R-LIME: Rectangular Constraints and Optimization for Local Interpretable Model-agnostic Explanation Methods*

Genji Ohara[1][0000−1111−2222−3333], Keigo Kimura[2,3][1111−2222−3333−4444], and Mineichi Kudo[3][2222−−3333−4444−5555]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. In recent years, complex machine learning models such as deep neural network or random forests are used in many situations because of their high accuracy. However, complexity of the models leads to difficulty for introducing them in sensitive context because users cannot understand the reason why the models output the specific results. We propose a new method to generate explanations about the complex classifier and its particular output, that is called R-LIME. R-LIME approximizes a complex decision boundary by a linear classifier locally and maximizes the approximation region as long as the accuracy of the linear model is higher than given threshold.

Keywords: Interpretable machine learning · Local surrogate model

1 Introduction

In recent years, complex machine learning models such as deep neural network or random forests are used in many situations because of their high accuracy. However, complexity of the models leads to difficulty for introducing them in sensitive context because users cannot understand the reason why the models output the specific results. To solve this problem, many explanation methods have been proposed.

In this paper, we focus on local and model-agnostic explanation methods. LIME (Local Interpretable Model-agnostic Explanation) [1] is one of the most popular explanation methods. LIME approximizes a complex decision boundary by a linear classifier locally. But users cannot evaluate generality of LIME explanations because it does not provide the approximation region. To solve this problem, we propose a new method, R-LIME. R-LIME approximizes a complex

* Supported by organization x.

decision boundary by a linear classifier locally and maximizes the approximation region as long as the accuracy of the linear model is higher than given threshold. Moreover, R-LIME provides the approximation region as conjunction of predicates, that is called anchor [2].

2 Related Works

2.1 LIME

2.2 Anchor

2.3 BELLA

3 Proposed Method

3.1 Overview

3.2 Algorithm

4 Experiments

4.1 Qualitative Evaluation

4.2 Quantitative Evaluation

5 Future Works

6 Conclusion

6.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

Subsequent paragraphs, however, are indented.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

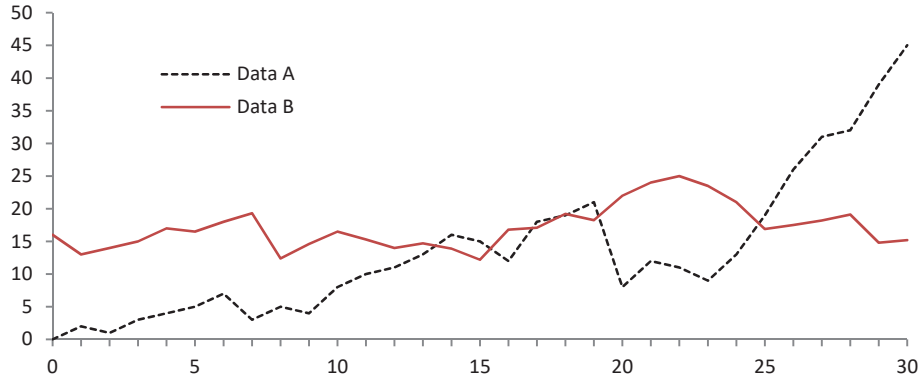

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.
Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].

Acknowledgements Please place your acknowledgments at the end of the paper, preceded by an unnumbered run-in heading (i.e. 3rd-level heading).

References

1. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
2. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1), 1527–1535 (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11491, https://ojs.aaai.org/index.php/AAAI/article/view/11491

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017