The data subject should have the right [...] to obtain an explanation of the decision reached [...] and to challenge the decision.
Correctional Offender Management Profiling for Alternative Sanctions

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.

https://en.wikipedia.org/wiki/COMPAS_(software)
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida’s Broward County, and other jurisdictions.

https://en.wikipedia.org/wiki/CMPAS_(software)
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida’s Broward County, and other jurisdictions.

https://en.wikipedia.org/wiki/COMPAS_(software)
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida’s Broward County, and other jurisdictions.

https://en.wikipedia.org/wiki/COMPAS_(software)
A Machine Learning Primer

Features

\[ X = [X_1, \ldots, X_d]; \]
\[ x \in \mathbb{X}, \text{ e.g., } \mathbb{X} = \mathbb{R}^d \]

Measurements

\[ \hat{X} = [\hat{X}_1, \ldots, \hat{X}_d]; \]
\[ x \in \mathbb{X} \]

Labels

\[ Y \text{ with } y \in \mathbb{Y}, \text{ e.g., } \mathbb{Y} = \{0, 1\} \]

Hypothesis class

\[ H = \{h_1, \ldots, h_M\}, h_m: \mathbb{X} \mapsto \mathbb{Y} \]

Loss function

\[ L(h, x, y), \text{ e.g., zero-one loss.} \]

Training set

\[ S = \{(x_i, y_i)\}_{i=1}^N \text{ sampled iid from } P(\mathbb{Y}, \mathbb{X}) \]

Learning algorithm

\[ A(S, H, l) \]

Bayes-optimal hypothesis

\[ h^* \in H \text{ with (empirical) risk } L(S)(h^*) \]

Predictions

\[ \hat{Y} = h^*(\hat{X}) \]

Interventions do \{\[X\] (on features) and do \[\hat{X}\] (on measurements)\}
Features $\mathbf{X} = [X_1, \ldots, X_d]$; $x \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
A Machine Learning Primer

- Features $\mathbf{X} = [X_1, \ldots, X_d]$; $\mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\mathbf{x} \in \mathcal{X}$
Features $X = [X_1, \ldots, X_d]$; $x \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$

Measurements $\hat{X} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\hat{x} \in \mathcal{X}$

Labels $Y$ with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$

Interventions do \{X\} (on features) and do \{\hat{X}\} (on measurements)
A Machine Learning Primer

- Features $\mathbf{X} = [X_1, \ldots, X_d]$; $\mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\mathbf{x} \in \mathcal{X}$
- Labels $\mathcal{Y}$ with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
A Machine Learning Primer

- Features $\mathbf{X} = [X_1, \ldots, X_d]; \mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_d]; \mathbf{x} \in \mathcal{X}$
- Labels $\mathcal{Y}$ with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}, h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \mathbf{x}, y)$, e.g., zero-one loss.
A Machine Learning Primer

- Features $\mathbf{X} = [X_1, \ldots, X_d]; \mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_d]; \mathbf{x} \in \mathcal{X}$
- Labels $\mathcal{Y}$ with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \mathbf{x}, y)$, e.g., zero-one loss.
- Training set $\mathcal{S} = \{(x^i, y^i)\}_{i=1}^N$ sampled iid from $P(Y, X)$
Features $\mathbf{X} = [X_1, \ldots, X_d]; \mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$

Measurements $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_d]; \mathbf{x} \in \mathcal{X}$

Labels $\mathcal{Y}$ with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$

Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m: \mathcal{X} \mapsto \mathcal{Y}$

Loss function $L(h, x, y)$, e.g., zero-one loss.

Training set $S = \{(x^i, y^i)\}_{i=1}^N$ sampled iid from $P(\mathcal{Y}, \mathbf{X})$

Learning algorithm $A(S, \mathcal{H}, l)$
A Machine Learning Primer

- Features $\mathbf{X} = [X_1, \ldots, X_d]$; $\mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\mathbf{x} \in \mathcal{X}$
- Labels $\mathcal{Y}$ with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \mathbf{x}, y)$, e.g., zero-one loss.
- Training set $\mathcal{S} = \{(\mathbf{x}^i, y^i)\}_{i=1}^N$ sampled iid from $P(\mathcal{Y}, \mathcal{X})$
- Learning algorithm $A(\mathcal{S}, \mathcal{H}, l)$
- Bayes-optimal hypothesis $h^* \in \mathcal{H}$ with (empirical) risk $L(\mathcal{S})(h^*)$
A Machine Learning Primer

- Features $\mathbf{X} = [X_1, \ldots, X_d]$; $\mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_d]$; $\mathbf{x} \in \mathcal{X}$
- Labels $\mathcal{Y}$ with $y \in \mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}$, $h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, \mathbf{x}, y)$, e.g., zero-one loss.
- Training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ sampled iid from $P(Y, \mathbf{X})$
- Learning algorithm $A(S, \mathcal{H}, l)$
- Bayes-optimal hypothesis $h^* \in \mathcal{H}$ with (empirical) risk $L(S)(h^*)$
- Predictions $\hat{\mathcal{Y}} = h^*(\hat{\mathbf{X}})$
A Machine Learning Primer

- Features $X = [X_1, \ldots, X_d]; \ x \in \mathcal{X}, \text{ e.g., } \mathcal{X} = \mathbb{R}^d$
- Measurements $\hat{X} = [\hat{X}_1, \ldots, \hat{X}_d]; \ x \in \mathcal{X}$
- Labels $Y$ with $y \in \mathcal{Y}, \text{ e.g., } \mathcal{Y} = \{0, 1\}$
- Hypothesis class $\mathcal{H} = \{h_1, \ldots, h_M\}, \ h_m : \mathcal{X} \mapsto \mathcal{Y}$
- Loss function $L(h, x, y), \text{ e.g., zero-one loss.}$
- Training set $S = \{(x^i, y^i)\}_{i=1}^N$ sampled iid from $P(Y, X)$
- Learning algorithm $A(S, \mathcal{H}, l)$
- Bayes-optimal hypothesis $h^* \in \mathcal{H}$ with (empirical) risk $L(S)(h^*)$
- Predictions $\hat{Y} = h^*(\hat{X})$
- Interventions do $\{X = ()\}$ (on features) and do $\{\hat{X} = ()\}$ (on measurements)
Example: The Permutation Feature Importance Score

Idea: Assess how removing each individual features affects model performance.

\[
PFI(X_i) = L(h^*, \{X \setminus X_i, \tilde{X}_i\}, Y) - L(h^*, X, Y)
\]

with \( \tilde{X}_i \sim P(X_i) \) and \( \tilde{X} \perp \perp \{X \setminus X_i, Y\} \).

The xAI/IML Zoo

xAI/IML is a Causal Problem
xAI/IML is a Causal Problem
xAI/IML is a Causal Problem
xAI/IML is a Causal Problem
xAI/IML is a Causal Problem

World

\[ X_1 \rightarrow \hat{X}_1 \]
\[ X_2 \rightarrow \hat{X}_2 \]
\[ X_3 \rightarrow \hat{X}_3 \]
\[ X_4 \rightarrow \hat{X}_4 \]

\[ h^*(.) \rightarrow \hat{Y} \]
xAI/IML is a Causal Problem

\[ h^*(.): \hat{X}_1, \hat{X}_2, \hat{X}_3, \hat{X}_4 \rightarrow \hat{Y} \]
xAI/IML is a Causal Problem
Structural Causal Models (SCMs)

For a given set of variables $X = \{X_i\}_{i=1}^N$, a structural causal model (SCM) is defined by $X_i = f_i(p_{ai}, \epsilon_i)$ with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the parents $p_{ai} \subset X \setminus X_i$ chosen such that the corresponding graph contains no cycles.

Example:

SCM

$X_1 = \epsilon_1$

$X_2 = \epsilon_2$

$X_3 = X_1 \cdot X_2 + \epsilon_3$

DAG

$X_1 \rightarrow X_2 \rightarrow X_3$

Data

$\epsilon \sim p(\epsilon)$

$x_i = f_i(p_{ai}, \epsilon_i)$

$x \sim p(x)$

Def.: $X_i$ is a cause of $X_j$, iff there exist values of $X_i$ and $X_j$ such that $p(x_j | do\{x_i\}) \neq p(x_j)$. 

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a structural causal model (SCM) is defined by

$$X_i = f_i(\text{pa}_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the parents $\text{pa}_i \subset \mathcal{X} \setminus X_i$ chosen such that the corresponding graph contains no cycles.

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^{N}$ a structural causal model (SCM) is defined by

$$X_i = f_i(pa_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^{N}$ exogenous noise terms and the parents $pa_i \subset \mathcal{X}\backslash X_i$ chosen such that the corresponding graph contains no cycles.

Example:

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(p_a_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the *parents* $p_a_i \subset \mathcal{X}\setminus X_i$ chosen such that the corresponding graph contains no cycles.

Example:

$$\begin{align*}
SCM \\
X_1 &= \epsilon_1 \\
X_2 &= \epsilon_2 \\
X_3 &= X_1 \cdot X_2 + \epsilon_3
\end{align*}$$

Structural Causal Models (SCMs)

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a *structural causal model* (SCM) is defined by

$$X_i = f_i(\text{pa}_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the *parents* $\text{pa}_i \subset \mathcal{X}\backslash X_i$ chosen such that the corresponding graph contains no cycles.

Example:

**SCM**

$$
\begin{align*}
X_1 &= \epsilon_1 \\
X_2 &= \epsilon_2 \\
X_3 &= X_1 \cdot X_2 + \epsilon_3
\end{align*}
$$

**DAG**

$$
\begin{align*}
X_1 &\rightarrow X_2 \\
&\downarrow \\
&\downarrow \\
&X_3
\end{align*}
$$

For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a \textit{structural causal model} (SCM) is defined by

$$X_i = f_i(\text{pa}_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the \textit{parents} $\text{pa}_i \subset \mathcal{X}\setminus X_i$ chosen such that the corresponding graph contains no cycles.

Example:

SCM

\begin{align*}
X_1 &= \epsilon_1 \\
X_2 &= \epsilon_2 \\
X_3 &= X_1 \cdot X_2 + \epsilon_3
\end{align*}

DAG

\begin{align*}
X_1 &\rightarrow X_2 & \rightarrow X_3
\end{align*}

Data

\begin{align*}
\epsilon &\sim p(\epsilon) \\
x_i &= f_i(\text{pa}_i, \epsilon_i) \\
x &\sim p(x)
\end{align*}
For a given set of variables $\mathcal{X} = \{X_i\}_{i=1}^N$ a structural causal model (SCM) is defined by

$$X_i = f_i(pa_i, \epsilon_i)$$

with $\{\epsilon_i\}_{i=1}^N$ exogenous noise terms and the parents $pa_i \subset \mathcal{X}\setminus X_i$ chosen such that the corresponding graph contains no cycles.

Example:

<table>
<thead>
<tr>
<th>SCM</th>
<th>DAG</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 = \epsilon_1$</td>
<td>$X_1$, $X_2$</td>
<td>$\epsilon \sim p(\epsilon)$</td>
</tr>
<tr>
<td>$X_2 = \epsilon_2$</td>
<td>$\downarrow$</td>
<td>$x_i = f_i(pa_i, \epsilon_i)$</td>
</tr>
<tr>
<td>$X_3 = X_1 \cdot X_2 + \epsilon_3$</td>
<td>$\downarrow$</td>
<td>$x \sim p(x)$</td>
</tr>
</tbody>
</table>

Def.: $X_i$ is a cause of $X_j$, iff there exist values of $X_i$ and $X_j$ such that $p(x_j|\text{do}\{x_i\}) \neq p(x_j)$.

Causal reasoning with SCMs

Causal factorization:

\[ P(X) = P(X_2 | X_1, X_3, X_4) P(X_1 | X_4) P(X_3 | X_4) P(X_4) \]

Interventions are represented by the do-operator, e.g.,

\[ P(X_2 | \text{do}(X_1 = x_1)) = P(X_2 | X_1 = x_1, X_3, X_4) P(X_3 | X_4) P(X_4) \]

(J. Pearl, *Causality*. 2000.)
Causal reasoning with SCMs

Causal factorization:

\[ P(\mathbf{X}) = P(X_2|X_1, X_3, X_4)P(X_1|X_4)P(X_3|X_4)P(X_4) \]

(J. Pearl, *Causality*. 2000.)
Causal reasoning with SCMs

Causal factorization:

\[ P(\mathbf{X}) = P(X_2|X_1, X_3, X_4)P(X_1|X_4)P(X_3|X_4)P(X_4) \]

Interventions are represented by the do-operator, e.g.,

\[ P(\mathbf{X}|\text{do}(X_1 = x_1)) = P(X_2|X_1 = x_1, X_3, X_4)P(X_3|X_4)P(X_4). \]

(J. Pearl, *Causality*. 2000.)
Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

- **Causal Markov Condition (CMC):** Every node in $X$ is conditionally independent of its nondescendents given its parents.

- **Faithfulness:** There are no further independence relations among the nodes in $X$ beyond those implied by d-separation.

- **d-separation:** Let $A, B, C$ non-intersecting subsets of $X$. $A$ and $B$ are d-separated given $C$ iff for all nodes on the path where the arrows meet head-to-tail ($\rightarrow \rightarrow$) or tail-to-tail ($\leftarrow \rightarrow$) the node is in $C$, for all nodes where the arrows meet head-to-head ($\rightarrow \leftarrow$) neither the node or any of its descendants are in $C$.

Assuming the CMC and faithfulness, $dSep(A, B | C) \iff A \perp \perp B | C$. 
Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in \( \mathcal{X} \) is conditionally independent of its nondescendents given its parents.
Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in $X'$ is conditionally independent of its nondescendents given its parents.

Faithfulness: There are no further independence relations among the nodes in $X'$ beyond those implied by d-separation.
Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in $\mathcal{X}$ is conditionally independent of its nondescendents given its parents.

Faithfulness: There are no further independence relations among the nodes in $\mathcal{X}$ beyond those implied by d-separation.

d-separation: Let $A, B, C$ non-intersecting subsets of $\mathcal{X}$. $A$ and $B$ are d-separated given $C$ iff

- for all nodes on the path where the arrows meet head-to-tail ($\rightarrow . \rightarrow$) or tail-to-tail ($\leftarrow . \rightarrow$) the node is in $C$,

- for all nodes where the arrows meet head-to-head ($\rightarrow . \leftarrow$) neither the node or any of its descendants are in $C$. 
Causal inference

How can we infer the structure of the DAG from the data it generates? We need concepts and assumptions that link the structural with the observational world:

Causal Markov Condition (CMC): Every node in $X$ is conditionally independent of its nondescendants given its parents.

Faithfulness: There are no further independence relations among the nodes in $X$ beyond those implied by d-separation.

d-separation: Let $A, B, C$ non-intersecting subsets of $X$. $A$ and $B$ are d-separated given $C$ iff

- for all nodes on the path where the arrows meet head-to-tail ($\rightarrow . \rightarrow$) or tail-to-tail ($\leftarrow . \rightarrow$) the node is in $C$,
- for all nodes where the arrows meet head-to-head ($\rightarrow . \leftarrow$) neither the node or any of its descendants are in $C$.

Assuming the CMC and faithfulness, $dSep(A, B|C) \iff A \indep B|C$. 
Example

The chain $X_1 \rightarrow X_2 \rightarrow X_3$.

$X_1 \not\perp \perp X_3$.

The fork $X_1 \leftarrow X_2 \rightarrow X_3$.

$X_1 \not\perp \perp X_3$.

The collider $X_1 \rightarrow X_2 \leftarrow X_3$.

$X_1 \not\perp \perp X_3$. 

$X_2$. 


Example

The chain

\[ X_1 \rightarrow X_2 \rightarrow X_3 \]
Example

The chain

\[ X_1 \rightarrow X_2 \rightarrow X_3 \]

The fork

\[ X_1 \leftarrow X_2 \rightarrow X_3 \]
Example

The chain
\[ X_1 \rightarrow X_2 \rightarrow X_3 \]

The fork
\[ X_1 \leftarrow X_2 \rightarrow X_3 \]

The collider
\[ X_1 \rightarrow X_2 \leftarrow X_3 \]
Example

The chain

$X_1 \rightarrow X_2 \rightarrow X_3$

$X_1 \not\perp\!\!\!\!\!\perp X_3$

The fork

$X_1 \leftarrow X_2 \rightarrow X_3$

The collider

$X_1 \rightarrow X_2 \leftarrow X_3$
Example

The chain

\[ X_1 \rightarrow X_2 \rightarrow X_3 \]
\[ X_1 \not\perp \perp X_3 \]
\[ X_1 \perp \perp X_3 | X_2 \]

The fork

\[ X_1 \leftarrow X_2 \rightarrow X_3 \]

The collider

\[ X_1 \rightarrow X_2 \leftarrow X_3 \]
Example

The chain
\[ X_1 \rightarrow X_2 \rightarrow X_3 \]
\[ X_1 \not\perp X_3 \]
\[ X_1 \perp X_3 \mid X_2 \]

The fork
\[ X_1 \leftrightarrow X_2 \rightarrow X_3 \]
\[ X_1 \not\perp X_3 \]

The collider
\[ X_1 \rightarrow X_2 \leftrightarrow X_3 \]
Example

The chain
\[ X_1 \rightarrow X_2 \rightarrow X_3 \]
\[ X_1 \not\perp X_3 \]
\[ X_1 \perp X_3 \mid X_2 \]

The fork
\[ X_1 \leftarrow X_2 \rightarrow X_3 \]
\[ X_1 \not\perp X_3 \]
\[ X_1 \perp X_3 \mid X_2 \]

The collider
\[ X_1 \rightarrow X_2 \leftarrow X_3 \]
Example

The chain
\[ X_1 \rightarrow X_2 \rightarrow X_3 \]
\[ X_1 \perp X_3 \]
\[ X_1 \perp X_3 \mid X_2 \]

The fork
\[ X_1 \leftarrow X_2 \rightarrow X_3 \]
\[ X_1 \perp X_3 \]
\[ X_1 \perp X_3 \mid X_2 \]

The collider
\[ X_1 \rightarrow X_2 \leftarrow X_3 \]
\[ X_1 \perp X_3 \]
\[ X_1 \perp X_3 \mid X_2 \]
Example

The chain
\[ X_1 \rightarrow X_2 \rightarrow X_3 \]
\[ X_1 \not\perp \! \! \! \perp X_3 \]
\[ X_1 \perp \! \! \! \! \perp X_3 | X_2 \]

The fork
\[ X_1 \leftarrow X_2 \rightarrow X_3 \]
\[ X_1 \not\perp \! \! \! \perp X_3 \]
\[ X_1 \perp \! \! \! \! \perp X_3 | X_2 \]

The collider
\[ X_1 \rightarrow X_2 \leftarrow X_3 \]
\[ X_1 \not\perp \! \! \! \perp X_3 \]
\[ X_1 \perp \! \! \! \! \perp X_3 | X_2 \]
xAI/IML is a Causal Problem
xAI/IML is a Causal Problem

\[ X_1 \]

\[ X_2 \]

\[ X_3 \]

\[ X_4 \]

\[ Y \]
xAI/IML is a Causal Problem

- $X_1 \rightarrow \hat{X}_1$
- $X_2 \rightarrow \hat{X}_2$
- $X_3 \rightarrow \hat{X}_3$
- $X_4 \rightarrow \hat{X}_4$
- $Y$
xAI/IML is a Causal Problem

\[ X_1 \rightarrow \hat{X}_1 \rightarrow h^*(\cdot) \rightarrow \hat{Y} \]

\[ X_2 \rightarrow \hat{X}_2 \rightarrow h^*(\cdot) \rightarrow \hat{Y} \]

\[ X_3 \rightarrow \hat{X}_3 \rightarrow h^*(\cdot) \rightarrow \hat{Y} \]

\[ X_4 \rightarrow \hat{X}_4 \rightarrow h^*(\cdot) \rightarrow \hat{Y} \]

\[ Y \]

\[ \hat{Y} \neq P(Y|\text{do}\{X_1 = x_1 + \delta x_1\}) \]

\[ \hat{Y} \neq P(Y|\text{do}\{X_3 = x_3\}) \]
xAI/IML is a Causal Problem

\[ X_1 \xrightarrow{\text{World}} \hat{X}_1 \xrightarrow{h^*(.)} \hat{Y} \]

\[ X_2 \xrightarrow{\hat{X}_2} h^*(.) \]

\[ X_3 \xrightarrow{\hat{X}_3} h^*(.) \]

\[ X_4 \xrightarrow{\hat{X}_4} h^*(.) \]

\[ Y \xrightarrow{\hat{Y}} \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_1\} \neq P(Y \mid \text{do}\{X_1\}) \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_3\} \neq P(Y \mid \text{do}\{X_3\}) \]

\[ X_1 \perp \perp X_2 \mid X_3 \]

\[ \hat{Y} \perp \perp \hat{X}_2 \text{ (if } h^* \text{ is optimal) but } \hat{Y} \not\perp \perp \hat{X}_2 \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_1\} \neq P(Y \mid \text{do}\{X_1\}) \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_3\} \neq P(Y \mid \text{do}\{X_3\}) \]

\[ X_1 \xrightarrow{X_2} \hat{X}_1 \xrightarrow{h^*(.)} \hat{Y} \]

\[ X_2 \xrightarrow{\hat{X}_2} h^*(.) \]

\[ X_3 \xrightarrow{\hat{X}_3} h^*(.) \]

\[ X_4 \xrightarrow{\hat{X}_4} h^*(.) \]

\[ Y \xrightarrow{\hat{Y}} \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_1\} \neq P(Y \mid \text{do}\{X_1\}) \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_3\} \neq P(Y \mid \text{do}\{X_3\}) \]

\[ X_1 \perp \perp X_2 \mid X_3 \]

\[ \hat{Y} \perp \perp \hat{X}_2 \text{ (if } h^* \text{ is optimal) but } \hat{Y} \not\perp \perp \hat{X}_2 \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_1\} \neq P(Y \mid \text{do}\{X_1\}) \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_3\} \neq P(Y \mid \text{do}\{X_3\}) \]
xAI/IML is a Causal Problem

World

\( X_1 \rightarrow \hat{X}_1 \)

\( X_2 \rightarrow \hat{X}_2 \)

\( X_3 \rightarrow \hat{X}_3 \)

\( X_4 \rightarrow \hat{X}_4 \)

\( Y \)

Model

\( h^*(.) \rightarrow \hat{Y} \)

\( P(\hat{Y} | \text{do}\{\hat{X}_1 = \hat{x}_1 + \delta x_1\}) \neq P(Y | \text{do}\{X_1 = x_1 + \delta x_1\}) \)

\( P(\hat{Y} | \text{do}\{\hat{X}_3 = \hat{x}_3\}) \neq P(Y | \text{do}\{X_3 = x_3\}) \)
xAI/IML is a Causal Problem

\[ X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5 \rightarrow Y \]

\[ \hat{X}_1 \rightarrow \hat{X}_2 \rightarrow \hat{X}_3 \rightarrow \hat{X}_4 \rightarrow h^*(.) \rightarrow \hat{Y} \]

World

Model

\( P(\hat{Y} \mid \text{do}\{\hat{X}_3\}) \neq P(Y \mid \text{do}\{X_3\}) \)
xAI/IML is a Causal Problem

\[ Y \perp X_2 | X_3 \Rightarrow \hat{Y} \perp \hat{X}_2 \text{ (if } h \text{ is optimal) but } \hat{Y} \not\perp X_2 \]
xAI/IML is a Causal Problem

\[ \hat{Y} \perp \perp \hat{X}_2 \text{ (if } h \text{ is optimal) but } \hat{Y} \not\perp \perp X_2 \]

\[ P(\hat{Y}|\text{do}\{\hat{X}_1 = \hat{x}_1 + \delta_{x_1}\}) \neq P(Y|\text{do}\{X_1 = x_1 + \delta_{x_1}\}) \]
xAI/IML is a Causal Problem

\[ X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5 \rightarrow Y \]

\[ \hat{X}_1 \rightarrow \hat{X}_2 \rightarrow \hat{X}_3 \rightarrow \hat{X}_4 \rightarrow h^*(.) \rightarrow \hat{Y} \]

\[ Y \perp X_2 \mid X_3 \Rightarrow \hat{Y} \perp \hat{X}_2 \text{ (if } h \text{ is optimal) but } \hat{Y} \not\perp X_2 \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_1 = \hat{x}_1 + \delta x_1\}) \neq P(Y \mid \text{do}\{X_1 = x_1 + \delta x_1\}) \]

\[ P(\hat{Y} \mid \text{do}\{\hat{X}_3 = \hat{x}_3\}) \neq P(Y \mid \text{do}\{X_3 = x_3\}) \]
A Taxonomy of xAI/IML

What object is explained?

- The prediction $\hat{Y}$?
- The target variable $Y$?
- Their relationship $R$ (as measured by the loss)?

On what level is the object explained?

- Associations $X = x$?
- Model-level interventions $do(\hat{X} = x)$?
- World-level interventions $do(X = x)$?
What object is explained?
What object is explained?
- The prediction \( \hat{Y} \)?
What object is explained?

- The prediction $\hat{Y}$?
- The target variable $Y$?
What object is explained?

- The prediction \( \hat{Y} \) ?
- The target variable \( Y \) ?
- Their relationship \( R \) (as measured by the loss) ?
What object is explained?

- The prediction $\hat{Y}$?
- The target variable $Y$?
- Their relationship $R$ (as measured by the loss)?

On what level is the object explained?
A Taxonomy of xAI/IML

What object is explained?
- The prediction $\hat{Y}$?
- The target variable $Y$?
- Their relationship $R$ (as measured by the loss)?

On what level is the object explained?
- Associations $X = x$?
What object is explained?
- The prediction $\hat{Y}$?
- The target variable $Y$?
- Their relationship $R$ (as measured by the loss)?

On what level is the object explained?
- Associations $X = x$?
- Model-level interventions $\text{do}(\hat{X} = x)$?
A Taxonomy of xAI/IML

What object is explained?
- The prediction $\hat{Y}$?
- The target variable $Y$?
- Their relationship $R$ (as measured by the loss)?

On what level is the object explained?
- Associations $X = x$?
- Model-level interventions $\text{do} (\hat{X} = x)$?
- World-level interventions $\text{do} (X = x)$?
### Nine Perspectives on Model and Data

<table>
<thead>
<tr>
<th>$\hat{Y}$</th>
<th>$R$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$do(X = x')$</td>
<td>Does the model's mechanism rely on gender?</td>
<td></td>
</tr>
<tr>
<td>$X = x'$</td>
<td></td>
<td>Which biomarkers are correlated with the disease?</td>
</tr>
<tr>
<td>$do(X = x')$</td>
<td>What can I do to get accepted?</td>
<td>Does the diagnosis model also work in a different hospital?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What can we do to treat the disease?</td>
</tr>
</tbody>
</table>
Contribution I: Improvement-Focused Causal Recourse (ICR)

Recourse Setting
Recourse Setting

• suppose a ML model rejects your request (job application, loan application, hospital admission, ....)

• recourse recommendations tell you what to do to get accepted
Recourse Setting

• suppose a ML model rejects your request (job application, loan application, hospital admission, ….)

• recourse recommendations tell you what to do to get accepted

• existing methods: counterfactual explanations (CE) [Wachter et al.], causal recourse (CR) [Karimi et al.]
Problem

Terminology
Problem

• existing methods (CE, CR) may suggest to game the predictor

Terminology

gaming: tricking the predictor into falsely believing that one is qualified

\[ \hat{y} = 1 \text{ but } y = 0 \]
Problem

• existing methods (CE, CR) may suggest to game the predictor

• recourse should guide towards both acceptance and improvement

Terminology

**gaming**: tricking the predictor into falsely believing that one is qualified

\[ \hat{y} = 1 \text{ but } y = 0 \]

**acceptance**: reverting the model’s decision

\[ \hat{y} = 1 \]

**improvement**: reverting the underlying real-world state

\[ y = 1 \]
**Illustrative Example**

**Goal:** predict CoViD risk to decide whether you are allowed to enter a hospital (without testing for CoViD)

https://riskcalc.org/COVID19/
Illustrative Example

Counterfactual Explanations (CE)
[Wachter et al.]

Causal Recourse (CR)
[Karimi et al.]

Improvement-Focused CR (ICR)
[Koenig et al.]

viewpoint

CoViD

sympt.

vacc.

\( \hat{Y} \)

What is the minimal \( \text{do}(X_I = x') \) such that \( \hat{y} = 1 \)?

idea

exemplary explanation

“If you plug lower symptom values into the model, the prediction is favorable.”

What is the most cost-efficient such that?
\( \text{do}(a) \)

\( \hat{y} = 1 \)

\( y = 1 \)
Illustrative Example

Counterfactual Explanations (CE) [Wachter et al.]

Causal Recourse (CR) [Karimi et al.]

Improvement-Focused CR (ICR) [Koenig et al.]

viewpoint

CoViD

vacc.

sympt.

\( \hat{Y} \)

What is the minimal \( \text{do}(X_f = x') \) such that \( \hat{y} = 1 \)?

idea

exemplary explanation

“If you plug lower symptom values into the model, the prediction is favorable.”

acceptance \( (\hat{y} = 1) \)

not guaranteed
Illustrative Example

Counterfactual Explanations (CE)  
[Wachter et al.]

- **viewpoint**
  - vacc.
  - CoViD
  - sympt.
  - \( \hat{Y} \)

Causal Recourse (CR)  
[Karimi et al.]

- **viewpoint**
  - vacc.
  - CoViD
  - sympt.
  - \( \hat{Y} \)

<table>
<thead>
<tr>
<th>idea</th>
<th>What is the minimal ( do(X_I = x') ) such that ( \hat{y} = 1 )?</th>
<th>What is the most cost-efficient ( do(X_I = x') ) such that ( \hat{y} = 1 )?</th>
</tr>
</thead>
<tbody>
<tr>
<td>exemplary explanation</td>
<td>“If you plug lower symptom values into the model, the prediction is favorable.”</td>
<td>“If you treat your symptoms (take cough syrup), the model will accept you.”</td>
</tr>
</tbody>
</table>

| acceptance (\( \hat{y} = 1 \)) | not guaranteed | yes |
Illustrative Example

Counterfactual Explanations (CE) [Wachter et al.]

Causal Recourse (CR) [Karimi et al.]

**viewpoint**
- vacc.
- symp.

**idea**
- What is the minimal \(do(X_I = x')\) such that \(\hat{y} = 1?\)

**exemplary explanation**
- “If you plug lower symptom values into the model, the prediction is favorable.”
- “If you treat your symptoms (take cough syrup), the model will accept you.”

**acceptance** (\(\hat{y} = 1\))
- not guaranteed
- yes

**improvement** (\(y = 1\))
- not guaranteed
- not guaranteed
### Illustrative Example

<table>
<thead>
<tr>
<th>Counterfactual Explanations (CE) [Wachter et al.]</th>
<th>Causal Recourse (CR) [Karimi et al.]</th>
<th>Improvement-Focused CR (ICR) [Koenig et al.]</th>
</tr>
</thead>
<tbody>
<tr>
<td>viewpoint</td>
<td>idea</td>
<td></td>
</tr>
<tr>
<td>CoViD $Y$</td>
<td>What is the minimal $do(X_I = x')$ such that $\hat{y} = 1$?</td>
<td>What is the most cost-efficient $do(X_I = x')$ such that $\hat{y} = 1$?</td>
</tr>
<tr>
<td>sympt. $\hat{Y}$</td>
<td></td>
<td>What is the most cost-efficient $do(X_I = x')$ such that $y = 1$?</td>
</tr>
<tr>
<td>vacc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>acceptance $(\hat{y} = 1)$</td>
<td>not guaranteed</td>
<td>yes</td>
</tr>
<tr>
<td>improvement $(y = 1)$</td>
<td>not guaranteed</td>
<td>not guaranteed</td>
</tr>
<tr>
<td>example explanation</td>
<td>“If you plug lower symptom values into the model, the prediction is favorable.”</td>
<td>“If you treat your symptoms (take cough syrup), the model will accept you.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“If you get vaccinated, you will decrease your CoViD risk and thus be accepted.”</td>
</tr>
<tr>
<td>acceptance $(\hat{y} = 1)$</td>
<td>not guaranteed</td>
<td>yes</td>
</tr>
<tr>
<td>improvement $(y = 1)$</td>
<td>not guaranteed</td>
<td>not guaranteed</td>
</tr>
</tbody>
</table>

---

**What is the most cost-efficient such that?**

$do(X_I = x')$  
\[ \hat{y} = 1 \]

---

**What is the minimal such that?**

$do(X_I = x')$  
\[ y = 1 \]
The Methods & The 9 Perspectives

<table>
<thead>
<tr>
<th>$\hat{Y}$</th>
<th>$R$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$do(X = x')$</td>
<td>CEs</td>
<td></td>
</tr>
<tr>
<td>$X = x'$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$do(X = x')$</td>
<td>CR</td>
<td>ICR</td>
</tr>
</tbody>
</table>
ICR: Optimization Problem

For a target improvement probability $\bar{\gamma}$, some cost function $c$ and pre-recourse observation $x^{pre}$, the ICR action $a := do(X_{Ia} := \theta_{Ia})$ is found by optimizing

$$\argmin_a c(a; x^{pre}) \text{ s.t. } \gamma(a; x^{pre}) \geq \bar{\gamma};$$

with $\gamma(a, x^{pre})$ being the improvement probability for action $a$ and an individual with characteristics $x^{pre}$. 
How to Define Improvement Confidence $\gamma$?

**Goal:** For an action $do(a)$, estimate probability of improvement while taking as many features as possible into account.

$$\gamma_{sub} := P(Y_{post} = 1| x_{pre I_a}, do(a))$$
How to Define Improvement Confidence $\gamma$?

**Goal:** For an action $do(a)$, estimate probability of improvement while taking as many features as possible into account.

\[ \gamma \]

*individualized*  
(structural causal model known):

---

*estimation based on* structural counterfactuals

---

*precision* takes all features into account

---

*definition*  
\[ \gamma^{ind} := P(Y^{post} = 1 | x^{pre}, do(a)) \]
How to Define Improvement Confidence $\gamma$?

**Goal:** For an action $do(a)$, estimate probability of improvement while taking as many features as possible into account.

<table>
<thead>
<tr>
<th></th>
<th>individualized</th>
<th>subpopulation-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(structural causal model known):</td>
<td>(only causal graph known):</td>
</tr>
<tr>
<td><strong>estimation based on</strong></td>
<td>structural counterfactuals</td>
<td>conditional average treatment effect</td>
</tr>
<tr>
<td><strong>precision</strong></td>
<td>takes all features into account</td>
<td>only takes features that are not affected by the action into account</td>
</tr>
<tr>
<td><strong>definition</strong></td>
<td>$\gamma^{ind} := P(Y^{post} = 1 \mid x^{pre}, do(a))$</td>
<td>$\gamma^{sub} := P(Y^{post} = 1 \mid x^{pre}_{G_a}, do(a))$</td>
</tr>
</tbody>
</table>
How to Achieve Acceptance ($\hat{y} = 1$)?
How to Achieve Acceptance ($\hat{y} = 1$)?

**Intuition:** Classifiers remain accurate under ICR actions → acceptance ensures improvement.
How to Achieve Acceptance ($\hat{y} = 1$)?

**Intuition:** Classifiers remain accurate under ICR actions → acceptance ensures improvement.

Why?
How to Achieve Acceptance ($\hat{y} = 1$)?

**Intuition:** Classifiers remain accurate under ICR actions → acceptance ensures improvement.

Why?

- ICR only recommends interventions on causes
How to Achieve Acceptance ($\hat{y} = 1$)?

**Intuition:** Classifiers remain accurate under ICR actions → acceptance ensures improvement.

Why?

- ICR only recommends interventions on causes
- intervening on causes does not affect $P(Y \mid X)$
How to Achieve Acceptance ($\hat{\gamma} = 1$)?

**Intuition:** Classifiers remain accurate under ICR actions $\rightarrow$ acceptance ensures improvement.

Why?

- ICR only recommends interventions on causes
- intervening on causes does not affect $P(Y \mid X)$

$\rightarrow$ the classifier is *stable* w.r.t. ICR actions
How to Achieve Acceptance ($\hat{y} = 1$)?

**Intuition:** Classifiers remain accurate under ICR actions $\rightarrow$ acceptance ensures improvement.

Why?

- ICR only recommends interventions on causes
- Intervening on causes does not affect $P(Y \mid X)$

$\rightarrow$ the classifier is *stable* w.r.t. ICR actions
Interpretable Machine Learning
A Guide for Making Black Box Models Explainable

@ChristophMolnar

Summary: A Causal Perspective on IML

Pitfalls to Avoid when Interpreting Machine Learning Models

Gunnar König

Alex Markham

Abstract
Modern implementations of machine learning (ML) models include high-dimensional performance metrics and visualization techniques that provide model interpretability, but can have strong weaknesses if applied incorrectly. We illustrate pitfalls of ML model interpretation such as bad model generalization, dependent factors, feature interactions or simplified causal in- terpretations. Our paper addresses ML practitio- ners by warning awareness of pitfalls and pointing out solutions for correct model interpretation, as well as ML researchers by discussing open issues for further research.

1. Introduction
Traditionally, researchers have used parametric models, e.g., linear models, to conduct inference. However, a noticeable shift has happened over the last years towards more non-parametric and non-linear ML models. Models such as random forests, boosting or neural networks often require interpretable models or many prediction tasks, as most ML models handle feature interactions and non-linear effects automatically (Fernández-Delgado et al., 2014). Many disciplines benefit from the predictive performance of ML models and abound scientific questions using ML interpretation techniques. Examples of such efforts include understanding why ML models make certain predictions (Molnar et al., 2020), ranging concept covers in various domains (Ortiz et al., 2019), and investigating the fairness of the trained models (Krieger et al., 2020), understanding wildlife disease (Poulin-Brodeur et al., 2019), improving crop yield (Chabalovski et al., 2020; Zhang et al., 2019), inferring behavior from smartphones (Gulliford et al., 2019), and analyzing 4G for network behavior (Pooch Quintero et al., 2022). Practitioners are usually interested in the global effects that features have on the outcomes and their importance for correct predictions. For certain model choices, e.g., linear models or decision trees, feature importance or impact scores can be inferred from the learned parameters and model structure. In contrast, complex non-linear models that, e.g., do not have interpretable parameters, make it more difficult to extract such knowledge. Therefore, interpretation methods are necessary to understand the relationship between features and outcomes. Prominent techniques for global feature effects include the partial dependence plots (PDP) (Friedman et al., 1999), averaged local effects (ALE) (Apley & Zhu, 2016) and individual conditional expectation (ICE) (Goldstein et al., 2015). A common feature importance technique is the permutation feature importance (RF) (Breiman, 2001; Fisher et al., 2018; Cano-Belén et al., 2019). This paper will mostly focus on pitfalls of global interpretation techniques when the full functional relationship underlying the data is to be analyzed. Out of scope for this discussion are “local” interpretation methods such as (LIME) (סלקאן et al., 2016) or con- textual explanations (Wachter et al., 2017, Dandl et al., 2020), where individual predictions are to be explained – usually to enable decisions to individuals.

Contribution
We review pitfalls of global model-