Learning Cost-Effective and Interpretable Treatment Regimes for Judicial Bail Decisions

Himabindu Lakkaraju
Department of Computer Science
Stanford University
himalv@cs.stanford.edu

Cynthia Rudin
Department of Computer Science
Duke University
cynthia@cs.duke.edu

1 Introduction

Decision makers, such as judges, make crucial choices regarding judicial bail decisions on a daily basis. Such decisions typically involve careful assessment of the subject's (or defendant's) condition, analyzing the costs associated with the possible actions, and the nature of the consequent outcomes. Further, there might be costs associated with the assessment of the subject's condition itself. For instance, a judge deciding if a defendant should be granted bail studies the criminal records of the defendant, and enquires for additional information (e.g., defendant's personal life or economic status) if needed. She then recommends a course of action (or treatment)[1] that trades off the risk with granting bail to the defendant (the defendant may commit a new crime when out on bail) with the cost of denying bail (adverse effects on defendant, or defendant's family, cost of jail to the county).

Decision makers often leverage personal experience to make decisions in these contexts, without considering data, even if massive amounts of it exist. Machine learning models could be of immense help in such scenarios – but these models would need to consider all three aspects discussed above: predictions of counterfactuals, costs of gathering information, and costs of treatments. Further, these models must be interpretable in order to create any reasonable chance of a human decision maker actually using them. In this work, we address the problem of learning such cost-effective, interpretable treatment regimes from observational data.

Prior research addresses various aspects of the problem at hand in isolation. For instance, there exists a large body of literature on estimating treatment effects [5, 13, 4], recommending optimal treatments [1, 15, 6], and learning intelligible models for prediction [10, 8, 11, 2]. However, an effective solution for the problem at hand should ideally incorporate all of the aforementioned aspects. Furthermore, existing solutions for learning treatment regimes neither account for the costs associated with gathering the required information, nor the treatment costs. The goal of this work is to propose a framework which jointly addresses all of the aforementioned aspects.

We address the problem at hand by formulating it as a task of learning a decision list that maps subject characteristics to treatments such that it: 1) maximizes the expectation of a pre-specified outcome when used to assign treatments to a population of interest 2) minimizes costs associated with assessing subjects' conditions and 3) minimizes costs associated with the treatments themselves. We propose a novel objective function to learn a decision list optimized with respect to the criteria listed above. We show that the proposed objective is NP-hard. We then optimize this objective by modeling it as a Markov Decision Process (MDP) and employing a variant of the Upper Confidence Bound for Trees (UCT) strategy which leverages customized checks for pruning the search space effectively. Our results on a real world dataset comprised of judicial bail decisions demonstrate the effectiveness of the proposed solution.

[1] In this paper, we use the word *treatment* to represent a course of action or a choice. For instance, in the context of bail decisions, releasing a defendant on a cash bond is one of the possible treatments.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Our Framework

First, we formalize the notion of treatment regimes and discuss how to represent them as decision lists. We then propose an objective function for constructing cost-effective treatment regimes.

Input Data and Cost Functions. Consider a dataset $\mathcal{D} = \{(\mathbf{x}_1, a_1, y_1), (\mathbf{x}_2, a_2, y_2) \cdots (\mathbf{x}_N, a_N, y_N)\}$ comprised of N independent and identically distributed observations, each of which corresponds to a *subject* (individual), potentially from an observational study. Let $\mathbf{x}_i = \left[x_i^{(1)}, x_i^{(2)}, \cdots x_i^{(p)}\right] \in [\mathcal{V}_1, \mathcal{V}_2, \cdots \mathcal{V}_p]$ denote the *characteristics* of subject i. $\mathcal{V}_f$ denotes the set of all possible values that can be assumed by a characteristic $f \in \mathcal{F} = \{1, 2, \cdots p\}$. Each characteristic $f \in \mathcal{F}$ can either be a binary, categorical or real valued variable. In the bail setting, example characteristics include defendant's age, previous criminal records, mental health status etc., Let $a_i \in \mathcal{A} = \{1, 2, \cdots m\}$ and $y_i \in \mathbb{R}$ denote the *treatment* assigned to subject i and the corresponding *outcome* respectively. We assume that y_i is defined such that higher values indicate better outcomes. For example, the outcome associated with a defendant can be regarded as a score capturing the aftermath of the judge's decision. If a defendant is released on bail and he/she commits a new crime when out on bail, then the corresponding outcome score will be lower than the score assigned to a scenario where the defendant is on his/her best behavior (commits no new crimes, appears at all further court dates) when out on bail.

It can be much more expensive to determine certain subject characteristics compared to others. For instance, a defendant's age can be easily retrieved from his/her records. On the other hand, determining if he/she is mentally ill currently requires more comprehensive medical diagnosis, and is therefore more expensive in terms of monetary costs, time and effort required both from the defendant as well as the judicial system. We assume access to functions $d : \mathcal{F} \rightarrow \mathbb{R}$, and $d' : \mathcal{A} \rightarrow \mathbb{R}$ which return the cost of determining any characteristic in $\mathcal{F}$, and the cost of each treatment $a \in \mathcal{A}$ respectively.

Treatment Regimes. A treatment regime is a function that takes as input the characteristics of any given subject $\mathbf{x}$ and maps them to an appropriate treatment $a \in \mathcal{A}$. We employ *decision lists* to express treatment regimes as they tend to be intelligible. A decision list is an ordered list of rules embedded within an if-then-else structure. A treatment regime expressed as a decision list π is a sequence of $L+1$ rules $[r_1, r_2, \cdots, r_{L+1}]$. The last one, r_{L+1}, is a default rule which applies to all those subjects who do not satisfy any of the previous L rules. Each rule r_j (except the default rule) is a tuple of the form (c_j, a_j) where $a_j \in \mathcal{A}$, and c_j represents a *pattern* which is a conjunction of one or more predicates. Each predicate takes the form (f, o, v) where $f \in \mathcal{F}$, $o \in \{=, \neq, \leq, \geq, <, >\}$, and $v \in \mathcal{V}_f$ denotes some value v that can be assumed by the characteristic f (Eg.,'Age $\geq 40 \wedge$ Gender=Female"). We define an indicator function, *satisfy*(x_i, c_j), which returns a 1 if x_i satisfies c_j and 0 otherwise.

The rules in π partition the dataset $\mathcal{D}$ into $L+1$ groups: $\{\mathcal{R}_1, \mathcal{R}_2 \cdots \mathcal{R}_L, \mathcal{R}_{\text{default}}\}$. A group $\mathcal{R}_j$, where $j \in \{1, 2, \cdots L\}$, is comprised of those subjects that satisfy c_j but do not satisfy any of $c_1, c_2, \cdots c_{j-1}$. This can be formally written as:

$$\mathcal{R}_j = \left\{ \mathbf{x} \in [\mathcal{V}_1 \cdots \mathcal{V}_p] \mid \text{satisfy}(\mathbf{x}, c_j\} \wedge \bigwedge_{t=1}^{j-1} \neg \, \text{satisfy}(\mathbf{x}, c_t) \right\}. \quad (1)$$

The treatment assigned to each subject by π is determined by the group that he/she belongs to. More formally,

$$\pi(\mathbf{x}_i) = \sum_{l=1}^{L} a_l \, \mathbb{1}(\mathbf{x}_i \in \mathcal{R}_l) + a_{\text{default}} \, \mathbb{1}(\mathbf{x}_i \in \mathcal{R}_{\text{default}}) \quad (2)$$

Similarly, the cost incurred when we assign a treatment to the subject i (*treatment cost*) according to the regime π is given by:

$$\phi(\mathbf{x}_i) = d'(\pi(\mathbf{x}_i)) \quad (3)$$

where the function d' (defined previously) takes as input a treatment $a \in \mathcal{A}$ and returns its cost.

We can also define the cost incurred in assessing the condition of a subject i (*assessment cost*) as per the regime π. Note that a subject i belongs to the group $\mathcal{R}_j$ if and only if the subject does not satisfy the conditions $c_1 \cdots c_{j-1}$, but satisfies the condition c_j (Refer to Eqn. 1). This implies that the assessment cost incurred for this subject i is the sum of the costs of all the characteristics that appear in $c_1 \cdots c_j$. If $\mathcal{N}_l$ denotes the set of all the characteristics that appear in $c_1 \cdots c_l$, the assessment cost of the subject i as per the regime π can be written as:

$$\psi(\mathbf{x}_i) = \sum_{l=1}^{L} \left[\mathbb{1}(\mathbf{x}_i \in \mathcal{R}_l) \times \left(\sum_{e \in \mathcal{N}_l} d(e) \right) \right]. \tag{4}$$

Objective Function. We first formalize the notions of expected outcome, assessment, and treatment costs of a treatment regime π with respect to the dataset $\mathcal{D}$.

The quality of the regime π is partly determined by the expected outcome when all the subjects in $\mathcal{D}$ are assigned treatments according to π. The higher the value of such an expected outcome, the better the quality of the regime π. There is, however, one caveat to computing the value of this expected outcome – we only observe the outcome y_i resulting from assigning $\mathbf{x}_i$ to a_i in the data $\mathcal{D}$, and not any of the counterfactuals. To address this problem, we compute the expected outcome of a given regime π using doubly robust estimation [12]:

$$g_1(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} \left[\frac{\mathbb{1}(a_i = a)}{\hat{\omega}(x_i, a)} \{y_i - \hat{y}(x_i, a)\} + \hat{y}(x_i, a) \right] \mathbb{1}(\pi(x_i) = a) \tag{5}$$

$\hat{\omega}(x_i, a)$ denotes the probability that the subject i with characteristics x_i is assigned to treatment a in the data $\mathcal{D}$. $\hat{\omega}$ represents the propensity score model. In practice, we fit a multinomial logistic regression model on $\mathcal{D}$ to learn this function. Our framework does not impose any constraints on the functional form of $\hat{\omega}$. $\hat{y}$ corresponds to the outcome regression model and is learned in our experiments by fitting a linear regression model on $\mathcal{D}$ prior to optimizing for the treatment regimes.

The assessment cost of a subject i w.r.t. the regime π is given in Eqn. 4. The expected assessment cost across the entire population can be computed as:

$$g_2(\pi) = \frac{1}{N} \sum_{i=1}^{N} \psi(\mathbf{x}_i). \tag{6}$$

The treatment cost for a subject i who is assigned treatment using regime π is given in Eqn. 3. The expected treatment cost across the entire population can be computed as:

$$g_3(\pi) = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i). \tag{7}$$

The smaller the expected assessment and treatment costs of a regime, the more desirable it is in practice.

Given the observational data $\mathcal{D}$ and a set of all possible combinations of candidate rules $C(\mathcal{L})$, our objective is to maximize $g_1(\pi)$, and minimize $g_2(\pi)$ and $g_3(\pi)$:

$$\arg\max_{\pi \in C(\mathcal{L}) \times \mathcal{A}} \lambda_1 g_1(\pi) - \lambda_2 g_2(\pi) - \lambda_3 g_3(\pi). \tag{8}$$

The λ's in Eqn. 8 are non-negative weights that scale the relative influence of terms in the objective. The above objective function is NP-Hard (refer Appendix for more details [9]).

2.1 Optimizing the Objective

We optimize our objective by modeling it as as a Markov Decision Process (MDP) and then employing Upper Confidence Bound on Trees (UCT) algorithm to find a treatment regime which maximizes Eqn. 8. We also propose and leverage customized checks for guiding the exploration of the UCT algorithm and pruning the search space effectively.

Markov Decision Process Formulation Our goal is to find a sequence of rules which maximize the objective function in Eqn. 8. To this end, we formulate a fully observable MDP such that the optimal policy of the posited formulation provides a solution to our objective function.

A fully observable MDP is characterized by a tuple $(\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathbf{R})$ where $\mathbf{S}$ denotes the set of all possible states, $\mathbf{A}$ denotes the set of all possible actions, $\mathbf{T}$ and $\mathbf{R}$ represent the transition and reward functions respectively. Below we define each of these in the context of our problem. *State Space.* Conceptually, each state in our state space captures the effect of some partial or fully constructed decision list. To illustrate, let us consider a partial decision list with just one rule "if Age $\geq 40 \wedge$ Gender = Female, then T1". This partial list induces that: (i) all those subjects that satisfy the condition of the rule are assigned treatment T1, and (ii) Age and gender characteristics will be required in determining treatments for all the subjects in the population.

To capture such information, we represent a state $\tilde{s} \in \boldsymbol{S}$ by a list of tuples $[(\boldsymbol{\tau}_1(\tilde{s}), \sigma_1(\tilde{s})), \cdots (\boldsymbol{\tau}_N(\tilde{s}), \sigma_N(\tilde{s}))]$ where each tuple corresponds to a subject in $\mathcal{D}$. $\boldsymbol{\tau}_i(\tilde{s})$ is a binary vector of length p defined such that $\tau_i^{(j)}(\tilde{s}) = 1$ if the characteristic j will be required for determining subject i's treatment, and 0 otherwise. Further, $\sigma_i(\tilde{s})$ captures the treatment assigned to subject i. If no treatment has been assigned to i, then $\sigma_i(\tilde{s}) = 0$.

Note that we have a single start state $\tilde{s}_0$ which corresponds to an empty decision list. $\boldsymbol{\tau}_i(\tilde{s}_0)$ is a vector of 0s, and $\sigma_i(\tilde{s}_0) = 0$ for all i in $\mathcal{D}$ indicating that no treatments have been assigned to any subject, and no characteristics were deemed as requirements for assigning treatments. Furthermore, a state $\tilde{s}$ is regarded as a terminal state if for all i, $\sigma_i(\tilde{s})$ is non-zero indicating that treatments have been assigned to all the subjects.

Actions. Each action can take one of the following forms: 1) a rule $r \in \mathcal{L}$, which is a tuple of the form (pattern, treatment). Eg., (Age$\geq$40 $\wedge$ Gender=Female, T1). This specifies that subjects who obey conditions in the pattern are prescribed the treatment. Such action leads to a non-terminal state. 2) a treatment $a \in \mathcal{A}$, which corresponds to the default rule, thus this action leads to a terminal state.

Transition and Reward Functions. We have a deterministic transition function which ensures that taking an action $\tilde{a} = (\tilde{c}, \tilde{t})$ from state $\tilde{s}$ will always lead to the same state $\tilde{s}'$. Let U denote the set of all those subjects i for which treatments have already been assigned to be in state $\tilde{s}$ i.e., $\sigma_i(\tilde{s}) \neq 0$ and let U^c denote the set of all those subjects who have not been assigned treatment in the state $\tilde{s}$. Let U' denote the set of all those subjects i which do not belong to the set U and which satisfy the condition $\tilde{c}$ of action $\tilde{a}$. Let Q denote the set of all those characteristics in $\mathcal{F}$ which are present in the condition $\tilde{c}$ of action $\tilde{a}$. If action $\tilde{a}$ corresponds to a default rule, then $Q = \emptyset$ and $U' = U^c$. With this notation in place, the new state $\tilde{s'}$ can be characterized as follows: 1) $\tau_i^{(j)}(\tilde{s}') = \tau_i^{(j)}(\tilde{s})$ and $\sigma_i(\tilde{s}') = \sigma_i(\tilde{s})$ for all $i \in U$, $j \in \mathcal{F}$; 2) $\tau_i^{(j)}(\tilde{s}') = 1$ for all $i \in U^c$, $j \in Q$; 3) $\sigma_i(\tilde{s}') = \tilde{t}$ for all $i \in U'$.

Similarly, the immediate reward obtained when we reach $\tilde{s}'$ by taking $\tilde{a} = (\tilde{c}, \tilde{t})$ from the state $\tilde{s}$ can be written as:

$$\frac{\lambda_1}{N} \sum_{i \in U'} o(i, \tilde{t}) - \frac{\lambda_2}{N} \sum_{i \in U^c, j \in Q} d(j) - \frac{\lambda_3}{N} \sum_{i \in U'} d'(\tilde{t})$$

where o is defined in Eqn. 5, d and d' are cost functions for characteristics and treatments respectively.

UCT with Customized Pruning The basic idea behind the Upper Confidence Bound on Trees (UCT) [7] algorithm is to iteratively construct a search tree for some pre-determined number of iterations. At the end of this procedure, the best performing policy or sequence of actions is returned as the output. Each node in the search tree corresponds to a state in the MDP state space and the links in the tree correspond to the actions. UCT employs the UCB-1 metric [3] for navigating through the search space.

We employ a UCT-based algorithm for finding the optimal policy of our MDP formulation, though we leverage customized checks to further guide the exploration process and prune the search space. Recall that each non-terminal state in our state space corresponds to a partial decision list. We exploit the fact that we can upper-bound the value of the objective for any given partial decision list. The upper bound on the objective for any given non-terminal state $\tilde{s}$ can be computed by approximating the reward as follows: 1) all the subjects who have not been assigned treatments will get the best

Bail Dataset

# of Data Points	86152
Characteristics & Costs	age, gender, previous offenses, prior arrests, current charge, SSN (cost = 1) marital status, kids, owns house, pays rent addresses in past years (cost = 2) mental illness, drug tests (cost = 6)
Treatments & Costs	release on personal recognizance (cost = 20) release on conditions/bond (cost = 40)
Outcomes & Scores	no risk (score = 100), failure to appear (score = 66) non-violent crime (score = 33) violent crime (score = 0)

Table 1: Summary of datasets.

possible treatments without incurring any treatment cost 2) no additional assessments are required by any subject (and hence no additional assessment costs levied) in the population. The upper bound on the incremental reward is thus:

$$\text{upper bound}(U^c) = \lambda_1 \frac{1}{N} \sum_{i \in U^c} \max_t o(i,t).$$

During the execution of UCT procedure, whenever there is a choice to be made about which action needs to be taken, we employ checks based on the upper bound of the objective value of the resulting state. Consider a scenario in which the UCT procedure is currently in state $\tilde{s}$ and needs to choose an action. For each possible action $\tilde{a}$ (that does not correspond to a default rule[2]) from state $\tilde{s}$, we determine the upper bound on the objective value of the resulting state $\tilde{s}'$. If this value is less than either the highest value encountered previously for a complete rule list, or the objective value corresponding to the best default action from the state $\tilde{s}$, then we *block* the action $\tilde{a}$ from the state $\tilde{s}$. This state is provably suboptimal.

3 Experimental Evaluation

Here, we discuss the detailed experimental evaluation of our framework. First we analyze the outcomes obtained and costs incurred when recommending treatments using our approach. Then, we qualitatively analyze the treatment regime produced by our framework.

If Gender=F **and** Current-Charge =Minor Prev-Offense=None **then** RP

Else if Prev-Offense=Yes **and** Prior-Arrest =Yes **then** RC

Else if Current-Charge =Misdemeanor **and** Age$\leq$ 30 **then** RC

Else if Age$\geq$ 50 **and** Prior-Arrest=No, **then** RP

Else if Marital-Status=Single **and** Pays-Rent =No **and** Current-Charge =Misd. **then** RC

Else if Addresses-Past-Yr$\geq$ 5 **then** RC

Else RP

Figure 1: Treatment regime for bail data; RP refers to milder form of treatment: release on personal recognizance, and RC is release on conditions/bond which is comparatively harsher.

Dataset. Our dataset consists of information pertaining to bail decisions of about 86K defendants (see Table 1). It captures information about various defendant characteristics such as demographic attributes, past criminal history (cost = 1), personal (cost = 2) and health related information (cost = 6) for each of the 86K defendants. Further, the decisions made by judges in each of these cases: release on personal recognizance (cost = 20) / release on conditions/bond (cost = 40), and the corresponding outcomes (e.g., if a defendant committed another crime when out on bail) are also available. The characteristics that were harder to obtain were assigned higher costs compared to the ones

[2]We can compute exact values of objective function if the action is a default rule because the corresponding decision list is fully constructed.

	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7
IPTL	77.6	14.53	35.23	8.57	9
MCA	73.4	19.03	35.48	12.03	-
OWL (Gaussian)	72.9	28	35.18	13	-
OWL (Linear)	71.3	28	34.23	13	-
Human	69.37	-	33.39	-	-

Table 2: Results for Treatment Regimes. Our approach: CITR; Baselines: IPTL, MCA, OWL; Human refers to the setting where judges assigned treatments (or made decisions).

that were readily available. Similarly, the treatment that placed a higher burden on the defendant and/or the system (release on conditions/bond) was assigned a higher cost. When assigning scores to outcomes, undesirable scenarios (e.g., violent crime when released on bail) received lower scores.

Baselines. We compared our framework to the following state-of-the-art treatment recommendation approaches: 1) Outcome Weighted Learning (OWL) [17] 2) Modified Covariate Approach (MCA) [14] 3) Interpretable and Parsimonious Treatment Regime Learning (IPTL) [16]. While none of these approaches explicitly account for treatment costs or costs required for gathering the subject characteristics, MCA and IPTL minimize the number of characteristics/covariates required for deciding the treatment of any given subject. OWL, on the other hand, often uses all the characteristics available in the data when assigning treatments.

Quantitative Analysis. We analyzed the performance of our approach CITR (Cost-effective, Interpretable Treatment Regimes) with respect to various metrics such as: average outcome obtained (Avg. Outcome), average assessment and treatment costs (Avg. Assess Cost, Avg. Treat Cost), average no. of characteristics (Avg. # of Characs.) used to determine treatment of any given defendant, and number of rules in the rule list (List Len). These results are shown in Table 2. It can be seen that the treatment regimes produced by our approach results in better average outcomes with lower costs. It is also interesting that our approach produces more concise lists with fewer rules compared to the baselines. While the treatment costs of all the baselines are similar, there is some variation in the average assessment costs and the outcomes. IPTL turns out to be the best performing baseline in terms of the average outcome, average assessment costs, and average number of characteristics. The last line of Table 2 shows the average outcomes and the average treatment costs computed empirically on the observational data. These statistics (in the last line) represent the outcomes and costs corresponding to the decisions made by human judges. It is interesting to note that the regimes learned by algorithmic approaches perform better than human experts.

Qualitative Analysis. The treatment regime produced by our approach on the bail dataset is shown in Figure 1. The constructed regime is able to achieve good outcomes without even using the most expensive characteristics such as mental illness tests and drug tests. Personal information characteristics, which are slightly more expensive than defendant demographics and prior criminal history, appear only towards the end of the list and these checks apply only to 21.23% of the population. It is interesting that the regime uses the defendant's criminal history as well as personal and demographic information to make recommendations. For instance, females with minor current charges (such as driving offenses) and no prior criminal records are typically released on bail without conditions such as bonds or checking in with the police. On the other hand, defendants who have committed crimes earlier are only granted conditional bail.

4 Conclusions

In this work, we proposed a framework for learning cost-effective, interpretable treatment regimes from observational data. To the best of our knowledge, this is the first solution to the problem at hand that addresses all of the following aspects: 1) maximizing the outcomes 2) minimizing treatment costs, and costs associated with gathering information required to determine the treatment 3) expressing regimes using an interpretable model. We modeled the problem of learning a treatment regime as a MDP and employed a variant of UCT which prunes the search space using customized checks. We demonstrated the effectiveness of our framework on real world data comprising of bail decisions.

References

[1] E.-M. Abulesz and G. Lyberatos. Novel approach for determining optimal treatment regimen for cancer cemotherapy. *International journal of systems science*, 19(8):1483–1497, 1988.

[2] J. Bien and R. Tibshirani. Classification by set cover: The prototype vector machine. *arXiv preprint arXiv:0908.2284*, 2009.

[3] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.

[4] J. A. Dorresteijn, F. L. Visseren, P. M. Ridker, A. M. Wassink, N. P. Paynter, E. W. Steyerberg, Y. van der Graaf, and N. R. Cook. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *Bmj*, 343:d5888, 2011.

[5] R. B. D'Agostino. Estimating treatment effects using observational data. *Jama*, 297(3):314–316, 2007.

[6] A. Fan, W. Lu, R. Song, et al. Sequential advantage selection for optimal treatment regime. *The Annals of Applied Statistics*, 10(1):32–53, 2016.

[7] L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[8] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, 2016.

[9] H. Lakkaraju and C. Rudin. Learning cost-effective treatment regimes using markov decision processes. https://arxiv.org/abs/1610.06972, 2016.

[10] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[11] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2012.

[12] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[13] J. J. McGough and S. V. Faraone. Estimating the size of treatment effects: moving beyond p values. *Psychiatry (1550-5952)*, 6(10), 2009.

[14] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

[15] M. P. Wallace and E. E. Moodie. Personalizing medicine: a review of adaptive treatment strategies. *Pharmacoepidemiology and drug safety*, 23(6):580–585, 2014.

[16] Y. Zhang, E. B. Laber, A. Tsiatis, and M. Davidian. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904, 2015.

[17] Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.