Explaining deep neural networks using counterfactuals

Mansi Shivani
Advisor: Dr. V. Susheela Devi

Indian Institute of Science

mansishivani@iisc.ac.in

May 27, 2024

Overview

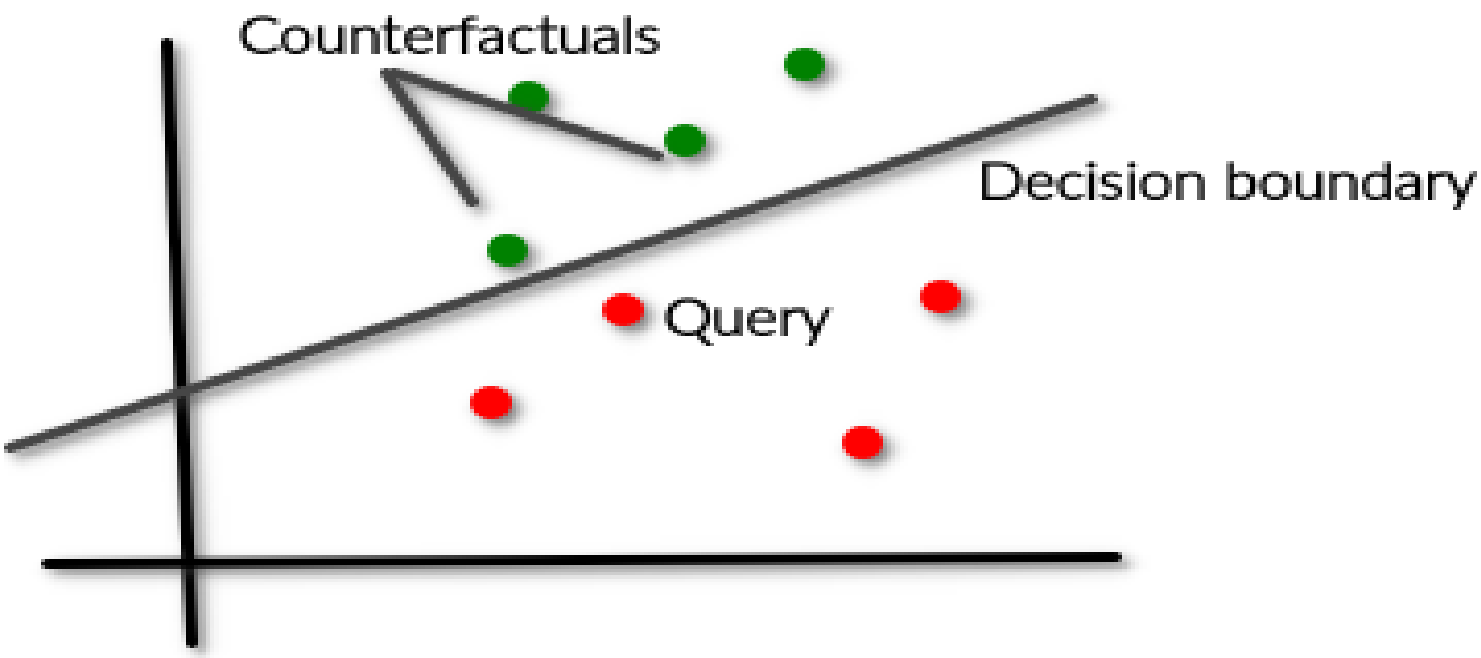
Explainable AI

- Explaining the outcomes of complex machine learning models is a perquisite for establishing trust between the machines and users. As humans increasingly rely on neural networks to process large amounts of data and make decisions, it is crucial to develop solutions that can interpret the predictions of such models.
- Explaining the outcomes of a model can help reduce bias and contribute to improvements in model design, performance, and accountability by providing beneficial insights into how models behave

Counterfactuals I

- A useful method for explaining single predictions of a model are counterfactual explanations.
- Counterfactuals provide explanations in the form of "if these features had different values, your outcome would have been positive".
- For people who receive adverse decision from a model is is important to know why they have not been given a positive outcome, either to understand the decision making process or to assess their actionable options to change the outcome.
- The task of finding a counterfactual explanation is then to find some X' that is in some way related to the original instance X but leading to a different prediction $y^{'}$

Counterfactuals II



Related work I

Many methods of CF generation treat it as an optimization problem.

1. A simple loss function as given by Watcher et al. 2018 would be: $L(x, x^{'}, y^{'}, \lambda) = \lambda(f(x^{'}) - y^{'})^2 + d(x, x^{'})$ But they do not handle categorical features.
2. S. Dandl et al. 2020 describe a more well-rounded loss function that takes care of a few more things than the previous one by adding terms for sparsity.
3. Diverse Counterfactual Explanations (DiCE) by Mothilal et al. 2020 aims at generating a set of counterfactual explanations that are diverse and actionable. They define the loss function over the entire set instead of a single counterfactual instance.

Related work II

Apart from treating CF generation as an optimization problem, a few novel approaches have been explored.

1. FACE: Feasible and Actionable Counterfactual Explanations by Poyiadzi et al. 2020 aims at finding high-density paths of change. It solves CF generation as a graph problem using a shortest path algorithm over the dataset.
2. Model-Agnostic Counterfactual Explanations for Consequential Decisions (MACE) by Karimi et al. 2020 propose a novel algorithm that solves a sequence of satisfiability problems, where both the distance function (objective) and predictive model (constraints) are represented as logic formulae. It maps the nearest CF problem into a sequence of satisfiability (SAT) problems and uses standard SMT (satisfiability modulo theories) solvers as black-box to solve it.

Algorithm

We propose a two part approach drawing from the concept of coordinate descent that generates valid, feasible and actionable counterfactuals while effectively handling categorical features.

Algorithm 1: Generate counterfactuals

Result: A set of k counterfactuals CF

x = original input;

k = no. of counterfactuals;

D = dataset used to train the model;

f = model used to make the decision;

CF = {};

Freeze all numerical features;

Find k closest examples in D to input x w.r.t constraints;

Put it in CF;

Unfreeze numerical features;

Freeze categorical features;

Algorithm

for $i \leftarrow 1$ **to** k **do**

cf init = CF[i];

$x^{'}$ = cf init;

L = d(x,$x^{'}$) + (1 - f($x^{'}$));

optimize($x^{'}$ w.r.t L);

CF[i] = $x^{'}$;

Algorithm

1. First, we find k instances in the dataset D that as closest to our input query with respect to categorical variables only. The instances chosen are subject to some criteria. They should belong to the desired class. In addition, they should follow certain constraints such as immutability or non-actionability of certain features.
2. Once we have our set, we use each instance as a starting point of the gradient descent optimization problem. We run the optimization algorithm for a few iterations (2-5) calculating loss only with respect to the numerical variables this time.
3. The final instances make our counterfactuals. We inverse transform the dataset and display the results.

Our contributions I

1. Most existing algorithms do not handle categorical features while others use distance measures meant for numerical features on label encoded dataset. Our algorithm is the first one that enables use of appropriate and statistically sound distance metrics for categorical features.
2. Handling numerical and categorical distances differently frees the categorical distance measure from the constraint of being differentiable.
3. We enforce constraints by having two attributes of 'Actionability' and 'Mutability' for each feature.
4. Finding initial points from the dataset instead of random assignment leads to faster convergence helps us stay true to the data distribution. Starting from points that are true to the distribution and follow the required constraints makes the resulting counterfactuals feasible and actionable.

Dataset and model used I

Attribute	Type
age	continuous
workclass	categorical
fnlwgt	continuous
education	categorical
education-num	continuous
marital-status	categorical
occupation	categorical
relationship	categorical
race	categorical
sex	categorical
capital-gain and gain	continuous
hours-per-week	continuous
native-country	categorical

Table 1:Attributes of the Adult Income dataset

Dataset and model used II

1. The dataset used for training a sequential model is a balanced Adult income dataset.
2. The dataset has 14 attributes consisting of both numerical and categorical features.
3. It has 7500 positive samples and 7500 negative samples. The model gives an accuracy of 82%

Input to the algorithm

	age	workclass	fnlwgt	education	education_num	...	cap_gain	cap_loss	hrs_per_week	native_country
0	20	Private	170091	Some-c...	10	...	0	0	10	United...

1 rows × 15 columns

confidence score of positive outcome

0.0019922063

Figure 1 : Input to the algorithm

Output generated by the algorithm

	age	workclass	fnlwgt	education	...	cap_loss	hrs_per_week	native_country	Confidence
0	37	Private	194745	Bachelors	...	100	44	United...	0.913220
1	49	Private	207770	HS-grad	...	64	44	United...	0.790139
2	35	Private	135873	Assoc-voc	...	28	39	United...	0.749209
3	32	Private	136955	Some-c...	...	0	28	United...	0.999907

Figure 2 : Generated counterfactuals

Evaluation of CFs I

1. Evaluation of generated counterfactual explanations is an issue of main concern. Unfortunately, despite an increasingly expanding literature of counterfactual explanations, no uniform set of evaluation methods has been adopted so far.
2. Most publications do not evaluate their frameworks while others go for subjective evaluation. Recent papers have tried to include objective evaluation. But there are no clear standard metrics on which existing and new methods can be compared in a fair manner.

Validity

Validity is simply the percentage of counterfactual examples that fall in the desired class.

$$Validity\% = \frac{|\{c \text{ in C s.t} f(c) > 0.5|\}}{k}$$

Evaluation of CFs II

Proximity

It measures how close the counterfactuals are to the input.

$$Proximity = -\frac{1}{k}\sum_{i=1}^{k} d(c_i, x)$$

Sparsity

Sparsity measures the number of features in which a change is suggested by the counterfactual. In reality, it may be challenging to make a vast number of changes to get the desired outcome. CFs with fewer and more minor changes are preferable.

$$Sparsity = 1 - \frac{1}{kd}\sum_{i=1}^{k}\sum_{l=1}^{d} I_{ci \neq xi}$$

Evaluation of CFs III

Diversity

Different things may work for different users, and hence diversity ensures that we have something for everyone. The metric can be defined as the mean of summed distance between each pair or the mean of sparsity.

$$Diversity = \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{i=j+1}^{k} d(c_i, c_j)$$

$$Count - diversity = \frac{1}{k(k-1)d} \sum_{i=1}^{k-1} \sum_{i=j+1}^{k} \sum_{l=1}^{d} I_{ci \neq xi}$$

Evaluation of CFs IV

Results

Results			
Metric	**k=2**	**k=6**	**k=10**
Validity	95.7%	97.2%	98.3%
Proximity	-15.6	-14.8	-13.4
Sparsity	0.52	0.50	0.48
Diversity	20.52	17.43	14.27
Count Diversity	0.535	0.534	0.539

Table 2 : Metrics for different values of k

For all the metrics, higher values are preferable.
Our algorithm performs well with respect to validity and sparsity.
Mothilal et al. 2020 report similar values for sparsity and count diversity.
They get a near perfect validity%.

Distance measures for categorical variables I

Overlap

Overlap dissimilarity measure is one of the most common distance measures used in the literature due to its simplicity. In overlap distance measure, a distance value of 0 is assigned for matches and 1 is assigned for mismatches.

$$D(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k \neq Y_k \\ 0 & \text{otherwise} \end{cases}$$

Distance measures for categorical variables II

Eskin

It assigns a weight of $\frac{2}{nk^2}$, where n_k is the number of different values the k^{th} feature takes.

$$D(X_k, Y_k) = \begin{cases} \frac{2}{nk^2} & \text{if } X_k \neq Y_k \\ 0 & \text{otherwise} \end{cases}$$

This measure gives more weight to mismatches that occur on attributes that take many values.

Distance measures for categorical variables III

Goodall

This distance measure always gives one for mismatches but takes occurrence values into account for matches. $p(x_i)$ is the fraction of records in which the i^{th} feature takes on the value of x_i

$$D(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k \neq Y_k \\ p(x_i)^2 & \text{otherwise} \end{cases}$$

It is a distance measure that evaluates diagonal entries only.

Distance measures for categorical variables IV

Comparison of CFs generated using different distance measures

Results			
Metric	**Overlap**	**Eskin**	**Goodall**
Validity	95.6%	96.8%	100%
Proximity	-13.6	-13.0	-57.7
Sparsity	0.51	0.50	0.22
Diversity	15.7	16.0	28.6
Count Diversity	0.53	0.50	0.71

Table 3 : Metrics for different categorical distance measures

Overlap is a naive distance measure often used for categorical features. Eskin measures distance for off-diagonal entries and goodall measures distance for diagonal entries only. From the table we can infer that , diagonal measuring distance measures give good validity and diversity but not good proximity.

Comparison with random initialization and MO I

To understand the importance of initial point, we compare results between our algorithm and if we just used random points from dataset as our initial points.
The technique of finding the counterfactuals from the dataset itself using some distance measure is called Minimum Observable CF.
Parameters used are k=4 and 4 iterations of gradient descent

Results			
Metric	**RandomCF**	**MO**	**Our algorithm**
Validity	67.5%	65.5%	99.3%
Proximity	-22.4	-24.8	-13.9
Sparsity	0.25	0.43	0.52
Diversity	20.9	17.4	16.2
Count Diversity	0.63	0.58	0.53

Table 4 : Comparison with some basic algorithms

Comparison with random initialization and MO II

	Positive class probability of init CF	Positive class probability of final CF
0	0.0575	0.910
1	0.0637	0.790
2	0.6582	0.749
3	0.9900	0.990

Figure 3 : Model prediction before and after gradient descent optimization

Comparison with random initialization and MO III

1. We show the importance of finetuning the initial samples using descent algorithm. There is a huge bump in classification confidence.
2. The reason for this is that no model has a 100% accuracy. A sample's true label and model prediction may not always match. A good counterfactual has to balance between the data distribution and model prediction. Our two step approach tries to do that by choosing starting point close to the query but then also optimizing with respect to prediction score.

Conclusion and Future Work

1. Counterfactuals fulfill most of the criteria of a good explanation and have many practical applications.
2. Future work concerns deeply analyzing the different features that will generate sensible and diverse counterfactual explanations, making them robust and scalable, present them in a way best suitable for every individual user taking into account the feasibility in practice.
3. The proposed algorithm is capable of encoding simple constraints that the data must follow. However,there are inherent causal relationships among the attributes. It is essential to ensure that those also hold. Finding a way to incorporate causal constraints is essential.

References I

1. Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." arXiv preprint arXiv:2006.11371 (2020).
2. Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL and Tech. 31 (2017): 841.
3. Dandl, Susanne, et al. "Multi-Objective Counterfactual Explanations." arXiv preprint arXiv:2004.11165 (2020).
4. Verma, Sahil, John Dickerson, and Keegan Hines. "Counterfactual Explanations for Machine Learning: A Review." arXiv preprint arXiv:2010.10596 (2020).
5. Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.

References II

6. Poyiadzi, Rafael, et al. "FACE: feasible and actionable counterfactual explanations." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.
7. Karimi, Amir-Hossein, et al. "Model-agnostic counterfactual explanations for consequential decisions." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.
8. Joshi, Shalmali, et al. "Towards realistic individual recourse and actionable explanations in black-box decision making systems." arXiv preprint arXiv:1907.09615 (2019).
9. Stepin, Ilia, et al. "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence." IEEE Access 9 (2021): 11974-12001.
10. Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Adult Income. http://archive.ics.uci.edu/ml/datasets/Adult

The End