Artificial Intelligence for Economics

Prof. Adway Mitra

Artificial Intelligence

Indian Institute of Technology Kharagpur

Week – 08

Lecture - 40

Lecture 40 : Bias, Fairness, Ethics and Interpretability in AI

Hello everyone, welcome to this course on artificial intelligence on economics. I am Adway Mitra, an assistant professor at Indian Institute of Technology, Kharagpur. Today is the last lecture of this course and today we are going to talk about bias, fairness, ethics and interpretability in AI especially in the context of economics. So, today we will start with some ethical issues of AI, we will discuss the issues of bias and fairness as well as local and global explainability of AI models and we will also briefly discussed about the environmental impact of AI ML and the necessity of what is known as green ai so first of all what is ethics in machine learningso like the the question of ethics may arise in the practice when we are buildingoperational systems that involve ai and ml the question of ethics might appear at at many different levels in many different formats so first of all ethics in data collection When we are because like we know that machine learning models are generally data hungry that is the more we feed data into it the better it performs. But in our desperation to like feed more data we may end up like choosing or extracting data from various people or various agencies through unethical ways and also distributing them for like for purposes which may turn out to be like which need not be beneficial for them. So, basically these results in privacy concerns especially when we are talking about acquisition sensitive personal data.

So, it is possible that like if some personal data can be extracted from me, then I like it may help me in the sense that the various AI services which I am availing let us say recommendation systems and so on. Like they may perform better, but I myself may not be very comfortable in disclosing those information like to the world. So, like this is a question of ethics. Next, machine learning can be applied in areas of potential societal impacts like in which case it is possible to have ethical dilemmas also especially where critical decisions are involved like because we know that machine learning based systems are not necessarily perfect.

they like even the best machine learning system is not going to have 100 percent accuracy. It will have some some it there will always be some probability of error. Now, in many cases these kinds of errors can have very significant impact on the general well being of the society. So, when we are deploying such AI-ML systems for like that is like in these kinds of context, then the question of ethics becomes very important. For example, when we are talking about an autonomous vehicle, so the autonomous vehicle like if the sensors associated with it make some wrong decision due to some failure of some machine learning model at some level, then it can cause a collision.

Or when we are considering the policing or judicial system there are wrong decision may result in an innocent person being framed and so on and so forth. So, like these are like these are kinds of applications I mean very critical applications where the use of AI-ML always has to be monitored because there is always the various ethical questions involved in it. That is whatever decision the AI-ML system is taking whether like if it is incorrect then of course, there like the it can have grave consequences. Even if it is correct from one point of view there are such in many sensitive situations it may not necessarily be the ethical decision in all cases. So, like a so, like these are matters which we need to keep like keep in mind when we are deploying AI-ML systems for these kinds of applications.

So, like basically the guidelines will be avoid using sensitive attributes to make predictions and make probabilistic predictions with uncertainty estimates to flag uncertain cases. So, that like if you if the system is not sure it should always be able to convey that uncertainty and also there should always be robust validation of these decisions with a human in the loop that. So, that is whenever it is the AI system is giving a some kind of a prediction in a very critical application like the ones we discussed there should always be a human in the loop to cross check that and give feedback to the system. So, that the system may learn to take better decisions in the future and also we it is necessary to have regular performance testing and model updation as and when necessary. Now, what are the different kinds of issues of bias or fairness that can arise in machine learning as we already mentioned machine learning can take some wrong decisions.

Now, when we say wrong like one one kind of wrong might be like simply classify that is confusing one class with another and so on and so forth, but it might happen that there may be certain underlying biases in which like the like one kind of mistakes is more likely than another kind of mistakes and so on. Now that is like it is entirely possible that like there are multiple in a classification problem there are multiple classes. Some of the classes are often predicted by the model, but the other classes are rarely predicted. Now, so like we can say that here a bit of a fairness issue might be involved that means that if there are now the reason why this kind of bias is happening in the AI system may lie in the kind of data with which the model was trained. So, it is possible that there are like the data was itself bias that is there are very few training examples from some classes.

So, the model has not learnt to recognize examples from those classes. So, hence when the model is deployed it may not be able to recognize those classes. So, that can be is of course, solved by approach in recent approaches in machine learning such as one short learning, zero short learning and so on and so forth. but it can be but the nature of the bias can be slightly different also for example the bias can arise from the bias training data or from flawed assumptions or the design choices made during the algorithm development and unfairness can disproportionately affect certain groups of people and it can reinforce social inequalities and create disparate impacts in areas of hiring lending criminal justice and so on And this is this kind of fairness issues can emerge especially if the certain variables are correlated with the protected attributes like say race or gender. Let us consider a few examples of this kind of bias.

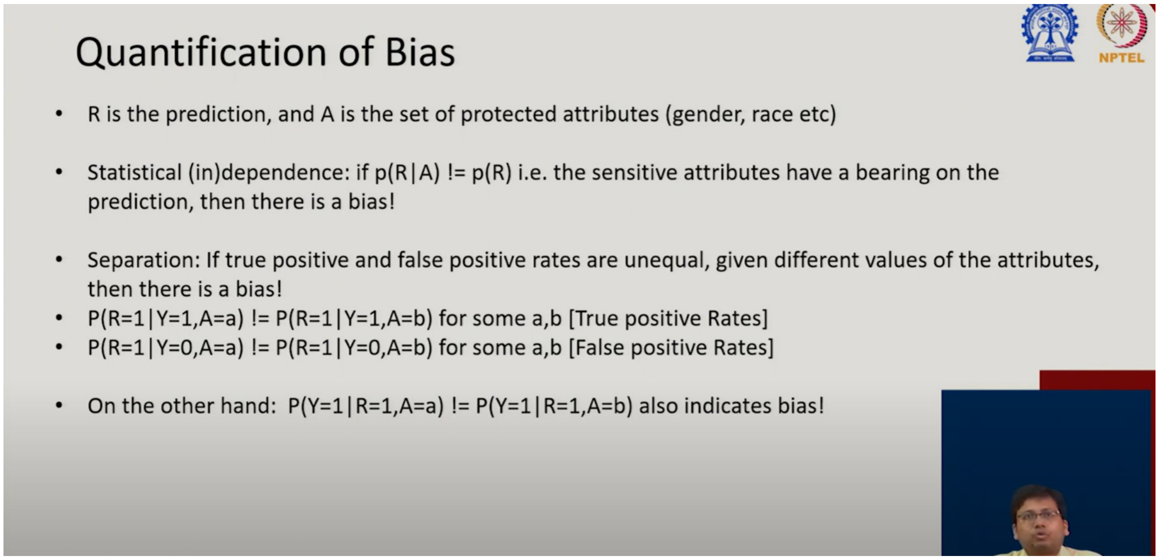
So, let us say we have trained a decision tree to automatically shortlist applicants for a post that is the there is a post there are some requirements and then lots of applicants have filed for as applications for that post. So, the decision tree will according to some hierarchy of features it will scan each of the applicants and decide whether a that person should be shortlisted or not. Now it might be that in like when we are training the decision tree like the decision tree has only seen data from the past and so based like based on that the decision tree has learnt which are the attributes based on which it may I mean I mean people may have been recruited in the past and accordingly it will recruit in the future also. So now it might happen that in the past only men have been selected for that particular post. Now this may have happened due to like mindset of the people earlier, but we do not want that to be perpetuated.

Now, the decision tree by itself is of course, not like is not going to do gender discrimination, but before because it has been trained on data which is basically biased that is like in the past training data it might be that. 80 or like maybe it more than 90 percent of the women candidates have been rejected while maybe only 50 percent of the men candidates had been rejected. So, based on that the decision tree inadvertently ends up rejecting the or tending to reject female applicants. So, like there was already a bias in the data, but that is the machine learning model has learnt to perpetuate that data even I mean that bias even though it was not explicitly trained to do so. Or consider one more recent example.

So let us consider a vaccine distribution system. Now it is the system predicts who needs the vaccine most urgently. Now it might have been that in the past the vaccine had been taken up mostly by those people who could afford it. So, when we are training this model, we are training it by who purchased the vaccine and who did not purchase it. Now, if it happens that mostly rich people had purchased the vaccine so far, so then it will inadvertently end up learning economic status of a person as a determining factor whether they should be given the vaccine or not.

So, and as a result of that it is the when the model like decides whom it will give the vaccine next, it may be giving it only to rich people because that is what it has seen happening in the past. So, these are examples where the I mean the algorithm not only preserves, but it may even perpetuate the biases that were already present I mean the unfortunate unintended biases which were present in the data. So, from the model designers perspective everything is fine because they do have not designed the model so as to make it biased towards any group. So, they may say that like it is the models are neutral, but in reality it is not because data on which the model is being trained that data cannot be called as neutral. So, now how do we understand whether a particular model is biased or not or whether they it has some fairness issues or not.

So, like the let us consider r as the prediction by the model y as the as the desired value of I mean the desired value of whatever is being predicted and a let us consider as a set of protected attributes such as gender and race. What do we mean by protected attributes? I mean these are the attributes on which I do not want the decision to be taken. I do not want like these attributes to be taken into account while making the predictions. Now, when we consider statistical dependence or



Quantification of Bias

- R is the prediction, and A is the set of protected attributes (gender, race etc)
- Statistical (in)dependence: if p(R|A) != p(R) i.e. the sensitive attributes have a bearing on the prediction, then there is a bias!
- Separation: If true positive and false positive rates are unequal, given different values of the attributes, then there is a bias!
- P(R=1|Y=1,A=a) != P(R=1|Y=1,A=b) for some a,b [True positive Rates]
- P(R=1|Y=0,A=a) != P(R=1|Y=0,A=b) for some a,b [False positive Rates]
- On the other hand: P(Y=1|R=1,A=a) != P(Y=1|R=1,A=b) also indicates bias!

independence. So, we can consider $p(R|A)$ and $p(R)$ the probability distribution of a particular predicted value.

given the its sensitive attributes and we compare it to the case where the sensitive attributes are not given. Now if we see that these two distributions are different from each other that basically means that the model models prediction is somehow dependent on these sensitive attributes which is not what we want. So there is a bias. So like so one way to test for bias is to check whether these two variables R and A whether they are independent or not. another thing is another idea is that of separation now let us consider along with r and a let us consider y which is the desired outcome that is like whether a particular candidate should be hired or not and r is like the machines prediction whether that candidate it recommends that candidate to be hired or not and a as I said at the of sensitive attributes like gender or race.

So, in this case we can consider what is a true positive rate. So, let us say that y equal to 1 that is I want this person to be hired and its attributes are a. Now, what is the probability that the algorithm will also recommend that person to be hired. So, that is one thing which we are calculating and we are comparing it with the same quantity that is the machines prediction provided these sensitive attributes are changed. Now if these two are different that means that so like this is like this is basically known as the true positive rates true it is called true positive because like I want it to this y equal to y to be 1 and r is also 1 that is I want the prediction to be 1 and the model is also predicting 1.

So it is a case of a true positive. but the probability of the true positive may end up depending on the value of the sensitive attributes. Similarly, in case of these false positive also let us say y equal to 0 that is I want the model to predict 0 ideally and the model is predicting 1 by mistake. but the probability of this false prediction it seems to be dependent on the attributes. So, this is like if these have these two are in both of these are indicative of the bias that is to say the model is more likely to make a mistake when the attributes take certain values.

So, again coming back to the gender discrimination case. So, it is possible that that for male also there is some probability that the model can make a mistake that is I want the person to be hired, but the model says that he should not be hired. So, there is a probability that this kind of mistake can happen, but maybe in case of women the probability of this mistake is much higher that is even if I want this particular lady to be hired it is possible that the machine recommends that they should not be hired. because of the bias in the data on which it has been trained.

So, this is like in the. So, these are some of the measures by which we can quantitative measures by which we can understand whether a particular model is biased or not. Now, suppose it is biased then how do we mitigate the bias. So, first of all we have to detect the bias that we already discussed. Now, mitigation of bias largely depends on data on the quality of the data. That is we have to somehow have to ensure equitable and unbiased decision making by providing high quality diverse and representative training data.

Now in some cases that may not be possible like as I said like it might be that the past data is biased anyway because we have done things in the past which we now do not approve of. So, whatever is the past data that now somehow have to be changed or somehow have to be overridden. Now, how do we do that? So, like it has to like we have to use some kind of pre-processing techniques on the data or maybe we have to generate some synthetic data to counterbalance the biased data which is present which is actually present. and so on and so forth. And another possibility is to change the loss function of the models themselves.

We like instead of like just maximizing the true positive rate which most models try to do anyway, we can also try to minimize the disparity between the true positive rates for different values of the protected attributes. Like in this case we see that bias is happening when the true positive rate or the false positive rates are changing for different Now, here my aim will be that to minimize this disparity regardless of what this like this attributes are if the model like is going to be right let its probability of giving correct answer be independent of the attributes. So, like we can enforce these kinds of constraints into the through using the loss function, so that the model learns not only to maximize the true positive rate, but also minimize the difference between the true positive rates for the different protected attributes. Now, the question comes of like specifically in the domain of economics where how does this fair and ethical AI come into picture. So, first fair AI systems can promote market efficiency by ensuring that opportunities and resources are distributed equitably instead of like perpetuating the existing biases which we like just an example of what we already discussed.

Another thing is like when we are considering these recommender systems, so will a consumer trust in the recommender system may depend on the perception of fairness based on the kinds of products being recommended. Like if I feel that is the kind that is I am like being always being recommended a kind of products which are of only one type rather than and which does not suit me, then I may feel that it is unfair to me in some sense. That is it is suggesting to me things which are not necessarily I mean especially if I am like I belong to a minority group of customers whose requirements are different from most of the others in that. But I am always suggested products which are according to the needs of majority of the people instead of like people like me who are in the minority then I may feel that the I mean recommender system is in some sense it is discriminating against me it is not hearing my voice or it is not taking me into consideration so I will consider it as a issue of fairness so I may reject that recommendation engine another issue is that of social policy making so like once again when there are this I mean when there are resources to be allocated to different parts of the society for different types of social welfare based activities again it is important to use a fair ai systems so that the distribution is always fair now fair does not always mean equal distribution it means distribution according to the needs of the different people that is those who need more should get more those who need less should get less and so on and so forth now like that is when we are calculating who needs more who needs less like this may be different from who has more and who has less and however the data might be biased because typically those who it might be that the in the past those who had more had received more and those who had less had received less even though that is exactly opposite of what is intended so the data will be biased so we need this the concepts of fairness and so on to mitigate these kinds of biases and furthermore like in case of a market where there are many competitors Now, use of fair AI systems can help to create a level playing field. Now, let us talk about one more aspect that of interpretability of AI models.

So, like we have seen earlier that many AI and ML models they have very high accuracy in making predictions. They can sometimes match or maybe in some task even better the human performances. But the problem remains that many AI and ML models are essentially black boxes. Like for example, consider a huge neural network having large number of hidden layers and so on. Let us consider the example of chat GPT.

I mean chat GPT we have all seen that it does a great work in predicting or it is in answering all my questions. But I may be interested to ask why it is giving this particular answer or why is it so good if that is the case if that is the question then it is very difficult to answer it because it is a very complex system. It like it may involve not millions, but billions maybe even trillions of parameters. Now, if those parameters have somehow been their values had been somewhat different, then the question is would it still have done better done as well or would its performance have been improved or would it be worse. So, with the these questions is not easy to answer because the model is not necessarily very interpretable.

So, now, when it is now we cannot ask now there are certain kinds of ml or ai models which are like more interpretable but their performance is not good and then the and the reverse is also true so this if you look at this figure it like you can see that So, models like linear models or rule based models etcetera like they have they are they are very much explainable that is we by looking at the model we can understand what the model is doing, but their performance is not necessarily the best on many like data sets which are used in the real world their performance is not necessarily good even though they may do well on only on synthetic data. But if we consider models such as deep learning or various ensemble based models their performance is found to be very high empirically, but their explainability is quite less that is we cannot say we cannot really explain why they are doing as well. So, when like this question of this explainability or interpretability this becomes particularly important in various decision support systems like when some critical decisions are taken based on the AI predictions or AI-ML model predictions that is the so like this in the case of any safety critical decision like we may not be satisfied only with the decision I may want to understand why the decision was taken in the first place like if you look at this cartoon. So, like here there is an ML model which is used for some smart grid applications basically to indicate like to predict the power allocations to different users based on real times. So, so like the model has been trained on past data and based on it based on that data it is making some predictions.

But now the question is like the that is the system which runs on this smart grid applications. It may be subject to various kinds of checks and balances, there may be an auditor who wants to check whether the ML model is it is fair and legal that is like we have already discussed the notion of fairness. Now when we have a very complicated model it is not easy to understand whether let us say that it is found to be unfair that is or it has some kind of bias the kind of biases which we discussed earlier. the false positive rates and so on.

Let us say that it it is biased. Now the question will arise who is responsible for that bias is it the that the model is fine, but the training data is faulty or is it that the model itself has some bias. Now if the model itself has bias it is very difficult to find that because the because the model is very complex it is like a black box with millions of parameters which no one can understand. So, like for this kind of situation we need an interpretable models. So, so like so that that is why we need some kind of explanations which are generated by the interpretable ML. I mean where we are not only making a predictions, but we are also somehow giving an explanation of why we made that kind that particular prediction.

So, now the question is what kinds of explanations are these going to be. There are now so that brings us to the vast literature of interpretability methods. So they are like interpretability methods are broadly of two types local or global. So a local model it tries to explain every single prediction as a as a single case. So maybe it will somehow tell us like which of the attributes where given most importance while making the prediction and so on.

While a global interpretable method it will explain the overall model that is the model by itself may be a very complicated neural network with millions of parameters, but somehow it will give me some kind of abstraction which will like help me understand it. As I mentioned there are certain models which are like more interpretable in nature. So, if I have a model which is less interpretable in nature like a deep neural network. may be a global interpretability methods may try to approximate that model for me by building its building a some sort of a decision tree or something. So, that by looking at it I get a feel of what are the kind of thing like what are the at least what are the factors that this neural network is looking at.

There are other dichotomies when it comes to these interpretability methods like say purposes of interpretability. So like some purposes it might be to create a wide box models that is we want to that is we have a complex model but we want to approximate it by a like a simple and hence interpretable model. Or it might be we want post hoc explanations that is it makes a prediction and then it gives me like some explanation about why it that made that particular explanation and so on and so forth. Like so that is typically this is done like in there are various well known ways of doing it. So, one is rule based methods, rule based models that is like let us say we have a like a model which is accurate, but not very interpretable.

Now, the question is can the models outcomes be explained by simple rules by decision trees. Now, these rules will of course, be different from different parts of the feature space. Otherwise like if the if it is the same I mean same decision tree worked for the entire feature space then of course, I would not have built the complex model in the first place. I would have just gone with the simple interpretable decision tree, but probably the feature space is far too complex to be represented by a decision tree. So, we that is why we chose the neural network in the first place, but now what we are saying is can we divide the feature space such that in different parts of the feature space maybe a decision tree is going to be enough.

So, that is like like that that is the approach of rule based models. Another is post hoc explanations that is can like for every predictions can be visualize which parts of the inputs were important in driving the final predictions that is like the so if you remember the linear regression or the Shapley value based approach in that case also like it was a post hoc expansion based methods that is whenever it is we have a in case of the Shapley value methods we provide some Shapley we associate a Shapley value with every single feature for every data point the like the Shapley values of the features they indicate which towards which direction that feature is driving the prediction is it driving towards the correct value or is it driving towards the wrong value or is it induce is one value of feature inducing a prediction. positive anomaly or is it inducing a negative anomaly and so on. So, like there is a significant amount of work in this global and local interpretations will not go into the details of this. And one last topic which I want to mention before I wrap up this lecture and this course is AI and climate change.

So, as we already discussed much of many of the task related to ai and ml in these days especially those related to computer vision natural language processing etcetera we discussed several of such approaches in the last two lectures and as I said many of these approaches are now done using neural networks which are typically like have a large number of hidden layers with lots of parameters So, estimating all those parameters requires a lot of computations involving GPUs and like maybe more advanced sometimes super computers also and so on. Now, if we have so much computing then that obviously leads to huge energy requirements and as a consequence also huge carbon emissions. So, some people actually did a study of how much carbon emissions are carried out by these a state of the art AI systems and the results were quite mind boggling. And it was found that like if we consider the amount of carbon emitted in training a transformer with about 213 million parameters like we discussed transformers in the last lecture. So, that is a state of the art model for natural many natural language processing part.

So, now if you consider the emissions by an average human being in one year as you can see it is like just a tiny fraction of the carbon emitted by the by such a transformer model and this is just for training the model I mean for searching the right architecture and the right value of the parameters. not for not necessarily for deploying the model which also will require some task. So, we will require some amount of computations. So, you can understand like how much carbon inefficient these models are and so and we can also understand that if as we go for deploying these models more and more they can then these models can pose as very serious risk to the general climate at a time when the earth is facing the threat of climate change and that is mostly due to anthropogenic emissions that is emissions by human activities now we are already trying to cut down on various activities or switch to more energy efficient methods now when we when it comes to applying these kinds of methods they are they are already very much we can say that they are quite like polluting methods and they themselves consume a very huge amount of energy so the like it is very important to embrace what is known as green ai which where the task is not only to develop energy efficient hardware which is the task of like engineer like electrical and electronics engineers but also for computer and machine learning scientists also it is important to come up with ideas how pre trained models can be used and Like for every single task we should not have to train our own models, we should be able to like borrow models trained by others and somehow utilize them for our own tasks and this requires us to make important research contributions in domain adaptation and so on and so forth. So, with that we come to the conclusion of this lecture just to wrap up.

So, wrong predictions by AI and ML models can have grave social and individual consequences existing bias in data can be learned by models to amplify these prejudices bias can be quantified and removed through suitable loss functions. best performing models often lack interpretability of results and that is why we need global and local explanations to understand the working of these kinds of black box models. And finally, reusing pre-trained AI models is important to curb the AI model emissions and that is why we like it is not enough to have AI, but we need green AI. So, these are the main points of today's lecture and with that we wind up not only today's lecture, but also this course. So, I hope the like in this throughout this course from the lectures of myself and the other professors Driptho Bakshi and Palash Dey hopefully you have got a wide range of ideas about how various AI various aspects of AI can be used in various domains of economics and I hope in in the coming days you will work on this topic and you will contribute more to this so it was a pleasure having having all of you with us hope you enjoyed the experience as well so all of you stay well and take care and maybe someday we will meet bye everyone good luck