Causal modeling, inference, and machine learning

Louis Wehenkel,

Institut Montefiore,
Department of Electrical Engineering and Computer Science,
Université de Liège, Liège, Belgium.

Advanced Machine Learning Course 2020

Overview

- Probabilistic dependence versus cause-effect relations
- A quick reminder about Bayesian networks
- Functional causal models
- Learnability of cause-effect relations
- Exploiting cause-effect relations in machine learning

Probabilistic dependence versus cause-effect relations

Probabilistic independence/dependence ($\perp, \not\perp$)

- Random variables $X, Y, Z, \ldots$ (continuous or discrete), and their values $x, y, z\ldots$
- Densities, probability mass functions, distributions: $P(x, y), P(x|z)\ldots.$
- (Conditional) (in)dependence of random variables: $X \perp Y$, $X \not\perp Y$, $X \perp Y|Z\ldots$
- Meaning of $X \not\perp Y$: $P(y|x) \neq P(y)$ (and also $P(x|y) \neq P(x)$).
- E.g.: $Wet \not\perp Rain$ (and also $Rain \not\perp Wet$)

Cause-effect relations ($\hookrightarrow, \hookleftarrow, \not\hookrightarrow, \not\hookleftarrow$) and interventions (aka $do[X = x]$)

- $X \hookrightarrow Y$, $X \hookrightarrow Y \hookrightarrow Z$, $X \hookrightarrow Y \hookleftarrow Z$ ...
- E.g.: $Rain \hookrightarrow Wet$, or $Sprinkler \hookrightarrow Wet \hookleftarrow Rain$, and $Wet \hookrightarrow Slipery\ldots$
- Meaning of $X \hookrightarrow Y$: $P(y|do[X = x]) \neq P(y)$.
- E.g.: $Rain \hookrightarrow Wet$ (but $Wet \not\hookrightarrow Rain$).

Probabilistic dependence versus cause-effect relations

Probabilistic independence/dependence ($\perp, \not\perp$)

- Random variables $X, Y, Z, \ldots$ (continuous or discrete), and their values $x, y, z\ldots$
- Densities, probability mass functions, distributions: $P(x, y), P(x|z)\ldots.$
- (Conditional) (in)dependence of random variables: $X \perp Y$, $X \not\perp Y$, $X \perp Y|Z\ldots$
- Meaning of $X \not\perp Y$: $P(y|x) \neq P(y)$ (and also $P(x|y) \neq P(x)$).
- E.g.: $Wet \not\perp Rain$ (and also $Rain \not\perp Wet$)

Cause-effect relations ($\hookrightarrow, \hookleftarrow, \not\hookrightarrow, \not\hookleftarrow$) and interventions (aka $do[X = x]$)

- $X \hookrightarrow Y$, $X \hookrightarrow Y \hookrightarrow Z$, $X \hookrightarrow Y \hookleftarrow Z$ ...
- E.g.: $Rain \hookrightarrow Wet$, or $Sprinkler \hookrightarrow Wet \hookleftarrow Rain$, and $Wet \hookrightarrow Slipery\ldots$
- Meaning of $X \hookrightarrow Y$: $P(y|do[X = x]) \neq P(y)$.
- E.g.: $Rain \hookrightarrow Wet$ (but $Wet \not\hookrightarrow Rain$).

Reichenbach's common-cause principle [1]

(it relates probabilistic dependence and causal relationships among variables)

This principle states (in the form we use it in this lecture) that

$$\text{If } X \not\perp Y \text{ then } \exists Z : [X \hookleftarrow Z \hookrightarrow Y] \wedge [X \perp Y | Z].$$

For example, we can pretend that $X \perp Y$, as soon as we exclude that X and Y could have a common cause!

NB: principle includes the cases where

- either $X \equiv Z$: then we have $X \not\perp Y$ and $X \hookrightarrow Y$
- or $Y \equiv Z$: then we have $X \not\perp Y$ and $Y \hookrightarrow X$
- or both $X \equiv Z$ and $Y \equiv Z$: then we have trivially $X \equiv Y$

Reichenbach's common-cause principle [1]

(it relates probabilistic dependence and causal relationships among variables)

This principle states (in the form we use it in this lecture) that

$$\text{If } X \not\perp Y \text{ then } \exists Z : [X \hookleftarrow Z \hookrightarrow Y] \wedge [X \perp Y | Z].$$

For example, we can pretend that $X \perp Y$, as soon as we exclude that X and Y could have a common cause!

NB: principle includes the cases where

- either $X \equiv Z$: then we have $X \not\perp Y$ and $X \hookrightarrow Y$
- or $Y \equiv Z$: then we have $X \not\perp Y$ and $Y \hookrightarrow X$
- or both $X \equiv Z$ and $Y \equiv Z$: then we have trivially $X \equiv Y$

Reminder: Bayesian networks viewed as probabilistic models

- A Bayesian network is a DAG (directed acyclic graph) representing a set of distributions that are compatible with the factorization given by the graph.
- E.g. the DAG $X \to Z \leftarrow Y$, encodes the set of distributions $P(x, y, z)$ that satisfy

$$P(x, y, z) = P(x)P(y)P(z|x, y), \forall x, y, z$$

- The $d-$separation criterion allows to graphically infer all conditional independence statements that are satisfied by all compatible distributions.
- E.g. the DAG $X \to Z \leftarrow Y$ represents the single statement $X \perp Y$.
- Theorem: two DAGs are observationally equivalent (i.e. they have the same sets of compatible distributions) iff they have the same skeleton and set of $v-$structures.
- NB: In order to stress the difference of the type of assumptions expressed by such graphs with those of causal models, we use the symbol $\to$ instead of $\hookrightarrow$.

Reminder: Bayesian networks viewed as probabilistic models

- A Bayesian network is a DAG (directed acyclic graph) representing a set of distributions that are compatible with the factorization given by the graph.
- E.g. the DAG $X \rightarrow Z \leftarrow Y$, encodes the set of distributions $P(x, y, z)$ that satisfy

$$P(x, y, z) = P(x)P(y)P(z|x, y), \forall x, y, z$$

- The $d-$separation criterion allows to graphically infer all conditional independence statements that are satisfied by all compatible distributions.
- E.g. the DAG $X \rightarrow Z \leftarrow Y$ represents the single statement $X \perp Y$.
- Theorem: two DAGs are observationally equivalent (i.e. they have the same sets of compatible distributions) iff they have the same skeleton and set of $v-$structures.
- NB: In order to stress the difference of the type of assumptions expressed by such graphs with those of causal models, we use the symbol $\rightarrow$ instead of $\hookrightarrow$.

Reminder: Bayesian networks viewed as probabilistic models

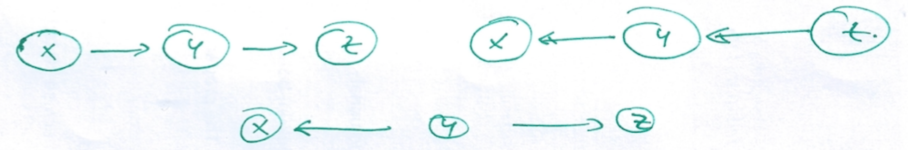
- A Bayesian network is a DAG (directed acyclic graph) representing a set of distributions that are compatible with the factorization given by the graph.
- E.g. the DAG $X \rightarrow Z \leftarrow Y$, encodes the set of distributions $P(x, y, z)$ that satisfy

$$P(x, y, z) = P(x)P(y)P(z|x, y), \forall x, y, z$$

- The $d-$separation criterion allows to graphically infer all conditional independence statements that are satisfied by all compatible distributions.
- E.g. the DAG $X \rightarrow Z \leftarrow Y$ represents the single statement $X \perp Y$.
- Theorem: two DAGs are observationally equivalent (i.e. they have the same sets of compatible distributions) iff they have the same skeleton and set of $v-$structures.
- NB: In order to stress the difference of the type of assumptions expressed by such graphs with those of causal models, we use the symbol $\rightarrow$ instead of $\hookrightarrow$.

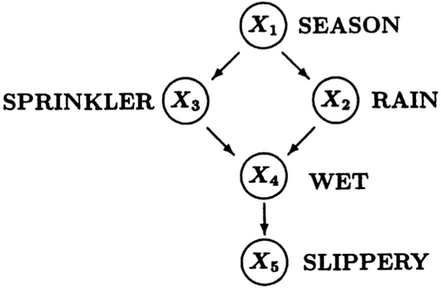
Reminder: comments about Bayesian networks

- Perfectness, stability, DAG-isomorphism
  - See Chapter 1 of [2], and rethink about the double fair coin flipping problem
- Examples of observationally equivalent DAGs



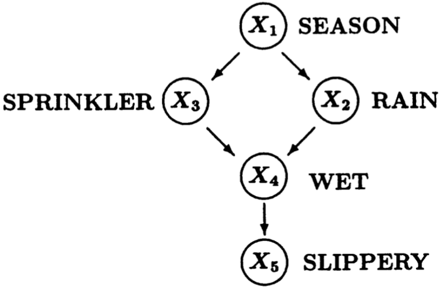
- Algorithmic advantages of PGMs
  - Sparse models (sample complexity, memory requirements)
  - Efficient inference, and learning
- Limits of learning PGMs from observational data only

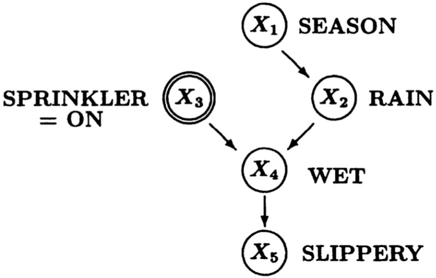
Causal (interpretation of) Bayesian networks



In other words, in addition to pure probabilistic inference (conditioning, a.k.a. guessing in the presence of evidence), we also allow the use of the graphical structure to model the effect of interventions.

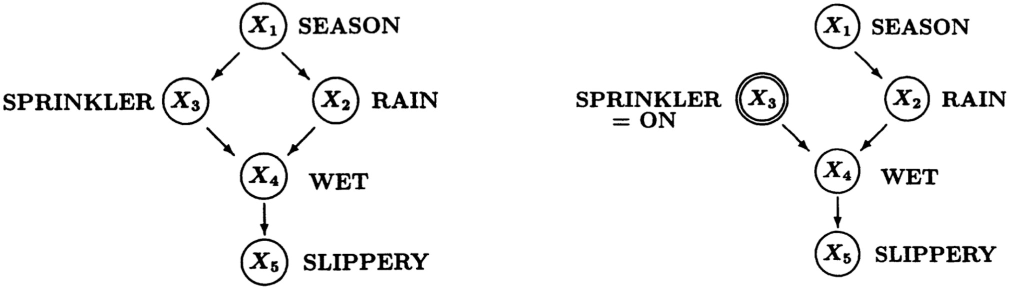
Causal (interpretation of) Bayesian networks





In other words, in addition to pure probabilistic inference (conditioning, a.k.a. guessing in the presence of evidence), we also allow the use of the graphical structure to model the effect of interventions.

Causal (interpretation of) Bayesian networks



In other words, in addition to pure probabilistic inference (conditioning, a.k.a. guessing in the presence of evidence), we also allow the use of the graphical structure to model the effect of interventions.

Functional causal models

(a.k.a. Structural equation models, and Structural causal models)

- A set of vars $X_1, \ldots, X_p$ whose relations we want to model
- For each var X_i a functional assignment equation:

$$x_i := f_i(x_1, \ldots, [x_i], \ldots x_p, u_i)$$

  where $[\cdot]$ means that its argument is not allowed, and where the U_i denote suitably chosen noise variables.
- We consider the subclass of Markovian functional causal models:
  - the directed graph over $X_1, \ldots, X_p$ induced by the assignment equations is acyclic;
  - the noise variables $U_1, \ldots, U_p$ are mutually independent random variables.
  - Given the marginals $P(u_1), \ldots, P(u_p)$, such a model induces a joint distribution $P(x_1, \ldots, x_p, u_1, \ldots, u_p)$, and hence $P(x_1, \ldots, x_p)$ by marginalization.
  - $P(x_1, \ldots, x_p)$ factorises according to the DAG induced by the set of functions f_i.

Functional causal models

(a.k.a. Structural equation models, and Structural causal models)

- A set of vars $X_1, \dots, X_p$ whose relations we want to model
- For each var X_i a functional assignment equation:

$$x_i := f_i(x_1, \dots, [x_i], \dots x_p, u_i)$$

  where $[\cdot]$ means that its argument is not allowed, and where the U_i denote suitably chosen noise variables.
- We consider the subclass of Markovian functional causal models:
  - the directed graph over $X_1, \dots, X_p$ induced by the assignment equations is acyclic;
  - the noise variables $U_1, \dots, U_p$ are mutually independent random variables.
  - Given the marginals $P(u_1), \dots, P(u_p)$, such a model induces a joint distribution $P(x_1, \dots, x_p, u_1, \dots, u_p)$, and hence $P(x_1, \dots, x_p)$ by marginalization.
  - $P(x_1, \dots, x_p)$ factorises according to the DAG induced by the set of functions f_i.

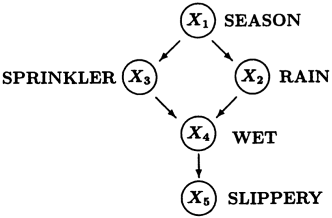
Example of a functional causal models

The sprinkler example (see Chapter 1 of [2]):

$$
\begin{aligned}
x_1 &:= u_1 && \text{SEASON} \in \{\text{spring}, \text{summer}, \text{fall}, \text{winter}\} && (1)\\
x_2 &:= f_2(x_1, u_2) && \text{RAIN} \in \{\text{true}, \text{false}\} && (2)\\
x_3 &:= f_3(x_1, u_3) && \text{SPRINKLER} \in \{\text{on}, \text{off}\} && (3)\\
x_4 &:= f_4(x_2, x_3, u_4) && \text{WET} \in \{\text{true}, \text{false}\} && (4)\\
x_5 &:= f_5(x_4, u_5) && \text{SLIPPERY} \in \{\text{true}, \text{false}\} && (5)
\end{aligned}
$$

where $u_2, \ldots, u_5 \in \{\text{normal}, \text{trigger}, \text{inhibit}\}$.

e.g. $f_5(x_4, u_5) \equiv (x_4 \vee [u_5 = \text{trigger}]) \wedge \neg[u_5 = \text{inhibit}]$



Modeling interventions with Functional causal models

Intervention: means changing the mechanism that determines the value of a variable.

- A simple intervention: if we want to express the fact that we force the sprinkler to be on, this can modelled by replacing (3) by

$$x_3 := \text{on}.$$

- A more sophisticated intervention: randomizing the status of the sprinkler, can be modelled by replacing (3) by

$$x_3 := u_3' \in \{\text{on}, \text{off}\}.$$

- Another intervention: using the sprinkler when it doesn't rains, can be modeled by replacing (3) by

$$x_3 := f_3'(x_2, u_3).$$

Modeling interventions with Functional causal models

Intervention: means changing the mechanism that determines the value of a variable.

- A simple intervention: if we want to express the fact that we force the sprinkler to be on, this can modelled by replacing (3) by

$$x_3 := \text{on}.$$

- A more sophisticated intervention: randomizing the status of the sprinkler, can be modelled by replacing (3) by

$$x_3 := u'_3 \in \{\text{on}, \text{off}\}.$$

- Another intervention: using the sprinkler when it doesn't rains, can be modeled by replacing (3) by

$$x_3 := f'_3(x_2, u_3).$$

Modeling interventions with Functional causal models

Intervention: means changing the mechanism that determines the value of a variable.

- A simple intervention: if we want to express the fact that we force the sprinkler to be on, this can modelled by replacing (3) by

$$x_3 := \text{on}.$$

- A more sophisticated intervention: randomizing the status of the sprinkler, can be modelled by replacing (3) by

$$x_3 := u_3' \in \{\text{on}, \text{off}\}.$$

- Another intervention: using the sprinkler when it doesn't rains, can be modeled by replacing (3) by

$$x_3 := f_3'(x_2, u_3).$$

Modeling interventions with Functional causal models

Intervention: means changing the mechanism that determines the value of a variable.

- A simple intervention: if we want to express the fact that we force the sprinkler to be on, this can modelled by replacing (3) by
$$x_3 := \text{on}.$$
- A more sophisticated intervention: randomizing the status of the sprinkler, can be modelled by replacing (3) by
$$x_3 := u_3' \in \{\text{on}, \text{off}\}.$$
- Another intervention: using the sprinkler when it doesn't rains, can be modeled by replacing (3) by
$$x_3 := f_3'(x_2, u_3).$$

Counterfactual reasoning with Functional causal models

What is counterfactual reasoning: given some 'evidence' we want to see what would have possibly happened if some things had been done differently.

- For example, in the "Wet floor example":
  - we have observed that SLIPPERY = true (we call this the "Evidence")
  - and we want to know $P(\text{SLIPPERY} = \text{true})$ (we call this the "Query")
  - if we had forced SPRINKLER = off (we call this the "Action")
- General procedure for counterfactual reasoning:
  - **Abduction:** determine the joint $P(u_1, \ldots, u_p|e)$ by using probabilistic inference over the intact model while incorporating the evidence e, and modify the model by replacing the original $P(u_1, \ldots, u_p)$ by this $P(u_1, \ldots, u_p|e)$.
  - **Impose action:** change the equation(s) of the model to reflect the hypothetical situation we want to analyze.
  - **Make counterfactual prediction:** use probabilistic inference over the resulting model to answer the query.

Counterfactual reasoning with Functional causal models

What is counterfactual reasoning: given some 'evidence' we want to see what would have possibly happened if some things had been done differently.

- For example, in the "Wet floor example":
  - we have observed that SLIPPERY = true (we call this the "Evidence")
  - and we want to know P(SLIPPERY = true) (we call this the "Query")
  - if we had forced SPRINKLER = off (we call this the "Action")
- General procedure for counterfactual reasoning:
  - **Abduction:** determine the joint $P(u_1, \ldots, u_p | e)$ by using probabilistic inference over the intact model while incorporating the evidence e, and modify the model by replacing the original $P(u_1, \ldots, u_p)$ by this $P(u_1, \ldots, u_p | e)$.
  - **Impose action:** change the equation(s) of the model to reflect the hypothetical situation we want to analyze.
  - **Make counterfactual prediction:** use probabilistic inference over the resulting model to answer the query.

Learnability of causal models

Suppose that we are given a sample $\left((x_1^i, \ldots, x_p^i)\right)_{i=1}^n$ i.i.d. from some functional causal model over the observed variables $X_1, \ldots, X_p$.

- How to infer the functional links f_i, when the structure is already given?
- How to infer the structure (or a part of it) of a causal model?
- What kind of experiments to conduct in order to gather additional data, so as to enable structure learning, or so as to merely speed-up learning?
- Given several data-sets under several experiments and/or under several different environments, how to use these to infer structure and functional links?

Learnability of causal models

Suppose that we are given a sample $\left((x_1^i, \ldots, x_p^i)\right)_{i=1}^n$ i.i.d. from some functional causal model over the observed variables $X_1, \ldots, X_p$.

- How to infer the functional links f_i, when the structure is already given?
- How to infer the structure (or a part of it) of a causal model?
- What kind of experiments to conduct in order to gather additional data, so as to enable structure learning, or so as to merely speed-up learning?
- Given several data-sets under several experiments and/or under several different environments, how to use these to infer structure and functional links?

Learnability of causal models

Suppose that we are given a sample $\left((x_1^i, \ldots, x_p^i)\right)_{i=1}^n$ i.i.d. from some functional causal model over the observed variables $X_1, \ldots, X_p$.

- How to infer the functional links f_i, when the structure is already given?
- How to infer the structure (or a part of it) of a causal model?
- What kind of experiments to conduct in order to gather additional data, so as to enable structure learning, or so as to merely speed-up learning?
- Given several data-sets under several experiments and/or under several different environments, how to use these to infer structure and functional links?

Learnability of causal models

Suppose that we are given a sample $\left((x_1^i, \ldots, x_p^i)\right)_{i=1}^n$ i.i.d. from some functional causal model over the observed variables $X_1, \ldots, X_p$.

- How to infer the functional links f_i, when the structure is already given?
- How to infer the structure (or a part of it) of a causal model?
- What kind of experiments to conduct in order to gather additional data, so as to enable structure learning, or so as to merely speed-up learning?
- Given several data-sets under several experiments and/or under several different environments, how to use these to infer structure and functional links?

Learnability of causal models

Suppose that we are given a sample $\left((x_1^i, \ldots, x_p^i)\right)_{i=1}^n$ i.i.d. from some functional causal model over the observed variables $X_1, \ldots, X_p$.

- How to infer the functional links f_i, when the structure is already given?
- How to infer the structure (or a part of it) of a causal model?
- What kind of experiments to conduct in order to gather additional data, so as to enable structure learning, or so as to merely speed-up learning?
- Given several data-sets under several experiments and/or under several different environments, how to use these to infer structure and functional links?

Cause-effect models over two variables...

Suppose that we only have two variables X and Y.

(NB: this viewpoint could be the result of splitting in some way $X_1, \ldots, X_p$ in two parts $X = (X_{l_1}, \ldots, X_{l_k})$ and $Y = (X_1, \ldots, X_p) \setminus X$.)

and that we want merely to infer from observational data whether $X \hookrightarrow Y$ or $Y \hookrightarrow X$.

(Assuming that we can exclude other Z options.)

In other words, we want to test $H_0 \left\{ \begin{array}{ccc} x & = & u_1 \\ y & = & f_y(x, u_2) \\ & \textit{with} & U_1 \perp U_2, \end{array} \right.$

with respect to the alternative $H_1 \left\{ \begin{array}{ccc} y & = & u'_1 \\ x & = & f_x(y, u'_2) \\ & \textit{with} & U'_1 \perp U'_2, \end{array} \right.$

given a dataset $\left((x^i, y^i)\right)_{i=1}^{n}$.

... Cause-effect models over two variables

Without specifying any information about $P(u_i)$, $P(u'_i)$, f_x and f_y the problem of chosing among H_0 and H_1 can not be solved effectively (see e.g. Chapter 4 of [3]).

… Cause-effect models over two variables

Without specifying any information about $P(u_i)$, $P(u_i')$, f_x and f_y the problem of chosing among H_0 and H_1 can not be solved effectively (see e.g. Chapter 4 of [3]).

Notice that this question is similar to asking wether $P(x)P(y|x)$ or $P(y)P(x|y)$ is the right way of modelling the relation between X and Y.

Without specifying any information about $P(u_i)$, $P(u_i')$, f_x and f_y the problem of chosing among H_0 and H_1 can not be solved effectively (see e.g. Chapter 4 of [3]).

Notice that this question is similar to asking wether $P(x)P(y|x)$ or $P(y)P(x|y)$ is the right way of modelling the relation between X and Y.

A possible way out, is to make assumptions about the family of functions used to model f_x and f_y, and the family of noise distributions $P(u_i)$ and $P(u_i')$.

… Cause-effect models over two variables

Without specifying any information about $P(u_i)$, $P(u_i')$, f_x and f_y the problem of chosing among H_0 and H_1 can not be solved effectively (see e.g. Chapter 4 of [3]).

Notice that this question is similar to asking wether $P(x)P(y|x)$ or $P(y)P(x|y)$ is the right way of modelling the relation between X and Y.

A possible way out, is to make assumptions about the family of functions used to model f_x and f_y, and the family of noise distributions $P(u_i)$ and $P(u_i')$.

Another way out is to accept the idea that experiments should be carried out to help deciding about H_0 with respect to H_1.

Cause-effect models over more than two variables

If enough data is available, the skeleton and set of V-structures can be inferred from observational datasets only.

The resulting (essential) graph is semi-directed in general.

This essential graph can be used to construct queries and experiments in order to further direct it.

It is an active research field.

How to automatize reasoning about how the world works?

Leveraging principles of causal modelling in machine learning

- Pure observational supervised learning: if we know that $X \hookrightarrow Y$, rather try to model $P(y|x)$ directly, rather than modeling $P(x, y)$ and doing inference.
- Transportability and transfer learning: understanding the causal relations among variables helps to formulate more 'stable' models, which can be learned in a more robust way.
- Active learning, reinforcement learning, development of algorithms for handling the exploitation versus exploration dilemma in an intelligent way, from diverse datasets.
- The modelling of (causal) mechanisms yielding missing values, and using these models in the context of learning.

This is not black magic but topics for research and engineering.

References

[1] H. Reichenbach, The Direction of Time. Univ. of California Berkeley Press, 1956.

[2] J. Pearl, Causality. Cambridge university press, 2009. [Online]. Available: http://bayes.cs.ucla.edu/jp_home.html

[3] J. Peters, D. Janzing, and B. Schölkopf, Elements of causal inference: foundations and learning algorithms. MIT press, 2017. [Online]. Available: https://mitpress.mit.edu/books/elements-causal-inference

[4] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," The Journal of Machine Learning Research, vol. 14, no. 1, pp. 3207–3260, 2013.

Tutorials and Talks on the WEB:

- Mini course on Causality at MIT - Jonas Peters, YouTube 2018
- Judea Pearl and Elias Bareinbaum - Various talks, YouTube
- Actual Causality: A Survey - Joseph Halpern, YouTube 2018