

[FOR XIXI: Do you think that this belongs to background, too?]

For this measure, we evaluate the likelihood of a counterfactual trace by determining whether a counterfactual leads to the desired outcome or not. For this purpose, we use the predictive model, which returns a prediction for each counterfactual sequence. As we are predicting process outcomes, we typically predict a class. However, it is often difficult to force a deterministic model to produce a different class prediction. Therefore, we can relax the condition by maximising the prediction score of the desired counterfactual outcome[1]. If we compare the difference of the counterfactual prediction score with the factual prediction score, we can determine an increase or decrease. Ideally, we want to increase the likelihood of the desired outcome. We refer to this value as *delta*. However, the binary case introduces some noteworthy considerations.

Within this task setting, we have to consider multiple cases. First, prediction score which is typically limited to a domain within 0 and 1, which we can interpret as probability distribution. Hence, if the model score is 0.6, then the model has a confidence of 60% that the input can be categorised as belonging to class 1. For instance, within a medical process we could say, the model is 75% confident that the patient can be cured. Conversely, there's a 25% percent confidence that the process instance belongs to class 0. We can make decisions by using a threshold. Typically this threshold is lies at 50%. Hence, we determine that a patient can be cured if the model's confidence exceeds 0.5¹. If we want to determine a soft version of the delta measure, we have to take this decision threshold into account.

We identify 2 cases:

- Case 1: A counterfactual generator *flips* the prediction score to the opposite side of the decision threshold. Then, we archieve our general aim and the difference between the scores is a direct indicator of the counterfactual's success.
- Case 2: A counterfactual does not change the factual decision. For instance, when the counterfactual and the factual prediction score for a patients recovery chance are below 0.5 or both are above 0.5. Then we have to consider the whether the is moving towards the desired outcome or away from it.

[2.1] If the prediction for the factual decides an outcome of 0 but the predictions score for the counterfactual is even lower, then we did not change the prediction at all. In fact, we increase the chance of the factual outcome. That situation is worse than what we desire. For instance, a patient would not want to pursue a counterfactual situation

in which his odds of recovery are worse than his current.

[2.2] In contrast, if a prediction model's score leads to an outcome of 0 but the counterfactual returns a higher prediction score than for the factual predictions score, a patient might still be interested in the counterfactual. In some situations even a small improvement is desirable.

The subcases of 2 go in both ways. Hence, we have to incorporate each case differently in the delta score.

¹Obviously, the domain of the application decides where this threshold lies. One can always argue that a confidence of 51% is close to randomly guessing.