Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

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 permitiran 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 dirigido[1], 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 Laplaciana[2], definida como $\mathbf{L} = \mathbf{D} - \mathbf{A}$, donde D es una matriz diagonal cuyos valores indican la cantidad de conexiones que tiene cada nodo.

[1] no se distingue el orden de la conexión entre dos nodos
[2] Utilizada mucho en física y en procesamiento de señales

$$D_{ii} = \sum_{j=1}^{n} A_{ij} \tag{2}$$

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:

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

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 normalizada[4] 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].

$$\mathbf{L}\mathbf{v} = \lambda \mathbf{v} \tag{4}$$

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, permite establecer un criterio aproximado para cortar la red en dos minimizando la cantidad de aristas que se cortan.

[3] Teorema de Perron-Frobenius

[4] Denominada matriz estocástica, donde el mayor autovalor vale uno

Covarianza y correlación

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

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

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.

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

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 expresar como:

$$\mathbf{C} = \frac{\mathbf{X}^t\mathbf{X}}{n - 1} \tag{7}$$

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:

$$\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^t \tag{8}$$

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 similaridad[5] a una matriz de $\mathbb{R}^{m \times m}$ que computa una función sobre cada par de datos ij.

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

Un caso sencillo de computar es utilizando el producto interno.

$$\mathbf{D} = \mathbf{X}\mathbf{X}^t \tag{10}$$

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 administrador[6], 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.

Tareas

1. Método de la potencia con deflación

[5] dependiendo el caso también se llama de disimilaridad o distancia
[6] nodos 1 y 34 respectivamente en nuestros datos
[7] *ego network*

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.

1.2 Realizar tests para verificar la implementación del método en casos donde los autovalores y autovectores sean conocidos se antemano. También, puede resultar de utilidad contrastar con alguna librería de cálculo numérico (Numpy, Scipy).

2. Club de Karate

2.1 Computar la centralidad de autovector de la red del Club de Karate. Cuáles son los nodos más centrales? Presentar el vector de forma normalizada.

2.2 Computar todos los autovectores de la matriz Laplaciana de la red del Club de Karate. 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.

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. Proponer alguna forma para compararlas o utilizar alguna de estas opciones. Comparar las redes mediante la correlación de las matrices de adyacencia estiradas (*flatten*). Comparar utilizando la correlación entre las listas de autovalores.

3.3 Realizar el procedimiento anterior para suficientes valores de umbral y buscar el valor óptimo que genera la red de atributos más similar al grafo de amistades.

3.4 Filtrar la matriz de atributos utilizando PCA: Computar la matriz de covarianza, luego los autovectores. Seleccionar k componentes principales y proyectar los datos al nuevo espacio. Repetir el punto anterior para varios valores de k y varios valores 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.

Puntos opcionales (no obligatorios)

- Analizar la cantidad de iteraciones necesarias por el método de la potencia para lograr cierta convergencia, y su comportamiento para los distintos autovalores.
- Dada una matriz de covarianza $\mathbf{C} = \mathbf{X}^t\mathbf{X}$, encontrar una relación entre sus autovalores y sus autovectores con los de la matriz $\hat{\mathbf{C}} = \mathbf{X}\mathbf{X}^t$.

- Comparar la performance de computar las componentes $\mathbf{V}$ de PCA mediante SVD $\mathbf{X} = \mathbf{U\Sigma V}^t$ o mediante $\mathbf{X^T X}/(n-1) = \mathbf{C} = \mathbf{V}^t\mathbf{DV}$

Fecha de entrega

- Formato Electrónico: Domingo 23 de Octubre de 2022, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto `[TP2 Grupo N]` seguido de la lista de apellidos de los integrantes del grupo separados por punto y coma ;.

  `Ejemplo: [TP2 Grupo 3] Lennon; McCartney; Starr; Harrison`

- Se ruega no sobrepasar el máximo permitido de archivos adjuntos de 20MB. Tener en cuenta al realizar la entrega de no ajuntar bases de datos disponibles en la web, resultados duplicados o archivos de backup

Importante: El horario es estricto. Los correos recibidos después de la hora indicada no serán considerados.

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.