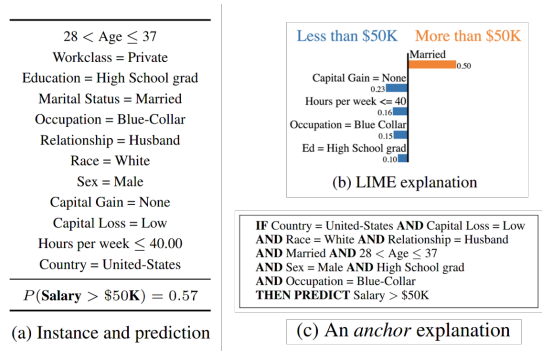
Quest: A Query-driven Explanation Framework for Black-Box Classifiers on Tabular Data

Nadja Geisler

Technical University of Darmstadt (TU Darmstadt), Department of Computer Science, Hochschulstraße 10, 64289 Darmstadt, Germany

Keywords
XAI, post-hoc explanation, model-agnostic, classification, black box



(a) Instance and prediction (b) LIME explanation (c) An *anchor* explanation

Figure 1: LIME/Anchors explanations near the decision boundary in the UCI adult dataset (from [1]).

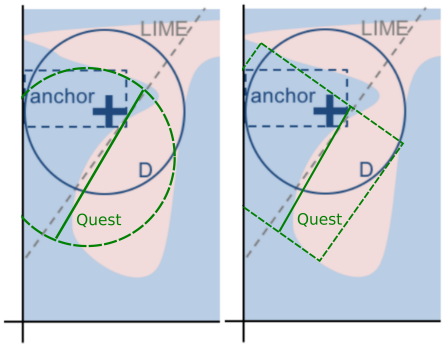


Figure 2: Toy examples of areas for LIME, Anchors & `Quest` near the decision boundary (adapted from [1])

Explainability efforts are well established in the ML and AI communities by now, with local, model-agnostic approaches currently being the tool of choice in information retrieval [2] and search [3] as well as many other areas. One major challenge in the field is the lack of sophisticated approaches for tabular/relational data, as opposed to text or images. Generic approaches, e.g. feature importance, limit expressiveness and readability.

LIME [4] still remains the basis of many approaches for local, model-agnostic explanations. It was adapted to tabular data to serve as a baseline for Anchors [1]. Both approaches use a white-box model (surrogate) to approximate a black-box model locally in order to explain its decisions. However, their resulting explanations are limited: LIME (as implemented by the authors) focuses on feature importance. Anchors are derived as simple predicates to create if-then rules. Figure 1 shows examples for the UCI adult data set [5].

As an alternative we suggest `Quest`, a framework for query-driven post-hoc explanations of individual classifier decisions on tabular data.

Query-driven explanation First, we introduce a more expressive representation of explanations consisting of query predicates. A custom set of common query predicates is extremely expressive while still compact. Queries such as `capital-gain > capital-loss` or `rel == 'married' AND children > 1` also have the benefit of being easily converted into the WHERE clause of a SQL statement, to be executed directly on any relational database. Still more importantly, using queries to explain black-box model behavior within local boundaries has the advantage of explaining not only why a model produced an outcome but also why not!

An explanation produced by `Quest` can be thought of as boundaries and a decision surface that separates classes within the boundaries. Samples on one side of the decision surface within the boundaries form the result set of a query Q. Alongside Q ("Why?") stands $\overline{Q}$ ("Why not?") with its result set covering samples on the opposite side of the decision surface. This consideration of queries as explanations gives the user an intuitive way of thinking about the local neighborhood, supports generalization on the user's side and keeps the focus on the data instead of the surrogate model. The combination of

DESIRES 2021 – 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy
nadja.geisler@cs.tu-darmstadt.de (N. Geisler)
https://www.dm.tu-darmstadt.de/ (N. Geisler)
0000-0002-5245-6718 (N. Geisler)

Q and $\overline{Q}$ ensures an explanation from both sides of the decision surface, something established systems for this task lack. Query representation can be limited to a small number of operators without loss of expressiveness and normalized to facilitate comparison and elimination of duplicates. This is achieved through application-specific, user-defined functions for complex relationships and reducing logical operators while retaining functional completeness. An example would be the disjunctive normal form (DNF) using only AND, OR, and NOT operators.

We impose a complexity budget on $Q/\overline{Q}$ to ensure users are able to understand them well. This can be adapted to the target group. Complexity of queries can be thought of as the number and type of predicates making it easily computable and comparable.

Framework approach We now need to determine $Q/\overline{Q}$ for a given data point such that it best explains the behavior of the black-box model in the local neighborhood within a fix complexity budget. To leverage the flexibility of the query language, a very complex approach of generating queries suitable to a wide range of scenarios (numerical/categorical attributes, data distributions, sparsity, dependencies between attribute values, noise level, ...) would be necessary. A framework approach gives us the opportunity to be flexible and extensible but still conceptionally straight forward.

We suggest a selection mechanism over several explanation classes to minimize drawbacks of individual approaches and produce a good fit for the input data. Classes not applicable to the given data/scenario can be eliminated immediately while the further process is essentially a hyper-parameter search, covering the decision between explanation classes as well as their respective parameters. Strategies like pruning or successive halving can be applied after starting with several instances of applicable explanation classes for the given data point.

We propose three exemplary classes that vary in expressiveness (i.e., complexity and form of the representations they produce) as well as other properties:

Decision trees make robust candidates that can be applied to numerical and categorical attributes, can capture disjoint areas, and work with any condition type.

Adaptations of clustering techniques, using a suitable distance metric and constraints on labels intuitively fit the task of grouping instances.

A linear model on a reduced feature set could be used for local relations between attributes.

Each class could produce a different "form" of neighborhood, different from the rigidity of LIME/Anchors. Figure 2 shows a toy example for LIME, Anchors and `Quest` for a point near the decision boundary. Note that `Quest` boundaries (green, dashed) differ, left being distance-based, right linear. The decision boundary is linear with one attribute depending on the other in both cases. The explanation classes allow for endless extensibility, as quality metrics used for the selection process are defined on the query representation they all share. Within explanation classes, we suggest imposing a hard complexity restraint (that could be adapted to context upfront) and then optimizing for accuracy.

We compare explanations (candidates produced by `Quest` and baselines such as LIME/Anchors) regarding accuracy, coverage (area and/or proportion of original samples) and class balance within an explanation.

Acknowledgments

This research and development project is/was funded by the German Federal Ministry of Education and Research (BMBF) within the "The Future of Value Creation – Research on Production, Services and Work" program (funding number 02L19C150) and managed by the Project Management Agency Karlsruhe (PTKA). The author is responsible for the content of this publication.

References

[1] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018, pp. 1527–1535.

[2] M. Verma, D. Ganguly, Lirme: Locally interpretable ranking model explanation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1281–1284. URL: https://doi.org/10.1145/3331184.3331377. doi:10.1145/3331184.3331377.

[3] J. Singh, A. Anand, Exs: Explainable search using local model agnostic interpretability, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 770–773. URL: https://doi.org/10.1145/3289600.3290620. doi:10.1145/3289600.3290620.

[4] M. T. Ribeiro, S. Singh, C. Guestrin, "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, 2016, pp. 1135–1144.

[5] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.