Artificial Intelligence for Economics

Prof. Adway Mitra

Artificial Intelligence

Indian Institute of Technology Kharagpur

Week – 04

Lecture - 17

Lecture 17 : Interventional Causality and Attribution

Hello everyone, welcome to this course on machine from artificial intelligence for economics. I am Adway Mitra, an assistant professor at Indian Institute of Technology Kharagpur and today we are starting our lecture 17 whose topic is Interventional Causality and Attribution. So, in the last lecture which is lecture 16 we introduced the concept of causality, we showed you how causality is different from correlation and we also discussed one simple notion of causality which is known as the Granger causality which is frequently used in economics. So, today also we will continue to discuss other aspects of causality. We will start with randomized control trials which is a very like a very useful and successful tool in the domain of economics. We will also deal with the concepts of structural causal model, shapely value based feature attribution and double machine learning.

So, in other words today we will see how some of the predictive machine learning models which we discussed earlier how they can be used in the context of causality. So, first of all what is interventional causality? So,like let us say a policy is proposed and we want to see the impacts of that policy whether they are desired or undesired. So, in the last lecture also we discussed this example that is let us say that there is a vaccine which claims to cure a disease. the disease and we want to understand whether that claim is correct or not.

That is if I administer that vaccine or that medicine to a person then can be necessarily say that they are like or can we say that they are with confidence that there is a significant chance of that patient recovering as opposed to if the medicine was not given. Or let us say does a school scholarship prevent dropouts. So, the like this is a policy measure. Let us say that some government or some agency is promising to give scholarships to school students so that they do not drop out. And now can we say that that will have its desired effect or not, will it actually reduce the number of dropouts or is that not the case.

So, like the naive approach to this problem is of course, to consider data in the form of intervention comma target and just test for causality using the Granger's approach. So, let us say that we have like a time series of this intervention where the intervention was applied on a person or on a in a region whatever the case might be and also the value of the target variable that is the like whether they how many people recovered from the disease or how many people dropped out from the school we have a time series of that also and the intervention also and then we try to find if there is a like the granger causal relation between these two time series So, as we had discussed earlier also Granger causality is has several drawbacks one of which is that it does not account for confounders. So, it might entirely happen that yes we do see that like they are like it seems that there is some sort of a Granger causal relation between these two variables. but we will never be sure that there is no other variable at play which is like we can say that let us say we can express like y as a linear function of x I mean the past values of x. but there may be some other variable z.

So, such that y is the that is y can be expressed as the linear function of the values of z and it may also be that x itself is also a linear function of z. So, when we are being able to express y as a linear function of x it may be entirely because of the presence of that variable z which is linearly related with both x and y. and so in that case that z is the confounding variable, but we are missing it entirely. Instead we are just satisfied to find that there exists a linear relation between y and x and we are just saying that one granger causes y. So, to get rid of these kinds of problems.

So, obviously, it is not enough to deal with to have just this kind of data. That is Granger causality in a sense ultimately is still that is we have earlier discussed the difference between correlation and causation. Now one idea which or one frequent criticism of Granger causality which we hear is that it is not really different from correlation. That is any machine learning algorithm is basically correlational, it is not causational. That is like when we are using some predictors or features to predict the value of the output variable y, we are actually looking for causal I mean we are actually looking for correlations only in the data, we are not looking for causal relationships.

So, the success of this Granger causality whether it is linear or it is neural Granger causality as we also discussed in the previous lecture is not a foolproof guarantee of a actual causal relations between them. So, to then what can be a full proof way of understanding finding causal relations. The best possible approach is to perform controlled experiments, where you apply the intervention while every keeping every other condition as unchanged. and then you observe the impact on the target variable. And like if you like that is let us say that you are you have a process going on which is very much controlled and you just pull a single lever in that process you without disturbing anything else that pulling a single lever is like making your interventions.

Now, it may happen and then you observe that what are the changes that happen in the system as a result of your pulling the lever. It may happen that you are pulling the lever triggers of chain of other responses which finally, impact the target variable which you are looking at either in a positive way or in a negative way. So, if that happens then we can say there is a causal relation between the two things. That is if I intervene that is in this case it is not a case of the experiment proceeding the process proceeding as it was earlier. It was proceeding in a certain way then you made a disruption you manually pulled a trigger to kind of shake up the system and as a result the system changed its state variable and the conditional variable and that change of state somehow impacted the target variable also. So, like if this happens then that is we can say the most convincing proof of the existent of a causal relation between the intervention variable and the outcome variable. However, in most cases this kind of a control experiment is not possible. Say like let us say in the case of like medicine itself or if we are trying to say that make a causal statement that say smoking causes cancer. Like how will you make sure that like it so it you may do an experiment where you are giving the medicine to some people and not giving the medicine to our other people, but then how does it how do you make sure that like the all no other condition is disrupted.

So, like if we could do such a controlled experiment then of course, it would be the great. So, that is known as what is known called as randomized control trials. So, it is randomized control trial is basically it is an approximation of the controlled of this kind of control experiments. So, the it is a principled way to pull out the causal effects of a treatment that can let us say that can take k possible values. So, then we divide the population homogeneously into k parts and then apply the different values of the treatment variable to each part.

So, let us say the treatment variable x it takes k values x 1 x 2 up to x k. So, then what happens is you have a population that population you divide into k groups the first group like for the first group you set x equal to x 1 for the second group you set x equal to x x 2 and so on and so forth. But the important and then you see the impacts of it that is like you see that is in each of the subpopulations into which you have done this experiment. You see you measure the value of y in all the subpopulations and you see whether you are seeing any change across the different subpopulations. Is it that in the first like the first group x where x equal to x 1 you see that y taking a particular value y 1.

In the last group which where x equal to x k we see are we entirely different value of y let us call that as y k. Are you seeing this kind of thing or are you seeing that the value of y remains more or same across all the groups. So, like this is the broad idea of this randomized control trials. However, the key is thing here is that the process of breaking up the population the initial population into k different groups this has to be done homogeneously. That is we have to say that each of these subpopulations these k subpopulations each of them are perfectly are perfect representations of the original population.

That is like let us say that the original population had let us just say 50 males and 50 females. In that case each of these subpopulations they should also have 50 males and 50 females. So, that we cannot say that the change in the value of y was primarily due to the change in the gender ratio not as a result of the this treatment variable x. And not only gender, but like various other attributes which may be relevant to the process with respect to age. Let us say that in the original sample say let us say 10 percent of the people were below the age of 10 in that case each of these subpopulations should also have the same property and so on.

In other words that is we have when we are like doing the intervention we have to make sure that the populationsare I mean no other thing changes in the population except for the value of x. So, like even so even if we are not actually carrying on the experiment in the in a controlled way we divide the population into homogeneous groups. So, that no like we can be we can expect that I mean the only noticeable difference between the between these different groups is the value of x there is the apart from this the populations otherwise seem to be to be I mean to be identical to each other in terms of all I mean statistical properties of all other variables. So, this is the broad idea of randomized control trials. So, this randomized control trial has been an extremely successful idea in economics.

In fact, the famous Indian origin economist Abhijit Banerjee and his wife a professor Esther Dufflow, they received the Nobel Prize in economics in 2019like they both of them have done extensive amount of work in this randomized control trials. So, some of the possible applications of randomized control trials in economics like in pretty much every branch of economics it can be useful let us say in case of developmental economics. So, then let us say the interventions can be something like a cash transfer scheme or the microfinance schemes or educational initiatives etcetera and we see there the impacts on let us say health care in low and middle income countries. So, in the like if we want to do this experiment we will divide this like we will identify some countries which are otherwise quite similar to each other in terms of most attributes such as demographics and things like that. And then one in one of those countries this scheme some let us say some cash transfer scheme is started and in the other country it is not started.

Then we see what is like like what some particular impact variable let us say these the general well being of the population let us say per capita GDP or things like that or like overall health of the people is it necessarily improving in the first country compared to the second country or do we not see any such significant improvement. So, the like that is one possible example. Then in case of labor economics we may be using RCTs to see the impacts of let us say some job training programs or minimum wage changes or employment subsidies etcetera. These are the possible values of the intervention variables and possible values of the target variable. like their employment status whether they are that is they can shift to whether first of all they can get some employment whether they can shift to better paying jobs and like other dynamics of the labour market.

In case of educational economics the in these impacts I mean the like the interventions can be something like vouchers for school students teacher training programs like adoption of different teaching technologies and so on. and the possible target variables to on which they they are these the these policies impacts may be observed are say things like their scores in some standardized tests or the rates at which they graduate and like in case of India. whether the school dropout rate decreases or not in case of health environmental economics financial economics also we can see like we can have various we may be interested in various kinds of interventions and like and see the possible impacts of those interventions on different outcome variables which are relevant to understanding the overall health of the system or overall environment of the country overall finance of the like different people in the system and so on and so forth . So, this is the like broad idea of causality. Now if you want to put this idea into a mathematical framework.

So, let us so how do you so let us say we carry out this kind of a randomized control trial. Now the question is how do we interpret the result or that is like I was saying that I change the value of x and see whether the value of y changes keeping everything else as unchanged. So, like even if y changes then then also I like how like how do I quantify the change. So, I may be interested in calculating this kind of a probability distribution of P of y given x. that is if x takes such and such values then what will be the possible values of y I want to build a probability distribution on it.

Now, this probability distribution this is an observational distribution that is you may have seen various example situations where the value of x is something and the value of y is something else. So, accordingly you have built a distribution. But that is like we can say that like that does not indicate the causal relation between x and y. The causal relation between x and y if it has to be quantified it requires a different kind of thing like it has to see that if we make we have to understand what will be the value of y if we make an intervention on x. So, what is so, in other words if I write this do of x which means that I make a disruption in the system and set the forcibly set the value of x to some other value.

So, that is very different from the original thing where x may be changing according to a natural process and y may also be changing according to a natural process both of which are related to the dynamics of the system which may involve various other variables we some of them we can be confounders also and based on that we calculate this distribution. Here however, we are not allowing the x to vary naturally, I am forcing x to a particular value irrespective of all other things in the system. So, like this model indicates the like in a nice way. That is let us say that like let us say that this is the intervention variable and this is the outcome variable. So, like in the normal situation there may be a confounder which influences both the x and the y.

Now, when you are making an intervention you are changing the I mean the value of x to some particular value of your choice irrespective of the confounder. So, you are basically breaking the relation between the confounder and the x so that this is what the system becomes now. So, like whenever you are making this kind of an intervention you are kind of actually you are changing the system. So,it is really this system which you should be using to calculate the or to understand the causal impact of y on x not this system. Now, like this example can also these set of graphics can also make it make the situation clearer.

So, let us say that there is a causal relation between x and y that is x is we can say is the cause and y is the effect and let us say that no confounder or anything like that is involved. So, in this case there is no difference between the these the observational conditional distribution P of y given x and the this interventional distribution P of y given do of x. That is even if you manually change the value of x to some other value the I mean the probability distribution or the conditional distribution of y will remain unchanged because it depends only on x and no other thing. So, whatever value of x you set to the value of y will also shift accordingly. Now, if it happens that let us say the causal relation was reverse that is y is the cause and x is the effect and in that case you like it will turn out that y given do of x is nothing, but P of y itself which means that like that is if you make an intervention on x then y is not going to change.

If you make the intervention on y then x would have changed because that is what the how the they are causally related, but x is your intervention variable. So, if you change x then y is not going to change. So, P of y given do x is nothing, but P of y itself. and this is the situation where the confounder is present.

So, there in the original. So, there is a z which influences both the intervention variable x and target variable y. Now, here what you are doing is you are forcing x to a particular value irrespective of z. So, you are breaking the relation between x and z the same as what happened in this case also. So, in this case y probability of y given to x is again it becomes like the probability distribution of y itself.

Now, y still continues to depend on z. So, the like when I say p of y like that in some sense it marginalizes over the possible values of z, but when I am talking when I am considering x that does not involve y anymore. So, the problem with this approach, this approach is like if we could calculate all these kinds of distributions these like y given do of x and things like that, then that would have conclusively solved the problem of causal analysis for me. The problem here is that many of these experiments are not actually feasible. Say for example, I want like let us say I want to understand whether smoking causes cancer or not. So, in that case like I may be tempted to do a randomized control trial where I like may force some that is I divide a population homogeneously into two groups.

I force the everyone in the first first group to smoke and I like forcibly prevent everyone from the in the second group from smoking and we see that if people in the first group are getting cancer and the people of the second group are not getting cancer. Like if we could do this kind of experiment then we could say whether there is a causal relation between smoking and cancer or not, but obviously this is an extremely unethical experiment to do that is we cannot like force some people to smoke and at the risk of them getting cancer. If this is an unethical experiment some other experiments might be simply impossible. Let us say I have a claim like if certain region of Pacific Ocean gets heated up that causes decrease of rainfall in India. Now, there is no possible way in which I can artificially heat up that region of Pacific Ocean and see whether there rainfall over India is increasing.

So, this is not unethical, but it is simply infeasible. So, not all I mean if we could do interventions then that would have been the best way of studying causality, but not all intervention based studies are possible. So, in that case what can we do the like the other approach is just whatever data is available you just make same or somehow utilize that existing data to like make the like to understand the claims of causality. Now, the problem with this is the presence of something known as or the requirement of something known as counterfactuals.

So, let us again consider a situation. So, let us we know that temperature is rising all over the world everyone knows that everyone agrees with that also there is no way to deny it. Now let us say that scientist is claiming that human induced climate change is the cause of rising temperature. That is because let us say humans are urbanizing they are like cutting trees, they are building cities, they are building cars and vehicles as a result of all that this temperature change is happening. Now, there is a let us say there is a climate change denier who is saying that no there is true that temperature is high, but that would have happened anyway it is not because of human influence after a few years again that temperature will come down also. Now the scientist will try to challenge the climate change denier by saying that if industrialization had not happened then we would not be seeing the high temperature.

But then the climate change denier will say that but industrialization has happened like it is easy for you to say this because there is no way of disproving it because it is something like a counterfactual. I mean industrialization has happened how can we say what would be the temperature if industrialization has not happened. So, this is an example of a what is known as a counterfactual. So, we like since we neither have any observations nor we can do any interventions to know about the counterfactual scenarios. So, what can we do? So, one possible way is so like the so what is basically what is the counterfactual our counterfactual is basically such an is a situation like or a particular data point which is not present in our data set.

So, let us say I in our data set has like I am considering the two variables x and y I have lots of observations of both x and y. But now the question which I am asking is if x took a different value which is not present in the data set then what what the value of y would be. So, this is the question which is being asked, but it turns out that I am I am not unable to answer this question because the I mean if I do not have the value of x in the data set then how will I answer like say what would be the value of y in that case. So, like it is something it is a we can say it is a limitation of the data which we have. So, how do we get rid of this limitation? The answer is if we could generate some like if we could simulate data then we could fill up the gaps that are present in our data set.

That is like if you could build something like a generative model which allows us to create fake data of x and y while making sure that it is consistent with the data which is already present then this problem might have been solved. Then I like any like for any value of x which you query I could have said that what could be possible value of y. I could even say that if industrialization had not happened then what could be possible values of the temperature. So, this requires the presence of this what is known as the structural causal model that is I would actually in that case write down or I would be able to express the relation between x and y as well as the different other variables which may be relevant to the systems in the form of equations like this. These equations they can be either deterministic or probabilistic that is either I can write variables like w, z, y etcetera as some deterministic function of the intervention variables x or maybe the some of the intermediate variables like w or I could express these as some kind of conditional probabilistic probability distributions.

If so like I we can call it as the structural causal model. So, we but the problem is that from where will we get the this kind of structural causal model it is not directly available I mean it is not available to us. So, one possible way in which we can try to construct these structural causal models is with the help of machine learning models. that is we can actually express y equal to f of x where f is any of the predictive model which we have seen earlier in during the hour when we are discussing with supervised learning such as linear regression or linear classification or decision tree or something like that. So, like we so, like we so, y is the of course, the target variable x is like a like I can consider x as only the treatment variable which we are interested in or we can also throw in all other possible influencing factors including possible confounders into the system andtry to express y as a like as a some function of that.

f for example, if f is a neural network any arbitrary function can also be expressed by it provided it can be answered fully ok. So,now this brings us to the concept of feature contributions in like the contributions of different features in case of a predictive model. So, let let us say that like I have a predictive model to which like the different features I mean like as I mentioned the different predictors are present in that in the are being provided as inputs to that model and I am trying to understand the value of y. So, now that predictive model in general it can be like anything it can be a neural network or it can be an arbitrarily complicated neural function. But later for the for the benefit of or for the ease of my understanding, let me express y as a linear function of all these possible predictors ok.

Even if the relation between these is I mean not a linear one, let me just write it in this way in which I will try to look at these coefficients and try to understand what roles they are playing. So, so like if I take the expectation of these. So, I let me just imagine these the target variable as well as all the different predictors as random variables. So, that I can take their expectations in general by expectation I simply may mean the mean values of all of these across the large data set and I take the differences. So, that I can see that Δ_y that is what is Δ_y? Δ_y is obtained by $y-E[y]$.

So, what does that mean? It means that a particular observation of y how by what quantity it is differing from the expected or the mean value of y. And we can write it like because of this relation we can write it as $a_1\Delta_{x_1}+a_2\Delta_{x_2}$ and so on. what is Δ_{x_1} is by how much this predictor x_1 has changed in this example compared to its usual value or by how much has x_2 changed in this example compared to its usual value and so on and so forth. So, like here I am basically like what I am trying to do is Δ_y is the change that has happened in the target variable and I am trying to attribute this change to changes in each of the predictors and I am trying to understand which like the. So, some of the predictors $x_1, x_2,$ etcetera some of them may have let us say that in this data point.

the value of y is higher than what it usually is and like let us like take one example. Let us say that the actual the prediction let we are considering the bicycle rentals in a particular day. So, let us say that on a particular day it is found that only these many bicycles should be rented out while on an average day some 4500 these many bicycles are rented out. So, clearly there is a huge difference. So, now what caused this difference which are the different factors are causing the difference.

Now, it might be that the we have all the possible predictors which are known to influence the number of bicycle rentals. So, which of those features took some unusual value on that particular day because of which the this the sales change. So, drastically I mean the why change so, drastically are all of these all of those predictors changing in one particular way or some of them changing in different ways also. So, like we so, we try to answer this question using a concept known as Shapley value. Now, this Shapley value is a concept which is borrowed from game theory it aims to quantify the contribution of each member in a team.

So, it is calculated in this way and like I am not going to the mathematics of these Shapley values. So, like, but basically what it tries to give us is like like if a particular member of the team was not present. then what would have been the result of the I mean overall performance of the team. So, or rather what to like what part of the teams overall performance can be attributed to that particular player. Now, when we are talking about this difference it could either be a positive difference or a negative difference.

It might be that if the if that player was not part of the team then the overall team would have done better which basically means that that that player underperformed. or it could be that or it could be that the person over performed and as a result of it the team performed better than usual. It may even happen that the person himself over the player himself performed better than what he usually does, but the team overall performance was still low because some of the other players underperformed. So, the roles the like so these like in this case also we use the concept of Shapley value. The Shapley values can be calculated by a certain formula certain I mean these formula are actually quite difficult to calculate numerically.

So, because of which there are certain approximation algorithms like this where I will not go into the details of this algorithm, but it allows us to calculate the Shapley value. So, corresponding to so if we have observations of both like all x and y, then we can apply these algorithms called SHAP to file like make to like to assign the shapely values to each of these features or each of these predictors in every for every single data point. So, that we can just examine every data point see whether the y was better than expected or worse than expected and which of and in any of the cases we can even try to attribute that positive or negative change of y compared to its mean value to the individual features. Like if we can say that like in this case we can like this example of bicycle rentals we can say that they are like there were certain factors which were actually positive which were favorable on that day like this first factor that is the temperature was quite good. So, it was an ideal day for cycling the month was also like a month in which many people do like to cycle that is October when it is neither very hot or neither very nor very cold.

But there were some other factors like thunderstorms which were unfavorable that is like people do not like to rent bicycles on a stormy day. So, that is some factor which worked against us. There is also one concept known as the like the double lasso or the double machine learning which becomes important for us in this case. So, like as I said earlier like we are trying to express the outcome variable y in terms the predictor I mean the treatment variable x and we also have the possible other variables d let us say the which includes the these confounders and other things as well as some random factors. So, I just write down y as a linear function of the all the variables as we are doing earlier also.

So, but we in this case we make a separation between the treatment variable x and the other variables d which can include some confounders also. Now, because they are confounders I write this an additional equation which means that x itself has a relation with the like these other variables d. So, remember that d is something which impacts both x and y. So, that is why we are writing it like this.

Now, so the both we have two regression equations here. So, we can try to solve some linear regression to find out like so that will help us to find to solve the value of beta 1 and beta 2. So, if I could especially I am interested in this beta 2. So, beta 2 is like basically telling us the change or that is how much of the change of can be brought about by x that is if x changes by amount delta x then the change of y will be beta 2 times delta x right. So, that like so in a sense beta 2 is the strength of the causal relation between the intervention variable x and the target variable y. Now, as it turns out that this regression problem is not easy to do or I mean we can solve it, but the result we will get will not be a consistent result because it like there is a something known as an identifiability problem.

Now, because both of these are like that is like you because of the presence of this d you will in both of these equations you will never be able to correctly or accurately estimate β_2 it I mean the that is the β_1 and β_2 in a sense will get fused with each other because of the presence of this relation. So, now, there is one approach which is known as the double machine learning. So, there is like based on that there is this theorem known as the Frisch-Wohr-Lowell theorem which actually allows us to break down the above problem into 2 different into 3 different regression problems and the solution to which will give us the β_2. So, the first problem is you first solve the say the second problem that is you regress d on x and using the usual regression linear regression and you get the necessary values.

Now you calculate the residual that is the d hat is what you could is the part of d which you were able to predict using x. And then you also regress y on x I am sorry in this there is a slight change in notation he in this case d is the intervention variable and x is the confounder variable that is I mean a notational like abuse of notation in this case. So, you first regress the intervention variable on x on the confounder variable and you calculate the residual that is what part of the intervention can be predicted by the confounders. then you also regress y on x that is you and then you get the residual w hat which is y minus y hat means what part of this thing of the outcome can be predicted from the like from the confounders.

And finally, you are left withw hat and v hat. So, v hat is that part of the which is not dependent on the counter on the confounders and w hat that is the part of the outcome which is not also dependent on the confounders. So, now you regress w on w hat on v hat and the result which you get is the beta 1 which you are looking for. So, like so that is the process of double machine learning. So, if you are if you do this approach then you will be able to actually calculate how much this y the target variable how much of it actually depends on the intervention variable instead of the confounder variables. So, to conclude so, we discussed about interventional causality where the we are trying to estimate the impact of a treatment on a target variable through controlled experiments.

Now, randomized control trials they are an important part of economics. structural causal models they help us to generate counterfactual data, but and these can be approximated by predictive models. Now, Shapley values these are concept borrowed from game theory which we will discuss in the coming lectures of this course. It is something it can be used to attribute the outcome to the different factors of a predictive model that is whenever the outcome variable it behaves differently from its mean value we can attribute that its deviation to different parts of the to different influencing variables. And finally, double ML we can we can use to separate the impacts of confounders while mentioning the by measuring the impact of the interventions on the outcome. So, with this we come to the end of this sub topic of causality in the coming lectures we will deal with some other topics. So, see you then till then all of you please stay well and take care see you again bye.