CENTRE FOR DEVELOPMENT
OF
APPLIED AI LAB.
(CDAAIL)

CounterFactual Models

Prof. SK Udgata and team.

CENTRE FOR DEVELOPMENT OF APPLIED AI LAB.
(CDAAIL)

Contents :-

1) Types of Models based on Interpretability :-

1) GlassBox Models :- “Glass-Box models” are interpretable due to their structures and are completely exposed to the users. Eg :- Linear Models, Decision Trees.
2) BlackBox Models :- “Black-Box models” do not disclose anything about the internal design, structures of implementations. Eg :- Deep Neural Networks, SVM (Support Vector Machine), Random Forests, Gradient Boosting.

For Model Explainability the below Python Libraries we have used.

1) LIME Explanation :- Gives Local Explanations,
2) SHAP Explanation :-Gives Global Explanations

2) What are CounterFactual models and Explanations ?

1) “Countering the Facts of observed outputs(Effects)”, by making some changes in inputs(Cause).
2) A counterfactual explanation of a prediction describes the smallest changes to the feature values that changes the prediction to an observed output in users favour.
3) “If I would not have smoked, I might not be getting cancer”.

   “If I had studied harder, I would have passed the exam”.

4) By using the CounterFactual models we make the BlackBox models to Glass-Box by explaining the reasons for the results.

   “What features need to be changed and by how much to flip a model’s prediction?” (i:e to reverse an unfavorable outcome).

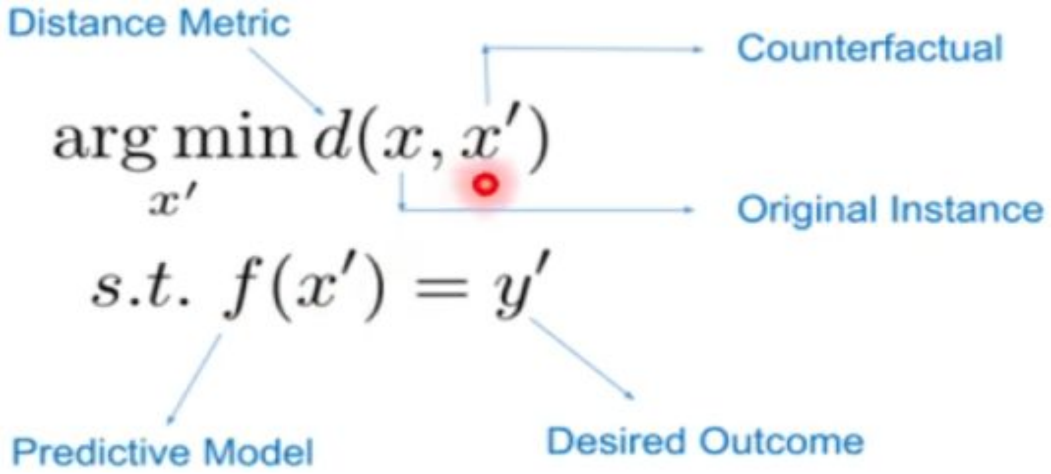
3) CounterFactual Explanations and Generative Adversarial Network(GAN).

1) Counterfactuals concepts are from (GANs) which are generative models, they create new data instances that resembles our training data and so CFS.
2) CFs is one of the(minimum) solution from all the set of the solutions.

4) Mathematical Modelling of CounterFactual Models.

Mathematically this is a optimization problem and our objective is to find x’ which is the counterfactual sample that changes the prediction of our black box model to a target class y’.

1) Our purpose is to find f(x’) that change the prediction to desirable class y’.
2) We know f(x)=y & f(x’)=y’.
   We need to find the i/p x’ so that it will fall into y’.



Choice of distance metric dictates what kinds of counterfactual are chosen.

Feasibility of CounterFactuals Models. [**Usten et al 2019**]

1) It is not always possible to change some of features eg :- Race, Gender or Color etc.

Take 2: Feasible and Least Cost Counterfactuals

$$\arg\min_{x'} d(x, x') \quad s.t.\ f(x') = y'$$

$$\arg\min_{x' \in \boxed{\mathcal{A}}} \boxed{cost}(x, x') \quad s.t.\ f(x') = y'$$

- $\mathcal{A}$ is the set of feasible counterfactuals (input by end user)
  - E.g., changes to race, gender are not feasible

1) We take the feasible solutions from the a set A, which is the set of all the counterfactuals inputs.
2) Cost function tells that how much it is difficult to go from x→ x' ?
3) The 'd' is the Manhattan distance. `Manhattan(A,B)= |x1-x2| + |y1-y2|`

Problem in the CounterFcatuals Generated. **[Usten et el 2019]**

1) The counterfactuals are biased against Age, Gender, Race.
2) Some CounteFactuaks are not feasible to act upon these features. As one cannot act upon these features.
3) So, our Algo should generate the feasible counterfactuals to act upon.
4) They used a strategy where x’ should be pick up from the set of feasible counterFactuals
5) Developed a Strategy where end user should inputs a set o counterfactuals and x’ should be chosen from that set of feasible counterfactuals.
6)

Finding 3 :- Causally Feasible Counteactuals or Attributes Dependency or Interactions among attributes. [**Mahajan et al**]

1) It is important to account for feature interactions when generating counterfactuals.
2) They are considering new distance metric d_causal(x,x’) , they suggested to use the SCM (Structure Causal Model) to define this new distance metric
3) This underlying causal models captures the feature interactions.
4)

Global CounterFactual Explanations : [**Rawal et al 2020**]

1) Useful for regulators to ensure that there should be not any bias against a particular community or race in AI model before implementing it on real-life situations.
2) Global counterfactual explanations are useful to know the behavior of the overall model.
3) They aggregate the Local CounterFcatuals to generate the Global counterFactuals.
4) For certain demography the model is asking to change lot more features than others.
5) So, Global counterfactuals are useful to find out any bias in model.

Dataset Description :-

1) “BirthsFinal Data for 2022”. Published by “**Centres for disease control and prevention (CDC)** . (U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES).
2)