Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220553

The Application of Machine Learning Algorithm and Causal Inference in Analyzing Student Performance

Ganyu Li[1,6],Yuetong Lin[2], Zhuotong Zhong[3], Bryan Zhang[4] and Kaiwen Zheng[5]

[1]Guangzhou Institute of Science and Technology, Guangzhou, 510540, China
[2]ShenShen College of International Education, Shenzhen, 518017, China
[3]Walnut High School, CA, 91789-2535, United States
[4]Servite High School, CA, 92801-1952, United States
[5]The Affiliated High School of South China Normal University, Guangzhou, 510630, China

[6] 914997964@qq.com

Abstract. The causal inference is the process of determining the causality and correlation between different variables in a system. This paper mainly focuses on the introduction of theories of causal inference analysis and application of causal inference on student performance data. The adjustment formula, backdoor criterion and front-door criterion are three different methods to analyze different causal diagrams which can be used to quantify and analyze the causal effect of students' choices on performance and school support. Taking advantage of logistic regression model, counterfactual inference can utilize students' experience to predict the possibility of result under different choice conditions. The Natural Direct Effect in the path-specific effect is used to study the direct effect of sensitive attribute gender on the acquisition of school support, determining the fairness of decision-making of the school support.

Keywords: casual inference, casual diagram, counterfactual inference, path-specific effect, logistic regression.

1. Introduction

Students' performance in school is difficult to quantify and analyze, their choices and ideas in school curriculum activities are getting more and more attention, and these performances can often affect the education they receive.

This paper mainly focuses on the application of causal inference to student performance data. Then find out whether there is causality and correlation in students' different choices, and whether these behaviors can affect the school support. These causalities can be shown concisely in the form of causality diagram model. Taking advantage of the adjustment formula, front-door criterion and backdoor criterion [1], the most influential causality will be extracted, and the obtained results will give unneglectable information that is easier to detect and analyze than the data table. Then, counterfactual inference can utilize students' experience to predict the possibility of result under different choice conditions. Last but not least, the Natural Direct Effect [2] in the path-specific effect [3] implies that

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2/20220553

the particular causal path will become the focus of research, other unimportant causal paths are treated with baselines.

In practical application, the students' data are binarized and put into the logistic regression algorithm, which makes the students' gender, learning willingness and access to extra support clearly quantified and displayed.

2. The causal inference diagram models

2.1. The adjustment formula

Here is some of the notation that were being used. The uppercase letter like X represent a variable, and a lowercase letter like x represent a value of the variable.

Finding a relationship between two endogenous variable or descendent could create a misleading outcome since there is another variable that are influencing the two endogenous variables. When this happen, this is the time this is the time that you will need to use adjustment formula. The do operator "do(X=x)" or "do(x)" was a notation that will be used in adjustment formula. [1] Where we are forcing x to become a value that we want, so means there will be no variable or situation that X could be influence.

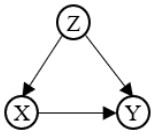

Figure 1. A casual diagram example

In Fig.1 the adjustment formula will compute the relationship between x and y when the value of z change. This is the adjustment formula:

$$\mathrm{P}(\mathrm{Z}=\mathrm{z}|\mathrm{do}(\mathrm{Y}=\mathrm{y}))=\sum_{z}\mathrm{P}(\mathrm{Z}=\mathrm{z}|\mathrm{Y}=\mathrm{y},\mathrm{a}=\mathrm{A})\mathrm{P}(\mathrm{a}=\mathrm{A}) \quad (1)$$

This formula (1) will find the probability of y when the value of x was given. The benefit of using this formula is since you are using the do operator, there is no variable that could influence x. So, the diagram after the do operator looks like this:

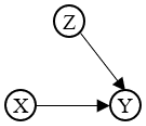

Figure 2. Another casual diagram example

Looking at Fig.2, you will know that both X and Z could affect Y, therefore we want to find the probability of Y when both X and Z are given in the adjustment formula. Then we are going times this with P(Z=z) which is the probability of Z is equal to z, since we want to know the ratio that z occupy in the data.

2.2. The backdoor criterion

$$P\big(Y=y\big|do(X=x)\big)=\sum_{z}(P(Y=y|X=x,Z=z)P(Z=z)) \quad (2)$$

The Backdoor Criterion builds off the Adjustment Formula, but solves the problem of having unobserved parents. When finding the effect of a variable that has no parents/unobserved parents to condition on and without intervening with the data, Backdoor Criterion is used to find an alternative variable to condition on. Backdoor Criterion is also an easy method to check if the causal effect is calculatable. Using Backdoor Criterion, a "backdoor path" can be found. If a variable is found along the backdoor path, then it can be considered an alternative to condition on. If multiple variables are on the path, then any one of them can be conditioned on. If the path includes a collider, then in addition to the

DOI: 10.54254/2755-2721/2/20220553

child of the collider, its parents must also be included. The descendants of those parents can also be included, and more than two variables can be conditioned on at once.

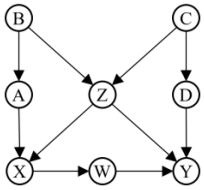

Figure 3. Casual diagram model

Z must be conditioned on along with its parents B, C or its descendants A, D. Therefore, to find the effect of X on Y, conditioning on Z,B; Z,C; Z,A; Z,D; Z,B,C,A,D; or any other combination is valid.

Every Backdoor Path must have a variable Z which has an edge pointing to X and is not a descendent of X, and Z blocks the path between other variables which connect to Y. The Backdoor Path in the graph below (fig. 4) assuming the objective is finding the causal effect of X on Y is X -> Z -> W -> Y. The reason is because X is a descendent of Z, and Z blocks the path between X and W, which is important because W also points to Y.

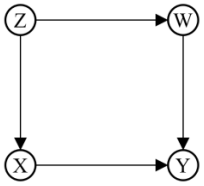

Figure 4. Casual diagram model

In figure 3, there are a total of four valid Backdoor Paths. The simplest would be X ->Z->Y. Another possible path is X->A->B->Z->C->D->Y. Because A,B,C,D is present on one of the paths, it is a valid variable to condition on, while Z is present in every path so it must be conditioned on.

2.3. The front-door criterion

As discussed, the Backdoor criterion provides a method for identifying sets of covariates when considering nonexperimental data. However, the Backdoor criterion is not a universal solution; something more is needed for the conditions that do not suffice the Backdoor criterion: the Front-door criterion.

To illustrate, consider an example of the relationship between smoking and lung cancer as Figure 5, attempting to determine the effect smoking has on lung cancer.

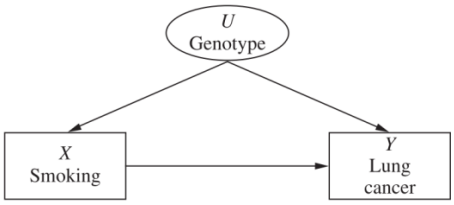

Figure 5. example

As shown in Figure 5, the variable that needs adjustment is X, with an unobserved variable U. However, U cannot be directly adjusted for it is unobserved and variable X cannot be adjusted either. In this case, the Front-door criterion is introduced to properly adjust X. To resolve this problem, a new variable Z is introduced as follows to block from X<-U->Y (see Figure 6).

DOI: 10.54254/2755-2721/2/20220553

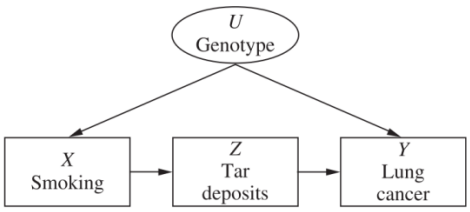

Figure 6. example

The formula functions (3) just as the Backdoor criterion.

$$P\left(y\middle|do(x)\right) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \quad (3)$$

2.4. The counterfactual inference

Counterfactual inference refers to taking the facts that have happened in the past as experience and inferring the different results after modifying the conditions.

$$Y_x(u) = Y_{M_x}(u) \quad (4)$$

The (u) in the (4) is called counterfactual. It is derived from Mx modified by equation M, that is, the right part of (4).

$$Y = aX + U \quad (5)$$

This equation (5) can be regarded as the experience of counterfactual. The coefficient a can be estimated by the overall model. The letter Y is the variable we are concerned about, which can be calculated from a single sample when inferring. During the modifying, replace X with the observed data x, and then calculate the unique U of the sample data. Finally the modified equation Mx can be obtained.

It should be noted that when the sample variable x we focus on is observed to be 2, if we want to know how the result will change if the variable x is 4. The X in M (which is (5)) is 2, the X in the modified equation Mx (which is the right part of (4)) is 4, and the modified result is the left part of (4), $Yx(u)$.

3. Experiments

Our dataset is based on the [4] in the internet. It has 33 attributes and 395 samples, it's all about student performance. In this paper we focus on some of these attributes, and established the causality and causal diagram model.

3.1. Application for the adjustment formula

A causal relationship we want to find in our data is the relationship between the want to take higher education and will you get extra educational support. Fig.7 is a causal diagram that represents the relationship where G represents gender, H represents higher education and S represents extra educational support.

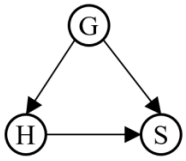

Figure 7. Casual diagram

To find out the relationship we need to apply the adjustment formula. (6) is the equation that we want to calculate.

$$P(S = 1|do(H = 1)) \quad (6)$$

DOI: 10.54254/2755-2721/2/20220553

Where S=1 represents the student getting extra educational support, and H=1 represents the student wanting to take a higher education. After calculating this question we could find out what is the probability that a student could get extra educational support if every student wants a higher education.

$$\sum_g P(S = 1|H = 1, G = g)P(G = g) \quad (7)$$

The equation (7) is the simplified version after applying the adjustment formula. Using frequency to plug in this equation we could get an actual value.

Table 1. Results of Data

	Wants higher education	*Don't want a higher education*
Male	14 out of 172 get extra educational support (8%)	1 out of 16 get extra educational support (6%)
Female	36 out of 204 get extra educational support (18%)	0 out of 4 get extra educational support (0%)
Total	50 out of 375 get extra educational support (13%)	1 out of 20 get extra educational support (5%)

Table 1 is an organized table using frequency of data. This table expresses data in many ways. There are three variables in the table, which are gender, wants of higher education, and did the student get extra educational support. The probability of this data could apply directly into the equation and find a result.

The results from Table 1 indicate that if you want to take a higher education you have about a 13% chance to get extra educational support. To compare with this result we also find out the probability that you will get extra educational support if students don't want a higher education.

$$\begin{aligned} & P(S = 1|do(H = 0) \\ &= \sum_g P(S = 1|H = 0, G = g)P(G = g) \\ &= 0.0295 \end{aligned} \quad (8)$$

The equation (8) is a similar process where H=0 represent the student don't want to take a higher education. This result means that if the student doesn't want a higher education it will have about 3% of chance to get a higher education. So, if the student wants to have a higher education the student will have a chance to get extra educational support as a result of this experiment.

3.2. Application for the backdoor criterion

For the sake of convenience, the variables will be assigned letters:

- School support = U,
- Higher education = H
- Address = A
- Paid for class = P

DOI: 10.54254/2755-2721/2/20220553

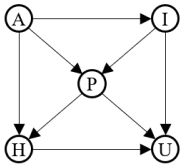

Figure 8. Casual diagram

To apply Backdoor Criterion, an example consisting of students in a math class will be used. The objective of the graph below is to find the effect pursuing higher education has on whether a student seeks educational support.

Other aspects of a student are also considered, those being the address which is either rural or urban, whether a student has internet, and if a student has paid for extra classes in math. The paid variable will have to be used along with either address or internet because paid is a collider, but for this experiment, higher->address->paid->school support will be used.

$$P\left(U\middle|do(H)\right) = \sum_{a,p} P(U|H,A,P)P(A,P)$$
$$= 0.1342$$

Students who want to pursue a higher education (H=1) got 0.1342.

$$P\left(U\middle|do(H)\right) = \sum_{a,p} P(U|H,A,P)P(A,P)$$
$$= 0.0299$$

Students who did not want to pursue a higher education (H=0) got 0.0294.

In conclusion, students who wanted a higher education would more likely get support. While this conclusion may be obvious, it shows that the Backdoor Criterion has in fact worked.

3.3. Application for the front-door criterion

Consider a relationship between the extra support from family and the grade of the student. The sets of variables are defined as follows: $s \in S$ as Boolean variable denoting whether the student's family provided extra support, $g \in G$ as a numeric variable range from 0 to 1 denoting the final grade of the student, with 0 indicates the original grade is in range [0,12] and 1 in range (12,20], and F as an unmeasured variable indicating the pressure to study from the family. The causal graph would be Figure 9.

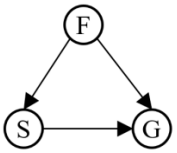

Figure 9. Casual diagram model

As one may notice, it is impossible to block all spurious paths from S to G: variable F is unobserved. A new measurement is thereby introduced, $p \in P$ as a Boolean variable denoting if the student has extra paid courses within the subjects (Math or Portuguese). The edited causal graph would be as below.

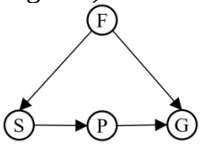

Figure 10. The edited casual diagram model

In Figure 10, the Front-door Criterion is satisfied by 1. P intercepts all directed path from S to G; 2. There is no unblocked path from S to P; 3. All backdoor paths from P to G are blocked by S. Also, it is

DOI: 10.54254/2755-2721/2/20220553

to be noted that P(X) is equal to P(X=x), as would $P(X_0)$ = $P(X=X_0)$ and X_0 is true. The formula functions as the following

$$P\left(g\middle|do(s)\right) = \sum_{p} P(p|s) \sum_{s_1} P(g|s_1, p)\, P(s_1)$$
$$= 0.3358212$$

From the experiment, it is concludable that the direct causal effect from S to G is very large. The proportion changes as much as 0.3358, proving that family support indeed has great influence on student's performance on grade.

The probability of having family support impact the grade of the student's grade is as high as 33.58%.

3.4. Application for counterfactual inference

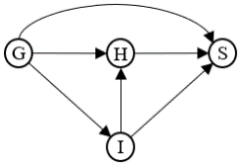

Figure 11. Casual diagram model

In Fig.11, the letter G is sex, which means the gender of student, the values of the attribute are F and M. The letter H is higher, which means whether the student wants to take higher education, the values of the attribute are yes and No. The letter I is internet, which means whether the student can access the internet at home, there are also Yes and No. The letter S is support, which means whether student can get extra educational support, there are also Yes and No.

In particular, when we need to apply the machine learning algorithm to this casual model, we will binarize F to 0, M to 1, Yes to 1, No to 0.

Because the data are binary, the logistic regression algorithm is used to predict the results. The causal diagram used in the experiment is shown in Fig. 11, and according to [1] there are three steps to compute the counterfactuals. The following four formulas show how each factor is estimated from the data, notation σ represents the calculation of SIGMOD function after logical regression.

$$G = U_g$$
$$I = \sigma_i + U_i$$
$$H = \sigma_h + U_h$$
$$S = \sigma_s + U_s$$

The formation of the formula is determined by the casual diagram Fig. 11. For example, the formula of I is determined by G - > I in Fig9, and the formula of H is determined by G - > H & I - > H, which shows how logical regression estimates the probability according to the casualty of the causal diagram.

Now the performance of a student A is measured as G = 0, I = 0.49, H = 0.68, S = 0.65. The first step of computing the counterfactual is to calculate $U_g = 0$, $U_i = 0.51$, $U_h = 0.32$, $U_s = 0.35$ according to the formula, which is the unique variable belonging to student A. The second step is to modify the formula, set the I intervention to 0, and observe the performance of student A if he doesn't access the Internet at home, as shown in Fig 12. Because I is intervened, the causal path G - > I is truncated.

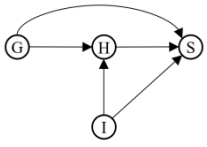

Figure 12. Modified casual diagram

DOI: 10.54254/2755-2721/2/20220553

In the third step, according to the measured unique variable U, the results are G = 0, I = 0, H = 0.93, S = 0.8. The comparison results show that counterfactual S is 0.15 higher than the practical data, which means that if student A doesn't access the Internet at home, he will be more likely to get the support of the school.

4. Path-specific effect

4.1. Introduction

We may be interested in the effect of an attribute along particular causal pathways, so path-specific effect (Pearl 2001) can be introduced. Path-specific Effect is defined [5] as given a causal path set π, the π-specific effect of the value change of X from x_0 to x_1 on Y=y through π is given by (9).

$$PE\pi(x_1, x_0) = P\left(y_{x_1|\pi, x_0|\pi'}\right) - P(y_{x_0}) \tag{8}$$

Where $P\left(y_{x_1|\pi, x_0|\pi'}\right)$ represents the distribution of Y where the intervention of $do(x_1)$ is transmitted through π while the intervention $do(x_0)$ is transmitted through other paths π' except of π, and$P(y_{x_0})$represents the distribution of Y where the intervention of $do(x_0)$ is transmitted along all causal paths from X to Y.

For the given causal graph in Fig.8, we may wish to find the direct effect of G(gender) on S(school supplement) only through the path G → Y which is called natural direct effect (NDE), but we are not interested in paths G → H → S, G → I → S, G → I → H → S. Along the path G → Y, we treat the sensitive attribute G as g, while along other pathways, variables behave as if the treatment variable G is set to g', i.e. H(g'), I(g'). Thus, the NDE of G to S in Fig.1 can be formulated as (10).

$$NDE = P\left(S\left(g, H(g'), I(g')\right)\right) - P(S(g')) \tag{9}$$

From [2], this NDE is equivalent to (11).

$$\sum_{H,I}(P[S|g, H, I] - P[S|g', H, I]) * P[H, I|g'] \tag{10}$$

4.2. Experiment

To find the NDE of being female on the result of getting school supplement, we can draw a frequency table from the data set and calculate the result:

$$NDE = \sum_{H,I}(P[S = 1|G = 0, H, I] - P[S = 1|G = 1, H, I) * P[H, I|G = 1]$$
$$= 0.079596$$

This means that the intervention on G from 1 to 0 which is from male to female can increase the probability to get school supplement by 7.96%.

We can also fine the NDE of being male on the result of getting school supplement by estimating by frequency:

$$NDE = \sum_{H,I}(P[S = 1|G = 1, H, I] - P[S = 1|G = 0, H, I) * P[H, I|G = 0]$$
$$= -0.093362$$

Which shows that the intervention on G from 0 to 1 (being a male) would decrease the probability of getting school supplement by 9.34%.

Compared with the expectation of getting school supplement which is 12.9%, the NDE of gender (7.96% and -9.34%) is significant to the result of S.

DOI: 10.54254/2755-2721/2/20220553

In conclusion, a female student is more likely to get school supplement than a male student, so the school supplement project is unfair with respect to gender.

5. Conclusion

In order to study what factors can most affect students' access to school support, this paper uses the adjustment formula, backdoor criterion, front-door criterion, counterfactual inference and path-specific effect in causal inference. Under the application of different methods, varied conclusions are drawn from different angles, which can help students understand and improve their performance in school. Taking this paper as an example, if there are any data models that are difficult to quantify and analyze in the future, we can try to comprehensively use causal inference and machine learning algorithms to get more specific results.

References

[1] Causal Inference in Statistics:A Primer; Judea Pearl, Madelyn Glymour, Nicholas P. Jewell.
[2] Fair Inference on Outcomes; Razieh Nabi, Ilya Shpitser; Computer Science Department; Johns Hopkins University {rnabiab1@, ilyas@cs}.jhu.edu
[3] Path-Specific Counterfactual Fairness; Silvia Chiappa; DeepMind London; csilvia@google.com.
[4] Fairness Datasets, Student Performance Dataset, 2021. [Online]. Available: https://ashryaagr.github.io/Fairness.jl/dev/datasets/#Student-Performance-Dataset
[5] Yongkai Wu; Lu Zhang; Xintao Wu; Hanghang Tong; PC-Fairness: A Unified Framework for Measuring Causality-based Fairness; Accepted as a poster to NeurIPS 2019; https://arxiv.org/abs/1910.12586