Landmarks in Case-based Reasoning: From Theory to Data

Wijnand VAN WOERKOM [a,1], Davide GROSSI [b,c,d], Henry PRAKKEN [a,e] and Bart VERHEIJ [b]

[a] *Department of Information and Computing Sciences, Utrecht University, The Netherlands*

[b] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands*

[c] *Amsterdam Center for Law and Economics, University of Amsterdam, The Netherlands*

[d] *Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands*

[e] *Faculty of Law, University of Groningen, The Netherlands*

Abstract. Widespread application of uninterpretable machine learning systems for sensitive purposes has spurred research into elucidating the decision making process of these systems. These efforts have their background in many different disciplines, one of which is the field of AI & law. In particular, recent works have observed that machine learning training data can be interpreted as legal cases. Under this interpretation the formalism developed to study case law, called the theory of precedential constraint, can be used to analyze the way in which machine learning systems draw on training data – or should draw on them – to make decisions. These works predominantly stay on the theoretical level, hence in the present work the formalism is evaluated on a real world dataset. Through this analysis we identify a significant new concept which we call landmark cases, and use it to characterize the types of datasets that are more or less suitable to be described by the theory.

Keywords. precedential constraint, artificial intelligence, interpretability, law, landmark cases

1. Introduction

Much present-day research is focussed on making artificial intelligence (AI) more transparent. This is partially in response to mounting concerns that uninterpretable algorithms, so-called 'black box' AI, are making high-impact decisions, such as those with legal, social, or ethical consequences, in an unfair or irresponsible manner. A prominent example of such a system is the proprietary software Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), developed by Northepointe, Inc. for automatic risk assessment of various forms of recidivism, which has seen nationwide use in the United States [1]. Allegations by [1] that COMPAS racially discriminates in its decision making

[1] Corresponding author; e-mail: w.k.vanwoerkom@uu.nl.

process have led to a host of followup research and discussions. The COMPAS developers have published a response [2] and others have pointed to flaws in the original analysis by ProPublica [3,4], but as [3] points out, regardless of the veracity of the allegations, the situation is symptomatic of the larger problem that the black box nature of these systems is obstructing independent assessment of bias.

Many different kinds of solutions have been proposed, among which those to make the AI inherently more transparent [5], to formulate appropriate regulations [6], and to monitor the systems and measure bias [7]. The line on which the present work builds is that of *post hoc* interpretability methods, in which the black-box system is analysed after it has been trained and little to no access to the way it functions is assumed [8,9,10].

There are in turn many types of post hoc explanation methods. We will focus on a particular branch originating from the intersection of AI & law, based on *case-based reasoning* (CBR). The idea of a CBR explanation of a decision is to provide an analogy between it and relevant training examples. Proponents of this approach, such as [11], argue that explanations of this form are natural to humans as we are well acquainted with reasoning by analogy, are simple in form, and draw on real evidence in the sense that in most applications the training examples serve as a gold standard that the black box adheres to. Two recent examples of this approach from AI & law are [12] and [13].

The method of [13] is based on the theory of *precedential constraint*, introduced by Horty in [14], which is a formal framework developed to describe the *a fortiori* reasoning process underlying case law, i.e., to which extent does a body of precedents constrain a decision in a new case? The key idea of [13] behind applying this theory is that the training data used by most modern machine learning systems for binary classification, which consists of rows of data for a set of features together with a binary target variable, can be interpreted as the fact situations of legal cases together with their verdicts. Using this 'training examples as cases' interpretation [13] uses the theory of precedential constraint as the theoretical foundation for building a post hoc explanation algorithm.

In this work we test the degree to which training examples can be thought of as legal cases, by instantiating the model of precedential constraint on the COMPAS dataset published in [1] and subsequently evaluating various statistics. This data is of interest to us for two reasons. First of all it concerns real world data rather than fictional data. Secondly, it is highly relevant to the concerns that drive explainable AI research, as automated decisions may be made on the basis of such data which have a big social impact. As such it is representative of the situations to which our explanation methods may be applied. For the evaluation we are interested in the *consistency* percentage, which can be thought of as the degree to which the data obeys the precedent set by other examples. Through this analysis we find that an important role is played by what we shall refer to as *landmark* cases; those cases that set a new precedent with respect to the other cases. We find that in the case of the COMPAS data, a relatively small number of these landmarks force the decision of almost all other cases.

The rest of this work is structured as follows. We begin in Section 2 by summarising the theory of precedential constraint, to the extent that we will need it to do our analysis. In Section 3 we introduce the concept of a landmark case in this setting. Then, in Section 4, we apply the theory to a selection of the COMPAS dataset, and some variations upon it. We discuss the results of this analysis in Section 5, and conclude with a summary and some closing thoughts in Section 6.

2. The Formalism of Precedential Constraint

In [14,15] a framework was introduced to formally describe the *a fortiori* reasoning process underlying case law. This is the type of law arising from the rules or principles used in deciding previous cases called *precedents*. Since courts must decide similarly in similar situations, we can say precedents constrain future decision making. This theory underlies the explanation method introduced in [13], and is what we will be evaluating on a recidivism dataset in Section 4. As such we summarise it here before proceeding.

2.1. Dimensions, Preferences, and Cases

In order to describe the *fact situation* of a case we use what are called *dimensions* in the AI & law literature, which are formally just partially ordered sets, i.e. a set together with a reflexive, antisymmetric, and transitive relation. The idea is that we specify a situation as we would a point in space: by specifying its values, the coordinates, in each of the dimensions. The partial order of a dimension serves to indicate the preference which its elements have towards a particular outcome.

Formally a dimension $(d, \preceq)$ is as a set d with a partial order $\preceq$ on d. The model assumes there is a set D of dimensions $\{(d_1, \preceq_1), \ldots, (d_n, \preceq_n)\}$, and a *fact situation* is then a choice function on $\{d_1, \ldots, d_n\}$, i.e. a function F such that $F(d_i) \in d_i$ for $1 \leq i \leq n$. A fact situation can be decided for either of two *outcomes*, or, *sides*, 0 or 1. We will denote an unspecified side with the variable s, and its *opposite* outcome by $\bar{s} := 1 - s$. A *case* is now a fact situation F together with an outcome s, written as $F : s$. When we want to refer to the value of a case $c = F : s$ in a dimension $d \in D$ we may write $c(d)$ instead of $F(d)$. A set of cases CB is called a *case base*.

The order $\preceq$ of a dimension d specifies the relative preference the elements of d have towards a particular outcome. More specifically, if $v \prec w$ for $v, w \in d$ this means w prefers outcome 1 relative to v, and conversely v prefers outcome 0 relative to w. Usually we want to compare preference towards an arbitrary outcome s, so to do this we define for any dimension $(d, \preceq)$ the notation $\preceq_s := \preceq$ if $s = 1$ and $\preceq_s := \succeq$ if $s = 0$. Note that by definition we have $\preceq_s = \succeq_{\bar{s}}$.

Example 2.1. To give some intuition for these definitions we consider a running example of recidivism data. For the sake of continuity we will describe the dimensions as they appear in the data in Section 4. Convicts are described along three dimensions: age $(d_{\texttt{Age}}, \preceq_{\texttt{Age}})$, the number of prior offenses $(d_{\texttt{Priors}}, \preceq_{\texttt{Priors}})$, and sex $(d_{\texttt{Sex}}, \preceq_{\texttt{Sex}})$. Age and number of priors have the natural numbers as possible values, so $d_{\texttt{Age}} := \mathbb{N}$ and $d_{\texttt{Priors}} := \mathbb{N}$. The values for sex are $d_{\texttt{Sex}} := \{\mathrm{M}, \mathrm{F}\}$. The outcome for this domain is a judgement of whether the person is at high (1) or low (0) risk of recidivism. The associated orders are as follows; for age it is the 'greater-than' order on the natural numbers; for the number of priors we take the 'less-than' order; and for sex we take $F \prec_{\texttt{Sex}} M$:

$$\begin{aligned}(d_{\texttt{Age}}, \preceq_{\texttt{Age}}) &:= (\mathbb{N}, \geq),\\ (d_{\texttt{Priors}}, \preceq_{\texttt{Priors}}) &:= (\mathbb{N}, \leq),\\ (d_{\texttt{Sex}}, \preceq_{\texttt{Sex}}) &:= (\{\mathrm{M}, \mathrm{F}\}, \{(\mathrm{F}, \mathrm{F}), (\mathrm{M}, \mathrm{M}), (\mathrm{F}, \mathrm{M})\}).\end{aligned}$$

Beware of the confusion that can arise when the order is the opposite of what one would expect, for instance we have $40 \preceq_{\texttt{Age}} 20$ because $40 \geq 20$.

2.2. The Forcing Relation

The principle of *stare decisis* states that similar cases must be decided similarly, and so the outcome of any particular case will set a precedent which future decision making should abide by. Put differently, precedent constrains future decision making. This is the phenomenon which the theory of precedential constraint tries to model, and it does so principally through the *forcing* relation, introduced in [15, Definition 12].

Definition 2.1. Given fact situations F and G we say G is *at least as good* as F for an outcome s, denoted $F \preceq_s G$, if it is at least as good for s on every dimension d:

$$F \preceq_s G \quad \text{if and only if} \quad F(d) \preceq_s G(d) \text{ for all } d \in D.$$

If moreover $c = F : s$ is a previously decided case we say that c *forces* the decision of G for s. A case base CB forces the decision of G for s if it contains a case that does so.

This relation models *a fortiori* reasoning: if a fact situation F has been decided for s and we encounter a new situation $G \succeq_s F$, then G should also be decided for s. To illustrate this we consider our running example. Suppose that we have classified a female, 30 years of age, with a total of 5 priors, as being high-risk. Then, certainly, a 24 year-old male with 10 prior offenses should be classified high-risk as well, according to the way we ordered these dimensions.

Since the forcing relation is defined componentwise on the dimensions, and all the dimensions are partially ordered, the forcing relation is a partial order as well.

2.3. Consistency

In addition to making decisions about new fact situations on the basis of a case base and the forcing relation, we can consider the degree to which the cases within a case base are *consistent* or *inconsistent* with each other, relative to the forcing relation. For instance, if a case base CB contains two cases $F : s$, $G : t$, such that $F \preceq_s G$ we can have either $s = t$, in which case the decision of G for t is consistent with that of F for s, or we can have $s \neq t$, in which case the decision of G for t is inconsistent with that of F for s.

Definition 2.2. A case $F : s$ is said to be *inconsistent* with respect to a case base CB when deciding F for $\bar{s}$ is forced by CB, and *consistent* otherwise. A case base is said to be consistent when all of its cases are, and inconsistent otherwise.

The notion of consistency defined here is a strong criterion on case bases, and can be expected to fail for an actual case base. It is therefore more interesting to consider *consistency* of a case base as the relative frequency of consistent cases within it.

3. Landmark Cases

We now focus on a special kind of cases that we call *landmark cases*, a notion that to the best of our knowledge is new in the literature. The motivating idea is that when a case has its outcome forced by another, it is – by transitivity of the forcing relation – rendered superfluous as a precedent. As such the most salient cases are those that do not have their outcome forced by another case; these are what we call landmarks.

Definition 3.1. Cases in a case base *CB* which are minimal with respect to the forcing relation $\preceq_s$ are called *landmark* cases; those that are not are called *ordinary*.

This formal notion is connected to the informal idea of a landmark case, in the sense that a formal landmark case does not have its decision forced by another precedent. In that sense it represents new legal ground, and the decision maker has used its discretion, going beyond what is decided by other cases. In the following, we show how our formal notion of landmark cases can be applied to the analysis of an actual data set of cases.

Among landmarks we can further quantify impact by the number of cases of which they force the outcome. This leads us to define two sets that are of particular interest.

Definition 3.2. Given a case base *CB* and an outcome s we define the set L_s of cases with outcome s that force the outcome of the greatest number of other cases in *CB*:

$$L_s := \operatorname*{argmax}_{F : s \in CB} |\{G : t \in CB \mid F \preceq_s G\}|. \tag{1}$$

When L_s is a singleton we write l_s for its sole element.

It follows from transitivity of the forcing relation that the cases in L_0 and L_1 are landmarks as in Definition 3.1, but in general they will not contain all landmarks.

4. Evaluating the Model on Recidivism Data

As an application and evaluation of the theoretical framework we use the COMPAS recidivism dataset from [1], which contains information on convicts and whether they recidivated within two years after being arrested for an initial charge. The idea behind our evaluation is as follows. We first preprocess the data to extract the features that will correspond to dimensions in our formal sense. Then, we determine the orders for these dimensions by fitting a logistic regression model on the data and looking at the signs of the resulting coefficients; positive means the order for that dimension is the usual one on the number line; negative means the order is the reverse of what it usually is. Having assigned the dimension orders we look at the resulting consistency percentages of the COMPAS data and two variations on it, as well as information on the landmark cases.

Our investigation will be similar to the one in [13, Section 6]. However there are two important differences. Firstly, we use real world data which is representative of the intended domain of the explanation methods that we wish to develop. In contrast, the data used in [13] appears to be largely synthetic; the mushroom dataset [16] is listed as containing *hypothetical* samples of mushrooms; the churn dataset [17] is described as data about a *fictional* telecommunications company; and the admission dataset [18] had values "*entered manually with no specific pattern. It was random assignment.*"[2] Secondly, we use logistic regression coefficients instead of Pearson correlation coefficients in order to determine the dimension orders. These different approaches can result in very different orders, and we will argue why logistic regression seems to be the better choice.

[2] See the comments of the admission dataset's author at `https://archive.fo/yIvUN`.

Table 1. An overview of the COMPAS features of interest. A comprehensive overview of the meaning of all the features is lacking in [1], hence this is only our best attempt at an interpretation.

Feature	Description	Order
`Age`	Age of the convict at the time of the COMPAS assessment.	Descending.
`Sex`	Gender as specified when the convict was arrested, can take on the values 'Male' or 'Female'.	Female $\prec$ Male.
`ChargeDegree`	Indicates whether the charge that led to the assessment was a felony (F) or a misdemeanor (M).	M $\prec$ F.
`DaysInJail`	The number of days the convict spend in jail for the crime, computed by comparing (and rounding down) the number of days between the `c_jail_in` and `c_jail_out` fields.	Ascending.
`DaysInCustody`	The number of days the convict spend in custody, computed in the same way as `DaysInJail` but with the `c_custody_in` and `c_custody_out` fields.	Ascending.
`Priors`	The number of offenses committed prior to the one that led to the COMPAS assessment. This is computed as the sum of the `juv_fel_count`, `juv_misd_count`, `juv_other_count`, and `priors_count` fields in the original dataset.	Ascending.
`Label`	The label, indicating whether there was "*a criminal offense that resulted in a jail booking and took place after the crime for which the person was* COMPAS *scored . . . within two years after the first*" [1].	N/A.

4.1. Data Preprocessing

Before analyzing the COMPAS data we preprocess it. In particular, we discard features that are not of interest, delete rows that do not have values for the remaining features, create new features on the basis of old ones, and finally delete duplicate rows. Below follows a more detailed description of the steps taken.

First we discard features that are not of interest. For instance, many of the features in the original dataset pertain to the COMPAS system, but presently we are only interested in the data describing the convicts and whether they recidivated or not, not in the COMPAS system itself. Some features are of interest to us but are not in the right format. For instance, the two columns `c_jail_in` and `c_jail_out` together tell us how many days the convict spend in prison, but are represented in a date format, so we replace them with a new `DaysInJail` feature holding the number of days spend in prison. A complete overview of the resulting features and their meaning can be found in Table 1.

Lastly we remove any rows that do not have values for any of the relevant features, or which occur more than once in the data. This last step is necessary for our landmark analysis; a case c may be a landmark, but if there is a second case d with exactly the same fact situation and outcome as c but not *equal* to c, then neither c nor d are landmarks.

We are then left with a total of 5873 rows and we will henceforth refer to that set when we say 'COMPAS dataset'. In addition we will look at two variations on that set. The first we will call the 'small COMPAS dataset', which is obtained from the COMPAS dataset by omitting all features except `Age` and `Priors`, and then deleting all duplicates. The second we call the 'CORELS dataset', and is obtained by changing the labels in the small COMPAS dataset according to the recidivism prediction rule found by CORELS as described in [19, Figure 1].

4.2. Determining the Preference Orders

The main difficulty with making a precedential constraint model for a particular domain lies in determining the orders for the dimensions. For instance, in our example with recidivism data we have an age dimension, and to determine its respective order is to say whether the elderly are more likely to recidivate than the young, or vice versa. Knowledge engineering techniques and statistical methods can be used for this purpose. For instance, for the age dimension, much has been written on the interplay between age and recidivism, the conclusion of which is summarized by the adage that "*people age out of crime*," meaning that as people age they become decreasingly likely to recidivate. Another option is to look at statistical trends in the data, for instance, by considering the sign of the Pearson correlation between age and recidivism. If it is positive, we say that likelihood of recidivism increases with age, and if it is negative, we say it decreases.

For our evaluation we apply the statistical method. We will use the same underlying idea as used by [13], which is to use a function c that associates each numerical feature x with a *coefficient* $c(x)$ indicating the degree to which the values of x favour outcome 1. If $c(x)$ is positive we order the values of x with the usual 'less-than' order $\leq$ on the number line, and if it is negative we order it using the 'greater-than' order $\geq$; so more precisely $\preceq := \leq$ if $c(x) \geq 0$ and $\preceq := \geq$ if $c(x) < 0$.

If x is categorical we cannot apply c directly so we use *dummy variables*. More specifically, if x is a categorical feature which can take the possible (unordered) values $v_1, \ldots, v_n$, then we introduce for each value v_i a dummy variable d_{v_i} which is a binary feature indicating whether $x = v_i$. Then we define $v_i \preceq v_j$ if and only if $c(d_{v_i}) \leq c(d_{v_j})$.

For the present work we define c using *logistic regression*. Supposing we have features $x_1, \ldots, x_n$ the logistic model has parameters $\beta_0, \ldots, \beta_n$, and models the probability that a given sample belongs to class 1 by the formula

$$p(x_1, \ldots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}. \tag{2}$$

We find appropriate values for the β parameters using the scikit-learn implementation of a maximum likelihood estimation with default parameters [20], and after this is done we can simply put $c(x_i) := \beta_i$. The resulting orders on the features are listed in Table 1.

4.3. The Pearson Correlation of Dummy Variables

In [13] the function c is defined using the Pearson correlation. However this approach seems to work poorly with categorical features, and we will now look at why this is. This will involve some mathematical details that are not relevant to the further analysis of the COMPAS data, so the reader may skip ahead to Section 4.4.

Given n samples $(x_1, y_1), \ldots, (x_n, y_n)$ of binary variables x and y, the estimate of the Pearson correlation $r(x, y)$ between x and y is given by

$$r(x, y) = \frac{n\mathbb{1}_{xy} - \mathbb{1}_x \mathbb{1}_y}{\sqrt{n\mathbb{1}_x - \mathbb{1}_x^2}\sqrt{n\mathbb{1}_y - \mathbb{1}_y^2}}, \tag{3}$$

where $\mathbb{1}_x$ is the number of times x takes value 1 in the samples, $\mathbb{1}_y$ the number of times y takes value 1, and $\mathbb{1}_{xy}$ the number of times x and y both take value 1.

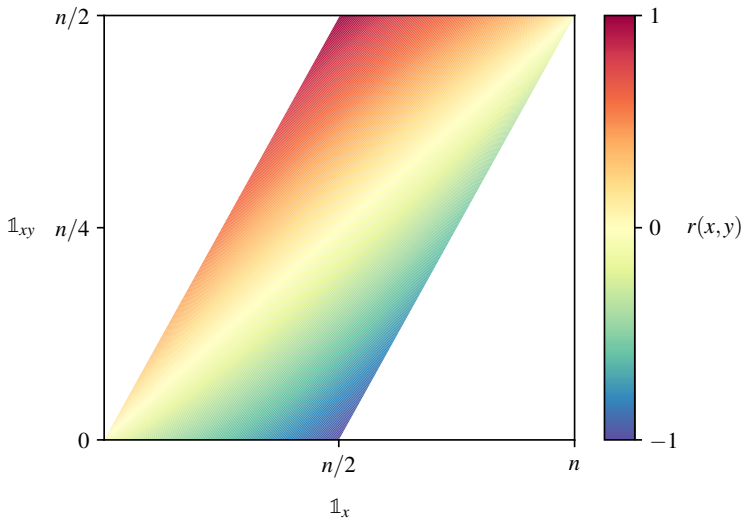

Figure 1. A plot of Eq. 3, the Pearson correlation coefficient for binary vectors x and y, for a fixed value of $n := 400$, and with $\mathbb{1}_y := n/2$. Note the constraints that $\mathbb{1}_x + \mathbb{1}_y - n \leq \mathbb{1}_{xy} \leq \mathbb{1}_x$.

In order to get a sense of how this function behaves we plot its values for a fixed n and with $\mathbb{1}_y := n/2$, see Figure 1. This plot shows that when $\mathbb{1}_x$ is relatively low, or relatively high, the range of $r(x,y)$ (as a function of $\mathbb{1}_{xy}$) is not $[-1,1]$ but some restricted interval near 0. More precisely, writing $s(x) := \sqrt{n\mathbb{1}_x - \mathbb{1}_x^2}$, we can calculate that for $0 \leq \mathbb{1}_x \leq n/2$ the range of $r(x,y)$ is $[-\mathbb{1}_x/s(x), \mathbb{1}_x/s(x)] \approx [-\mathbb{1}_x/(n/2), \mathbb{1}_x/(n/2)]$. This is undesirable when x is a dummy variable, as then $\mathbb{1}_x$ simply indicates the number of times the original categorical feature took the value which the dummy variable represents, i.e. the number of samples we have of that class.

Example 4.1. Let us consider an example to illustrate this point. The original COMPAS data includes a `Race` variable, with possible values including 'Asian' and 'Caucasian'. The value Asian occurs much less often than Caucasian (0.4% against 34%), meaning that the value of $\mathbb{1}_x$ for the dummy variable for `Race` = Asian is much lower than that for the `Race` = Caucasian variable. As a result, its Pearson correlation must land in a very small interval around 0, while the one for Caucasian has almost the full range available. Indeed, the order for the `Race` dimension on the basis of the Pearson correlation method puts Caucasian in the last position (i.e. comparatively least prone to recidivate), and Asian a little over halfway in the order. To compare this with a measure that does not place such great importance on the number of samples that we have of each race, we consider the relative frequency $\mathbb{1}_{xy}/\mathbb{1}_x$ of recidivism within that class. The picture is now the opposite of what we see with Pearson correlation, with Asian ending lowest in the ranking (at 28% prevalence) and Caucasian a little over halfway (at 40% prevalence).

As mentioned we opt to use logistic regression rather than the Pearson correlation or some other function $f(x,y)$ such as the relative frequency $\mathbb{1}_{xy}/\mathbb{1}_x$. There are two reasons for this. Firstly, logistic regression does not consider the vectors x and y in isolation but also takes into account the other variables in the data. Secondly, this method of determining the orders comes with a decision boundary which (if the dimensionality is low enough) can be visualized, and we will make use of this later on.

Table 2. On the left is a summary of the forcing relation on the COMPAS dataset, and the impact of the landmarks l_0 and l_1 defined in Definition 3.2. On the right is a concrete description of l_0 and l_1. Notice that they are archetypal examples of the *opposite* class that they belong to; l_0 is a young male with many priors, yet did not recidivate; while l_1 is an older female with no priors, but did recidivate.

	Label 0	Label 1	Total
Consistent	76	397	473
Inconsistent	2783	2617	5400
Forced by l_0	2271	1765	4036
Forced by l_1	2296	2700	4969
Landmarks	70	18	88

d	$l_0(d)$	$l_1(d)$
`Age`	23	49
`Sex`	Male	Female
`ChargeDegree`	F	M
`DaysInJail`	70	0
`DaysInCustody`	70	0
`Priors`	11	0

4.4. Results

Having selected the dimensions, assigned their orders, and constructed the case base, we can now evaluate various statistics. Keep in mind that all the results that follow are stated relative to the orders that we have assigned; different orders result in different statistics.

4.4.1. On the COMPAS *Data*

We start by looking at the consistency percentage, i.e. the relative frequency of cases which do not have their outcome disputed by the forcing relation on the case base. We find the COMPAS dataset is only 8% consistent, see Table 2. This low percentage is caused by a small number of landmarks, outliers in the data that one would expect to have the opposite label of the one they received. Of those landmarks the l_0 and l_1 cases defined by Definition 3.2 are most impactful. They are shown in Table 2. In Figure 2 an overview of the collective impact of the landmarks is shown.

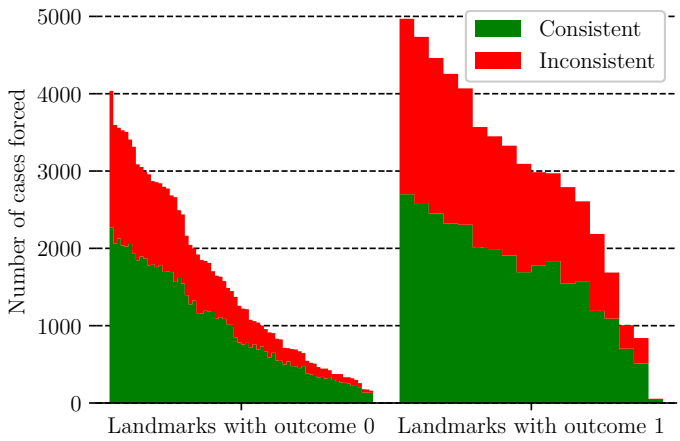

Figure 2. A visualization of the impact of the landmarks in the COMPAS data. Each vertical bar represents one landmark and shows the number of cases of which it forces the decision. The green area depicts the portion of those cases that have the same outcome as that landmark, and the red area the portion of cases which do not. Note that the leftmost bars are – by definition – the l_0 and l_1 cases of Definition 3.2, respectively.

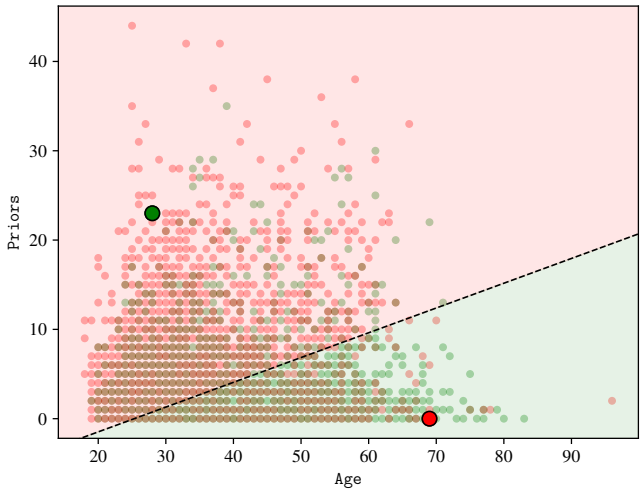

Figure 3. An illustration of the small COMPAS dataset and its l_0 and l_1 landmarks (Definition 3.2), indicated by the enlarged circles. The green dots indicate cases with outcome 0 and the red dots those with outcome 1. The dotted line is the decision boundary associated with the logistic regression coefficients.

Remark 4.1. The notions of landmark and outlier, while similar, are not the same; a landmark need not be an outlier (cf. Figure 4) and an outlier need not be a landmark (for instance when there is an outlier even further across the best fit decision boundary).

4.4.2. On the Small COMPAS *Data*

High dimensional data is difficult to visualize, so in order to get a better view of these results we repeat our analysis on a subset of the data with only the two most predictive variables – `Age` and `Priors`. We dub the resulting data the small COMPAS dataset. The resulting order on the variables remains the same as in the larger version. This lets us visualize the data, the decision surface of our logistic model, and the landmarks l_0 and l_1, see Figure 3 for the resulting plot. The l_0 and l_1 cases highlight the cause for the inconsistency: there are many cases that lie on the opposite side of the decision boundary for their class, and so their 'forcing cones' contain many cases of the opposite class.

4.5. On the CORELS *Data*

The preceding results have shown that the model of precedential constraint is a poor fit on the COMPAS data. This makes sense intuitively, because when someone of a certain age and with some number of priors recidivates, we cannot expect this to set a precedent which future convicts will abide by. For example, when an elderly lady with no prior offenses recidivates, this will have very little influence on the behaviour of convicts thereafter. In other words, the process underlying recidivism does not respect precedence.

This type of reasoning should be more suited to our running example from Section 2 in which we judge *risk* of recidivism. When a person is assigning low or high risk of recidivism, we would expect this assignment to obey the a fortiori principle, if it agrees with our notion of what makes someone more or less likely to recidivate.

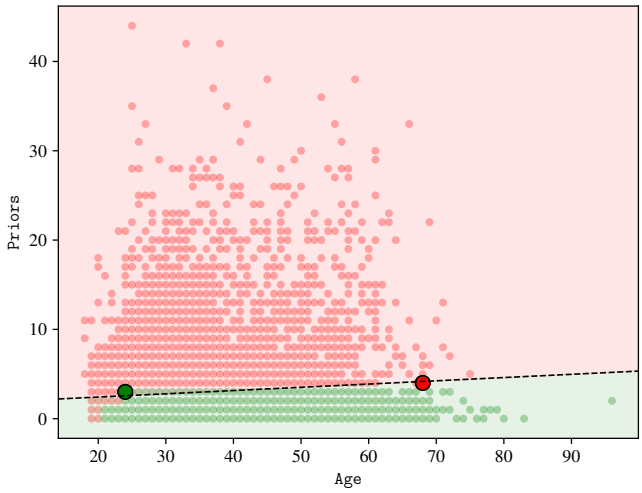

Figure 4. The small COMPAS dataset from Section 4.4.2, but with its labels replaced according to the CORELS rule shown in Eq. 4. The data is now better linearly separable. As a result the landmarks are close to the decision boundary and do not cause inconsistency.

To test this hypothesis we change the labels of the small COMPAS data according to a sensible risk assessment rule, listed in [19, Figure 1]. It was mined from the original COMPAS data using the Certifiably Optimal Rule Lists (CORELS) algorithm, introduced in [19]. This rule is listed below, with the only modification being that we omit the clause that $x.\texttt{Sex} = \text{Male}$ from the first case distinction, since we have omitted the `Sex` dimension for the sake of visualizability:

$$x.\texttt{Label} := \begin{cases} 1 & \text{if } 18 \leq x.\texttt{Age} \leq 20, \\ 1 & \text{if } 21 \leq x.\texttt{Age} \leq 23 \text{ and } 2 \leq x.\texttt{Priors} \leq 3, \\ 1 & \text{if } 3 < x.\texttt{Priors}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Changing all labels according to this rule, and then removing duplicates, results in a new dataset that will be referred to as the CORELS data.

Now we can again fit our model to this data and visualize the l_0 and l_1 landmarks of Definition 3.2, along with the decision boundary, see Figure 4 for the resulting plot. As expected this rule does satisfy the a fortiori principle, and as a result the consistency is very high (in fact the dataset is fully consistent). The l_0 and l_1 landmarks give a good sense of where the decision boundary is located.

5. Discussion

In all, these results suggest that we can think of the phenomenon of inconsistency in two ways. The first is the mathematical view that the theory of precedential constraint

contains a linearity assumption, and that the consistency percentage is a measure of the degree to which the data is linearly separable. Of each class, the landmarks are then those cases which lie furthest in the direction of the best fit linear decision boundary, and the farther they cross it the more inconsistency they cause. The second is the semantic view that it tells us to what degree the labelling process relies on a fortiori reasoning, or the degree to which we can expect precedent to be obeyed. If this is the case, then the landmarks are those cases that most reveal the nature of the underlying labelling process.

Our results suggest that the presence of a small number of landmarks that force the decision of the rest is what we can expect of an average dataset, because in general a partial order will have far fewer minimal elements than that it will have elements in total. Two factors that can influence this is the number of dimensions, and the way in which we order them. For instance if we have a dimension with more than two values and we order them so that they are all incomparable, it will immediately become impossible for any case to force the outcome of another, and so every case becomes a landmark.

6. Conclusion

In the recent work [13] an explainable AI method was developed on the basis of the theory of precedential constraint. In the present work we evaluated how suitable this theory is to model the kind of data we might encounter in cases that require explainability methods. As an example of such a situation we chose the COMPAS data from [1]. We fitted the model on this data and some variations upon it, and analysed the results in the sense that we measured their consistency percentages, and looked at the structure of the forcing relation on cases. Through this analysis, and the use of the concept of landmark cases, we extrapolate from these results a characterization of the type of datasets that are fit to be described by the theory of precedential constraint. This characterization can be viewed mathematically, as consistency indicating the degree of linear separability of the data; or viewed semantically, as consistency indicating the degree to which the process generating the data respects precedence, or depends on a fortiori type reasoning.

These results raise several questions which may be addressed in future research. Firstly, there is the question to what degree our results are relative to the method we used to determine the dimension orders. In the framework as outlined by [15] these orders may be partial but in our case we only use linear orders. What are the situations in which we might want to make elements of a dimension incomparable, and how would the presence of incomparable pairs affect our findings regarding consistency? Closely related is a second question regarding the method of determining these orders. In this work we used logistic regression, but other ways of doing this are conceivable. What are the differences between these approaches, and what should the measure of success be? Especially in a setting where the theory of precedential constraint underlies an explanation method, as is the case for the method developed in [13], it might be better to use the black box that is under examination to determine these orders, because in that case we are not so interested in what *we* think the orders should be but what the black box thinks they should be. Thirdly, since the theory of precedent is in some sense itself a data-driven model, we can ask how it compares to other such models. The model on the CORELS data, shown in Figure 4, suggests it functions as a type of decision tree, but an obvious difference is that it is not always capable of classifying an arbitrary unseen case, and that it may give conflicting classifications (even if the case base is consistent).

Acknowledgements

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022.

References

[1] Larson J, Mattu S, Kirchner L, Angwin J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. 2016.

[2] Dieterich W, Mendoza C, Brennan T. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc. 2016;7(4).

[3] Rudin C, Wang C, Coker B. The Age of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review. 2020 3;2(1). Available from: https://hdsr.mitpress.mit.edu/pub/7z10o269.

[4] Barenstein M. ProPublica's COMPAS Data Revisited. arXiv preprint arXiv:190604711. 2019.

[5] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15. Available from: https://doi.org/10.1038/s42256-019-0048-x.

[6] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law. 2017;7(2):76-99.

[7] Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y. Measuring Bias in Contextualized Word Representations. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. Florence, Italy: Association for Computational Linguistics; 2019. p. 166-72.

[8] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135-44.

[9] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv JL & Tech. 2017;31:841.

[10] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: International Conference on Machine Learning. PMLR; 2017. p. 1885-94.

[11] Nugent C, Cunningham P. A Case-Based Explanation System for Black-Box Systems. Artificial Intelligence Review. 2005;24(2):163-78.

[12] Čyras K, Satoh K, Toni F. Explanation for case-based reasoning via abstract argumentation. In: Baroni P, Gordon T, Scheffler T, Stede M, editors. Computational Models of Argument. IOS Press; 2016. p. 243-54.

[13] Prakken H, Ratsma R. A top-level model of case-based argumentation for explanation: Formalisation and experiments. Argument & Computation. 2021;Preprint:1-36.

[14] Horty JF. Rules and reasons in the theory of precedent. Legal Theory. 2011;17(1):1–33.

[15] Horty J. Reasoning with dimensions and magnitudes. Artificial Intelligence and Law. 2019;27(3):309-45.

[16] Schlimmer J. UCI Machine Learning Repository; 1987. Available from: https://archive.ics.uci.edu/ml/datasets/mushroom.

[17] IBM. Telco customer churn dataset, part of the Base Samples of Cognos Analytics; 2019. Available from: https://www.kaggle.com/blastchar/telco-customer-churn.

[18] Acharya MS, Armaan A, Antony AS. A Comparison of Regression Models for Prediction of Graduate Admissions. In: 2019 International Conference on Computational Intelligence in Data Science (ICCIDS); 2019. p. 1-5.

[19] Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C. Learning Certifiably Optimal Rule Lists for Categorical Data. Journal of Machine Learning Research. 2018;18(234):1-78. Available from: http://jmlr.org/papers/v18/17-716.html.

[20] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.