TECNOLÓGICO DE MONTERREY

CAMPUS GUADALAJARA

Evidencia 1. Artículo de investigación - Redes bayesianas caso discreto

Por

Paola Enríquez Reyes

A01741055

Luis Jesús Castillo Goyenechea

A01275697

Erik Ernesto Ocegueda Sambrano

A01639729

Juan Pablo Bernal Lafarga

A01742342

Análisis de métodos de razonamiento e incertidumbre

Profesor Javier Edgardo Garrido Guillén

(Grupo 101)

20 de Agosto de 2023

ABSTRACT

Las Redes Bayesianas (RB) son herramientas poderosas para modelar sistemas probabilísticos complejos. En contextos discretos, donde las variables toman valores finitos, las RB estructuran la captura de dependencias y causalidades entre variables. Esta primera parte del artículo plasma el planteamiento inicial de la situación problema empleando Redes Bayesianas. Examina estrategias de aprendizaje de estructura y parámetros desde datos, destacando exploración de modelos y regularización. Resalta su utilidad para representar y razonar relaciones probabilísticas en escenarios complejos y nos permitirá dar con una conclusión y resultados completos para el caso del comportamiento de los datos respecto al medio de transporte y sus influyentes.

INTRODUCCIÓN

Las Redes Bayesianas discretas han emergido como un enfoque altamente efectivo, en la búsqueda de herramientas para modelar y analizar sistemas en los que la incertidumbre y la dependencia entre variables. Estas redes se basan en dos conceptos básicos: la teoría de probabilidad y la teoría de grafos. La

relación de estos elementos proporciona lo necesario para representar y comprender la compleja interacción entre variables aleatorias con dominios discretos.

En una Red Bayesiana discreta, las variables se representan como nodos en un grafo dirigido acíclico, donde las aristas representan relaciones de dependencia condicional entre las variables. Cada nodo en el grafo representa una variable aleatoria, y las aristas indican la influencia que una variable tiene sobre otra. La ventaja clave de esta representación radica en su capacidad para capturar relaciones causales y condicionales de manera intuitiva y visual.

La inferencia en Redes Bayesianas discretas se basa en el Teorema de Bayes, que permite actualizar y estimar probabilidades condicionales a medida que se obtienen nuevos datos o información. La propagación de creencias se logra mediante el algoritmo de propagación de mensajes, que permite calcular eficientemente probabilidades posteriores a lo largo de las relaciones en el grafo.

El aprendizaje en Redes Bayesianas discretas se enfoca en dos aspectos principales: la estructura del grafo y los parámetros de probabilidad. El aprendizaje de la estructura implica determinar qué variables están directamente relacionadas y cómo se conectan en el grafo, a menudo utilizando algoritmos de búsqueda que consideran medidas de dependencia y verosimilitud. Por otro lado, el aprendizaje de los parámetros implica estimar las distribuciones de probabilidad condicional para cada variable dada su información contextual y observada.

La utilidad de las Redes Bayesianas discretas abarca una amplia gama de aplicaciones. En el ámbito médico, pueden utilizarse para el diagnóstico de enfermedades a partir de síntomas observados, en sistemas expertos, pueden ayudar a razonar y tomar decisiones basadas en conocimiento incierto, en la planificación y toma de decisiones, pueden proporcionar una herramienta valiosa para evaluar escenarios bajo diferentes condiciones, entre otros.

Es en este caso que dentro de este curso se realizará este proyecto en torno a la situación problema otorgada por el profesor. Comenzando por el planteamiento de una serie de preguntas importantes relacionadas con las bases de datos relacionadas mayormente a los medios de transporte.

MÉTODOS

Dentro de la metodología utilizada en esta evidencia, se buscó cumplir con puntos específicos dentro de la situación problema. Por lo tanto, se dará a conocer los procedimientos y pasos llevados a cabo para abordar el análisis de los datos y construir una Red Bayesiana que represente las relaciones de interés entre las variables.

Preparación de Datos

Se seleccionaron aleatoriamente cuatro queries del conjunto previamente definido. Estas queries sirvieron como base para la construcción de la Red Bayesiana. Los datos correspondientes a estas queries se extrajeron y prepararon para su análisis.

Creación de la Variable de Medio de Transporte

Con el objetivo de determinar el medio de transporte más utilizado por cada persona en la encuesta, se construyó una variable que representa esta información. Utilizando las respuestas a las preguntas p1a_1 a p1a_22, se realizó un proceso de consolidación y cálculo para identificar el medio de transporte predominante en cada caso.

Eficiencia y Seguridad del Transporte Público

Para complementar el análisis, se obtuvo información relevante sobre la eficiencia y seguridad del transporte público a partir de las preguntas p17_1 y p17_4. Estas respuestas se consideraron para evaluar el impacto de estas variables en las decisiones de transporte.

Propuestas de Estructuras de DAG

Se generaron al menos dos estructuras diferentes de Grafos Acíclicos Dirigidos (DAG) que representan las dependencias entre las variables de interés. Estas estructuras se basaron en el conocimiento previo y las relaciones entre las variables utilizadas.

Construcción de la Red Bayesiana

Utilizando las estructuras de DAG propuestas, se procedió a la construcción de la Red Bayesiana. Cada nodo en el grafo representa una variable y las aristas indican las relaciones de dependencia entre ellas.

Selección de la Mejor Estructura de la Red

Se evaluó la adecuación de cada estructura de la Red Bayesiana a los datos. Se analizó cuál de las propuestas se ajusta mejor a los patrones observados en la información y se argumentó la elección basada en medidas de ajuste y coherencia con el contexto e investigación.

Se revisó la pertinencia de las relaciones representadas en la DAG seleccionada. Se consideró si las conexiones entre las variables reflejan de manera precisa las interacciones y dependencias que se esperaría encontrar en la realidad.

Aplicación del Algoritmo Hill-Climbing

Se implementó el algoritmo de búsqueda "hill-climbing" para obtener la mejor estructura de la DAG que se ajuste a los datos. Se exploraron diferentes configuraciones de la Red Bayesiana con el objetivo de refinar la representación de las relaciones entre las variables. Además,

cabe recalcar que en la realización de esta evidencia hubo uso activo de herramientas computacionales de programación, siendo el lenguaje R el principal aliado.

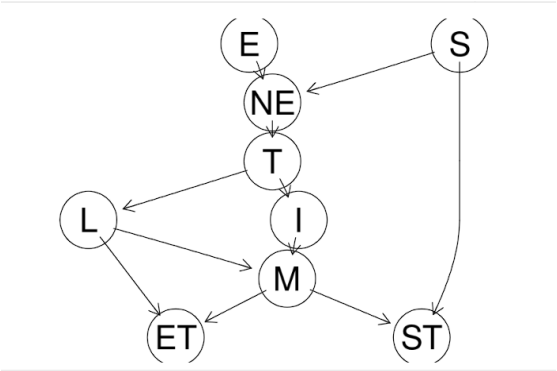
Evaluación de la Estructura Propuesta

Se analizó la estructura resultante de la aplicación del algoritmo hill-climbing. Analizando el sentido y concordancia con el conocimiento previo y las expectativas sobre cómo las variables están interconectadas en el contexto del estudio.

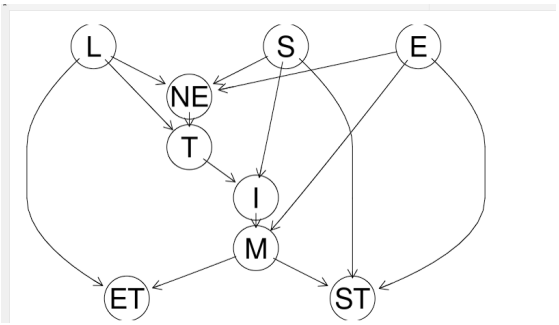
En conjunto, estos métodos permitieron explorar y analizar las relaciones entre las variables de interés, construir una Red Bayesiana que representa estas relaciones y evaluar la coherencia de la estructura propuesta con los datos y el contexto subyacente.

RESULTADOS

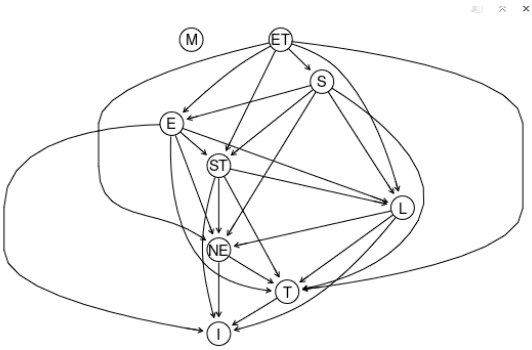
Primera dag propuesta



Segunda dag propuesta



Estructura generada por algoritmo Hill-Climbing.



¿Creen que tiene sentido esta estructura? ¿Por qué?

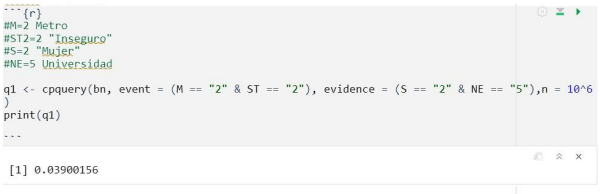
La estructura no tiene sentido, pues está relacionando todas las variables menos la variable que nos interesa que, en este caso, sería el vector de transportes. Es por eso por lo que para responder las queries utilizamos la dag2 obteniendo los siguientes resultados:

```
```{r}
score(dag1, data = data_dag, type = "aic")
score(dag2, data = data_dag, type = "aic")
```

[1] -10711.78
[1] -11228.58
```

Queries:

¿Qué probabilidad hay de que las mujeres que estudian una carrera universitaria sean violentadas usando el metro?



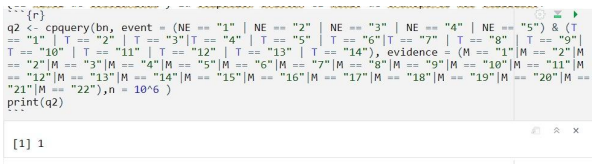
```r
#M=2 Metro
#ST2=2 "Inseguro"
#S=2 "Mujer"
#NE=5 Universidad

q1 <- cpquery(bn, event = (M == "2" & ST == "2"), evidence = (S == "2" & NE == "5"),n = 10^6
)
print(q1)
```

```
[1] 0.03900156
```

El resultado obtenido en esta query fue de 0.03900156 lo que implica que hay un 3.900156% de probabilidad de que una mujer que estudia una carrera universitaria sea violentada usando el metro

¿El nivel de escolaridad y la ocupación afectan el medio de transporte más utilizado?

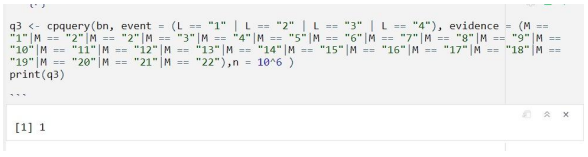


```r
q2 <- cpquery(bn, event = (NE == "1" | NE == "2" | NE == "3" | NE == "4" | NE == "5") & (T
== "1" | T == "2" | T == "3"|T == "4" | T == "5" | T == "6"|T == "7" | T == "8" | T == "9"|
T == "10" | T == "11" | T == "12" | T == "13" | T == "14"), evidence = (M == "1"|M == "2"|M
== "2"|M == "3"|M == "4"|M == "5"|M == "6"|M == "7"|M == "8"|M == "9"|M == "10"|M == "11"|M
== "12"|M == "13"|M == "14"|M == "15"|M == "16"|M == "17"|M == "18"|M == "19"|M == "20"|M ==
"21"|M == "22"),n = 10^6 )
print(q2)
```

```
[1] 1
```

El resultado obtenido en esta query fue de 1. Este valor es interpretado como un “si” debido a que en la query nos preguntan que tiene como respuesta un “si” o un “no”

¿Son más populares algunos métodos de transporte sobre otros en base al tamaño de una localidad?



```r
q3 <- cpquery(bn, event = (L == "1" | L == "2" | L == "3" | L == "4"), evidence = (M ==
"1"|M == "2"|M == "2"|M == "3"|M == "4"|M == "5"|M == "6"|M == "7"|M == "8"|M == "9"|M ==
"10"|M == "11"|M == "12"|M == "13"|M == "14"|M == "15"|M == "16"|M == "17"|M == "18"|M ==
"19"|M == "20"|M == "21"|M == "22"),n = 10^6 )
print(q3)
```

```
[1] 1
```

El resultado obtenido en esta query fue de 1. Al igual que en el caso de la query anterior nos preguntan algo que se responde con una positiva o una negativa, por

¿Cuál es la probabilidad en que una persona de sexo masculino siendo profesionista use comúnmente el colectivo ?

```r
#M=5 "Colectivo"
#S=1 "Hombre"
#T=1 "Profesionista"

q4 <- cpquery(bn, event = (M == "5"), evidence = (S == "1" & T == "1"),n = 10^6 )
print(q4)
```

```
[1] 0.1395531
```

El resultado obtenido en esta query fue de 0.1395531 lo que implica que existe un 13.95531% de probabilidad de que un hombre siendo profesionista use comúnmente el colectivo.

DISCUSIÓN

La generación de DAGs es un proceso esencial en el análisis de datos causales, y R es una herramienta ampliamente utilizada para la realización de esta evidencia. Sin embargo, como se demostró en este estudio, la calidad de la DAG generada inicialmente puede influir significativamente en la capacidad de responder a preguntas específicas. Los investigadores deben ser conscientes de que, en algunos casos, la DAG inicial generada por R puede no ser suficiente para modelar adecuadamente las relaciones causales en los datos. Por lo tanto, es esencial considerar la posibilidad de ajustar la DAG o explorar alternativas para garantizar resultados más precisos.

En este caso vimos que la DAG generada por R no fue adecuada para responder a ciertas queries específicas. Sin embargo, al crear DAGs alternativas, logramos mejorar significativamente la capacidad de nuestro modelo para proporcionar respuestas precisas y relevantes. Esto subraya la importancia de la supervisión y el ajuste manual en el proceso de generación de

DAGs, y destaca la necesidad de adaptar las herramientas y métodos disponibles para satisfacer las demandas específicas de análisis de datos causales en situaciones particulares, es necesario el que analicemos a profundidad los resultados conseguidos con ayuda de las herramientas computacionales para así notar las inconveniencias que se pueden presentar.

Bibliografía

Data Science. (2021). Redes bayesianas. *DATA SCIENCE*. https://datascience.eu/es/matematica-y-estadistica/redes-bayesianas/

José Carlos Santiesteban Rojas, Dianet Utria Pérez, Carlos Enrique Hernández Reyes. (2012). Definición de Redes Bayesianas y sus aplicaciones. Revista Vinculando. https://vinculando.org/articulos/redes-bayesianas.html

Libretexts. (2022). 13.5: Teoría de redes bayesianas. *LibreTexts Español*. https://espanol.libretexts.org/Ingenieria/Ingenier%C3%ADa_Industrial_y_de_Sistemas/Libro%3A_Din%C3%A1mica_y_Controles_de_Procesos_Qu%C3%ADmicos_(Wolf)/13%3A_Estad%C3%ADsticas_y_antecedentes_probabil%C3%ADsticos/13.05%3A_Teor%C3%ADa_de_Redes_Bayesianas

Nodo Red bayesiana. (2021, 17 agosto). IBM. Recuperado 16 de agosto de 2023, de https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-bayesian-network-node

Manuel, M. A. J. (2013). *Herramienta para la extracción de redes bayesianas predictivas a partir de bases de datos temporales*. http://tesis.ipn.mx/handle/123456789/11263