machine learning algorithms. Further it acts as a deterrent to ethically minded companies that might want to measure the biases in their algorithms.

Disparate Treatment

Suppose a company predicts the sensitive feature and uses this as an input to its model. Should this be considered disparate treatment?

What about the case where the employer implements an algorithm, finds out that it has a disparate impact, and uses it anyway? Doesn't that become disparate treatment? No it doesn't and in fact, somewhat surprisingly, deciding not to apply it upon noting the disparate impact could result in a disparate treatment claim in the opposite direction[4]. We'll return to this later. Okay, so what about disparate impact?

[4]In fact, I met a data scientist at a conference, who was working for a financial institution, that said her team was trying to predict sensitive features such as race and gender in order to measure bias in their algorithms.

Disparate impact

In order to establish a violation, it is not enough to simply show that there is a disparate impact, but it must also be shown either that there is no business justification for it, or if there is, that the employer refuses to use another, less discriminatory, means of achieving the desired result. So how much of an impact is enough to warrant a disparate impact claim? There are no rules here only guidelines. The Uniform Guidelines on Employment selection procedures from the Equal Employment Opportunity Commission (EEOC) provides a guideline that if the selection rate from one protected group is less than four fifths of that from another, it will generally be regarded as evidence of adverse impact, though it also states that the threshold would depend on the circumstances.

Assuming the disparate impact is demonstrated, the issue becomes proving business justification. The requirement for business justification has softened in favour of the employer over the years; treated as "business necessity"[1] earlier on and later interpreted as "business justification"[2]. Today, it's generally accepted that business justification lies somewhere between the extremes of "job-relatedness" and "business necessity". As a concrete example of disparate impact and taking the extreme of job-relatedness - the EEOC along with several federal courts have determined that discrimination on the sole basis of a criminal record to be a violation under disparate impact unless the particular conviction is related to the role, because Non-White applicants are more likely to have a criminal conviction.

For a machine learning algorithm, business justification boils down to the question of job-relatedness of the target variable. If the target variable is improperly chosen, a disparate impact violation can be established. In practice however the courts will accept most plausible explanations of job-relatedness since not accepting it would set a precedent that it is determined discriminatory. Assuming the target variable to be proven job-related then, there is no requirement to validate the model's ability to predict said trait, only a guideline which sets a low bar (a statistical significance test showing that the target variable correlates with the trait) and which the court is free to ignore.

Assuming business justification is proven by the employer, the final burden then falls on the plaintiff to show that the employer refused to use a less discriminatory "alternative employment practice". If the less discriminatory alternative would incur additional cost (as is likely) would this be considered refusing? Likely not.

While on the surface, disparate impact might seem like a solution, the current framework of a weak business justification (in terms of a plausible target variable) and the employer refusing an alternative employment practice with no requirement to validate the model offers little resolve. Clearly there is need for reform.

1.2.4 Anti-classification versus anti-subordination

Just as the meaning of fairness is subjective so is the interpretation of anti-discrimination laws. At one extreme, anti-classification holds the weaker interpretation, that the law is intended to prevent classification of people based on protected characteristics. At the other extreme, anti-subordination defines the stronger stance, that anti-discrimination laws exist to prevent social hierarchies, class or caste systems based on protected features and, that it should actively work to eliminate them where they exist. An important ideological difference between the two schools of thought is in the application of positive discrimination policies. Under anti-subordination principles, one might advocate for affirmative action as a means to bridge gaps in access to employment, education, pay and other such pursuits, that are a direct result of historical systemic discrimination against particular groups. A strict interpretation of the anti-classification principle would prohibit such actions. Both ideologies have been argued and upheld in landmark cases.

In 2003, the Supreme Court held that a student admissions process that favours "under-represented minority groups" does not violate the Fourteenth Amendment[3], provided it evaluated applicants holistically at an individual level. The same year, the New Haven Fire Department administered a two part test in order to fill 15 openings. Examinations were governed in part by the City of New Haven. Under the city charter, civil service positions must be filled by one of the top three scoring individuals. 118 (White, Black and Hispanic) fire fighters took the exams. Of the resulting 19 candidates who scored highest on the tests and

could the considered for the positions, none were Black. After heated public debate and under threat of legal action either way, the city threw out the test results. This action was later determined to be a disparate treatment violation. In 2009, the court ruled that disparate treatment could not be used to avoid disparate impact without sufficient evidence of liability of the latter[4]. This landmark case was the first example of conflict between the two doctrines of disparate impact and disparate treatment or anti-classification and anti-subordination.

Disparate treatment seems to align well with anti-classification principles, seeking to prevent intentional discrimination based on protected characteristics. In the case of disparate impact, things are less clear. Is it a secondary 'line of defence' designed to weed out well masked intentional discrimination? Or is its intention to address existing inequalities? One can draw parallels here with the 'business necessity' versus 'business justification' requirements discussed earlier.

1.2.5 Future legislation

In May 2018, the European Union (EU) brought into action the General Data Protection (GDPR) a legal framework around the protection of personal data of EU citizens. The framework is divided into binding and non-binding recitals. The regulation sets provisions for processing of data in relation to decision making, described as 'profiling' under recital 71[6]. Though currently non-binding, it provides an indication of what's to come. The recital talks specifically about having the right not to be subject to decisions based solely on automated processing. It specifically talks about credit applications, e-recruiting and any system which analyses or predicts aspects of a persons performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements. The recital also talks about requirements around using "appropriate mathematical or statistical procedures" to prevent "discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation".

In April 2019, the Algorithmic Accountability Act was proposed. The bill requires specified commercial entities to conduct impact assessments of automated decision systems and specifically states that assessments must include evaluations and risk assessment in relation to "accuracy, fairness, bias, discrimination, privacy, and security" not just for the model output but for the training data. The bill has cosponsors in 22 states and has been referred to the Committee on Commerce, Science, and Transportation for review. These examples are clear indications that the issues of fairness and bias in automated decision making systems are on the radar of regulators.

1.3 The problem with data

The problem of distinguishing correlation from causation is an important one in identifying bias. To demonstrate the danger of making causal inferences from observational data and stress the importance of subject matter knowledge and rigorous analysis in making such determinations, we discuss a well known and particularly relevant example of Simpson's paradox[10], also known as the reversal paradox and Yule-Simpson effect.

1.3.1 Simpson's paradox

In 1973, University of California, Berkeley received approximately 15,000 applications for the fall quarter[10]. At the time it was made up of 101 departments. 12,763 applications reached the decision stage. Of these 8442 were male and 4321 were female. The acceptance rates for the applicants were 44% and 35% respectively (see Table 1.3).

On the face of it, it seems a likely case of discrimination against women. Indeed, a χ^2 hypothesis test for independence between the variables (gender and application acceptance) reveals that the probability of observing such a result or worse, assuming they are independent, is 6×10^{-26}. A strong indication that they are not independent and therefore bias in favour of male applicants. Since admissions are determined by the individual departments, it's worth trying to understand which departments might be responsible. We focus

Table 1.3: Graduate admissions data from Berkeley (fall 1973).

Gender	Admitted	Rejected	Total	Acceptance Rate
Male	3738	4704	8442	44.3%
Female	1494	2827	4321	34.6%
Aggregate	5232	7531	12763	41.0%

on the data for the six largest departments, shown in Table 1.4. Here again we see a similar pattern. There appears to be bias in favour of male applicants, and a χ^2 test shows that the probability of seeing this result under the assumption of independence is 1×10^{-21}. It looks like we have quickly narrowed down our search.

Table 1.4: Graduate admissions data from Berkeley (fall 1973) for the six largest departments.

Gender	Admitted	Rejected	Total	Acceptance Rate
Male	1198	1493	2691	44.5%
Female	557	1278	1835	30.4%
Aggregate	1755	2771	4526	38.8%

Figure 1.1 shows the acceptance rates for each department by gender, in decreasing order of acceptance rates. Performing χ^2 tests for each department reveals the only department where there is strong evidence of bias is A, but the bias is in favour of female applicants. The probability of observing the data for department A, under the assumption of independence, is 5×10^{-5}. So what's going on? Figure 1.2 shows the application

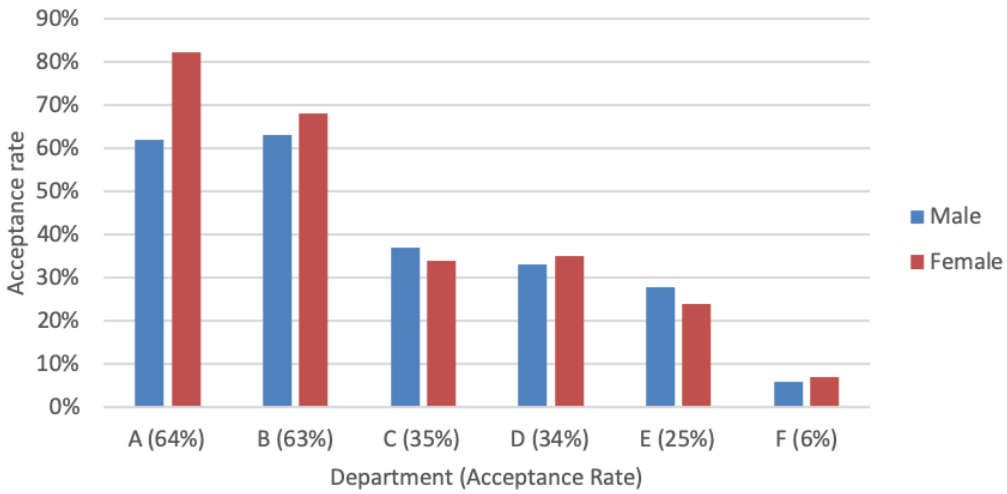


Figure 1.1: Acceptance rate distributions by department for male and female applicants.

distributions for male and female applicants for each of the six departments. From the plots we are able to see a pattern. Female applicants are more often applying for departments with a lower acceptance rate. In other words a larger proportion of the women are being filtered out overall, simply because they are applying to departments that are harder to get into.

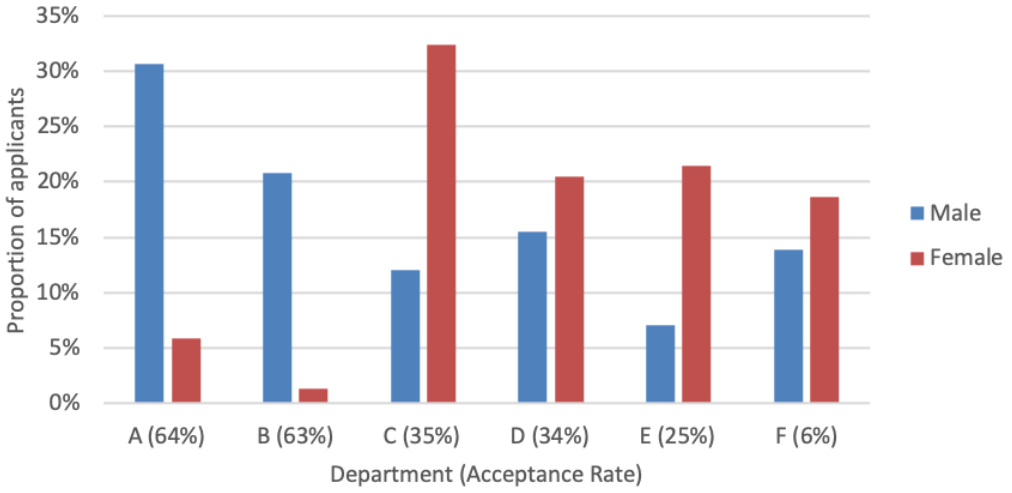


Figure 1.2: Application distributions by department for male and female applicants.

This is a classic example of Simpson's Rule - where an observable relationship between two categorical variables (in this case gender and acceptance) disappears or reverses after controlling for one or more other variables (in this case department). Simpson's Rule is a special case of so called association paradoxes (where the variables are categorical, and the relationship changes qualitatively), but the same rules also apply to continuous variables. The marginal measure of association (e.g. correlation) between two variables need not be bounded by the partial measures of association after controlling on one or more variable. Although Edward Hugh Simpson famously wrote about the paradox in 1951, it was not discovered by him. In fact, it was reported by George Udny Yule as early as 1903 and in the case of the association paradox for continuous variables, that was demonstrated by Karl Pearson in 1899.

Let's discuss another quick example. A 1996 follow-up study on the effects of smoking recorded the mortality rate for the participants over a 20 year period. They found higher mortality rates among the non-smokers, 31.4% compared to 23.9% which, in itself, might imply a considerable protective affect from smoking. Clearly there's something fishy going on. Disaggregating the data by age group showed that the mortality rates were higher for smokers in all but one of them. Looking at the age distribution of the populations of smokers and non-smokers, it's apparent that the age distribution of the non-smoking group is more positively skewed, that is, they are older on average. This concords with the rationale that non-smokers live longer - hence the difference in age distributions of the participants.

1.3.2 Causality

In both the above examples, it is the disaggregated data that contains salient information and enables us to understand the true nature of the relationship between the variables of interest. As we shall see in this section, this need not be the case. To show this, we discuss two examples. In each case, the data is identical but the meaning of the variables is not. The examples are those Simpson gave in his original 1951 paper[20].

Suppose we have three binary variables, A, B and C, and we are interested in understanding the relationship between A and B given a set of 52 data points. The contingency tables[5] for variables A and B are shown in Figure 1.5, first for all the data points and the stratified by C. The first table indicates that A

[5]Each cell of a contingency table shows the number of examples in the dataset satisfying the conditions given in the corresponding row and column headers. The final row and column typically show totals, in our case (in Figure 1.3) we display the rates instead for convenience.

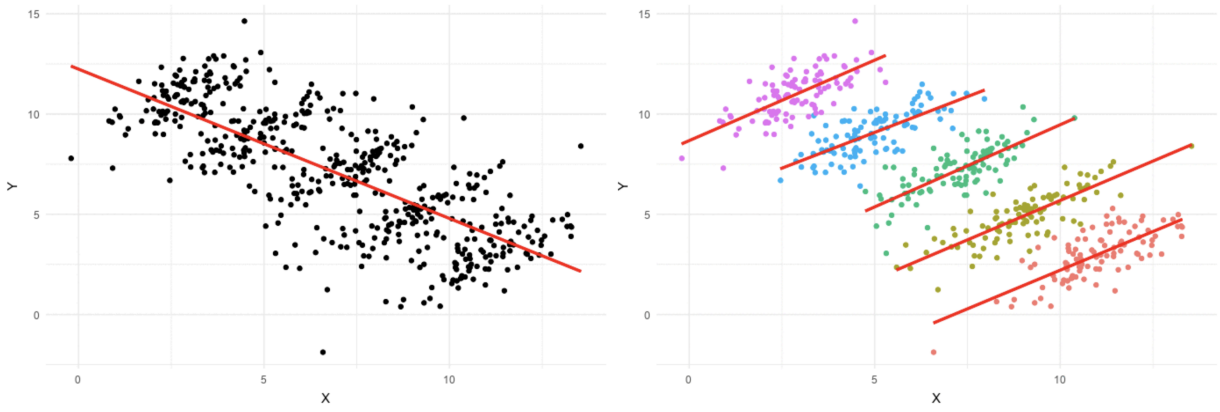


Figure 1.3: Visualisation of Simpsons Paradox. Wikipedia.

and B are unconditionally (i.e. marginally) independent (since changing the value of one variable does not change the distribution of the other). The next two tables suggest A and B are conditionally dependent given C. Which distributions give us the most relevant understanding of the association between A and B? To show that it depends on the context, we consider two different examples.

Table 1.5: Contingency tables[5] for variables A, B and C.

All data points

	$A=1$	$A=0$
$B=1$	20	6
$B=0$	20	6
Rate	50%	50%

$C=1$

	$A=1$	$A=0$
$B=1$	5	3
$B=0$	8	4
Rate	38%	43%

$C=0$

	$A=1$	$A=0$
$B=1$	15	3
$B=0$	12	2
Rate	56%	60%

Example 1: Pack of cards

Suppose the population is a pack of cards. It so happens that baby Milen has been messing about with the cards and made some dirty in the process.

- A tells us if the card is plain ($A=1$) or royal (King, Queen, Jack; $A=0$).
- B tells us if the card is black ($B=1$) or red ($B=0$).
- C tells us if the card is dirty ($C=1$) or clean ($C=0$).

In this case, it's clear we are interested in the aggregated data since the cleanliness of the cards has no bearing on the joint distribution of A and B.

Example 2: Treatment effect on mortality rate

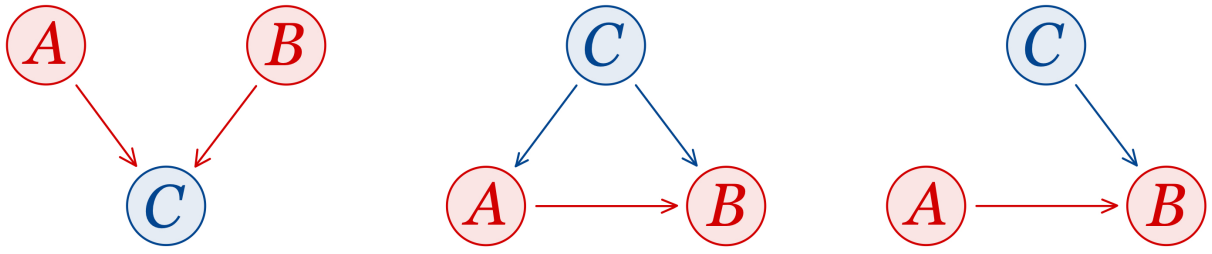
Next, suppose that the data relates to the results of medical trials for a drug on a potentially lethal illness.

- A tells us if the subject was treated ($A=1$) or not ($A=0$).
- B tells us if the subject died ($B=1$) or recovered ($B=0$).
- C tells us if the subject was male ($C=1$) or female ($C=0$).

In this case the disaggregated data shows that the drug reduces the mortality rate for both male and female participants and the effect is obscured by aggregating the data.

Back to causality

The key difference between these examples is the causal relationship between the variables rather than the statistical structure of the data. In the first example, the variable C is a 'colliding' variable, in the second case it is a 'confounding' variable. Figure 1.4 shows the causal relationships between the variables in the two cases. The causal diagram in Figure 1.4 a) shows the variables A, B and C for the first example; card



(a) Colliding variable C (b) Confounding variable C (c) Prognostic variable C

Figure 1.4: Causal diagrams for A, B and C when C is a colliding, confounding and prognostic variable.

type (plain or royal), colour (black or red) and cleanliness (dirty or clean) respectively. The arrows exist from A to C and B to C because apparently, baby Milen had a preference for royal cards over plain and red cards over black. Conditioning on a collider C generates an association between A and B, even if they are unconditionally independent. This common effect is often observed as selection bias.

The causal diagram in Figure 1.4 b) shows the variables A, B and C for the second example; treat (yes or no), died (yes or no) and gender (male or female) respectively. The arrows exist from C to A because men were less likely to receive treatment and from C to B because men were also less likely to die. The arrow from A to B represents the effect of treatment on mortality which is observable only by conditioning on gender. Note that there are two sources of association in opposite directions between variables A and B (treatment and death). A positive association due to the differing effect by gender and a negative association due to the efficacy of the treatment. The two effects cancel each other out when the data is aggregated.

We see through the discussion of these two examples that statistical reasoning is not sufficient to be able to determine which of the distributions (marginal or conditional) are relevant. Doing so requires subject matter knowledge regarding the causal structure of the the variables. Note that the above conclusions in relation to colliding and confounding variables does not generalize to complex time varying problems with dynamic variables and confounders.

1.3.3 Collapsibility

We have demonstrated the importance of causality in interpreting data and dissecting Simpson's Paradox. But there is another factor involved in its manifestation, that is, the nature of the measure of association in question. Suppose that in the study of the efficacy of the treatment, men and women were equally likely to be treated. This removes the causal relationship between variables A and C (treatment and gender), and variable C becomes prognostic rather than confounding. See Figure 1.4 c). In this case the decision as to which distributions are most relevant would depend only on the target population under discussion. In the absence of the confounding variable in our study one might reasonably expect the marginal measure of association to be bounded by the partial measures of association. Such intuition is correct in the case where the measure of association is collapsible (that is, it can be expressed as the weighted average of the partial measures), but not otherwise. Some examples of collapsible measures of association are the risk ratio and

risk difference. The odds ratio however is not collapsible. We'll return to these measures of association in chapter 3.

1.4 What's the harm?

In this section we discuss some of the broader harms caused related to machine learning technologies.

1.4.1 The illusion of objectivity

One of the most concerning things about the machine learning revolution, is perception that these algorithms are somehow objective (unlike humans), and are therefore a better substitute for human judgement. This viewpoint is not just a belief of laymen but an idea that is also projected from within the machine learning community. There are often financial incentives to exaggerate the efficacy of such systems. It is important to be cautious in describing machine learning algorithms as objective. Data is produced by a necessarily subjective set of decisions (how and who to sample, how to group events or characteristics, which features to collect). Modelling also involves making choices about how to process the data, what class of model to use and how to measure success. Finally, even if our model is calibrated to the data well, it says nothing about the distribution of errors across the population. The consistency of algorithms in decision making compared to humans (who individually make decisions on a case by case basis) is often described as a benefit, but it's their very consistency that makes them dangerous - capable of discriminating systematically and at scale.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a "case management system for criminal justice practitioners". The system, produces recidivism risk scores. It has been used in New York, California and Florida, but most extensively in Wisconsin since 2012, at a variety of stages in the criminal justice, from sentencing to parole. The documentation for the software describes it as an "objective statistical risk assessment tool".

In 2013, Paul Zilly was convicted of stealing a push lawnmower and some tools in Barron County, Wisconsin. The prosecutor recommended a year in county jail and follow-up supervision that could help Zilly with "staying on the right path." His lawyer agreed to a plea deal. But Judge James Babler upon seeing Zilly's COMPAS risk scores overturned the plea deal that had been agreed on by the prosecution and defense, and imposed two years in state prison and three years of supervision. At an appeals hearing later that year, Babler said "Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months"[8]. In other words the judge believed the risk scoring system to hold more insight that the prosecutor who had personally interacted with the defendant.

1.4.2 The ethics of classification

The appeal of classification is clear. It creates a sense of order and understanding. It enables us to formulate problems neatly and solve them. An email is spam or it's not. An x-ray shows tuberculosis or it doesn't. A treatment was effective or it wasn't. There are lots of useful applications of classification. We think of taxonomies as objective categorisations, but they are not. They are snapshots in time, representative of the culture and biases of the creators. The very act of creating a taxonomy, gives life to some categories while erasing others. Classifying people inevitably has the effect of reducing them to labels; labels that can result in people being treated as members of a group, rather than individuals; labels that can linger for much longer than they should. The Dewey Decimal System was developed in the late 1800's and widely adopted in the 1930's to classify books. Until 2015, it categorised homosexuality as a mental derangement.

Classification of people in particular has a dark history that continues today. From the 1930's until the second world war, machine classification systems were used by Nazi Germany to process census data in order to identify and locate Jews, determine what property and businesses they owned, find anything of value that could be seized and finally to send them to their deaths in concentration camps. Classification systems have often been entangled with political and social struggle across the world. They have been used extensively in many parts of the world to enforce racial segregation and social hierarchies that determined everything from where people could live and work to whom they could marry. In 2019 it was estimated that some half a

million Uyghurs (and other minority Muslims) are being held in internment camps in China without charge for the purposes of 'countering extremism' and promoting 'social integration'.

Recent papers on detecting criminality"[24] and sexuality[23] and ethnicity[22] from facial images have sparked controversy in the academic community. The latter in particular looks for facial features that identify among others, Chinese Uyghurs. Physiognomy (judging character from the physical features of a persons face) and phrenology (judging a persons level of intelligence from the shape and dimensions of their cranium) have historically been used as pseudo-scientific tools of oppressors, to prove the inferiority races and justify subordination and genocide. If machine gaydar doesn't seem like that big a deal it's worth remembering that homosexuality is still illegal in over 70 countries, some of which enforce the death penalty. The authors of the paper do state that their "findings expose a threat to the privacy and safety of gay men and women", nevertheless it is not without merit to ask if some technologies should be built at all.

1.4.3 The filter bubble

Many believed the internet would breath new life into democracy. The decentralization, decreased cost and increased accessibility of information would result in greater distribution of power and flatter social structures. In this new era people would be able to share ideas and organise grass roots movements at a scale that would accelerate progress. Some of these ideas have been realised to an extent but the increased accessibility and volume of data has created new problems. The amount of information available to us through the internet is overwhelming. Email, blog posts, Twitter, Facebook, Instagram, Linked In, What's App, You Tube, Netfix, TikTok and more. Today there are seemingly endless ways and places for us to communicate and share information. This barrage of information has resulted in what has been described as the attention crash. There is simply too much information for us to attend to all of it meaningfully. The mechanisms through which we can acquire new information that demands our attention too have expanded. We carry our smart phones everywhere we go and sleep beside them. For many people there is hardly a moment when we are unplugged and inaccessible. Our focus shifts from text message, to email to blog post in the blink of an eye. Media producers themselves have adapted their content in order to accommodate our new shorted attention spans.

With so much information available it's easy to see the appeal of automatic filtering and curation. And of course, how good would said system really be if it didn't take into account our personal tastes and preferences? So what's the problem?! Over the last decade, personalisation has become entrenched in the systems we interact with day to day. Targeted advertising was just the beginning. Now it's not just the trainers you browsed once that follow you around the web until you buy them, it's everything. Since 2009, Google has returned personalised results every time one queries their search engine, so two people who enter the same text don't necessarily get the same result. Some 40% of Americans under thirty get their news through social networking sites such as twitter and Facebook but this may be happening without you even knowing. Since 2010, it's not the Washington Post that decides which news story you see in the prime real estate that is the top right hand corner of their home page, it's facebook - the same goes for the New York Times. So the kinds of algorithms that once determined what we spent our money on now determine our very perception of the world around us.

Ignoring, for a moment, the fact that having the power to shape people's perception of the world, in just a few powerful hands is in itself a problem. A question worth pondering on is what kind of citizens people who only ever see things they 'like', would make. As Eli Pariser put it in his book The Filter Bubble, "what one seems to like may not be what one actually wants, let alone what one needs to know to be an informed member of their community or country". There was a time when people believed the internet would make the world smaller. Anyone, regardless of their background, could be our next door neighbour. In some senses personalisation does the exact opposite. It risks us all living in a room full of mirrors, where we only ever hear the voices of people who see the world as we do, being deprived of differing perspectives. Of course we have always lived in our own filter bubble in some respects but the thing that has changed is that now we don't make the choice and often don't even know when we are in it. We don't know when or how decisions are made about what we should see. We are more alone in our bubbles than we have ever been before.

Social capital is created by the interpersonal bonds we build in shared identity, values, trust and reci-

procity. It encourages people to collaborate in order to solve common problems for the common good. There are two kinds of social capital, bonding and bridging. Bonding capital is acquired through development of connections in groups that have high levels of similarity in demographics and attitudes - the kind you might build by, say going to church. Bridging capital is created when people from different backgrounds (race, religion, class) connect - something that might happen at a town hall meeting say. The problem with personalisation is that by construction it reduces opportunities to see the world through the eyes of people who don't necessarily look like us. It reduces bridging capital and that exactly the kind of social capital we need to solve wider problems that extend beyond our narrow and short term self interests.

1.4.4 Disinformation

In June 2016, it was announced that Britain would be leaving the EU. 33.5 million people voted in the referendum of which 51.9% voted to leave. The decision that will impact the UK for, not just a term, but generations to come rested on less that 2% of voters. Ebbw Vale is a small town in Wales where 62% of the electorate (the largest majority in the country) voted to leave. The town has a history in steel and coal dating back to the late 1700's. By the 1930's the Ebbw Vale Steelworks was the largest in Europe by volume. In the 1960's it employed some 14,500 people. But, towards the end of the 1900's, after the collapse of the UK steel industry, the town suffered one of the highest unemployment rates in Britain. What was strange about the overwhelming support to leave was that Ebbw Vale was perhaps one of the largest recipients of EU development funding in the UK. A £350m regeneration project funded by the EU replaced the industrial wasteland left behind when the steelworks closed in 2002 with The Works (a housing, retail and office space, wetlands, learning campus and more). A further £33.5 in funding from the European Social Fund paid for a new college and apprenticeships, to help young people learn a trade. An additional £30 million for a new railway line, £80 million for road improvements and shortly before the vote a further £12.2 million for other upgrades and improvements were all from the EU.

When journalist Carole Cadwalladr returned to the small town where she had grown up to report on why residents had voted so overwhelmingly in favour of leaving the EU, she was no less confused. It was clear how much the town had benefited from being part of the EU. The new road, train station, college, leisure centre and enterprise zones (flagged an EU tier 1 area, eligible for the highest level of grant aid in the UK), everywhere she went she saw signs with proudly displayed EU flags saying so. So she wandered around town asking people and was no less perplexed by their answers. Time and time again people complained about immigration and foreigners. They wanted to take back control. But the immigrants were nowhere to be found, because Ebbw Vale had one of the lowest rates of immigration in the country. So how did this happen? How did a town with hundreds of millions of pounds of EU funding vote to leave the EU because of immigrants that didn't exist? In her emotive TED talk[7], Carole shows images of some the adverts on Facebook, people were targeted with as part of the leave campaign (see Figure 1.5). They were all centred around a lie - that Turkey was joining the EU.

Most people in the UK saw adverts on buses and billboards with false claims that the National Health Service (NHS) would have an extra £350 million a week if we left the EU. Those adverts circulated in the, open for everyone to see, making it possible to debate and debunk them in the mainstream media. The same cannot be said for the adverts in Figure 1.5. They were targeted towards specific individuals, as part of an evolving stream of information displayed in their Facebook 'news' feed. The leave campaign paid Cambridge Analytica (a company that had illegally gained access to the data of 87 million Facebook users) to identify individuals that could be manipulated into voting leave. In the UK, spending on elections in the is limited by law as a means to ensure fair elections. After a nine month investigation, the UK's Electoral Commission confirmed these spending limits had been breached by the leave campaign. There are ongoing criminal investigations into where the funds for the campaign originate (overseas funding of election campaigns is also illegal) but evidence suggests ties with Russia. Brexit was the precursor to the Trump administration winning the US election just a few months later that year. The same people and companies used the same strategies. It's become clear that current legislation protecting democracy is inadequate. Facebook, was able to quietly profit from politically motivated money without recognizing any responsibility to disclose the transactions or carry out due diligence on where the money came from. Five years later, the full extent of

Figure 1.5: Targeted disinformation adverts shown on Facebook[7].

the disinformation campaign on Facebook has yet to be understood. Who was shown what and when, how people were targeted, what other lies were told, who paid for the adverts or where the money came from.

Since then deep learning technology has advanced to the point of being able to pose as human in important ways that risk enabling disinformation not just through targeted advertising but machines impersonating humans. GANs can fabricate facial images, videos (deepfakes) and audio. Advancements in language models (Open AIs GPT-2 and more recently GPT-3) are capable of creating lengthy human like prose given just a few prompts. Deep learning now provides all the tools to fabricate human identities and target dissemination of false information at scale. There are growing concerns that in the future, bots will drown out actual human voices.

1.4.5 Harms of allocation

An allocative harm happens when a system allocates or withholds an opportunity or resource. Systems that approve or deny credit allocate financial resources; systems that decide who should and should not see adverts for high paying jobs allocate employment opportunities and systems that determine who will make a good tenant allocate housing resources. Harms of allocation can lead us to challenge the justice and fairness of specific determinations and outcomes. They happen as a result of discrete decisions at a given point in time, the immediate impact of which can be quantified. The methods in this book are designed to measure and mitigate harms of allocation in machine learning systems. Increasingly however, machine learning systems are not just affecting us through allocation but are shaping our view of the world and society at large by deciding what we do and don't see. Harms that are far more difficult to quantify.

1.4.6 Harms of representation

Harms of representation occur when systems enforce the subordination of groups through characterizations that affect the perception of them. In contrast to harms of allocation, harms of representation have long-term effects on attitudes and beliefs. They create identities and labels for humans, societies and their cultures.

Harms of representation don't just affect our perception of each other, they affect how we see ourselves. They are difficult to formalise and in turn difficult to quantify but the effect is real.

The surgeon's dilemma

A father and his son are involved in a horrific car crash and the man died at the scene. But when the child arrived at the hospital and was rushed into the operating theatre, the surgeon pulled away and said: "I can't operate on this boy, he's my son". How can this be?

Did you figure it out? How long did it take? There is, of course, no reason why the surgeon couldn't be the boy's mother. If it took you a while to figure out, you're not alone. More than half the people presented with this riddle struggle, and that includes women. The point of this riddle is to demonstrate the existence of unconscious bias. Representational harms are insidious. They silently fix ideas in peoples subconscious about what people of a particular gender, nationality, faith, race, occupation and more, are like. They affect our perception of world and they set boundaries for ourselves and other people. In her 2017 NIPS keynote, Kate Crawford described five types of harms of representation:

Stereotyping

Stereotyping occurs through excessively generalised portrayals of groups. In 2016 the Oxford English Dictionary was publicly criticised[16] for employing the phrase "rabid feminist" as a usage example for the word rabid. The dictionary included similarly sexist common usages for other words like shrill, nagging and bossy. But even before this, historical linguists observed that words referring to women undergo pejoration (when the meaning of a word deteriorates over time) far more often than those referring to men[19]. Examples include words such as 'mistress' (once simply the female equivalent of 'master' meaning a woman with authority), 'hussy' (once a neutral term describing the head of a household), 'madam' (once simply the female equivalent of 'sir', a woman of high rank) and the list goes on.

Unsurprisingly then, gender stereotyping is known to be a problem in natural language processing systems. In 2016 Bolukbasi et al. showed that word embeddings exhibited familiar gender biases in relation to occupations[11]. By performing arithmetic on word vectors, they were able to uncover relationships such as

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

In 2017 Caliskan et al. found that Google Translate contained similar gender biases.[12] In their research they found that "translations to English from many gender-neutral languages such as Finnish, Estonian, Hungarian, Persian, and Turkish led to gender-stereotyped sentences". So for example when they translated Turkish sentences with genderless pronouns: "O bir doktor. O bir hemişre." the resulting English sentences were: "He is a doctor. She is a nurse." They performed these types of tests for 50 occupations and found that the stereotypical gender association of the word almost perfectly predicted the resulting pronoun in the English translation.

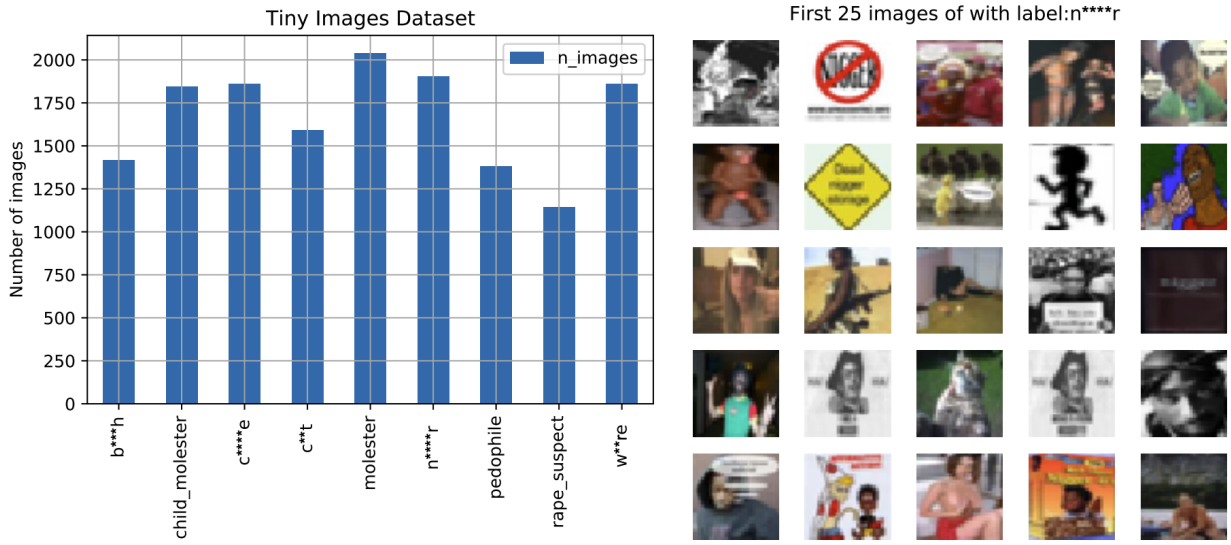
Recognition

Harms of recognition happen when groups of people are in some senses erased by a system through failure to recognise. In her TED Talk, Joy Buolamwini, talks about how as an undergraduate studying computer science she worked on social robots. One of her projects involved creating a robot which could play peek-a-boo, but she found that her robot (which used third party software for facial recognition) could not see her. She was forced to borrow her roommate's face to complete the project. After her work auditing several popular gender classification packages from IBM, Microsoft and Face++ in the project Gender Shades[5] in 2017 and seeing the failure of these technologies on the faces of some of the most recognizable Black women of her time, including Oprah Winfrey, Michelle Obama, and Serena Williams, she was promted to echo the words of Sojourner Truth in asking "Ain't I a Woman?". Harms of recognition result in failures to see the humanity in people.

Denigration

In 2015, much to the horror of many people, it was reported that Google Photos had labelled a photo of a Black couple as Gorillas. It's hard to find the right words to describe just how offensive an error this is. It demonstrated how a machine, carrying out a seemingly benign task of labelling photos, could deliver an attack on a person's human dignity.

In 2020, an ethical audit of several large computer visions datasets[17], revealed some disturbing results. TinyImages (a dataset of 79 million 32 x 32 pixel colour photos compiled in 2006, by MIT's Computer Science and Artificial Intelligence Lab for image recognition tasks) contained racist, misogynistic and demeaning labels with corresponding images. Figure 1.6 shows a subset of the data found in TinyImages. The problem,



(a) Class-wise counts of the offensive classes

(b) Samples from the class labelled `n****r`

Figure 1.6: Subset of data in TinyImages exemplifying toxicity in both the images and labels[17].

unfortunately, does not end here. Many of the datasets used to train and benchmark, not just computer vision but natural language processing tasks, are related. Tiny Images was compiled by searching the internet for images associated with words in WordNet (a machine readable, lexical database, organised by meaning, developed at Princeton), which is where TinyImages inherited its labels from. ImageNet (widely considered to be a turning point in computer vision capabilities) is also based on WordNet and, Cifar-10 and Cifar-100 were derived from TinyImages.

Vision and language datasets are enormous. The time, effort and consideration in collecting the data that forms the foundation of these technologies (compared to that which has gone into advancing the models built on them), is questionable to say the least. Furthermore a dataset can have impact beyond the applications trained on it, because datasets often don't just die, they evolve. This calls into question the technologies that are in use today, capable of creating persistent representations of our world, and trained on datasets so large they are difficult and expensive to audit.

And there's plenty of evidence to suggest that this is a problem. For example, people have found that Google searches were more likely to return personalised advertisements for arrest records searches that were suggestive of an arrest record[21]. Suggestive in the sense that they claim to have arrest records specifically for the name that you searched, regardless of whether they do in reality have them. As well as resulting in allocative harms for people applying for jobs for example, this is denigrating. Google's Natural Language API for sentiment analysis also is known to have problems. In 2017, it was assigning negative sentiment to sentences such as "I'm a jew" and "I'm a homosexual" and "I'm black"; neutral sentiment to the phrase

"white power" and positive sentiment to the sentences "I'm christian" and "I'm sikh".

Under-representation

In 2015, the New York Times reported, that "Fewer women run big companies than men named John", despite this Google's image search still managed to under-represent women in search results for the word "CEO". Does this really matter? What difference would an alternate set of search results make? A study the same year found that "people rate search results higher when they are consistent with stereotypes for a career, and shifting the representation of gender in image search results can shift people's perceptions about real-world distributions."[15].

Ex-nomination

Ex-nomination occurs through invisible means and affects people's views of the norms within societies. It tends to happen through mechanisms which amplify the presence of some groups and suppress the presence of others. The cultures, beliefs, politics of ex-nominated groups over time become the default. The most obvious example is the ex-nomination of Whiteness and White culture in western society, which might sound like a bizarre statement - what is White culture? But such is the effect of ex-nomination, you can't describe it, because it is just the norm and everything else is not. Richard Dyer in his book White examines the reproduction and preservation of whiteness in visual media over five centuries, from the depiction of the crucifixion to modern day film. It's should hardly come as a surprise then that facial recognition software might not see black faces and more often incorrectly identify the gender of black women. Or that an image of generative model called Pulse converted a pixelated picture of Barack Obama, into a high-resolution image of a white man.

The ex-nomination of White culture is evident in our language too, in terminology like whitelist and white lie. If you look up white in dictionary and or thesaurus and you'll find words like innocent and pure, light, transparent, immaculate, neutral. Doing the same for the word black on the other hand, returns very different associations, dirty, soiled, evil, wicked, black magic, black arts, black mark, black humour, blacklist and black is often used as a prefix in describing disastrous events. A similar assessment can be made for gender with women being under-represented in image data and feminine versions of words more often undergoing pejoration (when the meaning or status of a word deteriorates over time). Consider the words mistress (once simply the female equivalent of master, now used to describe a woman in a relationship with a man married to another); madam (once simply the female equivalent of sir, now also used to describe a woman who runs a brothel); hussy (once a neutral term for the head of a household, now used to describe a woman who has sexual relationships with multiple partners); and governess (female equivalent of governor, later used to describe a woman responsible for the care of children).

Members of ex-nominated groups experience a kind of privilege that it is easy to be unaware of. It is a power that comes from being the norm. They have advantages that are not earned, outside of their financial standing or effort, that the 'equivalent' person outside the ex-nominated group would not. Their hair type, skin tone, accent, food preferences and more are catered to by every store, product, service and system and it cost less to access them; they see themselves represented in the media and are more often represented in a positive light; they are not subject to profiling or stereotypes; they are more likely to be treated as individuals rather than as representative of (or as exceptions to) a group; they are more often humanised - more likely to be be given the benefit of the doubt, treated with compassion and kindness and thus recover from mistakes; they are less likely to be suspected of crimes; more likely to be trusted financially; they have greater access to opportunities, resources and power and are able to climb financial, social and professional ladders faster. The advantages enjoyed by ex-nominated groups accumulate over time and compound over generations.

Summary

Machine learning

- A model is a simplified representation of the real world. Machine learning models are trained on historic data and best capture the dense part of the distribution. When faced with rare or unprecedented events they will struggle to perform well. Obtaining adequately rich and relevant data is a major limitation for most machine learning models.
- At almost every major life event, going to university, getting a job, buying a house, getting sick, decisions are increasingly being made by machines. By construction, these models encode existing societal biases. They not only proliferate but are capable of amplifying them and are easily deployed at scale. Understanding the shortcomings of these models and ensuring such technologies are deployed responsibly are essential if we are to safeguard social progress.

Discrimination, bias, fairness and ethics

- Building machine learning systems ethically is not about finding the perfect answer every time but rather expanding our perspectives on the technology we develop. It's looking for the cracks before deploying systems, preventing the foreseeable failures and doing the best we can on the ones we didn't see coming.
- The standard, approach to training a model, which is essentially to minimise the aggregate error on the training data, is loosely justified in a utilitarian sense, in that we choose the decision process which maximises accuracy (minimises the probability of error) for the greatest number of people (everyone in the training data).
- Principles of Justice as Fairness:

  1. **Liberty principle:** Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which is compatible with the same scheme of liberties for all;
  2. **Equality principle:** Social and economic inequalities are to satisfy two conditions:
     (a) **Fair equality of opportunity:** The offices and positions to which they are attached are open to all under conditions of fair equality of opportunity;
     (b) **Difference principle** They must be of the greatest benefit to the least-advantaged members of society.

  The principles of justice as fairness are ordered by priority so that fulfilment of the liberty principle takes precedence over the equality principles and fair equality of opportunity takes precedence over the difference principle. In contrast to utilitarianism, justice as fairness introduces a number of constraints that must be satisfied for a decision process to be fair. Applied to a machine learning one might interpret the liberty principle as a requirement of some minimum accuracy level (maximum probability of error) to be set for all members of the population, even if this means the algorithm is less accurate overall. Parallels can be drawn here in machine learning where there is a trade-off between fairness and utility of an algorithm.
- Anti-discrimination laws were born out of long-standing, vast and systemic discrimination against historically oppressed and disadvantaged classes. Such discrimination has contributed to disparities in all measures of prosperity (health, wealth, housing, crime, incarceration) that persist today.
- Legal liability for discrimination against protected classes may be established through both disparate treatment and disparate impact. Disparate treatment (also described as direct discrimination in Europe) refers to both formal differences in the treatment of individuals based on protected characteristics, and the intent to discriminate. Disparate impact (also described as indirect discrimination in Europe) does not consider intent but is concerned with policies and practices that disproportionately impact protected classes.
- Just as the meaning of fairness is subjective, so too is the interpretation of anti-discrimination laws. Two conflicting interpretations are anti-classification and anti-subordination. Anti-classification is a weaker

interpretation, that the law is intended to prevent classification of people based on protected characteristics. Anti-subordination is the stronger interpretation that anti-discrimination laws exist to prevent social hierarchies, class or caste systems based on protected features and, that it should actively work to eliminate them where they exist.

Association paradoxes

- Identifying bias in data can be tricky. Data can be misleading. An association paradox is a phenomenon where an observable relationship between two variables disappears or reverses after controlling for one or more other variables. In such cases in order to know if the marginal or partial relationships are relevant, one must understand the causal nature of the relationships. Association paradoxes can also occur for non-collapsible measures of association. Collapsible measures of association are those which can be expressed as the weighted average of the partial measures.

Harms of unfair bias

- It is important to be cautious in describing machine learning algorithms as objective. Algorithms trained on data are exposed to bias since data is produced by a necessarily subjective set of decisions. The consistency of algorithms in decision making compared to humans (who make decisions on a case by case basis) is often described as a benefit, but it's their very consistency that makes them dangerous - capable of discriminating systematically and at scale.
- Classification creates a sense of order and understanding. It enables us to formulate problems neatly and solve them. But classifying people can also have negative consequences too. It inevitably has the effect of reducing people labels; labels that can result in people being treated as members of a group, rather than individuals.
- Personalisation algorithms that shape our perception of the world in a way that covertly mirror our beliefs can have the effect of decreasing bridging capital which is important in solving global problems.
- Targeted political advertising and technologies that enable machines to impersonate humans are powerful tools that can be used as part of orchestrated campaigns of disinformation that manipulate perceptions at an individual level and yet at scale. They are capable of causing great harm to political and social institutions and pose a threat to security.
- An allocative harm happens when a system allocates or withholds an opportunity or resource. Harms of representation occur when systems enforce the subordination of groups through characterizations that affect the perception of them. In contrast to harms of allocation, harms of representation have long-term effects on attitudes and beliefs. They create identities and labels for humans, societies and their cultures. Harms of representation affect our perception of each other and even ourselves. Harms of representation are difficult to quantify. Some types of harms of representation are, stereotyping, (failure of) recognition, denigration, under-representation and ex-nomination.

References

[1] Griggs v. Duke Power Co., 401 U.S. 424, 1971. https://en.wikipedia.org/wiki/Griggs_v._Duke_Power_Co.

[2] Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 1989. https://en.wikipedia.org/wiki/Wards_Cove_Packing_Co._v._Atonio.

[3] Grutter v. Bollinger, 539 U.S. 306, 2003. https://en.wikipedia.org/wiki/Grutter_v._Bollinger.

[4] Ricci v. DeStefano, 557 U.S. 557, 2009. https://en.wikipedia.org/wiki/Ricci_v._DeStefano.

[5] *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, volume 81. Proceedings of Machine Learning Research, 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[6] General Data Protection Regulation (GDPR): (EU) 2016/679 Recital 71, May 2018. https://gdpr-info.eu/recitals/no-71/.

[7] *Facebook's role in Brexit - and the threat to democracy*. TED, 2019. https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, March 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[9] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *Calif Law Rev.*, 104:671–732, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

[10] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187, Issue 4175:398–404, February 1975. https://science.sciencemag.org/content/187/4175/398.

[11] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. https://arxiv.org/abs/1607.06520.

[12] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186, April 2017. https://researchportal.bath.ac.uk/en/publications/semantics-derived-automatically-from-language-corpora-necessarily.

[13] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, April 2016. https://www.bloomberg.com/graphics/2016-amazon-same-day/.

[14] President Lyndon B. Johnson. Speech to a joint session of congress on march 15, 1965. *Public Papers of the Presidents of the United States*, I, entry 107:281–287, March 1965. http://www.lbjlibrary.org/lyndon-baines-johnson/speeches-films/president-johnsons-special-message-to-the-congress-the-american-promise.

[15] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal representation and gender stereotypes in image search results for occupations. *ACM*, 2015. https://www.csee.umbc.edu/~cmat/Pubs/KayMatuszekMunsonCHI2015GenderImageSearch.pdf.

[16] Emer O'Toole. A dictionary entry citing 'rabid feminist' doesn't just reflect prejudice, it reinforces it. *The Guardian*, January 2016. https://www.theguardian.com/commentisfree/2016/jan/26/rabid-feminist-language-oxford-english-dictionary.

[17] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision?, 2020. https://arxiv.org/abs/2006.16923.

[18] John Rawls. *Justice As Fairness: a Restatement*. Harvard University Press, Cambridge, Mass., 2001. (1921-2002).

[19] David Shariatmadari. Eight words that reveal the sexism at the heart of the english language. *The Guardian*, January 2016. https://www.theguardian.com/commentisfree/2016/jan/27/eight-words-sexism-heart-english-language.

[20] E Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, Series B, 13:238–241, March 1951. https://www.jstor.org/stable/2984065.

[21] Latanya Sweeney. Discrimination in online ad delivery. *SSRN*, 2013. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240.

[22] C. Wang, Q. Zhang, W. Liu, Y. Liu, and L. Miao. Facial feature discovery for ethnicity recognition. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018. https://espace.curtin.edu.au/handle/20.500.11937/71484.

[23] Y. Wang and M. Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 2018. https://www.gsb.stanford.edu/faculty-research/publications/deep-neural-networks-are-more-accurate-humans-detecting-sexual.

[24] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images, 2017. https://arxiv.org/abs/1611.04135.

Chapter 2

Ethical development

This chapter at a glance

- The machine learning cycle - feedback from models to data
- Bias measurement and interventions in the model development life cycle
- A taxonomy of common causes of bias
- Model governance and ethical risk management

In this chapter, we transition to a more systematic approach to understanding the problems and solutions of fairness in decisions making systems. In later chapters we will look at different measures of unfairness and mitigation techniques but before we discuss and analyse these methods we review some more practical aspects of responsible model development. None of the interventions to mitigate bias that we will talk about in this book are capable of fixing a poorly formulated, discriminatory machine learning problem so we'll start by looking at the machine learning cycle and discuss the importance of how a model is used. We'll discuss where in the machine learning model development cycle bias metrics and interventions fit and we'll classify the most common causes of bias, identifying the parts of the workflow to which they correspond. A model in itself is not the source of unfair or illegal discrimination, models are developed and deployed by people as part of a process. In order to address the problem of unfair bias we need to look at the whole system, not just the model. The methods we discuss in this book for measuring and mitigating bias won't remedy negligent deployment or management of a machine learning system. Where models can be harmful we should expect to have processes in place that aim to avoid common, foreseeable or catastrophic failures. We'll discuss how to take a proactive rather than reactive approach to managing risks associated with models.

In this chapter, we will present problems and interventions schematically. We provide a set of references for building, reviewing and monitoring machine learning solutions that aim to avoid the common pitfalls that result in biased models. We take a high enough view that the discussion remains applicable to many machine learning applications. The specifics of the framework, can be tailored for a particular application. Indeed the hope is that the resources in this chapter can be used as a starting point for data science teams that want to develop their own set of standards. Together we will progress towards thinking critically about the whole machine learning cycle, development, validation and monitoring of machine learning systems. By the end of this chapter we will have a clearer picture of what due diligence in model development and deployment (with respect to fairness and discrimination) might look like, from a practical perspective.

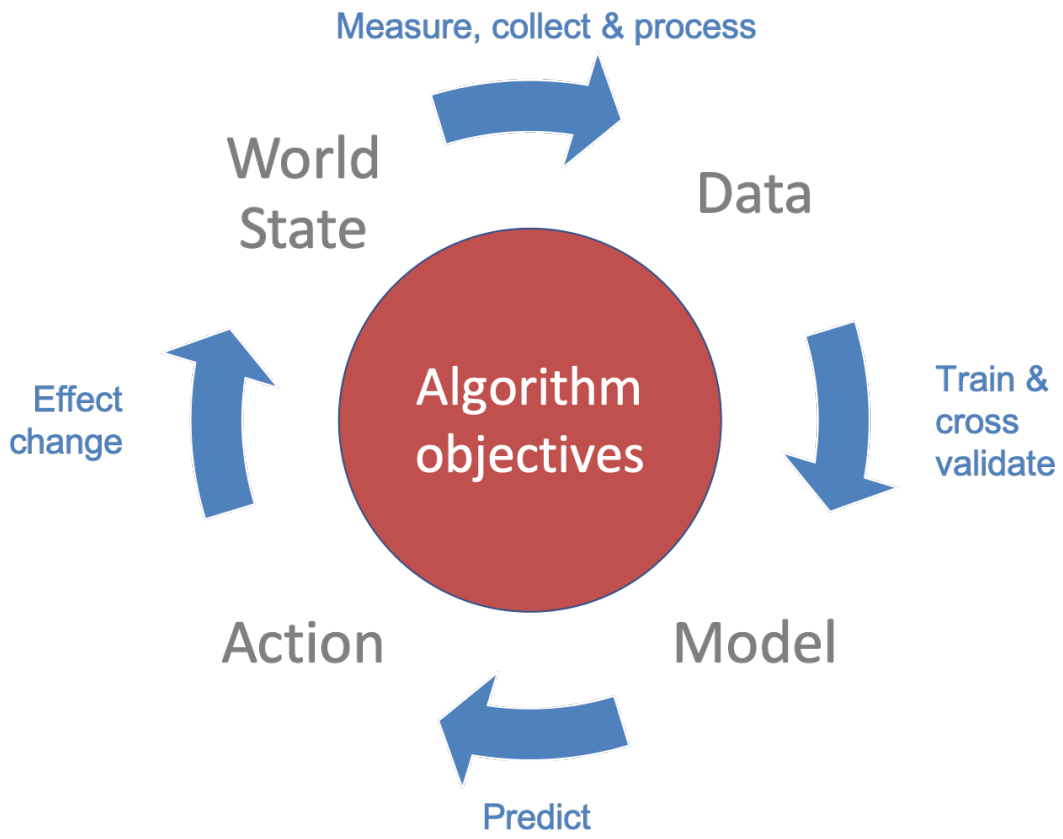


Figure 2.1: The machine learning cycle

2.1 Machine Learning Cycle

It's important to realise that machine learning solutions can have longterm and compounding effects on the world around us. In this section we analyse the machine learning cycle in a variety of different examples to breakdown the mechanisms that determine the nature and magnitude of the effect. In Figure 2.1 we present a high-level depiction of the interaction between a machine learning solution and the real world. At the core of our cycle is some set of objectives. These can be achieved in a myriad of different ways. The translation of these objectives, into a tractable machine learning problem, consists of a series of subjective decisions; what data we train a model on, what events we predict, what features we use, how we clean and process the data, how we evaluate the model and the decision policy are all choices. They determine the model we create, the actions we take and and finally the cycle we end up with.

The most familiar parts of the cycle to most developers of machine learning solutions are on the right hand side; processing data, model selection, training and cross validation and prediction. Each action taken on the basis of our model prediction creates a new world state, which generates new data, which we collect and train or model on, and around it goes again. The actions we take based on our model predictions define how we use the model. The same model used in a different way can result in a very different feedback cycle.

Notice that the world state and data are distinct nodes in in the cycle. Most machine learning models rely on the assumption that the training data is accurate, rich and representative of the population, but this is often not the case. Data is a necessarily subjective representation of the world. The sample may be biased, contain an inadequate collection of features, subjective decisions around how to categorise features into groups, systematic errors or be tainted with prejudice decisions. We may not even be able to measure

the true metric we wish to impact. Data collected for one purpose is often reused for another under the assumption that it represents the ground truth when it does not.

Figure 2.1 shows the most complete version of the cycle which might not always be the case. In some cases the cycle might be broken. An example might be if one used a dataset to train a model, made decisions based on it and never re-trained the model. Though our model may continually impact the world state, resulting changes are not fed back into the model. The implicit assumption being that the world is static and that taking actions based on the model output does not alter the world state.

2.1.1 Feedback effects

In cases where the ground truth (target variable) assignment systematically disadvantages certain classes, actions taken based on predictions from models trained on the data can reinforce the bias and even amplify it. Similarly, decisions made on the basis of results derived from machine learning algorithms trained on data that under or over-represents certain classes, can have feedback effects that further skew the representation of those classes in future data. The cycle of training on biased data (which justifies inaccurate beliefs), taking actions in kind, and further generating data that reinforces those biases can become a kind of self-fulfilling prophecy. The good news is that just as we can create pernicious cycles that exaggerate disparities, we can create virtuous ones that have the effect of reducing disparities. Let's take two examples to illustrate the point.

In the United States, predictive policing has been implemented by police departments in several states including California, Washington, South Carolina, Alabama, Arizona, Tennessee, New York and Illinois. Such algorithms use data on the time, location and nature of past crimes, in order to determine how and where to patrol and thus improve the efficiency with which resources are allocated. The aim on the face of it is a noble one - to deter crime from happening - prevention over cure, so to speak. However there is a major flaw with these algorithms. The data used to train these algorithms is not of where crimes occurred (such data does not exist), but rather where there have been previous arrests. A proxy target variable (arrests) is used in absence of available data for the desired target variable (crime). Racial disparities in policing in the US is a well publicised problem (see Figure 2.2). In 2015, an analysis by The Hamilton Project found that at the state level, Blacks were 6.5 times as Whites to be incarcerated for drug-related crimes[2]. Taking actions based on predictions from an algorithm trained on arrest data will amplify existing disparities between under and over-policed neighbourhoods which correlate with race.

As a comparative example, let's consider car insurance. It is well publicised that car insurance companies discriminate against young male drivers as they are deemed to be at higher risk of being involved in accidents.

Age discrimination in car insurance

Take a moment to think about why this is legal or considered fair (despite age and gender being legally protected characteristics in the countries where these insurance companies operate).

Insurance companies act on risk predictions by determining the price of insurance at an individual level - the higher the risk, the more expensive the cost of insurance. What is the feedback effect of this on the data? Of course young men are disadvantaged by having to pay more, but one can see how this pricing structure acts as an incentive to drive safely. It is in the drivers interest to avoid having an accident that would result in an increase in their car insurance premiums. For a high risk driver in particular, an accident could potentially make it prohibitively expensive for them to drive. The feedback effect on the data would be to reduce the disparity in incidents of road traffic accidents among high and low risk individuals.

Along with the difference in the direction of the feedback effects in the examples given above, there is another important distinction to be made in terms of the magnitude of the feedback effect. This is related to how much control the institution making decisions based on the predictions, has over the data. In the predictive policing example the data is entirely controlled by the police department. They decide where to police, who to arrest determining who is and isn't in the data. They produce the training data, in its

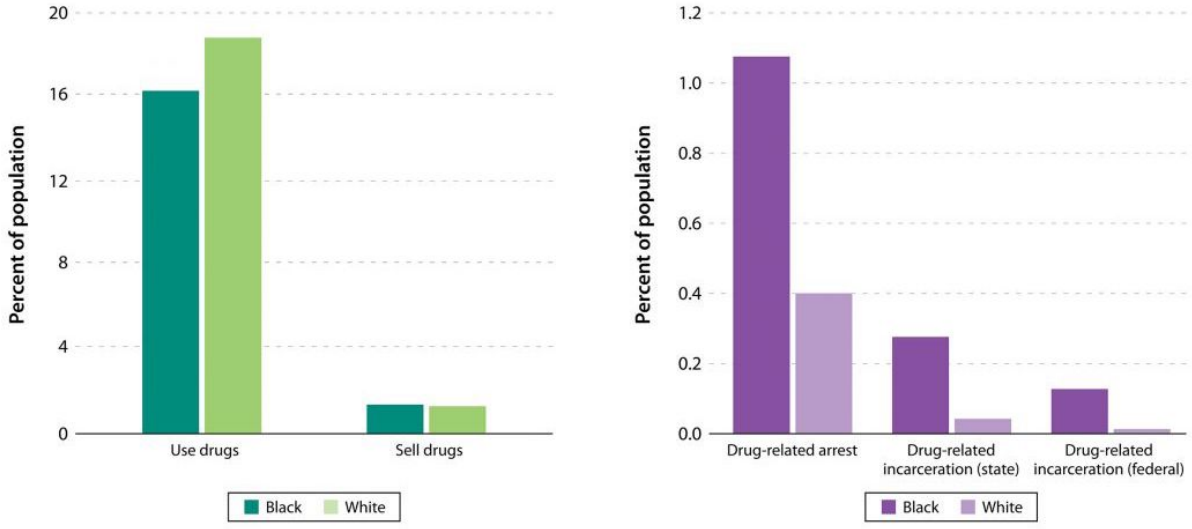


Figure 2.2: Rates of drug use and sales compared to criminal justice measures by race[2].

entirety, as a result of their actions. Consequently, feedback effect of acting on predictions based on the data is strong and capable of shifting the entire distribution of data generated in the future. Insurance companies by comparison, have far less influence over the data (consisting individuals involved in road traffic accidents). Though they can encourage certain driving behaviours through pricing, they do not ultimately determine who does and does not have an accident, even if that insurance company dominates the market. As such, feedback effects of risk-related pricing in car insurance are likely to be less strong in comparison.

2.1.2 Model application and feedback

A crucial part of responsible model development is understanding and communicating its limitations - what the model can and (perhaps more importantly) cannot be used for. Clearly defining what a model will be used for is an important part of enabling effective and focussed analysis and testing of it. In addition considering the cases for which the model is not suitable and documenting them can prevent models from being used inappropriately.

The idea that the use case for a product, tool or model should be well understood before release; that it should be validated and thoroughly tested for that use case and further that the potential harms caused (even for unintended uses) should be mitigated is not novel. In fact, many industries have safety standards set by a regulatory body that enshrine this idea in law. The motor vehicle industry has a rich history of regulation aimed at reducing risk of death or serious injury from road traffic accidents that continues to evolve today. In the early days, protruding knobs and controls on the dash would impale people in collisions! It was not until the 1960s that seatbelts, collapsing steering columns and head restraints became a requirement. Since then, safety testing and requirements have continued to expand (albeit slowly more recently thanks to reduced funding of regulatory bodies) to including rear brake lights, a variety of impact crash tests, ISOFIX child car seat anchors among others. There are many more such examples across different industries but it is perhaps more instructive to consider an example that involves the use of models.

Derivatives are financial contracts (products) that result in payments dependent on future events. The details, such as amounts and events that lead to payments are described in the contract. For example, a contract that gives the holder the right, but not obligation, to buy 100 units of S&P 500 (the underlying asset) in a years time (the expiry), at a price fixed today (strike price), is the simplest example of a derivative known as a call option. Derivatives pricing can get complicated quickly as the events which result in payments

become more elaborate. In derivatives markets, it is known that models are product specific. A model that is suitable for pricing one financial instrument will not necessarily be appropriate to price another, just because the underlying asset (in our example the S&P 500) is the same (even though it's the underlying assets behaviour that is being modelled). A derivatives pricing models, for a variety of reasons often only capture limited characteristics of the underlying instrument. For this reason, banks that trade derivatives validate models specifically for the instruments they will be used to price and document their testing. Furthermore they must track their product inventory (along with the model used to price each contract) in order to be able to show that they are not using models to price products for which they have not been approved (validated as part of their internal model review process).

Though machine learning models are not (yet) regulated in the same way, it's easy to see the parallels. Clear consideration of the use case is not just about making sure that the model performs well enough for that use case. How a model is used ultimately determines the actions that are taken off the back of the resulting predictions and thus the nature of the feedback that model has on future data. To illustrate the value of use case specific testing, we return to COMPAS[12]. The system was not designed to be used in sentencing. Tim Brennan (the co-founder of Northpointe and co-creator of its COMPAS risk scoring system) testified at Paul Zilly's appeal hearing that they "wanted to stay away from the courts". Documentation[18] for the software (dated 2015) describes it as a risk and needs assessment and case management system. It talks about it being used "to inform decisions regarding the placement, supervision and case management of offenders" and probation officers using the recidivism risk scales to "triage their case loads", it does not anywhere discuss its use in sentencing.

Could this model, developed as a case management tool for probation officers be useful in determining the length of an offenders sentence? Napa County, California, uses a similar risk scoring system in the courts. What's interesting is that a Superior Court Judge, who trains other judges in evidence-based sentencing, cautions colleagues in their interpretation of the scores. He outlines a concrete example of where the model falls short. He says "A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job. Meanwhile, a drunk guy will look high risk because he's homeless. These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be."[12]

So, let's look at the feedback of the model on the data for a variety of different use cases. Propublica's review looked at recidivism risk for more than 10,000 criminal defendants in Broward County, Florida[11]. Their analysis found the distributions of risk scores for Black and White defendants to be markedly different, with White defendents being more likely to be scored low-risk - see Figure 2.3.

Comparing predicted recidivism rates for over 7,000 of the defendants with the rate that actually occurred over a two-year period, they found the accuracy of the algorithm in predicting recidivism for Black and White defendants to be similar (59% for White and 63% for Black defendants), however the errors revealed a different pattern. They found that Blacks were almost twice as likely as Whites to be labelled as higher risk but not actually re-offend. The errors for White defendants were in the opposite direction; while being more likely to be labelled as low-risk, they more often went on to commit further crimes. See Table 2.1. Let's assume for a moment that the disparity in recidivism risk between Black and White defendants to be

Table 2.1: COMPAS comparison of risk score errors for White versus Black defendants

Error type	White	Black
Labelled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labelled Lower Risk, But Did Re-Offend	47.7%	28.0%

accurate, i.e. that Black defendants do in fact re-offend more often than White defendants, let's consider the feedback effect on the racial disparity for a range of different use cases for the model.

In the courts, when the COMPAS recidivism risk score has been used to determine sentence length as a means to reduce crime - the higher the risk, the longer the sentence. What's the effect on the disparity? Current research suggests that "The longer and harsher the prison sentence – in terms of less freedom,

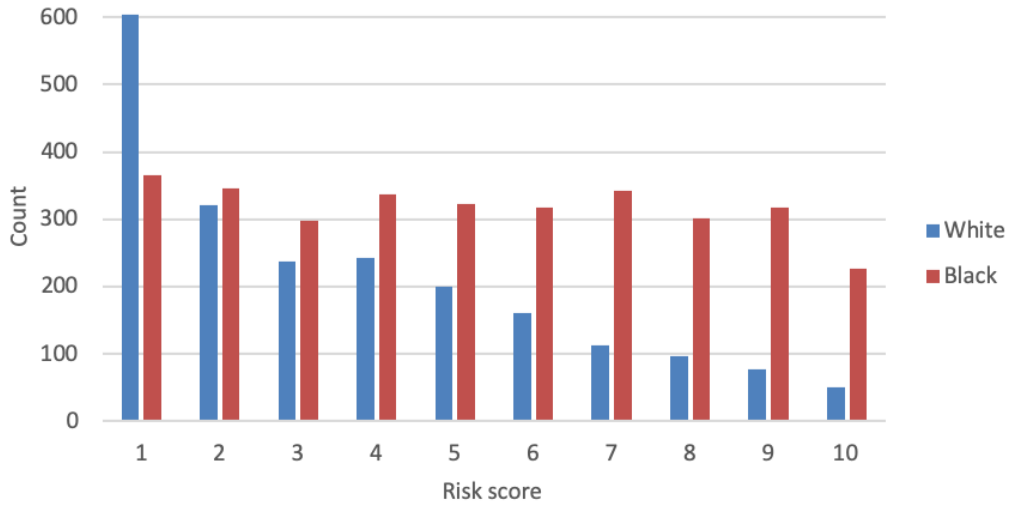


Figure 2.3: Comparison of recidivism risk scores for White and Black defendants[11]

choice and opportunity for safe, meaningful relationships – the more likely that prisoners' personalities will be changed in ways that make their reintegration difficult and that increase their risk for re-offending"[10]. The feedback effect of subjecting high-risk defendants to longer sentences would be to further exacerbate the racial disparity, since Black defendants would more often suffer longer sentences resulting in greater difficulty reintegrating into society and increase the liklihood of re-offending upon release. What about reducing crime? What does the research say about that? It is the certainty, rather than severity of punishment that acts as a deterrent to crime[17]. Long-term sentences are particularly ineffective for drug crimes as drug sellers are easily replaced in the community[14]. On balance, excessive incarceration has negative consequences for public safety because finite resources spent on prison are diverted from policing, drug treatment, preschool programs, or other interventions that might produce crime-reducing benefits.

Let's consider another use case for recidivism risk scoring. The US has the highest rate of incarceration in the world, at 0.7% of the population[19]. It's higher than countries with authoritarian governments, those that have recently been locked in civil war and those with murder rates more than twice that in the US. Comparing with countries that have stable democratic governments, the incarceration rate in the US is more than 5 times that of its closest peer - the UK. The US spends $57 billion a year on housing more than 2.2 million people in prison[13], almost half of which are private companies that spend significant sums on lobbying the federal government for policies that would further increase incarceration. Some have advocated for the use of risk scores in sentencing in order to reduce the rate of incarceration, the idea being that if the risk scores are low then defendants can be spared prison time. What would the feedback effect on the observed racial disparity for this use case be? The action is a little different, we 'reward' low-risk rather than 'punish' high-risk scoring defendants, but the feedback effect would be similar - to increase the existing racial disparity since White defendants will more often be spared prison time.

Finally, what if the software was used as a way to distribute limited rehabilitation resources, allocating them to those defendants that that were deemed to be at the highest risk of re-offending and thus the most in need of help. What would the feedback effect on the racial disparity in this case be? It's easy to see that using the model in this way would reduce the disparity. Of course we have made numerous simplifying assumptions in our analysis of the feedback on the disparity; that rehabilitation consistently reduces the risk of recidivism (regardless of the crime), that recidivism risk is calculated on a long time horizon, that the relationship between sentence length and recidivism is monotonic. Without getting into the weeds, the point here is simply that the same model can have very different feedback cycle if used in a different way.

How a model is used is important. The question to ask is, does the action taken on the back of the model serve to push extremes to the centre, or push them further apart? What you have to understand to answer the question, will depend on the specifics of the problem.

2.2 Fairness and bias interventions

We will get in to the details of a range of methods for measuring and mitigating bias and unfairness in the chapters that follow, but first we take a look at where in the machine learning model development workflow these metrics and interventions fit. Figure 2.4 depicts the model development, deployment and monitoring life cycle at a high level.

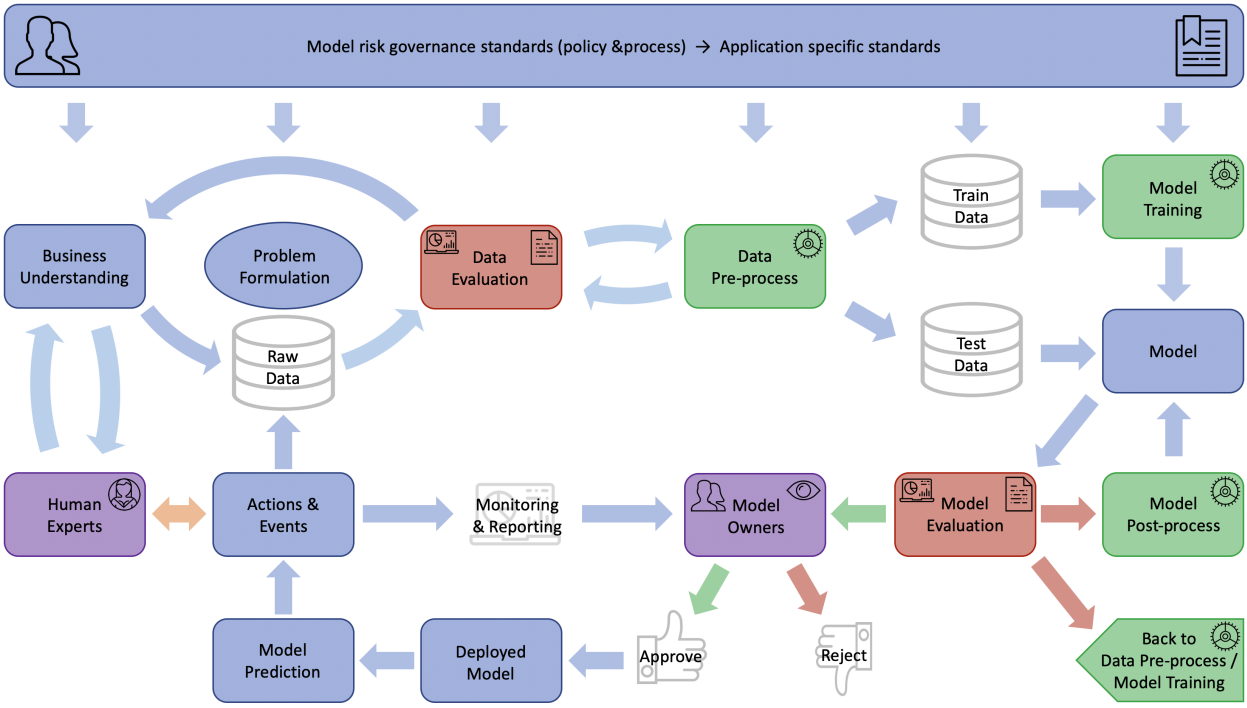


Figure 2.4: Fairness aware machine learning system development, deployment and management workflow. Note, the term 'human experts' is context dependent here, in relation to business understanding it is any person with valuable insight or perspective on the product, in relation to decision making it will be a person capable of adjudicating on the evidence to determine the best action.

2.2.1 Metrics

Bias and fairness metrics are essentially calculated on data. There are two stages at which we'll be interested in measuring bias and or fairness.

Model input The training data. Node labelled *data evaluation* in Figure 2.4.

Model output The predictions produced by our model. Node labelled *model evaluation* in Figure 2.4

It's worth highlighting at this point that these two need not be the same. That is, bias and fairness metrics calculated on the training data will generally be different to those calculated on the model output

for a variety of reasons. By comparing the two, we can evaluate how well the model is replicating the distributions it's trained on, specifically the biases in the data.

2.2.2 Mitigation techniques

There are essentially three stages at which one can intervene to mitigate bias when developing a machine learning model and we categorise them accordingly:

Pre-processing techniques modify the historical data on which the model is trained, the idea being that fair/unbiased data will result in a fair/unbiased model once trained. Node labelled *data pre-process* in Figure 2.4.

In-processing techniques alter the training process or objective in order to create model with fairer/less biased predictions. Node labelled *model training* in Figure 2.4.

Post-processing techniques take a trained model and modify the output such that the resulting predictions are fairer/less biased. Node labelled *model post-process* in Figure 2.4.

2.2.3 Can a model be biased?

A common belief shared by many well known and respected professional machine learning scholars and practitioners is that bias comes from the data not the model. We've spoken about numerous examples of biased models in this book already. So what do people mean when they say this? In more theoretical disciplines a model is interpreted as being the parametric form (so that different values of the parameters don't change the model, for example, the term *linear model* describes a family of models). More practical disciplines view a model more as a black box - provided with input, the model returns output (if the parameters change, the output changes and so must the model). From a practical perspective then it's clear that a model can discriminate since if the data discriminates so does the model (assuming the model fits the data reasonably well).

The notion that bias is an artefact of data, rather than a model is at best counterproductive and at worst misleading. It implies that models and data are independent when they are not. Model development is an iterative process. The modelling choices we make will depend on the data, and model results will in turn influence our choices regarding the training data. Treating the data and model as independent entities diminishes the responsibility of model developers in addressing the problem of biased and unfair algorithms and ignores the very practical nature of developing models to solve real world problems. If we are developing a model as a means to understand the world then it may make sense for utility to be our sole objective because the goal is simply to find out how well the model fits observations. For an application that will be used to make predictions that have an impact in the real world, the objectives should extend beyond utility to other metrics. Those other metrics will depend on the problem, in part 2 of this book we'll explore a range of metrics that measure notions such as fairness and diversity.

2.3 Common causes of bias

There are many ways in which machine learning solutions can produce biased predictions. In this section we present a taxonomy of common causes with examples. In addition we relate the causes in the taxonomy to the corresponding stages of the model development and deployment life cycle so that it may be used as a reference when designing machine learning solutions in order to avoid common pitfalls. We follow the taxonomy described in the excellent paper by d'Alessandro et. al.[], giving additional examples where valuable. Table 2.2 summarises the taxonomy of common causes of bias in a machine learning system.

Figure 2.5 shows the causes of bias in the context of the model development and deployment workflow, indicating both the stages of the workflow to which they relate along with their categorisation within the

Table 2.2: Taxonomy of common causes of bias in machine learning models[6].

Component	Root Cause	Issue Type	Issue Description
Model Issue	Data Issue	Discrimination in data	
		Sample Bias	Under-representation of protected class
			Over-representation of protected class
		Low support	
	Misspecification	Target variable	Target variable subjectivity
			Proxy target variable learning
			Heterogeneous target variable
		Features	Inclusion of protected features without control variables
			Inclusion of protected feature proxies (redlining)
		Cost function	Failure to specify asymmetric error costs
			Omitted discrimination penalties
System / Process Issue	Failure to validate	Data appropriateness and preparation	
		Modelling approach, implementation and evaluation	
	Failure to monitor	Poor feedback loop	
	Non-deference to human expert		

taxonomy. The central row of Figure 2.5 shows a flow diagram representing the model development and

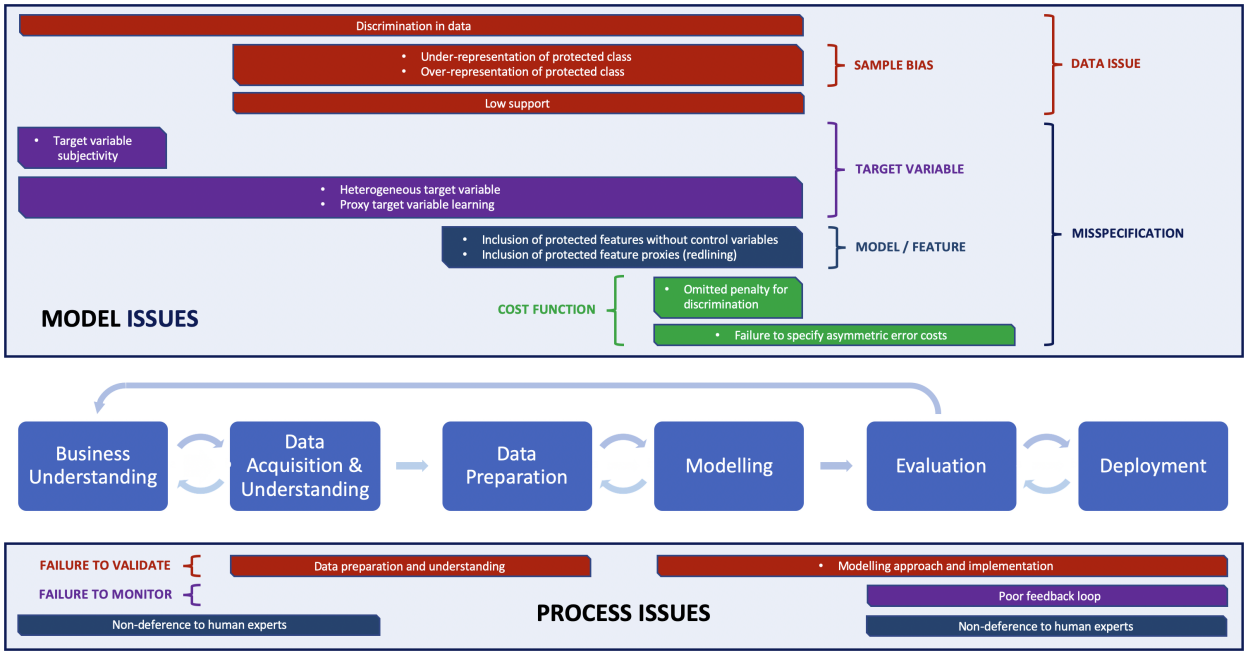


Figure 2.5: Taxonomy of common causes of bias in machine learning models together with the stages of the model development and deployment life cycle they relate to.

deployment life cycle (at a higher level than that shown in Figure 2.4). For the purposes of the taxonomy, we distinguish the machine learning model from the larger machine learning system, of which it is a component.

1. **The machine learning model** is defined as the function mapping f from features $(\boldsymbol{X}, \boldsymbol{Z})$ to predictions $\hat{Y}$.

2. **The machine learning system** includes the wider infrastructure, processes and policies around development, deployment and monitoring of the model.

From Table 2.2 we can see that we categorise the causes of bias as one of two types, those that relate directly to the model (shown in the box above the flow diagram in Figure 2.5) and those that arise as a result of failures in the model development and deployment process (shown in the box below the flow diagram in 2.5). The causes of bias within each of these categories are further displayed in boxes, the width of which indicate the stages of the workflow to which they relate. Curly brackets are used to indicate the categorisation of the cause in the taxonomy. In the sections that follow we go through the taxonomy of common causes of bias and provide examples.

2.3.1 Modelling issues

In this section we discuss common causes of discrimination that relate directly to the model. We categorise these as originating from one of two sources:

1. **Data issues** refer to bias that arises as a direct result of issues with the data
2. **Misspecification** refers to issues relating to misspecification of the underlying problem in the modelling of it.

The latter is an extension of the notion of model misspecification in statistics where the functional form of a model does not adequately reflect reality.

Data issues

Within data issues we organise potential sources of bias into three types

1. **Discrimination in data:** Disparities between protected classes, in either outcomes or accuracy and completeness are observed in the data.
2. **Sample Bias:** The data is not representative of the population - protected classes are under or over-represented.
3. **Low support:** Minority classes naturally have fewer data points to train on.

1. Discrimination in data The most obvious way that bias can enter a machine learning model is through the data it is trained on. Discrimination against protected classes is a part of our history and a reality of our society and the data reflects this. One of the goals of training is to in fact replicate the distribution of the data it is trained on. If the training data contains biases we naturally would expect these to propagate through to the predictions of the resulting algorithm. The most obvious way in which this is seen is through a disparity in outcomes between different subgroups of the population. Take medical data where racial and gender disparities in diagnosis and treatment are well publicised as the *health gap*. In particular, there is a growing body of research across the US and Europe that exposes systematic undertreatment and misdiagnosis of pain in women and Black patients.

- A 1990 study that looked at medication administered to postoperative coronary artery bypass graft patients differed significantly by gender. Men were more frequently administered pain medication compared to female patients and that female patients were more frequently administered sedative medication[5].
- A 2008 study found that women were less likely to receive opioid analgesia than men when reporting acute abdominal pain at an emergency department[15].
- A 2001 paper surveys extensive research showing that despite evidence that women report more severe, frequent and longer duration of pain than men, they are treated less aggressively for it[9].
- Research in the US in 2012 showed that Black patients are 22% less likely than Whites to be given pain medication and 29% less likely that Whites to be treated with opioids (despite prescription drug abuse being more prevalent among White Americans).

- A study in 2015 showed that this disparity in pain treatment extended to children with White children being three times more likely to be treated with opioids for appendicitis.
- In a 2016 US study, 200 medical students and residents were given a series of statements about biological differences between White and Black people (such as "Black people's skin is thicker than White people's skin") and asked to say whether they were true or false. Half the respondents marked false statements as true. Later when given case studies for Black and White patients reporting pain, those participants that held those beliefs rated the Black persons pain lower and made less accurate treatment recommendations[8].

A more subtle way in which discrimination can manifest in data is where the accuracy and completeness of records are correlated to sensitive features. This can happen for example, where there are geographic disparities in services provided by institutions (which tend to be correlated with race) or where institutions systematically fail to produce accurate and timely records for protected groups. A good example of the latter is again provided by medical data. Misdiagnosis leads to longer lags between the reporting of symptoms and accurate diagnosis. Given the above we might expect greater delays in accurate diagnosis for protected classes. A study of 12,000 rare disease patients across Europe in 2009 which found significant diagnosis delays for women compared to men[1]; for example 12 compared to 20 months for Crohn's disease (despite being more common among women) and 16 compared to 4 years for Ehlers-Danlos syndrome. Systematic delays in diagnosis for protected groups mean that for any given snapshot in time, the medical records for more frequently misdiagnosed groups are less accurate.

2. Sample Bias Another, way in which the training data can result in biased predictions is where there is an imbalance in the prevalence of classes of sensitive features in the training data. All too often training data is assumed to be representative of the population that the model will serve, where it is not. Both under and over representation can be disadvantageous.

Under-represented classes are exposed to higher error rates a problem which arises as a result of 'low support', that is a smaller pool of data points to train the model on. Let's briefly examine why this might happen. In training we are trying to find the set of model parameters which minimizes some aggregate measure of model error on training data (the 'cost' function). If the contribution to the cost from majority classes dominates (which it will without intervention) the algorithm is naturally incentivised to focus learning characteristics of majority classes as a means of reducing the cost faster. Majority classes being more richly represented by the data results in the model being better able to generalize to them.

In 2017, Joy Buolamwini and Timnit Gebru conducted a seminal study Gender Shades[3] which found that systems sold by IBM, Microsoft, and Face++ had as high as a 34.4% accuracy gap in gender classification, between lighter-skinned males and darker-skinned females. Error rates on lighter-skinned males did not exceed 1%. While we can't be sure without seeing the data these algorithms were trained on, a lack of adequate representation of darker-skinned females in the training data is more than likely at least a contributing factor to such disproportionate error rates. Dataset bias (an over reliance on a single dataset for testing and benchmarking), is a well known issue in computer vision tasks.

One of the drivers behind big data initiatives is the plummeting cost of collection and storage data. Companies and institutions are able to train models that better target individuals, reducing costs and boosting profits. However, often data collection methods fail to adequately represent historically disadvantaged classes of people who are often less engaged in certain data generating ecosystems. A good example of this, given by Barocas & Selbst[4] is that of the phone app Street Bump, which was developed by the City of Boston to reduce the cost and time taken to find (and consequently repair) pot holes. The app uses data generated by the accelerometers and GPS of Boston residents' smart phones as they drive. Once a pothole is located it is automatically added to the city's system to schedule a repair. One can see easily see how this method of data collection might fail to adequately capture data from poorer neighbourhoods, where car and smart phone ownership are less prevalent; neighbourhoods which probably correlate with race and are already likely to suffer from lack of investment.

Over-representation of classes can also result in bias through increased scrutiny. Predictive policing discussed earlier provides an example of this. But in practice any process (algorithmic or otherwise) which seeks to identify negative behaviour in which disproportionate resource is allocated to some subgroup will

result in disproportionately more instances of that negative behaviour being observed among members of that group. The result is induced correlation in the data between the negative behaviour and over-scrutinised class even if in reality there is none. If an algorithm that seeks to identify negative behaviours is used to determine where to allocate resource to identifying those behaviours in the future, an implicit assumption is made that where no observation was made the negative behaviour did not occur. If observations are not uniformly distributed across the population, the result will be a pernicious cycle that continually amplifies the association between the negative behaviour and members of the over-scrutinised class.

3. Low support It is worth briefly mentioning that the problem of low-support need not be a result of under-representation of classes. This is a particular problem for individuals belonging to multiple disadvantaged classes for example Black women.

Misspecification

We categorise mispecification as one of three types, based on which aspect of the modelling they relate to

1. **Target variable:** the events our model predicts
2. **Features:** the features we use in our model
3. **Cost function:** how we evaluate our model in training

1. Target variable We discuss three common issues related to the target variable definition:

- Target variable subjectivity
- Proxy target variable learning
- Heterogeneous target variable

One of the challenges in developing a machine learning is the translation of the underlying problem by definition of a target variable - something which can be observed, measured and recorded accurately. While there are relatively uncontentious examples that machine learning solutions lend themselves well to (spam detection in email or on-base and slugging percentage for player valuation in Major League Baseball). Often however, this translation is non-trivial and rather subjective. It requires the model developer to interpret some problem and translate objectives and requirements into a target variable that can be measured or is accessible in order to set up a tractable machine learning problem.

For the purpose of illustration, consider the problem of determining which candidates from a pool would make 'good' employees. How do we determine what a good employee looks like? What might our target variable be? Do they produce results faster? Do they stay with the company longer? Do they take less annual leave? Perhaps they have better annual performance ratings? Each of the above variables would likely exhibit correlations with protected classes that would result in different kinds of biases infiltrating our algorithm. The data might show for example that men tend to stay with the company longer but this might simply be because the workplace is more hostile towards women and bear no relation to how good the employee actually is.

Of course one of the problems in the above example (in addition to the subjectivity in the choice of target variable) is that, like with the predictive policing example, data on the variable we want doesn't really exist so we use a proxy that we believe to be correlated with the variable we wish to affect. In 2018, Amazon was forced to scrap a recruitment tool it spent four years developing. The algorithm rated resumes of potential employees and was trained on 10 years worth of resumes submitted by job applicants. The exact details of the algorithm were not publicised but based on what we know about the training data, it is likely that the proxy variable they used was some measure of how the candidates had performed in the hiring process, and potentially beyond, in the past.

Another common issue is the use of a heterogeneous target variable, where a range of different events are coarsely grouped into a single outcome. This might happen for example where the event of particular interest is rare and by including more events in the target the predictive accuracy of the model increases as

it has more data to learn from. D'Alessandro et. al[6] provide a useful example in predictive policing where the model developer is initially interested in predicting violent crime but ends up incorporating petty crimes (which happen much more frequently) in the target variable in pursuit of a more accurate model.

2. Features We discuss two issues related to feature selection:

1. Inclusion of protected features without control variables
2. Inclusion of protected feature proxies (redlining)

In an ideal world we would train a machine learning model on a sufficiently large dataset consisting of a rich set of features that actually influence the target variable rather than simply being correlated to it. More often than not, the reality is rather different. Comprehensive data can be expensive and difficult to collect. Factors that influence the target variable might not be easily measured or be measurable at all, while data containing more erroneous indicators might simply be cheaper to obtain or more readily available. This is a common way in which bias against protected classes can enter our model.

We discuss two cases. In the first we include a protected feature because it appears to be predictive of the target variable. Of course using protected characteristic when building a model would invariably come with disparate treatment liability so is not a problem one is typically faced with but we take this opportunity to recall the importance of controlling for confounding variables, in drawing conclusions about relationships between features from observational data (see section 1.3).

The second (more common) case is one where we do not use protected features in the model but use features which are predictive of them. Take another example in the context of hiring. Historically employers have taken the reputation of the university that applicants graduated from as a strong indicator of the calibre of the candidate. But many of the most reputable universities have very low rates of non-White/Asian students in attendance. A hiring process which is strongly influenced by the university from which the applicant graduated, can erroneously disadvantage racial groups that are less likely to have attended them. While the university an applicant graduated from, might correlate to some degree with success in a particular role it is not in itself the driver. An algorithm that directly takes into account the skills and competencies required for the role would be more predictive and simultaneously less biased. Given the cost of collecting comprehensive data, one might argue that higher error rates for some classes would be financially justified (rational prejudice).

3. Cost function We discuss two issues related to the cost function specification:

1. Failure to specify asymmetric error costs
2. Omitted discrimination penalties

A critical consideration in how we specify our model is the cost function. It is how we evaluate our model in training and essentially determines the model (parameters) we end up with. The cost function can be interpreted as an expression of our model objectives and so provides a natural route to addressing discrimination concerns. A common failure in the design of classification models is proper accounting of the costs of the different types of classification errors (false negative versus false positives). If the harm caused by the different types of misclassification are asymmetric, the cost matrix should reflect this asymmetry.

More broadly (for both regression and classification), it is important to consider the contribution from each sample in the training data to the cost function in training. Upsampling (or simply upweighting, depending on the learning algorithm you are using) is a valuable tool to keep in mind and can alleviate a number of the issues discussed above, that are common sources of bias. Let's take the issue of low support. By upsampling minority classes, one can increase the importance of reducing errors for those data points, relative to other more abundant classes, during learning. Though it's worth noting that it cannot resolve issues relating to a lack of richness of representation for classes with low support. Another case in which upsampling can help is that discussed in relation to definition of a heterogeneous target variable. By upsampling data points that correspond to the primary event of interest (violent crime in the example we discussed above), one can again increase the importance of the model fitting to those data points.

For an algorithm that solves a problem in a regulated domain, it would make sense for the absence of discrimination to be a model objective along with utility. This can be achieved by use of a penalty term in the cost function which relates to discrimination in the resulting predictions (just as we have terms that relate to the error or overfitting). Essentially the idea is similar to that of regularisation to avoid overfitting. We introduce an additional hyper-parameter to tune, which represents the strength of the penalty for discrimination in our cost. We will discuss this and upsampling in more detail when we discuss bias mitigation techniques, in part three of the book.

2.3.2 System / process issues

In this section we discuss failures that relate to the larger machine learning system (rather than more directly to the model). Any institution (no matter how big or small) that uses models in a setting where there are real world consequences of the model being wrong should have processes in place to ensure the risks are assessed, understood and mitigated. We talk about the kinds of processes that form part of a responsible development and deployment in the section that follows but cover them at a high level here for completeness. We consider three ways in which process interventions play a critical role in avoiding bias in machine learning systems.

1. **Validation** and testing of the approach, data and model pre-deployment
2. **Monitoring** of the model post-deployment
3. **Keeping human experts in the loop** in development and post-deployment

1. Failure to validate

We discuss two important aspects of the modelling which should be appropriately validated pre-deployment

1. Data appropriateness and preparation
2. Modelling approach, implementation and evaluation

A common failure that results in biased algorithms is inadequate review and validation of a machine learning systems pre-deployment. Testing for discrimination should be built into the model/product development process. Interpretation and translation of ethical standards into data/model/product requirements, consideration of the resulting machine learning cycle and determining appropriate metrics should be an integral part of the problem formulation and solution design.

An assessment should be made with regards to how appropriate the data is for the model use case. Understanding the provenance of the data (who collected it, how it was collected and for what purpose) is important. Is it representative of the population the model built on it intends to serve? Exploratory data analysis (EDA) should include understanding if there is bias and or discrimination in the data. In particular understanding how is the target variable distributed for different subgroups of the population and what the nature of the resulting machine learning cycle might be for the intended and unintended use cases. Is there strong correlation between protected features and other variables? Data preparation methods should also be analysed for potential biases they may introduce. Methods for measuring and mitigating bias as part of data preparation may be used and should also be analysed during development and model review.

The process of testing and analysing model output for performance should also include corresponding analysis for discrimination and fairness. How are predictions and errors distributed for different subgroups of the population? How does the model output distribution differ from the training data? What are the limitations of the model? What should it not be used for and why? Again methods for measuring and mitigating bias may also be used and should be properly analysed.

2. Failure to monitor

Post-deployment monitoring is an important part of responsible model development and deployment. Analysis should not stop once the model is deployed. Decisions on what to monitor and necessary feedback

mechanisms should be determined during development. It's important to understand if the model is performing in line with expectations (based on pre-deployment testing and analysis). Are there changes in distributions? Is the data coming out of the model more or less biased than the data going in? This should be of particular concern where the actions taken based on predictions determine the composition of future data. Mechanisms for model monitoring should account for asymmetries in information. Post-deployment monitoring encompasses a range of other processes (in addition to direct tracking of the model described above) such as risk materiality tracking and periodic re-reviews. We will discuss in more detail later in the chapter.

3. Non-deference to human experts

In section 1.3 we discussed the importance of domain knowledge in interpreting data to avoid erroneous conclusions about relationships between variables. We highlight the need for consultation with human experts at the development stage in understanding the problem and data. In addition, different stakeholders in the risk of a model will have different perspectives and it is of important to understand these in assessing ethical risks.

Given that models are simplified representations of real world systems and we know that they will make errors, responsible development should build in processes for anticipating and dealing with those cases and where necessary deferring to the judgement of a human expert.

2.4 Responsible development and deployment

In the previous section we talked about some of the most common causes of bias in machine learning algorithms and linked them to the relevant parts of the machine learning model development and deployment cycle so we can take timely action to avoid them. In this section we take another step towards the solution. We discuss what a fairness aware aware development, deployment and management workflow for a machine learning system (MLS) looks like in more detail. For the most part, the workflow is not much different to one that is concerned with effective (machine learning specific) model risk management. One that acknowledges that models are fallible and accordingly has standards, processes and monitoring in place to prevent forseeable failures and mitigate the associated risks. The main difference is that we consider bias and ethical risk as key components of the risks that must be managed. Of course model performance (accuracy) is an important part of being fair (it's hard to think of a model that is no better than guessing as fair) but viewing model evaluation through an ethical lens requires a more holistic assessment of the system, it's purpose, reliability and impact; not just for the business but for all who are exposed to or affected by it.

In Figure 2.4, overarching the whole process is a set of model governance standards. These define the full details of the workflow and will depend on the particular application being developed, among other things. Below we will describe the kinds of activities that happen within the workflow that play a pivotal role in responsible development. Some of the activities in the flow diagram use icons to highlight their importance in creating fairer, more reliable machine learning systems. For the purpose of discussing elements of the workflow in Figure 2.4, we shall divide it into three components:

1. Model governance standards
2. Pre-deployment
3. Post-deployment

Each component encompasses a subset of nodes in the workflow. The composition of each of these components will vary greatly for different applications and organisations. For example, one would expect that an application that advises recidivism risk for sentencing should go through a much more robust testing and validation process than a photo filter that figures out where to put animated ears given a facial image; and a company like Google will have a very different risk profile, resources and infrastructure compared to say a start-up with six people with one deployed model. Nevertheless, there are components of responsible development, deployment and management of models that can be applied to MLS. We discuss these here.

2.4.1 Model governance standards

The concept of model governance, though relatively new in machine learning circles, is not a new one. For financial institutions (which depend on vast numbers of proprietary models), model governance is a central part of model risk management and there are well established frameworks for handling it. In addition, the regulatory landscape of financial modelling is also considerably more mature. Given this, it is instructive to look at how such institutions manage model risk.

So what does responsible and ethical MLS development and deployment look like? The biggest difficulty in answering this question is that there is no one size fits all answer. The answer to the question will depend on a whole multitude of factors and can be approached from many different perspectives, to mention a few:

- **Domain:** Different domains will have different legal and ethical concerns for example employment or housing versus entertainment or targeted advertising.
- **The number and complexity of the models being used by the business:** An organisation that uses hundreds of models and composes them to create new products would benefit greatly from infrastructure and methodologies for measuring the materiality of the associated risks that would enable prioritisation of work related to mitigating them. In contrast, for a business based on a single model, this would be less of a concern.
- **Cost of errors:** Where the potential harm caused by model errors is high, pre-deployment testing will need to be extensive and prescribed. Well defined and mandatory processes will play an important role - checklists, contingency planning, detailed logging so post-mortems can be conducted and more.

Given this, how does one approach the problem of responsible development? Step zero is to create a set of model governance standards. The purpose of model governance standards is to clearly define and communicate what responsible model development and deployment looks like for your specific application, use case, domain, business, principles and values. It essentially documents and communicates the why, who, what and how of your model risk management approach. What are the kinds of questions we might want our model governance standards to answer?

- **Why is the work important?** What kinds of events are you trying to avoid? What legislation is the company subject to? What are the consequences of failures? What are the values of the company that you want to protect?
- **Who is responsible?** Who are the stakeholders? Who is accountable for managing the various identified risks?
- **What are they responsible for?** What are their roles? What kind of expertise are required to understand and manage the risks? What are the questions each stakeholder must answer? What are the responsibilities of those experts at the various stages of the model development and deployment life cycle? What authority do they have in relation to determining if the model is fit for deployment? Who decides what?
- **How do you manage the risk?** What are the rules, processes and requirements that ensure the companies values are maintained, people are treated fairly, the legal requirements are fulfilled and the model risks are appropriately managed? How do the stakeholders work together? For example some roles might need to be independent while others work alongside one another. How is the materiality of different risks measured and tracked? Numbers of predictions? Numbers of individuals? Consequences? Revenue? What are the requirements around training data (documentation, review, storage, privacy, consent and such)? What are the requirements around modelling (documentation, testing, explainability and such)? What are the processes and requirements around proposing, reviewing, testing, deploying, monitoring model related risks? For example, frequency of risk reviews, forums for discussion and monitoring, required logging. What are the processes and requirements in place for (specific foreseeable types of) model failures? Are there stakeholder specific templates or check-lists that ensure particular questions get answered at specific points in the model development and deployment life cycle?

The list of questions above is by no means exhaustive but a good starting point. Creating a set of model governance standards is about planning. Machine learning systems can be complicated and have many points of failure: problem formulation, the data collection, data processing, modelling, implementation, deployment. The only way to reduce the risk of failures is to be organised, deliberate and plan for them. Creating a set of standards does exactly that. Where the systems we build have real world consequences, the preparation, planning and process around development, review, analysis, deployment and monitoring of them should reflect that. Ensuring that the right questions get asked at the right time, knowing who is responsible for answering them and fixing things when they go wrong is a core part of developing and deploying models ethically.

Compliance

The benefits of having excellent model governance standards with well defined goals, processes, roles and responsibilities won't be realised if in practice they are not followed. In large organisations, this can be a challenge. The role of internal audit is to provide objective feedback on the risks, systems, processes and compliance at an executive level. From a model governance perspective the role of internal audit is to ensure that there are good processes in place and that the processes are being followed. Internal audit's role is entirely independent of the business all the way upto the executive level. All functions within the business are required to provide cooperate with internal audit and provide unfettered access to whatever information they request. Internal audit does not contribute to the improvement of or compliance to processes directly. Their role is to observe assess and report back to senior leadership. In risk management circles, internal audit are considered to be the third line of defence. We shall talk about the first and second lines shortly.

2.4.2 Pre-deployment

In this section we discuss four activities as part of creating a candidate model for deployment.

- **Problem formulation:** Translating a business objectives into a tractable machine learning problem.
- **Model development:** Developing a machine learning solution.
- **Model validation:** Independent validation of the proposed machine learning solution.
- **Model approval:** Process around accepting (or rejecting) a machine learning solution for deployment.

Problem formulation

Problem formulation is the first key step in developing a machine learning solution and an especially pivotal one in ethical risk assessment. The problem formulation stage plays perhaps the largest role in what the end product will actually be. It is the stage at which the model objectives, requirements, target variable and training data are determined and it is the stage at which the most important ethical question (whether the model should be built at all) must be answered.

A thorough examination of ethical issues demands consideration of a diversity of voices, which is well known to be lacking in technology. This is the stage at which it is important to consider who is affected by the technology, consult with them and ensure their views are understood and incorporated in the understanding of the problem and design of a potential solution. Who are the human experts? By that we mean, those who would have valuable insight and opinions on the potential harms of the model you plan on building? Who does the model advantage and who does it disadvantage? Want to use machine learning to help manage diabetes? What are the interests of the health insurance company funding the development? Have you consulted with diabetics in addition to specialist physicians? What are their concerns? What is the problem from the different perspectives? Would a model be able to help or are there simpler solutions? How and for who?

It is also the stage at which the materiality of the risk should be assessed. What's the worst that can happen? How likely is such a failure? How many people would be exposed to the model? As part of problem formulation one should examine the machine learning cycle in the context of the biases in the data

and consider the nature (direction and strength) of the feedback of resulting actions on future data. It's important to consider other ways in which the model might be used (other than that intended) and the corresponding feedback cycle in those cases. How the model might be misused?

In any system that is vulnerable to errors and less than optimal subjective decisions, pre-deployment independent review is a well established method of preventing costly foreseeable failures. Whether it's a completely new solution built from scratch or a modification to an existing solution that's being deployed, an independent review process is an important element of responsible model development. Below we describe the responsibilities of two separate roles, model development (designing a solution) and the model validation (critical assessment of the solution risks).

Model development

The model developers role is to translate the business problem into a tractable machine learning problem and create a initial model solution. They will work with the business and receive input from other necessary domain experts relevant to the application to develop a possible solution. This will include tasks such as acquiring and interpreting data that is relevant for the problem, determining a target variable, model objectives, performance measures, fairness measures and more.

The model developer should document the solution. Documentation should include for example, descriptions of the data (data provenance - how and for what purpose it was collected, by whom) and model, motivation behind subjective decisions that were made to arrive at the solution (how to process the data, what features to use/ignore, model type, cost function, sample weights, bias and success metrics), known data/model issues, how the model was tested, what it's limitations are, what it should and should not be used for with justification. The documentation of the model should provide enough detail to be able to re-implement the model, reproduce results and justify the solution approach. Documentation can be standardised using application specific templates. Recent research discusses standardisation of documentation specifically for publicly released datasets[7] and machine learning models[16].

In terms of preventing failures, the model developer would be considered the first line of defence. The responsibility of developing a model responsibly and ethically lies, in the first instance, with them. The model developer should aim to create a model they believe to be production ready. The model developer should follow the processes and fulfil the requirements specified in the model governance standards. These should include ethical as well as performance standards which must be met. If there is an accuracy requirement for example, has the performance been tested for different subgroups of the population? Have conjunctions of protected characteristics also been tested?

Model validation

As part of the pre-deployment process, the model should be reviewed. The model validator will also be a data scientist but their role is different to that of the model developer. Where the developers primary role is to develop a solution to the business problem, the role of a model validator is to critique the solution. The model validator will identify and expose issues with the problem formulation, data and data processing. They will verify the model performance metrics (error, bias, fairness), look for model weaknesses and demonstrate them through testing. This might involve independently re-implementing the model (to expose implementation errors), measuring the performance on another dataset or comparing results with a different problem formulation or model for example. The model validator might devise mitigation strategies for identified risks. Such strategies might include setting model usage limits or additional monitoring requirements. They might for example identify cases when human review is required or reject the proposed solution entirely if the risk is not acceptable. The role of the reviewer could be thought of as something akin to the hacker but with the advantage of having access to the model documentation (provided by the model developer). The model reviewer in pre-deployment acts as something of a gate keeper. The model reviewer must also document their analysis, testing and critique and recommendations regarding the solution.

The model review process acts as the second line of defence. To be effective, the model reviewer's role must be independent of the model developer's to some extent. What does independence mean? We mentioned

the distinct goals of their roles and this is important. The validator should not drive the development of a solution approach or model but instead focus on critique. In reality it's easy to see that the iterative nature of model development will mean that the criticism of the solution may get rolled into it's development, blurring the lines between critique and collaboration. For example, if the validator has suggestions for modifications to the model that improve it, it might make sense for the model developer to implement them. From an efficiency perspective, it might make sense for the solution to be reviewed at several critical stages of the development process making the overall process seem more collaborative - if there's a problem with the data that was missed, the developer might want to know before going on to build and train a model on it. One of the challenges then is how to preserve independence between the roles, and ensure that the value of having adversarial criticism in preventing failures, is not lost in collaboration. How best to preserve independence will depend on the specifics and is something that should be considered within the model governance standards. In a bank, the model developer and validator are required (by the regulator) to serve under different business functions (the trading desk versus risk management). They have different reporting lines and might be required to work in physically separate locations.

Deployment approval process

In any responsible development process where there are pre-deployment safeguards in place, one should expect there to be a process for approving deployment and that as part of that process, some proposals will be 'rejected'. We use the terminology of 'approve' and 'reject' in Figure 2.4 as a way to emphasise ownership of the decision, but clarify that 'reject' doesn't necessarily mean that the project gets abandoned and all the work that went into developing the solution is wasted. It will, in practice, often just mean that more information is required before a decision can be made or some other blocker has yet to be addressed. In a well designed development process, one would expect that a solution that would be rejected for deployment altogether, doesn't make it that far in the process. Model owners will be aware of what's in development and related issues.

So who are these model owners? As discussed above, one feature of the model governance standards is to specify accountability and this should include specifying who the model owners are. We discussed the role of a model validator as being something of a gatekeeper but there are often many people involved in the development and deployment of a machine learning system and the model governance standards should specify which of them plays what role in deciding when a solution is ready to be deployed, that is, who owns which model risk. Each of the model owners will have different (potentially conflicting) concerns. Apart from the model developer and validator, model owners might include

- The business (who will confirm the their problem has been solved and requirements met, in some situations this role might be filled by a product manager) and that they understand and are happy with the risks it presents,
- Domain experts that may have had input in the development of the solution (legal or application specific council) and/or may be responsible for dealing with cases for which the model is deemed inappropriate (a radiologist for a pneumonia detector for example),
- Engineering who might be responsible for ensuring that infrastructure (data collection, storage, post-deployment monitoring and reporting for example) requirements can be met.

As discussed earlier, different organisations and applications will carve up responsibilities differently. Above we outline the kinds of concerns model owners might have for illustration purposes. In effect, the model owners represent the different stakeholders of the risk associated with the model and collectively they are accountable, though for potentially differing aspects of it. They will together determine if the model is ready to be deployed in production. They will also be responsible for monitoring the model post-deployment, periodic re-review of the risks and failure post-mortums that determine what changes are required when issues arise, including amendments to the model governance standards themselves. The model governance standards might be interpreted as a contract between the model owners that describes their commitments, individually and collectively in managing the risk.

2.4.3 Post-deployment

Managing risk related to a model does not stop once the model has been deployed, far from it. In many ways, it's just the beginning. The pre-deployment review of the model is just the first. Thereafter periodic re-reviews of the model are a means to catch risks that may have been missed the first time around. Are people using the model in ways that were not expected? The frequency of re-reviews will depend on the risk level of the model/application in question. As model usage increases so does the associated risk. The materiality of the risk should be monitored and reported to the model owners. The world is dynamic and the risk associated with models evolves with it. Deployed models should be monitored (according to the metrics identified during reviews) to understand if they are behaving in line with expectations, or if distributions are shifting.

Processes and procedures in the event of failures should be specified as part of the model governance standards, in particular what steps should be taken by which model owner. One of the issues with machine learning solutions is that when there are failures (say, a photo or sentence is labelled in an offensive way), the easiest response is an ad hoc rule based approach to 'fixing' the very specific issue that occurred - "if this, then don't do that". But this kind of action doesn't address the root of the problem. Such an approach should likely only be the first step in taking remedial action. A failures should prompt a full re-review. Having a more robust process around dealing with failures when they occur, should mean that not only is action is taken in a timely manner, but also that meaningful changes are made as a result of them. A post-mortem should focus on understanding the weaknesses of the model governance process (not the failure of individuals) that contributed to it and appropriately prioritise any changes required to remedy them.

Summary

Machine learning cycle

- Machine learning solutions can have long-term and compounding effects on the world around us. Figure 2.1 illustrates the interaction between a machine learning solution and the real world. At the core of our cycle is some set of objectives which can be achieved in a myriad of different ways. The translation of these objectives into a tractable machine learning problem, requires a series of subjective choices. Choices around what data to train the model on, what events to predict, what features to use, how to clean and process the data, how to evaluate the model and what the decision policy should be will all determine the model we create, the actions we take and ultimately the cycle we end up with.
- Data is a necessarily subjective representation of the world. The sample may be biased, contain an inadequate collection of features, subjective decisions around how to categorise features into groups, systematic errors or be tainted with prejudice decisions. We may not even be able to measure the true metric we wish to impact. Data collected for one purpose is often reused for another under the assumption that it represents the ground truth when it does not.
- In cases where the ground truth (target variable) assignment systematically disadvantages certain classes, actions taken based on predictions from models trained on the data are capable of reinforcing and further amplifying the bias.
- Decisions made on the basis of results derived from machine learning algorithms trained on data that under or over-represents certain classes can have feedback effects that further skew the representation of those classes in future data.
- The actions we take based on our model predictions define how we use the model. The same model used in a different way can result in a very different feedback cycle.
- The magnitude of the feedback effect will depend how much control the institution making decisions based on the predictions, has over the data the training data.
- Just as we can create pernicious machine learning cycles that exaggerate disparities, we can also create virtuous ones that have the effect of reducing disparities. Therefore it's important to consider the whole machine learning cycle when formulating a machine learning problem

Fairness and bias interventions

- Figure 2.4 depicts the model development, deployment and monitoring life cycle at a high level. Bias and fairness metrics are essentially calculated on data. There are two stages at which we'll be interested in measuring bias and or fairness; the data going into our model (training data) and the data coming out of it (the predictions produced by our model). In Figure 2.4 these nodes are labelled as data evaluation and model evaluation.
- There are essentially three stages at which one can intervene to mitigate bias when developing a machine learning model labelled *data pre-process*, *model training* and *model post-process* in Figure 2.4. We categorise them accordingly:
    - **Pre-processing techniques** modify to the historical data on which the model is trained, the idea being that fair/unbiased data will result in a fair/unbiased model once trained.
    - **In-processing techniques** alter the training process or objective in order to create model with fairer/less biased predictions.
    - **Post-processing techniques** take a trained model and modify the output such that the resulting predictions are fairer/less biased.

Common causes of bias

- Table 2.2 summarises the taxonomy of common causes of bias in a machine learning system.
- Figure 2.5 shows common causes of bias in the context of the model development and deployment workflow, indicating both the stages of the workflow to which they relate and their categorisation within the taxonomy.
- For the purposes of the taxonomy, we distinguish the machine learning model from the larger machine learning system, of which it is a component.
    1. **The machine learning model** is defined as the function mapping f from features $(\boldsymbol{X}, \boldsymbol{Z})$ to predictions $\hat{Y}$.
    2. **The machine learning system** includes the wider infrastructure, processes and policies around development, deployment and monitoring of the model.

Modelling issues

We categorise common causes of discrimination that relate directly to the model as originating either from the data or misspecification of the problem:

1. **Data issues** refer to bias that arises as a direct result of issues with the data
    - **Discrimination in data:** Disparities between protected classes, in either outcomes or accuracy and completeness are present in the data.
    - **Sample Bias:** The data is not representative of the population - protected classes are under or over-represented.
    - **Low support:** Minority classes naturally have fewer data points to train on.
2. **Misspecification** refers to issues relating to misspecification of the underlying problem in the modelling of it. These can be related to different aspects of the modelling.
    - **Target variable:** defining a target variable in order to translate objectives to a tractable machine learning problem can be a non-trivial and subjective task that can introduce biases. Common issues that arise when defining a target variable include choosing a heterogeneous target variable that coarsely groups different events together or a proxy variable in place of the true target variable we wish to predict.

- **Features:** there are two possibilities here, including protected features without confounding variables and more commonly, including proxies for protected features.
- **Cost function:** Common issues include failure to specify asymmetric error costs in the cost matrix and failing to penalise discrimination.

System and process issues

Common system and process failures that aide in the creation of biased algorithms include

1. **Insufficient validation** and testing of the approach, data and model with respect to bias and fairness pre-deployment
2. **Poor monitoring** of the model post-deployment
3. **Failure to keep human experts in the loop** during development and post-deployment

Responsible development and deployment

Model governance standards

- Machine learning systems can be complicated and have many points of failure: problem formulation, the data collection, data processing, modelling, implementation, deployment. The only way to reduce the risk of failures is to be organised, deliberate and plan for them. Creating a set of standards does exactly that. They make sure the right questions get asked at the right time and that there is clarity around who is responsible for what.
- The purpose of creating a set of model governance standards is to clearly define and communicate what responsible model development and deployment looks like for your specific application, domain, business, principles and values. It essentially documents and communicates the why, who, what and how of your model risk management approach.
    - **Why is the work important?** What kinds of events are you trying to avoid? What are the consequences of failures? What are the values of the company that you want to protect?
    - **Who is responsible?** Who are the stakeholders? Who is accountable for managing the various identified risks?
    - **What are they responsible for?** What are their roles/expertise? What authority do they have in relation to determining if the model is fit for deployment?
    - **How do you manage the risk?** What are the policies, processes and requirements that ensure the companies values are maintained, people are treated fairly, the legal requirements are fulfilled and the model risks are appropriately managed? How do the stakeholders work together?
- In large companies that carry lots of model risk it can be difficult to ensure there is consistency in standards of due diligence in model development and deployment across the board. The role of internal audit is to provide independent and objective feedback on the risks, systems, processes and compliance at an executive level. From a model governance perspective they determine if that there are good processes in place and that the processes are being followed. From a risk management perspective internal audit is considered to be the third line of defence.

Pre-deployment

- **Problem formulation:** Translating a business problem into a machine learning one.
    - The problem formulation stage plays a pivotal role in what the end product will actually be. It is the stage at which the model objectives, requirements, target variable and training data are determined and it is the stage at which perhaps the most important ethical question (whether the model should be built at all) must be answered.

- Consider who is affected by the technology, consult with them and ensure their views are understood and incorporated in the understanding of the problem and design of a potential solution.
- Assess the materiality of the risk. What's the worst that can happen? How likely is such a failure? How many people are exposed to the model?
- Examine the machine learning cycle in the context of the biases in the data and consider the nature (direction and strength) of the feedback of resulting actions on future data.
- Consider other ways in which the model might be used (other than that intended) and the corresponding feedback cycle in those cases. How the model might be misused?

- An independent review process is an important element of responsible model development. This means that pre-deployment there are two separate data science roles, model development (designing a solution) and the model validation (critical assessment of the solution).
- **Model development:** The model developers role is to translate the business problem into a tractable machine learning problem and create a model solution.
  - The model developer will work with the business and receive input from other necessary domain experts relevant to the application to develop a possible solution.
  - The model developer should document the solution. Documentation should include descriptions of the data and model, justification of the approach, known issues and limitations, model testing (biases as well as performance), what the model should not be used for and why. Templates are a good way of standardising documentation.
  - In terms of preventing failures, the model developer should be considered to be the first line of defence. The responsibility of developing a model responsibly and ethically lies, in the first instance, with them.
- **Model validation:** The role of a model validator is to criticise the proposed solution.
  - The model validator will identify and expose issues with the problem formulation, data and data processing. They will verify the model performance metrics (error, bias, fairness), look for model weaknesses and demonstrate them through testing. They may also devise mitigation strategies for identified risks.
  - The role of the reviewer might be thought of as a hacker but with the advantage of having access to the model documentation (provided by the model developer). They also act as a gate keeper.
  - The model reviewer must also document their analysis, testing and critique and recommendations regarding the solution.
  - The model reviewer should be viewed as the second line of defence.
- **Model approval:** The model owners collectively determine if a solution is ready for deployment.
  - Model owners act as the final stage gate keepers before deployment. They will each have been involved in different aspects of the development and deployment of the machine learning system.
  - In effect, the model owners represent the different stakeholders of the risk associated with the model and collectively they are accountable, though for potentially differing aspects of it.
  - They will also be responsible for monitoring the model and risk materiality post-deployment and ensuring that periodic re-review, failure processes and post-mortems occur and are effective.
  - The model governance standards might be interpreted as a contract between the model owners that describes their commitments, individually and collectively in managing the risk.

Post-deployment

- **Monitoring of deployed models:** The world is dynamic and the risk associated with models evolves with it. Deployed models should be monitored to understand if they are behaving in line with expectations. The metrics which should be reported to model owners should be identified pre-deployment by the model developer and validator.

- **Risk materiality tracking:** As model usage increases so does the associated risk. As part of monitoring, metrics that give an indication of the risk associated with the model is should be reported to the model owners.
- **Periodic re-review:** The pre-deployment independent review of the model is just the first. Thereafter, periodic re-reviews of the model are a means to catch risks that may have been missed the first time around. The frequency of re-reviews will depend on the risk level of the model/application in question.
- **Failure event process:** Processes and procedures in the event of failures should be specified as part of the model governance standards, in particular what steps should be taken by which model owner. Having a robust process around dealing with failures when they occur should mean that action is taken in a timely manner and that meaningful changes are made as a result of them.
- **Failure post-mortems:** A post-mortem should focus on understanding the weaknesses of the model governance process (not the failure of individuals) that contributed to it and appropriately prioritise any changes required to remedy them.

References

[1] The voice of 12,000 patients: Experiences and expectations of rare disease patients on diagnosis and care in europe, 2009. https://www.eurordis.org/IMG/pdf/voice_12000_patients/EURORDISCARE_FULLBOOKr.pdf.

[2] Rates of drug use and sales, by race; rates of drug related criminal justice measures, by race, 2015. Source: BLS n.d.c.; Carson 2015; Census Bureau n.d.; FBI 2015.

[3] *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, volume 81. Proceedings of Machine Learning Research, 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[4] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *Calif Law Rev.*, 104:671–732, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

[5] Karen L. Calderone. The influence of gender on the frequency of pain and sedative medication administered to postoperative patients. *Sex Roles*, 23:713–725, 1990. https://link.springer.com/article/10.1007/BF00289259.

[6] Brian d' Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134, June 2017. https://arxiv.org/abs/1907.09013.

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2020. https://arxiv.org/abs/1803.09010.

[8] Kelly M. Hoffman, Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301, 2016. https://www.pnas.org/content/113/16/4296.

[9] Diane E. Hoffmann and Anita J. Tarzian. The girl who cried pain: A bias against women in the treatment of pain. *SSRN*, 2001. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=383803.

[10] Christian Jarrett. How prison changes people. *BBC Future*, May 2018. https://www.bbc.com/future/article/20180430-the-unexpected-ways-prison-time-changes-people.

[11] Jeff Larson. Propublica analysis of data from broward county, fla. Technical report, ProPublica, March 2016. https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb.

[12] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, March 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[13] Bryan Lufkin. The myth behind long prison sentences. *BBC Future*, May 2018. https://www.bbc.com/future/article/20180514-do-long-prison-sentences-deter-crime.

[14] Marc Mauer. Long-term sentences: Time to reconsider the scale of punishment. *The Sentencing Project*, November 2018. https://www.sentencingproject.org/publications/long-term-sentences-time-reconsider-scale-punishment/#:~:text=There%20are%20several%20reasons%20for,sentences%20add%20little%20to%20the.

[15] Esther H. Chen MD, Frances S. Shofer PhD, Anthony J. Dean MD, Judd E. Hollander MD, William G. Baxt MD, Jennifer L. Robey RN, Keara L. Sease MaEd, and Angela M. Mills MD. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Academic Emergency Medicine*, 15:414–418, May 2008. https://pubmed.ncbi.nlm.nih.gov/18439195/.

[16] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, January 2019. http://dx.doi.org/10.1145/3287560.3287596.

[17] Daniel S. Nagin. Deterrence in the twenty-first century: A review of the evidence. *Crime and Justice*, 42, May 2018. https://www.journals.uchicago.edu/doi/abs/10.1086/670398.

[18] Northpointe. *Practitioners Guide to COMPAS Core*, March 2015. https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf.

[19] Peter Wagner and Wendy Sawyer. States of incarceration: The global context. *Prison Policy Initiative*, June 2018. https://www.prisonpolicy.org/global/2018.html.

Chapter 3

Group Fairness

This chapter at a glance

- Group fairness metrics
- Using AIF360 to compute group fairness metrics
- Incompatibility of group fairness criteria

The term *group fairness* is used to describe a class of metrics that are used to measure discrimination or bias in a given decision process (algorithmic or human). In this chapter we will introduce the different types of group fairness metrics in a structured way. We will compare and analyse the different types of metrics, in terms of their meaning and implications. For classification problems, we'll derive results that show how the various fairness metrics discussed, can in fact be incompatible in certain cases; that is to say, they cannot be satisfied simultaneously except in some degenerate cases. By the end of this chapter, we will have a deep understanding of the various group fairness metrics, thus enabling us to make educated choices about which metrics to (and not to) use for any given problem. In addition to discussing group fairness, we'll get started with AIF360. We'll introduce it and use it to compute the group fairness metrics in Jupyter Notebook on a dataset. Let's get started!

We begin with an overview. Group fairness metrics all stem from the same high level notion of fairness; the idea that some *property* should be balanced (or equal) across different *subgroups* of a population. The *subgroups* are determined by the values of *protected characteristics* such as gender or race. We also describe these as *sensitive features*. Partitions of the population could be defined by a single protected characteristic or logical conjunctions of multiple sensitive features. For example, if we were considering both race and gender simultaneously, one group of the partition might be Black women, another White men, and so on. The *property* we'll be interested in balancing will be some statistical measure; the particular kind, will depend on our beliefs about what fairness should mean in the context of the problem. Group fairness criteria can be broadly classified into two types; those defined by comparing *outcomes* across groups and those that compare *errors*. In the former case, for a binary classifier (that either accepts or rejects individuals), we would compare acceptance rates; for a regression problem, we might look at the the mean predicted target value. In the latter case, for a binary classifier, we might be more interested in false positive or false negative errors (depending on which kind are more advantageous to the individual); for a regression problem we might be interested in understanding if the process systematically over or under estimates for one group over another and how much by.

Let's look at some concrete examples for classification and discuss a few different interpretations of the definition in each case. Given a process that determines which job applicants make it to the interview stage, *balanced outcomes* across gender would require the probability of being invited to interview (acceptance

rate), be the same, regardless of gender. Expressed in a different way, we believe that the acceptance rate should be independent of gender. What about *balanced errors*? We might believe that for a fair decision process, the false discovery rate (the rate at which we incorrectly choose to interview individuals[1]) should be the same for all genders. That is, if we are not unjustly biased in accepting applicants of a particular gender, the false discovery rates should be the same for them all. Or again, put another way, we believe that we are fairly choosing who to interview if the false discovery rate is independent of gender.

In general group fairness criterion and measures can be derived from independence constraints on the joint distributions of the non-sensitive features X, sensitive features, Z, the target feature Y and predicted target $\hat{Y}$. For a summary of the notation and conventions we use in this book go to page v. Note that for a continuous classification model to be fair for all thresholds, we would replace $\hat{Y}$ with P in the above statement. For brevity (and because it makes the equations a little easier to read) we will express all constraints in terms of $\hat{Y}$, but keep in mind that for a classification model we might want to instead impose it on the score P.

3.1 Balanced outcomes

First we look at fairness constraints that impose independence between the outcome $\hat{Y}$ or target variable Y (depending on if we are interested in assessing the fairness of the data coming out of, or going into, our model respectively) and the sensitive feature, Z. We will consider two extremes, one where the variables are unconditionally independent and the other where the variables are conditionally independent given the other features X. To return to our example of determining which applicants make it to the next round of interviews, in the first case our fairness constraint requires that the acceptance rate be independent of gender while our second constraint requires that the acceptance rates be independent of gender, all else being equal (also known as the twin test). We will call the criteria independence and conditional independence respectively. Independence can be viewed as addressing disparate impact, since we are only interested in the relationship between the outcome and sensitive feature. Conditional independence has been interpreted as addressing disparate treatment, since if it is not satisfied, it establishes the sensitive feature as being the cause for the disparity in outcomes (assuming X and Z are the only model inputs). We summarise these fairness criteria in Table 3.1.

Table 3.1: Fairness constraints on outcomes.

Independence	Conditional Independence
$\hat{Y} \perp Z$	$(\hat{Y}\|X) \perp Z$

3.1.1 Independence

Of all the fairness criteria, independence is the most well known and imposes the strongest constraint. It requires the predicted target variable to be (unconditionally) independent of the sensitive feature. In other words, the distribution of the predicted target variable should be the same for all values of the sensitive feature,

$$\hat{Y} \perp Z \quad \Rightarrow \quad \mathbb{P}(\hat{y}|z) = \mathbb{P}(\hat{y}).$$

Note that we were in fact looking at independence criterion for the 1973 Berkeley admissions example in section 1.3. Imposing independence is a strong expression of the view that fairness is equality. It might be interpreted as the notion that abilities (or features) in all groups are, or should be, equally distributed; the belief that observed differences in the distributions in training data are a manifestation of unfair discrimination, errors in data collection, or both, rather than inherent differences in the abilities of people belonging to one group or another.

[1]For a reminder of confusion matrix metrics, see section B.1 of appendix B

Below we will define a range of fairness metrics, all derived from the notion of independence. Notice that independence imposes a constraint on only two random variables - the predicted target $\hat{Y}$ and sensitive feature Z. In the equations that follow, we provide metrics that quantify the fairness of our model output $\hat{Y}$, but we could equally well replace the predicted target variable $\hat{Y}$, with the actual target variable Y to assess the fairness of our data under the same criterion instead.

Mutual information, denoted I, is popular in information theory for measuring dependence between random variables.

$$I(\hat{Y}, Z) = \sum_{\hat{y}\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} \mathbb{P}(\hat{y}, z) \log \frac{\mathbb{P}(\hat{y}, z)}{\mathbb{P}(\hat{y})\mathbb{P}(z)}. \tag{3.1}$$

It is equal to zero, if and only if the joint distribution of Z and $\hat{Y}$ is equal to the product of their marginal distributions. Therefore, two variables which have zero mutual information are independent (equation (C.4)). The **normalised prejudice index**[5] divides mutual information by a normalising factor so that the resulting value falls between zero and one:

$$r_{\text{npi}} = \frac{I(\hat{Y}, Z)}{\sqrt{H(\hat{Y})H(Z)}}, \tag{3.2}$$

where

$$H(Y) = -\sum_{y\in\mathcal{Y}} \mathbb{P}(y) \log \mathbb{P}(y), \tag{3.3}$$

is the entropy. We have provided the formula for discrete random variables, for continuous variables we simply replace the sums with integrals.

Exercise: Normalised prejudice index

Write a function that takes two arrays y and z of categorical features and returns the normalised prejudice index.

1. Compute the probability distributions $\mathbb{P}(y)$, $\mathbb{P}(z)$ and $\mathbb{P}(y, z)$. Note that these can be thought of as the frequency with which each event occurs.
2. Compute the entropies $H(y)$ and $H(z)$ shown in equation (3.3) and use these to compite the normalising factor, $\sqrt{H(y)H(z)}$.
3. Compute the mutual information $I(z, y)$ shown in equation (3.1) and divide by the normalising factor.

You can test your implementation against scikit-learn's:
sklearn.metrics.normalized_mutual_info_score.

A simple relaxation of independence requires only the mean predicted target variable (rather than the full distribution) to be equal for all values of the sensitive feature. Assuming the sensitive feature to be binary, that is,

$$\mathbb{E}(\hat{Y}|Z = 1) = \mathbb{E}(\hat{Y}|Z = 0).$$

A popular measure derived from this for regression problems is called the **mean difference** which (as the name suggests) looks at the difference between the mean predictions for different values of the sensitive feature Z,

$$d = \mathbb{E}(\hat{Y}|Z = 1) - \mathbb{E}(\hat{Y}|Z = 0).$$

Taking the simplest example of discrete binary classifier where we have a binary sensitive feature. We can write the requirement of independence as,

$$\mathbb{P}(\hat{Y} = 1|Z = 1) = \mathbb{P}(\hat{Y} = 1|Z = 0).$$

This requirement goes by many names in research literature - **demographic parity**, **statistical parity** and **parity impact** among others. With this criterion for fairness we can quantify the disparity by looking at the difference (as with mean difference) or the ratio of the probabilities for each sensitive feature. We can calculate the metrics from the 2×2 contingency table[2] shown in Table 3.2. In bio-medical sciences, the **risk**

Table 3.2: Contingency table[2] for prediction against the sensitive feature.

	$\hat{Y}=1$	$\hat{Y}=0$	Total
$Z=1$	n_{11}	n_{10}	$n_{Z=1}$
$Z=0$	n_{01}	n_{00}	$n_{Z=0}$
Total	$n_{\hat{Y}=1}$	$n_{\hat{Y}=0}$	n

difference:

$$d = \mathbb{P}(\hat{Y}=1|Z=1) - \mathbb{P}(\hat{Y}=1|Z=0) = \frac{n_{11}}{n_{Z=1}} - \frac{n_{01}}{n_{Z=0}},$$

measures the impact of treatment (or risk factors), Z on outcome, $\hat{Y}$. In discrimination literature, it has been described as the **discrimination score** and **statistical parity difference** among others. Note that if $\hat{Y}=1$ is the advantageous outcome and $Z=1$ is the advantaged group, we would expect d to be non-negative[3]. The algorithm is fair when $d=0$. The further from zero, the more unfair. A modified version of this metric is the **normalised difference**[10] which divides the difference by,

$$d_{\max} = \min\left\{\frac{\mathbb{P}(\hat{Y}=1)}{\mathbb{P}(Z=1)}, \frac{\mathbb{P}(\hat{Y}=0)}{\mathbb{P}(Z=0)}\right\} = \min\left\{\frac{n_{\hat{Y}=1}}{n_{Z=1}}, \frac{n_{\hat{Y}=0}}{n_{Z=0}}\right\}, \tag{3.4}$$

thus ensuring the normalised difference is bounded between plus and minus one.

Exercise: Statistical parity difference maximum

Show that

$$d_{\max} = \min\left\{\frac{\mathbb{P}(\hat{Y}=1)}{\mathbb{P}(Z=1)}, \frac{\mathbb{P}(\hat{Y}=0)}{\mathbb{P}(Z=0)}\right\}.$$

Alternatively, we could instead take the ratio as a measure of discrimination:

$$r = \frac{\mathbb{P}(\hat{Y}=1|Z=1)}{\mathbb{P}(\hat{Y}=1|Z=0)} = \frac{n_{11}}{n_{Z=1}} \Big/ \frac{n_{01}}{n_{Z=0}}.$$

In biomedical sciences this measure is called the **risk ratio** and is used to measure the strength of association between treatment (or risk factors), Z, and outcome, $\hat{Y}$. It has been described in discrimination aware machine learning literature as the **impact ratio** or **disparate impact ratio**. The algorithm is fair if $r=1$. The further from one r is, the more unfair. The Equal Employment Opportunity Commission (EEOC) have used this measure in their guidelines for identifying discrimination in employment selection processes[4]. As a rule of thumb, the EEOC determine that a company's selection system is having an adverse impact on a particular group if the selection rate for that group is less than four-fifths (or 80%) that of the most advantaged group, that is, the impact ratio is less than 0.8 where $Z=0$ is the most advantaged group (for which the acceptance rate is the highest).

[2]Each cell of a contingency table shows the number of examples in the dataset satisfying the conditions given in the corresponding row and column headers with totals in the final row and column.

[3]We'll see later that this is not the case in the AIF360 implementation where $Z=1$ is the unprivileged group and $Z=0$ is the privileged group

The **elift ratio**[8] is similar to the impact ratio but instead of comparing acceptance rates for protected groups to each other, we compare them to the overall acceptance rate:

$$r_{\text{elift}} = \frac{\mathbb{P}(\hat{Y} = 1|Z = 0)}{\mathbb{P}(\hat{Y} = 1)}.$$

In theory, any measure of association suitable for the data types can be used as a metric to understand the magnitude of discrimination in our data or predictions. The **odds ratio** (popular in natural, social and biomedical sciences) is the ratio of the odds of a positive prediction for each group. We can write it as:

$$r_{\text{odds}} = \frac{\mathbb{P}(\hat{Y} = 1|Z = 1)\mathbb{P}(\hat{Y} = 0|Z = 0)}{\mathbb{P}(\hat{Y} = 0|Z = 1)\mathbb{P}(\hat{Y} = 1|Z = 0)} = \frac{n_{11}n_{00}}{n_{10}n_{01}}.$$

The odds ratio is equal to one when there is no discrimination. Recall that the odds ratio is not a collapsible measure (see section 1.3.3).

Exercise: Odds ratio

Show that the odds ratio is always greater than or equal to one in the case where $\hat{Y} = 1$ in the advantaged outcome and $Z = 1$ is the privileged group.

A nice feature of independence metrics is they can be evaluated on both the data and the model. A common problem in machine learning is that existing biases in the data can be exaggerated if protected groups are minorities in the population. By comparing bias metrics for the data with those of our model output, we can understand if our model is inadvertently introducing biases that do not originate from the data.

It should be intuitive that independence can only be satisfied naturally by a model if the target variable Y and sensitive feature Z are independent. If this is not the case then satisfying independence for your model will not permit the theoretically 'perfect' solution $\hat{Y} = Y$, should your model be able to achieve it. We would naturally expect that the stronger the relationship between the sensitive feature and target, the greater the trade-off between fairness and utility in satisfying independence criterion.

A major shortcoming of independence (discussed in section 1.3) is that it doesn't consider that there may be confounding variables. It assumes that all relevant features are held by all protected groups equally and where there are differences it assumes unfairness and passes the task of correcting for it, to the decision maker. Consider a simple hypothetical example where there are discrepancies between credit card approval rates for men and women at the population level which disappear once you control for (the confounding variable) income. It could be argued then that the real issue of fairness here appears to be the fact that women generally earn less than men. If the lender was to enforce independence between gender and its loan approval rate, say by setting less strict income requirements for women than men, this might feasibly lead to higher default rates among women. Clearly a less than desirable solution which, arguably, doesn't address the actual underlying problem. In fact it might be argued that enforcing independence could lead to less fair outcomes, on an individual level, in the sense that a man and woman who were the same in all other features would receive different outcomes as a result of enforcing independence in this way. We'll talk about individual fairness in the next chapter.

Suppose we want to measure the relationship between the sensitive feature and outcome using one of the above metrics. A natural solution to the problem of confounding variables is to control for them by conditioning on them (if you have them in your dataset, that is). Of course you need to know which variables to control for. Next, we consider the extreme case where we condition on all other variables.

3.1.2 Conditional Independence

Here we discuss the criterion which requires that predicted target variable is conditionally independent of the sensitive feature given all other features; that is,

$$(\hat{Y}|X)\perp Z \quad \Rightarrow \quad \mathbb{P}(\hat{y}|z,x) = \mathbb{P}(\hat{y}|x).$$

Suppose we wish to establish a causal connection between the decision or outcome and an individual's membership in some protected group. Typically in a decision process there are a number of unobserved variables, which makes proving a causal connection difficult. Take a job interview for example, the factors that determine who gets hired are typically subjective and often not even recorded, (as is typically the case in decisions which involve human judgement). In the case where a decision is made purely on the basis of an algorithm and there are no unobserved variables, making this connection becomes trivial. We simply perform a so called 'twin test'. We imagine a 'counterfactual' world in which for every individual in this world (say John Doe) there exists an 'identical twin' in the counterfactual world which differers only by the protected feature of interest (Jane Doe). We then simply compare outcomes for the two individuals (John and Jane). If the outcomes are different, we have established the individual's membership in the protected group (in our example, gender) as the sole reason for it.

Taking this approach to establishing cause with a model is pretty straight forward. We simply conduct a randomized experiment. The individuals for which we check the model output, need not exist, we can simply fabricate them and determine what the resulting model prediction is. Doing the twin test for a dataset (i.e. where you do not have access to the algorithm, only the decisions/predictions) is less trivial since the conterfactual twin for any given example need not exist in the data and we have no way of producing them without the algorithm. In addition, for any given point in the non-sensitive feature space the number of data points will likely be too small to justify the use of statistical methods in establishing cause. Barring this issue, using the counterfactual approach to establishing the fairness of our model, we can consider all the metrics we have above with independence as our fairness criterion but conditioned on X as well as Z. So for example we define the **causal mean difference** as

$$d = \mathbb{E}(\hat{Y}|Z=1, X=x) - \mathbb{E}(\hat{Y}|Z=0, X=x).$$

and the **observed mean difference** as

$$d = \mathbb{E}(Y|Z=1, X=x) - \mathbb{E}(Y|Z=0, X=x).$$

Multiple papers have described this as a means to establish disparate treatment liability in an algorithmic decision process because it exposes differing treatment of individuals based on protected class membership. Note that while it would be sufficient to demonstrate disparate treatment, (as discussed in section 1.2.3) it is not necessary. Using protected features in the algorithm would be enough to result in disparate treatment liability in the US, the actual impact of using the feature is irrelevant.

3.1.3 Introduction to AIF360

Now that we have covered some measures of fairness, let's dive into calculating them. In this book we are going to use IBM's AI Fairness 360 (AIF360). AIF360 is currently the most comprehensive open source library available for measuring and mitigating bias in machine learning models. The Python package includes an extensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models many of which we will cover in this book. The system has been designed to be extensible, adopted software engineering best practices to maintain code quality, and is well documented. The package implements techniques from at-least eight published papers and includes over 71 bias detection metrics and nine bias mitigation algorithms[2]. These techniques can all be called in a standard way, similar to scikit-learn's fit/transform/predict paradigm.

In this section we're going to use AIF360 to calculate some of the metrics we've talked about in the previous section as a means to get started working with it. For calculating the metrics we've talked about

so far, using AIF360 might seem to add unnecessary overhead as they are reasonably straightforward to code up directly once you have your data in a Pandas DataFrame. But remember, the library contains implementations of more complicated metrics and bias mitigations algorithms that we'll cover later on in this book. Before we can use the library, we need to install it. Instructions are provided in Appendix A.

Statlog (German Credit Data) Data Set

The Jupyter Notebook, `mbml_german.ipynb`, contains an example calculating some of the above fairness metrics on both a dataset and model output. It uses the Statlog (German Credit Data) Data Set, in which one thousand loan applicants are classified as representing 'good' or 'bad' credit risks based on features such as loan term, loan amount, age, gender, marital status and more.

Exercise: Statlog (German Credit Data) Data Set

Sections 1-3 in the Jupyter Notebook, `mbml_german.ipynb`, load the data and perform some exploratory data analysis (EDA), looking at correlation heat maps (using a variety of different measures of association) and comparing distributions of the target for different values of the features. Open the notebook and run the code up to section four. You should be able to answer the following questions by working through the notebook.

1. What proportion of the population is classified as male/female?
2. What proportion of the population have good credit vs bad?
3. How many continuous variables are there? What are they? Do any of them appear to be related? If so how?
4. How many categorical variables are there? What are they? Do any of them appear to be related? If so how?

Some EDA here?

Calculating independence metrics

In order to calculate our metrics on the data using AIF360, we must have it in the correct format; that is, in a Pandas DataFrame (`data_df`) containing only numeric data types. In code listing 3.1, we calculate the rate at which male and female applicants are classified as being good credit risks (`base_rate`) along with the difference (`mean_difference`) and the ratio (`disparate_impact`) of these rates.

Listing 3.1: Calculating independence metrics for the data using AIF360

```
# Create a DataFrame to store results in
outcomes_df = pd.DataFrame(columns=['female', 'male',
                                    'difference', 'ratio'],
                           index=['data', 'model',
                                  'train data', 'train model',
                                  'test data', 'test model'])

# Define privileged and unprivileged groups
privileged_groups = [{'sex_male':1}]
unprivileged_groups = [{'sex_male':0}]

# Create an instance of BinaryLabelDataset
data_ds = BinaryLabelDataset(df = data_df,
    label_names = ['goodcredit'],
```

```
    protected_attribute_names = ['sex'])

# Create an instance of BinaryLabelDatasetMetric
data_metric = BinaryLabelDatasetMetric(data_ds,
    privileged_groups = privileged_groups,
    unprivileged_groups = unprivileged_groups)

# Compute the metrics with data_metric and store them in outcomes_df
outcomes_df.at['data', 'female'] = data_metric.base_rate(privileged=0)
outcomes_df.at['data', 'male'] = data_metric.base_rate(privileged=1)
outcomes_df.at['data', 'difference'] = data_metric.mean_difference()
outcomes_df.at['data', 'ratio'] = data_metric.disparate_impact()
```

In the notebook we look at these metrics on both the data and the model output for three different sets of the data (the full dataset, the train set and the test set) with two different models (one trained on the full dataset and another trained only on a subset of the data - the training set). In code listing 3.1, we create a DataFrame to display the results in (`outcomes_df`) and populate the first row of it. First we define our privileged and unprivileged groups.

Defining privileged and unprivileged groups

The format for these is a list of dictionaries. Each dictionary in the list defines a group, the key being a feature and the value being the value of the feature for members of the group. The key, value pairs in the dictionaries are joined with an intersection (AND operator) and the dictionaries in the list are joined with a union (OR operator). So for example,

```
[{'sex': 1, 'age>=30': 1}, {'sex': 0}]
```

corresponds to individuals such that,

```
(data_df['sex']==1 AND data_df['age>=30']==1)  OR (data_df['sex']==0)
```

Next we create a `BinaryLabelDataset` object (`data_ds`) which in turn is used to create a `BinaryLabelDatasetMetric` object (`data_metric`). We then calculate the fairness metrics from `data_metric` and store the results in `outcomes_df`.

Exercise: Multiple sensitive features

Calculate independence metrics (base rates, difference and ratio) for the full dataset in the case where the privileged group is males age 30 and over, and the unprivileged group is females under the age of 30. Do this two ways, using AIF360 and using Pandas. Compare your results to make sure they match.

Once we have trained a model and made predictions, similar code can be written to calculate independence metrics on the model predictions for the full dataset. Code listing 3.2 shows how we do this using the predictions from the trained model `clf`.

Listing 3.2: Calculating independence metrics for the model using AIF360

```
# Create a DataFrame with the features and model predicted target
model_df = pd.concat([X, pd.Series(clf.predict(X), name='goodcredit')],
        axis=1)

# Create an instance of BinaryLabelDataset
```

```
model_ds = BinaryLabelDataset(df = model_df,
    label_names = ['goodcredit'],
    protected_attribute_names = ['sex_male'])

# Create an instance of BinaryLabelDatasetMetric
model_metric = BinaryLabelDatasetMetric(model_ds,
    privileged_groups = privileged_groups,
    unprivileged_groups = unprivileged_groups)

# Compute the metrics with model_metric and store them in outcomes_df
outcomes_df.at['model', 'female'] = model_metric.base_rate(privileged=0)
outcomes_df.at['model', 'male'] = model_metric.base_rate(privileged=1)
outcomes_df.at['model', 'difference'] = model_metric.mean_difference()
outcomes_df.at['model', 'ratio'] = model_metric.disparate_impact()
```

Table 3.3 shows the results of the calculations stored in `outcomes_df` from the notebook. From Table 3.3 we note some variation in the rates at which men and women are predicted to present good credit risks for the model versus the data. In particular, the model acceptance rates are higher for both male and female applicants than those observed in the data. There are particularly big differences when we compare results for the test data versus the model on the test data (test model), which is not surprising since the mean difference and impact ratio for the train data and test data are markedly different. In addition we are aware that our model is overfitting. Without intervention, our model appears to be reducing the bias present in the data for the test set (as measured by our independence metrics).

Table 3.3: Acceptance rates for the Statlog (German Credit) Data Set.

	Female	Male	Difference	Ratio
Data	0.648	0.723	-0.0748	0.897
Model[a]	0.674	0.749	-0.0751	0.900
Train data	0.659	0.719	-0.0601	0.916
Train model[b]	0.667	0.731	-0.0647	0.911
Test data	0.607	0.741	-0.1345	0.819
Test model[b]	0.705	0.820	-0.1152	0.860

[a]Model trained on the full dataset.
[b]Model trained on the train dataset only.

Exercise: Twin test

Implement the twin test (described in section 3.1.2) for the model trained on the full dataset. Calculate the causal mean difference between male and female applicants using 2000 data points (1000 male and 1000 female applicants) i.e. the full dataset together with the 'twin' of the opposite gender.

3.2 Balanced errors

In this section we learn about fairness criteria which seek to balance errors across groups, rather than outcomes. The fundamental assumption here is that the training data is fair; the target variable is the ground truth variable we wish to affect, the data is accurate and representative of the population and the features directly impact the target. Assuming we have said data, for our model to be fair, we require the errors to be distributed similarly for different subgroups of the population (defined by the values of sensitive features). Expressed differently, we want the errors to be independent of protected characteristics, that is,

$(\hat{Y} - Y) \perp Z$. We discussed earlier in the chapter how independence and conditional independence constraints have been interpreted as avoiding disparate impact and treatment respectively. Analogously, balanced error criterion have been described as avoiding **disparate mistreatment**[9].

3.2.1 Regression

A relaxation of this criterion balances the mean error for the groups (rather than comparing the full distributions). **Balanced residuals**[3] takes the difference of the mean errors as a measure of fairness:

$$d_{\text{err}} = \mathbb{E}(\hat{y} - y | Z = 1) - \mathbb{E}(\hat{y} - y | Z = 0),$$

or written more explicitly,

$$d_{\text{err}} = \frac{1}{n_0} \sum_{i|z_i=0} (y_i - \hat{y}_i) - \frac{1}{n_1} \sum_{i|z_i=1} (y_i - \hat{y}_i).$$

Here $d_{\text{err}} = 0$ would be considered fair.

3.2.2 Classification

For a classification problem the most obvious relaxation would be to ensure equal error rates (or equivalently accuracy) for all groups. As an example, recall the project Gender Shades we discussed in section 1.4, that audited several commercial gender classification packages measured their accuracy for different protected groups. To derive a measure of fairness from this criterion we could (as before) take the difference or the ratio; both of these are implemented in AIF360. The **error rate difference** is given by

$$d_{\text{err}} = \mathbb{P}(\hat{Y} \neq Y | Z = 1) - \mathbb{P}(\hat{Y} \neq Y | Z = 0).$$

Again here $d_{\text{err}} = 0$ would be considered fair. The **error rate ratio** is given by

$$r_{\text{err}} = \frac{\mathbb{P}(\hat{Y} \neq Y | Z = 1)}{\mathbb{P}(\hat{Y} \neq Y | Z = 0)}$$

in which case $r_{\text{err}} = 1$ would be considered fair.

A binary classification model can make two different types of errors (false positives and false negatives), one of which will often be more desirable than the other. Table B.2 in appendix B summarises the different types of error rates for a binary classification model that we might want to balance. The differences and ratios of all of these metrics can be found in AIF360s ClassificationMetric class.

Balancing errors (or equivalently performance metrics) across groups can be broken down into two separate criteria, described as **separation** and **sufficiency**[1]. Each of these criteria can be defined as a conditional independence constraint on the joint distributions of the sensitive features, Z, the target feature Y and predicted target $\hat{Y}$. We summarise the these in Table 3.4. In the sections that follow we shall see how

Table 3.4: Fairness constraints on errors.

Separation	Sufficiency
$\hat{Y} \perp (Z \mid Y)$	$Y \perp (Z \mid \hat{Y})$

each of these criteria correspond to balancing error rates along the columns (conditioning on Y) or the rows (conditioning on $\hat{Y}$) of the confusion matrix (see Table B.2 in appendix B). In the former case, the criterion requires the false negative and false positive rates to be the same for all groups; in the latter case, the false discovery rate and false omission rate to be the same for all groups. Let's start with separation.

Separation

Separation requires the predicted target variable to be independent of the sensitive feature, conditioned on the target variable, that is, $\hat{Y} \perp (Z|Y)$. We can say that the predicted target $\hat{Y}$, is 'separated' from the sensitive feature Z, by the target variable Y. The corresponding graphical model for separation criteria is shown in Figure 3.1. Essentially we are saying that for a fixed value of the target variable, there should be

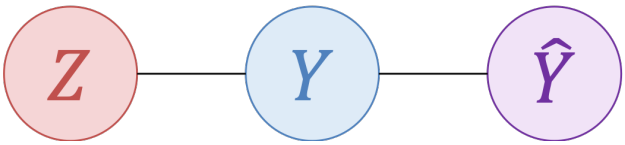


Figure 3.1: Graphical model for separation.

no difference in the distribution of the predicted target variable, for different values of the sensitive feature. That is,

$$\mathbb{P}(\hat{y}|y, z) = \mathbb{P}(\hat{y}|y).$$

Unlike independence, separation, allows for dependence between the predicted target variable and the sensitive feature but only to the extent that it exists between the actual target variable and the sensitive feature.

Once again let's take the simplest example of discrete binary classifier where we have a single sensitive binary feature. We can write this requirement as two conditions,

$$\mathbb{P}(\hat{Y} = 1|Z = 1, Y = 1) = \mathbb{P}(\hat{Y} = 1|Z = 0, Y = 1),$$
$$\mathbb{P}(\hat{Y} = 1|Z = 1, Y = 0) = \mathbb{P}(\hat{Y} = 1|Z = 0, Y = 0).$$

Recall that $\mathbb{P}(\hat{Y} = 1|Y = 1)$ is the true positive rate (TPR) of the classifier and $\mathbb{P}(\hat{Y} = 1|Y = 0)$ is the false positive rate (FPR). We see then that separation requires the true positive rate to be the same for all values of the sensitive feature and the false positive rate to be the same for all values of the sensitive feature. Note that the true positive rate is balanced if and only if the false negative rate is balanced, so thinking in terms of error metrics only, separation requires the false negative and false positive rates to be balanced. This fairness criterion is most well known as **equalised odds**[6].

Implemented in IBM's fairness library, AIF360, are two related metrics. The **average odds difference** measures the magnitude of unfairness as the average of the difference in true positive rate and false positive rate, that is,

$$d_{\text{av-odds}} = \frac{1}{2}[TPR_{Z=0} - TPR_{Z=1} + FPR_{Z=0} - FPR_{Z=1}].$$

The **average odds error** measures the magnitude of unfairness as the average of the absolute difference in true positive rate and false positive rate, that is,

$$d_{\text{av-odds-err}} = \frac{1}{2}[|TPR_{Z=0} - TPR_{Z=1}| + |FPR_{Z=0} - FPR_{Z=1}|].$$

A relaxed version of equalised odds, called **equal opportunity**[6], requires only the true positive rates to be the same across all groups (assuming a positive prediction is the more advantageous ourcome). A metric which uses this as a criterion to measure unfairness in AIF360 is **equal opportunity difference** which takes the difference in true positive rates across groups, that is,

$$d_{\text{eq-op}} = TPR_{Z=0} - TPR_{Z=1}.$$

Exercise: Fair equality of opportunity

Can you see how the metric *equal opportunity* relates to the second principle of justice as fairness discussed in section 1.2.1?

Sufficiency

Sufficiency requires the sensitive feature Z and target variable Y to be independent, conditional on the predicted target variable $\hat{Y}$, that is, $Y \perp (Z|\hat{Y})$. We can say that the predicted target $\hat{Y}$ is 'sufficient' for the sensitive feature Z. That is to say, given $\hat{Y}$, Z provides no additional information. The corresponding graphical model for sufficiency criteria is shown in Figure 3.2. Comparing sufficiency to separation we

Figure 3.2: Graphical model for sufficiency.

note that Y and $\hat{Y}$ are reversed in the graphical model and conditional independence constraint. It should hopefully be straightforward to see then that sufficiency requires the false omission rate and false discovery rate (shown in Table B.2 of appendix B) to be balanced across protected groups.

Exercise: Sufficiency

Show that sufficiency is satisfied if and only if the false omission rate and false discovery rate are equal for all groups.

There are some nice properties of separation and sufficiency criteria. Note that unlike balanced outcome criteria they do not preclude the theoretically 'perfect' solution, $\hat{Y} = Y$. The criteria also preclude large differences in error rates for different groups that are typical when disadvantaged classes are minorities suffering from low support. It's worth reiterating that unlike independence, separation and sufficiency criteria assume that the relationship between Y and Z prescribed by the training data is fair, thus only make sense if the target variable is reliable as the ground truth. In such cases, balancing error criteria allow flexibility in the choice which types of errors are important to equalize, based on the human cost. For example, in pretrial risk assessment we might choose to prioritise balancing false positive rates if we believe that it is preferable to set free a guilty defendant than incarcerate an innocent one. As another example, let's take the infamous NYPD stop-and-frisk program where pedestrians were stopped, interrogated and searched on 'reasonable' suspicion of carrying contraband. In this case we might want to ensure false discovery rates are balanced across groups to ensure we are not disproportionately targeting particular minority groups.

Exercise: Stop-and-frisk

- Why might we choose to balance false discovery rates for stop-and-frisk, rather than say false omission, false negative or false positive rates?
- Is it fair to only balance false discovery rates?
- How might we go about measuring the false omission rate if we wanted to check if it was also balanced?

Of our two fairness criteria, separation and sufficiency, the latter imposes a weaker constraint on our model. To understand why we explore another interpretation of sufficiency which intuitively explains why, in many cases, it is satisfied implicitly through the training process[7]. Let us look at sufficiency criteria in terms of the classification score P,

$$\mathbb{P}(Y = 1|P = p, Z = 1) = \mathbb{P}(Y = 1|P = p, Z = 0) \quad \forall p$$

We say that a classifier score is calibrated if

$$\mathbb{P}(Y = 1|P = p) = p \quad \forall p.$$

Essentially, this is the requirement that the proportion of data points assigned the score p, which did in fact have a positive outcome $Y = 1$, should be equal to the score p. The score p can then be interpreted, at the population level, as the probability that the a positive prediction $\hat{Y} = 1$ would be correct[4].

From the definitions above we can see that if our classifier scores are calibrated for all groups, sufficiency is automatically satisfied. Conversely, if our model satisfies sufficiency but not calibration by group, we can calibrate our model score through a simple transformation. We simply pick a value for Z, $Z = 1$ say, and then calculate the mapping,

$$\mathbb{P}(Y = 1|P = p, Z = 1) = f(p).$$

We then transform all our scores to new scores (which satisfy calibration by group) by applying the inverse mapping $f^{-1}(P)$.

3.2.3 Back to AIF360

In order to calculate balanced error metrics with AIF360, we need to create an object of type `ClassificationMetric`. Returning to our example working with the German Credit Data, code listing 3.3 calculates a series of balanced error metrics and populates the DataFrame `errors_df` with them. Recall that `data_ds` and `model_ds` were created in code listings 3.1 and 3.2 respectively; `privileged_groups` and `unprivileged_groups` were defined in the former code listing.

Listing 3.3: Calculating balanced error metrics with AIF360

```
# Create a DataFrame to store results in
errors_df = pd.DataFrame(columns=['female', 'male',
                                  'difference', 'ratio'],
                         index=['ERR', 'FPR', 'FNR', 'FDR', 'FOR'])

# Create an instance of ClassificationMetric
clf_metric = ClassificationMetric(data_ds,
    model_ds,
    privileged_groups = privileged_groups,
    unprivileged_groups = unprivileged_groups)

# Compute the metrics with clf_metric and store them in errors_df
# Error rates for the unprivileged group
errors_df.at['ERR', 'female'] = clf_metric.error_rate(privileged=False)
errors_df.at['FPR', 'female'] =
    clf_metric.false_positive_rate(privileged=False)
errors_df.at['FNR', 'female'] =
    clf_metric.false_negative_rate(privileged=False)
errors_df.at['FDR', 'female'] =
    clf_metric.false_discovery_rate(privileged=False)
errors_df.at['FOR', 'female'] =
    clf_metric.false_omission_rate(privileged=False)

# Error rates for the privileged group
errors_df.at['ERR', 'male'] = clf_metric.error_rate(privileged=True)
errors_df.at['FPR', 'male'] =
```

[4]For the score to be interpretable as this probability at the individual level, we would need to satisfy the stronger criteria $P = \mathbb{E}[Y|X]$

```
    clf_metric.false_positive_rate(privileged=True)
errors_df.at['FNR', 'male'] =
    clf_metric.false_negative_rate(privileged=True)
errors_df.at['FDR', 'male'] =
    clf_metric.false_discovery_rate(privileged=True)
errors_df.at['FOR', 'male'] =
    clf_metric.false_omission_rate(privileged=True)

# Differences in error rates
errors_df.at['ERR', 'difference'] = clf_metric.error_rate_difference()
errors_df.at['FPR', 'difference'] =
    clf_metric.false_positive_rate_difference()
errors_df.at['FNR', 'difference'] =
    clf_metric.false_negative_rate_difference()
errors_df.at['FDR', 'difference'] =
    clf_metric.false_discovery_rate_difference()
errors_df.at['FOR', 'difference'] =
    clf_metric.false_omission_rate_difference()

# Ratios of error rates
errors_df.at['ERR', 'ratio'] = clf_metric.error_rate_ratio()
errors_df.at['FPR', 'ratio'] = clf_metric.false_positive_rate_ratio()
errors_df.at['FNR', 'ratio'] = clf_metric.false_negative_rate_ratio()
errors_df.at['FDR', 'ratio'] = clf_metric.false_discovery_rate_ratio()
errors_df.at['FOR', 'ratio'] = clf_metric.false_omission_rate_ratio()

display(errors_df)
```

The DataFrame `error_df` is shown in Table 3.5. This time we just look at the metrics for the model trained

Table 3.5: Error metrics for the Statlog (German Credit Data) Data Set.

Error metric[a]	Female	Male	Difference	Ratio
ERR	0.246	0.180	0.066	1.37
FPR	0.458	0.472	-0.014	0.97
FNR	0.108	0.078	0.030	1.39
FDR	0.250	0.152	0.098	1.65
FOR	0.235	0.296	-0.061	0.79

[a]We abbreviate error rate (ERR), false positive rate (FPR), false negative rate (FNR), false discovery rate (FDR) and false omission rate (FOR). See appendix section B.1 for detailed descriptions of confusion matrix metrics.

on the training set and calculated on the test set. We note that the overall error rate is 37% higher for female applicants. The false negative rate is 39% higher for female applicants, that is for female applicants we more often incorrectly predict that they represent bad credit risks when they are in fact good credit risks. We also note that the false discovery rate is 65% higher for female applicants which means that when we do predict women to be credit worthy they are more often not. The false omission rate is 21% lower for female applicants which means we are more often correct when we predict that they are not credit worthy. Our findings are not surprising given the difference in prevalence of credit worthy male and female applicants between our training and test sets shown in Table 3.3.

Recall that when we compared fairness metrics under the independence criterion, it appeared that our model was reducing the level of bias in the data. Note that comparing balanced error metrics (in addition to independence metrics) gives us a richer understanding of the behaviour of our model in relation to protected

groups.

3.3 Incompatibility between fairness criteria

So far in this chapter we have learned a range of different group fairness criteria and seen how each of them can be viewed as imposing different restrictions on the joint distributions of our variables X, Z, Y and $\hat{Y}$. In this section we will show that these fairness criteria in some cases are restrictive enough to mean that satisfying multiple fairness criteria is impossible, except in some degenerate cases. In proving incompatibility between the fairness criteria given about we'll use various rules of probability. These are summarised in Appendix C.

3.3.1 Independence versus Sufficiency

Independence versus Sufficiency

Independence ($Z \perp \hat{Y}$) and sufficiency ($Z \perp Y | \hat{Y}$) can only be simultaneously satisfied if the sensitive feature, Z and the target variable $\hat{Y}$ are independent ($Z \perp Y$).

To see this consider the conditional distribution $\mathbb{P}(z|y, \hat{y})$. Applying independence criterion, equation (C.5), followed by the product rule in equation (C.2),

$$Z \perp \hat{Y} \quad \Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z|y) = \frac{\mathbb{P}(z, y)}{\mathbb{P}(y)}. \tag{3.5}$$

Applying sufficiency, equation (C.6), followed by independence, equation (C.5), gives,

$$Z \perp Y | \hat{Y} \quad \Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z|\hat{y}) = \mathbb{P}(z). \tag{3.6}$$

Equating equations (3.5) and (3.6) and then rearranging gives,

$$\mathbb{P}(z|y) = \mathbb{P}(z)\mathbb{P}(y).$$

Thus, Z and Y must be independent.

3.3.2 Independence versus Separation

Independence versus Separation

In the case that Y is binary, independence ($Z \perp \hat{Y}$) and separation ($Z \perp \hat{Y} | Y$) criteria can only be simultaneously satisfied if either $\hat{Y} \perp Y$ or $Y \perp Z$.

To see this we start by using the sum rule in equation (C.1), and then the product rule in equation (C.2) to give,

$$\mathbb{P}(\hat{y}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}, y) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y)\mathbb{P}(y). \tag{3.7}$$

Doing the same again but conditioning on Z, we have

$$\mathbb{P}(\hat{y}|z) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y, z)\mathbb{P}(y|z).$$

Since $\hat{Y} \perp Z$ we can rewrite this as

$$\mathbb{P}(\hat{y}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y)\mathbb{P}(y|z). \tag{3.8}$$

Equating equations (3.7) and (3.8) and rearranging gives,

$$\sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y) \left[\mathbb{P}(y) - \mathbb{P}(y|z)\right] = 0 \tag{3.9}$$

If we assume Y is binary, then

$$\mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = 0).$$

Substituting in equation (3.9) we can show that

$$[\mathbb{P}(\hat{y}|Y = 0) - \mathbb{P}(\hat{y}|Y = 1)][\mathbb{P}(Y = 0) - \mathbb{P}(Y = 0|z)] = 0,$$

which is true if and only if

$$\mathbb{P}(\hat{y}|Y = 0) = \mathbb{P}(\hat{y}|Y = 1) \quad \Rightarrow \quad \hat{Y} \perp Y,$$

or,

$$\mathbb{P}(Y = 0) = \mathbb{P}(Y = 0|z) \quad \Rightarrow \quad Y \perp Z.$$

3.3.3 Separation versus Sufficiency

Separation versus Sufficiency I

In the case where all events in the joint distribution of Z, Y and $\hat{Y}$ have non zero probability, separation ($Z \perp \hat{Y}|Y$) and sufficiency ($Z \perp Y|\hat{Y}$) can only be simultaneously be satisfied if the sensitive feature, Z is independent of both the target variable Y and the predicted target $\hat{Y}$, that is if $Z \perp Y$ and $Z \perp \hat{Y}$.

To see this consider the conditional distribution $\mathbb{P}(z|y, \hat{y})$. Applying separation and sufficiency criteria gives,

$$\mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z|y) = \mathbb{P}(z|\hat{y}). \tag{3.10}$$

Substituting (3.10) into the product rule (C.2) gives,

$$\mathbb{P}(z, y) = \mathbb{P}(z|\hat{y})\mathbb{P}(y). \tag{3.11}$$

Substituting equation (3.11) into the sum rule (C.1) gives,

$$\mathbb{P}(z) = \sum_{y \in \mathcal{Y}} \mathbb{P}(z|\hat{y})\mathbb{P}(y)$$

Provided all events have non-zero probability, we know that $\mathbb{P}(z|\hat{y})$ is not some trivial function of Y and we can move $\mathbb{P}(z|\hat{y})$ outside of the summation. Thus we have,

$$\mathbb{P}(z) = \mathbb{P}(z|\hat{y}) \tag{3.12}$$

and Z and $\hat{Y}$ must be independent. Equating equations (3.10) and (3.12) tells us that Z and Y must also be independent.

Separation versus Sufficiency II

In the case where Y is binary, separation and sufficiency can only be satisfied simultaneously if the sensitive feature is independent of the target variable, or the model has an accuracy of 100% ($\hat{Y} = Y$) or 0% ($\hat{Y} = 1 - Y$).

Consider the case where Y is binary. Separation requires all groups to have the same true positive rate (recall or TPR) and the same false positive rate (FPR). On the other hand, sufficiency requires all groups to have the same positive predictive value (precision or PPV) and the same negative predictive value (NPV). Then under separation and sufficiency, we can write the positive and negative predictive values in terms of the true positive and false positive rates as follows:

$$PPV = \frac{pTPR}{pTPR + (1-p)FPR} \tag{3.13}$$

and

$$NPV = \frac{(1-p)(1-FPR)}{p(1-TPR) + (1-p)(1-FPR)} \tag{3.14}$$

where $p = \mathbb{P}(Y = 1)$.

Exercise: Predictive values

Prove the results given in equations (3.13) and (3.14)

Denote $p_z = \mathbb{P}(Y = 1|Z = z)$ then we can show from equations (3.13) and (3.14) that for any distinct pair of groups $Z = a$ and $Z = b$ for both separation and sufficiency to hold we must have

$$FPR(p_a - p_b)TPR = 0 \tag{3.15}$$

and

$$(1 - FPR)(p_a - p_b)(1 - TPR) = 0 \tag{3.16}$$

respectively.

Exercise: Separation versus sufficiency

Show that for separation and sufficiency to hold equations (3.15) and (3.16) must hold for for any pair of groups $Z = a$ and $Z = b$.

Equations (3.15) and (3.16) can only be simultaneously satisfied in 3 cases:

1. $p_a = p_b \, \forall \, a, b$ in which case $Y \perp Z$,
2. $FPR = 0$ and $TPR = 1$ in which case $Y = \hat{Y}$,
3. $FPR = 1$ and $TPR = 0$ in which case $Y = 1 - \hat{Y}$.

Summary

Group fairness

- The term group fairness is used to describe a series of metrics that all stem from the same high level idea; the notion that some property should be balanced or equal across different subgroups of a population, where the subgroups are determined by the values of some protected characteristic such as gender or race.
- In general group fairness criterion and measures can be derived from independence constraints on the joint distributions of the non-sensitive features X, sensitive features, Z, the target feature Y and predicted target $\hat{Y}$.
- Group fairness criteria can be broadly classified into two types; those seeking to balance outcomes across groups and those balancing errors.

Balanced Outcomes

Independence

- The term *group fairness* is used to describe a class of metrics that are used to measure discrimination or bias in a given decision process.
- Independence imposes the requirement that the predicted target variable be independent of the sensitive feature.
- Independence can be viewed as addressing disparate impact, since we are only interested in the relationship between the outcome and sensitive feature.
- Independence is a strong expression of the view that fairness is equality. It might be interpreted as the notion that abilities (or features) in all groups are, or should be, equally distributed; the belief that observed differences in the distributions in training data are a manifestation of unfair discrimination, errors in data collection, or both, rather than inherent differences in the abilities of people belonging to one group or another.
- A nice feature of independence metrics is they can be evaluated on both the data and the model. A common problem in machine learning is that existing biases in the data can be exaggerated if protected groups are minorities in the population. By comparing independence metrics for the data and with those of our model output we can understand if our model is inadvertently introducing biases that do not originate from the data.
- If the target variable Y and sensitive feature Z are not independent then satisfying independence for your model will not permit the theoretically 'perfect' solution $Y = \hat{Y}$. We would naturally expect that the stronger the relationship between the sensitive feature and target, the greater the trade-off between fairness and utility in satisfying the independence criterion.
- A major shortcoming of independence is that it doesn't consider that there may be confounding variables. It assumes that all relevant features are held by all protected groups equally and where there are differences it assumes unfairness and passes the task of correcting for it to the decision maker.

Conditional independence

- Conditional independence imposes the requirement that the predicted target variable be conditionally independent of the sensitive feature, given all other features.
- Conditional independence has been interpreted as addressing disparate treatment, since it exposes differing treatment of individuals based on protected class membership. In reality, while it would be sufficient to demonstrate disparate treatment, it is not necessary. Using protected features in the algorithm would be enough to result in disparate treatment liability in the US, the impact of using the feature is irrelevant.
- In the case where a decision is made purely on the basis of an algorithm and there are no unobserved variables, we can perform a 'twin test' to establish disparate treatment. We conduct a randomized experiment and calculate the causal mean difference. If the value is non-zero, we have established the existence of disparate treatment.

Balanced errors

- Balanced error criteria assume that the relationship between the target variable and sensitive feature prescribed by the training data is fair so only make sense if the target variable is reliable as the ground truth. Under this assumption that our data is fair, for our model to be fair, we require errors to be distributed similarly for different subgroups of the population (defined by the values of sensitive features).
- Ensuring balanced errors has been described as avoiding disparate mistreatment.
- For a regression model balanced residuals takes the difference of the mean errors for each group as a measure of fairness.

- For a classification problem we could could use the error rate difference or the error rate ratio as a measure of fairness.
- Unlike balanced outcome criteria, balanced error criteria do not preclude the theoretically 'perfect' solution, $\hat{Y} = Y$.

Separation

- Separation requires the predicted target variable to be independent of the sensitive feature, conditioned on the target variable
- Separation, allows for dependence between the predicted target variable and the sensitive feature but only to the extent that it exists between the actual target variable and the sensitive feature.
- For a binary classification model, separation requires both the false negative and false positive rates to be balanced across groups. This criterion is known as equalised odds
- Equal opportunity criterion requires only the true positive rates to be the same across all groups (assuming a positive prediction is the more advantageous outcome).

Sufficiency

- Sufficiency requires the sensitive feature and target variable to be independent, conditional on the predicted target variable.
- For a binary classification model, sufficiency requires both the false omission rate and false discovery rates to be balanced across protected groups.
- Sufficiency is is a weaker model constraint compared to separation as it is often satisfied implicitly through the training process.

Incompatibility between fairness criteria

- Independence ($Z \perp \hat{Y}$) and sufficiency ($Z \perp Y | \hat{Y}$) can only be simultaneously be satisfied if the sensitive feature Z, and the target variable $\hat{Y}$, are independent ($Z \perp Y$).
- In the case that Y is binary, independence ($Z \perp \hat{Y}$) and separation ($Z \perp \hat{Y} | Y$) criteria can only be simultaneously satisfied if either $\hat{Y} \perp Y$ or $Y \perp Z$.
- Separation ($Z \perp \hat{Y} | Y$) and sufficiency ($Z \perp Y | \hat{Y}$) can only be simultaneously be satisfied if the sensitive feature, Z is independent of both the target variable Y and the predicted target $\hat{Y}$, that is if $Z \perp Y$ and $Z \perp \hat{Y}$.
- In the case where Y is binary, separation and sufficiency can only be satisfied simultaneously if the sensitive feature is independent of the target variable, or the model has an accuracy of 100% ($\hat{Y} = Y$) or the model has an accuracy of 0% ($\hat{Y} = 1 - Y$).

AIF360

- To use AIF360 to calculate fairness metrics we need our data in a Pandas DataFrame which has only numeric data types to create a `BinaryLabelDataset` object.
- Balanced outcome fairness metrics are methods of the `BinaryLabelDatasetMetric` class.
- Balanced error fairness metrics are methods of the `ClassificationMetric` class.
- Privileged and unprivileged groups are defined in the format of a list of dictionaries. Each dictionary in the list defines a group, the key being a feature and the value being the value of the feature for members of the group. The key, value pairs in the dictionaries are joined with an intersection (AND operator) and the dictionaries in the list are joined with a union (OR operator).

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[2] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. https://arxiv.org/abs/1810.01943.

[3] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, December 2013. https://www.researchgate.net/publication/261637367_Controlling_Attribute_Effect_in_Linear_Regression.

[4] U.S. Equal Employment Opportunity Commission. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *Federal Register*, 44(43), March 1979. https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines.

[5] Kazuto Fukuchi, Jun Sakuma, and Toshihiro Kamishima. Prediction with model-based neutrality. *IEICE TRANS. INF. & SYS.*, E98-D(8), August 2015. https://www.kamishima.net/archive/2015-t-ieice-print.pdf.

[6] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016. https://arxiv.org/abs/1610.02413.

[7] Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning, 2019. https://arxiv.org/abs/1808.10013.

[8] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, August 2008. https://www.researchgate.net/publication/221654695_Discrimination-aware_data_mining.

[9] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, April 2017. http://dx.doi.org/10.1145/3038912.3052660.

[10] Indre Zliobaite. On the relation between accuracy and fairness in binary classification, 2015. https://arxiv.org/abs/1505.05723.

Appendices

Appendix A

Installing AIF360

1. In this book we will use Python in Jupyter notebooks from the Anaconda Python distribution platform. If you don't already have it download and install it.

2. Create an environment named `mbml`. Using the command line interface (CLI):

   ```
   \$ conda create --name mbml python=3.7
   ```

3. Activate your new environment:

   ```
   $ conda activate mbml
   ```

4. This book is a work in progress. As part of analysing the metrics and methods it uses code that is not yet available with the library[1]. Once it is merged, you will just be able to just pip install the aif360 library. Until then you must clone this fork of AIF360:

   ```
   $ git clone https://github.com/leenamurgai/AIF360.git
   ```

5. Download the notebook `mbml_german.ipynb` from Manning's GitLab repository and save it in the "AIF360/examples" folder.

6. You should now be able to open and run the notebook from the CLI as you usually would:

   ```
   $ jupyter notebook mbml_german.ipynb
   ```

[1] If you're interested, here is the open pull request.

Appendix B

Performance Metrics

B.1 Confusion Matrix Metrics

B.1.1 Performance Metrics

Table B.1: Summary of performance metrics for a binary classifier

		Ground Truth		
		$y = 1$	$y = 0$	Metric
Prediction	$\hat{y} = 1$	True Positive	False Positive Type I Error	Positive Predictive Value[a] $\mathbb{P}(\hat{y} = y \mid \hat{y} = 1)$
	$\hat{y} = 0$	False Negative Type II Error	True Negative	Negative Predictive Value $\mathbb{P}(\hat{y} = y \mid \hat{y} = 0)$
Metric		True Positive Rate[b] $\mathbb{P}(\hat{y} = y \mid y = 1)$	True Negative Rate $\mathbb{P}(\hat{y} = y \mid y = 0)$	Accuracy $\mathbb{P}(\hat{y} = y)$

[a] Positive Predictive Value = Precision
[b] True Positive Rate = Recall

B.1.2 Error Metrics

Table B.2: Summary of error rate types for a binary classifier

		Ground Truth		
		$y = 1$	$y = 0$	Error Rate Type
Prediction	$\hat{y} = 1$	True Positive	False Positive Type I Error	False Discovery Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 1)$
	$\hat{y} = 0$	False Negative Type II Error	True Negative	False Omission Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 0)$
Error Rate Type		False Negative Rate $\mathbb{P}(\hat{y} \neq y \mid y = 1)$	False Positive Rate $\mathbb{P}(\hat{y} \neq y \mid y = 0)$	Error Rate $\mathbb{P}(\hat{y} \neq y)$

Appendix C

Rules of probability

C.1 Discrete random variables

C.1.1 Sum rule

$$\mathbb{P}(x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(x, y) \tag{C.1}$$

C.1.2 Product rule

$$\mathbb{P}(x, y) = \mathbb{P}(y|x)\mathbb{P}(x) \tag{C.2}$$

C.1.3 Bayes' rule

$$\mathbb{P}(y|x) = \frac{\mathbb{P}(x|y)\mathbb{P}(y)}{\mathbb{P}(x)} \tag{C.3}$$

C.1.4 Independence

For $X \perp Y$

$$\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y) \tag{C.4}$$
$$\mathbb{P}(y|x) = \mathbb{P}(y) \tag{C.5}$$

C.1.5 Conditional Independence

For $X \perp Y|Z$ (or equivalently $Y \perp X|Z$ by symmetry)

$$\mathbb{P}(x|y, z) = \mathbb{P}(x|z) \tag{C.6}$$

and by symmetry,

$$\mathbb{P}(y|x, z) = \mathbb{P}(y|z) \tag{C.7}$$

Using the product rule followed by C.7 we also have,

$$\begin{aligned} \mathbb{P}(x, y|z) &= \mathbb{P}(y|x, z)\mathbb{P}(x|z) \\ &= \mathbb{P}(y|z)\mathbb{P}(x|z) \end{aligned} \tag{C.8}$$

C.2 Continuous random variables

C.2.1 Sum rule

$$f_X(X) = \int f_{X,Y}(x, y)\,\mathrm{d}y \tag{C.9}$$

C.2.2 Product rule

$$f_{X,Y}(x, y) = f_{Y|X}(x, y) f_X(x) \tag{C.10}$$

C.2.3 Bayes' rule

$$f_{Y|X}(x, y) = \frac{f_{X|Y}(x, y) f_Y(y)}{f_X(x)} \tag{C.11}$$

C.2.4 Independence

For $X \perp Y$

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \tag{C.12}$$
$$f_{Y|X}(x, y) = f_Y(y) \tag{C.13}$$

C.2.5 Conditional Independence

For $X \perp Y|Z$ (or equivalently $Y \perp X|Z$ by symmetry)

$$f_{X|Y,Z}(x, y, z) = f_{X|Z}(x, z) \tag{C.14}$$

and by symmetry,

$$f_{Y|X,Z}(x, y, z) = f_{Y|Z}(y, z) \tag{C.15}$$

Using the product rule followed by C.15 we also have,

$$\begin{aligned} f_{X,Y|Z}(x, y, z) &= f_{Y|X,Z}(x, y, z) f_{X|Z}(x, z) \\ &= f_{Y|Z}(y, z) f_{X|Z}(x, y) \end{aligned} \tag{C.16}$$

Appendix D

Solutions to exercises

D.1 Chapter 3 Exercises

D.1.1 Balanced outcomes

Exercise: Normalised prejudice index

Write a function that takes two arrays y and z of categorical features and returns the normalised prejudice index.

1. Compute the probability distributions $\mathbb{P}(y)$, $\mathbb{P}(z)$ and $\mathbb{P}(y, z)$. Note that these can be thought of as the frequency with which each event occurs.
2. Compute the entropies $H(y)$ and $H(z)$ shown in equation (3.3) and use these to compite the normalising factor, $\sqrt{H(y)H(z)}$.
3. Compute the mutual information $I(z, y)$ shown in equation (3.1) and divide by the normalising factor.

You can test your implementation against scikit-learn's:
sklearn.metrics.normalized_mutual_info_score.

Listing D.1: Calculating the normalised prejudice index

```
# Import the necessary classes
import pandas as pd
import scipy.stats as ss

def normalised_mutual_information(x, y):
    """normalised mutual information between x and y"""

    # Compute the probability distributions
    px   = x.value_counts(normalize=True)
    py   = y.value_counts(normalize=True)
    pxy  = pd.Series(zip(x,y)).value_counts(normalize=True)

    # Compute the normalising factor
    norm = math.sqrt( ss.entropy(px) * ss.entropy(py)  )

    # Compute mutual information, divide by the normalising factor
```

```
    # and return the result
    return sum([p * math.log(p / (px[xy[0]] * py[xy[1]]))
                for xy, p in p_xy.items()]) / norm
```

Exercise: Statistical parity difference maximum

Show that

$$d_{\max} = \min\left\{\frac{\mathbb{P}(\hat{Y}=1)}{\mathbb{P}(Z=1)}, \frac{\mathbb{P}(\hat{Y}=0)}{\mathbb{P}(Z=0)}\right\}.$$

We can write statistical parity difference as

$$d = \mathbb{P}(\hat{Y}=1|Z=1) - \mathbb{P}(\hat{Y}=1|Z=0).$$

Let's rewrite this with advantaged and disadvantaged outcomes and groups to make it more concrete,

$$d = \mathbb{P}(y^+|z^+) - \mathbb{P}(y^+|z^-) = \frac{\mathbb{P}(y^+,z^+)}{\mathbb{P}(z^+)} - \frac{\mathbb{P}(y^+,z^-)}{\mathbb{P}(z^-)} \leq \frac{\mathbb{P}(y^+)}{\mathbb{P}(z^+)}.$$

This maximal value occurs when

$$\mathbb{P}(y^+,z^+) = \mathbb{P}(y^+) \quad \text{and} \quad \mathbb{P}(y^+,z^-) = 0;$$

that is, when all members of the advantaged class, receive the advantaged outcome. We can also write,

$$\begin{aligned} d = \mathbb{P}(y^+|z^+) - \mathbb{P}(y^+|z^-) &= \mathbb{P}(y^-|z^-) - \mathbb{P}(y^-|z^+) \\ &= \frac{\mathbb{P}(y^-,z^-)}{\mathbb{P}(z^-)} - \frac{\mathbb{P}(y^-,z^+)}{\mathbb{P}(z^+)} \leq \frac{\mathbb{P}(y^-)}{\mathbb{P}(z^-)}. \end{aligned}$$

Here the maximal value occurs when

$$\mathbb{P}(y^-,z^-) = \mathbb{P}(y^-) \quad \text{and} \quad \mathbb{P}(y^-,z^+) = 0;$$

that is, when all members of the disadvantaged class, receive the disadvantaged outcome. Thus,

$$d_{max} = \min\left\{\frac{\mathbb{P}(y^+)}{\mathbb{P}(z^+)}, \frac{\mathbb{P}(y^-)}{\mathbb{P}(z^-)}\right\}.$$

Note that,

$$\frac{\mathbb{P}(y^+)}{\mathbb{P}(z^+)} = \frac{\mathbb{P}(y^-)}{\mathbb{P}(z^-)} \quad \Leftrightarrow \quad \mathbb{P}(y_+) = \mathbb{P}(z_+);$$

that is, when all members of the advantaged class, receive the advantaged outcome and all members of the disadvantaged class, receive the disadvantaged outcome.

Exercise: Odds ratio

Show that the odds ratio is always greater than or equal to one in the case where $\hat{Y} = 1$ in the advantaged outcome and $Z = 1$ is the privileged group.

$$r_{\text{odds}} = \frac{\mathbb{P}(\hat{Y}=1|Z=1)\mathbb{P}(\hat{Y}=0|Z=0)}{\mathbb{P}(\hat{Y}=0|Z=1)\mathbb{P}(\hat{Y}=1|Z=0)}.$$

Let $\hat{Y} = 1$ be the advantaged outcome and $Z = 1$ be the privileged group, then we can write,

$$r_{\text{odds}} = \frac{\mathbb{P}(\hat{y}_+|z_+)\mathbb{P}(\hat{y}_-|z_-)}{\mathbb{P}(\hat{y}_-|z_+)\mathbb{P}(\hat{y}_+|z_-)}$$

In this case, since $\mathbb{P}(\hat{y}_+|z_+) > \mathbb{P}(\hat{y}_+|z_-)$ and $\mathbb{P}(\hat{y}_-|z_-) > \mathbb{P}(\hat{y}_-|z_+)$, the numerator is always greater than the denominator and the odds ratio will be greater than one.

D.1.2 Balanced errors

Exercise: Sufficiency Show that sufficiency is satisfied if and only if the false omission rate and false discovery rate are equal for all groups.

Sufficiency implies

$$\mathbb{P}(y|\hat{y}, z) = \mathbb{P}(y|\hat{y}).$$

For the simplest case of a binary classifier where we have a single sensitive binary feature. We can write this requirement as two conditions,

$$\mathbb{P}(Y = 1|Z = 1, \hat{Y} = 1) = \mathbb{P}(Y = 1|Z = 0, \hat{Y} = 1),$$
$$\mathbb{P}(Y = 1|Z = 1, \hat{Y} = 0) = \mathbb{P}(Y = 1|Z = 0, \hat{Y} = 0).$$

Recall that $\mathbb{P}(Y = 1|\hat{Y} = 1)$ is the positive predictive value (PPV) of the classifier and $\mathbb{P}(Y = 1|\hat{Y} = 0)$ is the false omission rate (FOR). We see then that sufficiency requires the positive predictive value to be the same for all values of the sensitive feature and the false omission rate to be the same for all values of the sensitive feature. Note that the positive predictive value is balanced if and only if the false discovery rate is balanced, so thinking in terms of error metrics only, separation requires the false discovery and false omission rates to be balanced.

D.1.3 Incompatibility of fairness criteria

Separation versus Sufficiency

Exercise: Predictive values

Prove the results given in equations (3.13) and (3.14).

We want to write the positive and negative predictive values (PPV and NPV respectively) in terms of the true positive, false positive and acceptance rates (TPR, FPR and p respectively). We start by looking at some relationships between the elements of a confusion matrix shown in table D.1. where $n = TP + FP + FN + TN$ denotes the total number of data points. Using the equations in the final row of the table we can write,

$$pTPR = \frac{TP}{n}, \qquad (1-p)FPR = \frac{FP}{n},$$
$$p(1-TPR) = \frac{FN}{n}, \qquad (1-p)(1-FPR) = \frac{TP}{n}.$$

Finally, we can substitute these into our expressions for PPV and NPV in the right hand column of table D.1 to find the relationships in equations (3.13) and (3.14).

$$PPV = \frac{pTPR}{pTPR + (1-p)FPR}$$
$$NPV = \frac{(1-p)(1-FPR)}{p(1-TPR) + (1-p)(1-FPR)}.$$

Table D.1: Confusion matrix

		Ground Truth		
		$y = 1$	$y = 0$	
Prediction	$\hat{y} = 1$	True Positive TP	False Positive FP	$PPV = \frac{TP}{TP + FP}$
	$\hat{y} = 0$	False Negative FN	True Negative TN	$NPV = \frac{TN}{FN + TN}$
		$TPR = \frac{TP}{TP + FN}$ $1 - TPR = \frac{FN}{TP + FN}$ $p = \frac{TP + FN}{n}$	$FPR = \frac{FP}{FP + TN}$ $1 - FPR = \frac{TN}{FP + TN}$ $1 - p = \frac{FP + TN}{n}$	

Exercise: Separation versus sufficiency

Show that for separation and sufficiency to hold equations (3.15) and (3.16) must hold for for any pair of groups $Z = a$ and $Z = b$.

For separation to hold the true positive and false positive rates must be constant across all values of the sensitive features. For sufficiency to hold the positive and negative predictive values must be constant across all values of the sensitive features. For brevity we shal use a subscript to denote conditioning on Z, for example $p_z = \mathbb{P}(Y = 1|Z = z)$. For both separation and sufficiency to hold, we must have

$$\begin{aligned}
& PPV_a = PPV_b \\
\Leftrightarrow \quad & \frac{p_a TPR}{p_a TPR + (1 - p_a)FPR} = \frac{p_b TPR}{p_b TPR + (1 - p_b)FPR} \\
\Leftrightarrow \quad & p_b TPR[p_a TPR + (1 - p_a)FPR] = p_a TPR[p_b TPR + (1 - p_b)FPR] \\
\Leftrightarrow \quad & p_b TPR(1 - p_a)FPR = p_a TPR(1 - p_b)FPR \\
\Leftrightarrow \quad & TPR(p_b - p_a)FPR = 0
\end{aligned} \tag{D.1}$$

Similarly, we must also have,

$$\begin{aligned}
& NPV_a = NPV_b \\
\Leftrightarrow \quad & \frac{(1 - p_a)(1 - FPR)}{p_a(1 - TPR) + (1 - p_a)(1 - FPR)} = \frac{(1 - p_b)(1 - FPR)}{p_b(1 - TPR) + (1 - p_b)(1 - FPR)} \\
\Leftrightarrow \quad & (1 - p_b)(1 - FPR)[p_a(1 - TPR) + (1 - p_a)(1 - FPR)] \\
& = (1 - p_a)(1 - FPR)[p_b(1 - TPR) + (1 - p_b)(1 - FPR)] \\
\Leftrightarrow \quad & (1 - p_b)(1 - FPR)p_a(1 - TPR) = (1 - p_a)(1 - FPR)p_b(1 - TPR) \\
\Leftrightarrow \quad & (1 - FPR)(p_b - p_a)(1 - TPR) = 0
\end{aligned} \tag{D.2}$$