Learning Causal Influence

CS F320 Foundations of Data Science

—

Under the guidance of

Basabdatta Bhattacharya

Dept of CSIS

BITS Pilani
K K Birla Goa Campus

Group Members - G3 (Report)

Harsh Gujarathi	***2020A7PS1712G***
Yash Khanna	***2020A7PS1713G***
Satyam Bansal	***2020A7PS0171G***
Vaibhav Jaiswal	***2020A7PS1379G***
Dhruv Rohira	***2020A7PS1725G***
Krishanu Shah	***2020A7PS1728G***

Index

S.No.	Title	Page No.
1	Index	2
2	Introduction	3
3	Methodology	3
4	Calculating Probabilities and Random Walk	6
5	Understanding Sample's Characteristics	7
6	GES Search (Greedy Equivalence Search)	8
7	Conclusion	9
8	Relevant Code	10
9	References	10

Introduction

Research in the technical domain involves several deductions and relationships between the variables or the attributes of the sample population on which the experiments are carried out.

Say we are studying a sample population of 1000 rats, and we have to find if there is a possible relationship between two characteristics or attributes shown by rats (take characteristics S and L). To find an underlying relation, we have to manipulate one variable and determine if it affected the values seen by the other variable.

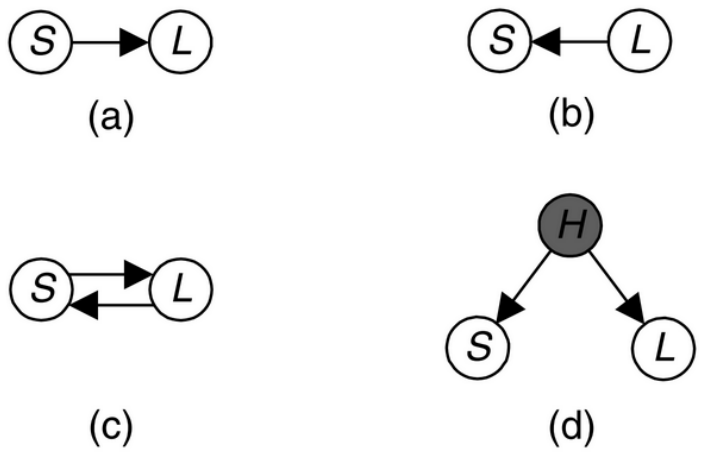

Fig 1. All possible relationships (dependencies) between the variables S and L.
Note: H here represents the hidden variable.

We first identify the population of entities we wish to consider. Our variables are features of these entities. We then sample several entities from the population. For every entity selected, we manipulate the value of X so that each possible value is given to the same number of entities.

After the value of X is set for a given entity, we measure the value of Y for that entity. The more the resultant data shows a dependency between X and Y, the more the data supports that X causally influences Y.

Our project is based on learning the relationships between variables and attributes of a sample population, how they are connected using Causal Directed Acyclic Graphs (DAGs), and what we can learn from them and deduce relationships without performing active experiments using only passive data.

Methodology

Based on the algorithm given above, we infer that After the value of X is set for a given entity; we measure the value of Y for that entity. The more the resultant data shows a dependency between X and Y, the more the data supports that X causally influences Y.

Q. What is a Causal Directed Acyclic Graph(DAG)?

A **Causal Directed Acyclic Graph(DAG)** G is a DAG with an edge from X to Y, implying that X is the cause of Y.

The manipulation of a variable can be represented by a variable M external to the system being studied. Suppose X's manipulation is seen by variable M, we say that the relation between M and X is deterministic. This means that there exists a one-to-one relationship between M and X, for specific values of M, we have unique values of X.

Considering a manipulation experiment, we will say that X causes Y, or we suspect causation between X and Y if an arrow starts from X to Y in the causal graph.

For a detailed explanation, refer to the image below,

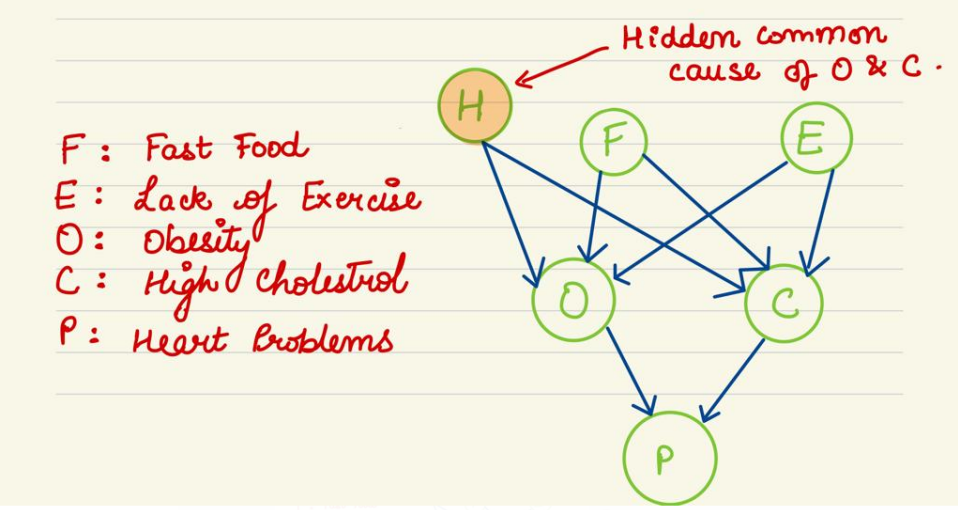

Fig 2. According to the graph, we expect that for a set containing (X, Y) where X causes Y. The set contains {(F, O),(F, C),(E, O),(E, C),(O, P),(C, P)}.

Note: We are taking H as a hidden cause here for O and C to satisfy the Markov Condition.

Besides those conditions already stated for the causal Markov assumption to hold, the following additional conditions must be satisfied for the **Causal Faithfulness Assumption** to hold:

1. There should not be any more conditions on the causal relationships..
2. We cannot draw an edge from X to Y if for every causal path from X to Y there is a causal mediary in the set of observed variables.

Q. What are Causal Networks?

A **Causal Network** is a Causal DAG that follows the causal Markov assumption, and the following conditions hold:

1. X is a direct cause of Y, there exists no variable W in the set such that if we knew the values of W, then change in X no longer changes Y.
   We will use the notation I(X, Y|Z) to denote that X is independent of Y conditional on Z, saying that if Z is given, then no matter how many times we change the probability distribution of Y, there will be no change in the probability distribution of X (X is independent of Y).
   This is also known as the Markov condition in causal Directed Acyclic Graph, giving rise to several variable independencies in the graph.

2. No causal feedback loops are present, no hidden variables are present, and all hidden variables are accounted for
   Note: in Fig 1, there are Hidden Variables in the graph. Hence, it is not a causal Network. We will be explaining a causal network below.

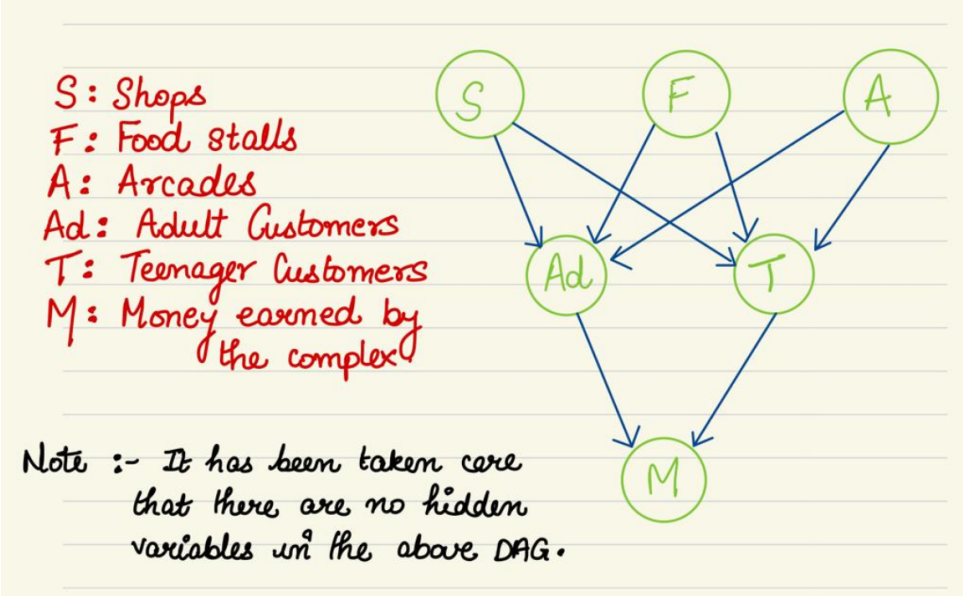

Fig 3. Shopping Complex simulation. We are taking the number of shops, food stalls, and arcades, causing changes in the number of adult and teenage customers, finally affecting the money earned by the complex.

By using the causal DAG, we deduce if there are possible causation relations between two variables present in the problem. For the example above, the dependency set includes {(S,Ad),(S,T),(F,Ad),(F,T),(A,Ad),(A,T),(Ad,M),(T,M)}.
As this represents a causal network, it will be following the Markov condition giving us the independency set {(Ad,{T, M}|{S, F, A}),(T,{Ad, M}|{S, F, A}),(M,{S, F, A}|{Ad, T})}.

(Ad,{T, M}|{S, F, A}) mean that Ad is independent of the variables in {T, M} provided values in the variable set {S, F, A}.

Causal Embedded Faithfulness Assumption

If we assume the observed probability distribution P of the variables is embedded faithfully in a causal DAG containing the variables, we say we are making the causal embedded faithfulness assumption.

Suppose we have a probability distribution P of the variables in a set V, V is a subset of W, and G is a DAG whose set of nodes is W. Then P is embedded faithfully in W if all and only the conditional independencies in P are entailed by the Markov condition applied to W and restricted to the nodes in V.

Calculating probabilities and Random Walk

As all the relations present in the causal networks are based on conditional probabilities, we are demonstrating the process how the probabilities are calculated.
We will have a node matrix that will define the probability of the variable jumping from one node to another.

In the example shown, for the 2D Array denoted by,

```
array([[0. , 0.5, 0.5],
       [0.3, 0. , 0.7],
       [0.5, 0.5, 0. ]])
```

we mean that the probability of the point in a random walk to go from node 1 to node 1, 2, and 3 is 0, 0.5, and 0.5, respectively. Similarly, we will calculate the conditional probability for the walk to node 2 and node 3. That is, P(Node 2 | Node 1) = 0.5, and so on.

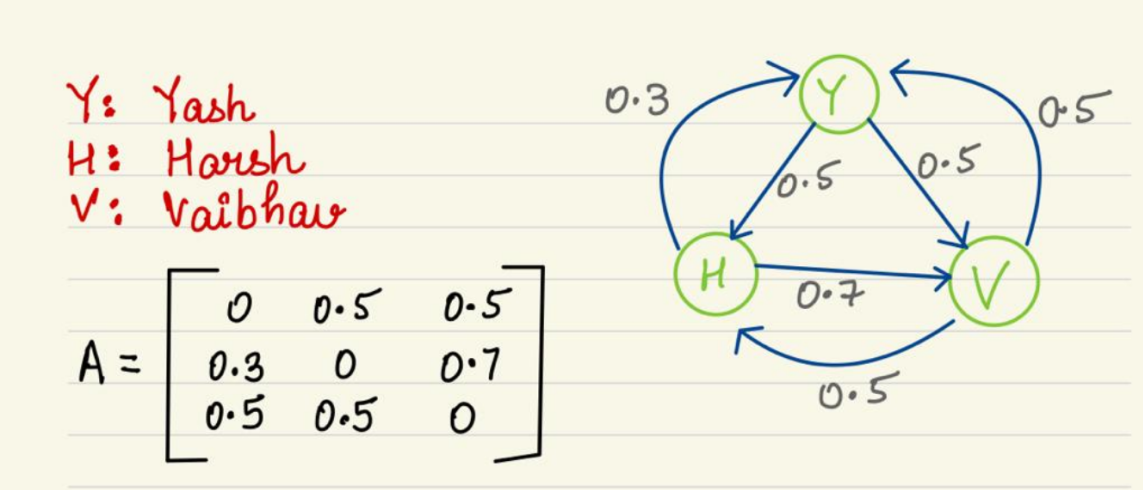

Fig 4. We define any random walk as the jump from one node to a subsequent set of nodes.

An example of random walk for 14-15 jumps,
Yash Khanna ---> Harsh Gujarathi ---> Vaibhav ---> Harsh Gujarathi ---> Vaibhav ---> Harsh Gujarathi ---> Vaibhav ---> Yash Khanna ---> Harsh Gujarathi ---> Vaibhav ---> Yash Khanna ---> Harsh Gujarathi ---> Vaibhav ---> Yash Khanna ---> Vaibhav ---> stop

Multiplying the 2D array defined by itself give us the probability of reaching the destination node after that many steps. This means that A^2 denotes the probability of reaching after 2 steps.
Result: With continuous multiplications, the probabilities reach specific values. This means that, the probabilities converge to specific points as the number of iterations tends to infinity.

```
A^n =
 [[0.28888889 0.33333333 0.37777778]
 [0.28888889 0.33333333 0.37777778]
 [0.28888889 0.33333333 0.37777778]]

π =  [0.28888889 0.33333333 0.37777778]
```

Fig 5. A^n with n tending to infinity gives us.

Understanding Sample's Characteristics

In this section, we will be explaining an example, we will be designating all the characteristics with nodes present in the graph and will be using the conditional probabilities to form the relations between the nodes to get the final Causal Network.
Taking on an example, we are considering the following characteristics as the nodes of the causal network:-

Node	Explanation
age	Age of the Person
hours-per-week	Working hours per week clocked by the person
hasGraduateDegree	Is the person graduated or not
inRelationship	Is the person in relationship or not
isWhite	Is the person white or not
isFemale	Gender of the person
greaterThan50k	Salary greater than 50000 dollars or not

We are checking about the relationship between the person's job credentials with their personal characteristics like Gender, Relation status and Race and tend to find a relation between all the nodes present with the following dataset.

	age	hours-per-week	hasGraduateDegree	inRelationship	isWhite	isFemale	greaterThan50k
0	39	40	0	0	1	0	0
1	50	13	0	1	1	0	0
2	38	40	0	0	1	0	0
3	53	40	0	1	0	0	0
5	37	40	1	0	1	1	0
...	...	...	...	...	...	...	...
32556	27	38	0	0	1	1	0
32557	40	40	0	1	1	0	1
32558	58	40	0	0	1	1	0
32559	22	20	0	0	1	0	0
32560	52	40	0	0	1	1	1

For calculating the relation between the attributes, we are calculating for each of them, the mean, standard deviation, and minimum-maximum and printing the distribution as well.

Causal Model function with arguments (Y, D, X), where Y is the variable that is used as the primary variable or indicator that will be influenced, D is treated variable (meaning we are using and fixing the value of variable in D) while it will be finding the influence of X with the Y variable. We are checking the connection between all the variables this way.

At the end of the code, we print the causal estimate of the model and print the causal network with all the connections between the variables present inside it.

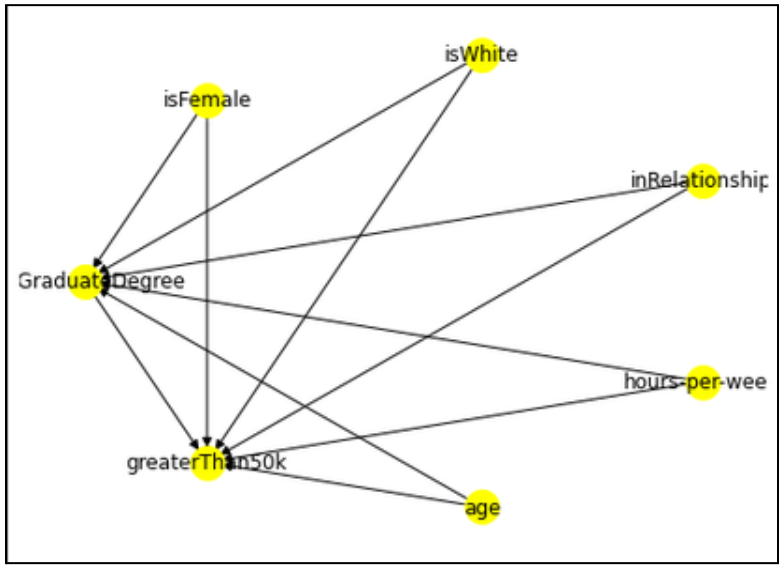

Fig 6. Graphical representation of causal relations for the dataset.

GES Search (Greedy Equivalence Search)

A propensity score is a conceptually simple statistical tool that allows researchers to make more accurate causal inferences by balancing non-equivalent groups that may result from using a non-randomized design.

Using GES tool present in the library **causallearn.search.ScoreBased.GES** and calculate the probabilities with respect to all the nodes present in the dataset and use that to calculate the causal relationships between the nodes. We are using **BIC (Bayesian Information Criterion)** score in our algorithm.

The algorithm recursively searches for the edges present in the final graph and represents the answer in the form of an adjacency matrix (for a 3-variable dataset) like

```
[[ 0  0 -1]
 [ 0  0 -1]
 [ 1  1  0]]
```

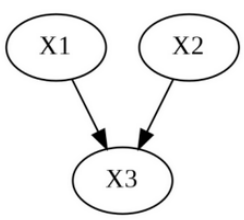

Fig 7. Here in row x and column y -1 denotes the edge going out of node x and coming in node y the opposite happens in the case of 1.

Refer to the code in the repository for further examples.

Conclusion

In any research problem, we need to find the causal relationships between several random variables/characteristics of sample populations. We need to calculate their respective probabilities and generate causal networks for the same.

In this report, we have taken the **Census Income Data Set** and have created the Causal Network for the dataset. The network denotes the causal relationship between the variables and characteristics present in the dataset and how they influence each other.
We are using the libraries like **CausalModel** from **causalinference** that help us to check the relation between the variables present in the dataset.

The score-based approach that is Greedy Equivalence Search (GES), can also be used as an alternative to calculate the causal relations. Though, this involves pre-calculation of conditional probabilities of the variables that are given as input for the GES algorithm.

The main purpose of the report is to provide a method for calculating the relations between the parameters present in the dataset given as an input.

Relevant Code

1. Random Walk Github Repository: https://github.com/Causal-Inferance/Causal-Network-and-Random-Walk
2. Causal Inference using Propensity Score: https://github.com/Causal-Inferance/Causal-Inferance-propensity-score
3. Propensity Score Dataset: https://github.com/Causal-Inferance/propensity-score-dataset
4. Causal Networks and their explanation: https://github.com/Causal-Inferance/Learning-Causal-Inferance
5. GES Search Code (Greedy Equivalence Search): https://github.com/Causal-Inferance/Causal-Inferance-GES

References:

1. Neapolitan, R.E., Jiang, X. (2006). A Tutorial on Learning Causal Influence. In: Holmes, D.E., Jain, L.C. (eds) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, vol 194. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-33486-6_2
2. For random walk code:
    a. NumPy:- https://numpy.org/doc/
    b. scipy.linalg:- https://docs.scipy.org/doc/scipy/reference/linalg.html
    c. GitHub:- https://github.com/Causal-Inferance/Causal-Network-and-Random-Walk/blob/main/Causal_Network_and_Random_Walk.ipynb
3. For understanding sample characteristic code
    a. causal inference:- https://causalinferenceinpython.org/
    b. scipy.stats:- https://docs.scipy.org/doc/scipy/reference/stats.html
    c. matplotlib.pyplot:- https://matplotlib.org/stable/tutorials/introductory/pyplot.html
    d. seaborn:- https://seaborn.pydata.org/
4. Huang, B., Zhang, K., Lin, Y., Schölkopf, B., & Glymour, C. (2018, July). Generalized score functions for causal discovery. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1551-1560).
5. Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov), 507-554.
6. GES Search (Greedy Equivalence Search):
    a. causal-learn:- https://causal-learn.readthedocs.io/en/latest/index.html
    b. causal inference:- https://causalinferenceinpython.org/
    c. matplotlib:- https://matplotlib.org/stable/index.html
7. Talebi, S. (2022) *Causal effects via propensity scores*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/propensity-score-5c29c480130c (Accessed: December 4, 2022).
8. *Welcome to causal-learn's documentation!* (no date) *causal*. Available at: https://causal-learn.readthedocs.io/en/latest/index.html (Accessed: December 4, 2022).
9. *UCI Machine Learning Repository: Census Income Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/census+income (Accessed: December 4, 2022).
10. *NetworkX documentation NetworkX*. Available at: https://networkx.org/ (Accessed: December 4, 2022).