732A75 Advanced Data Mining
TDDD41 Data Mining - Clustering and Association Analysis
Lecture 9: Exercises and Causal Discovery

Jose M. Peña
IDA, Linköping University, Sweden

Outline

- Content
  - Exercises
  - Causality in a nutshell
  - Randomized controlled trial
  - Correlation is not causation
  - Local causal discovery algorithm
  - Summary
- Literature
  - Cooper, G. F. A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. Data Mining and Knowledge Discovery 1, 203-224 (1997).
  - Silverstein, C., Brin, S., Motwani, R. and Ullman, J. Scalable Techniques for Mining Causal Structures. Data Mining and Knowledge Discovery 4, 163-192 (2000).

Exercises

- Run the Apriori algorithm on the following transactional database with minimum support equal to one transaction. Explain step by step the execution.

Transaction id	Items bought
1	A, B, C
2	X, Y, Z
3	A, Y, C
4	X, B, Z

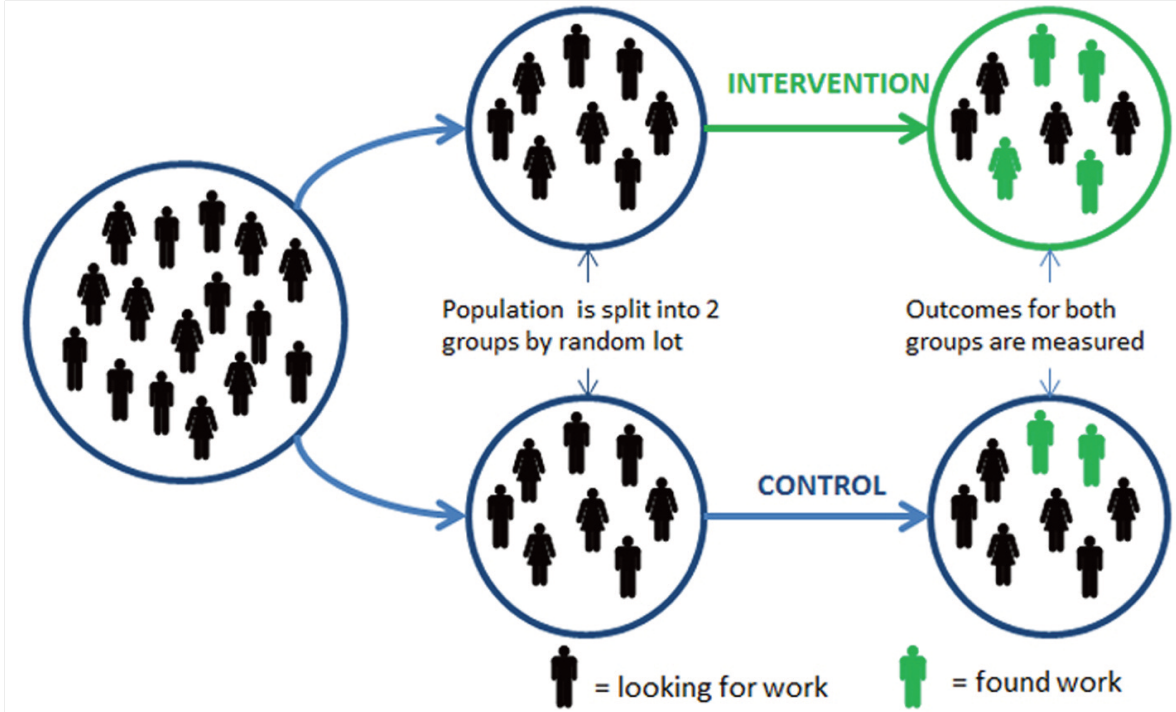
- Repeat the exercise above with the following additional constraint: Find the frequent itemsets that contain the item A. Make clear when and how the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.
- Let the items A, B, C, X, Y and Z have a price of respectively -3, -2, -1, 1, 2 and 3 units. Repeat the exercise above with the following additional constraint: Find the frequent itemsets whose range is smaller than 3 (recall that the range is the price of the most expensive item minus the price of the cheapest item). Make clear when and how the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.
- Repeat the exercises above with the FP grow algorithm.
- Apply the Simple algorithm to the frequent itemset XBZ on the database above in order to find association rules with confidence greater than 0.5.

Causality in a Nutshell

- Consider the following scenario:
  - Buying eggs typically makes customers buy bread and butter as well.
  - Customers typically buy eggs and milk together.
- In other words:
  - Eggs causes bread and butter.
  - Milk and eggs are statistically dependent but not necessarily causally related.
- Then, the following association rules should have good confidence:
  - $milk, eggs \rightarrow bread, butter$
  - $milk \rightarrow bread, butter$
- However, intervening/forcing/making customers buy milk (e.g. by giving them a discount) may not increase the sales of bread and butter (the opposite may actually happen).
- The problem lies in that association rules represent statistical rather than casual relationships.
- Association analysis is an exploratory rather than a confirmatory task.
- We say that X causes Y if $p(Y|do(X = x)) \neq p(Y|do(X = x'))$ for some x and x' values of X. In other words, intervening on X affects Y.
  - Intervention: Fix the value of a variable (for the whole population) so that it is no longer governed by its natural causes.
  - Observation: Focus on the subpopulation that attains a particular value for a variable.
  - It may be that $p(cholesterol|do(exercise)) \neq p(cholesterol|exercise)$.

Randomized Controlled Trial

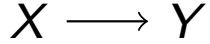
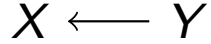
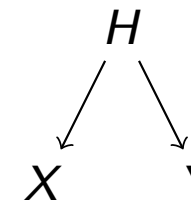
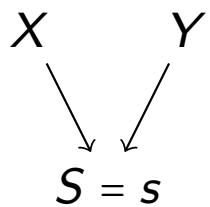
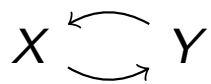
- Does the new job regulation create more jobs ? That is, $NJR \rightarrow MJ$?



- Gold standard for causal discovery, but it is not always feasible, e.g. the treatment/intervention may be too costly or prohibited due to ethical considerations.
- What can we do then to discover causal relationships ? Use observational data.
  - Table D with N samples (rows) of V random variables (columns).
  - All entries in D are filled, i.e. no missing data.
  - Not necessarily transactional data, e.g. the variables may be discrete, continuous, or ordinal.

Correlation is not Causation

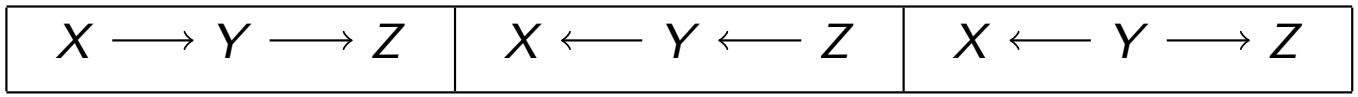
- Assume that X and Y are dependent, i.e. $p(X, Y) \neq p(X)p(Y)$. Then, this may be due to

Causal	Causal	Confounding
		
Selection bias	Feedback	Combinations
		...

- If we know that there are not hidden variables and X is not caused by Y (e.g. $X = \textit{gender}$ or X temporally precedes Y), then $X \to Y$ is the only explanation.
  - Note that X is not necessarily a direct cause of Y, i.e. there may be mediators.

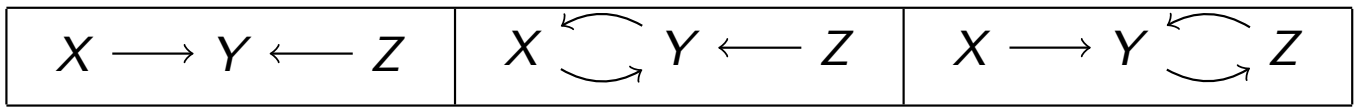
Correlation is not Causation

- Assume that X and Y are dependent, Y and Z are dependent, and X and Z are independent given Y, i.e. $p(X, Z|Y) = p(X|Y)p(Z|Y)$ for all Y with $p(Y) > 0$. Then, this may be due to



  or confounding, selection bias, feedback loops or combinations thereof.

- However, never due to



- If we know that there are not hidden variables and X is not caused by Y, then $X \to Y \to Z$ is the only explanation.
- If we know that there are not hidden variables and Y is not caused by X or Z, then $X \leftarrow Y \to Z$ is the only explanation.
- Assume that X and Y are dependent, Y and Z are dependent, X and Z are independent, and X and Z are dependent given Y. If we know that there are not hidden variables, then $X \to Y \leftarrow Z$ is the only explanation.
  - If hidden variables may be present, then $X \leftarrow H_1 \to Y \leftarrow H_2 \to Z$ is also a possible explanation.

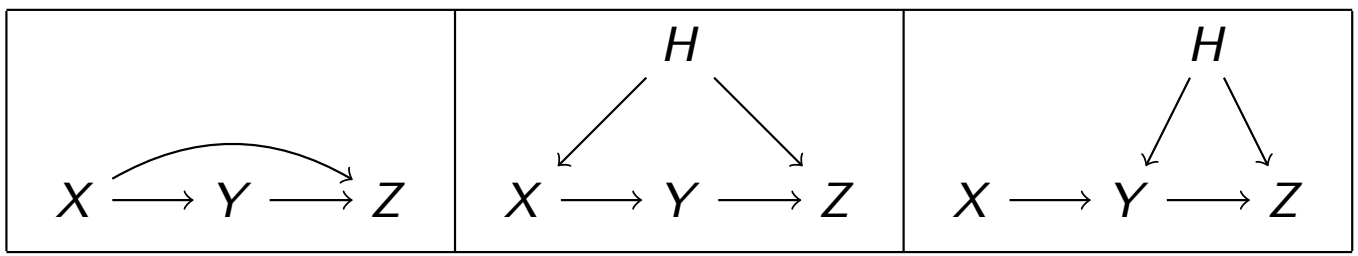
Local Causal Discovery Algorithm

- If we know that some variable is uncaused by the other observed variables and there is no selection bias (there may be confounding though), then the following algorithm returns correct causal relationships.

Algorithm: LCD(D, V, X)
Input: An observational database D, a set of variables V, and a variable X which is not caused by any other variable in V.
Output: Causal relationships discovered in D.

1 for all $Y \in V \setminus \{X\}$ do
2 if X and Y are dependent in D then
3 for all $Z \in V \setminus \{X, Y\}$ do
4 if Y and Z are dependent in D then
5 if X and Z are independent given Y in D then
6 output $Y \to Z$ and $p(Z|Y)$ as estimated from D

- Note that $p(Z|do(Y)) = p(Z|Y)$ in this case.
- Note that the LCD algorithm is correct but not complete, e.g. it misses



Summary

- Association rules represent statistical rather than casual relationships.
- Association analysis is an exploratory rather than a confirmatory task.
- However, it is possible to learn causal relationships from observational data, e.g. the LCD algorithm.