Algorithmic Recourse

Giovanni De Toni
giovanni.detoni@unitn.it

Advanced Topics in Machine Learning and Optimization
30th November 2022

Why do we need explanations?

- **Recidivism risk** (Dressel & Farid [8])
- **University admissions** (Waters & Miikkulainen [30])
- **Rejecting/Accepting a job applicant** (Liem et al. [15])
- **Prescribing medications and treatments** (Yoo et al. [31])
- Many others...

Why do we need explanations?

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission (Caruana et al. [4])

Why do we need explanations?

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission (Caruana et al. [4])

Predict pneumonia risk based on user features (and thus the need for hospitalization).

Why do we need explanations?

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission (Caruana et al. [4])

Predict pneumonia risk based on user features (and thus the need for hospitalization).

Rule discovered by an **RBL model**:

$$HasAsthma(x) \Rightarrow LowRisk(x)$$

Why do we need explanations?

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission (Caruana et al. [4])

Predict pneumonia risk based on user features (and thus the need for hospitalization).

Rule discovered by an **RBL model**:

$$HasAsthma(x) \Rightarrow LowRisk(x)$$

Are neural networks safer to use?

Why do we need explanations?

Why do we need explanations?

- **Example-based explanations**
  - Prototype and criticism (Kim et al. [14])
- **(Local/Global) Model-agnostic explanations**
  - SHAP (Lundberg & Lee [16])
  - LIME (Ribeiro et al. [21])
- **Counterfactual explanations** (Wachter et al. [29])
- **Interpretable Models** (e.g., decision trees, linear models)
- See surveys on the topic (Adabi & Berrada [1])

Why do we need explanations?

- **Example-based explanations**
  - Prototype and criticism (Kim et al. [14])
- **(Local/Global) Model-agnostic explanations**
  - SHAP (Lundberg & Lee [16])
  - LIME (Ribeiro et al. [21])
- **Counterfactual explanations** (Wachter et al. [29])
- **Interpretable Models** (e.g., decision trees, linear models)
- See surveys on the topic (Adabi & Berrada [1])

These methods mostly target machine learning practitioners and researchers!

Explainability as ”right to an explanation”

In reality, a user wants to know how to act to appeal to or change a potentially negative decision.

We need to consider *”explanations as a means to help a data-subject act rather than merely understand”* [29]

It is also defined as a requirement by the GDPR [27]

Algorithmic Recourse

Definition 1 (Algorithmic Recourse, adapted from [25])

Algorithmic recourse is the systematic process of reversing unfavourable decisions by algorithms and bureaucracies across a range of counterfactual scenarios.

Definition 2

A **counterfactual explanation** (CFE) is a statement about ”how the world would have (had) to be different for a desirable outcome to happen”.

We are usually interested in **nearest counterfactual explanations**, the most similar instances of the feature vector that change the prediction of the classifier.

Counterfactual Explanations

$$\mathbf{x} := \{x_0, \ldots, x_n\} \quad x \in \mathcal{X}$$
$$h : \mathcal{X} \rightarrow \{0, 1\}$$
$$d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

Counterfactual Explanations

$$
\begin{aligned}
\mathbf{x} &:= \{x_0, \ldots, x_n\} \quad x \in \mathcal{X} \\
h &: \mathcal{X} \rightarrow \{0, 1\} \\
d &: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{x}^* = &\operatorname*{arg\,min}_{\mathbf{x}'} \quad d(\mathbf{x}, \mathbf{x}') \\
\text{s.t.} \quad &h(\mathbf{x}) \neq h(\mathbf{x}^*)
\end{aligned} \tag{1}
$$

Counterfactual Explanations

- CFEs are **model-agnostic**
- CFEs do not need to be instances from the training data
- CFEs are **human-friendly** explanations
  - Both **contrastive** and **selective**
- CFEs are relatively easy to find (e.g., minimizing a loss function)

Counterfactual Explanations

Wachter et al. [29] provides a loss function to learn CFEs.

$$\mathcal{L}(\mathbf{x}, \mathbf{x}', y', \lambda) = \lambda(h(\mathbf{x}') - y')^2 + d(\mathbf{x}, \mathbf{x}')$$

$$d(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{n} \frac{|x_i' - x_i|}{MAD_i} \tag{2}$$

Counterfactual Explanations

Wachter et al. [29] provides a loss function to learn CFEs.

$$\mathcal{L}(\mathbf{x}, \mathbf{x}', y', \lambda) = \lambda(h(\mathbf{x}') - y')^2 + d(\mathbf{x}, \mathbf{x}')$$
$$d(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{n} \frac{|x_i' - x_i|}{MAD_i} \tag{2}$$

$$\mathbf{x}^* = \arg\min_{x' \in \mathcal{X}} \max_{\lambda \in \mathbb{R}} \quad \lambda(h(\mathbf{x}') - y')^2 + d(\mathbf{x}, \mathbf{x}') \tag{3}$$

Counterfactual Explanations

There are already many research works on how to build CFEs:

- **Multi-objective Counterfactual Explanations** [5]
- **Counterfactual Explanations under uncertainty** [24]
- **MACE** (Karimi et al. [13])
- **LORE** (Guidotti, Monreale, Ruggieri, Pedreschi, et al. [9])
- **DICE** (Mothilal et al. [17])
- **FACE** (Poyiadzi et al. [19])
- Many surveys on the topic. See Guidotti, Monreale, Ruggieri, Turini, et al. [10]

Counterfactual Explanations

- Given $\mathbf{x} \in \mathcal{X}$, there exists multiple $\mathbf{x}^*$ (Rashomon Effect)
- CFEs are not **actionable**
- CFE optimization does not consider the **feasibility**
- Prior works ignore the **causal relationship** between features.

Counterfactual Interventions

- Actionable **sequence of actions** instead of a CFE
- It defines a **cost** to mimic the **user's effort** for each action
- It considers **causal relationships** between features
- **Minimize** the cost of the sequence, such that $h(\mathbf{x}) \neq h(\mathbf{x}')$
- **Same properties** of counterfactual explanations (CFEs).

Counterfactual Interventions II

$$\mathbf{x} := \{x_0, \ldots, x_n\} \quad x \in \mathcal{X}$$

$$h : \mathcal{X} \to \{0, 1\}$$

$$a \in \mathcal{A} \qquad C : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$$

Counterfactual Interventions II

$$\mathbf{x} := \{x_0, \dots, x_n\} \quad x \in \mathcal{X}$$
$$h : \mathcal{X} \to \{0, 1\}$$
$$a \in \mathcal{A} \qquad C : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$$

$$\begin{aligned} I^* &= \arg\min_{I \in \mathcal{I}} \sum_{t=1}^{T} C(a_t, \mathbf{x}_t) \\ \text{s.t.} \quad I^* &= \{a_t\}_{t=1}^{T} \\ \mathbf{x}_t &= I(\mathbf{x}_{t-1}) \\ h(I(\mathbf{x}_0)) &\neq h(\mathbf{x}_0) \end{aligned} \tag{4}$$

Counterfactual Interventions II

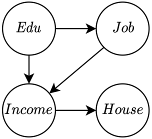
$$\mathbf{x} := \{x_0, \dots, x_n\} \quad x \in \mathcal{X}$$
$$h : \mathcal{X} \to \{0, 1\}$$
$$a \in \mathcal{A} \qquad C : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$$

$$\begin{aligned} I^* &= \arg\min_{I \in \mathcal{I}} \sum_{t=1}^{T} C(a_t, \mathbf{x}_t) \\ \text{s.t.} \quad I^* &= \{a_t\}_{t=1}^{T} \\ \mathbf{x}_t &= I(\mathbf{x}_{t-1}) \\ h(I(\mathbf{x}_0)) &\neq h(\mathbf{x}_0) \end{aligned} \tag{4}$$

Algorithmic Recourse is an **NP-Hard problem**.

Causality

(A)

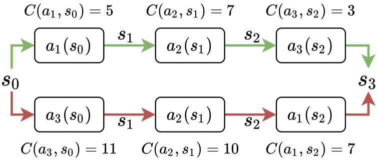


a_1 : get_degree(bachelor)

a_2 : change_job(developer)

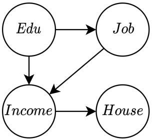
a_3 : change_house(buy)

(B)



Causality

(A)

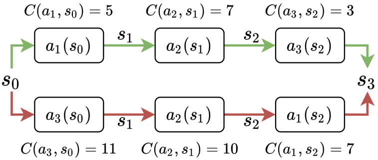


a_1 : get_degree(bachelor)

a_2 : change_job(developer)

a_3 : change_house(buy)

(B)



Theorem 3 (Adapted from 3.2 in [22])

Unless we are intervening on variables without descendants, algorithmic recourse can be guaranteed only if the structural equations are known, no matter the amount or the type of available data.

Counterfactual Interventions

- **Recourse in linear classification** (Spangher et al. [23])
- **SYNTH** (Ramakrishnan et al. [20])
- **CSCF** (Naumann & Ntoutsi [18])
- **FastAR** (Verma et al. [26])
- **FARE** (De Toni, Lepri, & Passerini [6])
- See several surveys on the topic (e.g., Karimi et al. [12])

Counterfactual Interventions

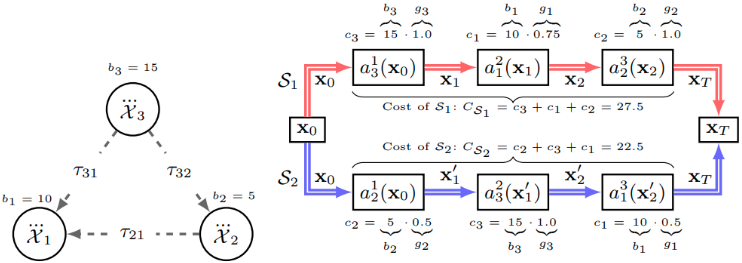
- **Recourse in linear classification** (Spangher et al. [23])
- **SYNTH** (Ramakrishnan et al. [20])
- **CSCF** (Naumann & Ntoutsi [18])
- **FastAR** (Verma et al. [26])
- **FARE** (De Toni, Lepri, & Passerini [6])
- See several surveys on the topic (e.g., Karimi et al. [12])

CSCF (Naumann & Ntoutsi [18])

$$\min_{\mathcal{S}}\Big(\underbrace{o_1}_{\text{Sequence cost}}, \underbrace{o_2}_{\text{Gower's distance}}, \underbrace{o_{2+1}, \ldots, o_{2+h}, \ldots, o_{2+d}}_{\text{Feature tweaking frequencies}}\Big)$$

$$\text{s.t. } f(\mathbf{x}_T) = \texttt{accept} \quad \text{and} \bigwedge_{(a_i, v_i) \in \mathcal{S}} \mathbb{C}_i$$

CSCF (Naumann & Ntoutsi [18])



(a) Feature relationship graph $\mathcal{G}$ (b) Different sequences $\mathcal{S}_1$ (red) and $\mathcal{S}_2$ (blue)

Fig. 2. For simplicity, the $\tau(\cdot)$ functions in (a) are based on binary conditions: $\tau_{32} = 1.0$ if $\mathcal{X}_3 := \texttt{US}$, else 0.5. $\tau_{31} = 0.5$ if $\mathcal{X}_3 := \texttt{US}$, else 1.0. $\tau_{21} = 0.5$ if $\mathcal{X}_2 \geq \texttt{BSc}$, else 1.0. As a reference, the action efforts b_i are provided above each feature in (a).

CSCF (Naumann & Ntoutsi [18])

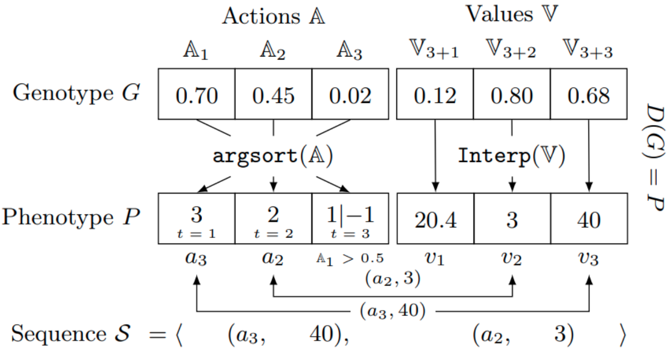


Fig. 3. Anatomy and representation of the solution decoding.

CSCF (Naumann & Ntoutsi [18])

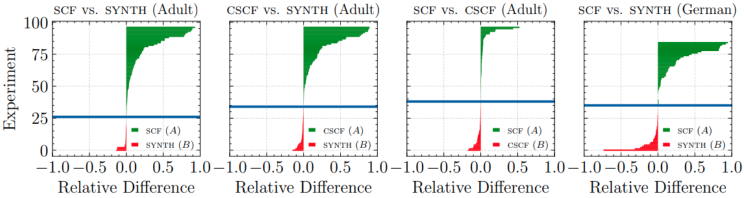
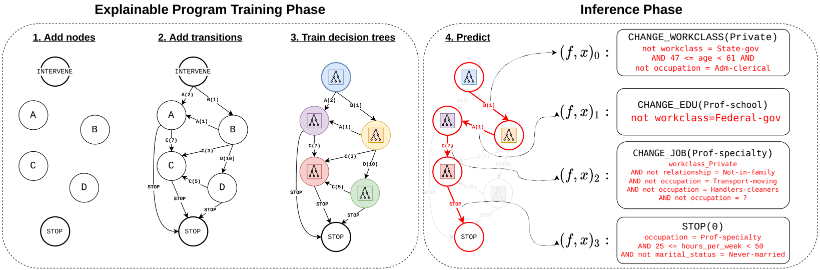


Fig. 4. Relative minimal sequence cost (o_1) differences between the three methods for both datasets and solutions with $T \leq 2$. It is computed as: $(B - A)/\max\{A, B\}$.

Counterfactual Interventions

- Current methods rely on **optimization techniques**
- **Run them ex-novo** for each user (might be a costly process)
- Fail to explain **why** we are suggesting each intervention (Barocas et al. [2])
- **Limitations of CFE-based recourse** (Karimi et al. [12])

FARE (De Toni, Lepri, & Passerini [6])



Future Directions

- **Jointly train end-to-end models providing both predictions and interventions.** See VCNet (Guyomard et al. [11])

Future Directions

- **Jointly train end-to-end models providing both predictions and interventions**. See VCNet (Guyomard et al. [11])
- **Human-in-the-Loop Counterfactual Intervention Generation**. Eliciting user preferences over the actions (De Toni, Viappiani, et al. [7])

Future Directions

- **Jointly train end-to-end models providing both predictions and interventions**. See VCNet (Guyomard et al. [11])
- **Human-in-the-Loop Counterfactual Intervention Generation**. Eliciting user preferences over the actions (De Toni, Viappiani, et al. [7])
- **Validation with real-users of counterfactual interventions** See ”One counterfactual does not make an explanation” (Butz et al. [3])

Future Directions

- **Jointly train end-to-end models providing both predictions and interventions**. See VCNet (Guyomard et al. [11])
- **Human-in-the-Loop Counterfactual Intervention Generation**. Eliciting user preferences over the actions (De Toni, Viappiani, et al. [7])
- **Validation with real-users of counterfactual interventions** See "One counterfactual does not make an explanation" (Butz et al. [3])
- **Fairness of Algorithmic Recourse**. See "On the fairness of causal algorithmic recourse" (von Kügelgen et al. [28]).

Questions?

References I

Adabi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. *IEEE Access*, *6*, 52138–52160.

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 80–89).

Butz, R., Hommersom, A., Barenkamp, M., & van Ditmarsch, H. (n.d.). One counterfactual does not make an explanation.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In (p. 1721–1730). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2783258.2788613` doi: 10.1145/2783258.2788613

References II

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In *International conference on parallel problem solving from nature* (pp. 448–469).

De Toni, G., Lepri, B., & Passerini, A. (2022). Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis. *arXiv preprint arXiv:2201.07135*.

De Toni, G., Viappiani, P., Lepri, B., & Passerini, A. (2022). Generating personalized counterfactual interventions for algorithmic recourse by eliciting user preferences. *arXiv preprint arXiv:2205.13743*.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*.

References III

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Guyomard, V., Fessant, F., Guyet, T., Bouadi, T., & Termier, A. (2022). Vcnet: A self-explaining model for realistic counterfactual generation. In *Proceedings of the european conference on machine learning and principles and practice of knowledge discovery in databases (ecml pkdd).*

Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 353–362).

Karimi, A.-H., Von Kügelgen, J., Schölkopf, B., & Valera, I. (2020). Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, *33*, 265–277.

References IV

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, *29*.

Liem, C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607–617).

References V

Naumann, P., & Ntoutsi, E. (2021). Consequence-aware sequential counterfactual generation. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 682–698).

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). Face: feasible and actionable counterfactual explanations. In *Proceedings of the aaai/acm conference on ai, ethics, and society* (pp. 344–350).

Ramakrishnan, G., Lee, Y. C., & Albarghouthi, A. (2020). Synthesizing action sequences for modifying model decisions. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 5462–5469).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " 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* (pp. 1135–1144).

References VI

Schölkopf, B., & von Kügelgen, J. (2022). From statistical to causal learning. *arXiv preprint arXiv:2204.00607*.

Spangher, A., Ustun, B., & Liu, Y. (2018). Actionable recourse in linear classification. In *Proceedings of the 5th workshop on fairness, accountability and transparency in machine learning.*

Tsirtsis, S., De, A., & Rodriguez, M. (2021). Counterfactual explanations in sequential decision making under uncertainty. *Advances in Neural Information Processing Systems*, *34*, 30127–30139.

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 284–293).

References VII

Verma, S., Hines, K., & Dickerson, J. P. (2022). Amortized generation of sequential algorithmic recourses for black-box models. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 8512–8519).

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, *10*(3152676), 10–5555.

von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., & Schölkopf, B. (2022). On the fairness of causal algorithmic recourse. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 9584–9594).

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, *31*, 841.

Waters, A., & Miikkulainen, R. (2014). Grade: Machine learning support for graduate admissions. *Ai Magazine*, *35*(1), 64–64.

References VIII

Yoo, T., Ryu, I., Lee, G., Kim, Y., Kim, J., Lee, I., ... Rim, T. (2019). *Adopting machine learning to automatically identify candidate patients for corneal refractive surgery. npj digit med 2: 59.*