Landmarks in Case-based Reasoning: From Theory to Data

Wijnand VAN WOERKOM[a], Davide GROSSI[b, c, d], Henry PRAKKEN[a, c], 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.

Motivation and Theoretical Background - Precedential Constraint

In order to describe the fact situation of a case we use what are called dimensions in the AI & law literature [1], which are formally just partially ordered sets, i.e. a set together with a reflexive, antisymmetric, and transitive relation.

Definition A *dimension* is a partially ordered set $(d, \preceq)$. We assume there is a set D of such dimensions. The orders of the dimensions indicate the relative preferences their values have towards either of two outcomes 0 and 1. This means that for values $v, w \in d$ and $v \prec w$, then w prefers outcome 1 relative to v, and conversely v prefers outcome 0 relative to w.

Example To give some intuition for these definitions we consider a running example of recidivism risk prediction. The dimensions we use are age, sex, and number of prior offenses. The definition of D is as follows:

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

The choice of these orders mean that we consider younger males with many priors to be more likely to recidivate than older females with many priors.

Definition A *fact situation* is a choice function on D, i.e. a function F with $F(d) \in d$ for each $d \in D$. A fact situation paired with an outcome $s \in \{0, 1\}$ is a *case*. A case base *CB* is a set of cases. For a case *p = (F, s)* we write *p(d)* instead of *F(d)*.

Example Consider two cases p, q in our recidivism example, both judged high risk (meaning paired with outcome 1):

$$\begin{aligned} p(\texttt{Age}) &= 45, & p(\texttt{Priors}) &= 4, & p(\texttt{Sex}) &= \mathrm{M},\\ q(\texttt{Age}) &= 50, & q(\texttt{Priors}) &= 5, & q(\texttt{Sex}) &= \mathrm{M}.\end{aligned}$$

Definition The way in which precedent constrains future decision making is modelled by the *forcing relation* on cases. For a case p with outcome 1 and a fact situation F we say p forces the outcome of F for outcome 1 if $p(d) \preceq q(d)$ for all $d \in D$.

Example In our example above the outcome of the fact situation of q is forced for 1 by the case p.

Landmark cases

In this work we bring attention to 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 Cases in a case base *CB* which are minimal with respect to the forcing relation are called *landmark 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 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*:

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

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

Automatically Determining Dimension 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.

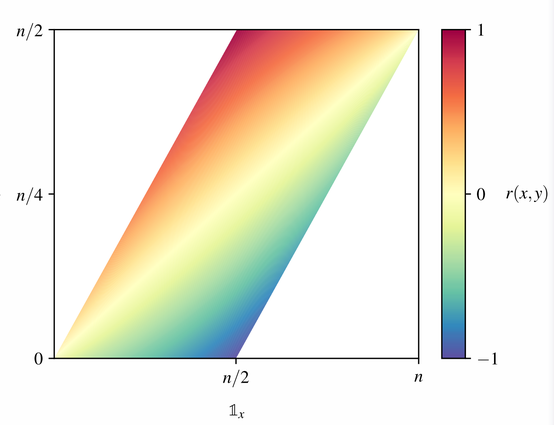
A general method was proposed in [2] which uses a function c which associates each numerical feature x to a coefficient *c(x)* indicating the degree to which higher values favor outcome 1. If *c(x)* is positive then the higher the value, the higher the preference for outcome 1; if *c(x)* is negative then the lower the value, the higher the preference for outcome 1. This lets us assign the dimension order for a numerical feature by $\preceq := \leq$ 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 a dummy variable d_v which is a binary feature indicating whether $x = v$. Then we define $v_i \preceq v_j$ if and only if $c(d_{v_i}) \leq c(d_{v_j})$.

In [2] the value of c for a numerical feature is given by its Pearson correlation in the data. In this work we instead opt to use logistic regression, by fitting a logistic model to the features and then letting c assign a feature x to its corresponding coefficient in the logistic model

We opt to do it this way because the Pearson correlation approach seems to work poorly with categorical features. This is because for binary variables (such as the dummy variables this approach relies on) is given by the equation below:

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

The graph shown on the right illustrates the issue: too much emphasis is placed on class prevalence.



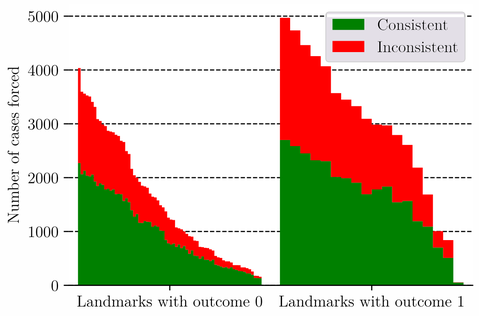
Evaluating the Model on Recidivism Data

As an application and evaluation of the theoretical framework we use the COMPAS recidivism dataset from [3], which contains information on convicts and whether they recidivated within two years after being arrested for an initial charge. We selected a subset of the features and used our method to automatically determine the dimension orders, as shown here.

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.

We then considered the degree to which the cases in the case base are consistent, by measuring the relative frequency of cases that do not have the opposite outcome they received forced by the rest of the case base. It turns out the case base is highly inconsistent, as the graph below demonstrate. This is caused entirely by a relatively small number of landmarks.

Out of the 5873 cases only 473 were consistent, which gives a consistency percentage of 8%. The dataset contains just 88 landmarks, of which just 2 force the vast majority of the outcomes of other cases, as illustrated by the graph to the right. Each bar rerpresents one landmarks, and the height of the bar indicates the number of cases forced by that landmark. The leftmost are the l_0 and l_1 cases defined earlier. The l_0 case is a 23 year old mate with many priors who committed a felony crime but did not recidivate. The l_1 case is a 49 year old female with no priors who committed a misdemeanor, but did recidivate. So we see they are archetypal examples of the *wrong* class.



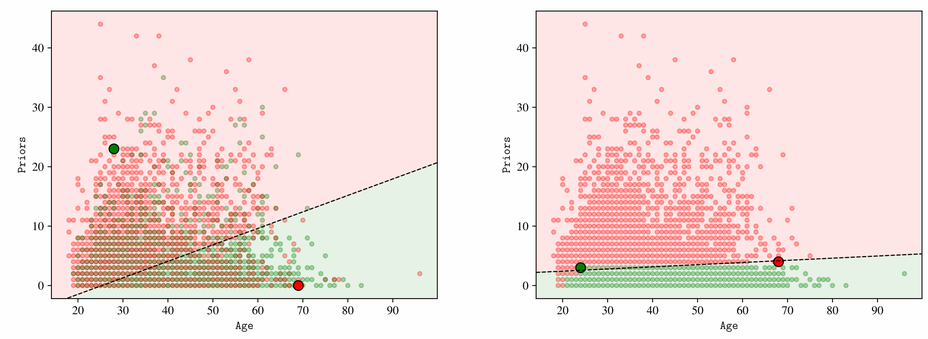
Inconsistency, Linear Separability, and Adhering to Precedence

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 number of priors. The dimension orders remain the same as in the complete dataset. This lets us visualize the data, the decision surface of our logistic model, and the landmarks l_0 and l_1, see the left plot below. 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.

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. This type of reasoning should be more suited to our running example from earlier in which we *judge risk* of recidivism. To test this hypothesis we change the labels of the age and priors compas data according to a sensible risk assessment rule, listed in [4]:

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

Repeating our analysis again yields the graph on the right below. 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 landmarksarks give a good sense of where the decision boundary is located.



Conclusion

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.

References

[1] Horty J. Reasoning with dimensions and magnitudes. Artificial Intelligence and Law. 2019;27(3):309-45.
[2] Prakken H, Ratsma R. A top-level model of case-based argumentation for explanation: Formalisation and experiments. Argument & Computation. 2021;Preprint:1-36.
[3] Larson J, Mattu S, Kirchner L, Angwin J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. 2016.
[4] 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.

Acknowlegements

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, https://hybrid-intelligence-centre.nl, grant number 024.004.022.

NWO Hybrid Intelligence Utrecht University