

The metric-aware kernel-width choice for LIME

Aurelio Barrera-Vicent^{1,*}, Eduardo Paluzo-Hidalgo² and Miguel A. Gutiérrez-Naranjo¹

¹Dept. Computer Science and Artificial Intelligence, University of Seville, Spain

²Dept. Applied Mathematics I, University of Seville, Spain

³institution, address, country

Abstract

Local Interpretable Model-Agnostic Explanations (LIME) are a well-known approach to provide local interpretability to Machine Learning models. LIME uses an exponential smoothing kernel based on the kernel width value, which defines the width of the local neighbourhood. In this paper, we study the influence of the distances for these local explanations, and we explore the choice of kernel width to guarantee a fair performance comparison between the distances.

Keywords

XAI, LIME, Kernel width, Explainability

1. Introduction

The gap between the recent development of AI models and their social use has led the scientific community to develop a new research area called Explainable Artificial Intelligence (XAI) (see, e.g., [1, 2, 3, 4, 5, 6] for some comprehensive surveys). According to the literature, XAI must contain a set of techniques to provide clear, intelligible, trustworthy, and interpretable explanations of the decisions, predictions, and reasoning processes made by AI models. From a technical point of view, there are many criteria for building a taxonomy of XAI methods [7]: Model agnostic vs. model specific; intrinsic vs. post hoc; etc. One of them considers whether the explanation is *local* or *global* [8].

One of the key points in these local explanations is the meaning of *local*. In this way, the sense of proximity among points in the training dataset of the model plays a central role in local explanations, and hence the choice of the distance considered is crucial in order to find plausible explanations. In this paper, we consider one of the most widely used XAI methods, the well-known method LIME [9] and we consider the influence of considering several metrics on it and how the stability and adherence of the model [10] perform when the dimension of the dataset grows and study the definition of a fair parameter (the kernel width) for a trustworthy

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.


✉ aurbarvic@alum.us.es (A. Barrera-Vicent); epaluzo@us.es (E. Paluzo-Hidalgo); magutier@us.es (M. A. Gutiérrez-Naranjo)

🌐 <https://personal.us.es/epaluzo/> (E. Paluzo-Hidalgo); <https://www.cs.us.es/~naranjo/> (M. A. Gutiérrez-Naranjo)

🆔 0009-0002-0165-3110 (A. Barrera-Vicent); 0000-0002-4280-5945 (E. Paluzo-Hidalgo); 0000-0002-3624-6139 (M. A. Gutiérrez-Naranjo)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

performance.

2. Related work

LIME is a XAI model to provide interpretability for individual instances¹. Each local explanation is a set of feature-value pairs that determine which features provide a greater contribution to the prediction, together with a numeric value that quantifies this contribution. In the literature, we can find several studies of LIME focussing on different improvements on the original algorithm as [11, 12, 13, 14, 15, 16, 10]. LIME algorithm tries to maximize the local fidelity of the explanations approximating the model to be explained M by a simpler model while having a low complexity (in terms of human readability) of the interpretable model. In LIME, by default the simpler model is a Ridge linear model. The input of LIME is a model M and an individual data sample x to be explained. To generate a dataset to train the Ridge linear model, firstly, a *interpretable representation* of the dataset is computed. The dataset is then discretized (by default in quartiles) and samples z around the binary representation of x are drawn weighted by the proximity measure between an instance z and x in the binary representation. Generally, the proximity measure is defined as an exponential kernel $\Pi_x(z) = \exp(-D(x, z)^2/\tau^2)$ where τ is the kernel width and D the chosen distance. Let us remark that the kernel width defines the locality of the model. Finally, a Ridge regression model is trained on the generated perturbed data. To quantify the stability of the LIME explanations, in [17], the authors proposed the CSI metric, which measures the similarity between the coefficients in different repetitions of the LIME algorithm. Roughly speaking, for each feature, using a Gaussian distribution of the coefficients, 95% confidence intervals are computed. Then, the intersection of the confidence intervals is binary encoded and the value is normalized. Finally, the mean of all the values obtained is computed as a measure of concordance of the specific feature's coefficients among the different LIME repetitions.

3. Relation between metrics

The reliability of the Euclidean distance to capture the intuition of proximity in high-dimensional metrics spaces has been widely studied. For example, in [18] the authors explain how using traditional metrics $L_k(x, y) = \sum_{i=1}^n |x_i^k - y_i^k|^{1/k}$ in high-dimensionality problems leads to a loss of the notion of proximity. This is a major concern in problems where proximity plays an important role. LIME uses a metric to measure the distance between binary vectors whose components are 0's and 1's. In this context, the Hamming distance can be expressed as $\text{Hamming}(x, y) = \sum_{i=1}^n |x_i - y_i|$ which is exactly L_1 . In the case of binary vectors, the only difference between applying the Hamming or Euclidean distance is the square root, given that $1^2 = 1$ and $0^2 = 0$. Furthermore, the maximum in the Hamming distance between two binary sequences in \mathbb{R}^n is n , but the maximum distance between two binary sequences in \mathbb{R}^n in the Euclidean distance is \sqrt{n} , as one is the square root of the other. In the LIME implementation, the default value given to the kernel width is $0.75 \cdot \sqrt{n}$, which can be seen as 75% of the maximum

¹Along this paper, we refer exclusively to explanations of tabular data.

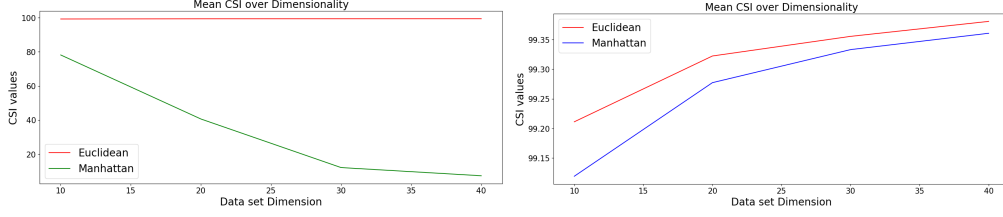


Figure 1: On the left, the Euclidean and the Manhattan distance CSI values are shown when using the predefined kernel width. On the right, both distance CSI values are shown when using the predefined kernel for the Euclidean distance and the proposed one for the Manhattan distance. The experiment was repeated 10 times and the results are the mean values.

value given by Euclidean distances over an n -dimensional space. Using the same idea, we define a custom kernel width for the Manhattan distance as $0.75 \cdot n$. Analogously, a kernel width comparable to the Euclidean metric performance (using the custom definition) for each L_k is given by $kw = 0.75 \cdot n^{1/k}$.

4. Experimentation

4.1. Kernel-width selection towards a fair comparison

In this experiment, we study the appropriate kernel width to be used depending on the metric. Firstly, different synthetic data sets composed of 500 samples of different dimensions (from 10 to 40 with a step of 10) were generated. Secondly, for each dimension, a Random Forest classifier was trained using 90% of the dataset as a training set. Thirdly, LIME was used to explain the remaining 10% of the data set (used as a test set) using all the attributes available. The *CSI* metric is computed for the coefficient stability [10]. The last step was computed for different choices of the kernel width. Specifically, for the Euclidean distance, we used the 75% of the maximum possible distance, and for the Manhattan distance, both the 75% of the maximum possible distance and the 75% of the square root of the maximum possible distance were applied. In Figure 1, the mean values of the CSI coefficients are provided for the different dimensions.

4.2. Comparison between the two distances

In this second experiment, we compare the explanations obtained using LIME using both distances. The explanations provided by LIME are an ordered list of the attributes based on its contribution. The dataset used is the UCI ML Breast Cancer Wisconsin (Diagnostic) dataset composed of 569 samples of dimension 30 for binary classification. The dataset was split into training and test set in a proportion 90 – 10 and a Random Forest classifier was trained. Then, explanations were computed for the remaining 10%, obtaining a vector of the 30 attributes ordered by importance. This was done using the predefined kernel width for the Euclidean distance, and both the predefined and the proposed kernel width for the Hamming distance. The vectors were compared as follows: Given two vectors x, y with length the number of

Distance	kernel width	CSI	
		Mean	Variance
Euclidean	$0.75 \times \sqrt{n}$	99.39	0.1
Manhattan	$0.75 \times \sqrt{n}$	9.48	8.53
Manhattan	$0.75 \cdot n$	99.35	0.12

kernel width	Max	Mean	Min
$0.75 \cdot \sqrt{n}$	23.93	8.69	0.28
$0.75 \cdot n$	1.37	0.19	0

Table 1

On the left, CSI values for the different choices of distances and kernel width. Higher values of CSI mean better stability. On the right, similarity measure between the order of importance of the attributes given by both distances for the two choices of the kernel width. The results depicted in the table are the mean values for the test set.

features in the dataset, for each feature we compute the difference between the coordinates of the feature in x and y . Finally, the mean of these differences is computed. This provides a similarity measure in the order of importance of the attributes. In Table 1 (left), the coefficients CSI are shown, depicting the stability reached by LIME using the different kernel widths. Let us remark that the conclusions of Experiment 1 are also achieved. We can see that the high values are reached for the Euclidean distance using the predefined kernel width and the Manhattan distance using the proposed one. In Table 1 (right), the similarity measure between the order of the features computed using LIME is provided. Therefore, it is shown that both distances have similar performance, providing similar relevance of the same features.

5. Conclusions

In the foundations of Machine Learning, it is well-known that the Euclidean distance loses the proximity notion for high dimensions, while the Hamming distance performs better in that context. However, this fact is not considered in the standard use of LIME and hence, undesirable explanations can be obtained. In this paper, we have studied the relationship between the Euclidean and the Hamming distances in this context, concluding its similarity. Experimentally, we have shown that the stability of the LIME algorithm converges and that the resultant order of importance of the features is similar when using the Euclidean and the Hamming distance if the kernel width is adapted to the chosen distance.

Acknowledgments

Partially supported by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement no. 101070028, and national projects PID2019-107339GB-I00 and TED2021-129438B-I00 funded by MCIN/AEI/ 10.13039/501100011033 and NextGenerationEU/PRTR. The content reflects the views of the authors only.

References

- [1] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), *IEEE Access* 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.
- [2] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115. URL: <https://doi.org/10.1016/j.inffus.2019.12.012>. doi:10.1016/j.inffus.2019.12.012.
- [3] Y.-N. Chuang, G. Wang, F. Yang, Z. Liu, X. Cai, M. Du, X. Hu, Efficient XAI techniques: A taxonomic survey, 2023. doi:10.48550/ARXIV.2302.03225.
- [4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Comput. Surv.* 51 (2019) 93:1–93:42. URL: <https://doi.org/10.1145/3236009>. doi:10.1145/3236009.
- [5] A. Jacovi, Trends in explainable AI (XAI) literature, 2023. doi:10.48550/ARXIV.2301.05433.
- [6] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, *Knowledge-Based Systems* 263 (2023) 110273. doi:10.1016/j.knosys.2023.110273.
- [7] C. Molnar, *Interpretable Machine Learning*, 2 ed., Independently published, 2022. URL: <https://christophm.github.io/interpretable-ml-book>, accessed on 2023-04-14.
- [8] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, 2018. doi:10.48550/ARXIV.1808.00033.
- [9] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, 2016. doi:10.48550/ARXIV.1602.04938.
- [10] G. Visani, E. Bagli, F. Chesani, OptiLIME: Optimized LIME explanations for diagnostic computer algorithms, *CoRR abs/2006.05714* (2020). URL: <https://arxiv.org/abs/2006.05714>. arXiv:2006.05714.
- [11] M. R. Zafar, N. M. Khan, Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems, 2019. doi:10.48550/ARXIV.1906.10263.
- [12] S. M. Shankaranarayana, D. Runje, Alime: Autoencoder based approach for local interpretability, 2019. doi:10.48550/ARXIV.1909.02437.
- [13] S. Shi, Y. Du, W. Fan, An extension of lime with improvement of interpretability and fidelity, 2020. doi:10.48550/ARXIV.2004.12277.
- [14] S. Mishra, B. L. Sturm, S. Dixon, Local interpretable model-agnostic explanations for music content analysis, in: S. J. Cunningham, Z. Duan, X. Hu, D. Turnbull (Eds.), *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, 2017, pp. 537–543. URL: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/216_Paper.pdf.
- [15] M. S. Kovalev, L. V. Utkin, E. M. Kasimov, Survlime: A method for explaining machine learning survival models, *Knowledge-Based Systems* 203 (2020) 106164. doi:10.1016/j.knosys.2020.106164.
- [16] D. Garreau, U. von Luxburg, Looking deeper into tabular lime, 2020. doi:10.48550/ARXIV.

2008.11092.

- [17] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, D. Capuzzo, Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models, *Journal of the Operational Research Society* 73 (2021) 91–101. URL: <https://doi.org/10.1080/01605682.2020.1865846>. doi:10.1080/01605682.2020.1865846.
- [18] C. C. Aggarwal, A. Hinneburg, D. A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: J. Van den Bussche, V. Vianu (Eds.), *Database Theory — ICDT 2001*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 420–434.