When causality matters for prediction: Investigating the practical tradeoffs

Robert E. Tillman       Peter Spirtes

Carnegie Mellon
Department of Philosophy       Machine Learning Department
College of Humanities and Social Sciences       School of Computer Science

NIPS 2008 Workshop on Causality: Objectives and Assessment
Causal Discovery

The Usual Setup:
- Unobserved data generating process
- i.i.d. sample
Causal Discovery

The Usual Setup:
- Unobserved data generating process
- i.i.d. sample

Objective:
- Learn structure, e.g. causal Bayesian network
The Usual Setup:

- Unobserved data generating process
- i.i.d. sample

Objective:

- Learn structure, e.g. causal Bayesian network

Assessment:

- Compare to “ground truth”, i.e. simulations, experimental studies, expert knowledge
Causal Discovery

The Usual Setup:
- Unobserved data generating process
- i.i.d. sample

Objective:
- Learn structure, e.g. causal Bayesian network

Assessment:
- Compare to “ground truth”, i.e. simulations, experimental studies, expert knowledge

Focus:
- Learn network models that accurately depict the data generating mechanism
The Standard Problem:
- “Target” variable associated with “predictor” variables
- i.i.d sample (training data)
The Standard Problem:
- “Target” variable associated with “predictor” variables
- i.i.d sample (training data)

Objective:
- Predict target from values of predictor variables
The Standard Problem:
- “Target” variable associated with “predictor” variables
- i.i.d sample (training data)

Objective:
- Predict target from values of predictor variables

Assessment:
- Compare predictions to known target values, i.e. testing data, cross validation
Prediction

The Standard Problem:
- “Target” variable associated with “predictor” variables
- i.i.d sample (training data)

Objective:
- Predict target from values of predictor variables

Assessment:
- Compare predictions to known target values, i.e. testing data, cross validation

Focus:
- Train classifier/regression model that minimizes loss function, e.g. makes accurate predictions
- Model need not resemble the true data generating mechanism, i.e. Naive Bayes
Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population
Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population

Causation and Prediction Challenge:

- Training data from unmanipulated population
Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population

Causation and Prediction Challenge:

- Training data from unmanipulated population
- (Structural) intervention is performed
- System stabilizes
Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:
- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population

Causation and Prediction Challenge:
- Training data from unmanipulated population
- (Structural) intervention is performed
- System stabilizes
- Draw i.i.d sample for predictors from manipulated population
- Predict target using predictor values from stabilized manipulated distribution
Causation and Prediction Challenge

Results:

- Participants used causal methods and methods which ignore causality
Causation and Prediction Challenge

Results:

- Participants used causal methods and methods which ignore causality.
- Some top-ranking participants did not use causal methods, i.e. support vector machines (for feature selection and classification).
- Other participants using causal methods did not do as well.
Causation and Prediction Challenge

Results:

- Participants used causal methods and methods which ignore causality.
- Some top-ranking participants did not use causal methods, i.e. support vector machines (for feature selection and classification).
- Other participants using causal methods did not do as well.

Questions:

- Is causality useful for standard prediction tasks?
Causation and Prediction Challenge

Results:
- Participants used causal methods and methods which ignore causality
- Some top-ranking participants did not use causal methods, i.e. support vector machines (for feature selection and classification)
- Other participants using causal methods did not do as well

Questions:
- Is causality useful for standard prediction tasks?
- Is it useful in practice?
Causation and Prediction Challenge

Results:
- Participants used causal methods and methods which ignore causality.
- Some top-ranking participants did not use causal methods, i.e., support vector machines (for feature selection and classification).
- Other participants using causal methods did not do as well.

Questions:
- Is causality useful for standard prediction tasks?
- Is it useful in practice?
- Is this a realistic scenario?
Results:
- Participants used causal methods and methods which ignore causality
- Some top-ranking participants did not use causal methods, i.e. support vector machines (for feature selection and classification)
- Other participants using causal methods did not do as well

Questions:
- Is causality useful for standard prediction tasks?
- Is it useful in practice?
- Is this a realistic scenario?

Possible Explanations:
- Sampling error, overfitting
- Parametric assumptions do not hold, i.e. linearity, Gaussianity
- Prediction for target is invariant under the manipulation.
Invariance of prediction under manipulations

Simple example:

Bayes optimal prediction for $Y$ is $P(Y|X)$
Simple example:

Bayes optimal prediction for $Y$ is $P(Y|X)$

- Manipulating $X$ does not change distribution of $P(Y|X)$, still Bayes optimal
- Prediction (once system stabilizes) is invariant under manipulation
Invariance of prediction under manipulations

Simple example:

Bayes optimal prediction for $Y$ is $P(Y|X)$

- Manipulating $Y$ does change distribution of $P(Y|X)$, $Y$ depends on manipulation
- Incorrect predictions in stabilized manipulated population
Predict CiliaDam
Parents of CiliaDam
When causality matters for prediction

Tillman and Spirtes

NIPS 2008 Workshop on Causality
Coparents (spouses) of CiliaDam
In a causal Bayesian network \( \mathcal{B} = \langle \mathcal{G}, P \rangle \) over variables \( V \), the Markov Blanket for \( X \in V \) is the minimal set of variables \( \text{MB}_X \subseteq V/\{X\} \) such that \( X \perp \!\!\!\!\perp V/\text{MB}_X | \text{MB}_X \).
In a causal Bayesian network \( \mathcal{B} = \langle \mathcal{G}, P \rangle \) over variables \( V \), the Markov Blanket for \( X \in V \) is the minimal set of variables \( \text{MB}_{X}^{G} \subseteq V / \{X\} \) such that \( X \perp \!\!\!\perp V / \text{MB}_{X}^{G} | \text{MB}_{X}^{G} \).

**Theorem (Pearl, 1988)**

The Markov blanket for \( X \) consists of the parents, children and coparents of \( X \) in \( \mathcal{G} \).
Interventions

Policy(\text{Smoker})=0

Policy(\text{Smoker})=1
Conditions for invariance of prediction under manipulations

Theorem (Prediction invariance)

In a causal Bayesian network \( B = \langle G, P \rangle \) over variables \( V \), let \( T \in V \) be a target, \( X \subseteq V \) a set of predictor variables, and \( Y \subseteq V \) the set of manipulated variables. If \( X \supseteq \text{MB}_T^G \) and \( \forall Y \in Y, Y \neq T \) and \( Y \notin \text{Children}(T) \), then prediction of \( T \) using \( X \) is invariant under the manipulation.

\[ \text{Income} \quad \text{Parent} \]

\[ \text{Smoker} \quad \text{Pollution} \]

\[ \text{CiliaDam} \quad \text{Genotype} \]

\[ \text{LungCapac} \quad \text{HeartDis} \]

\[ \text{BreathDis} \]
Conditions for invariance of prediction under manipulations

\[ P(T \mid X) = P(T \mid MB_T^G) \]
Conditions for invariance of prediction under manipulations

\[ P(T \mid X) = P(T \mid MB^G_T) \]
\[ = \frac{P(T, MB^G_T)}{\sum_T P(T, MB^G_T)} \]
Causation and Prediction

Invariance of prediction functions

Experimental Results

Conclusions

Conditions for invariance of prediction under manipulations

\[ P(T \mid X) = P(T \mid \text{MB}_T^G) \]
\[ = \frac{P(T, \text{MB}_T^G)}{\sum_T P(T, \text{MB}_T^G)} \]
\[ = \frac{\prod_{X \in T \cup \text{Children}(T) \cup \text{Parents}(T) \cup \text{Coparents}(T)} P(X \mid \text{Parents}(T))}{\sum_T \prod_{X \in T \cup \text{Children}(T) \cup \text{Parents}(T) \cup \text{Coparents}(T)} P(X \mid \text{Parents}(T))} \]

in the Markov blanket subgraph
Conditions for invariance of prediction under manipulations

\[
P(T \mid X) = P(T \mid \text{MB}_T^G) \\
= \frac{P(T, \text{MB}_T^G)}{\sum_T P(T, \text{MB}_T^G)} \\
= \frac{\prod_{X \in T \cup \text{Children}(T) \cup \text{Parents}(T) \cup \text{Coparents}(T)} P(X \mid \text{Parents}(T))}{\sum_T \prod_{X \in T \cup \text{Children}(T) \cup \text{Parents}(T) \cup \text{Coparents}(T)} P(X \mid \text{Parents}(T))}
\]

in the Markov blanket subgraph

\[
\ldots
\]

\[
= \frac{\prod_{X \in T \cup \text{Children}(T)} P(X \mid \text{Parents}(T))}{\sum_T \prod_{X \in T \cup \text{Children}(T)} P(X \mid \text{Parents}(T))}
\]
Correcting for manipulations

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathcal{V}$, let $T$ be a target and $Y \subseteq \mathcal{V}$ the set of manipulated variables. $P \left( T \mid MB_T^G(\text{Policy}(Y)) \right)$, is invariant under the manipulation of $Y$ if $\nexists Y \in Y$, such that $Y \in \text{Children}(T)$ and $Y$ is an ancestor of some $C \in \text{Children}(T) \cap \mathcal{V} / Y$. 

Policy(BreathDis) = 0
Correcting for manipulations

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $V$, let $T$ be a target and $Y \subseteq V$ the set of manipulated variables. $P \left( T \mid MB_T^{G(Policy(Y))} \right)$, is invariant under the manipulation of $Y$ if $\nexists Y \in Y$, such that $Y \in Children(T)$ and $Y$ is an ancestor of some $C \in Children(T) \cap V/Y$.

Policy(BreathDis) = 1
Causation and Prediction

Invariance of prediction functions

Experimental Results

Conclusions

Correcting for manipulations

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle G, P \rangle$ over variables $V$, let $T$ be a target and $Y \subseteq V$ the set of manipulated variables. $P \left( T \mid \text{MB}_T^G(\text{Policy}(Y)) \right)$, is invariant under the manipulation of $Y$ if $\nexists Y \in Y$, such that $Y \in \text{Children}(T)$ and $Y$ is an ancestor of some $C \in \text{Children}(T) \cap V/Y$. 

Policy(BreathDis) = 0
Correcting for manipulations

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $V$, let $T$ be a target and $Y \subseteq V$ the set of manipulated variables. $P \left( T \mid MB^{G(\text{Policy}(Y))}_T \right)$, is invariant under the manipulation of $Y$ if $\not\exists Y \in Y$, such that $Y \in \text{Children}(T)$ and $Y$ is an ancestor of some $C \in \text{Children}(T) \cap V/Y$. 

$\text{Policy}(\text{BreathDis}) = 1$

Make Correction!
Causation and Prediction

Invariance of prediction functions

Experimental Results

Conclusions

Model for experiments
Experiments

Method:

- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
Method:

- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of $T$ (including Markov blanket)
Experiments

Method:

- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of $T$ (including Markov blanket)
- Manipulate 0, . . . , 9 children of $T$ in addition
Experiments

Method:
- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of $T$ (including Markov blanket)
- Manipulate 0, . . . , 9 children of $T$ in addition
- Predict $T$ from manipulated distribution
Experiments

Method:
- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of $T$ (including Markov blanket)
- Manipulate 0, . . . , 9 children of $T$ in addition
- Predict $T$ from manipulated distribution

Hypotheses:
- Noncausal methods will be equivalent or better when no children are manipulated
Method:
- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of $T$ (including Markov blanket)
- Manipulate 0, ..., 9 children of $T$ in addition
- Predict $T$ from manipulated distribution

Hypotheses:
- Noncausal methods will be equivalent or better when no children are manipulated
- Causal methods will do increasingly better than noncausal methods as more children are manipulated
Differences between distributions

Squared difference between ground truth predictions for $T$ using unmanipulated and manipulated model
Prediction methods

Noncausal Methods:

- **LR-ALL**: linear regression using all predictors
- **LR-MB**: linear regression using only the Markov blanket
- **LASSO**: “least absolute shrinkage and selection operator”
- **SVR-RBF**: support vector regression using radial kernel
- **RVR-RBF**: relevance vector regression using radial kernel
Prediction methods

Noncausal Methods:

- **LR-ALL**  linear regression using all predictors
- **LR-MB**  linear regression using only the Markov blanket
- **LASSO**  “least absolute shrinkage and selection operator”
- **SVR-RBF**  support vector regression using radial kernel
- **RVR-RBF**  relevance vector regression using radial kernel

Causal Methods:

- **LR-MB/C**  linear regression with Markov blanket correcting for manipulated children
- **LR-MB/C**  linear regression with Markov blanket correcting for manipulated children and active paths to unmanipulated children
Total prediction error

0 Manipulated Nonchildren of $T$
5 Manipulated Nonchildren of $T$
Total prediction error

10 Manipulated Nonchildren of $T$
Nonlinear data

- Repeated previous simulations adding nonlinear dependencies
Nonlinear data

- Repeated previous simulations adding nonlinear dependencies
- Results so far inconclusive
- In general, nonparametric methods do best, though poor performance in all cases
Is causality relevant for prediction?
Conclusions

Is causality relevant for prediction?

- Yes, as long as a noncausal method is not invariant under the manipulation
Conclusions

Is causality relevant for prediction?

- Yes, as long as a noncausal method is not invariant under the manipulation
- But causality is needed to know noncausal methods are invariant
Conclusions

Is causality relevant for prediction?
- Yes, as long as a noncausal method is not invariant under the manipulation
- But causality is needed to know noncausal methods are invariant

In practice?
- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
Conclusions

Is causality relevant for prediction?

- Yes, as long as a noncausal method is not invariant under the manipulation
- But causality is needed to know noncausal methods are invariant

In practice?

- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
- Noncausal prediction may be frequently invariant under manipulations or only make small errors related to causality
Conclusions

Is causality relevant for prediction?
- Yes, as long as a noncausal method is not invariant under the manipulation
- But causality is needed to know noncausal methods are invariant

In practice?
- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
- Noncausal prediction may be frequently invariant under manipulations or only make small errors related to causality
- Advantages of nonparametric methods and methods which deal with overfitting well may cancel out errors related to causality
Conclusions

Is causality relevant for prediction?
- Yes, as long as a noncausal method is not invariant under the manipulation
- But causality is needed to know noncausal methods are invariant

In practice?
- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
- Noncausal prediction may be frequently invariant under manipulations or only make small errors related to causality
- Advantages of nonparametric methods and methods which deal with overfitting well may cancel out errors related to causality
- Many other variables involved, analysis incomplete
Conclusions

Is causality relevant for prediction?
- Yes, as long as a noncausal method is not invariant under the manipulation
- But causality is needed to know noncausal methods are invariant

In practice?
- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
- Noncausal prediction may be frequently invariant under manipulations or only make small errors related to causality
- Advantages of nonparametric methods and methods which deal with overfitting well may cancel out errors related to causality
- Many other variables involved, analysis incomplete

Future directions for causal discovery:
- Methods which deal with overfitting well
Conclusions

Is causality relevant for prediction?

- Yes, as long as a noncausal method is not invariant under the manipulation
- But causality is needed to know noncausal methods are invariant

In practice?

- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
- Noncausal prediction may be frequently invariant under manipulations or only make small errors related to causality
- Advantages of nonparametric methods and methods which deal with overfitting well may cancel out errors related to causality
- Many other variables involved, analysis incomplete

Future directions for causal discovery:

- Methods which deal with overfitting well
- Less restrictive parametric assumptions