Instructions for Paper Submissions to AISTATS 2023

Anonymous Author
Anonymous Institution

Abstract

The Abstract paragraph should be indented 0.25 inch (1.5 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The **Abstract** heading must be centered, bold, and in point size 12. Two line spaces precede the Abstract. The Abstract must be limited to one paragraph.

1 INTRODUCTION

XAI literature distinguishes between local and global interpretation methods (Molnar et al., 2020). Local methods explain a specific prediction, whereas global methods explain the entire model behavior. Global methods provide a universal explanation, summarizing the numerous local explanations into a single interpretable outcome (number or plot). For example, if a user wants to know which features are significant (feature importance) or whether a particular feature has a positive or negative effect on the output (feature effect), they should opt for a global explainability technique. Aggregating the individual explanations for producing a global one comes at a cost. In cases where feature interactions are strong, the global explanation may obfuscate heterogeneous effects (Herbinger et al., 2022) that exist under the hood, a phenomenon called aggregation bias (Mehrabi et al., 2019).

Feature effect forms a fundamental category of global explainability methods. The goal of the feature effect is to isolate the average impact of a single feature on the output. Feature effect methods suffer from aggregation bias because the rationale behind the average effect might be unclear. For example, a feature with zero average effect may indicate that the feature has no effect on the output or has a highly positive effect in some cases and a highly negative one in others.

Preliminary work. Under review by AISTATS 2023. Do not distribute.

There are two widely-used feature effect methods; Partial Dependence Plots (PDPlots) (Friedman, 2001) and Aggregated Local Effects (ALE)(Apley and Zhu, 2020). PDPlots have been criticized for producing erroneous feature effect plots when the input features are correlated due to marginalizing over out-of-distribution synthetic instances. Therefore, ALE has been established as the state-of-the-art feature effect method since it can isolate feature effects in situations where input features are highly correlated.

However, ALE faces two crucial drawbacks. First, it lacks a way to inform the user about potential heterogeneous effects that hide behind the average effect. In contrast, in Partial Dependence Plots (PDP), the heterogeneous effects can be spotted by exploring the Individual Conditional Expectations (ICE). Second, ALE requires an additional step, where the axis of the feature of interest is split in K fixed-size non-overlapping intervals, where K is a hyperparameter provided by the user. This splitting is done blindly, which can lead to inconsistent explanations.

In this paper, we extend ALE with a probabilistic component for measuring the uncertainty of the global explanation. The uncertainty of the global explanation expresses how certain we are that the global (expected) explanation is valid if applied to an instance drawn at random, or, in other words, the level of heterogeneous effects hidden behind the expected explanation. The uncertainty of the global explanation emerges from the natural characteristics of the experiment, i.e., the data generating distribution and the black-box function.

Furthermore, we automate the step of splitting the axis into non-overlapping intervals. The need for non-overlapping bins emerges from the ignorance of the data-generating distribution, enforcing all estimations to be based on the limited instances of the training set. Therefore, there is an implicit trade-off behind the formation of bins. Each bin must include enough instances for a robust estimation of the bin feature effect (expected value), and the uncertainty of the explanation (variance). On the other hand, each bin should include points with similar local effects. Therefore, we transform the bin splitting step into an unsupervised clustering problem, encoding the trade-off mentioned above in the objective function. We formally show that the objective of the clustering problem has lower-bound the (unavoid-

able) heterogeneity, i.e., the aggregated uncertainty of the global explanation. Therefore, we aim to find the optimal grouping of samples that adds the slightest uncertainty over the unavoidable heterogeneity. We finally solve the minimization problem by finding the global optimum using dynamic programming. Our method works out of the box without requiring any input by the user. We provide a theoretical and empirical evaluation of our method.

2 BACKGROUND AND RELATED WORK

3 THE U-DALE METHOD

3.1 Uncertainty Quantification

3.1.1 Methodology

3.2 Bin Spliting as a Clustering Problem

3.2.1 Methodology

3.2.2 Algorithms

4 SYNTHETIC EXAMPLES

5 REAL-WORLD EXAMPLES

6 GENERAL FORMATTING INSTRUCTIONS

The camera-ready versions of the accepted papers are 8 pages, plus any additional pages needed for references.

Papers are in 2 columns with the overall line width of 6.75 inches (41 picas). Each column is 3.25 inches wide (19.5 picas). The space between the columns is .25 inches wide (1.5 picas). The left margin is 0.88 inches (5.28 picas). Use 10 point type with a vertical spacing of 11 points. Please use US Letter size paper instead of A4.

Paper title is 16 point, caps/lc, bold, centered between 2 horizontal rules. Top rule is 4 points thick and bottom rule is 1 point thick. Allow 1/4 inch space above and below title to rules.

Author descriptions are center-justified, initial caps. The lead author is to be listed first (left-most), and the Coauthors are set to follow. If up to three authors, use a single row of author descriptions, each one center-justified, and all set side by side; with more authors or unusually long names or institutions, use more rows.

Use one-half line space between paragraphs, with no indent.

7 FIRST LEVEL HEADINGS

First level headings are all caps, flush left, bold, and in point size 12. Use one line space before the first level heading and one-half line space after the first level heading.

7.1 Second Level Heading

Second level headings are initial caps, flush left, bold, and in point size 10. Use one line space before the second level heading and one-half line space after the second level heading.

7.1.1 Third Level Heading

Third level headings are flush left, initial caps, bold, and in point size 10. Use one line space before the third level heading and one-half line space after the third level heading.

Fourth Level Heading Fourth level headings must be flush left, initial caps, bold, and Roman type. Use one line space before the fourth level heading, and place the section text immediately after the heading with no line break, but an 11 point horizontal space.

7.2 Citations, Figure, References

7.2.1 Citations in Text

Citations within the text should include the author's last name and year, e.g., (Cheesman, 1985). Be sure that the sentence reads correctly if the citation is deleted: e.g., instead of "As described by (Cheesman, 1985), we first frobulate the widgets," write "As described by Cheesman (1985), we first frobulate the widgets."

The references listed at the end of the paper can follow any style as long as it is used consistently.

7.2.2 Footnotes

Indicate footnotes with a number[1] in the text. Use 8 point type for footnotes. Place the footnotes at the bottom of the column in which their markers appear, continuing to the next column if required. Precede the footnote section of a column with a 0.5 point horizontal rule 1 inch (6 picas) long.[2]

7.2.3 Figures

All artwork must be centered, neat, clean, and legible. All lines should be very dark for purposes of reproduction, and art work should not be hand-drawn. Figures may appear at

[1] Sample of the first footnote.

[2] Sample of the second footnote.

the top of a column, at the top of a page spanning multiple columns, inline within a column, or with text wrapped around them, but the figure number and caption always appear immediately below the figure. Leave 2 line spaces between the figure and the caption. The figure caption is initial caps and each figure should be numbered consecutively.

Make sure that the figure caption does not get separated from the figure. Leave extra white space at the bottom of the page rather than splitting the figure and figure caption.

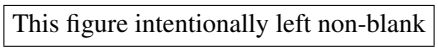

Figure 1: Sample Figure Caption

7.2.4 Tables

All tables must be centered, neat, clean, and legible. Do not use hand-drawn tables. Table number and title always appear above the table. See Table 1.

Use one line space before the table title, one line space after the table title, and one line space after the table. The table title must be initial caps and each table numbered consecutively.

Table 1: Sample Table Title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

8 SUPPLEMENTARY MATERIAL

If you need to include additional appendices during submission, you can include them in the supplementary material file. You can submit a single file of additional supplementary material which may be either a pdf file (such as proof details) or a zip file for other formats/more files (such as code or videos). Note that reviewers are under no obligation to examine your supplementary material. If you have only one supplementary pdf file, please upload it as is; otherwise gather everything to the single zip file.

You must use `aistats2023.sty` as a style file for your supplementary pdf file and follow the same formatting instructions as in the main paper. The only difference is that it must be in a *single-column* format. You can use `supplement.tex` in our starter pack as a starting point. Alternatively, you may append the supplementary content to the main paper and split the final PDF into two separate files.

9 SUBMISSION INSTRUCTIONS

To submit your paper to AISTATS 2023, please follow these instructions.

1. Download `aistats2023.sty`, `fancyhdr.sty`, and `sample_paper.tex` provided in our starter pack. Please, do not modify the style files as this might result in a formatting violation.
2. Use `sample_paper.tex` as a starting point.
3. Begin your document with

   ```
   \documentclass[twoside]{article}
   \usepackage{aistats2023}
   ```

   The `twoside` option for the class article allows the package `fancyhdr.sty` to include headings for even and odd numbered pages.
4. When you are ready to submit the manuscript, compile the latex file to obtain the pdf file.
5. Check that the content of your submission, *excluding* references, is limited to **8 pages**. The number of pages containing references alone is not limited.
6. Upload the PDF file along with other supplementary material files to the CMT website.

9.1 Camera-ready Papers

If your papers are accepted, you will need to submit the camera-ready version. Please make sure that you follow these instructions:

1. Change the beginning of your document to

   ```
   \documentclass[twoside]{article}
   \usepackage[accepted]{aistats2023}
   ```

   The option `accepted` for the package `aistats2023.sty` will write a copyright notice at the end of the first column of the first page. This option will also print headings for the paper. For the *even* pages, the title of the paper will be used as heading and for *odd* pages the author names will be used as heading. If the title of the paper is too long or the number of authors is too large, the style will print a warning message as heading. If this happens additional commands can be used to place as headings shorter versions of the title and the author names. This is explained in the next point.

2. If you get warning messages as described above, then immediately after `\begin{document}`, write

   ```
   \runningtitle{Provide here an
   alternative shorter version of the
   title of your paper}
   \runningauthor{Provide here the
   surnames of the authors of your
   paper, all separated by commas}
   ```

   Note that the text that appears as argument in `\runningtitle` will be printed as a heading in the *even* pages. The text that appears as argument in `\runningauthor` will be printed as a heading in the *odd* pages. If even the author surnames do not fit, it is acceptable to give a subset of author names followed by "et al."

3. The camera-ready versions of the accepted papers are 8 pages, plus any additional pages needed for references.

4. If you need to include additional appendices, you can include them in the supplementary material file.

5. Please, do not change the layout given by the above instructions and by the style file.

Acknowledgements

All acknowledgments go at the end of the paper, including thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support. To preserve the anonymity, please include acknowledgments *only* in the camera-ready papers.

References

References follow the acknowledgements. Use an unnumbered third level heading for the references section. Please use the same font size for references as for the body of the paper—remember that references do not count against your page length total.

References

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82:1059–1086, 2020. ISSN 14679868. doi: 10.1111/rssb.12377. The paper that proposed ALE plots.

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Repid: Regional effect plots with implicit interaction detection. 2 2022. URL `http://arxiv.org/abs/2202.07254`.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 8 2019. URL `http://arxiv.org/abs/1908.09635`.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges, 10 2020.