UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO PROBABILÍSTICO BAYESIANO DE INFERENCIA DE HIPERENLACES EN REDES CRIMINALES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

DIMITRI LOAIZA ARAYA

PROFESOR GUÍA:
RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:
PROFESOR 2
PROFESOR 3

Este trabajo ha sido parcialmente financiado por:
NOMBRE INSTITUCIÓN

SANTIAGO DE CHILE
2025

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA
POR: DIMITRI LOAIZA ARAYA
FECHA: 2025
PROF. GUÍA: NOMBRE PROFESOR

MODELO PROBABILÍSTICO BAYESIANO DE INFERENCIA DE HIPERENLACES EN REDES CRIMINALES

Hola buenas tardes a todos este es el abstract, espero que les guste.

Una frase de dedicatoria,
pueden ser dos líneas.

Saludos

Agradecimientos

Muchas gracias a todos.

Tabla de Contenido

Capítulo 1

Introducción

1.1. Motivación

El estudio de asociaciones o agrupaciones entre individuos es un tema de relevancia el análisis de redes sociales, con aplicaciones en diversas áreas, desde el marketing hasta la criminología. Cuando las asociaciones son de una cantidad variable de sujetos, es natural modelar el fenómeno como un hipergrafo: una generalización de los grafos donde los enlaces son de dos o más nodos.

En este contexto, un problema que ha ganado interés académico en los últimos años es la predicción de hiperenlaces; esto es, evaluar y proponer cuáles son los hiperenlaces más probables de de existir, ya sean futuros o faltantes, dados los patrones de asociación de los nodos en los hiperenlaces observados.

Nuestro interés por este problema nace desde un caso específico: la inferencia de participantes en un delito para el apoyo a la investigación penal en el contexto del proyecto Fiscal Heredia.

En la próxima sección daremos contexto sobre las aplicaciones de inteligencia artificial en la Fiscalía Nacional Chilena en el marco de Fiscal Heredia, para luego proponer un conjunto de características deseadas en un modelo de redes sociales para este contexto.

1.2. Fiscal Heredia

Explicar el origen y las diferentes aplicaciones.

Dado lo anterior, las características deseables del modelo planteado son:

1. Capacidad de aprender las preferencias de asociación de los nodos respecto a las características de la asociación (atributos de los hiperenlaces) como respecto a las características de los nodos con quienes se asocia.
2. Integración de un mecanismo de llegada de nuevos sujetos al mercado criminal a través de su participación en delitos con nodos ya pertenecientes a la red.
3. Capacidad de realizar inferencia sobre estos sujetos nuevos, los cuáles naturalmente tendrán poco historial para ajustar sus preferencias.

4. Capacidad de explicitar la incertidumbre sobre las inferencias realizadas, a manera de entregar información completa y transparente que apoye mejor el análisis criminal

1.3. Objetivos

Los objetivos para este trabajo de tesis son:

Objetivo General:

- Diseñar, desarrollar y evaluar un modelo bayesiano de inferencia de hiperenlaces que cumpla con las características deseables para su aplicación en el contexto de Fiscalía.

Objetivos Específicos:

- Proponer un diseño para el modelo basado en la literatura y los requisitos de la aplicación.
- Desarrollar el modelo en software especializado de manera que sea computacionalmente factible.
- Evaluar el rendimiento del modelo con conjuntos de datos de crímen y de marketing para explorar sus posibles usos.
- Proponer una breve metodología para su uso.

Capítulo 2

Revisión Bibliográfica

2.1. Inferencia Bayesiana

2.1.1. Definición

Como mencionan Gelman en [1], la 'inferencia Bayesiana es el proceso de ajustar un modelo de probabilidad a un conjunto de datos y resumir los resultados con una distribución de probabilidad sobre los parámetros del modelo y sobre cantidades no observadas como predicciones para nuevas observaciones'.

La primera pieza para hacer inferencia con este paradigma, es definir un modelo de probabilidad conjunta sobre los datos, y los parámetros que gobiernan la generación de los datos; esto es, definir:

$$p(\theta, y) = p(\theta)p(y|\theta)$$

Esto corresponde a definir un modelo de probabilidad generador de los datos, dados los valores de θ, y una distribución *a priori* de los parámetros de dicho modelo, la cuál luego será actualizada con la observación de los datos.

Para realizar la actualización de la distribución de θ, utilizamos la *regla de bayes*:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Donde el factor $p(y)$ corresponde a la distribución marginal de los datos en nuestro modelo. Este valor en la práctica no se utiliza pues corresponde a una constante.

2.1.2. Modelos Jerárquicos

Como se menciona en el capítulo 5 de [1], existen muchos casos donde los parámetros de nuestros modelos están relacionados por la estructura del modelo, implicando que la distribución conjunta de dichos parámetros debería reflejar dicha conexión.

Por ejemplo, en el caso del crímen, diferentes sujetos tienen diferentes

2.2. Hipergrafos y Predicción de Hiperenlaces

2.2.1. Definición de hipergrafo

Como mencionan Chen & Liu [2], los hipergrafos son una generalización de los grafos, donde los hiperenlaces, enlaces de los hipergrafos, pueden tener una cantidad arbitraria de nodos. Un hipergrafo se define como $H = \{V, E\}$ donde $V = \{n_1, n_2, ..., n_n\}$ es el conjunto de nodos y $E = \{e_1, e_2, \ldots, e_m\}$ es el conjunto de hiperenlaces, donde cada hiperenlace es un conjunto de nodos, o sea, $e_p \subseteq E \ \ \forall\, p \in 1, \ldots, m$.

Para representar matricialmente un hipergrafo, se utiliza una *matríz de incidencia* $H \in \mathbb{R}^{n \times m}$, donde cada fila representa un vector, cada columna un hiperenlace. En esta notación, el valor de $H_{j,i}$ es 1 si el nodo i pertenece al hipernlace j, y 0 si no.

Como ejemplo, a continuación mostramos una matriz de incidencia junto al hipergrafo que representa:

$$\begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \\ n_6 \end{array} \begin{array}{c} \begin{array}{ccc} e_1 & e_2 & e_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

Imagen

Para nuestro caso, tenemos además un vector de atributos para cada nodo e hiperenlace, w_i y z_j respectivamente, con $i \in 1, \ldots, n$ y $j \in 1, \ldots, m$.

2.2.2. Predicción de hiperenlaces

En [2], Chen & Liu realizan un análisis sistemático de la literatura sobre predicción de hipergrafos hasta el año 2022. En este articulo, se plantea el problema de la predicción de hiperenlaces como el aprendizaje de la función Ψ tal que para un hiperenlace potencial e se logre:

$$\Psi(e) = \begin{cases} \geq \epsilon & \text{si } e \in E \\ < \epsilon & \text{si } e \notin E \end{cases}$$

donde ϵ es un valor de corte para convertir un posible valor continuo de Ψ en valor binario. Esto puede ser visto como la detección de hiperenlaces no observados, en el caso de hipergrafos estáticos, o como la predicción de hiperenlaces futuros, en el caso de hipergrafos dinámicos.

Notamos que este planteamiento formula el problema de predicción de hiperenlaces como un problema de clasificación, una extensión natural de la predicción de enlaces en redes. Sin embargo, este planteamiento sufre del problema de clases desbalanceadas de manera extrema, pues la cantidad de hiperenlaces posibles para un hipergrafo con n nodos es 2^n [3] [4]. Para entender este problema, notemos que para un hipergrafo con 100 nodos, una cantidad

bastante modesta, tenemos 1.27×10^{30} posibles hiperenlaces.

Es por esto que para la predicción de hiperenlaces bajo esta formulación, se utiliza la técnica llamada *negative sampling*, una técnica para... (leer el paper de negative sampling para comentar un poco)

Este enfoque, aunque ampliamente utilizado en la literatura sobre predicción de hiperenlaces, es criticada por Yu et al en [5], mencionando que esta formulación asume que con un pequeño muestreo de los hiperenlaces negativos podemos generalizar la predicción sobre toda la población de hiperenlaces negativos.

Como alternativa, en [5] proponen un planteamiento donde: **dado** un hipergrafo $H = (V, E, X)$, con X la matriz de atributos de los nodos, un conjunto de nodos de consulta $Q \subset V$, un tamaño objetivo s, y una cantidad de soluciones objetivo k, **se busca obtener** hasta k hiperenlaces que pertenezcan al conjunto de hiperenlaces verdaderos positivos para nuestro hipergrafo, **sujeto a** que el conjunto de nodos de consulta estén dentro de los hiperenlaces propuestos. (este parrafo está casi igual al del paper citado, está bien?)

Este planteamiento se acerca nuestro caso de estudio más que la formulación como clasificación.

Para nuestro caso, el planteamiento a resolver es: dado un hipergrafo $H = (V, E, W, Z)$, con W la matriz de atributos de los nodos y Z la matriz de atributos de los hiperenlaces, queremos obtener los nodos pertenecientes a un nuevo hiperenlace del cuál observamos sus atributos y, opcionalmente, un subconjunto de los nodos del hiperenlace nuevo.

En términos de nuestro caso de estudio particular, queremos, en base a la información histórica de asociaciones criminales, predecir para un nuevo crímen del cuál observamos los atributos y opcionalmente un subconjunto de los sujetos que participan en el, el conjunto de sujetos completo que cometió el crimen.

Capítulo 3

Modelo Propuesto

3.1. Modelo Generativo Subyacente

Para plantear nuestro modelo, asumimos un modelo generativo subyacente, el cuál no nos interesa ajustar ni modelar en detalle, pero que es útil para el planteamiento del nuestro modelo de predicción de enlaces.

El algoritmo generativo de la hiperred se define como:

Algorithm 1 Hypergraph Generative Algorithm

Require: H_0, $p(e_{t+1}, a_{e_{t+1}}) = f(e_{t+1}, a_{e_{t+1}}|H_t, \Theta)$
  $t \leftarrow 0$
  while $t \leq T$ **do**
    Sample the next hyperedge from $p(e_{t+1}, a_{e_{t+1}})$
    $t \leftarrow t + 1$
  end while

Notación:

Los nodos de cada hiperenlace t se dividen entre nodos que ya eran parte de la red y nodos nuevos:

$e_t = e_{nuevos_t} \bigcup e_{antiguos_t}$

Los nodos de cada hiperenlace nuevo se dividen entre nodos conocidos y nodos desconocidos:

$e_{t+1} = e_{conocidos_{t+1}} \bigcup e_{desconocidos_{t+1}}$

Uniendo ambas notaciones, tenemos que un hiperenlace nuevo se descompone en:

$e_{t+1} = e_{antiguos,desconocidos_{t+1}} \bigcup e_{antiguos,conocidos_{t+1}} \bigcup e_{nuevos,conocidos_{t+1}} \bigcup e_{nuevos,desconocidos_{t+1}}$

X_S: Matriz de atributos de los nodos del conjunto S.

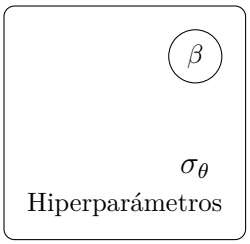
$p(e_{t+1}, a_{e_{t+1}}, X_{e_{nuevos_{t+1}}}) = f(e_{t+1}, a_{e_{t+1}} | H_t, \Theta)$

(1) $p(|e_{t+1}| = k \mid H_t, a_{e_{t+1}}, e_{conocidos_{t+1}}, \gamma)$

(2) $p(|e_{nuevos_{t+1}}| = l \mid H_t, |e_{t+1}|, a_{e_{t+1}}, e_{conocidos_{t+1}}, \theta)$

(3) $p(n_i \in e_{antiguos,desconocidos_{t+1}} \mid H_t, |e_{t+1}|, a_{e_{t+1}}, e_{conocidos_{t+1}}, \beta)$

(4) $p(X_{e_{nuevos,desconocidos_{t+1}}} \mid H_t, |e_{t+1}|, a_{e_{t+1}}, \lambda)$



3.2. Modelo Básico

Para nuestro primer acercamiento al modelo, consideraremos una cantidad fija de nodos I, y una cantidad de hiperenlaces observados J.

Definimos las siguientes variables aleatorias:

$x_{i,j}$: Participación del nodo i en el hiperenlace j.

z_j : Tamaño del hiperenlace j.

Y definimos los siguientes vectores de atributos:

$u_i^{t(j)}$: Atributos del nodo i en el momento de ocurrencia del hiperenlace j.

w_j : Atributos del hiperenlace j.

k_j : Identificador categórico del hiperenlace j.

Primero, planteamos los modelos probabilisticos para las dos variables aleatorias:

$$x_{i,j} \sim Bernoulli(logit^{-1}(u_{i,j}))$$

$$u_{i,j} = \alpha_{i,k_j} + \beta_i * w_j + \beta_{k_j} * u_i^{t(j)}$$

$$\beta_{k_j} \sim normal(0, 1)$$

$$\beta_i \sim normal(\mu, 1)$$

$$\mu \sim normal(0, 3)$$

Bibliografía

[1] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., y Rubin, D., Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2013, https://books.google.cl/books?id=ZXL6AQAAQBAJ.

[2] Chen, C. y Liu, Y.-Y., "A survey on hyperlink prediction", IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 11, p. 15034–15050, 2024, doi:10.1109/tnnls.2023.3286280.

[3] Patil, P., Sharma, G., y Murty, M. N., "Negative sampling for hyperlink prediction in networks", en Advances in Knowledge Discovery and Data Mining (Lauw, H. W., Wong, R. C.-W., Ntoulas, A., Lim, E.-P., Ng, S.-K., y Pan, S. J., eds.), (Cham), pp. 607–619, Springer International Publishing, 2020.

[4] Hwang, H., Lee, S., Park, C., y Shin, K., "Ahp: Learning to negative sample for hyperedge prediction", 2022, https://arxiv.org/abs/2204.06353.

[5] Yu, T., Lee, S. Y., Hwang, H., y Shin, K., " Prediction Is NOT Classification: On Formulation and Evaluation of Hyperedge Prediction ", en 2024 IEEE International Conference on Data Mining Workshops (ICDMW), (Los Alamitos, CA, USA), pp. 349–356, IEEE Computer Society, 2024, doi:10.1109/ICDMW65004.2024.00051.