Counterfactuals Basics

Badrinath Singhal

August 2021

In today's world a significant amount of decisions are been done by AI all over the world and with time the amount of decisions will only go up. AI are making many redundant decisions as well as few important decisions, for example some games are extensively using AI to improve experience, personal assistant are using AI to improve themselves so they can provide better services, people are trying to reduce workload of medical professionals by involving AI in diagnosis and treatment process and not to mention use of AI in drug discovery. It is evident that importance of AI will only go up with time and it will play a significant part in our lives in coming decades.

With increase in the use of AI, scrutiny of AI models are also on rise. If AI are making important decisions in people's lives, we want to know whether AI are unbiased, fair and understandable. Many recent research have found out some of the models that are widely publicized and are deeply involved in people's life are unbiased in some respect [1]. As the stakes get higher, AI needs to be more fair, unbiased and understandable and because of this research in the same are growing in recent years.

Most of the AI models are not interpretable i.e. not understandable to humans because of the nature of complexity. For example, a simple NN with 10 input layer, 10 hidden layer and 10 output layer contains 200 parameters. The values of these parameters are interdependent and make a decision, it's hard to understand how these 200 parameters work together to make a decision. As we move on to more complex problems the number of parameters only increases making it much harder to understand. Thus we need a better and a different way to understand how AI systems are making decisions for us.

This research is getting growing amount of attention in recent years, there have been many development which have been quite successful in explaining the AI systems to a certain extent. One such way is through counterfactual, Counterfactuals to a query is defined as such that it have a different result than the query with minimal understandable change.

For example, suppose there is a AI model which process the bank loan application and gives output as 'loan approved' or 'loan rejected'. If an applicant submits his/her application and the AI systems rejects the loan request, what minimum change the applicant have to do to his application so that the loan would be approved. It can be "If applicant have income of \$10000 more his/her loan would be approved" or "If applicant had applied for \$2000 less amount then the loan would be approved" or any other explanation. Although counterfactual should be such that it is practical for applicant to make the change, i.e. it cannot be "If applicant was 3 years younger his/her loan would be approved". Counterfactual would be minimum change for original query which results in a different decision than that of original query.

Counterfactuals have been widely researched for past few years, they can be used to investigate if the model have bias introduced or have helped user get the desired outcome from the system/models by making minimum achievable change to their query. Counterfactuals are a one of the great ways to understand the decision making of AI systems where they are making important decisions like bank loan approval.

There are few properties that counterfactuals should have so it can be regarded as good counterfactuals [2].

- Counterfactuals should be understandable
- Counterfactuals should be sparse, i.e. no. of changes should be as less and as close to original query as possible.
- Counterfactuals should be practical i.e. it shouldn't be unrealistic
- Counterfactuals decision should be different from that of it's original query.

There have been good amount of research which investigates counterfactuals deeply in many perspectives like how to generate good, sparse and diverse counterfactuals. Counterfactual generation have been tackled in different ways like optimization methods [3], generative methods [4][5], case based reasoning method [2] etc. Plenty of metrics have also been developed to evaluate the quality of counterfactuals [6].

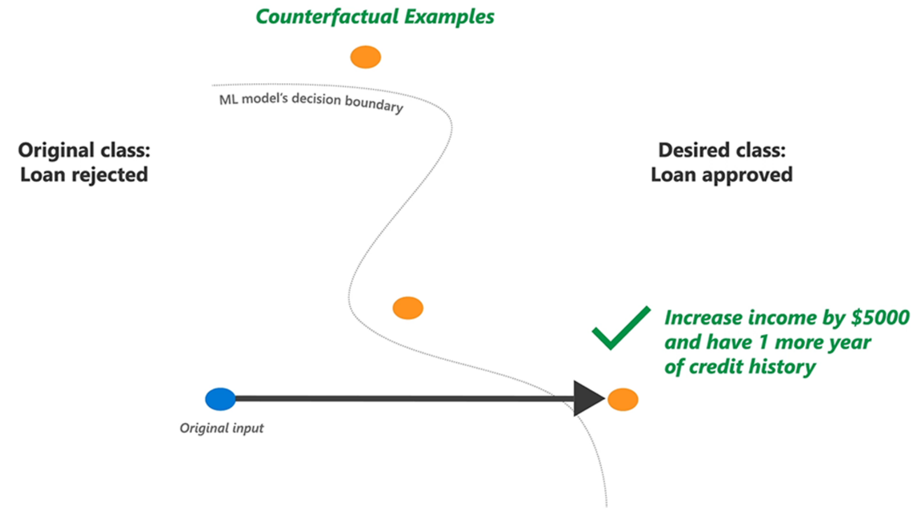

Figure 1: Example of counterfactuals (taken from [6]
)

Counterfactuals are good way to understand AI systems for people who are not involved in making those systems. Their use makes it an important are of research more than some mathematical methods to evaluate the counterfactual, user studies will provide much concrete information regarding the same.

END

References

[1] Racial Discrimination in Face Recognition Technology

[2] Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)

[3] COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

[4] Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems

[5] Interpretable Counterfactual Explanations Guided by Prototypes

[6] Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations