View Reviews

Paper ID 10778

Paper Title Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models

Reviewer #1

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The paper introduces a new way of calculating Shapley values for local model explanations. The authors propose "causal Shapley values," which differ from existing Shapley value methods in some respects. If I am correct in my understanding of the paper, the paper proposes using a different cooperative game for representing the model's dependence on subsets of features.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The paper presents a different perspective on how to integrate causal knowledge into Shapley value-based model explanations. The authors contrast their approach with another recent paper that takes a different approach to integrating causal knowledge, asymmetric Shapley values.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
I re-read this paper several times and found it very hard to follow. In particular, I could not find a part of the paper that clearly defined "causal Shapley values," which makes evaluating the paper rather difficult. I'm open to changing my opinion if the other reviewers disagree, but my impression is that this paper's clarity is a serious problem. I'll elaborate on this and a couple other thoughts below.

On the paper's clarity problems:
- The paper should be much more clear when describing its proposal, "causal Shapley values." Section 2 seemed to be the place where they were defined, and indeed Section 2 compares them with asymmetric Shapley values (so we should presumably know what they are by this point, right?). But an application of the Shapley value to Eq. 3 is precisely the marginal Shapley value, also known as the interventional Shapley value. So ultimately I couldn't tell whether this section introduces causal Shapley values or not.
- Section 2 says "interventional Shapley values conveniently simplify to marginal Shapley values." The discussion about the difference between the interventional approach and the marginal approach is a bit odd. It seems obvious, based on the definition of the do operator, that "conditioning by intervention" is equivalent to using the marginal distribution. Why are these initially presented as two different options? And why are distinctions drawn with Janzing et al. (lines 105-107), given that this approach does precisely the same thing?
- Other parts of the paper seem to clearly indicate that causal Shapley values are not an application of the Shapley value to Eq. 3. For example, Section 4 distinguishes between symmetric causal Shapley values and a variety of other Shapley variants. So we should certainly know causal Shapley values are by now, right?
- Section 6 says that the Shapley value estimation algorithms were modified to use the conditional interventional distribution, as given by Eq. 7. But the paper has no equation 7!
- In the end, my best guess was that causal Shapley values were somehow defined in Section 5. If so, the authors should have been much more clear about this, and probably should not have contrasted causal Shapley values with other variants throughout the paper. The authors might consider explicitly saying something along the lines of "we define causal Shapley values as follows," or "we define causal Shapley values as the Shapley values of the following cooperative game."

About the distinction between direct and indirect effects:

- This section (Section 3) requires substantial clarification. The extra term that's introduced to distinguish between direct and indirect effects should be explained in more detail. If I'm not mistaken, it seems to be equal to zero. The text below the math suggests that marginal Shapley values, which are equivalent to what has been decomposed here (the interventional Shapley value), has no indirect effect. Does that mean that the two terms on the second line are equal?
- Since this direct/indirect effect decomposition is hard to follow, the authors might consider writing the direct/indirect effect decomposition for conditional Shapley values as well, or at the very least putting it in the supplement, rather than just mentioning that such a decomposition can be made.
- If the conditional Shapley values permit this decomposition, why is an alternative preferred over them? Is there something wrong with how conditional Shapley values calculate direct and indirect effects?
- Overall, I could not tell what this section was trying to show. Was it just trying to show that marginal Shapley values have no indirect effect? Are the authors even talking about causal Shapley values here?

About the use of variable orderings in Section 5: using the terminology of causal chain graphs is new relative to the ASV paper, but the outcome of being able to being able to use partial causal orderings is not. ASVs use a very similar partial ordering concept in their experiments.

I did not find the empirical comparison of causal Shapley values with other Shapley variants (Figure 4) very convincing. This plot shows just two predictions with just two features' attributions, so this does not provide strong evidence in favor of causal Shapley values.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

The claims seem correct.

5. Clarity: Is the paper well written?

The paper is lacking in clarity. After re-reading it several times, I could not find the part of the paper that clearly defines their proposal, causal Shapley values. I'm open to re-evaluating the ideas in this paper if the authors can clarify their contribution, or if the other reviewers found the paper more comprehensible.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

The authors do a good job of citing plenty of relevant work.

7. Reproducibility: Are there enough details to reproduce the major results of this work?

Yes

9. Please provide an "overall score" for this submission.

4: An okay submission, but not good enough; a reject.

10. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

Yes

Reviewer #2

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

The paper describes a way to define causal feature relevance analysis. In contrast to the popular approach with Shapley values it averages only over orderings that are consistent with the causal order and uses Pearl's do-operator to obtain interventional instead of observational probabilities.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

The paper discusses issues with previous approaches convincingly. The suggested approach is at least worth to be discussed, although it may not be the final word on this complex topic (there probably isn't a final word). The decomposition of total effects into direct and indirect effect is inspiring, also the discussion of chain graphs with confounding and directional edges.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

The discussion of the decomposition of total into direct and indirect effect is interesting, but lacks systematic consideration since it is not mentioned how this generalizes to a larger number of paths. Given my comments below, the paper could be a bit more critical about its own concepts since quantifying influence is a difficult topic and every definition comes with its now weaknesses.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

As far as I can see the claims are correct.

5. Clarity: Is the paper well written?

The paper is mostly well-written. I liked particularly the interesting references to the psychological literature. Section 4 is unfortunately the weakest part in this regard, hard to say what its message is supposed to be. I found its last paragraph entirely confusing: why does sampling counterfactuals correspond to *averaging* over ordering in feature attribution quantification?

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

The only question remaining is regarding novelty compared to ref [6] (arXiv preprint).

7. Reproducibility: Are there enough details to reproduce the major results of this work?

Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

Every quantification of causal influences raises tons of questions. Here are some thoughts on the present one, which I would appreciate to be mentioned:

- although averaging over all orderings consistent with the DAG sounds natural, it comes with a serious conceptual issue, namely the discontinuity with respect to inserting arrows of zero strength. Assume X and Y are independent causes of Z. Then there are two possible orderings. Insert an arrow from X to Y with negligible impact, then there is only one ordering. Compare supplement S3 of 'Quantifying causal influences' by Janzing et al. for a similar problem

- issues with discontinuity at arrows with zero strength can be avoided by defining strength via the influence of the noise terms, see e.g. https://arxiv.org/abs/2007.00714
However, this solution comes with another conceptual problem: influence quantification then depends on the functional causal model and is no longer uniquely determined by interventional probabilities.

Regarding the criticism of ref [9]: it seems to me that it comes with a different scope, namely feature relevance quantification in the sense of understanding an algorithms rather than understanding causal relations in the real world. Maybe one could mention this difference. Note also that it appears at AISTATS, it's already online.

- I was struggling with the remark 'not every causal query need be identifiable (see e.g., [24])' if the DAG is known and all nodes are observed, which type of causal queries are not identifiable?

9. Please provide an "overall score" for this submission.
8: Top 50% of accepted NeurIPS papers. A very good submission; a clear accept.

10. Please provide a "confidence score" for your assessment of this submission.
4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #3

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The paper considers locally explaining predictions of a predictive model when the features are not independent. The paper proposes a method by which to compute Shapley values when we intervene on each feature, instead of conventional conditioning. The proposed approach considers access to the causal chain graph of the problem at hand, and derives Shapley values of the (intervened) features according to that. The proposed approach is tested on one dataset and one predictive model.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The problem and the approach are interesting.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
The theoretical and practical contributions of the paper are very limited.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
Yes, but only in limited settings.

5. Clarity: Is the paper well written?
Yes.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Yes.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
Overall:
The paper considers locally explaining predictions of a predictive model when the features are not independent. The paper proposes a method by which to compute Shapley values when we intervene on each feature, instead of conventional conditioning. The proposed approach considers access to the causal chain graph of the problem at hand, and derives Shapley values of the (intervened) features according to that. The proposed approach is tested on one dataset. The problem and the approach are interesting, but the theoretical

and practical contributions of the paper are limited. The theoretical contribution is limited to a special case, which limits applicability of the proposed approach. Further, the approach is only tested on one dataset with 7 features and only one trained ML model, i.e., XGBoost. This is not convincing enough. More methodologies, datasets, and a wider application setting must be considered. My assessment of this paper is a clear cut reject.

Major:
- The proposed methodology depends on cases where we can intervene on variables (do(x); Pearl's second rung on the ladder of causality; refer to the Book of Why). However, oftentimes, when one needs to provide an explanation for a predictive model, experimentation (do(x)) is not a viable option. One way of dealing with this is using counterfactual inference (Pearl's rung 3 in the ladder of causation). The proposed methodology would be much more convincing it if in addition to interventions, it would account for counterfactual inference. The authors could show this in their setting by answering the question: What would the predicted bike count be, if the season was spring, and the temperature was 20?

- The main contribution of the paper is summarized in Theorem 1. This theorem only holds for causal chain graphs, which are a special case, and make the contribution very limited. How would the proposed methodology work in other situations, such as when there are forks, (un)shielded colliders, etc.? Identifying the admissible set that would make the causal effect identifiable in these cases have been studied in the literature, but are not discussed in this paper. How would the proposed approach deal with such cases? The domain of applicability of the proposed methodology is very limited.
I appreciate the discussion of the chain, fork, confounder, and cycle options in Figure 1, but not all of them are not realistic in the case of the provided running example of the paper, i.e., bike rentals. How can temperature causally affect season in the fork? Perhaps, a better example could help.
The proposed approach is only tested on one real-world toy data with only 7 features according to the text. This does not show the scalability of the proposed method in practical examples.
It is okay to test the approach on an XGBoost model, but we should also see the results on at least one deep neural network too, because neural networks are a main concern among the XAI community.

- Causal relationships are asymmetric in nature. For example, if A causes B, then B might not necessarily cause A. It's not clear why the authors are calling causal values as symmetric.
Results of experiments in figure 4 are not convincing. If the temperature for Oct and Dec, in this example, are roughly the same, then the different in predicted bike count should be attributed to, perhaps, the seasonal difference, which is represented by cosine of year in this case. But the causal Shapley values have computed a lower effect (in magnitude) for cosine of year. Why is it the case? And why should we trust the explanation generated by the causal Shapley value instead of Asymmetric Shapley values.

Minor:
- There is inconsistency in references style. While many journal/conference names follow a capitalized first letter paradigm, such as Advances in Neural Information Processing Systems (reference 20), others do not follow this, such as Knowledge and information systems (reference 31). This inconsistency is apparent in many references. Careless writing, please fix.
- Line 230, clicking on "interventional conditional distribution (7)" takes me to Formula number 6, not 7.

9. Please provide an "overall score" for this submission.

3: A clear reject.

10. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

Yes

Reviewer #4

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

This paper proposes a causal approach to Shapley values, which are values that are used to explain what features of the input data to a model contributed to the model's output. By using a causal approach, dependencies between features in the input data can be dealt with, which was not possible before.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

This is a truly great paper. The contribution of this paper is outstanding, both from a conceptual point of view (letting explainable AI benefit from a causality-based conceptualization; I) and from the results (the computations are feasible and the provided real-world example is sensible). It is a step in the right direction and it is a noticeable step, not just an idea.

Secondly, this paper is incredibly well written. It was a delight to read it and I learned a lot in a short amount of time. This is the best paper I've read in a while. Reading it made up for the horrible paper I had to review earlier today.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

Of course one can always request more (more analyses, more datasets, more causal structures to test the approach on) but in my view, this paper is well-rounded and I didn't miss a single thing.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

I did not spot any mistakes.

I cannot vouch for the correctness of the most technical aspects of this work though, as I'm not an expert in Pearl's do-calculus or Shapley values.

5. Clarity: Is the paper well written?

Thank you for writing such a clear paper.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

Yes, perfectly.

7. Reproducibility: Are there enough details to reproduce the major results of this work?

Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

Suggestion for improvement: In Fig 3, bottom-left corner plots, make the grid lines are the same for the x- and y-axis, and perhaps change the x-axis tick labels to -500 and 500, too. That would make it easier to see how the range of the values differ (as emphasized in the main text). In the current form, I had to stare at it for a bit to convince myself that the scales are the same and it is not an effect of stretching the x-axis.

I'm excited to see in future work, how the comparison to human subjects' judgments will work out.

9. Please provide an "overall score" for this submission.

10: Top 5% of accepted NeurIPS papers. Truly groundbreaking work.

10. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes