Using Linear Approximations to Explain Complex, Blackbox Classifiers

USING LINEAR APPROXIMATIONS TO EXPLAIN COMPLEX, BLACKBOX CLASSIFIERS

ALEXIS ROSS

submitted in partial fulfillment of the requirements

for the degree of Bachelor of Arts in

Computer Science and Philosophy

Harvard University

April 2020

Abstract

Machine learning models have the potential to aid human decision-making in a variety of domains. However, many cannot be safely deployed because they are so complex that they are essentially "black boxes" to humans. Given this fact, the need for an independent method of explaining predictions made by such models arises. This thesis discusses how local linear approximations can be used to explain complex, blackbox classifiers. The first part of this thesis draws upon a philosophical account of causal explanation to argue that local linear approximations derived through sampling methods can be effective causal explanations of blackbox classifier predictions. The second part of this thesis proposes an original end-to-end framework for generating actionable counterfactuals that change classifier predictions. Empirical findings are presented which suggest that this method is a promising avenue for future work.

Thesis Structure

This paper is divided into two parts. Part I (Chapters 1-3) discusses the philosophical requirements of explanations of blackbox classifiers. This section builds heavily upon James Woodward's account of causal explanation (2003) to argue that linear approximations derived through sampling can function as effective causal generalizations of blackbox classifier predictions. Part II (Chapters 4-6) focuses on the application of linear approximations to obtaining actionable insight into non-linear classifiers: Specifically, I present original experimental findings suggesting that linear approximations can be used to generate actionable counterfactuals for non-linear classifiers.

Note: Part I is intended to fulfill the thesis requirements of the Philosophy department. Part II is intended to fulfill the thesis requirements of the Computer Science department.

Contents

Part I

Linear Approximations as Local, Causal Explanations

Chapter 1

Introduction

1.1 Motivation: Why do we need explanations?

Machine learning models have the capacity to aid human decision-making in a variety of domains, including medicine, banking, and criminal justice, and they are in many cases already in deployment. Such models are useful because they can learn patterns in large amounts of data–more data than a human could feasibly process to make predictions.

To illustrate the real-world utility of machine learning, suppose a model has been shown to accurately predict risk of death from pneumonia; specifically, suppose the model produces accurate predictions on 95% of a data sample held back from the data used to train the model. A doctor could use the model's prediction to confirm her own diagnosis of low risk in a new patient, especially if she has some uncertainty; she might reasonably think that the model may have learned through the thousands of scans it was trained on to pick up on very subtle features in scans that she may personally have missed.

A data-driven model would be especially useful to the doctor given that routine clinical judgement has been shown to poorly identify pneumonia severity. One 1996 study conducted in a New Zealand hospital found that clinical teams underestimated pneumonia severity compared to a set of rules based on criteria published by the British Thoracic Society in 1987. 20 out of the 250 studied patients with pneumonia died, and while the BTS rules categorized 19 of them as having severe pneumonia, medical staff identified

12 of them as having severe pneumonia, five as moderate, and two as mild (Neill et al., 1996).

Yet despite their potential to aid human decision-making by learning from large amounts of data, many machine learning models cannot currently be deployed in a safe manner. This is because many of the most accurate models are represented by such complex mathematical functions that they are essentially "black boxes": Even people who have full access to the computations underlying a given model often cannot understand its workings because of the number of its parameters. Such models are said to have low interpretability. Very complex, uninterpretable models cannot be safely used even when they have high accuracy because their opacity makes it difficult to determine whether the models have achieved high accuracy by learning meaningful, general patterns in the world or by exploiting problematic correlations in the data they were trained on, such as discriminatory correlations or patterns present only in the specific subset of data they were trained on that would not generalize to unseen data from another source. Even a model with 95% accuracy on a held back sample of data may not necessarily produce accurate predictions on new data if the new data is not adequately represented in the training data.

Consider the following real-world example that highlights the dangers associated with deploying black-box models: In 1997, Cooper et al. conducted a study to evaluate how machine learning models could be used to predict the probability of death for patients with pneumonia (Cooper et al., 1997). Such models, if successful, could be used to prioritize hospital treatment for high-risk patients. The study trained and evaluated different types of models, including logistic regression, a rule-based system, and neural networks, using data on 14,199 inpatients discharged from hospitals in the United States between 1987 and 1988.

Of these models, a neural network, the least interpretable kind of model achieved the highest accuracy. Yet, despite its accuracy, this neural network was never deployed

because of a discovery made by the researchers in another model: They found that the rule-based system had learned the rule that if a patient has asthma, the patient has a low risk of death from pneumonia. It turned out that this correlation between asthma and a low risk of death was an artifact that existed in the training data, since all patients with asthma were referred by their doctors to intensive care units and thus received intense care that lowered their risk of death. Since the neural network had been trained on the same data as the rule-based system, the researchers figured that it too would have learned this correlation and thus could not be actually used to aid medical decision-making.

Clearly, this correlation is not one we would want a deployed model to be using in making its predictions, given that a history of asthma actually increases a patient's risk of death from pneumonia. Yet fragile correlations like this one exist in data everywhere, and it is critical that we can rule out that a model is exploiting such a correlation before it is deployed for real-world use; without ruling out this possibility, we cannot know that our deployed model will make accurate predictions for novel data, which may not contain the same correlations present in its training data. With interpretable models, transparency of model logic comes built-in: In the described case, for instance, it was the interpretability of the rules in the rule-based system that allowed the researchers to determine that the models trained on the existing data could not be trusted. In contrast, with uninterpretable models such as neural networks, the logic of the models themselves is not interpretable. In order to answer questions such as whether someone with asthma would be more likely to die from pneumonia than someone without asthma, we need some an independent method of explaining model predictions.

While interpretable models trained on the same data can be used to understand training data artifacts to an extent, there is no guarantee that other uninterpretable models will be exploiting the same patterns in data. For instance, in the particular study discussed, researchers could infer that the neural network was likely untrustworthy models because it had been trained on the same data as the rule-based system, which was found to be

relying on a problematic correlation between asthma and probability of death; however, even if they had not found any problematic correlations being exploited by the rule-based system, it is possible that the neural network could have still learned other problematic patterns in the data. Determining whether it indeed had would require the use of some independent method of explaining what led to model predictions.

The need for such an explanation method is particularly important given that the most accurate models are often the least interpretable: Very complex machine learning models, such as neural networks with thousands of parameters, have more capacity to capture patterns in data and thus make more accurate predictions than more simple, interpretable models such as linear models and rule-based systems, which are limited in the complexities in data they can represent. Thus, though model deployers always have the choice of using an interpretable model, this choice comes at the cost of model accuracy and performance. Having a method of explaining even uninterpretable models would mean we could benefit from complex models' high capacity to represent data while also knowing whether and how they should be used. A model such as a neural network could be deployed, for instance, if doctors could have explanations of how features–such as asthma–contributed to a prediction–such as low/high-risk–for a given prediction. Then, if a doctor were deciding whether or not to send a patient home or send them to an intensive care unit, she could know not to trust the prediction.

There are additional cases where explanations of model predictions would be useful. In the domain of criminal justice, models that have been trained to predict recidivism rates, or a convicted criminal's likelihood of recommitting a crime, are currently being used in courtrooms to determine convicts, bond amounts and determine sentences. A judge who is deciding the sentence for a convict for whom a recidivism model gave a prediction of high risk would need an explanation of its prediction to ensure that it was not in some way discriminatory. If the prediction of high risk were driven by a legally protected attribute such as the convict's race, the judge may want to disregard the prediction. Thus, she

would find use in an explanation that could point out the extent to which features such as race causally influenced the model prediction.

In the financial industry, models are currently being used to predict loan candidates' likelihood of loan repayment. For a candidate denied a loan from a bank employing such models, an explanation of the model's prediction would give information to the denied candidate about what actions could be taken to change the model's prediction in the future.

These three different settings in which explanations would be helpful highlight the need for a method of explaining the predictions of black-box models if they are to be deployed. In this thesis, I consider what kinds of explanations we would want in each of these settings and develop a methodology for creating them; I will eventually argue that different sorts of explanations are useful in different contexts. The specific kind of model that will be considered in this thesis is a classifier, which assigns labels–such as low/high risk–to different input data. In Section 2, I discuss features of the kinds of explanations we would want and use these features to motivate adopting a particular kind of theory of explanations. In Section 3, I outline this theory, Woodward's counterfactual account of causal explanation. In Section 4, I argue for the use of linear models as explanations and contrast them with explanations consisting of counterfactual conditionals. In Section 5, I argue that different types of explanations are best suited for different contexts.

1.2 Desiderata for Explanations

The three outlined scenarios in which explanations of model predictions would be useful highlight three distinct uses of explanations–to determine whether a prediction was made in a discriminatory manner, whether a prediction was influenced by meaningless artifacts of data, and how a prediction can be changed in the future.

At the heart of each of these three kinds of explanations is the idea of causality: That is, a good explanation in each of these contexts would pick out the features that causally

influenced the prediction and describe their causal influence. With causal explanations, we could determine whether the causal influences on the model's predictions are trustworthy by comparing them with the causal structure of the phenomena being modeled in the world. Thus, the first requirement for explanations in the described cases is that they preserve and highlight causal relations between features of input data and model predictions.

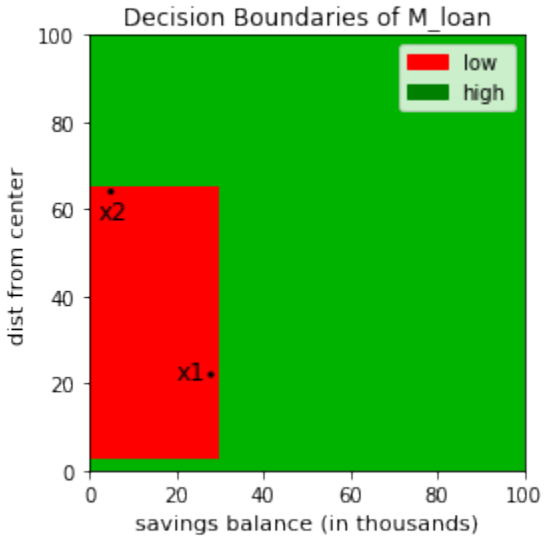
A second requirement is that the explanations be simple enough to comprehend. The reason why the mathematical computations underlying the model's predictions, which describe the causal relationships of interest between model features and output predictions, cannot themselves be the explanations is that they are too complex for people to grasp.

Lastly, we can distinguish between model explanations in terms of their scope. An explanation can apply to different subsets of the data, which can range in size. At one extreme, an explanation's scope could consist of only inputs very similar to a given input, or it could consist of all possible inputs a model could receive. It is necessary to distinguish between these types of explanations because the specificity of the explanations will vary greatly by scope. For instance, suppose we had a model–call it `M_loan`–that was trained to predict someone's likelihood of paying back a loan based on two features, `dist_from_center`, representing the distance a person lives from the city center in miles, and `savings_balance_thousands`, representing how many thousands of dollars they have in their savings account. Suppose the following logic described how the model made its predictions:

```
if dist_from_center < 3 or >= 65:
    return high
else if savings_balance_thousands > 30:
    return high
else:
    return low
```

Suppose this model made the following two predictions: Upon input `x1:`

`[savings_balance_thousands = 28, dist_from_center = 22]`, the model predicted `y1 = low`, and upon input `x2:[savings_balance_thousands = 5, dist_from_center = 64]` the model predicted `y2 = low`. The following figure depicts where `x1` and `x2` fall on the model's decision boundaries, or regions of input space partitioning inputs into different classes of predictions; the red region corresponds to the region of input space where the model predicts `low` likelihood, leading to a rejection of loan application, and the green region corresponds to the region of input space where the model predicts `high` likelihood, leading to an acceptance of loan application.



Good explanations of `y1` and `y2` would pick out different features as being causally influential. A good explanation of `y1` would highlight that `savings_balance_thousands` had a stronger causal influence on the model's prediction than did `dist_from_center`, since slight changes to `savings_balance_thousands` but not `dist_from_center` would result in a different model prediction; on the other hand, a good explanation of `y2` would pick out `dist_from_center`, rather than `savings_balance_thousands`, as a causal contributor.

These individual explanations of `y1` and `y2` are examples of *local* explanations, or explanations of particular predictions. They are useful in deciding what actions should be

taken in response to those particular predictions. The person represented by the instance `x1` who received a prediction of `low` might want to know, for instance, that it was his savings account balance, rather than how far he lived from city center, that most strongly influenced the model's assessment of his likelihood of loan repayment, since he could know based on this knowledge to focus on saving more money over the next few years; furthermore, this person would not find use in an explanation for `y2`, since he would not get insight into how to receive a loan from the knowledge that `dist_from_center` strongly influences the predictions for people who live far away from the city center. The explanation of `y1`, rather than an explanation of some other aspect of the behavior of `M_loan`, such as its behavior across other inputs, is what would be most of use to the person represented by `x1`. Thus, in this case, it is local explanations, rather than *global* explanations–or explanations of model behavior across many predictions–that are needed.

There are analogous uses of model explanations that specifically require local explanations in the contexts of medical diagnosis and recidivism prediction as well. A doctor who has been given a model's prediction of an asthmatic patient's likelihood of dying from pneumonia might want to know whether the patient's asthma influenced the prediction in a medically sound way; information about how a prediction for a very different patient–such as one of a different age–would not help answer this question. Similarly, a judge interested in determining whether a model's prediction of a convict's recidivism likelihood was influenced by a protected attribute, such as the person's race, would care about the attribute's causal influence on *that particular prediction*.

Not only do local explanations uniquely satisfy certain needs for explanations such as the ones just described, but also they provide the building blocks for global explanations. An explanation meant to apply to both `y1` and `y2` would have to build upon the individual explanations for `y1` and `y2` in some way, either by pointing out the different causal influences of the features for each of the two inputs or combining them in some way (such as through the use of an average). The same is true for explanations meant to apply to

more than two inputs. The question of how best to build global explanations from local explanations remains open: Does a feature count as a global causal contributor to model predictions if it is a local causal contributor to some, most, or all predictions? However, regardless of the answer to this question, since local explanations of features' causal influence play a critical role in determining their global causal influence, local explanations are valuable.

I do not deny that global explanations can provide useful insight into machine learning models; however, questions about the kinds of insight they provide and about how they should be constructed from local explanations are questions that I will put aside for this thesis. I will be limiting the scope of this thesis to local explanations of particular predictions rather than of general model behavior for the two reasons outlined: the usefulness provided by their specificity and the critical roles they play as building blocks of global explanations.

To summarize, in this thesis, I will be focusing on *local* explanations of particular predictions which are *causal* and *comprehensible*. I will ultimately argue that linear model approximations derived through sampling methods function as the appropriate local, causal, and comprehensible explanations; this argument makes use of an account of causal explanation offered by James Woodward. I will also discuss limitations of linear models and propose a way of dealing with these limitations, which appeals to certain properties determined by the contexts in which the explanations are being used.

Before outlining what such linear model explanations look like and arguing for their merit, I describe an account of causal explanation, which first lays out and builds upon an account of how to determine causality between variables. An account of causality which allows us to be precise about what it means for a feature to influence a model prediction is necessary to have an account of causal explanation which allows us to evaluate explanations by their ability to pick out causal relationships. As illustrated by the consideration above of `y1` and `y2`, we do not consider all features that a model receives as inputs to

be equal causal contributors to the model's predictions and thus to have equal weight in causal explanations. Thus, we need criteria for what exactly it is that makes certain features causal contributors and other features not. In the next section, I lay out the causal criteria of James Woodward and discuss how they can be applied to the endeavor of explaining features' causal influences on model predictions.

Chapter 2

An Account of Causal Explanation

In my thesis, I will be building upon counterfactual theories of causation, which derive causal notions from counterfactual conditionals. At the heart of such theories is the idea that the causal influence of one variable X on another variable Y can be determined by observing whether specific kinds of alterations to X produce changes in Y. This idea is able to explain our intuitions about the differing influences of `savings_balance_thousands` and `dist_from_center` on `y1` in the case of `M_loan`: `savings_balance_thousands` strongly influences `y1` while `dist_from_center` does not because slight alterations to `savings_balance_thousands` but not `dist_from_center` lead to a change in model prediction. The same reasoning can be applied to derive the stronger causal influence of `dist_from_center` than `savings_balance_thousands` on `y2`.

The specific theory I will be working with in this thesis is James Woodward's account (Woodward, 2003), which offers both an interventionist account of how to determine causation and an account of how to assess causal explanations in terms of interventions.

There are four notions central to Woodward's account of causal explanation: *generalizations*, *interventions*, *causal relationships*, and *testing interventions*. Woodward's high-level goal is to be able to account for what makes some *generalizations*–or relationships between changes in different variables–true causal explanations. To do so, he relies on the notion of *interventions*, which are experimental manipulations made on some variable X with respect to some target variable Y which can be used to derive the nature of the

causal relationships between X and Y. Causal relationships are derived through the use of interventions, and generalizations are evaluated through a specific kind of intervention called a testing intervention.

Woodward's view is that generalizations describe *true* causal relations if they are *invariant*, or stable and continue to hold, across a specific kind of intervention: *testing interventions*. Furthermore, he holds that among true generalizations, some have more explanatory power than others: A generalization's explanatory power is determined by the *range* of its invariance over interventions, including testing and non-testing interventions. It is important to note that explanatory power differs from the significance of a causal relationship between two variables: The latter is determined through interventions on just the two variables of the causal relationship, while the former is determined through interventions on any variables featuring in the generalization.

The requirements that Woodward lays out for a generalization to qualify as true and effective causal explanations are ones that I will eventually argue linear models derived from a sampling-based method satisfy. But before describing these requirements, I offer an explication of Woodward's interventionist account of causation, since Woodward's notion of interventions is a central building block in Woodward's account of how we should think about causal explanation.

2.1 An Intervention-Based Account of Causality

According to Woodward, what it means for a variable X to cause another variable Y is for there to be some *intervention* on X that would change the value of Y. An *intervention* on a variable X with respect to a variable Y is a special kind of experimental manipulation of X which makes it the case that a change in Y can only come about through the manipulation of X and not through some other causal route (94). Interventions on X with respect to Y can be used to determine the existence of a causal relationship between X and Y in the following way: If there is *some* intervention on X which produces a

change in Y, then Woodward holds that a causal relationship between X and Y exists.

Woodward offers the following example of an intervention (98): Researchers are investigating the causal relationship between treatment with a drug (T) and recovery from a disease (R). They have access to a population of subjects with the disease. Interventions that could be used to investigate the causal influence of the drug on recovery would be the administration or lack of administration of the drug to subjects; if the intervention of drug administration, which would set the value of T, did lead subjects to recover, a causal relationship between T and R could be inferred.

Not all manipulations of T would count as legitimate interventions, however. Woodward points out that if subjects learned whether they were administered the drug in the manipulations, the interventions could affect R independently of T through the effects of placebo, and so we would not be able to determine the causal effect of T on R (98). Woodward lists a series of technical requirements for a manipulation on X to count as an actual intervention; these requirements are meant to address non-intervention manipulations like the one described that may, because of the causal structure of the world, result in changes to Y that come about through a causal chain except the directed causal path from X to Y.

I now define these technical requirements and Woodward's notion of an intervention more precisely. Woodward introduces the notion of an *intervention variable* when defining interventions. I qualifies as an intervention variable for X with respect to Y if and only if:

I1) I causes X.

I2) I acts as a *switch* for all other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.

I3) Any directed path from I to Y goes through X.

I4) I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.

Woodward then provides the following definition of an *intervention*:

(IN) I's assuming some value $I = z_i$ is an intervention on X with respect to Y if and only if I is an intervention variable for X with respect to Y and $I = z_i$ is an actual cause of the value taken by X.

In the case of determining whether T has a causal effect on R, setting T through drug administration that involved telling subjects whether they received the drug would violate I3, since it would affect the value of R through a path that did not go through T–a path involving placebo effects (98). I2 ensures that any interventions on drug treatment entirely determine the value of T; so, for instance, if patients normally have the choice to take or not take the drug, the intervention breaks the connection between this voluntary choice and whether they receive the drug (97). And I4 is meant to ensure that the interventions on drug treatment are not correlated with other causes of recovery, which would be the case if patients receiving treatment had stronger immune systems than those not receiving treatment (97).

2.2 Making Interventions on Features

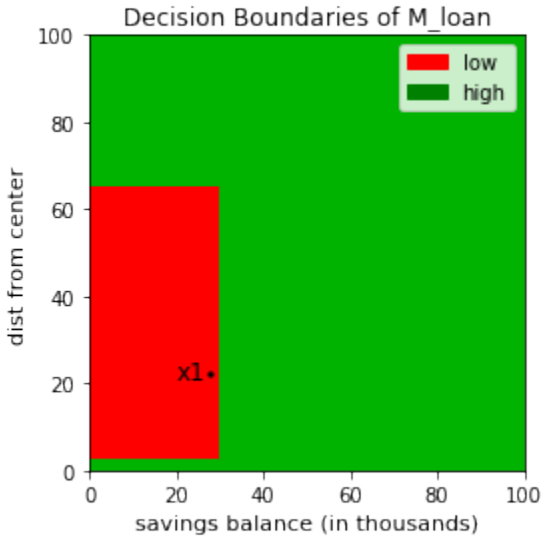
In this section, I discuss how interventions can be used to draw causal claims between input features and model predictions. I first offer a sketch of this application and show how manipulations of feature values qualify as Woodwardian interventions. I then argue that simple interventions cannot be used to derive all of our desired causal claims.

2.2.1 A Sketch

The technical requirements laid out by Woodward are met by the sorts of manipulations we would be making to determine causal relationships between input features and model predictions, since it is possible to make completely isolated manipulations of features that do not alter the causal relationships between other features and model predictions. In order to determine the causal influence of a feature F on a model's prediction P with a Woodwardian intervention, we could alter the value of the feature and see if the model's prediction changed. In such an alteration, the intervention variable I would be the alteration to the value of F of the original input. Such an alteration would meet I1, since the alteration would cause F to take on a certain value. It would meet I2, since it would entirely determine the value of F. Additionally, altering the value of F could not have an effect on P other than through the change in value of F, given that we are making a completely isolated change to the feature of interest; thus, this sort of alteration would meet I3. And lastly, this alteration would fulfill I4, since it is a controlled alteration and so is not correlated with any other cause of Z.

To make clear how such an intervention could be made, let's return to the toy model `M_loan`. Suppose we wanted to determine whether the variable `savings_balance_thousands` influenced prediction `y1 = low` for input `[x1: savings_balance_thousands = 28, dist_from_center = 22]`. Recall that the model has the following logic and decision boundary:

```
if dist_from_center < 3 or >= 65:
    return high
else if savings_balance_thousands > 30:
    return high
else:
    return low
```



Changing `savings_balance_thousands` from 28 to 31 would result in `y1` to change from `low` to `high`. This intervention would, on Woodward's account, enable us to conclude that `savings_balance_thousands` does causally influence the prediction, since what Woodward would require is that there be *some* intervention on `savings_balance_thousands` that leads to a change in model prediction to `high`. Thus, though there exist other interventions on `savings_balance_thousands` that would not lead to such a change in model prediction, such as changes setting it to 29 or 30, the existence of at least one intervention that leads to a change in model output is enough to derive a causal relationship between `savings_balance_thousands` and `y1`.

Furthermore, interventions can be used to compare the strengths of causes. In this particular example, interventions on `savings_balance_thousands` and `dist_from_center` could be used to conclude that `savings_balance_thousands` had a stronger causal influence than `dist_from_center` on `y1`. The minimum change in `dist_from_center` that would result in a different model prediction of `high` would be a change from 22 to 2, a change of 20 miles. On the other hand, the minimum required change in values for `savings_balance_thousands` is the change from 28 to 31 thousand dollars. In order to compare the magnitudes of these changes, we would normalize the feature val-

ues in these interventions using the means of each feature; this normalization equalizes the units of change we are comparing for different features. Suppose the mean `dist_from_center` across all instances in the training data were 50, and suppose the mean `savings_balance_thousands` were also 50 (in thousands). Then we would compare the magnitudes of values $22/50 - 2/50$ (for `dist_from_center` and $28/50 - 31/50$ (for `savings_balance_thousands`), or 0.4 and -0.06. Since $0.06 < 0.4$, we could conclude that `savings_balance_thousands` had more causal influence on the prediction `y1` than did `dist_from_center`. Comparisons across the minimum interventions on features that will result in changed model predictions can thus in this way be used to derive features' relative causal influence.

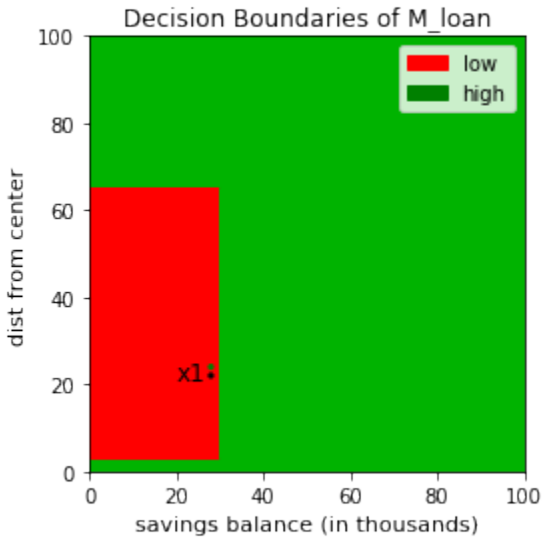
2.2.2 Limitations: Causal Irregularities and Overdetermination

At this point, I would like to address a natural question which may arise: Since interventions can be used to derive causal relations, why do we need generalizations such as linear model approximations to explain model predictions at all? It might be thought that the causal relationships derived from interventions on different features can function as causal explanations.

My answer to this question is the following: Though interventions can be used to determine whether causal relationships exist, single interventions cannot straightforwardly function as explanations in two sets of cases: cases where the causal relationships are subject to irregularities and cases of overdetermination. According to the view put forward, a causal relation between a feature and a prediction holds if there is *some* intervention on the feature that produces a change in prediction. I will show that this view fails to derive the appropriate causal relationships in two sets of cases: cases where the causal relationships are subject to irregularities and cases of overdetermination. Though there are ways to modify Woodward's account to derive the appropriate causal relationships using interventions in these cases–namely by aggregating information given by many different

interventions–this comes at the cost of cognitive convenience. Furthermore, I will later argue that linear models implicitly do this aggregation and are cognitively convenient to understand.

Causal Irregularities We want the causal claims that we draw about features and model predictions to pick up on underlying patterns in how features influence model predictions. Let us return to the endeavor of explaining the relative strengths of the causal influences of `dist_from_center` and `savings_balance_thousands` on `y1`. But suppose the decision boundaries of `M_loan` were slightly different so that it looked like the following:



Before, I suggested that we could conclude that `savings_balance_thousands` has a stronger causal influence than `dist_from_center` on `y1` by observing that the magnitude of the (normalized) minimum change to `savings_balance_thousands` required to change `y1` is smaller than the magnitude of the (normalized) minimal change to `dist_from_center` required. However, if the model were characterized by the current decision boundaries, this observation would not be true: Interventions of equal numerical magnitude on both `dist_from_center` and `savings_balance_thousands` would result

in changes to `y1`; specifically, changing `dist_from_center` from 22 to 23 would result in a change to `y1`. (And because the means of both features are assumed to be the same as described earlier, the magnitudes resulting from comparison with normalization are also equal.)

Yet I take it that our intuition in this case is that `savings_balance_thousands` still has a stronger causal influence than does `dist_from_center`, since the intervention changing `dist_from_center` from 22 to 23 is the *only* one that would cause a change. Besides this *single* exception of an intervention, it is true that small interventions on `dist_from_center` (which are the ones relevant for determining local causal relationships) do *not* lead to changes in `y1`. We would not want to say that there is the same sort of causal relationship between `dist_from_center` and `y1` as the one between `savings_balance_thousands` and `y1`, as is suggested by the equal magnitudes of the minimal prediction-changing interventions for both features.

As an analogy, consider the following example, inspired by one formulated by Dupré (1984): Suppose that in the real world, there is a very rare strain of strawberries, present in .1% of the world's strawberries, that when eaten strongly increases people's chances of getting stomach cancer. Suppose someone is trying strawberries for the first time. There are *some* strawberries, specifically .1% of all possible strawberries, that this person could eat which would causally influence the person's getting stomach cancer. But the person could eat almost all strawberries and not have her risk of getting stomach cancer change at all. In such a case, we would *not* conclude that the causal relationship between eating strawberries and getting stomach cancer is that eating strawberries causes stomach cancer, even though this causal claim would hold if certain rare strawberries were eaten. Rather, we would conclude that eating strawberries does *not* cause stomach cancer and that there are rare exceptions to this causal relationship.

Similarly, in the case described here, we would not want to conclude from the *single* local intervention on `dist_from_center` changing its value from 22 to 23 that there

is a causal relationship between `dist_from_center` and model predictions. We would instead want to conclude that `dist_from_center` does not locally causally influence `y1`, though there is an exception to this lack of causal relationship. We would thus want to distinguish between the causal claims drawn about the more regular causal relationship between `savings_balance_thousands` and `y1` and the less regular causal relationship between about `dist_from_center` and `y1` holding over only one local intervention.

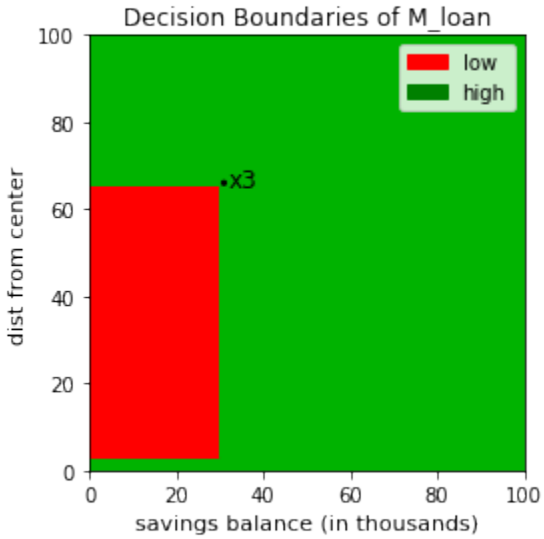
More generally, we want the causal claims we draw about the relationships between the features of `x1` and `y1` to reflect the underlying regularities of these relationships.[1] What this example highlights is that merely finding *some* intervention on a feature that produces a change in model prediction is *not* enough to conclude that the causal relationship between them is regular.

In order to derive causal relations with the appropriate regularities through interventions, we could modify Woodward's requirement for a feature to causally influence a model prediction that there be *some* intervention on a feature that will bring about a change in prediction to the requirement that there be a *range* of interventions to the feature that will bring about a change. In fact, this is precisely the proposal that David Lewis (2000) makes. But once we have information from ranges of interventions, we need some way of aggregating that information, which I will argue is just what linear models do.

Overdetermination. Whenever there are two or more features that are sufficient to cause a given result, the kinds of Woodwardian interventions on those features just discussed, which alter one of those features while keeping other features fixed, will fail to uncover the causes, since the other unaltered feature will still sufficiently cause the model prediction towards the original prediction.

For instance, suppose we wanted to explain `y3`, the prediction of `high` made by `M_loan` on instance `x3 = [savings_balance_thousands = 31, dist_from_center = 65]`, depicted below.

[1] A lengthier discussion of why causal robustness is desired can be found in Section 3.4.1.



This is a case of overdetermination since the values of both `savings_balance_thousands` and `dist_from_center` are independently sufficient to produce the prediction of `high`. The sufficiency of `savings_balance_thousands = 31` in causing the prediction can be seen by the fact that across all possible changes to other features–in this case the only other feature is `dist_from_center`–the prediction would still be `high`; that is, across all changes to the vertical position (`dist_from_center`) of `x3` that keep the horizontal position (`savings_balance_thousands`) of `x3`, fixed the prediction remains the same. Similarly, `dist_from_center = 65` is sufficient to bring about the prediction because the prediction would not change across any changes to the other features of `x3`, which here include just `x3`'s horizontal position (`savings_balance_thousands`).

Because the values of both `savings_balance_thousands` and `dist_from_center` are sufficient to bring about the prediction, they both causally influence the model's prediction. However, the simple approach of using single Woodwardian interventions to infer causal relationships would fail to uncover their causal influences. In order to determine the influence of `savings_balance_thousands`, we should, according to Woodward, make slight alterations to its value while keeping the values of other features fixed

and see whether there exists some alteration that leads to a change in model prediction. Yet, as just discussed, no such alterations would bring about a change, since the value of `dist_from_center` was also sufficient to bring about a change. Thus, the lack of an intervention on `savings_balance_thousands` that would cause a change in `y3` would appear to suggest the wrong conclusion–that `savings_balance_thousands` did not causally influence `y3`. Similarly, no intervention on `dist_from_center` would lead to a change in `y3`, since the value of `savings_balance_thousands` is also a sufficient cause of `y3 = high`. Thus, interventions would fail to highlight the causal influences of both `savings_balance_thousands` and `dist_from_center` on `y3`.

It is important that we can draw causal claims about causes in cases of overdetermination in all three uses of explanations outlined earlier: deciding future actions based on a prediction, determining whether a model made use of protected attribute in reaching a prediction, and determining whether a model prediction was influenced by meaningless artifacts in data.

As an illustration of the first case, suppose a person received an output from a loan repayment prediction model that predicted she had a low likelihood of paying back the loan and she wanted to figure out how to change the model's prediction in the future. As in the just discussed example where causal claims derived from Woodwardian interventions failed to highlight the causal influences of both `savings_balance_thousands` and `dist_from_center` on `y3`, if the prediction were sufficiently influenced by two features, an explanation that failed to uncover causes in cases of overdetermination would fail to highlight either of the two features, and she would be left with no information about what actions to take to change the model's prediction, even if such actions existed.

As an example of the second kind of use of explanation, suppose a judge is trying to determine whether a prediction made for convict's likelihood of recidivism was influenced by the convict's race. If the convict's race *and* the value of another feature, such as the nature of the convict's crime, were both sufficient causes of the prediction, an explanation

that failed to uncover causes in cases of overdetermination would fail to highlight that a protected attribute, race, did in fact influence the model's prediction. In fact, the explanation would fail to highlight either of the sufficient causes as causes.

Lastly, suppose a doctor wants to verify that a prediction of a patient's low risk of death from pneumonia was made in a medically sound way by making sure that the patient's asthma correctly increased the person's risk of death. Suppose the model actually did *not* learn the correct relationship between asthma and pneumonia-induced mortality and had in fact learned a relationship with an opposite causal direction than the true one in the real world, and so it in fact was the case that the person's asthma pushed the model's prediction towards a decreased risk of death. Now suppose the patient is old, and so the model's prediction had two sufficient causes: the patient's asthma and the patient's age. In such a case, an explanation that failed to highlight causally influential features in cases of overdetermination would fail to highlight the critical information that the patient's asthma actually influenced the prediction in a medically unsound way.

In order to use interventions to find causes in cases of overdetermination, we could consider expanding the notion of interventions to apply to combinations of features. For instance, we could intervene on both `savings_balance_thousands` and `dist_from_center` to see whether `y3` changes in response to pair-wise interventions. We might discover with such pair-wise interventions that changing `savings_balance_thousands` from 31 to 30 while also changing `dist_from_center` from 65 to 64 would lead to a change in model prediction. From this pair-wise intervention, in combination with the knowledge that no intervention on one of the two features results in a change to `y3`, we could conclude that both `savings_balance_thousands` and `dist_from_center` had causal influences on `y3`.

It seems like intervening jointly on multiple features in addition to intervening on single features gives us a way to address the issue raised about interventions' inability to uncover features that are causally relevant when there exist multiple sufficient causal features. However, it is unclear how to actually execute such joint interventions, especially

when the number of features is high. While intervening on pairs of features seems feasible, intervening on more than two features at a time is a computationally difficult task, given that there are many ways of intervening on multiple features–There are different groups of features that can be chosen, and for each, different ways of setting the values of each individual feature. It would be computationally intractable to execute *all* such interventions.

Furthermore, even if we did have a way of executing all, or most, of such interventions, it would still remain an open question how to derive causal claims from these interventions and their results. Introducing joint interventions into the mix of interventions to derive causal claims from only increases the need for aggregation of information across interventions, as well as the likelihood of finding causal irregularities of the sort previously discussed. Suppose we had a model that acted on 3 features instead of 2: `[f1, f2, f3, ..., f15]`, and suppose the values of `f1`, `f2`, and `f3` were each sufficient to bring about the prediction made on a particular instance. In order to use interventions to uncover each feature's causal sufficiency, we would need to obtain the following results: that no intervention on an individual feature brought about a change in model prediction, that no intervention on pairs of features brought about a change in model prediction, and that some intervention on all three features brought about a change in model prediction. But if there were one (and only one) intervention on `f1` and `f2` that caused a change in model prediction, would `f3` not count as a sufficient cause? It seems like this case would also be a causal irregularity that we would not want to affect the conclusion we drew about `f3`'s causal influence. What I am trying to highlight is that considering joint interventions only exacerbates the need for our explanations to rule out irregular instances of counterfactual dependence of model predictions on features.

At a high-level, the discussions of both causal irregularities and overdetermination highlight that we need to aggregate over interventions to get appropriate causal insight from them. But this need for aggregation is precisely what gives rise to the need for

generalizations, or relationships which relate changes in different variables and are meant to capture underlying causal patterns. I will argue that linear model explanations do this aggregation and thus can be used to derive causal relationships even in cases of overdetermination and causal irregularities. But before doing so, I turn to Woodward's account of generalizations.

2.3 An Invariance-Based Account of Explanations

Woodward seeks to explain what makes some generalizations–or relationships between changes in the values of one or more variable and changes in the values of another–qualify as legitimate causal explanations of the phenomena they model and others not. He wants to explain what it is that makes the law of gravity a true causal law, for instance (101). To do so, he appeals to his notion of interventions.

According to Woodward, "invariance under at least one testing intervention (on variables figuring in the generalization) is necessary and sufficient for a generalization to represent a causal relationship or to figure in explanations," where a *testing intervention* is a particular kind of intervention which changes an independent variable in a generalization enough so that the dependent variable in the generalization is predicted to change (250). By a generalization's *invariance*, Woodward means its reliability: That is, a generalization is *invariant* over an intervention or testing intervention if it produces the correct value for its dependent variable even when when that change is made.

The requirement that generalizations be invariant over *testing* interventions specifically ensures that the causal structure asserted by a generalization actually exists. To show how testing interventions specifically give us such insight, consider the following example offered by Woodward of an equation meant to describe the causal relationship between whether a light is on or off and the angular displacement of the light switch:

$$L = [q]$$

L is a variable that represents whether the light is on ($L = 1$) or off ($L = 0$), q represents the angular displacement of the switch in radians, and [] represents the floor function, which outputs the greatest integer less than or equal to its input, such that $[1.677] = 1$. According to this equation, the light switch turns on when the switch is at a position of 1 radian (about 57.3 degrees), or higher.

Suppose the position of the switch was at 15 degrees and we wanted to determine whether the equation were a true causal explanation of the relationship between the light's being on/off and the angular displacement of the light switch. In order to do so, according to Woodward, we would have to determine whether it was stable over a testing intervention on one of the variables featuring in the equation–in this case, q. An intervention that changed the switch position to 56 degrees would *not* be a testing intervention, since according to the equation the value of L is not predicted to change. On the other hand, an intervention that changed the switch position to 58 degrees *would* qualify as a testing intervention, since it would project that the light would turn on.

Woodward argues that the equation must be invariant over this latter sort of intervention–the one changing the switch position to 58 degrees–in order for the equation to be a true causal explanation because only under that circumstance would the equation capture the causal relationship between the value of L and the value of q. If we accept Woodward's characterization of causation in terms of interventions, what it means for q to causally influence L is for there to be some intervention on q for which L changes. The testing intervention setting q to 58 is predicted by the equation to be exactly this sort of intervention. Thus, the equation relating L and q must be invariant over this intervention to ensure that the causal relationship between L and q asserted in the equation actually exists–that is, that the causal relationships asserted by the equation capture the true causal relationships it seeks to model.

On the other hand, the intervention changing the switch position to 15 degrees does *not* suffice to make the equation causal, since the lack of change predicted by the inter-

vention would be compatible with a generalization other than the equation offered–the generalization asserting that the light switch is broken. If the light switch were actually broken, the equation could be invariant over–or give the correct prediction for–the intervention setting the value of q to 15 degrees (by properly predicting that the light would not turn on) while still not describing the actual lack of causal relationship between L and q. Thus, this intervention would not give insight into whether the equation captured the underlying causal relationship between q and L; we could only obtain this insight through an intervention for which the equation would predict a change. More generally, we can only know that a generalization captures true causal relations if it is invariant over change-predicting testing interventions.

Though invariance over a *single* testing intervention is what allows us to determine that a generalization describes a true causal relationship between variables, invariance over testing interventions has an additional purpose: It gives insight into the robustness of the asserted causal relationship. That is, testing interventions can be used to determine not only whether the causal structure asserted by a generalization holds at all (as just described), but also the extent to which that causal structure is stable. Consider Woodward's example of the generalization of the ideal gas law, which describes the relationships between volume V, pressure P, gas temperature T, the amount of substance of gas n (in moles), and R, the ideal gas constant:

$$PV = nRT$$

According to Woodward, this generalization is a true causal explanation because it is invariant under some testing interventions on T. However, it is not invariant under *all* interventions on T–If T is increased sufficiently, intermolecular forces between the gas molecules which are otherwise negligible become significant, and so the generalization will not produce the correct predictions for P or V. Only an intervention on T that set the temperature sufficiently high would allow us to determine that the causal structure

asserted by the ideal gas law breaks down at high temperatures. It is thus a range of testing interventions that allows us to determine *how* invariant a generalization is.

Woodward also acknowledges that invariance over *non*-testing interventions contributes to the explanatory power of generalizations. Better explanations will be invariant not only over a range of testing interventions, but also a range of other changes in variables. This is because generalizations that are more invariant can be applied to new situations beyond the specific "evidential context" in which it was discovered (296). For instance, the ideal gas law is a good explanation because it holds over a range of changes: It is invariant over many testing interventions setting the values of T, V, and n, and P and only breaks down when these testing interventions cross a certain threshold. The ideal gas law would have substantially less explanatory power if it only held for gases in a *particular* container.

This two-fold idea–that (1) a generalization's invariance over testing interventions determines whether it is a causal explanation at all and (2) the range of a generalization's invariance over both testing interventions and changes that are non-testing-interventions determines its explanatory power–will be critical to my argument that linear model approximations can be good explanations for a few reasons. (1) defines a requirement for linear models to function as good explanations of model predictions–that they are invariant over local testing interventions; this is something I will show. (2) provides a framework for creating criteria to choose between linear models, which I will outline later.

In the next section, I will show argue for the use of linear model approximations as appropriate explanations of model predictions. I will argue first that linear models meet the requirement outlined by (1) by showing that they are designed to be invariant over testing interventions. I will then argue that using linear models as explanations overcomes the shortcomings of using mere testing interventions. I will then discuss limitations of linear models and offer a way of dealing with such limitations, which involves using the framework provided by (2).

Chapter 3

Linear Approximations as Causal Generalizations of Model Predictions

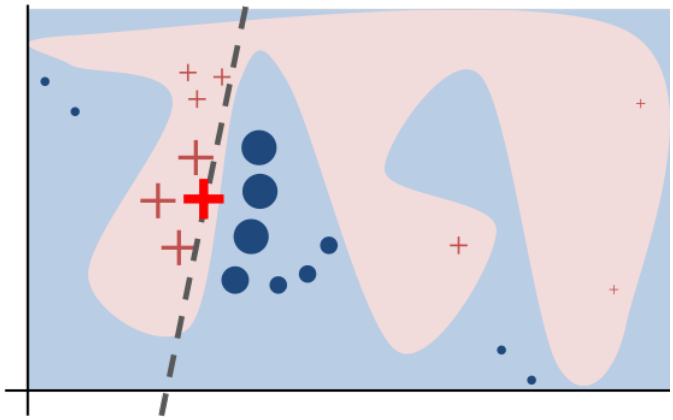
One useful class of generalizations that can function as explanations of complex models is the class of *linear* generalizations, which assign weights to input features and describe model predictions in terms of a weighted sums of these features. In fact, the idea of using linear approximations as explanations of model predictions has been explored in previous computational work. Most notably, Ribeiro et al. (2016) introduced a method called LIME, or local interpretable model-agnostic explanations, which produces linear model approximations using local sampling; they argue that such approximations can be thought of as explanations of model predictions.

In this chapter, I consider whether linear approximations derived through methods like LIME count as explanations in a *philosophical* sense. I build upon the account of causal explanation offered in Chapter 2 to argue that in many explanatory contexts, linear approximations derived through local sampling methods can function as effective causal generalizations. I also discuss the limitations of such linear approximations as explanations and ways they can be addressed. I begin in Section 3.1 with a discussion of what linear approximations derived through local sampling look like and show how they are designed to be locally invariant over testing interventions.

3.1 Local Invariance over Testing Interventions

LIME works in the following way: It samples points near the instance x being explained and weights them by their distance to x, where the distance metric used is a Euclidean distance metric based on feature values. A linear regression model is then fit to these weighted sampled points.[1]

The key point exploited by LIME and highlighted in the image below is that the decision boundaries–or regions of input space partitioning inputs into different classes based on model outputs–of even very complex, non-linear classifiers are often linear within small regions of space.[2] An example of a complex model's locally linear behavior and the learned linear approximation in the local linear region are shown in the figure below:



In this figure, the bold red cross represents the point of interest being explained, while the other red crosses and blue dots represent predictions of different classes made by the model. The axes correspond to two features of the input data, the red and blue regions correspond to the subspaces of the data for which the model's predictions are negative and positive respectively, and the bold red cross represents the prediction being explained. Samples' weights are shown by size of crosses/dots. The linear approximation obtained through LIME for the point of interest being explained is represented by the dashed line.

[1] For a more technical discussion of how LIME works, see Section 5.2.3.

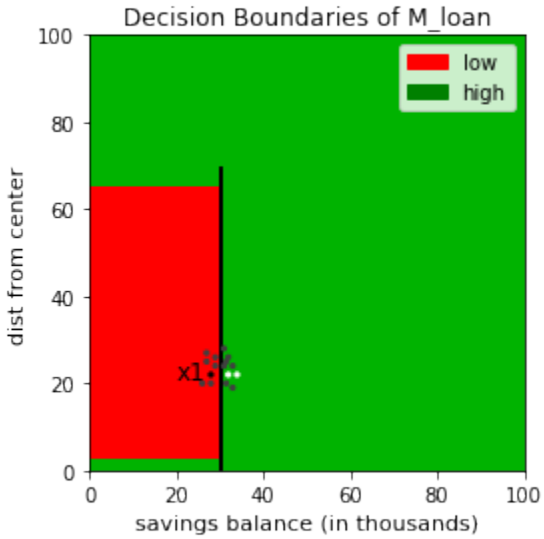
[2] For a more technical discussion of the local linearity of non-linear models' decision boundaries, see Section 5.1.

The linear approximation in this figure captures the patterns of counterfactual dependence of model outputs on input features. In other words, it is invariant over interventions, some of which are testing interventions. The alterations that are testing interventions in this image are those that consist of alterations to an individual feature that result in a change in model prediction. And we can see that the linear model is invariant over at least some of these testing interventions: There are testing interventions consisting of slight additions to the horizontal position of the red cross that would, according to the linear model, result in a different prediction (blue); these testing interventions also result in an actual change in model prediction. The linear model would similarly be invariant across interventions to the vertical position of the red cross–namely, those consisting of subtractions from the vertical position. Though there are also some testing interventions for which the linear approximation is *not* invariant, recall that Woodward's requirement is that there be only *some* interventions for which it is, not that the linear approximation be invariant over *all* testing interventions.

I claim that the above approximation's invariance over testing interventions is not coincidental: Rather, invariance over testing interventions is in fact built into linear approximations derived from sampling-based linear approximations like LIME. This is because such linear models are fit to points near a model's decision boundary, which are just those alterations which lead to a change in model's output.

Let us return to the example of `M_loan` to see how a linear approximation derived through sampling would be invariant over testing interventions, l. Depicted below are the equation and graphical depiction of the linear model explanation that would be returned for the prediction `y1`, where `sign()` represents the function that returns `green if > 0` and `red if <= 0`:.

```
y = sign(savings_account_balance - 30)
```



In the figure, the white dots represent the sampled points represent testing interventions, since they are changes to a single feature (specifically `dist_from_center`) that, according to the linear model, are expected to produce a change in model output. (Graphically, we can see that the linear model predicts changes in model output for these white dots since they are on the right side of the line.) My point here is that the white sampled points are points that the linear model was *designed* to produce the correct predictions for, since they were points that the linear model was fit to. In other words, the linear model was produced in a way that prioritized invariance over testing interventions (the white points). And if it is true that built into the process of fitting a linear model with a sampling-based method is an implicit prioritization of invariance over testing interventions, linear models learned through sampling can count as true causal explanations on Woodward's account of causal explanation.

The upweighting of testing interventions in linear approximations derived through LIME is implicit: Because points are weighted by distance to the original point, samples which only modify individual features, rather than ones that modify combinations of features, will have higher weights, since they will generally be 'closer' to the original instance than the result of an alteration to that feature and another. However, this upweighting

need not be implicit: The invariance of sampling-based linear approximations over testing interventions can increased through the *explicit* upweighting of testing interventions. This is only one of a few ways in which linear approximations derived through sampling can be improved upon to create better linear model explanations of model predictions. In Section 3.5, I offer some additional context-dependent considerations for creating better linear model explanations. I now turn argue for additional reasons that sampling-based linear approximations are effective explanations of model predictions. For the sake of simplicity, when I refer to linear approximations throughout the remainder of this chapter, I will be referring to approximations derived through the sampling method just described.

3.2 Comprehensibility

There is an additional reason why linear approximations can be useful explanatory generalizations: Their weights encode information about features' causal influences on model predictions and thus give rise to natural interpretations of the causal importance of various features (when they are applied to normalized feature values such that units of change across features are considered equal). A positive coefficient for a feature can be interpreted to mean that as the value of that feature increases, the model's prediction has a higher likelihood of being positive; conversely, a negative coefficient can be interpreted to mean a higher likelihood of being negative. Thus, the direction of causal influence of a feature can be inferred from its coefficient. Additionally, weights can be used to compare the relative strengths of different features: The higher a weight's magnitude, the stronger its causal influence.

The following would be an example of a linear model that defines a prediction of low/high risk of pneumonia (where `sign()` represents the function that returns `low if < 0` and `high if > 0`):

```
y = sign((4) * number_days_coughing + (-6) * has_asthma + (3) * is_smoker
+ (-2) * average_hours_slept)
```

In this linear model, `has_asthma` figures into causing the prediction to be low risk, given that `has_asthma` has a negative coefficient. Additionally, we can conclude that `has_asthma` has a stronger causal effect than `is_smoker` on the model prediction, since `has_asthma` has a coefficient with a larger magnitude (6) than does `is_smoker` (3).

Furthermore, though the kind of causal significance that can be derived from linear models' coefficients differs from the notion of causal significance discussed earlier, which defined the causal significance of a feature in terms of the minimal change on it required to change model outcome, significance defined in terms of coefficients very closely tracks the kind of significance defined in terms of minimal changes. If a linear model is invariant over testing interventions on features, which I will argue they are designed to be, then the feature with the minimal intervention that will result in a different model output is just the feature with the highest coefficient. This is because a change to the feature with the highest coefficient will make the largest mathematical difference and thus have the largest effect on the output of the linear model (and thus the output of the model, since we are assuming invariance).

To make this clear, take the linear model above to be the actual logic of the model, and suppose the model receives as input `x = [number_days_coughing = 12, has_asthma = 1, is_smoker = 1, average_hours_slept = 9]` and predicts `y = low`. (Again, we are assuming that the linear model acts on normalized feature values so that its coefficients are comparable across features; assume in this sake that the mean across the training data for each of the continuous features were: `[number_days_coughing: 10, average_hours_slept = 7]`; this means that the linear model is applied to the input x/mean, or `[savings_balance_thousands = 1.2, has_asthma = 1, is_smoker = 1, average_hours_slept = 1.286]`).

In this case, the minimal changes which could result in a changed model prediction for each of the features (with normalized versions on the right) would be:

`number_days_coughing`

- $12 \rightarrow 14.0, \quad 1.2 \rightarrow 1.4$

`has_asthma`

- $1 \rightarrow 0.87, \quad 1 \rightarrow 0.87$

`average_hours_slept`

- $9 \rightarrow 6.5, \quad 1.29 \rightarrow 0.93$

For this input, using the minimal changes derived from the coefficients of the linear model would get us the result that `has_asthma` had the most significant causal influence on the model prediction. This result tracks onto the result we can derive from the coefficients of the linear model, since `has_asthma` has the coefficient with the largest magnitude.

There *is* a limitation to the parallel I have just drawn between causal significance defined in terms of linear model coefficients and significance defined in terms of minimal changes. These notions only align if the notion of what defines *minimal* is a purely mathematical one. If we took into account what it means for an intervention to be possible, the minimal change to `has_asthma` that would change the model prediction would be $1 \rightarrow 0$, since having asthma is a binary feature. On this notion of minimal, influenced by real world possibility, `number_days_coughing` would be the most minimal change. If we were interested in deriving this causal insight, about `number_days_coughing`'s causal significance, merely comparing its linear coefficient with the coefficients over other features would not suffice.

More generally, what this example highlights is that the coefficients of features in linear models do not encode the same causal influences as do the minimal changes on those features when the distance metric determining minimality does not take into account real-world possibility. In cases where it is a feature's causal influence on this latter notion of causal influence that is desired, investigation beyond a comparison of coefficients is necessary.

However, coefficients of features in linear models do encode causal influences defined

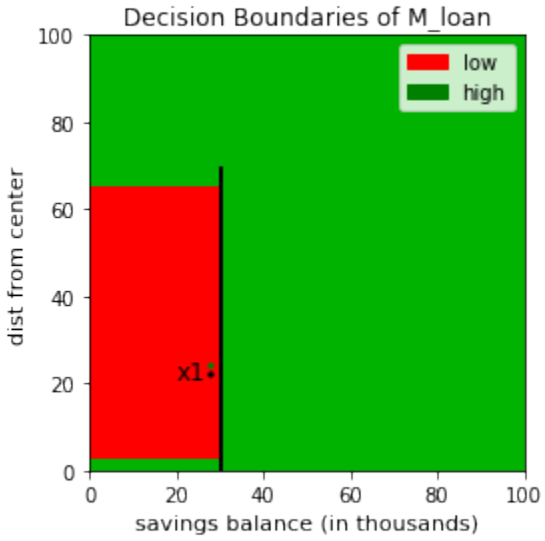
in terms of minimal changes when what is minimal is determined purely numerically. Furthermore, when what is of interest is how features causally influence model predictions over a range of interventions, and not just the minimal one, linear model coefficients encode the desired information and are easy to interpret. Thus, it is a general feature linear generalizations that they make the causal significance of different features easily comprehensible.

3.3 Addressing Causal Irregularities and Overdetermination through Aggregation

Not only can linear approximations derived through sampling be easily comprehended and function as true causal explanations according to Woodward's framework, I argue that they are good explanations because they can handle the failures of causal claims derived from single interventions. That is, they can derive the appropriate causal relationships between features and model outputs even in cases of causal irregularities and overdetermination.

In the earlier discussion of how interventions could be extended to derive the appropriate causal relationships in these cases, I established that considering ranges of interventions on features, including interventions on multiple features, would give enough information to lead to the appropriate insight, though we would need some way of aggregating information across those different interventions. My claim is that LIME linear explanations aggregate this information through the use of samples. Specifically, the sampled points that LIME linear regressions are fit to samples that represent the different sorts of interventions we would need to consider in cases of overdetermination–joint interventions–and causal irregularities–ranges of interventions, and their coefficients aggregate the causal relationships that hold across these interventions.

Causal Irregularities. Let's return again to the case of causal irregularity, which posed an issue for merely using Woodwardian interventions on single features. My argument there was that using a single intervention like Woodward requires to establish a causal relationship would lead us to conclude that `dist_from_center` causally influenced the prediction `y1` made for `x1`, even though our intuition would be that a good explanation of `y1` would *not* cite `dist_from_center` as a causally influential feature because the posited causal relationship between `dist_from_center` and `y1` is not regular across different interventions on `dist_from_center`. Thus, I concluded earlier, we need to consider ranges of interventions in order to draw regular causal claims.



The linear model explanation of `y1` would, in fact, not cite `dist_from_center` as a causally relevant feature, given that it would be fit to many points sampled around `x1`, the majority (all but one) of which would not model this dependence. More specifically, even though the irregularity (represented by the green point) reflects a counterfactual dependence that an increase in `dist_from_center` pushes the model prediction toward `low`, other sampled points that contained a slightly smaller or larger increase to `dist_from_center` (representing a range of other interventions to `dist_from_center` would *not* show this dependence, and so the linear model would reflect the majority of

these other points; at a maximum, the linear model depicted below would very slightly learn toward the left because of the single outlying point. By being designed to fit many samples (and thereby many different interventions), linear model explanations implicitly aggregate across interventions and thus do not encode causal relationships that are not robust.

Overdetermination. Linear explanations are also able to encode the correct causal influences of features in cases of overdetermination through samples. In the discussion earlier, I established that joint interventions on multiple features are needed to highlight the causal influences of causes when multiple sufficient causes are involved. My argument here is just that such joint interventions comprise some of the samples which LIME linear explanations are fit to.

Thus, in addition to the ranges of single-feature interventions which allow linear explanations to encode only somewhat robust causal relationships, joint interventions on multiple features also comprise the samples which LIME linear explanations are fit to.

3.3.1 Causal Robustness as a Feature, as a Bug

So far, I have assumed that the fact that sampling-derived linear approximations only encode robust causal relationships is a strength. I have made this assumption both in the context of arguing that their coefficients, which aggregate the causal influence of features across samples, can be interpreted as a measure of causal strength (Section 3.2) and in the context of arguing that the fact that they do not point out irregular causal relationships (Section 3.3) is a strength. In this section, I expand upon my argument for why it is in most cases desirable that linear approximations do not highlight causal irregularities. I also discuss the few explanatory contexts in which the fact that linear approximations encode only robust causal relationships is a weakness.

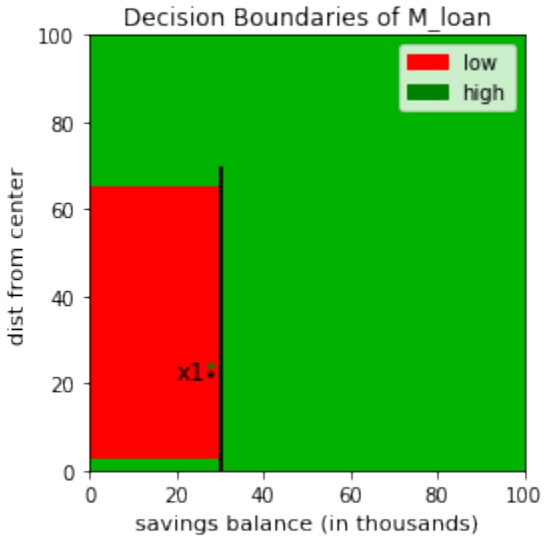
There are a few reasons why this is true. Firstly, as highlighted by the example of the rare, stomach-cancer-causing strawberries described in Section 2.2.2, we do not draw

causal claims about all variables in the real world that have mere non-zero chances of exhibiting a causal dependence. And if we do not draw such causal claims about real-world variables, why would we draw such claims about features and model predictions? We would not want our explanations to cite all features that have any chance of affecting model predictions as being causally relevant.

In addition to real-world causal irregularities, another reason for excluding causal claims about irregular causal relationships from our explanations of model predictions is that there exist imprecisions in the data models are trained on. For instance, values for the feature `dist_from_center` may be calculated from different city center points each time. Because of such imprecisions, the data itself may not contain completely robust causal relationships, and it is to be expected that models trained on such data may learn causal irregularities.

When we are using explanations of model predictions to determine whether the causal dependences of the model prediction on features reflects what real-world dependences are, irregular causal relationships arising from irregularities and imprecisions in the real world do not seem most of interest. Rather, it is regular causal relationships that are critical to investigate. For instance, suppose a doctor is trying to determine whether a given prediction made by a stomach cancer diagnosis model was causally influenced by features in a medically sound way. If a causal relationship between (a feature representing) patients' eating strawberries and model's predictions held across only very few interventions, this finding would be irrelevant for the doctor's decision of whether to trust the model prediction, given that it might actually reflect the rare causal irregularities that exist in the real world. If, on the other hand, that relationship held across many interventions on (the feature representing) patients' eating strawberries, the doctor would likely *not* want to trust the model. Thus, to determine whether a model prediction was reached in a medically sound way, the doctor would want to receive an explanation which only cites information about somewhat robust causal relationships between features and predictions.

It is also critical that our explanations of model predictions only encode robust causal claims when these causal claims are being used to decide future actions, since causal claims which hold across different interventions on features will lead to actions that more robustly will produce intended effects. For instance, if the person represented by `x1` in the figure below were seeking an explanation to figure out how to change the model's prediction in the future in order to receive a loan, she would want to know what features to change. This person would likely *not* want to receive an explanation that cited `dist_from_center` as having a strong causal influence on `y1`, since only one change in value of `dist_from_center` would result in a change to `y1`; taking actions based on an explanation that cited `dist_from_center` as having a causal influence on `y1` might mistakenly lead the person to think she should prioritize moving further from the city center, when really the best course of action for her would be to prioritize saving a bit more money. Thus, causal claims being used to act should encode robust causal relationships, since only these will only continue to hold over different possible actions.



In both of these cases, including information about causal irregularities would actually shroud understanding. This is because there are many irregular causal relationships, only holding over very few interventions, that can exist between features and model predic-

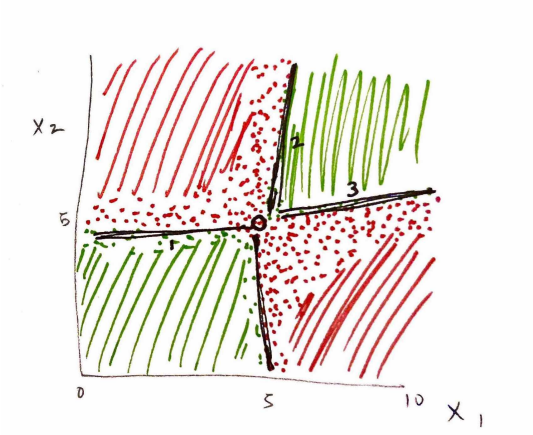
tions. It is this fact that makes linear approximations' limitation to only causally robust relationships a *feature*.

However, there are some cases where we in fact would want our explanations of model predictions to give insight into even very irregular causal dependences between features and output. In a context where an explanation is being used to determine whether a prediction was influenced by a protected attribute, it seems we *would* want information. For instance, suppose a model predicting recidivism likelihood gave a prediction of `high` to a black defendant. The existence of *any* prediction-changing intervention on race would suggest that there is some causal dependence of the model's prediction of recidivism likelihood on the feature race. In such a setting, even if there were only one such intervention–i.e. changing race from *black* to *white*–and all others–such as changing race from *black* to *asian*–we would want an explanation of the model prediction to point out the causal dependence of the model's prediction on the protected attribute. It is thus necessary to consider facts about the context in which an explanation is being used when determining how considerations about causal robustness should influence what features are explained as being causally relevant, and thus whether linear approximations will function as appropriate explanations.

3.4 Limitations of Linear Model Explanations

3.4.1 Non-Linear Model Effects

What are limitations of linear models derived from LIME as explanations of model predictions? Single local linear models do not function well as explanations for instances which lie very close to a nonlinear model effect, since slight interventions to the feature values of the instance might result in a new instance that lies in a part of the input space for which the model being explained uses very different logic. For instance, in the figure below, the circled instance is very close to a nonlinear model effect, where the model's decision function is characterized by a flip in signs.



There are four possible linear models that can be used to explain the circled point, each of which are numerically equally close to the instance. The following four linear models are candidates for the four lines pictured, where `sign()` represents the function that returns `green if < 0` and `red if > 0`:

1. `p = sign(-0.01 * x1 - x2 + 5)`
2. `p = sign(x1 - 0.1 * x2 - 5)`
3. `p = sign(-0.25 * x1 + x2 - 5)`
4. `p = sign(-x1 - 0.1 * x2 + 5)`

Each of these linear models has very different coefficients and thus encodes very different causal claims. In equations 1 and 3, `x2` has a coefficient with a much larger magnitude (1) than the coefficient of `x1` (0.01 and 0.25), suggesting that the causal influence of `x2` is much higher than the causal influence of `x1`–This is an observation which can also be seen in the graphical representations of the lines. In contrast, equations 2 and 4 give a larger coefficient, and thus more causal influence, to `x1` than to `x2`. The signs of the coefficients also vary across the linear models, suggesting that features have different directions of causal influence in some than in others. For instance, `x2` has a negative coefficient in equation 1, meaning that an increase in `x2` leads to an increased likelihood of a prediction of `red`; in contrast, `x2` has a positive coefficient in equation 3, meaning that an increase in `x2` leads to a decreased likelihood of a prediction of `red` and an increased likelihood of a prediction of `green`.

In cases such as this one, where a model's decision boundary around an instance being explained has nonlinear effects, a single linear model does not function as an explanation. We are thus stuck with the puzzle of figuring out what the proper explanation of a prediction such as `y3` should be, given that there appear to be multiple qualifying explanations.

3.4.2 Choosing Between Multiple Linear Models

It is important to note that the puzzle only arises if we are interested in drawing robust causal claims. There are facts of the matter about specific interventions on features `x1` and `x2`. However, it is indeterminate what general causal relationship holds between `x1`/`x2` and the model output in the local vicinity of the point, given that different interventions will reflect causal dependences with seemingly opposite directions. And it is this indeterminacy that gives rise to the puzzle of choosing between candidate linear models.

My proposal for how we should address the limitations of linear models in cases where a model's decision boundary has a non-linear effect is a two-fold one: We can and should in fact return multiple linear models as explanations in such cases, since the existence of multiple linear models that encode very different causal claims offers the insight that the local causal relationships between features and model predictions are indeterminate and will vary greatly over different interventions.

But in cases where we want to pick a single linear explanation of multiple candidate linear explanations, Woodward provides the resources in his theory for doing so. Specifically, we can build upon Woodward's idea that generalizations have different amounts of explanatory power depending on what changes they are invariant over. We can choose linear models which are invariant over certain desired changes, which are determined contextually. More specifically, what a linear model explanation will be used for should determine which changes to features it should be invariant across. When an explanation is being used to determine future actions based on a prediction, invariance across features

which are manipulable and likely to occur in the real world should be prioritized. When an explanation is being used to determine whether a prediction was influenced by a protected attribute, it should be invariant across changes to that protected attribute. And lastly, when an explanation is being used to determine whether a prediction was influenced by meaningless artifacts in the training data, invariance across the specific features that are suspected to be involved in meaningless artifacts should be prioritized.

Each of the four linear models `p1`, `p2`, `p3`, and `p4` count as true causal explanations because they are invariant over the testing interventions discussed earlier, but they are each invariant over different sorts of interventions: Models `p1` and `p3` are invariant specifically over interventions to `x2`, while models `p2` and `p4` are invariant over interventions to `x1`. When we are choosing between these models, we can choose based on which kind of invariance matters more for the current use. For instance, suppose `x1` represented a feature that was difficult to change in the real world, such as `savings_account_balance`. If `x2` on the other hand represented a more actionable feature, such as `dist_from_center`, models `p2` and `p4` might be preferred. We could further choose between models `p2` and `p4` by considering what ranges of interventions on `x2` (dist_from_center) were more likely: Suppose it was particularly difficult to move to one region of the city because of a shortage of homes available in that region, and suppose this region were closer to the city center than the individual in question currently lived. Then `p2` would be the preferred explanation, since it describes the causal relationships between `x2` and model outcome for actionable interventions–moving farther away from the city.

3.5 Chapter Summary

In conclusion, I have argued that linear approximations derived through sampling methods are effective local causal explanations of model predictions because they are implicitly locally invariant (and can be made to be explicitly so), comprehensible, and can be used to derive the desired causal claims in the cases of overdetermination and causal irregularities

because they encode robust causal claims. Though the robust causal insights that we get from linear approximations are useful in many contexts, I have also considered a few contexts in which interventions might be preferred for their ability to get insight into irregular causal relationships.

I have also suggested a few ways in which such methods can generate better explanations, which can be implemented through the weighting of sampled points. Firstly, as discussed in Section 3.1, locally sampled points representing testing interventions should be explicitly upweighted; furthermore, testing interventions which involve smaller alterations can be prioritized in this upweighting so that the causal claims linear approximations encode will be influenced by the notion of causal significance defined in terms of minimal changes. Secondly, we can improve upon the distance metric that is used to calculate weights of sampled points to incorporate contextual considerations, such as the real world feasibility of changes. For instance, if an explanation is being used to determine future actions, invariance over features which are actionable should be prioritized through an upweighting of sampled points with changes to those actionable features.

Part II

Linear Approximations for Actionable Insight into Blackbox Classifiers

Chapter 4

Obtaining Actionable Insight into Linear Models through Counterfactuals

In this chapter, I will show how linear approximations can be used to generate counterfactuals and thus gain actionable insight into model predictions. I introduce previous methods of generating counterfactuals for both non-linear and linear classifiers, which I will draw upon in Chapter 5.

4.1 An Introduction to Counterfactuals for Actionable Insight

One line of research in explainable machine learning has focused on selecting informative individual instances to give insight into the behavior of machine learning models. One primary motivation for such methods is that they give insight into the kinds of *actions* someone can take in order to obtain a different outcome.

Counterfactual explanations are one form of example-based explanations. A *counterfactual explanation* of an instance x and model prediction $f(x)$ refers to a set of minimal changes that can be made to x in order to change the model prediction. It describes the dependencies between features' values and model ouputs. For instance, suppose a loan applicant is denied a bank loan. A counterfactual explanation would look something like: *If you were one year older and were applying for $5,000 less in loan amount, you would*

have been approved for a loan. With this information, the loan applicant could apply for a smaller amount in loan when re-applying for a bank loan in the future. By enumerating the smallest changes to features that will bring about a different model prediction, counterfactual explanations can give practical, actionable insight into classifiers.

Suppose we are given a model f and an instance x, and we wish to find a counterfactual explanation for the point $f(x)$. In other words, we want to find the closest point x' that will bring about a different prediction, represented by y'. To find the desired x', some previous approaches have centered on minimizing various loss functions.

Wachter et al. (2017) propose minimizing the following loss:

$$L(x, x', y', \lambda) = \lambda \cdot \hat{f(x') - y')^2} + d(x, x')$$

$d(x, x')$ is a customizable function measuring the distance between x and x', for which they propose using an L_1 norm weighted by the inverse median absolute deviation of each feature k (MAD_k) in the set of training points P:

$$\text{MAD}_k = \text{median}_{j \in P}(|x_{j,k} - \text{median}_{l \in P}(x_{l,k})|)$$

$$d(x, x') = \sum_{k \in F} \frac{|x_k - x'_k|}{\text{MAD}_k}$$

Laugel et al. (2017) offer an additional approach, the Growing Spheres algorithm, which finds x' such that $f(x') = y'$ with a generative approach: They draw increasingly larger spheres around x and sample from those spheres until the model prediction for one of those sampled points is y'.

There are a couple of limitations in using either of these approaches to gain actionable insight into model predictions. The first is that the counterfactuals they return are not specifically constrained to only modify *actionable* features; a result, some counterfactuals might involve changes to features that cannot be changed with actions in the real world. For instance, a counterfactual that returned a change to a person's marital status or age

would not give practical insight into how a loan applicant denied a loan could reapply and be approved for a loan in the future. The second limitation is that a lack of actionable counterfactual returned by the methods of Wachter et al. (2017) and Laugel et al. (2017) do not guarantee that such actionable counterfactuals do not exist.

These limitations are what Ustun et al. (2019) seek to define in the optimization framework they propose for calculating what they call *recourse*, or the ability to change a decision of a model. I will describe their framework in the following section.

4.2 Generating Flipsets for Linear Classifiers

Motivated by the need for people to be able to gain actionable insight into model predictions and the limitations of previous methods in addressing this need, Ustun et al. (2019) propose an integer programming toolkit to measure the *recourse*, or person's ability to obtain a specific outcome from a pre-trained model, of pre-trained linear classifiers. They formulate an optimization problem that, given an instance x and an undesired outcome, searches over a grid of possible actions on actionable features and generates a *flipset*, or list of actionable changes to features that will lead to the desired outcome.

4.2.1 Optimization Framework

Here, I present their optimization framework and the integer program they use to solve a discretized version of the optimization problem.

Ustun et al. (2019) assume as inputs the following: a feature vector $x = [x_1, x_2, ..., x_d] \subseteq \mathbb{R}^{d+1}$; a linear classifier $f(x) = 1[\langle w, x \rangle \geq 0]$, where $w = [w_0, w_1, ..., w_d] \subset \mathbb{R}^{d+1}$ is a vector of coefficients and intercept (w_0); and a binary label $y \in \{-1, 1\}$, where $y = 1$ is assumed to be the desired outcome.

Given an instance with an undesired model prediction $f(x) = -1$, they form the following optimization problem to find an action a such that $f(x + a) = +1$:

$$\begin{aligned} \min \quad & \text{cost}(a; x) \\ \text{s.t.} \quad & f(x + a) = +1, \\ & a \in A(x) \end{aligned}$$

Here, $A(x)$ represents a set of feasible actions, where each action a is a vector $a = [0, a_1, ..., a_d]$ and each a_j is such that $x_j + a_j$ is a feature value that occurs in the training set. $cost(a, x)$ represents a function which can be used to prioritize some actions over others. Ustun et al. (2019) assume the following properties about the cost function: (i) i.e. no action incurs a cost of 0; (ii) actions that are larger have higher costs.

Ustun et al. (2019) note that solving this optimization problem gives the following guarantees: If the problem is *feasible* and a solution is returned, the optimal action $a*$ to change an undesired model outcome to a desired one is just the minimal-cost action returned by the framework. If it is *infeasible*, then the individual with features x cannot take any actions to change the received undesired outcome.

Ustun et al. (2019) formulate this optimization problem as an integer program (IP) such that it can be optimized with a solver[1]. Ustun et al. (2019) note that this IP formulation is desirable because it can search over different types of features (binary, ordinal, and categorical), can optimize a range of cost functions, and has a customizable action space, which can be used to limit the returned counterfactuals to involve only manipulations to actionable features. However, a limitation of the method proposed by Ustun et al. (2019) is that it is only applicable to *linear* classifiers. In the next chapter, I propose an extension of their framework to non-linear classifiers.

[1]For the specific integer program formulation, see the original paper by Ustun et al. (2019)

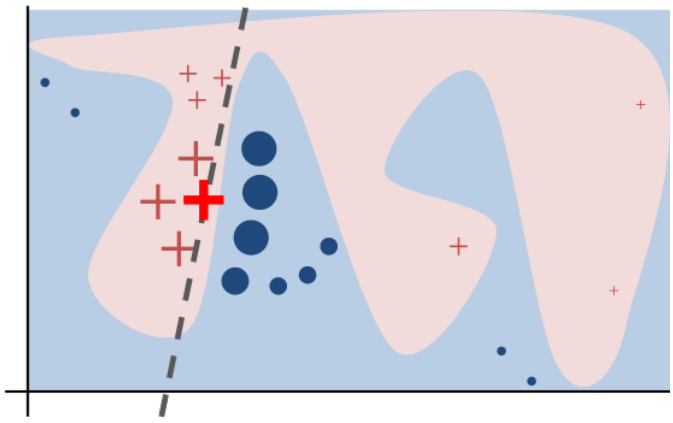
Chapter 5

Generating Counterfactuals for Non-linear Classifiers Using Linear Approximations

In this chapter, I propose an end-to-end approach for generating counterfactuals for non-linear classifiers. This framework uses local linear approximations of non-linear complex classifiers as input into the framework provided by Ustun et al. (2019) in order to measure classifiers' recourse and provide flipsets for non-linear models. The novelty in this approach consists of exploiting models' local linearity to be able to generate flipsets, or actionable counterfactuals, with the approach described in the previous section. I experiment with existing methods of creating linear approximations, LIME (Ribeiro et al., 2016) and MAPLE (Plumb et al., 2018) and show how they can be extended to produce counterfactuals. I call these extensions **Counterfactual-MAPLE** and **Counterfactual-LIME**. Furthermore, I propose a novel decision-tree-based approach called **Counterfactuals using Tree-based Local Neighborhoods**, or CoTLoN, which approximates non-linear classifiers with different neighborhoods of points in order to generate flipsets for individuals who receive unfavorable predictions.

5.1 Experimental Motivation: Non-Linear Classifiers are Locally Linear

Much previous work on creating local explanations of complex classifiers has made use of the idea of fitting simpler models in small neighborhoods (Ribeiro et al. (2018)). In particular, many existing methods exploit the local linearity of non-linear classifiers to create local explanations of model predictions (Ribeiro et al. (2016); Plumb et al. (2018)). The premise behind such work is that in small enough regions of a complex decision boundary, the decision boundary is linear. The following image created by Ribeiro et al. (2016) provides an example of this phenomenon:



Furthermore, for a particular class of non-linear classifiers–neural networks with piecewise activation functions–there has been previous work supporting this claim of local linearity. Hanin and Rolnick (2019) studied the *expressivity* of neural networks, as measured by their number of distinct linear regions, and found that the number of linear regions in deep neural networks with piecewise linear activations (such as `ReLU` and hard `tanh`) is far below its theoretical capacity. They mathematically prove that the average number of linear regions along one-dimensional partitions of neural networks at initialization grows linearly in the total number of neurons, rather than exponentially in depth. They also empirically validate that this result holds also for networks during training–The number of linear regions stays roughly constant. Additional work from Goodfellow et al. (2014) has

shown that it is neural networks' linear behavior in high-dimensional spaces that makes them vulnerable to various adversarial perturbations.

Because of non-linear classifiers' local linearity, it is possible to measure recourse and generate flipsets for predictions made by their predictions using the approach of Ustun et al. (2019). In what follows, I both exploit and validate this proposal experimentally.

5.2 Experimental Set-Up

The experiments I ran were intended to answer the following questions: Can linear approximations be used in conjunction with the framework for flipset generation proposed by Ustun et al. (2019) to generate flipsets for non-linear classifiers as well? The experiments were also designed to compare how well different methods of deriving linear approximations function for the purpose of returning flipsets for non-linear classifiers.

Let f represent a non-linear classifier trained on a dataset X, and let $\tilde{X}$ represent a set of instances for which the model gave a negative outcome. Let L represent a linear approximation method. The experiments I ran took the following form: For each instance $\tilde{x} \in \tilde{X}$ which received an undesired outcome of -1 from f (i.e. $f(\tilde{x}) = -1$), I generated a set of coefficients $w = [w_1, ..., w_d]$ and intercept w_0 using L. Then, I used the approximated linear coefficients w as input into the optimization framework of Ustun et al. (2019) to generate flipsets for the instance $\tilde{x}$. A maximum of 20 possible flipsets were able to be returned for each $\tilde{x}$. I measured both the accuracy of these flipsets and the portion of instances in $\tilde{X}$ for which flipsets were returned.

I experimented with model type (f), dataset (X), and linear approximation method L. Specifically, f was one of three different model types: a neural network, random forest classifier, and gradient boosted tree. X was one of two datasets: `COMPAS`, a dataset used for recidivism risk prediction, and `Adult`, an income prediction dataset based on census data. L was one of four main linear approximation methods–a baseline, a cluster-based approach, LIME (Ribeiro et al., 2016), and MAPLE (Plumb et al., 2018)–with some

variations. Each of the models, datasets, and linear approximation methods are described below.

Lastly, I ran 5 experiments for each combination of model, dataset, and linear approximation method. In Section 5.3, I report metrics averaged across these experiments.

5.2.1 Datasets

COMPAS. The Correctional Offender Management Profiling for Alternative Sanctions (`COMPAS`) is a dataset that was collected by Propublica (2016) and contains information about criminal defendants' likelihood of reoffending their crime. `COMPAS` includes features representing defendants' demographic information and information about their crime, and the target variable in this dataset is the score as output by the `COMPAS` algorithm. I used a processed version of this dataset that contained $6,907$ instances, each with $d = 7$ features, 3 of which were continuous and 4 of which were categorical. 0.5 of this dataset was used for training, and the rest was used for validation and flipset generation.

Adult. The `Adult` dataset from the UCI repository is a credit evaluation dataset, for which the target variable is an applicant's credit rating (Dua and Graff, 2017). This dataset includes features about an applicant's demographics, credit history, employment, and financial information. I used a processed version of this dataset that contained $d = 10$ features, 5 of which were categorical. This processed dataset contained $32,560$ examples.

5.2.2 Models

Three different types of models were trained and used as blackbox classifiers in the following experiments: neural networks, random forests, and gradient boosted trees. Each of the models were implemented using `sklearn`.

Neural Network. A multi-layer perceptron was trained using `sklearn`'s implementation of `MLPClassifier()`. This classifier had one hidden layer with 100 neurons and `relu`

activations.

Random Forest. A random forest classifier with 100 estimators was trained using `sklearn`'s implementation of `RandomForestClassifier()`.

Gradient Boosted Tree. A gradient boosted classifier with 100 estimators was trained using `sklearn`'s `GradientBoostingClassifier`.

5.2.3 Linear Approximation Methods

I now describe the linear approximations methods I experimented with. These methods differ not only in how the approximated linear coefficients are derived, but also in the number of unique sets of coefficients which are derived across all $\tilde{x} \in \tilde{X}$. Specifically, the baseline model derives a single set of coefficients shared by all $\tilde{x} \in \tilde{X}$, while the TSLC approach derives a pre-determined set of coefficients shared by $\tilde{X}$. The Counterfactual-LIME and Counterfactual-MAPLE approaches use LIME and MAPLE approximations respectively to derive a unique set of coefficients for each $\tilde{x}$.

Baseline. As a baseline, a single linear model was used to approximate the predictions made by the blackbox model f on X_{train}. The idea behind this baseline model is to treat the non-linear classifier as a single linear model. Using this approach, the same coefficients are used to calculate recourse for each $\tilde{x} \in \tilde{X}$. `sklearn`'s `Ridge()` was used to implement this baseline linear approximation. Specifically, the ridge regression model was fit to the probability estimates given by f for the favorable outcome $+1$.

CoTLoN. Counterfactuals using Tree-based Local Neighborhoods, or CoTLoN, generates a single linear approximation for different neighborhoods of points, which are determined through the use of a decision tree. To generate these neighborhoods, I trained a decision tree on the predictions made by f on X_{train}. Then, I trained a distinct lin-

ear approximation for each leaf node. All points reaching the same leaf node form a "neighborhood" of points.

I used `sklearn`'s `DecisionTreeRegressor()` to implement the decision tree and `sklearn`'s `Ridge()` to implement the linear model learned at each leaf node of the learned decision tree. Using this approach, the same linear coefficients are used for each cluster when calculating recourse for each $\tilde{x}$.

I determined the number of neighborhoods (and thus linear approximations) by setting the parameter `min_samples_leaf` given to `sklearn`'s `DecisionTreeRegressor()`, which determines the minimum number of samples at each leaf node of the learned decision tree. Specifically, I determined the value of `min_samples_leaf` through a fraction n_r of the number of training instances: `min_samples_leaf` $= n_r * len(X_{train})$, where n_r was one of [0.05, 0.1, 0.15].

Counterfactual-LIME. The Counterfactual-LIME approach to generating flipsets makes use of Local Interpretable Model-agnostic Explanations, or LIME, a method proposed by Ribeiro et al. (2016). LIME makes use of local sampling around an instance x to create a linear approximation of a classifier f around x. More specifically, given an instance x for which the prediction $f(x)$ is being explained, LIME samples $N = 5,000$ points in the following way: For continuous features, values of the sampled points are drawn from a normal distribution with features' mean and standard deviation across the training set. For categorical features, values are sampled with probabilities determined by the frequency with which each category appears in the training set. LIME then fits a weighted linear regression model to the probability estimates output by f for those sampled points. The weights are determined with the following exponential smoothing kernel:

$$\pi_x(z) = \sqrt{exp(-D(x,z)^2/\sigma^2)}$$

$D(x,z)$ represents the distance between points x and z. LIME's default distance metric is a Euclidean distance metric: $D(x,z) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_d - z_d)^2}$. σ

represents the kernel width, which LIME by default calculates as $k = 0.75$ times the square root of the number of features in x: $\sigma = \sqrt{d} * k$. I used the default implementation of LIME for tabular data[1], which uses `sklearn`'s implementation of `Ridge()`) as the default linear model but experimented with the value of k. k was one of [0.75, 0.25].

Using the Counterfactual-LIME approach, each $\tilde{x} \in \tilde{X}$ receives its own set of linear coefficients which are used when calculating recourses.

Counterfactual-MAPLE. The Counterfactual-MAPLE approach to generating flipsets relies on linear approximations derived through Model Agnostic suPervised Local Explanations, or MAPLE. MAPLE is a local linear modeling approach proposed by Plumb et al. (2018) that creates *supervised* neighborhoods around instances being explained. The idea behind MAPLE is to fit a tree ensemble to the training data and use this ensemble to identify the most relevant training points for a particular prediction. The default tree ensemble is a random forest with 200 estimators, implemented using `sklearn`'s `RandomForestRegressor()`. For each point x being explained, for each training point x', a similarity weight is calculated by counting the number of trees in which both x and x' occur in the same leaf node. The weight is calculated in the following way:

$$s(x, x') = \frac{1}{k} \sum_{k=1}^{K} \frac{c_k(x', x}{num_k(x)}$$

Here, $num_k(x) = \sum_{i=1}^{n} c_k(x_i, x)$ represents the number of training points in the same leaf node as x, and $c_k(x, x') = 1 \cdot \{leaf_k(x) = leaf_k(x')\}$, where $leaf_k(x)$ represents the index of the leaf node of tree k that contains x.

I used the default implementation of MAPLE.[2] Using the Counterfactual-MAPLE approach, each $\tilde{x} \in \tilde{X}$ receives its own set of linear coefficients which are used when calculating recourses.

[1] `https://github.com/marcotcr/lime/blob/master/lime/lime_tabular.py`
[2] `https://github.com/GDPlumb/MAPLE`

5.3 Results

For each experiment, I calculated the *projected flipset accuracy* returned using the linear approximation as input into the flipset generation framework. A flipset item, or action vector $[a_1, a_2, ..., a_d]$ is said to be *accurate* if it indeed leads to a flip in prediction to a desired outcome for the underlying model being approximated, i.e. $f(\tilde{x}) = -1$ but $f(\tilde{x} + a) = +1$. *Projected* flipset accuracy is a measure of what percentage of all potential flipsets that could have been returned were accurate: If a flipset was not returned, it was treated as *inaccurate* to account for the fact that some linear approximations returned highly accurate, but few, flipsets. *Projected flipset accuracy*, averaged across the 5 experiments I ran for each combination of model, dataset, and approximation method, are shown in Figure 5.1.

The main result this figure highlights is that all the examined approaches of deriving local linear approximations outperform the baseline in each experimental setting investigated: Local linear approximations lead to an increased ability to generate flipsets than the baseline approach of fitting a global linear model to the predictions made by f. This finding suggests that the end-to-end framework proposed in this thesis, of using local linear approximations to generate counterfactuals for non-linear classifiers, is a promising avenue of work. However, Figure 5.1 also highlights that there is much room for improvement on existing methods of deriving local linear approximations. In the next section, I discuss some of these avenues.

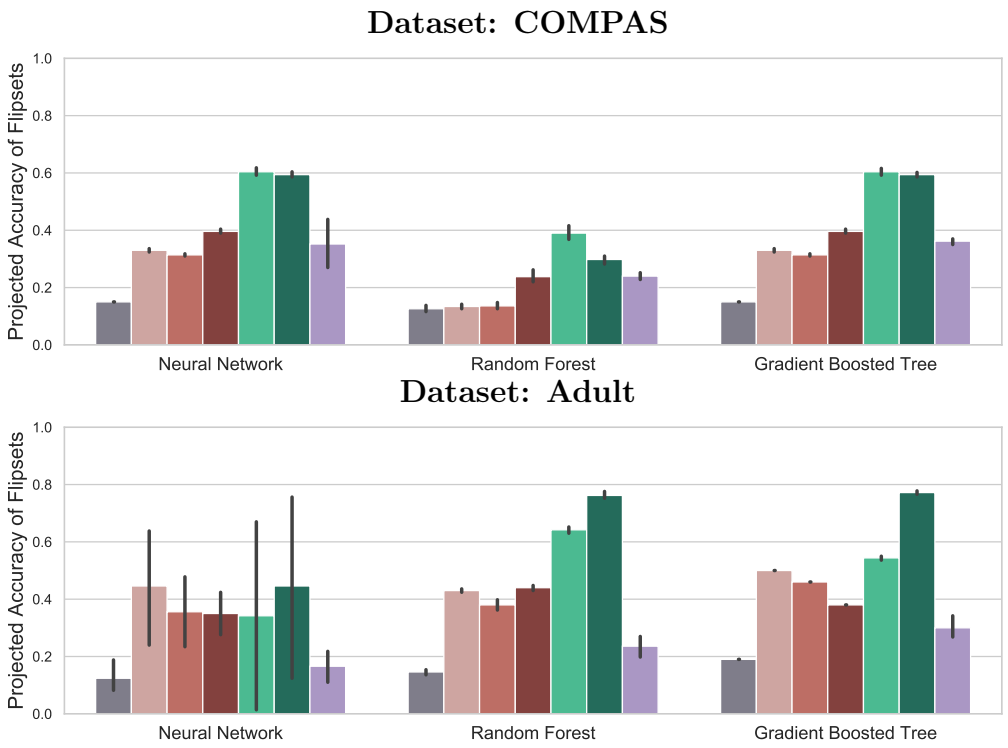


Figure 5.1: Results on the `COMPAS` and `Adult` datasets are shown for the following methods of deriving linear approximations: Baseline CoTLoN ($n_r = 0.01$) CoTLoN ($n_r = 0.05$) CoTLoN ($n_r = 0.15$) Counterfactual-LIME ($k = 0.75$) Counteractual-LIME ($k = 0.25$) Counteractual-MAPLE. Metrics averaged across 5 experimental runs in each setting are displayed.

5.4 Next Steps and Future Work

Of the linear approximation methods experimented with, Counterfactual-LIME obtained the highest projected recourse accuracies. One potential explanation for this is that it makes use of sampling, and thus linear models learned through LIME have more data to learn from.

However, there are a few inherent disadvantages of sampling-based methods: Firstly, sampling can result in points that are off-manifold and highly unrealistic. Secondly, the distance metric used to weight samples is ill-defined; there is no guarantee that the neighborhoods defined using Euclidean distance (the default distance metric used by LIME) are actually locally linear. The large effects of even slight variations in distance metrics can

be seen by the performance differences of LIME with a kernel width defined by $k = 0.75$ and LIME with a kernel width defined by $k = 0.25$ for the `Adult` dataset. There is no known way of knowing apriori which distance metric should be used for different datasets.

One avenue for future work would be to extend the non-sampling methods studied in this paper to see if the benefits of sampling–namely that it offers more data for accurate linear approximations can be learned–can be retained without the described disadvantages with the sampling used by LIME. One such way of doing this would be to extend the CoTLoN method proposed in this thesis in the following way: Use sampled points in the creation of the decision tree but not in the creation of linear approximations at each leaf node. Future experiments could test the hypothesis that such a method would lead to the generation of more accurate counterfactuals.

Bibliography

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. How we analyzed the compas recidivism algorithm. *Propublica*, 2016.

G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med*, 9(2):107–138, Feb 1997.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

John Dupré. Probabilistic causality emancipated. 1984.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. *ArXiv*, abs/1901.09021, 2019.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning, 2017.

David Lewis. Causation as influence. *The Journal of Philosophy*, 97:182–197, 04 2000. doi: 10.2307/2678389.

Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

A. M. Neill, I. R. Martin, R. Weir, R. Anderson, A. Chereshsky, M. J. Epton, R. Jackson, M. Schousboe, C. Frampton, S. Hutton, S. T. Chambers, and G. I. Town. Community acquired pneumonia: aetiology and usefulness of severity criteria on admission. *Thorax*, 51(10):1010–1016, Oct 1996.

Gregory Plumb, Denali Molitor, and Ameet Talwalkar. Model agnostic supervised local explanations, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1527–1535. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 10–19. ACM, 2019. doi: 10.1145/3287560.3287566. URL https://doi.org/10.1145/3287560.3287566.

Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explana-

tions without opening the black box: Automated decisions and the gdpr. *ArXiv*, abs/1711.00399, 2017.

James F. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.