



Introducción

El objetivo de este trabajo práctico es el de introducir algunas técnicas para el análisis de redes [1], a partir de dos casos de redes sociales. Una de estas redes está construida a partir de los lazos entre los miembros de un club de karate y la otra es la red de amigos de facebook de un usuario.

Para analizar estas redes, se van a utilizar técnicas basadas en la descomposición de matrices en autovectores y autovalores. Dichas técnicas permitirán caracterizar de distintas formas a las redes y entender su estructura.

Metodología

A continuación describimos ciertos conceptos que se van a utilizar para realizar las consignas.

Autovectores, Autovalores, Matriz de adyacencia y Matriz Laplaciana

La ecuación de autovectores y autovalores nos dice para una matriz cuadrada $\mathbf{A} \in \mathbb{R}^{n \times n}$, \mathbf{v} es autovector si se satisface:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{1}$$

para un escalar λ .

Por otro lado, un grafo finito G , definido en base a una lista de aristas que conectan nodos, puede representarse como una matriz \mathbf{A} , donde se cumple que los elementos toman el valor 1 si el nodo i está conectado con el j ($A_{ij} = 1$), y cero en caso contrario. Esta matriz se denomina matriz de adyacencia. Una propiedad es que para un grafo no dirigido¹, la matriz es simétrica.

Finalmente y importante para este TP, se denominan autovectores y autovalores de un grafo G a los que se obtienen a partir de su matriz de adyacencia \mathbf{A} . A partir de estas ideas, uno se podría preguntar qué relación tienen la estructura de la red con sus autovectores y autovalores.

Existe otra matriz que puede construirse a partir de la red cuya descomposición en autovectores permite realizar análisis sobre la estructura de la red. Esta es la matriz Laplaciana², definida como $\mathbf{L} = \mathbf{D} - \mathbf{A}$, donde \mathbf{D} es una matriz diagonal cuyos valores indican la cantidad de conexiones que tiene cada nodo.

¹no se distingue el orden de la conexión entre dos nodos

²Utilizada mucho en física y en procesamiento de señales

$$D_{ii} = \sum_{j=1}^n A_{ij} \quad (2)$$

Centralidad de autovector

En el análisis de redes, las medidas de centralidad refieren a ordenar los nodos en base a algún criterio que indique su importancia o influencia en la red. La medida más básica es la dada por la cantidad de conexiones que tiene un nodo. Esto se puede computar a partir de la matriz de adyacencia como en la ecuación 2.

Una medida de centralidad un poco mejor, es la que también considera la importancia de un nodo en base a la importancia que tienen los nodos con los que está conectado. Esto puede escribirse como una ecuación donde la importancia x_i depende de los otros nodos x_j :

$$x_i = \frac{1}{\lambda} \sum_{j=1}^n A_{ij} x_j \quad (3)$$

Este tipo de medidas de centralidad se conocen como centralidades de autovector, ya que la expresión es equivalente a la ecuación $\lambda \mathbf{x} = \mathbf{A} \mathbf{x}$. Si pedimos que la medida de centralidad tome valores positivos unicamente, puede mostrarse ³, que λ tiene que ser el mayor autovalor de la matriz de adyacencia [3]. La centralidad de Page Rank es una variante de una centralidad de autovector, donde se utiliza una matriz de adyacencia normalizada⁴ y se agrega la idea de caminante con saltos aleatorio.

Podemos decir, que realizar la descomposición en autovectores y autovalores de la matriz de adyacencia nos da cierta información sobre la estructura de la red, específicamente sobre los nodos más centrales.

Análisis de matriz Laplaciana

La descomposición en autovectores y autovalores de la matriz Laplaciana también brinda información sobre la red [4].

$$\mathbf{L} \mathbf{v} = \lambda \mathbf{v} \quad (4)$$

Se puede demostrar que para redes no dirigidas, \mathbf{L} es simétrica semidefinida positiva, por lo cual sus autovalores son iguales o mayores que cero ($\lambda \geq 0$). En particular nos interesa el autovalor más pequeño distinto de cero, el cual es llamado *conectividad algebraica*, y su correspondiente autovector (llamado vector de Fiedler). En muchos casos se puede usar el valor de la conectividad algebraica para estimar cuan bien está conectada toda la red de

³Teorema de Perron-Frobenius

⁴Denominada matriz estocástica, donde el mayor autovalor vale uno

forma global. Si vale 0 la red está partida en 2. Al mismo tiempo el vector de Fiedler permite establecer un criterio para cortar la red en dos partes. Para esto se toman los coeficientes del autovector (uno por cada nodo) y se consideran por un lado los valores positivos y por otro los negativos. Usualmente, esta forma de partir la red minimiza la cantidad de aristas que se cortan y genera dos particiones que estaban poco conectadas.

Covarianza y correlación

Dados dos vectores \mathbf{x} y \mathbf{y} definimos la covarianza como

$$Cov(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - \mu_x) \cdot (\mathbf{y} - \mu_y)}{n - 1} \quad (5)$$

Donde μ representa el valor medio del vector. Esta fórmula se puede interpretar como el producto interno de los vectores “centrados” y normalizada por la dimensión del vector menos uno.

Luego, definimos la correlación como una covarianza normalizada para que su rango sea entre -1 y 1.

$$Corr(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - \mu_x) \cdot (\mathbf{y} - \mu_y)}{\sqrt{(\mathbf{x} - \mu_x) \cdot (\mathbf{x} - \mu_x)(\mathbf{y} - \mu_y) \cdot (\mathbf{y} - \mu_y)}} \quad (6)$$

Esta expresión es equivalente al coseno del ángulo entre los vectores $\cos \theta_{\mathbf{xy}}$. Alta correlación implica vectores paralelos, con correlación positiva para vectores en la misma dirección y negativa en direcciones opuestas, y correlación nula implica vectores perpendiculares.

Si se consideran los vectores columna $\mathbf{x}_i - \mu_{x_i}$ formando la matriz \mathbf{X} , la covarianza entre todas las columnas se llama *matriz de covarianza* y se expresa como:

$$\mathbf{C} = \frac{\mathbf{X}^t \mathbf{X}}{n - 1} \quad (7)$$

Esta matriz es simétrica y semidefinida positiva por lo cual sus autovalores son no negativos.

Análisis de componentes principales

Consideremos una matriz de datos $\mathbf{X} \in \mathbb{R}^{m \times n}$ cuyas filas representan m vectores de dimensionalidad n . El análisis de componentes principales (PCA) busca encontrar un cambio de base de los vectores, de tal manera que las dimensiones se ordenen en componentes que explican los datos de mayor a menor. De forma análoga, el cambio de base busca ordenar las dimensiones según su varianza, de mayor a menor. Para hacer esto se realiza una descomposición en autovectores y autovalores de la matriz de covarianza de los datos:

$$\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^t \quad (8)$$

Donde \mathbf{V} es la matriz con los autovectores en las columnas y \mathbf{D} una matriz diagonal con los autovalores. Estos autovalores son la varianza que captura cada nueva dirección dada por los autovectores. La matriz \mathbf{V} permite transformar los datos \mathbf{X} en la nueva base mediante la operación \mathbf{XV} .

Dependiendo de la situación es posible reducir la dimensionalidad de los datos utilizando solo las k componentes principales que se deseen, es decir utilizando los primeros k vectores columna de la matriz \mathbf{V} .

Matriz de similaridad

Dado un conjunto de datos $\mathbf{X} \in \mathbb{R}^{m \times n}$ con m datos y n atributos, se define una matriz de similaridad⁵ a una matriz de $\mathbb{R}^{m \times m}$ que computa una función sobre cada par de datos ij .

$$D_{ij} = f(X_i, X_j) \quad (9)$$

Un caso sencillo de computar es utilizando el producto interno.

$$\mathbf{D} = \mathbf{X}\mathbf{X}^t \quad (10)$$

Redes Sociales

Para este trabajo práctico vamos a usar dos conjuntos de datos de redes sociales. El primero es conocida como la red del Club de Karate [5]. Es una red social de un club universitario estudiado durante 3 años, donde los 34 nodos representan miembros del club y las aristas interacciones entre los miembros fuera del club. Durante el estudio ocurrió un conflicto entre el instructor y el administrador⁶, el cual derivó en la separación del club en dos partes. Se incluye en los datos que bando tomó cada miembro. Se desea analizar la centralidad o influencia de los miembros y tratar de predecir a que grupo optó cada miembros después del conflicto a partir de la estructura de la red.

El segundo caso corresponde a una *red-ego*⁷ de Facebook[2]. Esta se construye considerando como nodos a todas las amistades de un usuario (ego) sin incluir al usuario como nodo. Luego, las aristas representan la relación de amistad entre los nodos. Los datos que vamos a utilizar también cuentan con una matriz de atributos donde hay un dato para cada nodo y una serie de atributos relacionados a categorías como trabajo, educación, ciudad natal, etc. La hipótesis que se maneja en este caso es si es posible establecer que dos nodos están conectados a partir de sus atributos compartidos. Para esto se puede computar la matriz de correlación o covarianza de los datos y establecer que dos nodos se conectan si superan un cierto valor.

⁵dependiendo el caso también se llama de disimilaridad o distancia

⁶nodos 1 y 34 respectivamente en nuestros datos

⁷ego network

Tareas

1. Método de la potencia con deflación

- 1.1 Implementar en C++ el método de la potencia con deflación para el cómputo de autovectores y autovalores. Como argumentos de entrada debe tomar un archivo de texto con la matriz, la cantidad de iteraciones y la tolerancia para la convergencia. Como salida debe entregar un archivo de texto con los autovalores y otro con los autovectores como columnas de una matriz. Si desean pueden utilizar *Eigen* para trabajar con las matrices, y *ctypes* o *pbind11* para interfacear con Python.
- 1.2 Realizar tests para verificar la implementación del método en casos donde los autovalores y autovectores sean conocidos de antemano. Probar casos con autovalores repetidos. Analizar la convergencia en casos con autovalores muy cercanos. Comparar los errores contra *Numpy*.

2. Club de Karate

- 2.1 Computar la centralidad de autovector de la red del Club de Karate utilizando su implementación. Cuáles son los nodos más centrales? Visualizar la red con el autovector y los valores del vector de forma normalizada.
- 2.2 Computar todos los autovectores de la matriz Laplaciana de la red del Club de Karate usando su implementación y comparar con *Numpy*. Analizar qué autovector predice mejor los grupos después del conflicto. Para esto medir el valor absoluto de la correlación entre cada autovector y el vector que indica el grupo que viene con los datos.

3. Ego-Facebook

- 3.1 Con los datos de Facebook, computar la matriz de similaridad con producto interno de la matriz de atributos, proponer un umbral u y construir un grafo conectando los nodos que superan ese umbral.
- 3.2 Queremos comparar la red de amistades original y la construida a partir de los datos. Para eso computaremos las siguientes métricas, 1) comparar las redes mediante la correlación de las matrices de adyacencia estiradas (*flatten*), 2) comparar utilizando la correlación entre las listas de autovalores (utilizar su implementación y validar con *Numpy*).
- 3.3 Realizar el procedimiento anterior para distintos valores de umbral y buscar el valor óptimo que genera la red de atributos más similar al grafo de amistades. Generar una figura que muestre como al variar el umbral existe un valor óptimo.
- 3.4 Filtrar la matriz de atributos utilizando PCA. Esto es, seleccionar k componentes principales y proyectar los datos al nuevo espacio para obtener una nueva matriz de atributos pero con k dimensionalidad. Repetir el punto anterior para varios valores de k . Obtener una figura donde se muestra para cada cantidad de componentes k como varía la métrica en función de u . Analizar los resultados.

En todos los casos es **obligatorio** fundamentar los experimentos planteados, proveer los archivos e información necesaria para replicarlos, presentar los resultados de forma conveniente y clara, y analizar los mismos con el nivel de detalle apropiado. En caso de ser necesario,

es posible también generar instancias artificiales con el fin de ejemplificar y mostrar un comportamiento determinado.

Forma y Fecha de entrega

- Se debe entregar un informe hecho en Latex de no más de 16 carillas. Este debe incluir una introducción al problema, la solución a las consignas mostrando pseudocódigos con una explicación de su funcionamiento. Figuras o tablas para visualizar resultados junto con sus interpretaciones. Las figuras y tablas tienen que estar referenciadas en el texto y tener una descripción. Una sección de conclusiones con un resumen y de lo mostrado y los principales resultados.
- Se debe entregar los códigos que generan los resultados, figuras, etc. Si se requieren procedimientos especiales para ejecutar el código deben ser descriptos.
- La entrega es el **Domingo 29 de Octubre 23:59h**. Para la entrega se facilitará un formulario online donde deberán adjuntar el informe y los archivos (incluir en el nombre de los archivos el nombre del grupo).
- Fecha reentregas: Viernes 1 de Diciembre hasta las 23.59 hs

Referencias

- [1] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.
- [2] Jure Leskovec and Julian McAuley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.
- [3] Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.
- [4] Ortrud R Oellermann and Allen J Schwenk. The laplacian spectrum of graphs.
- [5] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.