Interpretable Learning for Regression Using Causal Attention and Counterfactual Explanation

Projectplan

Kathrin Khadra

February 4, 2021

1 Literature Review

In the following section we will be exploring related work regarding this thesis. Generally, the methods presented are either model agnostic and able to perform on regression problems or address classification models. The properties of each paper can be found in Table 1. As we will be addressing regression Machine Learning models, we will be first looking at model-agnostic and regression competent methods.

1.1 Model-Agnostic or Regression Competent Methods

First in [PBM16], a causal model is described as a model which remains invariant under interventions (different changes of the environment). This definition is used to identify causal models under multiple other models via different experimental settings. The causal models will be identified by it's property of showing invariance in it's predicitive accuarcy during the experiments (experiments are different interventions for example). Thus, the method gives a confidence interval for the causality of the model. The results show that especially for Gaussian structural equation models the method can guarantee identifiability for the set of causal models. Furthermore, in [ZB18], counterfactual explanation is used to build fair decision-making systems. The measurements counterfactual direct (Ctf-DE), indirect (Ctf-IE), and spurious(Ctf-SE) effects are developed to eludicate the contrast of different decisions made by various discriminatory mechanisms. This is done by measuring transmission of change from stimulus to effect within the network using Ctf-DE, Ctf-IE and Ctf-SE. To test the measurements, different fairness constraints are defined and simulations, regarding detection, optimization and evaluation of decision making, on different discrimination analysis tasks are executed. In sum, the authors analyze the trade of between outcome as well as procedural fairness criteria and provide a measurable approach to build a fair decision making system. Both methods, as one can see in Table 1, address regression problems but do not concentrate on machine learning applications. Within this work, we will be addressing regression problems as well looking at using similar methods like counterfactual explanation.

Applicable for machine learning and model agnostic approaches will be discussed in the following paragraph. The framework LIME is build in [RSG16]. It is able to interpret the predictions of any classifier. That is done by training an interpretable model locally around the prediction. It is shown that interpretability is useful for multiple of models in trust-related tasks. (in text and image domains as well as both expert and non-expert users). In [SHG19], and approach using Counterfactual Explanation (CERTIFAI) is developed to test any machine learning model regrading robustness, transparency, fairness and interpretability. The method is model-agnostic and can be used on any type of input data. The paper introduces a genetic algorithm to generate the counterfactuals, and CERScore is developed to quantify the robustness of the tested machine learning model using the generated counterfactuals. The genetic algorithm brings flexibility to the model, through generating custom counterfactuals, and additionally CERScore is able to perform equivalently to similar methods. A model-agnostic causal explanation model (CXPlain), to estimate the impact of certain inputs of a machine learning model on the respective outputs, is derived in [SK19]. Here, Bootstrap ensembling is used to calculate the uncertainty regarding the feature importance of the model. This way a model with high-dimensional data can be made explainable by estimating the importance of each feature of the model. CXPlain's performance shows to be more accurate and faster than comparable model-agnostic algorithms in the field of estimating feature

importance. As we will be dealing with a regression machine learning problem, we can learn from the approaches above.

However, the papers mentioned so far deal with similar problems but the approaches they take are very different to our proposal. In the following paper [Caw08], as one can see in table 1, a regression problem is investigated by opening the black box. Hence looking at the inner state of the model. Causal and non-causal feature selections for regression problems is investigated. Here, causal feature selection uses inference of the Markov Blanket, direct causes and direct effects. The non-causal feature selection on the other hand is derived from logistic regression with Laplace prio based Bayesian regularisation. With those methods linear classifier, with a difference in the causal relationship of input and response variable between the training and test data, are evaluated. The causal feature selection method showed no significant enhancement regarding the predictive accuracy of the classifiers over a non-causal feature selection and/or using all features provided. This approach only accounts for classifiers and in contrast we will be looking at the internal state of our regression machine learning model.

Model-agnostic machine learning models, which are analyzed by opening the black box are discussed by [Boz02] and [STY16]. One way of making machine learning models explainable is to extract rules or interpretable models from them. In [Boz02], the authors extract C4.5 like Decision Trees from Neural Nets, in order to make the Neural Nets explainable. The high fidelity trees are extracted by a method, called DecText. Furthermore, DecText is also able to work with continuous features. As the current work only deals with MofN type Decision Trees, which are only suitable for MofN problems, the authors provide an alternative method for regular high dimensional real world problems. Another method to measure feature importance in machine learning is presented in [STY16]. Counterfactuals are generated by scaling down the original inputs and the counterfactual's gradient, called interior, is analyzed to assess to measure the importance of the respective feature. The algorithm is applied to an LSTM language model, GoogleNet for image object recoginition and a ligand-based virtual screening network with categorical features. With this method the authors where able to calculate interior gradients as easily as standard gradients. Furthermore, the algorithm is applicable to various DNNs and the feature importance yield to the prediction score. Although these approaches are model agnostic, we will be looking at the internal state of our machine learning model as well.

One goal for this thesis is to develop a model based method. This means that the algorithm should construct a causal neural net instead of analyzing an already trained model regarding its causality. Constructing a causal machine learning model solving a regression problem is discussed in the following. Using the classical statistical technique Cook's distance in large scale datasets to evaluate the influence of training samples in regression models, has been done in [WCZ$^+$16]. This means, that samples can be identified which have an extraordinary strong influence on the training and prediction of the model. In order to use Cook's distance on a large and high-dimensional data set the method influence sketching is introduced. Influence sketching applies random projections within the influence computations and is tested on a malware detection dataset (over 2 million executable files with almost 100,000 features). The authors showed that the impactful samples are very likely to be mislabeled. Additionally, the results show that deleting the samples, which where identified as highly influential by the method Influence sketching, brings the accuracy down to 90.24% from 99.47%. Deleting the same number of random samples only brings the accuracy down to 99.45%.

Overall looking at Table 1, one can see that existing papers mostly concentrate on analysing the causality of regression and model-agnostic classifier or machine learning models, after they have been trained already. However, some of them open the black box and look at the internal state of

the model. In contrast, the paper which constructs a causal model, rather than testing it after the training, does not analyze the internal state of that model. In our work, we are looking at bringing both together, which means to construct a causal machine learning model dealing with regression looking at the internal state of the model.

1.2 Classification Competent Methods

Apart from regression competent causality methods most of the literature concentrates on classification tasks so far. We will be looking at the most successful classification approaches that deal with machine learning in the following. In regards to Counterfactual Explanation, [KL17] have an interesting approach we can build on. Here, an influence function is used to compute the neural net input's effect on the output of the neural net. In this case linear regression and a CNN using image data are tested. The image data is altered, according to the influence function's parameter ϵ, in order to determine the inputs effect on the output. With the proposed method the authors where able to understand the model behavior, debug models, detect dataset errors, and create visually-indistinguishable training-set attacks. The interpretability of predictions using high-dimensional vectorized features with tabular data is derived in [LWL20]. Here explanation by intervention is used to eludicate the labels generated by the prediction. The framework GRACE generated 60% more accurate post-explanation decisions than that the competing baseline method LIME. Generally, the internal state of a machine learning model is quite intransparent. This can be an obstacle in interpreting the machine learning models. However in [KWG$^+$18], Concept Activation Vectors (CAVs) are used to explain the mentioned internal state. A testing framework with CAVs (TCAV) is developed, which uses directional derivatives to quantify the importance of certain features of the machine learning model for the result of the model. In order to test TCAV, image classification models are evaluated post-hoc and the results show that CAVs are able to give insights regarding the models predicition via the internal state. This can be done from standard image classifications to models solving specialized medical problems. Another method to provide explainability for classification tasks is counterfactual explanation. This method is addressed in [MST20]. In this work, a set of diverse actionable counterfactuals is generated to interpret any machine learning model. Additionally, metrics are provided to compare counterfactual-based methods to other local explanation methods. Tested on four real life datasets, the algorithm is able to outperform similar methods by approximating local decision boundaries through generating diverse counterfactual. In order to measure the influence inputs of a machine learning model have on the output, Quantitative Input Influence (QII) using interventions is introduced in [DSZ16]. QII is able to make decisions, made by a ML system on individuals or groups, transparent. Furthermore, the measurement is able to account for correlated inputs while quantifying the joint influence of inputs as well as the marginal influence of individual inputs (using principled aggregation measures) in each set. Apart from that, the issue of transparency potentially comprising privacy is addressed by applying differential privacy onto the transparency reports. Overall, QII is able to generate better explanations than comparable measurements, can be approximated efficiently and provides differential privacy. In order to measure the causal effect within a DNN, [GSK19] use the Causal Concept Effect (CaCE) as a basis of their method. This way mistakes caused by confounding can be evade. As a do-operator cannot be simulated effortlessly, simulating the CaCE has its challenges. That's why, Variational Auto Encoder (VAE) are used to develope the VAE-CaCE. The results show that the VAE-CaCE can measure the true concept causal effect for a various number of datasets. Looking at Table 1, the above approaches are able to deal with machine learning models but fail to deliver a model-based approach or open to black box of the models.

linke Spalte	rechte Spalte

Table 1: Your caption.

Two classification approaches which are model based, hence build causal models rather than analyzing trained models regarding causality are introduced in the following. Feature selection and extraction is addressed in [KSDV15]. The Mind the Gap Model (MGM) embeddeds interpretability criteria into the model design. This way, interpretability parameters can be tweaked and identifiable dimensions are generated, which can be used for data exploration and hypothesis generation. The method is able to outperform similar apporaches and obtain identifiable features on a range of diverse real-life datasets. In [KC17], the goal is to make the algorithm, determining the steering angle of an autonomous car, explainable. First visual attention is used to highlight image region which influence the models output. Then, Causal Attention is used to provide explainability by covering parts of the information, the identified regions, within the picture. The results show that the proposed framework does not degrade the the model performance and that the model demonstrates causal behavior (concentrating on features used by humans while driving). We will draw inspiration from the mentioned approaches regarding the model-based approach. However, we will also be looking at the internal state of the model rather than only at the input or output relationship.

[AP21, SGK17] and [HDR18] present approaches in the classification domain, which analyze the internal state of the machine learning model regarding its causlity. Another attention mechanism, sequential attention, is used in [AP21]. Here a deep tabular learning architecture, TabNet, is applied in order to select the most salient features. The results show that TabNet outperforms other similar methods on multiple non-performance-saturated tabular datasets. Furthermore, the method achieves interpretable feature attributions. An approach to take the activations of each neuron in a machine learning algorithm into account to interpret the model, is developed in [SGK17]. Called DeepLIFT, the approach backpropagates the influence of each neuron to every feature of the input. This way a contribution score of each neuron is generated by comparing the current activation to its reference activation. Through that the input with the largest contribution to the output can be identified. The approach is applied to Neural Nets trained with MNIST and genomic data. The results show that compared to other methods novel dependencies can be revealed within the model through DeepLIFT. Thus, DeepLIFT is able to outperform gradient-based methods. In [HDR18], causal machine learning models are build by using the salient concepts within CNNs. The human-understandable representation of network activations (generated by autoencoders) are used to extract the salient concepts. Through that a bayesian causal model that uses the extracted salient concepts is build to make the classification interpretable. The method is used on image classification to then subsequently identify and visualize features with significant causal influence and provides a novel approach to building causal models.

Overall, looking at the classification competent approaches, similar to the regression problems, they do not address both opening the black box and operating in a model-based manner. We will draw from the existing papers regarding the used methods, like causal attention, and alter them in a way so we can use them for regression problems. Furthermore, as mentioned before, we are aiming for a model-based and internal state approach rather than choosing either one of them.

References

[AP21] Sercan Arik and Tomas Pfister. TabNet: Attentive Interpretable Tabular Learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[Boz02] Olcay Boz. Extracting decision trees from trained neural networks. KDD '02, page 456–461. Association for Computing Machinery, 2002.

[Caw08] Gavin C. Cawley. Causal & non-causal feature selection for ridge regression. In *Workshop on the Causation and Prediction Challenge at WCCI 2008*, volume 3 of *Proceedings of Machine Learning Research*, pages 107–128. PMLR, 2008.

[DSZ16] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016.

[GSK19] Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CaCE). *CoRR*, abs/1907.07165, 2019.

[HDR18] Michael Harradon, Jeff Druce, and Brian E. Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *CoRR*, abs/1802.00541, 2018.

[KC17] Jinkyu Kim and John F. Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2961–2969. IEEE Computer Society, 2017.

[KL17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[KSDV15] Been Kim, Julie A Shah, and Finale Doshi-Velez. Mind the gap: A generative approach to interpretable feature selection and extraction. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2260–2268. Curran Associates, Inc., 2015.

[KWG+18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR, 2018.

[LWL20] Thai Le, Suhang Wang, and Dongwon Lee. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. *2020 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

[MST20] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual examples. In *ACM Conference on Fairness, Accountability, and Transparency*, January 2020.

[PBM16] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[RSG16] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics.

[SGK17] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[SHG19] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *CoRR*, abs/1905.07857, 2019.

[SK19] Patrick Schwab and Walter Karlen. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, volume 32, pages 10220–10230. Curran Associates, Inc., 2019.

[STY16] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Gradients of counterfactuals. *CoRR*, 2016.

[WCZ+16] Mike Wojnowicz, Ben Cruz, Xuan Zhao, Brian Wallace, Matt Wolff, Jay Luan, and Caleb Crable. "influence sketching": Finding influential samples in large-scale regressions. 11 2016.

[ZB18] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making - the causal explanation formula. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2037–2045. AAAI Press, 2018.