CPSC 6820 Data Science and Neural Networks, Final Project

Paper Reading

In this paper, the authors address these issues in terms of recourse, which can be defined as a person's ability to change model decisions by changing actionable input variables (e.g., income versus age or marital status). The ability to change model decisions by changing actionable input variables (e.g., income versus age or marital status). They proposed integer programming (IP) tools to ensure recourse in linear classification problems without model development. The authors show how the IP tool can inform stakeholders through experiments on credit scoring problems. The results show that recourse can be significantly affected by standard practices in model development and motivate the need to evaluate recourse in practice.

In machine learning, a recourse algorithm can be defined as the ability of a person to obtain the desired outcome from a fixed model. For example, in the setup of Figure 1, a person is denied a loan and seeks explanations and advice on how to proceed. This person has an annual salary (X1) of $75,000 and an account balance (X2) of $25,000, and the predictor grants a loan based on the binary output h = sign(X1 + 5 - X2 - $225,000). The existing methodology may identify the most recent counterfactual interpretation as another individual with an annual salary of $100 000 (+%33) or a bank balance of $30 000 (+%20), thus encouraging the individual to reapply when either of these two conditions is met. On the other hand, given that the action occurs in a world where job seekers save 30% of their salary (i.e., X2 B 3/10 - X1 + U2), a salary increase of only %14 to $85,000 would automatically result in an additional $3,000 in savings, with a net positive impact on the decision of the loan origination algorithm.

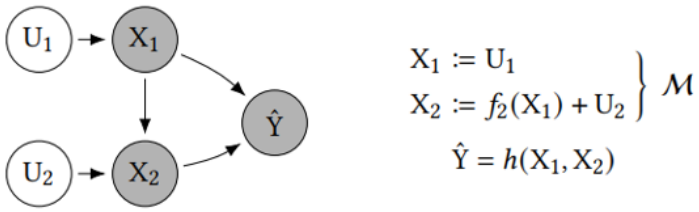

Figure 1: Illustration of an example causal generative process governing the world, showing both the graphical model, G, and the structural causal model, M. In this example, X1 represents an individual's annual salary, X2 is bank balance, and Y^ is the output of a fixed deterministic predictor h, predicting the eligibility of an individual to receive a loan.

The lack of recourse is often mentioned in calls for increased transparency and explainability in algorithmic decision-making. Yet, transparency and explainability do not provide meaningful protection with regards to recourse. In fact, even simple transparent models such as linear classifiers may not provide recourse to all of their decision subjects due to widespread practices in machine learning. These include:

• Choice of Features: A model could use features that are immutable (e.g., age ≥ 50), conditionally immutable (e.g., has_phd, which can only change from FALSE → TRUE), or should not be considered actionable (e.g., married).

• Out-of-Sample Deployment: The ability of a model to provide recourse may depend on a feature that is missing, immutable, or adversely distributed in the deployment population.

• Choice of Operating Point: A probabilistic classifier may provide recourse at a given threshold (e.g., yˆi = 1 if predicted risk of default ≥ 50%) but fail to provide recourse at a more stringent threshold (e.g., yˆi = 1 if predicted risk of default ≥ 80%).

• Drastic Changes: A model could provide recourse to all individuals but require some individuals to make drastic changes (e.g., increase income from $50K → $1M).

In this paper, the authors present tools to evaluate recourse for linear classification models, such as logistic regression models, linear SVMs, and linearizable rule-based models (e.g., rule sets, decision lists). The tools are designed to ensure recourse without interfering in model development. To this end, they aim to answer questions such as:

• Will a model provide recourse to all its decision subjects?

• How does the difficulty of recourse vary in a population of interest?

• What can a person change to obtain a desired prediction from a particular model?

The authors develop an efficient routine to solve this optimization problem, by expressing it as an integer program (IP) and handing it to an IP solver (e.g., CPLEX or CBC). And then they use the routine to create the following tools:

1. A procedure to evaluate the feasibility and difficulty of recourse for a linear classifier over its target population. Given a classifier and a sample of feature vectors from a target population, the procedure estimates the feasibility and difficulty of recourse in the population by solving the optimization problem for each point that receives an undesirable prediction. This procedure provides a way to check recourse in model development, procurement, or impact assessment.
2. A method to generate a list of actionable changes for a person to obtain a desired outcome from a linear classifier. Referring to this list as a flipset and present an example in Figure 2. In the United States, the Equal Opportunity Credit Act requires that any person who is denied credit is sent an adverse action notice explaining “the principal reason for the denial." It is well-known that adverse action notices may not provide actionable information. By including a flipset in an adverse action notice, a person would know a set of exact changes to be approved in the future.

The authors consider a standard classification problem where each person is characterized by a feature vector $x = [1, x_1 \ldots x_d] \subseteq X_0 \cup \ldots \cup X_d = X \subseteq R^d+1$ and a binary

label $y \in \{-1, +1\}$. It is assumed that they are given a linear classifier $f(x) = \text{sign}(\langle w, x \rangle)$ where $w = [w_0, w_1, \ldots, w_d] \subseteq R^d+1$ is a vector of coefficients and w_0 is the intercept. They denote the desired outcome as $\hat{y} = 1$, and assume that $\hat{y} = 1$ $[\langle w, x \rangle \geq 0]$. Given a person who is assigned an undesirable outcome $f(x) = -1$, and aim to find an action a such that $f(x + a) = +1$ by solving an optimization problem of the form.

Features to Change	Current Values		Required Values
n_credit_cards	5	⟶	3
current_debt	$3,250	⟶	$1,000
has_savings_account	FALSE	⟶	TRUE
has_retirement_account	FALSE	⟶	TRUE

Figure. 2. Hypothetical flipset for a person who is denied credit by a classifier. Each row (item) describes how a subset of features that the person can change to "flip" the prediction of the model from $\hat{y} = -1 \rightarrow +1$.

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

Here:

• $A(x)$ is a set of feasible actions from x. Each action is a vector $a = [0, a_1, \ldots, a_d]$ where $a_j \in A_j(x_j) \subseteq \{a_j \in R \mid a_j + x_j \in X_j\}$. It is said that a feature j is immutable if $A_j(x) = \{0\}$. And feature j is conditionally immutable if $A_j(x) = \{0\}$ for some $x \in X$.

• $\text{cost}(\cdot\,; x) : A(x) \rightarrow R+$ is a cost function to choose between feasible actions, or to measure quantities of interest in a recourse audit. It is assumed that cost functions satisfy the following properties: (i) $\text{cost}(0; x) = 0$ (no action ⇔ no cost); (ii) $\text{cost}(a; x) \leq \text{cost}(a + \epsilon x_j\,; x)$ (larger actions ⇔ higher cost). Solving (1) allows us to make one of the following claims related to recourse:

• If (1) is feasible, then its optimal solution $a*$ is the minimal-cost action to flip the prediction of x.

• If (1) is infeasible, then no action can attain a desired outcome from x. Thus, we have certified that the model f does not provide actionable recourse for a person with features x.

And here are three remarks to guarantee the model feasibility.

Remark 1. A linear classifier provides recourse to all individuals if it only uses actionable features and does not predict a single class.

Remark 2. If all features are unbounded, then a linear classifier with at least one actionable feature provides recourse to all individuals.

Remark 3. If all features are bounded, then a linear classifier with at least one immutable feature may deny recourse to some individuals.

Considering a discretized version of the optimization problem in (1), which can be expressed as an integer program (IP) and optimized with a solver. This approach has several benefits: (i) it can directly search over actions for binary, ordinal, and categorical features; (ii) it can optimize non-linear and non-convex cost functions; (iii) it allows users to customize the set of feasible actions; (iv) it can quickly find a globally optimal solution or certify that a classifier does not provide recourse.

The optimization problem in (1) can be expressed as an IP of the form:

$$\min \ \text{cost}$$

$$\text{s.t.} \ \ \text{cost} = \sum_{j \in J_A} \sum_{k=1}^{m_j} c_{jk} v_{jk} \tag{2a}$$

$$\sum_{j \in J_A} w_j a_j \geq -\sum_{j=0}^{d} w_j x_j \tag{2b}$$

$$a_j = \sum_{k=1}^{m_j} a_{jk} v_{jk} \qquad j \in J_A \tag{2c}$$

$$1 = u_j + \sum_{k=1}^{m_j} v_{jk} \qquad j \in J_A \tag{2d}$$

$$a_j \in \mathbb{R} \qquad j \in J_A$$

$$u_j \in \{0, 1\} \qquad j \in J_A$$

$$v_{jk} \in \{0, 1\} \qquad j \in J_A, \ k = 1, \ldots, m_j$$

To evaluate the cost and feasibility of recourse of a linear classifier by solving IP (2) for samples drawn from a population of interest. Formally, the auditing procedure requires: (i) the coefficient vector w of a linear classifier; (ii) feature vectors sampled from the target population $\{x_i\}_{i=1}^{n}$ where $f(x_i) = -1$. It solves the IP for each x_i to produce:

• an estimate of the feasibility of recourse (i.e., the proportion of points for which the IP is feasible);

• an estimate of the distribution of the cost of recourse (i.e., the distribution of cost(a ∗ i ; x_i) where ai ∗ is the minimal-cost action from x_i).

So, it is proposed that the maximum percentile shift:

$$\text{cost}(\boldsymbol{x} + \boldsymbol{a}; \boldsymbol{x}) = \max_{j \in J_A} \left| Q_j(x_j + a_j) - Q_j(x_j) \right|. \tag{3}$$

This cost function is well-suited for auditing because it produces an informative measure of the difficulty of recourse. If the optimal cost is 0.25, for example, then any feasible action must change a feature by at least 25 percentiles. In other words, there does not exist an action that flips the prediction by changing a feature by less than 25 percentiles.

The authors construct flipsets such as the one in Figure 2 using enumeration procedure that solves IP (2) repeatedly. In Algorithm 1, they present an enumeration procedure to produce a collection of minimal-cost actions that alter distinct subsets of features. The procedure solves IP (2) to recover a minimal-cost action a∗. Next, it adds a constraint to the IP to eliminate actions that alter the same combination of features as a∗. It repeats these two steps until it has recovered T minimal-cost actions or determined that the IP is infeasible (which means that it has enumerated a minimal-cost action for each combination of features that can flip the prediction from x).

It is proposed the total log-percentile shift:

$$\text{cost}(\boldsymbol{x} + \boldsymbol{a}; \boldsymbol{x}) = \sum_{j \in J_A} \log \left(\frac{1 - Q_j(x_j + a_j)}{1 - Q_j(x_j)} \right). \tag{4}$$

Algorithm 1 can be used to create an item in a flipset by listing the current feature values xj along with the desired feature values:

Algorithm 1 Enumerate T Minimal Cost Actions for Flipset

Input

IP — instance of IP (2) for coefficients $\boldsymbol{w}$, features $\boldsymbol{x}$, and actions $A(\boldsymbol{x})$

$T \geq 1$ — number of items in flipset

Initialize

$\mathcal{A} \leftarrow \emptyset$ — actions shown in flipset

repeat

$\boldsymbol{a}^* \leftarrow$ optimal solution to IP

$\mathcal{A} \leftarrow \mathcal{A} \cup \{\boldsymbol{a}^*\}$ — *add $\boldsymbol{a}^*$ to set of optimal actions*

$S \leftarrow \{j : a_j^* \neq 0\}$ — *indices of features altered by $\boldsymbol{a}^*$*

add constraint to IP to remove actions that alter features in S:

$$\sum_{j \notin S} u_j + \sum_{j \in S} (1 - u_j) \leq d - 1.$$

until $|\mathcal{A}| = T$ or IP is infeasible

Output: $\mathcal{A}$ — actions shown in flipset

The results of our audit in Figure 3, and present a flipset for a person who is denied credit by the most accurate classifier in Figure 4. As shown in Figure 3, tuning the $\ell 1$ -penalty has a minor effect on test error, but a major effect on the feasibility and cost recourse. In particular, classifiers with small $\ell 1$ -penalties provide all individuals with recourse. As the $\ell 1$ -penalty increases, however, the number of individuals with recourse decreases as regularization reduces the number of actionable features.

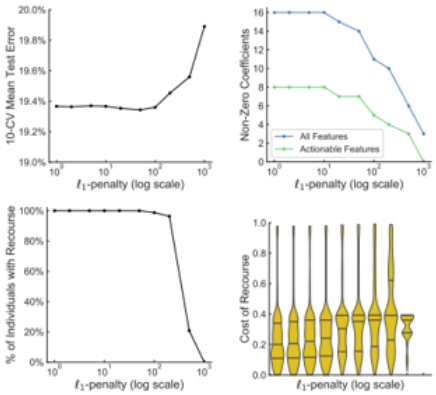

Figure. 3. Performance, sparsity, and recourse of $\ell 1$ -penalized logistic regression models for the credit dataset. It is shown the mean 10-CV test error (top left), the number of non-zero coefficients (top right), the proportion of individuals with recourse in the training data (bottom left), and the distribution of the cost of recourse in the training data (bottom right).

Features to Change	Current Values		Required Values
MostRecentPaymentAmount	\$0	⟶	\$790
MostRecentPaymentAmount	\$0	⟶	\$515
MonthsWithZeroBalanceOverLast6Months	1	⟶	2
MonthsWithZeroBalanceOverLast6Months	1	⟶	4
MostRecentPaymentAmount	\$0	⟶	\$775
MonthsWithLowSpendingOverLast6Months	6	⟶	5
MostRecentPaymentAmount	\$0	⟶	\$500
MonthsWithLowSpendingOverLast6Months	6	⟶	5
MonthsWithZeroBalanceOverLast6Months	1	⟶	2

Figure. 4. Flipset for a person who is denied credit by the most accurate classifier built for the credit dataset. Each item shows a minimal-cost action that a person can make to obtain credit.

The results of our audit in Figure 5 and show flipsets for a prototypical young adult in Figure 6. As shown, the median cost of recourse among young adults under the biased model is 0.66, which means that the median person can only flip their predictions by a 66 percentile shift in any feature. In comparison, the median cost of recourse among young adults under the baseline model is 0.14. These differences in the cost of recourse are less pronounced for other age brackets.

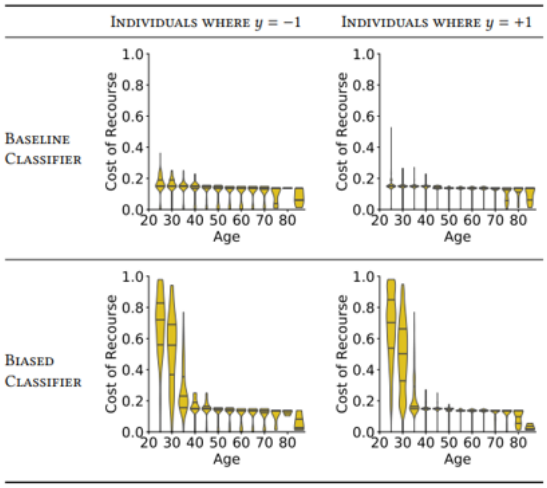

Figure. 5. Distributions of the cost of recourse in the target population for classifiers conditioned on the true outcome y. It is shown the distribution of the cost of recourse for the biased classifier (top) and the baseline classifier (bottom) for true negatives (left) and false negatives (right). The cost of recourse for young adults is significantly higher for the biased classifier, regardless of their true outcome.

Baseline Classifier

Features to Change	Current Values		Required Values
NumberOfTime30-59DaysPastDueNotWorse	1	⟶	0
NumberOfTime60-89DaysPastDueNotWorse	0	⟶	1
NumberRealEstateLoansOrLines	2	⟶	1
NumberOfOpenCreditLinesAndLoans	11	⟶	12
RevolvingUtilizationOfUnsecuredLines	35.89%	⟶	36.63%

Biased Classifier

Feature	Current Value		Required Value
NumberOfTime30-59DaysPastDueNotWorse	1	⟶	0
NumberOfTime60-89DaysPastDueNotWorse	0	⟶	1

Figure. 6. Flipsets for a young adult with Age = 28 under the biased classifier (top) and the baseline classifier (bottom). The flipset for the biased classifier has 1 item while the flipset for the baseline classifier has 4 items.

As shown in Figure 7, the cost of recourse can differ between males and females even when models ignore gender. These disparities can also be examined by comparing flipsets as in Figure 8, which shows minimal-cost actions for comparable individuals from each protected group.

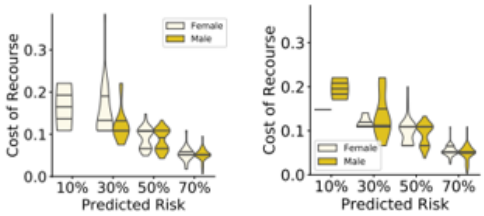

Figure. 7. Distribution of the cost of recourse for males and females with y = −1 (left) and y = +1 (right)

Female with $y_i = +1$ and $\Pr(y_i = +1) = 34.0\%$

Features to Change	Current Values		Required Values
LoanAmount	$7 432	⟶	$3 684
LoanDuration	36 months	⟶	25 months
CheckingAccountBalance ≥ 200	FALSE	⟶	TRUE
SavingsAccountBalance ≥ 100	FALSE	⟶	TRUE
HasGuarantor	FALSE	⟶	TRUE
LoanAmount	$7 432	⟶	$3,684
LoanDuration	36 months	⟶	23 months
LoanRateAsPercentOfIncome	2.00%	⟶	1.00%
HasTelephone	FALSE	⟶	TRUE
HasGuarantor	FALSE	⟶	TRUE
LoanAmount	$7432	⟶	$912
LoanDuration	36 months	⟶	7 months
HasTelephone	FALSE	⟶	TRUE

Male with $y_i = +1$ and $\Pr(y_i = +1) = 32.1\%$

Features to Change	Current Values		Required Values
LoanAmount	$15 857	⟶	$7 968
LoanDuration	36 months	⟶	32 months
CheckingAccountBalance ≥ 200	FALSE	⟶	TRUE
HasCoapplicant	TRUE	⟶	FALSE
HasGuarantor	FALSE	⟶	TRUE
Unemployed	TRUE	⟶	FALSE
LoanAmount	$15 857	⟶	$7 086
LoanDuration	36 months	⟶	29 months
CheckingAccountBalance ≥ 200	FALSE	⟶	TRUE
HasCoapplicant	TRUE	⟶	FALSE
HasGuarantor	FALSE	⟶	TRUE
LoanAmount	$15 857	⟶	$4 692
LoanDuration	36 months	⟶	29 months
CheckingAccountBalance ≥ 200	FALSE	⟶	TRUE
SavingsAccountBalance ≥ 100	FALSE	⟶	TRUE
LoanAmount	$15 857	⟶	$3 684
LoanDuration	36 months	⟶	21 months
HasTelephone	FALSE	⟶	TRUE

Figure. 8. Flipsets for a matched pair of individuals from each protected group. Individuals have the same true outcome y_i and similar levels predicted risk Pr ($y_i = +1$).

Code Reproduce

The Code link:
https://github.com/swang8record/Course_Project/blob/main/CPSC%206820%20Data%20Science%20and%20Neural%20Networks/Actionable%20Recourse%20in%20Linear%20Classification%20Reproduce.ipynb