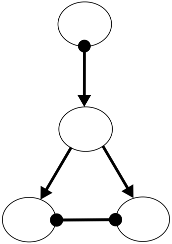
Probabilistic Graphical Models: Principles and Applications

Chapter 13: CAUSAL GRAPHICAL MODELS

L. Enrique Sucar, INAOE



Outline

Introduction

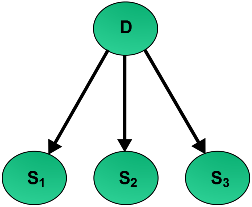
- Causality has to do with cause–effect relations; that is, identifying when there are two (or more) related phenomena, which is the *cause* and which is the *effect*
- Probabilistic models do not necessarily represent causal relations
- Consider the following two directed Bayesian networks: BN1: $A \rightarrow B \rightarrow C$ and BN2: $A \leftarrow B \leftarrow C$
- They represent the same conditional independence relations; however, if we define that a directed link $A \rightarrow B$ means A causes B, both models represent very different causal relations

Introduction

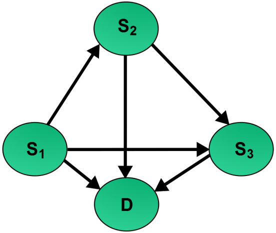
- There are several advantages to causal models, in particular, graphical causal models
- First, if we build a Bayesian network based on causal relations, the models tend to be easier to construct and understand (e.g., for communicating with domain experts), and are usually simpler
- Secondly, with causal models we can perform other types of reasoning that are not possible, at least in a direct way, with BNs:
  1. *interventions*, in which we want to find the effects of setting a certain variable to a specific value by an external agent (note that this is different from observing a variable)
  2. *counterfactuals*, where we want to reason about what would have happened if certain information had been different from what actually happened

Introduction

- For example, in the medical domain it is common to have a disease that produces several symptoms, which usually are conditionally independent given the disease
- If we represent this as a BN in the direction of causality, we obtain a simple, star structure
- An alternative structure requires links between the symptom variables, as these are conditionally independent given the disease but not marginally independent



(a)



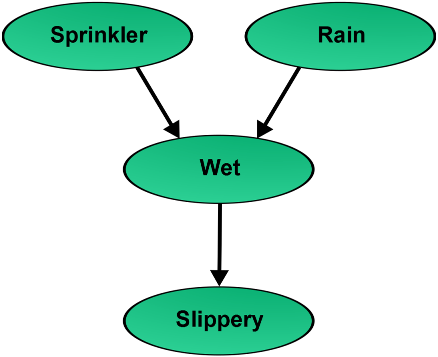
(b)

Causal Bayesian Networks

- A Causal Bayesian network (CBN) is a directed acyclic graph, G, in which each node represents a variable and the arcs in the graph represent *causal* relations
- $A \rightarrow B$ represents some physical mechanism such that the value of B is directly affected by the value of A
- This relation can be interpreted in terms of *interventions* –setting of the value of some variable or variables by an external agent
- Causal networks represent stronger assumptions than Bayesian networks, as all the relations implied by the network structure should correspond to causal relations

Example

- A simple example of a CBN – which basically encodes the following causal relations: (i) *Sprinkler* causes *Wet*, (ii) *Rain* causes *Wet*, (iii) *Wet* causes *Slippery*



Exogenous / Endogenous variables

- In a CBN, a variable which is a root node (variable with no parents) is called *exogenous*, and all other variables are *endogenous*.
- In the previous example, *Sprinkler* and *Rain* are exogenous variables, and *Wet* and *Slippery* are endogenous variables
- In a CBN if any of the parent variables of X, or any combination of them, is set to certain value via an intervention, this will have an effect on X

DO operation

- If a $P(\mathbf{X})$ is the joint probability distribution of the set of variables $\mathbf{X}$, then we define $P_y(\mathbf{X})$ as the distribution resulting from setting the value for a subset of variables, $\mathbf{Y}$, via an intervention
- This can be represented as $do(\mathbf{Y} = \mathbf{y})$, where $\mathbf{y}$ is a set of constants
- In the example, if we set the sprinkler to ON –$do(Sprinkler = ON)$, then the resulting distribution will be denoted as $P_{Sprinkler=ON}(Sprinkler, Rain, Wet, Slippery)$

Definition

- A CBN G is a directed acyclic graph over a set of variables $\mathbf{X}$ that is compatible with all the distributions resulting from interventions on $\mathbf{Y} \subseteq \mathbf{X}$, in which the following conditions are satisfied:
  1. The probability distribution $P_y(\mathbf{X})$ resulting from an intervention is Markov compatible with the graph G; that is, it is equivalent to the product of the conditional probability of each variable $X \in G$ given its parents: $P_y(\mathbf{X}) = \prod_{X_i} P(X_i \mid Pa(X_i))$.
  2. The probability of all the variables that are part of an intervention is equal to one for the value it is set to: $P_y(X_i) = 1$ if $X_i = x_i$ is consistent with $Y = y$, $\forall X_i \in \mathbf{Y}$.
  3. The probability of each of the remaining variables that are not part of the intervention is equal to the probability of the variable given its parents and it is consistent with the intervention: $P_y(X_i \mid Pa(X_i)) = P(X_i \mid Pa(X_i))$, $\forall X_i \notin \mathbf{Y}$.

Joint distribution

- Given the previous definition, the joint probability distribution can be calculated as a *truncated* factorization:

$$P_y(\mathbf{X}) = \prod_{X_i \notin \mathbf{Y}} P(X_i \mid Pa(X_i)) \tag{1}$$

such that all X_i are consistent with the intervention $\mathbf{Y}$

- Once all the parents of a variable X_i are set by an intervention, setting any other variable does not affect the probability of X_i:

$$P_{Pa(X_i),\mathbf{W}}(X_i) = P_{Pa(X_i)}(X_i) \tag{2}$$

such that $\mathbf{W} \cap (X_i, Pa(X_i)) = \emptyset$.

Causal reasoning

- Causal reasoning has to do with answering causal queries from a causal model, and in our case in particular from graphical causal models
- There are several types of causal queries we might consider, we will start by analyzing causal predictions, and then we will analyze counterfactuals

Prediction

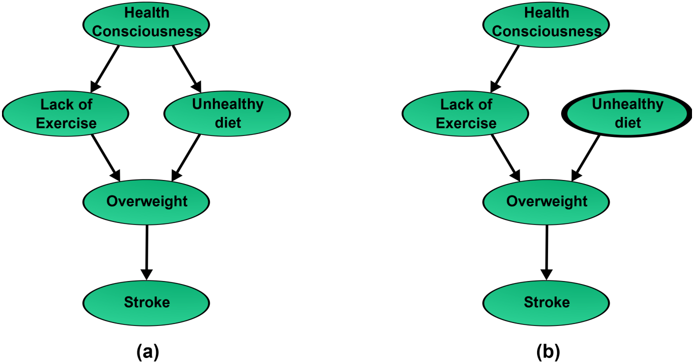
- Consider a causal Bayesian network which includes a set of variables: $X_G = \{X_C, X_E, X_O\}$; where X_C is a subset of *causes* and X_E is a subset of *effects*; X_O are the remaining variables
- What will the effect be on X_E when setting $X_C = x_C$?
- That is, we want to obtain the probability distribution of X_E that results from the intervention $X_C = x_C$:

$$P_C(X_E \mid do(X_C = x_C)) \tag{3}$$

Procedure

- To perform causal prediction with a CBN, G, the following procedure is followed:
  1. Eliminate all incoming arrows in the graph to all nodes in X_C, thus obtaining a modified CBN, G_r.
  2. Fix the values of all variables in X_C, $X_C = x_C$.
  3. Calculate the resulting distribution in the modified model G_r (via probability propagation as in Bayesian networks).

Example



- If we want to measure the effect of an “unhealthy diet” in the variable “stroke”, then we eliminate the link from “health consciousness” to “diet”
- Then we should set the value of “unhealthy diet” to *TRUE*, and by probability propagation obtain the distribution of “stroke

Interventions vs. Observations

- When is the distribution resulting from an intervention equal to the distribution resulting from an observation?
- In mathematical terms, is $P(X_E \mid X_C) = P_C(X_E \mid do(X_C = x_C))$?
- Both are equal if X_C includes all the parents of X_E and none of its descendants; given that any variable in a BN is independent of its non-descendants given its parents
- In other cases they are not necessarily equal, it will depend on other conditions

Counterfactuals

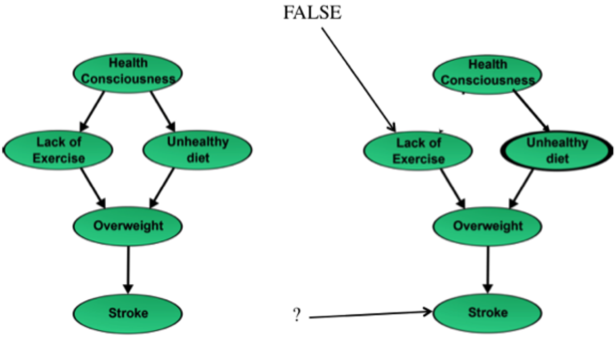
- Counterfactuals are a way of reasoning that we commonly perform in our life
- A typical counterfactual question would be: If some person suffered a stroke, would he still have suffered the stroke ("stroke"=TRUE) if he had exercised more ("lack of exercise"=FALSE)?

Inference

- Counterfactual inference involves 3 main steps:
  1. Abduction: modify the model in terms of the new evidence (in the example, modify the value of “stroke” to unknown).
  2. Action: perform the minimal intervention in the model according to the hypothetical evidence (in the example, set “lack of exercise” to FALSE, removing the link from “health consciousness” to “lack of exercise”).
  3. Prediction: perform probabilistic inference on the modified model and obtain the probability of the variable(s) of interest (in the example, perform probability propagation to estimate $P_{lack-exercise}(stroke \mid do(lack - exercise = FALSE))$.

Example

- *... will he had suffered the stroke if he had exercised more?*



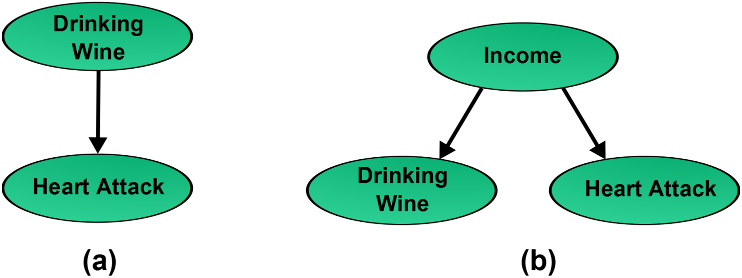
Learning Causal Models

- Learning causal models from data without direct interventions poses many challenges
- If we discover that there is a certain dependency between two variables, X and Y, we can not determine, without additional information, if X *causes* Y or vice versa
- Additionally, there could be some other factor that produces the dependency between these two variables

Example

- Consider that based on data we discover that people that drink wine tend to have less heart attacks
- Then we might be inclined to conclude that drinking wine tends to decrease the probability of a heart attack
- However, there might be another variable that produces this apparent causal relation, known as a *latent common cause* – it could be that both, wine drinking and heart attacks, have to do with the income level of the person, as persons with high income tend to drink more wine and at the same time have a lower probability of heart attack

Example



- A difficulty of learning causal relations lies in how to include in the model all the relevant factors

Assumptions

- In general, the following assumptions are used when learning the structure of causal networks:

  Causal Markov Condition: a variable is independent of its non-descendants given its direct causes (parents in the graph).

  Faithfulness: there are no additional independencies between the variables in the model that are not implied by the causal Markov condition.

  Causal Sufficiency: there are no common confounders of the observed variables in the model.

Bayesian Constraint-Based Causal Discovery

- An algorithm for learning causal networks
- This technique is an extension of the PC Bayesian network structure learning algorithm:
  1. Start with a completely connected graph and estimate the reliability of each causal link, $X - Y$, by measuring the conditional independence between X and Y. If a pair of variables are conditionally independent with a reliability above a certain threshold, then delete the edge between these variables.
  2. The remaining causal relations (undirected edges in the graph) are ordered according to their reliability. Then the edges in the graph are oriented starting from the most reliable relations, based on the conditional independence test for variable triples.

Reliaibility estimation

- To estimate the reliability of a causal relation, $R = X \rightarrow Y$, the algorithm uses a Bayesian score:

$$P(R \mid D) = \frac{P(D \mid M_R)P(M_R)}{P(D \mid M)p(M)} \tag{4}$$

- D is the data, M are all the possible structures, and M_R are all the structures that contain the relation R
- Thus, $P(M)$ denote the prior probability of a structure M and $P(D \mid M)$ the probability of the data given the structure
- Calculating this is very costly, so it is approximated by the marginal likelihood of the data given the structure

Maximal Ancestral Graphs

- Depending on the reliability threshold, the resulting network can have undirected edges, $-$, which means that there is not enough information to obtain the direction of the link, and bi-directed edges, $\leftrightarrow$, indicating that there is a common cofounder
- This type of structure is called a *Maximal Ancestral Graph* (MAG)
- The equivalence class for MAGs is a *Partial Ancestral Graph* (PAG)
- In a PAG there are three types of edges: directed, $\rightarrow$, and undirected, $-$, when these are consistent for all the graphs in the equivalence class; and those which are not consistent are marked with a circle, "o"

Applications

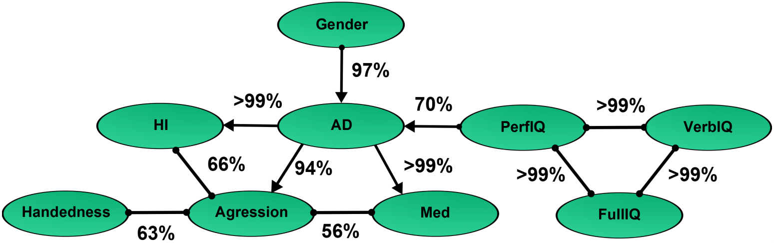
- There are many practical applications in which a causal model is useful. Just to mention a few:
  - Predicting the effects of certain interventions.
  - Learning causal graphical models from data.
  - Diagnosis of physical systems.
  - Generating explanations.
  - Understanding causal expressions.

Learning a Causal model for ADHD

- Data set that contains phenotypic information about children with Attention Deficit Hyperactivity Disorder (ADHD):
  1. Gender (male / female)
  2. Attention deficit score (continuous)
  3. Hyperactivity / impulsivity score (continuous)
  4. Verbal IQ (continuous)
  5. Performance IQ (continuous)
  6. Full IQ (continuous)
  7. Aggressive behavior (yes/no)
  8. Medication status (naïve / not naïve)
  9. Handedness (right / left)
- 223 subjetcs

PAG

- Using a reliability threshold of 50%, a network was obtained; which is represented as a parental ancestral graph



Interpretation

- Several interesting causal relations are suggested by the resulting network, some of which were already known from previous medical studies:
  - There is a strong effect from gender on the level of attention deficit.
  - The level of attention deficit affects hyperactivity / impulsivity and aggressiveness.
  - Handedness (left) is associated with aggressive behavior.
  - The association between performance IQ, verbal IQ and full IQ is explained by a latent common cause.
  - Only the performance IQ has a direct causal link with attention deficit.

Book

Sucar, L. E, *Probabilistic Graphical Models*, Springer 2015 – Chapter 13

Additional Reading - Causal Models

- Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, New York (2009)
- Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, MIT Press (2000)
- Wright, S.: Correlation and Causation. Journal of Agricultural Research. 20, 557–585 (1921)

Additional Reading - Inference and Learning

- Claassen, T., Heskes, T: A Bayesian Approach to Constraint Based Causal Inference. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), AUAI Press, pp. 207–216 (2012)
- Aitken, J.S.: Learning Bayesian Networks: Approaches and Issues. The Knowledge Engineering Review. 26(2), 99–157 (2011)
- Sokolova, E., Groot, P., Classen, T., Heskes, T.: Causal Discovery from Databases with Discrete and Continuous Variables. In: Probabilistic Graphical Models (PGM), pp. 442–457. Springer-Verlag (2014)