Intent-Aware Example-Based Explainability

Ikhtiyor Nematov

Université Libre de Bruxelles, Aalborg University
ikhtiyor.nematov@ulb.be

Abstract

In this PhD research, I aim to investigate the limitations inherent in current example-based explainability methods, develop solutions to address these deficiencies, and propose a novel direction for research that aligns example-based explainability with user needs and intent. The objectives of this research include: (a) Formalize existing evaluation metrics for local example-based explainability and propose new metrics to assess these methods from user perspectives; (b) Develop a framework for intent-aware local example-based explainability; and (c) Extend the framework for global example-based explainability and generate model-aware, faithful explanations.

Introduction

The failure of ML-based systems in numerous instances due to data errors, biases, and misalignment (Osoba and Welser IV 2017; Tashea 2017; Seymour et al. 2023; Burema et al. 2023) has driven researchers to develop explainability techniques. Various taxonomies for these methods exist, but a common classification is based on the type of explanation generated (Molnar 2022): *Model-based* methods involve creating interpretable surrogate models, such as linear models, to approximate the complex black-box ML model (Ribeiro, Singh, and Guestrin 2018; Silva et al. 2020). *Feature-based* methods focus on identifying important input features, such as words in text or parts of an image, that contribute the most to the prediction (Ribeiro, Singh, and Guestrin 2016; Fong, Patrick, and Vedaldi 2019; Ancona 2017). *Example-based* methods provide explanations for a specific target outcome by determining the importance of training samples (Ilyas et al. 2022; Garima et al. 2020; Koh and Liang 2017; Ghorbani and Zou 2019; Park et al. 2023), or offer a global overview of the model by identifying representative examples (Yeh et al. 2018; Kim, Khanna, and Koyejo 2016; Gurumoorthy et al. 2019).

Example-based explainability has several advantages. It is typically model-agnostic and provides easily understandable explanations. More importantly, by seeking to uncover causal relationships between training examples and model behavior, it can assist in model debugging and data cleansing (Hara, Nitanda, and Maehara 2019), and it is flexible for practitioners with varying levels of access to the model or data (Richardson et al. 2023).

In my research, I investigate both local and global example-based explainability, attempt to solve existing issues, and propose new visions. Firstly in local explainability, from a technical perspective, existing methods heavily depend on data quality, making their performance vulnerable to class outliers. These data points are wrongly labeled or exhibit ambiguity, making it hard for the model to generalize. Though irrelevant, they exhibit global influence and eventually compromise the explanations quality and relevance. Although some works tried to address this issue (et al. 2020) using loss-based elimination, we have demonstrated (Nematov et al. 2024b) that their solution is suboptimal. From a user perspective, they lack heterogeneity in their explanations, often addressing only the "why" (e.g., why it is a cat) and missing the "why not" (e.g. why it's not a dog). Moreover, the main goal of explainability is to help users or practitioners understand the model, yet existing methods do not consider the user *intent* when requesting an explanation. Users might have various intentions when seeking a local explanation. For instance, when the prediction is wrong (dog) the intent might be to investigate why the model made a mistake, it's preferred to observe both the data that was positively affecting and "pushing" the model to predict "dog" and the data that was negatively affecting and "pushing" the model to predict "cat". Typically, existing methods provide a single type explanation for all cases, which may not always be sufficient.

Global example-based explainability is less studied and all existing methods provide different forms of prototypes or representative samples. Prototypes can be real or imaginary samples. However, they use data itself and the ground truth while generating these prototypes (Kim, Khanna, and Koyejo 2016; Gurumoorthy et al. 2019), thus their explanation is not faithful to the model. Another use case of prototypes is when they serve as an inherently explainable model on their own, nearest prototype neighbor (Molnar 2022).

Research Objectives. In this PhD research, I aim to investigate the limitations inherent in current example-based explainability methods, develop solutions to address these deficiencies and propose a novel direction for research that aligns example-based explainability with user needs and in-

tent. The objectives of this research include:

1. Formalize existing evaluation metrics for local example-based explainability and propose new metrics to assess these methods from user perspectives.
2. Develop a framework for intent-aware local example-based explainability.
3. Extend the framework for global example-based explainability and generate model-aware, faithful explanations.

Local Example-based Explainability

Evaluation Metrics

At the beginning of the PhD I did extensive research on metrics to quantitatively evaluate explainability. It appears that for example-based methods there is a lack of such metrics. Thus, in my first paper (Nematov et al. 2024b), I formalized some general existing metrics such as relevance, and correctness, and proposed a new aspect, distinguishability which can be estimated using some metrics. In the following, I will introduce these metrics and discuss the results and observations of the paper. Let's call the sample to be explained the *explanandum*, the example-based explainer provides an *explanation* which consists of *training samples*.

Explainer Relevance is defined as the average similarity of explanandum to each training sample in its explanation. An explanation is *relevant* if it closely matches the explanandum. **Explainer correctness** is a notion from a controlled data check experiment where some rule is employed to label some chunk of the training data and then once the model has properly learned the rule while explaining the follower of that rule from unseen data the explainer must provide the training data that also followed the rule. **Explainer distinguishability** measures how specific an explainer's explanation is to the explanandum. One way to measure it is *explanation overlap* which is an expected Jaccard similarity between the explanations of two random explanandums. Another way is *actuve domain* which is the number of distinct training samples used by the explainer, normalized by the size of the training set. *Example popularity* is the probability of a training sample being chosen in the explanation. Higher distinguishability means lower overlap, bigger active domain, and lower popularity.

Results. Our analysis suggests that current example-based explainability techniques are susceptible to the influence of class outliers and thus perform poorly on the abovementioned metrics. Outright removal of outliers from explanations similar to (et al. 2020) may not always be advisable, as they can possess valuable insights and provide explanations for similar examples encountered by models.

Intent-aware explanations

In my second paper (Nematov et al. 2024a), I proposed a novel example-based explainability method, AIDE, that customizes its explanation according to user intent and provides diverse explanations from different angles. It also addresses the class outlier problem discussed in the previous section. First, AIDE uses proximity to filter explanation, considering influential samples, that are too distinct semantically from the point being explained, as outliers. It then selects four diverse quadrants as shown in Figure 1.

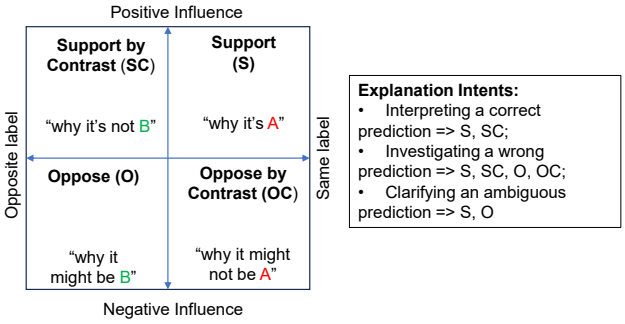

Figure 1: Intent-aware explanation with quadrants.

Support examples have a positive influence and the same label as the test instance. They positively affect the prediction. **Support by Contrast** examples have a positive influence but a different label. They explain the prediction by contrasting with similar examples of the opposite class. **Oppose** examples have a negative influence and different labels. They argue the test instance belongs to their class. **Oppose by Contrast** examples have a negative influence but the same label as the test instance. They argue the test instance does not fit the predicted class.

AIDE distinguishes 3 types of intent for the local explanation and selects specific quadrants for each: **1. Interpreting a correct prediction**: the user knows the prediction is correct and seeks to understand the model's reasoning. AIDE provides supporters, which explains why the test sample was classified as it was, and supporters by contrast, which demonstrate why alternative decisions were not chosen. **2. Investigating a wrong prediction**: the goal of the explanation is to understand the cause of an error, which may be due to mislabeled training samples or bias in the training data. AIDE helps track both types of errors. In the case of *mislabeled samples*, supporters are examined to identify potential errors, while opposers explain why the opposite label might be more suitable. For *bias in the training data*, where the model learns an extrinsic feature (e.g., associating snow with wolves (Besse et al. 2018)), AIDE examines all quadrants. If bias is present, it will appear in supporters but not in supporters by contrast. Opposers will negatively impact the prediction if they possess the biased feature, and the feature may also appear in opposers by contrast. **3. Clarifying the prediction for ambiguous samples**, AIDE reveals the labeling mechanism. For example, if an image contains both objects being classified, the classification rule influences the model. If the ground truth is correct, the model has learned this mechanism. AIDE identifies relevant, similarly ambiguous training samples that positively impact the prediction, acting as supporters.

Results. The quantitative evaluation demonstrated that our approach effectively addressed the outliers problem and outperformed state-of-the-art methods in correctness and continuity metrics. We also conducted a user study to assess

whether providing multiple quadrants improved understanding of predictions compared to only offering influential samples as in SOTA methods. Additionally, we explored how well the explanations met user intent. The results showed that 100% of users agreed that the explanations were relevant to their intent, and the same proportion found that providing contrastive examples helped see the model's decision boundary.

Global Example-Based Expalainability

After an extensive literature review on global example-based explainability, I am developing a method to generate prototypes that are faithful to the model, meaning they explain the model's predictions rather than just the data distribution. Our approach uses *local explanations* and *model predictions* to generate prototypes. The framework builds a *bipartite explanation graph* where the explanandum on one side is connected to training samples in its explanation. Since explanandums can have common training samples in their explanation, we utilize the adjacency matrix together with the semantic similarity of explanandum to find prototypes among them. We compare the faithfulness of these prototypes to the model against the existing prototype methods such as (Ilyas et al. 2022; Kim, Khanna, and Koyejo 2016). Once selected, these prototypes can be used as global explainers but also they can partition the data using the nearest prototype method, followed by cluster explanations. Cluster explanation can be selected by maximizing the coverage of local explanation of cluster members. In other words, we approximate the local explanation of all cluster members with a cluster explanation.

References

Ancona. 2017. A unified view of gradient-based attribution methods for Deep Neural Networks. In *Workshop on Interpreting, Explaining and Visualizing Deep Learning*. NIPS.

Besse, P.; Castets-Renard, C.; Garivier, A.; and Loubes, J.-M. 2018. Can Everyday AI be Ethical? Machine Learning Algorithm Fairness (english version).

Burema, D.; Debowski-Weimann, N.; von Janowski, A.; Grabowski, J.; Maftei, M.; Jacobs, M.; van der Smagt, P.; and Benbouzid, D. 2023. A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context. AIES '23, 705–714. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

et al., B. 2020. RelatIF: Identifying Explanatory Training Examples via Relative Influence. *PMLR*.

Fong, R.; Patrick, M.; and Vedaldi, A. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2950–2958.

Garima; Liu, F.; Kale, S.; and Sundararajan, M. 2020. Estimating training data influence by tracing gradient descent. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, 2242–2251. PMLR.

Gurumoorthy, K. S.; Dhurandhar, A.; Cecchi, G.; and Aggarwal, C. 2019. Efficient Data Representation by Selecting Prototypes with Importance Weights. In *2019 IEEE International Conference on Data Mining (ICDM)*, 260–269.

Hara, S.; Nitanda, A.; and Maehara, T. 2019. Data Cleansing for Models Trained with SGD. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 4215–4224.

Ilyas, A.; Park, S. M.; Engstrom, L.; Leclerc, G.; and Madry, A. 2022. Datamodels: Understanding Predictions with Data and Data with Predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 9525–9587. PMLR.

Kim, B.; Khanna, R.; and Koyejo, O. O. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1885–1894. PMLR.

Molnar, C. 2022. *Interpretable Machine Learning*. 2 edition.

Nematov, I.; Sacharidis, D.; Sagi, T.; and Hose, K. 2024a. AIDE: Antithetical, Intent-based, and Diverse Example-Based Explanations. arXiv:2407.16010.

Nematov, I.; Sacharidis, D.; Sagi, T.; and Hose, K. 2024b. The Susceptibility of Example-Based Explainability Methods to Class Outliers. arXiv:2407.20678.

Osoba, O. A.; and Welser IV, W. 2017. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

Park, S. M.; Georgiev, K.; Ilyas, A.; Leclerc, G.; and Madry, A. 2023. TRAK: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Richardson, B.; Sattigeri, P.; Wei, D.; Ramamurthy, K. N.; Varshney, K.; Dhurandhar, A.; and Gilbert, J. E. 2023. Add-Remove-or-Relabel: Practitioner-Friendly Bias Mitigation

via Influential Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 736–752. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Seymour, W.; Zhan, X.; Coté, M.; and Such, J. 2023. A Systematic Review of Ethical Concerns with Voice Assistants. AIES '23, 131–145. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Silva, J. M.; Gerspacher, T.; Cooper, M.; Ignatiev, A.; and Narodytska, N. 2020. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, volume 33, 1–11.

Tashea, J. 2017. Courts are using AI to sentence criminals. that must stop now.

Yeh, C.-K.; Kim, J.; Yen, I. E.-H.; and Ravikumar, P. K. 2018. Representer Point Selection for Explaining Deep Neural Networks. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.