Personalized Explanations

Maximilian Becker

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
maximilian.becker@kit.edu

Abstract

Machine learning systems are often hard to investigate and intransparent in their decision making . Explainable Artificial Intelligence (XAI) tries to make these systems more transparent. However, most work in the field focuses on technical aspects like maximizing metrics. The human aspects of explainability are often neglected. In this work, we present personalized explanations, which instead focus on the user. Personalized explanations can be adapted to individual users to be as useful and relevant as possible. They can be interacted with to give users the ability to engage in an explanatory dialog with the system. Finally, they should also protect user data to increase the trust in the explanation system.

1 Introduction

Artificial intelligence and machine learning have become extremely popular technologies that are widely used because of their many advantages. However, learned models like neural networks also have some major disadvantages, especially their lack in transparency. During training, the models learn correlations from the training data that enable them to make predictions on unseen data and take decisions. What exactly a model learned, what they pay attention to and how they make decisions is however hard to comprehend. The field

of explainable artificial intelligence (XAI) tackles this problem and tries to make learning systems more transparent. The goal is to create explanations for systems that help users or operators to better understand the models and their inner workings[2].

However, much of the research in the field is very technical, neglects the human aspect of explainability and only relies on researchers intuitions. Many works concentrate on technical aspects and try to maximize metrics that are not validated by user studies or grounded in psychology. Papers focused on users are mostly focused on user interfaces and not on the underlying algorithms[12, 10, 1].

In this work we present ideas for XAI methods that are better adapted to users to make them more useful and relevant. The first aspect is that methods should be individualized to the user to make them more helpful. Users should have the ability to customize an explanation in order to adapt it to their needs. Existing methods of interaction and individualization are presented in Section 2. A new approach for individualized and interactive explanations will be discussed in Section 3. Section 3.1 explains the approach and focuses on the individualization. However, users have to be able to interact with the system to get explanations they understand and are relevant for them. Methods of interaction with the new approach are shown in Section 3.2. Individualized explanations do however require personal data of the user in order to personalize the explanations. In order to build trust with the system, the user data has to be protected. Explanations can also help users to understand what kind of data is needed for a system to function properly. Concepts for data protection and data minimization will be shown in Section 4.

2 Background

There are different existing approaches in the literature to interact with XAI systems. The first one is to give the user the option to generate multiple explanations[15]. By generating multiple explanations the user gets different view points and has a better chance of understanding them. This can be done by generating multiple explanations of the same type or explanations of

different types. The user can also get the option to change input data[5]. By changing the instances that are explained by the system the user can get a better overview over the feature space and the behavior of the system. Other works evaluate the interaction with graphical representations or user interfaces[6]. They investigate how different visualizations or interfaces help users to understand the explanations. Another way to interact with an explanation system is through an interface using natural language processing[4]. This way the user can use natural language to interact with the system which makes it much more suited for end users with little technical knowledge. All these approaches leave the explanation system itself untouched and only build different user experiences around them. By interacting with an explanation system, the explanations will also be individualized on a basic level. However, explanations can also be individualized explicitly. DiCE[14] can generate counterfactual explanations that are diverse, meaning that different explanation instances are different from each other. The method can also be used to set feature constraints that are used to ensure feasibility of the explanations but can also be used to adapt explanations to individual users.

3 Individualized Explanations

Individualized explanations should be adaptable to the use case as well as the individual user. The explanations can be adapted by an admin or professional user or the end user of the system itself. Different aspects can be considered when adapting explanations. One aspect is general knowledge or world knowledge as well as knowledge about the use case in which a system is deployed. In a medical use case for example, other aspects are relevant compared to a financial use case. Different features are important in different contexts and different applications so the explanations have to reflect that. Explanations should also be adapted to the user group. Different user groups have different abilities and knowledge levels in machine learning and the application domain. Machine learning experts, domain experts and end users have different capabilities and require different explanations. However, the explanations should not only be adapted to the user group but also to the individual user itself. To achieve this,

different sources of background knowledge can be used (see Section 3.1) or the user can be given the ability to customize the explanations by herself (see Section 3.2).

3.1 Personalized Counterfactuals

Counterfactual explanations are a form of explanation for machine learning systems. A counterfactual describes an alternative state in which some changes were made that lead to a different outcome. In machine learning, the factual is an instance whose prediction from a model should be explained. The counterfactual is an instance with small changes that lead to a different classification. For example, if a credit application is rejected a counterfactual explanation could tell that the application would have been accepted if the credit amount was lowered by a certain amount[13].

Counterfactuals are a local explanation method, which means that they explain a single data point or decision of a model in contrast to global explanations which explain the behavior of the whole model. They are calculated by searching for the closest instance from the one that should be explained that changes the prediction of the classifier. This can be done by random sampling[3], using a gradient[9], formulating the problem as an optimization problem[16] or with genetic algorithms[14]. Counterfactuals originate from counterfactual thinking which people engage in regularly[12]. Thus people are already used to the concept which makes these explanations especially user friendly and suited for non technical end users.

It is however not obvious how different features with different value ranges and units should be treated when comparing counterfactual explanations. It is for example not possible to objectively compare what change in the credit amount equals to what change in the credit length. The idea behind personalized counterfactuals is to have a weighted distance metric to calculate the distance between a factual and different counterfactuals. The weights can be chosen by the user to represent her preferences. If for example the weight for the credit duration is low and the weight for the credit amount is high, a change in the credit amount will be penalized more and a change in the credit duration will be

preferred. This means that by changing the weights of the distance function the users can adapt the counterfactual explanations to their needs or preferences.

Besides the weighting of features, users have some more ways to personalize the explanations. Some features are unchangeable like a persons place of birth. Users can tell the system to ignore such features in the explanations. Other options would be to restrict the value range in which a feature can be changed or make it only changeable in one direction, like an age which can only get larger. If it is a multi-class classification problem the user can also specify the class the counterfactuals should be classified as.

In addition to options for single features, users also have the option to set global metrics to diversify the different counterfactual explanations shown to the user. These global metrics evaluate a set of different counterfactuals. By changing them, the user can get multiple similar explanations or more different ones. Users can also adjust a weight that measures in how many columns changes were made. With these global metrics, the user can configure the overall set of different counterfactual explanations

All these settings can be adjusted by an administrator to represent world knowledge or adapt the explanations to a specific use case. The administrator will set these options once for a specific application. In a second step, the end user can adjust all settings or a smaller subset in the application itself. This is done to personalize the explanations to the specific user.

After all the weights and metrics are set, the search for personalized counterfactuals is done with an evolutionary algorithm. Features of the original instance are randomly changed, excluding the ignored features. The new instances generated in this way are passed to the model to check if they are counterfactuals or if they are classified as the wanted class. Then, the distance from the original instance is measured by a weighted Euclidean distance using the previously specified weights. The set of instances is also evaluated by the global metrics. The best instances are chosen and changed again. This process is iterated until the rating by the metrics does not change anymore. This approach is completely model agnostic because it uses no internal information about the model that should be explained, like gradients. Only predictions of the model are used to check if instances are counterfactuals or if they are classified as the wanted class.

3.2 Interactive Explanations

An important aspect to adapt explanations to users is to enable users to interact with the system. This way, users can influence explanations and customize them to their needs, be it their knowledge level in a certain domain or their technical expertise. Explanations between humans are also often given in a form of dialog. The explainee can ask about things she does not understand or for details on the given explanation. The explainer can then give additional information, reformulate the given explanation or come up with a new one. Interaction and individualization go hand in hand because users are only able to individualize their explanations if they are able to interact with the system and users that can interact with an explanation system will try to get a better understanding by adapting the explanations to their needs. The personalized counterfactual explanations can also be interacted with in several ways. The first way is to change the weights for different features. This influences how much a feature is changed in the generated counterfactuals. Users can also exclude features from the search by marking them as unchangeable. This helps to only generate satisfiable explanations and not ones that are impossible or unrealistic by for example demanding to change a persons race. The next way to interact with the system is to adjust the global metrics that compare the set of generated counterfactuals. With these metrics, the users can get more or less diverse explanations and influence how sparse the explanations are, meaning how many features are changed. The final way is by changing the target class. With this setting, users can tell the system to generate explanations for a specific target class. By looking for counterfactuals with a specific target class, the user can see what changes to an instance are needed to reach the desired class.

4 Protection of Personal Data

The previous Sections showed how explanations can be personalized and interacted with to better meet the user's needs. However, personalization also has a drawback: it requires personal data from the users. Without some form of data about the user of an explanation system it is not possible to adapt explanations to the user. Personal data underlies the European Union's General

Data Protection Regulation (GDPR)[8] as well as the Artificial Intelligence Act[7]. These regulations demand the protection of personal data. One way to protect user data in any application is by the use of trusted computing methods. An approach to this is shown in section 4.1. Explanations can also be used to show users what influence sharing or not sharing some data has on a system. Users can see how system behavior changes with their decision and find a configuration that works for them. With such explanations users can make informed decisions on what data they want to share with a system and what data they want to keep private. This way users are able to minimize the data they have to share with a system. The idea for using XAI to explain the effect of sharing data is shown in Section 4.2.

4.1 Explainable AI and Trusted Computing

Users may be uncomfortable with sharing their data with a system that they do not understand and cannot trust. XAI can explain a systems behavior to a user but to get relevant explanations users often have to share their data first in order to get explanations that are relevant to their situation. In order to keep personal data safe and contribute to increasing the trust in the system trusted computing methods can be used. Two trusted computing technologies that are useful for this are Trusted Platform Modules (TPMs) or Trusted Execution Environments (TEEs). TPMs are trusted hardware modules that can verify the state of a system. TEEs are a separate part of the processor that enables secure data processing and is not accessible even by the operating system. These methods can be used to secure an explanation system and make it more trustworthy either by verifying that the system is in a trustworthy state with TPMs or by executing code on TEEs. The combination of these technologies can create trust in the system through trusted computing methods and trust in the underlying machine learning model through XAI.

4.2 Explainable AI for Data Minimization

Some applications require different kinds of data from users. For example, a health app may be interested in health data, location data and general information

about the user. However, users may not want to share all this data with an application. In order to let users make an informed decision about the data they want to share, they have to know how the system behavior changes if they decide to keep some of the data private. Here, we will present a concept on how XAI can be used to generate such explanations.

The idea is to use a combination of SHAP[11] and counterfactual explanations. SHAP is a XAI method building on Shapley values. It calculates a feature importance by omitting features from an instance and replacing them with values from random instances from the training set. The predictions of the model with the random feature values are averaged and compared to the result of the original instance. This way, the influence of the feature value on the original instance can be calculated. This idea can be combined with counterfactual explanations explained in Section 3.1. The combination of the two methods should be able to explain users how not sharing some of their data would influence the system behavior.

5 Summary

In this work, we presented some ideas for making explanations for AI systems more relevant to users. At first, methods for individualizing explanations were shown that make it possible to adapt explanations to individual users. Afterwards, existing principles of interaction with explanation systems were presented and it was shown how users can interact with personalized explanations. Interaction and individualization are interrelated because users have to interact with a system in order to individualize an explanation. Methods for protecting user data in the explanation process through trusted computing were shown. At the end, an idea on how to minimize the data a user has to share by providing explanations was presented.

References

[1] Amina Adadi and Mohammed Berrada. “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”. In: *IEEE access* 6 (2018), pp. 52138–52160.

[2] Nadia Burkart and Marco F Huber. “A survey on the explainability of supervised machine learning”. In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[3] Alberto Carlevaro et al. “Counterfactual Building and Evaluation via eXplainable Support Vector Data Description”. In: *IEEE Access* 10 (2022), pp. 60849–60861.

[4] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. “Crowd-based personalized natural language explanations for recommendations”. In: *Proceedings of the 10th ACM conference on recommender systems*. 2016, pp. 175–182.

[5] Michael Chromik. “reshape: A framework for interactive explanations in xai based on shap”. In: *Proceedings of 18th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET). 2020.

[6] Michael Chromik and Andreas Butz. “Human-XAI interaction: a review and design principles for explanation user interfaces”. In: *IFIP Conference on Human-Computer Interaction*. Springer. 2021, pp. 619–640.

[7] European Commission. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. URL: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206`.

[8] European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General*

Data Protection Regulation). URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[9] Chihcheng Hsieh, Catarina Moreira, and Chun Ouyang. "Dice4el: interpreting process predictions using a milestone-aware counterfactual approach". In: *2021 3rd International Conference on Process Mining (ICPM)*. IEEE. 2021, pp. 88–95.

[10] Mark T Keane et al. "If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques". In: *arXiv preprint arXiv:2103.01035* (2021).

[11] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[12] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[13] Christoph Molnar. *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. Lulu. com, 2020. URL: https://christophm.github.io/interpretable-ml-book/.

[14] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 607–617.

[15] Chelsea M Myers et al. "Revealing neural network bias to non-experts through interactive counterfactual examples". In: *arXiv:2001.02271* (2020).

[16] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.