Causal Inference with Graphical Neural Networks

Gwen Johnson Dylan Skinner Dallin Stewart
Jason Vasquez

April 12, 2024

1 Introduction

2 Graph Neural Networks

Originally proposed in 2005 by Mori et al. [1], Graph Neural Networks (GNNs) are a class of neural networks that operate on graph-structured data. They have gained significant attention in recent years due to their versatility in handling various types of graph data, including social networks, citation networks, biological networks, and more.

Unlike traditional neural networks, which operate on grid-like structures such as images or sequences, GNNs are specifically designed to capture and leverage the structural information present in graphs. One of their key strengths lies in their ability to learn meaningful representations of nodes in a graph, which can then be used for various downstream tasks such as node classification, link prediction, and graph classification. Their ability to capture and model complex relationships in graph data makes them invaluable tools for exploring and understanding real-world phenomena represented in graph form.

In our project, we are utilizing the power of GNNs to perform a supervised classification task. This means we are training a GNN on a labeled, tabular dataset, where each graph instance (or node) is associated with a target label.

3 do Operator

The **do** operator is a way to represent interventions in a causal model. It is a way to represent the effect of an intervention on a variable. As an example, consider the following model involving smoking.

If a person's fingernails (N) have turned yellow, this implies a higher probability that they are a heavy smoker (S) and hence have a higher probability of developing lung cancer (C). But, simply dyeing a persons fingernails yellow does not impact their probability of developing lung cancer.

So, in terms of **do** calculus, we can denote the process of setting a variable N to have a value *yellow* by $\mathbf{do}(N = yellow)$. We note that

$$P(C \mid N = \textit{yellow}) \neq P(C \mid \mathbf{do}(N = \textit{yellow})).$$

With this in mind, we now define the **do** operator.

Theorem 3.1 ([2]) *In a causal diagram Γ with nodes $X_1, \ldots, X_n$ and joint distribution $P(X_1, \ldots, X_n)$, the result of doing $X_i = x_i$ on the joint distribution is*

$$P(X_1, \ldots, X_n \mid \boldsymbol{do}(X_i = x_i)) = \frac{P(x_1, \ldots, x_n)}{P(x_i \mid \text{par}(x_i))} = \prod_{j \neq i} P(x_j \mid \text{par}(x_j)).$$

In this, we have $\text{par}(x_i)$ represent values of the parent nodes of $\text{PAR}(X_i)$ of X_i in Γ. The probabilities on the right hand side of the above equation are what we call *preintervention*. This means they use the original probabilities from the original model before doing $X_i = x_i$.

It is important to note that the above equation is how we calculate the probability of several events happening given one event has happened. What if we want to get the probability of a single event happening, given we do a single event? That leads to the following corollary.

Corollary 3.1.1 *If X and Y are random variables in a causal diagram Γ and $\text{PAR}(X)$ are the parents of X, then*

$$P(y \mid \boldsymbol{do}(x)) = \sum_{\text{par}} \frac{P(x,\, y,\, \text{par})}{P(x \mid \text{par})},$$

where the sum runs over all values par *that the variables $\text{PAR}(X)$ can take. If X has no parents, then*

$$P(y \mid \boldsymbol{do}(x)) = \frac{P(x,\, y)}{P(x)} = P(y \mid x).$$

Let us now consider a basic example to see how this works. Consider the following causal diagram in Figure 1.

In this diagram, we can see that A and C are both parents of B. So, for any values of x and b, Corollary 3.1.1 tells us that

$$P(X = x \,|\, \mathbf{do}(B = b)) = \sum_{\text{par}(b)} \frac{P(x,\, b,\, \text{par}(b))}{P(b \,|\, \text{par}(b))}$$

which, written out, is

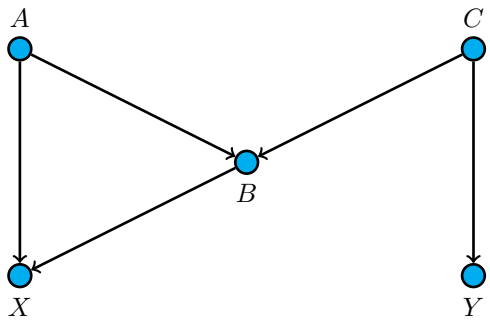

Figure 1: Basic causal diagram. Note it is in the form of a directed acyclic graph (DAG).

$$\sum_{\text{par}(b)} \frac{P(x,\, b,\, \text{par}(b))}{P(b \,|\, \text{par}(b))} = \sum_{a} \sum_{c} \frac{P(X = x\,,\, A = a\,,\, B = b\,,\, C = c)}{P(B = b \mid A = a,\, C = c)}.$$

By dependence of nodes only on their parents and the rules of probability, this turns into

$$\sum_{a} \sum_{c} \frac{P(X = x \mid A = a\,,\, B = b)P(B = b \mid A = a,\, C = c)P(A = a)P(C = c)}{P(B = b \mid A = a,\, C = c)},$$

which simplifies to

$$\sum_{a} \sum_{c} P(X = x \mid A = a\,,\, B = b)P(A = a)P(C = c).$$

Since there is only one instance where we are considering the probability with respect to c, we can simplify this to

$$\sum_{a} P(X = x \mid A = a\,,\, B = b)P(A = a),$$

which is our final answer.

While this introduction to the **do** operator might feel a bit abstract, it is the foundation of all current research in causal inference.

4 Data

For our project, we used the LUCAS0 dataset [3], which is a toy data set generated artificially by causal Bayesian networks with binary variables. The LUCAS0 dataset is a DAG with 11 nodes and 2000 training samples, where the DAG is represented as in Figure 2.

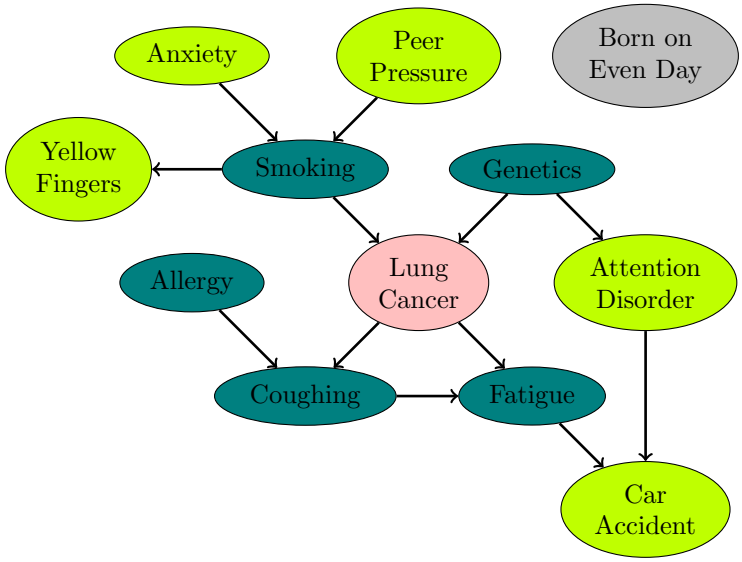

Figure 2: Basic causal diagram. Note it is in the form of a directed acyclic graph (DAG). Our target variable is shaded in pink, and the nodes in **teal** constitude the Markov blanket of the target variable.

Each node of the graph was is associated with a specific conditional probability that the creators of the dataset used to generate the data. These probabilities can be found in Table 1 in Appendix A.

5 do Calculus Result stuff (idk where this goes)

Let's consider the following conditional probabilities with **do** operators applied:

$$
\begin{aligned}
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{YF} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{PP} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{A} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{AD} = \mathrm{T})),\\
&P(\mathrm{LC} = \mathrm{T} \mid \mathbf{do}(\mathrm{CA} = \mathrm{T})),
\end{aligned}
$$

where LC is lung cancer, YF is yellow fingers, PP is peer pressure, A is anxiety, AD is attention disorder, and CA is car accident. We note that we chose this probability because it does not involve any variables that are in the Markov blanket of our target variable. Thus, we should expect that there is very little predictive power in this probability.

Plugging forward with **do** calculus, we can write $P(\text{LC} = \text{T} \mid \textbf{do}(\text{YF} = \text{T}))$ as (using corollary 3.1.1)

$$P(\text{LC} = \text{T} \mid \textbf{do}(\text{YF} = \text{T})) = \sum_{s \in \{T,F\}} \frac{P(\text{YF} = \text{T}, \text{LC} = \text{T}, \text{S} = s)}{P(\text{YF} = \text{T} \mid \text{S} = s)},$$

where S is smoking. Using the property that we can split the numerator into probabilities only involving parents (rewrite)

$$\sum_{s \in \{T,F\}} \frac{P(\text{YF} = \text{T} \mid \text{S} = s) P(\text{LC} = \text{T} \mid \text{S} = s, \text{G}) P(\text{S} = s \mid \text{A}, \text{PP})}{P(\text{YF} = \text{T} \mid \text{S} = s)}.$$

All of these fun conditionals lead us to

$$\sum_{s,a,g,p} \frac{P(\text{YF} = \text{T} \mid \text{S} = s) P(\text{LC} = \text{T} \mid \text{S} = s, \text{G} = g) P(\text{S} = s \mid \text{A} = a, \text{PP} = p)}{P(\text{YF} = \text{T} \mid \text{S} = s)},$$

where each s, a, g, p is summed over the set $\{T, F\}$.

6 Methodology

7 Experiments

7.1 Dataset Description

7.2 Experimental Setup

7.3 Results

8 Discussion

9 Conclusion

Acknowledgments

References

[1] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 729–734 vol. 2, 01 2005.

[2] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[3] Research Group for Lung Cancer Analysis. Lucas (lung cancer simple set) dataset. ETH Zurich Causality and Machine Learning Group, 2020. The LUCAS dataset contains CT scan images and annotations for lung cancer detection.

Appendix

A Creating the LUCAS0 Dataset

As described in the Data section, the LUCAS0 dataset is a toy dataset generated artificially by causal Bayesian networks with binary variables. The dataset consists of 11 nodes and 2000 training samples. The causal diagram for the LUCAS0 dataset is shown in Figure 2.

On the website, the authors mention that each node of the graph is associated with conditional probabilities which were used to generate the data. These probabilities are found in Table 1.

Conditional Probabilities	
Conditional	*Probability*
P(Anxiety=T)	0.64277
P(Peer Pressure=T)	0.32997
P(Smoking=T | Peer Pressure=F, Anxiety=F)	0.43118
P(Smoking=T | Peer Pressure=T, Anxiety=F)	0.74591
P(Smoking=T | Peer Pressure=F, Anxiety=T)	0.8686
P(Smoking=T | Peer Pressure=T, Anxiety=T)	0.91576
P(Yellow Fingers=T | Smoking=F)	0.23119
P(Yellow Fingers=T | Smoking=T)	0.95372
P(Genetics=T)	0.15953
P(Lung cancer=T | Genetics=F, Smoking=F)	0.23146
P(Lung cancer=T | Genetics=T, Smoking=F)	0.86996
P(Lung cancer=T | Genetics=F, Smoking=T)	0.83934
P(Lung cancer=T | Genetics=T, Smoking=T)	0.99351
P(Attention Disorder=T | Genetics=F)	0.28956
P(Attention Disorder=T | Genetics=T)	0.68706
P(Born an Even Day=T)	0.5
P(Allergy=T)	0.32841
P(Coughing=T | Allergy=F, Lung cancer=F)	0.1347
P(Coughing=T | Allergy=T, Lung cancer=F)	0.64592
P(Coughing=T | Allergy=F, Lung cancer=T)	0.7664
P(Coughing=T | Allergy=T, Lung cancer=T)	0.99947
P(Fatigue=T | Lung cancer=F, Coughing=F)	0.35212
P(Fatigue=T | Lung cancer=T, Coughing=F)	0.56514
P(Fatigue=T | Lung cancer=F, Coughing=T)	0.80016
P(Fatigue=T | Lung cancer=T, Coughing=T)	0.89589
P(Car Accident=T | Attention Disorder=F, Fatigue=F)	0.2274
P(Car Accident=T | Attention Disorder=T, Fatigue=F)	0.779
P(Car Accident=T | Attention Disorder=F, Fatigue=T)	0.78861
P(Car Accident=T | Attention Disorder=T, Fatigue=T)	0.97169

Table 1: Conditional probabilities for the nodes in the LUCAS0 dataset.