Integrating Machine Learning into Fair Inference

Haoyu Wang[1*], Hanyu Hu[2], Mingrui Zhuang[3] and Jiayi Shen[4]
[1]*Software School of Yunnan University, Kunming, China*
[2]*The Affiliated Tianhe School of Guangdong Experimental Middle School, Guangzhou, China*
[3]*Beijing NO.2 Middle School, Beijing, China*
[4]*Suzhou High School, Suzhou, China*

Keywords: Fairness, Machine Learning, Causal Inference, Path-Specific Effect, Natural Direct Effect, Counterfactual.

Abstract: With the boom of machine learning, fairness is an issue that needs to be concerned. The three main perspectives of this paper provide a thorough look at the fairness problem: First, we introduce a handy tool for causal inference, that is, causal graph, and apply formulas like adjustment formula, back-door formula, and front-door formula to see the effect of interventions, which can help with the fairness. Then some approaches to measure the fairness are introduced: natural direct path and path-specific effect. Finally, we use counterfactual inference further to study fairness with the help of causal graphs and integrate LFR, a model focusing on both group fairness and individual fairness.

1 INTRODUCTION

Nowadays, artificial intelligence is widely used in our lives. With the increasing use of automated decision-making systems, people are concerned about bias and discrimination in these systems. Since systems trained with the historical data will inherit the previous biases, we need to make a fair decision so that there are not unduly biased for or against protected subgroups in the population, such as the female, the elderly, and the ethnic minorities. The problem is deemed as fairness in machine learning. There are two crucial dimensions of fairness: group fairness and individual fairness. Group fairness ensures that the overall proportion of members in a protected group receiving positive or negative classification is identical to the proportion of the population as a whole. On the other hand, individual fairness achieves that any two similar individuals should be classified similarly.

Causal inference serves as a solution to fairness. Causality is prevalent in the universe. For example, the cure of a disease is due to using a specific drug. Machines can answer questions like whether this drug should be used to make a causal inference. Some causalities, however, may lead to discrimination on specific groups, damaging fairness. If gender is the cause of whether he/she gets the offer, there is no doubt that the employer biases against some particular gender, so this is unfair. In order to ensure fairness, machine learning systems developed to decide whether an employee can get the offer should not consider gender.

Machines are good at predicting probability, but it is difficult to predict results after intervening. Counterfactual, as its name indicates, captures notions of something that has not happened could happen with some conditions contrary to the fact. As a subset of causal inference, counterfactual inference appears to measure the fairness of machine learning systems based on causal inference. Counterfactuals are pretty common in our daily lives: every sentence in the subjunctive mood can be considered a counterfactual problem. When you hear your friend saying, "If I had done my assignment better, I would have got a better final score," you cannot immediately check whether this sentence is correct because there is no easy way to find someone with the same quality as him/her. Here "done my assignment better" is counterfactual because "your friend" signifies that he/she did not do homework well.

It is only recently that some researchers have considered this issue. Several papers have aimed to achieve group fairness, and some achieve individual fairness. Nabi and Shpitser have considered the problem of fair statistical inference on outcomes in a setting where we wish to minimize discrimination concerning a particular sensitive feature, such as

Wang, H., Hu, H., Zhuang, M. and Shen, J.
Integrating Machine Learning into Fair Inference.
DOI: 10.5220/0011908000003613
In *Proceedings of the 2nd International Conference on New Media Development and Modernized Education (NMDME 2022)*, pages 144-154
ISBN: 978-989-758-630-9

race or gender (Nabi, Shpitser, 2018). A paper has investigated real-world applications that have shown biases in various ways and listed different sources of biases that can affect AI applications (Binns 2018). Another paper draws on existing moral and political philosophy work to elucidate emerging debates about fair machine learning (Mehrabi, Morstatter, Saxena, Lerman, Galstyan, 2021). These papers have clearly illustrated the fairness in statistical studies and even provided some application scenarios that correlate with machine learning. However, they do not contain a systematic approach to integrating machine learning into the field of fairness.

The later parts of this paper are organized below: First, we introduce a handy tool for causal inference, that is, causal graph, and apply formulas like adjustment formula, back-door formula, and front-door formula to see the effect of interventions, which can help with the fairness. Then some approaches to measure the fairness are introduced: natural direct path and path-specific effect. Finally, we use counterfactual inference further to study fairness with the help of causal graphs and integrate LFR, a model proposed by Zemel, Wu, Swersky, Pitassi, and Dwork that focuses on both group fairness and individual fairness (Zemel, Wu, Swersky, Pitassi, Dwork, 2013). All the theories and experiments are based on the Community and Crime Dataset from the UCI repository (Acharya, Blackwell, Sen, 2016).

2 CAUSAL GRAPHS AND CAUSAL INFERENCE

2.1 Introduction to Causal Graphs

Causal graphs can be used in describing the causal relationship between attributes. As a visual model of causality between variables in a system, the causal graph makes it easier to draw realistic causal inferences, like doing exercises "causes" lower blood pressure. It plays a role by stimulating the identification of more potential confounding factors and the source of selection bias.

A causal graph is a directed acyclic graph (DAG), including a collection of nodes (also referred to vertices on some occasions) and directed edges. So the graph can be represented by $G = \{N, E\}$, where N is the set of nodes and E is the set of edges. An example of a causal graph is shown as Fig. 1:

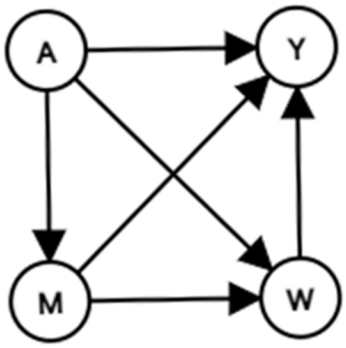

Figure 1: An example of a causal graph

In Fig. 1, we see that

$$N = \{A, M, W, Y\} \quad (1)$$

$$E = \{(A, M), (A, W), (A, Y), (M, W), (M, Y), (W, Y)\} \quad (2)$$

Each node in the graph represents a variable. We use solid nodes to represent observed variables and dashed nodes to represent unobserved variables. An edge indicates the causal effect between two variables, like (A, M), the edge directing from A to M, which means that A is the "cause" of M. Here we also say A is the parent node of M. Two nodes are adjacent if they are connected by an edge. There are paths between 2 nodes if they are connected by some sequences of edges. For example, A and Y are adjacent. From A to W, there are 2 paths: $A \to W$, $A \to M \to W$.

2.2 The Crime Dataset and Our Causal Graphs

In later parts of this paper, experiments are conducted on the Community and Crime Dataset, retrieved from the UCI repository (Acharya, Blackwell, Sen, 2016). Later we will call it Crime Dataset for short. It contains 1994 samples, and each of them contains 128 attributes. The first 4 attributes are *state, county, community, communityname*, which are nominal data, serving as the identifier of the sample and not for prediction. The 5th attribute is *fold*, whose values are integers ranging from 1 to 10, used for 10-fold cross validation. The 6th to the 127th attributes are social and socio-economic data that is plausible to do with the crime rate, such as *PolicPerPop* (police officers per 100K population). The last attribute is the goal attribute to be predicted: *ViolentCrimesPerPop* (total number of violent crimes per

100K population).

It is worth noting that the values of the 6th to the 128th attributes have been normalized into the decimal range from 0.00 to 1.00, using an unsupervised, equal-interval binning method. In this way, attributes retain their distribution and skew (for example, the population attribute has a mean value of 0.06 because most communities are small). The normalization preserves rough ratios of values within an attribute.

The dataset we used combines socio-economic data, law enforcement, and crime data from the 1990 U.S. Census. Data is described based on original values and used to predict the crime rate of specific communities in the United States. Besides, there are some sensitive attributes in the data about age, gender, and race in this dataset. For our goal of fairness, we try not to let these sensitive attributes decide the crime rate. However, it is inevitable to use these attributes for some causal inferences of other non-sensitive attributes.

For later experiments, we have drawn some causal graphs based on the Crime Dataset. They are shown as Fig. 2-Fig. 6:

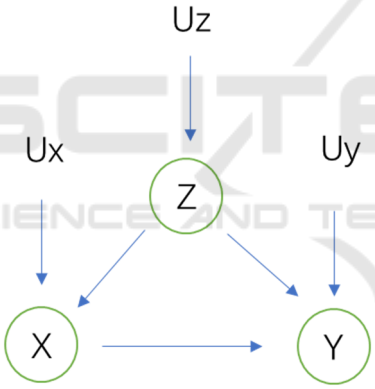

Figure 2: Causal Graph 1. X denotes the per capita income; Z denotes the percentage of people 16 and over who are employed; Y denotes the crime rate.

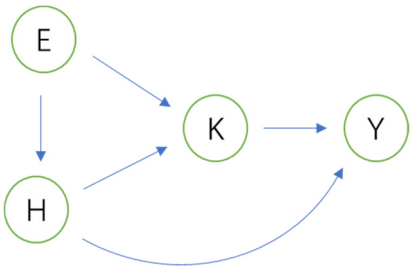

Figure 3: Causal Graph 2. H denotes percentage of people 25 and over with a bachelors degree or higher education; E denotes percent of people who do not speak English well; K denotes percentage of households with wage or salary income in 1989; Y denotes the crime rate.

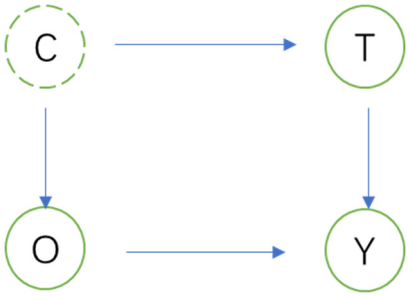

Figure 4: Causal Graph 3. O denotes the police operating budget; C denotes the commodity prices (unobserved); T denotes the percent of people using public transit for commuting; Y denotes the crime rate.

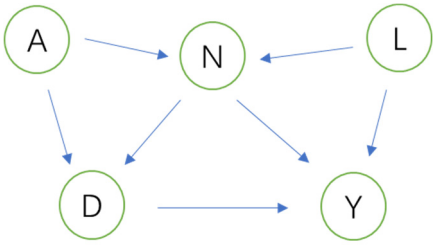

Figure 5: Causal Graph 4. A denotes percentage of kids born to never married; N denotes percentage of population who are divorced; L denotes percentage of people under the poverty level; D denotes percent of housing occupied; Y denotes the crime rate.

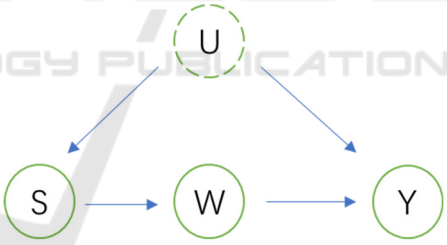

Figure 6: Causal Graph 5. S denotes number of different kinds of drugs seized; U - percent of people using drugs (unobserved);

2.3 Interventions on Causal Graphs

The ultimate goal of many statistical studies is to predict the effect of interventions. For example, we collect data on car accidents to find intervention factors to reduce the occurrence of car accidents; when we study new drugs, we intervene by asking patients to take drugs and observe the reaction of patients after taking drugs. When randomized controlled trials are not feasible, we often implement observational studies to obtain the relationship between variables by controlling specific data. Through this intervention, we can block the causal

relationship between some variables and analyze the impact of other variables.

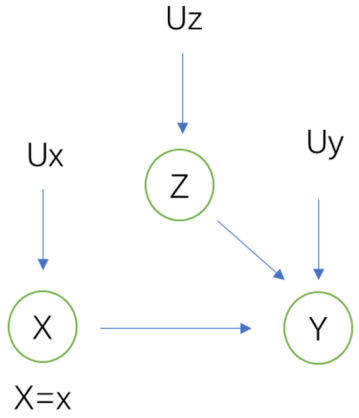

Figure 7: Causal Graph 1 after the intervention on X

In the case of Fig. 2, in order to determine the effect of X on Y, we simulate the intervention in the form of a graph surgery (as in Fig. 7 above, where X is controlled to be x, the manipulated probability is P_m. In the manipulated model of Fig. 7, the causal effect $P(Y = y|do(X = x))$ is equal to the conditional probability $P_m(Y = y|X = x)$. Combined with the primary attributes of probability and variables, we get a causal effect formula expressed by pre-intervention probability, known as adjustment formula:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|\, X = x,\, Z = z)\, P(Z = z) \quad (3)$$

There is another application of adjustment formula in Fig. 3. We want to gauge the effect of higher education (H) on crime rate (Y). We assume that people who have income are less likely to commit crimes. Using the same method as shown in (1), we get the following formula: is shown as belows

$$P(Y = y|do(H = h)) = \sum_k P(Y = y|\, H = h,\, K = k) \sum_E P(K = k,\, E = e,\, H = h)\, P(Z = z) \quad (4)$$

In the above discussion, we concluded that we should adjust for a variable's parents when we are trying to determine its effect on another variable. Nevertheless, often the variables have unobserved or inaccessible parents. In those cases, we use a simple test called the back-door criterion: given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the back-door criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X. If a set of variables Z satisfies the back-door criterion for X and Y, then the causal effect of X on Y is given by the back-door formula:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|\, X = x,\, Z = z)\, P(Z = z) \quad (5)$$

In Fig. 4, we are trying to gauge the effect of a police operating budget (O) on crime rate (Y). We have also measured people for public commuting (T), which has an effect on the crime rate. Furthermore, we know that commodity prices (C) affect both O and T, but it is an unobserved variable. Instead, we search for an observed variable that fits the back-door criterion from O to Y. We find that T, which is not a descendant of O, also blocks the back-door path $O \leftarrow C \rightarrow T \rightarrow Y$. Therefore, W meets the back-door criterion. By using the adjustment formula, we got the following formula:

$$P(Y = y|do(O = o)) = \sum_T P(Y = y|\, O = o,\, T = t)\, P(T = t) \quad (6)$$

Do operation can also be applied to some graph patterns that do not meet the back-door criterion to determine the causal effect that seems to have no solution at first. One such pattern, front-door, can identify the causal effect shown in Fig. 6, where the variable U is unobserved and hence cannot be used to block the back-door path from X to Y. A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if

1. Z intercepts all directed paths from X to Y.
2. There is no unblocked path from X to Z.
3. All back-door paths from Z to Y are blocked by X

This method can identify the causal effect in Fig. 6 through two consecutive applications of the back-door path. First, there is no back-door path from S to W. so we can immediately write the effect of S on W

$$P\big((W = w)\big|do(S = s)\big) = P(W = w|S = s) \quad (7)$$

then, the back-door path from W to Y, namely $W \leftarrow S \leftarrow U \rightarrow Y$, can be blocked by conditioning on X so that we can write the second formula like this

$$P(Y = y|do(W = w)) = \sum_S P(Y = y|\, S = s,\, W = w)\, P(S = s) \quad (8)$$

Now we chain together the two partial effects to obtain the overall effect of X on Y by summing all states' smaller z of capital Z, and we can get this. Through some changes in expression, we finally get the impact of the number of drugs on the crime rate.

$$P(Y = y|do(S = s)) = \sum_{w} P(Y = y|do(W = w))P(W = w|S = s) \quad (9)$$

2.4. Calculating the Interventions

Some formulas are deducted, though, yielding their values is another problem. In Crime Dataset, all the data for prediction and the outcome is continuous, which means that the probability with variables conditional on fixed values is insignificant. Also, it seems impossible to calculate the probability of a variable fixed on an exact value—instead, expectation matters. Nevertheless, machine learning and statistics help.

Let us start with an example: the formula for Fig. 2.

$$P(Y = y|do(X = x)) = \sum_{z} P(Y = y|X = x, Z = z)\,P(Z = z) \quad (10)$$

Since $P(Y = y)$ is hard to get, we convert it to

$$\mathbb{E}(Y|do(X = x)) = \sum_{Z} \mathbb{E}(Y|X = x, Z = z)P(Z = z) \quad (11)$$

In order to learn the effect on Y when we intervene X, we get the formula above based on the adjustment formula. Since Z is continuous, $\sum_z \mathbb{E}(Y|X = x, Z = z)\,P(Z = z)$ can be further converted to formulas with expectation. With the preliminary $X = x$, the formula means "given $X = x$ and a random selected Z, the expectation of Y", that is:

$$\mathbb{E}(Y|do(X = x)) = \sum_{Z} \mathbb{E}(Y|X = x, Z = z)P(Z = z) = \mathbb{E}(Y|X = x, Z) = \mathbb{E}(Y|X = x)\mathbb{E}(Z) \quad (12)$$

To get the expectation $\mathbb{E}(Y|X = x)$, we can train a model using the samples in the dataset that predicts Y with X. Amazingly, among many machine learning models, it is linear regression that best fit the causal relation from X to Y. The result of the prediction is shown in Fig. 8.

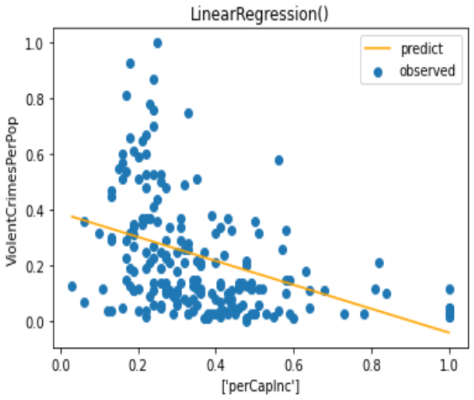

Figure 8: Diagram of the best model using *perCapInc* to predict *ViolentCrimesPerPop*

Assume that we are interested in $\mathbb{E}(Y|do(X = 0.5))$, then we input $X = 0.5$ and get predicted value 0.175, so $\mathbb{E}(Y|do(X = 0.5)) = 0.175$. $\mathbb{E}(Z)$ can be simply computed using the samples in the dataset, which is 0.501. As a result:

$$\mathbb{E}(Y|do(X = 0.5)) = \sum_{Z} \mathbb{E}(Y|X = 0.5, Z = z)P(Z = z) = \mathbb{E}(Y|X = 0.5, Z) = \mathbb{E}(Y|X = 0.5)\mathbb{E}(Z) = 0.175 \times 0.501 = 0.088 \quad (13)$$

The result indicates that, after exerting intervention $do(X = 0.5)$, Y is expected to be 0.088.

Formulas on other graphs can also be computed this way. The approach of computing intervention on continuous data is concluded as:

- Express the $P(Y = y|do(X = x))$ by expressions without do, using adjustment formula, back-door formula, front-door formula, and so on.
- Convert the probability expression to expectation, like $P(Y = y)$ to $\mathbb{E}(Y)$
- Calculate/Predict the $\mathbb{E}$:
  - The expectation of a single variable is the mean value of this variable in the dataset
  - For the expectation of compound expression like $\mathbb{E}(Y|X = x)$, build a model predicting Y using X, then input $X = x$ and use the predicted value

3 MEASURING THE FAIRNESS

3.1 Mediation and Direct Paths

Under some circumstances, we concentrate on the effect of one variable X on another variable Y in causal graphs. There may be many paths from X to Y, and some are direct while some are not. So the effect of X to Y includes the direct effect and the

indirect effect.

Mediation is encoded via a counterfactual contrast using a nested potential outcome of the form $Y(a, M(a'))$ (Nabi, Shpitser, 2018). Then a treatment like $X = a$ can be divided into two disjoint parts: one acts on Y but not M, and the other acts on M but not Y. Later, we will mainly focus on the former one, that is, the direct effect.

3.2 Natural Direct Effect

In causal mediation analysis, one quantity of interest is the natural direct effect (NDE). It is the impact of altering treatment underneath it while fixing the mediator to its unit-specific plausible value. The NDE compares the mean outcome, which is only directly influenced by the part of the treatment that will exert an effect on it, with the one under the placebo treatment (Binns 2018). Given $Y(a, M(a'))$, we define the following effects on the mean difference scale: the natural direct effect as

$$\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')]$$

which means for the outcome Y, A is set to a, and M is set to the value when A is set to a'.

3.3 Path-specific Effect

Path-specific effect (PSE) is a crucial indicator for evaluating mediation in the presence of multiple intermediaries.

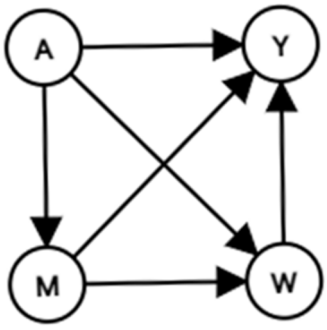

Figure 9: An example of a causal graph

From the graph Fig. 9, we can see there are four ways to go from A to Y: $A \to Y, A \to W \to Y, A \to M \to Y, A \to M \to W \to Y$. If we wish to evaluate the contribution of $A \to W \to Y$, with the presence of $A \to Y$, and $A \to M \to W \to Y$, effects along the path $A \to W \to Y$ is known as Path-specific effect. On the path of interest, A is set to the value a, and on other paths, A is set to the baseline value a'. With the concept, the path-specific effect from A to Y along the path $A \to W \to Y$ can be formulated by

$$\mathbb{E}[Y(a', W(M(a'), a), M(a'))] - \mathbb{E}[Y(a')]$$

We formalize the existence of discrimination as the existence of a particular path-specific effect. The reason why we use PSE is that when problems arise, such as gender or racial discrimination, we can issue conceptualization, make causal graphs according to the problems, and define a fair path from A (attribute about gender/race) to the outcome Y (crime rate), may be related to some media, or it is a direct-effect path, the problem will increase the PSE along these paths.

3.4 Using PSE in Our Graphs

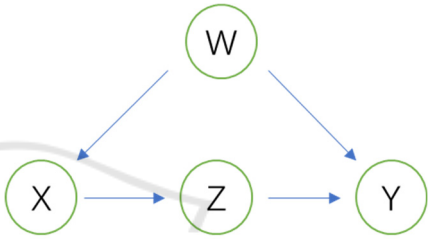

Figure 10: A causal graph we use to study the PSE. W denotes the percentage of people living in areas classified as urban; X denotes the percentage of the population that is African American; Z denotes median household income; Y denotes crime rate.

First, we focus on the causal graph Fig. 10.

It is inevitable to encounter sensitive variables in various data. When we find some paths that may be unfair, we can take some measures to avoid them. When we use Fig. 10 to estimate the impact of the variables on the crime rate, we may come to some discriminatory conclusions: the increase of African American income will reduce the crime rate. Obviously, the logical relationship between these two things is unfair. We will avoid this discrimination by choosing other paths or increasing fairness. This is where PSE works.

The path we are interested in is $W \to Y$.

PSE: $\mathbb{E}[Y(w) - \mathbb{E}[Y(w, Z(X(w'), w), X(w'))]]$

As W is set to the baseline w', X is represented with $X(w')$, Z is represented with $Z(X(w'), w)$, and Y is $Y(w, X(w'), Z(X(w'), w'))$. Changing the w' to w, since the baseline value will not have a great influence on the values we care about, so X is represented with $X(w)$, Z is represented with $Z(X(w'), w)$, and Y is $Y(w, X(w'), Z(X(w'), w))$.

Then we focus on another causal graph, above

mentioned Fig. 3.

The path we are interested in is $E \to H \to K \to Y$.

PSE: $\mathbb{E}[Y(e, H(e'), (K(H(e'), e))] - \mathbb{E}[Y(e, H(e'), e)]$

Since we want to evaluate the contribution on $E \to H \to K \to Y$, with the presence of $E \to K \to Y$, and $E \to H \to Y$, effects along the path $E \to H \to K \to Y$ is actually the path-specific effect. As E was set to the baseline value e', since baseline value will not affect the value of the path we are interested in, H will be represented with $H(e')$.

4 COUNTERFACTUAL INFERENCE

4.1 Introduction to Counterfactual Inference

Unlike the formulas introduced in the last sessions, which focus on the whole dataset with a large number of samples, the counterfactual inference is the study of the counterfactual effect of a single sample. But some preliminary to computing the counterfactual depends on the whole dataset, also.

In accord with Chapter 4 of *Causal Inference in Statistics: a Primer* (Pearl, Glymour, Jewell, 2016), we define $Y_{X=x}(u) = y$ as "Y would be y if X was x, with $\vec{u} = u$" where $\vec{u}$ is the vector of exogenous variables, like $\{u_x, u_y, u_z\}$ in this example. Assume that "your friend" had put 1.5 times of energy into the assignment. The answer to the grade example can be denoted as $Y_{Z=1.5Z}(u_x, u_y, u_z)$. Here X, Y, Z are given, through which we can calculate u_x, u_y, u_z.

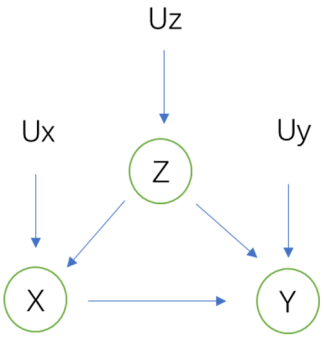

Figure 1: A causal graph of the grade example. Z: assignment; X: performance score; Y: final score

According to Fig. 11 which describes the counterfactual problem at the beginning, that your friend saying, "If I had done my assignment better, I would have got a better final score", we see that assignment can affect the grades both directly and indirectly. To quantify the effects, we assume that:

$$Z = u_Z \tag{14}$$

$$X = 2Z + u_X \tag{15}$$

$$Y = X + 3Z + u_Y \tag{16}$$

For the sample of "your friend" in the example, assume that $Z = 1, X = 3, Y = 5$, we can substitute these values into the equations above and yield $u_Z = 1, u_X = 1, u_Y = -1$. The answer became $Y_{Z=1.5}(1,1,-1)$. If $Z = 1$ was replaced with $Z = 1.5$, then:

$$X = 2Z + u_X = 2 \times 1.5 + 1 = 4 \tag{17}$$

$$Y = X + 3Z + u_Y = 5 + 3 \times 1.5 - 1 = 8.5 \tag{18}$$

Since Y would be 8.5 if Z was modified to 1.5, we can conclude that the final score of "your friend" would gain a 70% increase if he/she had put 1.5 times of energy into the assignment. In other words, after exerting a counterfactual effect, $Y_{Z=1.5}(1,1,-1) = 8.5$, while $Y = 5$ originally.

In later parts, we will discuss some approaches to compute the counterfactual based on some chosen attributes in Crime Dataset, then model the relations between them.

4.2 Model the Relations Using Machine Learning Methods

In causal graphs, each edge can be regarded as a relation between two variables. For a node Y in the graph with its parent nodes being $X_1, X_2, \cdots, X_n$ (in other words, for each integer i satisfying $i \in [1, n]$, there is an edge from X_i to Y), the relations can be modeled as $Y = f_Y(X_1, X_2, \cdots, X_n)$. If exogenous variable u_Y is into consideration, the equation will become $Y = f_Y(X_1, X_2, \cdots, X_n) + u_Y$. For example, on condition that $n = 5$, the relation among X_1, X_2, X_3, X_4, X_5 and Y is illustrated as Fig. 12.

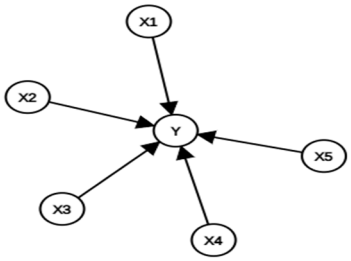

Figure 12: An example of such causal graphs: $n = 5$, and each of $X_1, X_2, \cdots, X_n$ directs to Y. In this situation, we represent Y as $Y = f_Y(X_1, X_2, \cdots, X_n) + u_Y$.

Although counterfactual inference focuses on the effect of a single sample, a large amount of samples in the dataset is required to train the models. When modeling the relations, exogenous variables like u_X, u_Y can be seen as the noise with a mean of 0. However, when computing the outcome with counterfactual assumptions, exogenous variables differ in different samples, which will be discussed in 4.3.

The simplest way to model the function is assuming they are linearly correlative: $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + u_Y$, like the example in 4.1. However, as we plot the relations between two variables, it is clear that the linear model cannot best fit the relation, which leads to a relatively high bias, as Fig. 13 shows.

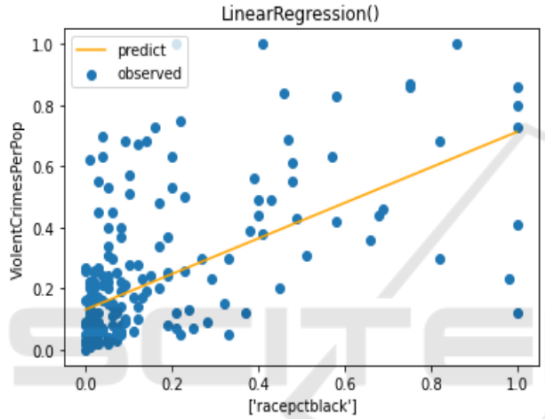

Figure 13: Adoption of linear regression on two attributes on Crime Dataset: X-axis is *racepctblack* (percentage of African Americans) and Y-axis is *ViolentCrimesPerPop* (total number of violent crimes per 100K population)

In our causal graphs with attributes from Crime Dataset, we tested 4 different machine learning models: Linear Regression, Decision Tree, Support Vector Regression, and Bayesian Ridge. For each model, we trained it using 10-fold cross validation, which provides a reasonable assessment of the performance of the model. Then we select the best model for each causal function according to the min-squared error (MSE) of the prediction. The function is represented as a node, all its parent nodes and the edges between them in the causal graph.

Take the causal graph Fig. 3 as an example. The results of model fitting are shown in Fig. 14-Fig. 16.

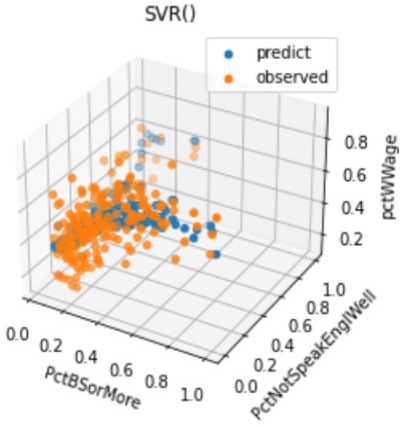

Figure 14: Scatter diagram of the best model using *PctBSorMore, PctNotSpeakEnglWell* to predict *pctWWage*

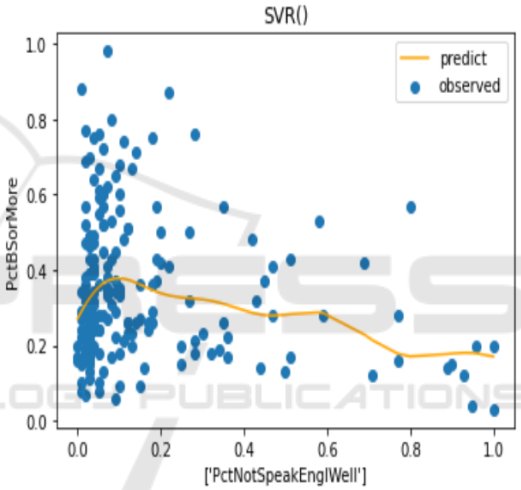

Figure 15: Diagram of the best model using *PctNotSpeakEnglWell* to predict *PctBSorMore*

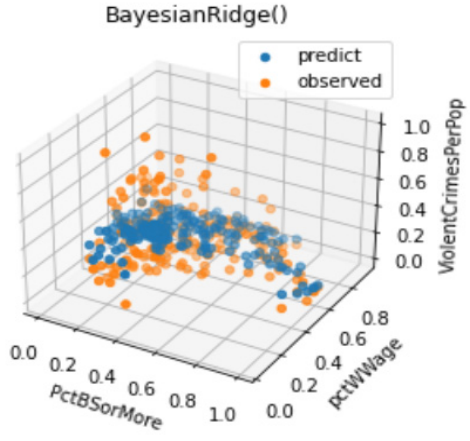

Figure 16: Scatter diagram of the best model using *PctBSorMore, pctWWage* to predict *ViolentCrimesPerPop*

From the results, we find that SVR (support vector regressor) best fits the former two relations, while bayesian ridge best fits the last relation. The other graphs are conducted the same things. We save the best models and apply them in later steps of counterfactual inference.

4.3 Compute the Counterfactuals

Chapter 4 of *Causal Inference in Statistics: a Primer* (Pearl, Glymour, Jewell, 2016) indicates that there are 3 steps to compute the counterfactual. Combined our work with the illustration of the book, we conclude that our steps are:

- Abduction: Use evidence of an actual sample to determine the value of exogenous variables U;
- Action: substitute the equations for the goal attribute Y with the interventional values $X = x$, resulting in the modified set of equations $Y_{X=x}(U)$;
- Prediction: compute the implied distribution on attributes except X based on U and models built in the last session, then the predicted value $\hat{Y}$ can represent $Y_{X=x}(U)$.

Our experiment focused on causal graph Fig. 10, manually selecting a community called Bethlehemtownship from Crime Dataset. Its values are $\{W = 0.43, X = 0.02, Z = 0.50, Y = 0.03\}$. We exerted counterfactual effect $X = 0.23$, since X denotes the percentage of the population that is African American with its 75% quantile being 0.23. By computing $Y_{X=0.23}(U)$, that is, "what if there were more African Americans in this community" we can judge whether the models trained in 4.2 discriminate against the specific race.

Table 1. Result of the counterfactual experiment

	X	W	Z	Y
Original sample	0.02	0.43	0.50	0.03
Sample after counterfactual effect	0.23 (presupposed)	0.43 (original)	0.29	0.17

According to Table 1 showing the results, we can say that, unfortunately, the models we trained are unfair. Since we simply adjust X with W unaltered, the predicted crime rate increased significantly. It is worth noting that mediator Z changes as well, which suggests that it is already unfair halfway to the outcome variable. But there are ways to tackle this problem, such as LFR introduced in the next session.

4.4 Application of Learning Fair Representations in Counterfactual Inference

Learning fair representations, abbreviated to LFR, is a machine learning-based model which takes fairness into consideration, both group fairness and individual fairness while assuring the accuracy of prediction at the same time (Zemel, Wu, Swersky, Pitassi, Dwork, 2013). LFR works on the dataset that is divided into protected group and unprotected group, and then it tries to attain the group fairness between the two groups.

To integrate LFR in our experiment for a prediction with better fairness, we adopted the criterion of LFR, that is:

$$L = a_z L_z + a_x L_x + a_y L_y \tag{19}$$

In this formula, a_z, a_x, a_y are hyperparameters mastering the tradeoff among L_z, L_x, L_y, which are three disparate measurements to be minimized: L_z measures the gap between the protected group and unprotected group in the prediction; L_x means the information loss in the prediction; L_y scales how inaccurate the prediction is, so the lower L_y is, the more accurate the model predicts. The detailed calculation of L_z, L_x, L_y is illustrated in *Learning fair representations* (Zemel, Wu, Swersky, Pitassi, Dwork, 2013).

For a node Y in our causal graph with parent nodes $X_1, X_2, \cdots, X_n$, if X_i is a sensitive attribute, then we separate protected and unprotected groups depending on the value of X_i, and train an LFR model to fit $Y = f_Y(X_1, X_2, \cdots, X_n)$.

In our experiment, including some sensitive attributes, like the one in 4.3 on Fig. 10, we can exceptionally adopt LFR on unfair paths, while fair paths are simply assembled the best model selected in 4.2. In this experiment, we adopted LFR on $X \to Z$, and the results are shown as below:

Table 2. Result of the counterfactual experiment, using LFR on $X \to Z$

	X	W	Z	Y
Original sample	0.02	0.43	0.50	0.03
Sample after counterfactual effect	0.23 (presupposed)	0.43 (original)	0.29	0.17
Sample after counterfactual effect (using LFR on $X \to Z$)	0.23 (presupposed)	0.43 (original)	0.39	0.05

Comparing the prediction of counterfactual effect by models integrating LFR and the prediction with no concern on fairness (see Table 2), it is explicit that LFR makes it a little fairer since Z (median household income) predicted by X is not that low. The outcome Y (total number of violent crimes per 100K population) is relatively low.

Moreover, we notice that the adopting LFR on $X \rightarrow Z$ makes predicted Z above its average of 0.36. If we divide the dataset into the protected group and unprotected group according to X (samples with $X > 0.23$ is divided into the protected group), we see that the mean of Z in the protected group is 0.24 (below the average), while the value in unprotected group is 0.40 (above the average). The predicted Z using LFR, interestingly, is close to 0.40.

5 CONCLUSION

In this paper, we took fairness in machine learning as a starting point since it became a social issue today. We chose Community and Crime Dataset because it has lots of attributes, including some sensitive ones, and we conducted experiments on it to explore some approaches to improve fairness. Our research was a glimpse of the world of fairness from three different perspectives.

In causal inference, we focused on the effect of intervening some variables on the outcome variable, which we denote as $P(Y = y|do(X = x))$, inspired by some previous work (Binns 2018). Since the effect of intervention is not observable, we need to convert expression with do to probability conditional on observable variables. The adjustment formula, back-door formula, and front-door formula are of great importance.

However, the dataset we chose contains mostly continuous data, making the probability of a concrete point meaningless. In our research, we innovatively proposed an approach that replaces P with $\mathbb{E}$, the expectation. For example, $P(Y = y|do(X = x))$ is equivalent to $\mathbb{E}\big(Y\big|do(X = x)\big)$. Then by either calculating directly or predicting with machine learning models, we can get the expectation of Y while intervening on X.

Later, we apply some measurements for fairness on our dataset: natural direct path and path-specific effect, proposed by Nabi and Shpitser (Nabi, Shpitser, 2018). They work when there are multiple paths from X, the variable we are interested in, to Y, the outcome variable. By shadowing the mediators between X to Y, we can learn the effect of X to Y specific to certain fair paths. For example, it is unfair to make gender directly affect the offer, but it is fair that gender influences the capabilities concerning the offer.

Finally, we studied the counterfactual inference. The goal of this part is computing $Y_{X=x}(u)$, which means the value Y would be if X was x, with exogenous variables $U = u$. The first step is to build models for edges in causal graphs, which signify the causal relationship between variables. We tried 4 different machine learning models: Linear Regression, Decision Tree, Support Vector Regression, and Bayesian Ridge, and trained each of them by 10-fold cross validation. Then we computed the counterfactual effect, according to the approaches introduced in Chapter 4 of *Causal Inference in Statistics: a Primer* (Pearl, Glymour, Jewell, 2016).

Since we found that building the models as mentioned above without concern on fairness may lead to discrimination on certain protected groups, we introduced learning fair representations to improve fairness. This model performs well on both group and individual fairness (Zemel, Wu, Swersky, Pitassi, Dwork, 2013). The results of our experiment showed that after integrating LFR on some counterfactual problems, the fairness was greatly improved while the accuracy remained at a relatively high level.

Indeed, there are many limitations in our current research. For PSE and NDE, we tried the same algorithm as the causal inference part (2), that is, replacing probability to expectation and calculating with the help of machine learning models. However, it did not work well because the prediction values overfocused the fairness criteria and had a significant error.

REFERENCES

Acharya, A., Blackwell, M., & Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. American Political Science Review, 110(3), 512-529.

Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency (pp. 149-159). PMLR.

Pearl, J., Glymour, M., and Jewell, N. Causal Inference in Statistics: a Primer. Wiley, 2016.

Nabi, R., & Shpitser, I. (2018, April). Fair inference on outcomes. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in

machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In International conference on machine learning (pp. 325-333). PMLR.