Aprendizaje Automático para Datos en Grafos

Laboratorio 3

Graciana Castro
4.808.848-2
gcastro@fing.edu.uy

Julian O'Flaherty
6.285.986-9
julian.o.flaherty@fing.edu.uy

1. Introducción

2. Grafos Erdös-Rényi

Los grafos *Erdös-Rényi*[1] (ER) son grafos aleatorios con un algoritmo de generación muy simple, donde a cada par de nodos se le asigna una arista con una probabilidad p. Pese a la simplicidad del algoritmo, los grafos ER tienen propiedades interesantes. Una de estas propiedades es que si

$$p > \frac{(1+\epsilon)ln(n)}{n} \qquad \epsilon > 0 \tag{1}$$

entonces la probabilidad de que el grafo sea conexo es practicamente 1. Analogamente, si

$$p < \frac{(1-\epsilon)ln(n)}{n} \qquad \epsilon > 0 \tag{2}$$

entonces la probabilidad de que el grafo sea conexo es practicamente 0.

En la figura 1 hacemos una validación empírica de esta propiedad, generando nueve grafos con el algoritmo de ER con 200 nodos, variando la probabilidad p alrededor del umbral de conectividad $p \approx \ln(n)/n \approx 0{,}02649$. Para cada valor de p se generan tres grafos distintos.

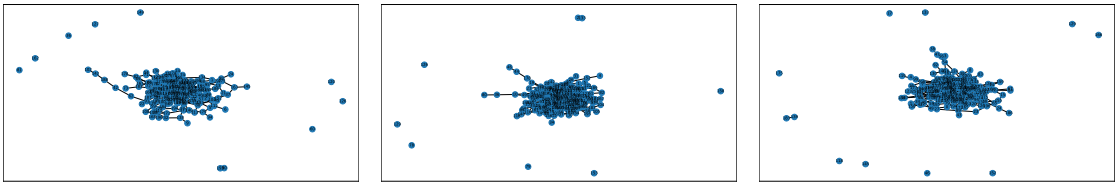
La primera observación es que para valores de p menores que el umbral de conectividad, el grafo es siempre disconexo, mientras que para valores de p mayores que el umbral de conectividad, el grafo es siempre conexo. Cabe recalcar que las condiciones de las cotas (1) y (2) son probabilísticas, por lo que puede existir una realización del grafo que no cumpla con la condición, solo que esto es altamente improbable. En el caso que p es igual al umbral de conectividad, no podemos afirmar ningún comportamiento, y se ve reflejado en que algunos de los grafos son conexos y otros no.

Otra característica a destacar, es que cuando el grafo ER no es conexo, las componentes que no pertenecen a la componente conexa más grande, son nodos aislados. Esto esta relacionado a la tendencia de los grafos ER a tener una componente gigante cuando $np > 1$, condición que cumplen las 3 probabilidades elegidas.

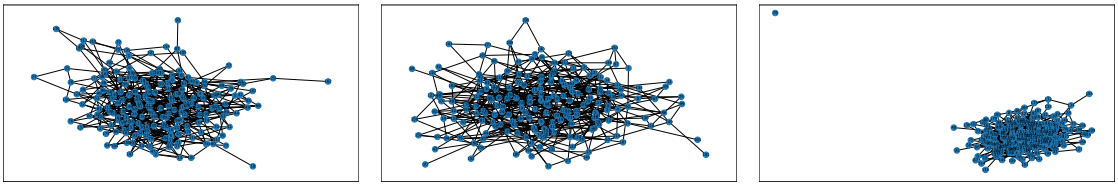
3. Grafos SBM

Los grafos *Stochastic Block Model* (SBM) son una extensión natural de los grafos ER, que busca solventar algunas de la limitaciones de los mismos. En particular, los grafos ER no son capaces de generar comunidades, que es una característica importante en muchos grafos reales. Los grafos SBM solucionan este problema trabajando con una matriz de probabilidad Q simétrica, donde Q_{ij} es la probabilidad de que un nodo del grupo i se conecte con un nodo del grupo j. La cantidad de nodos en cada grupos son parámetros del modelo n_i.

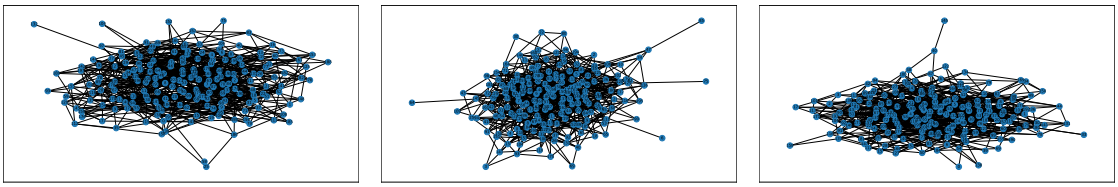
Analizemos el impacto de los vectores propios de la matriz Q en el grafo generado. Por simplicidad, trabajaremos con dos comunidades de igual tamaño $n_1 = n_2 = 50$. En la figura 2 se



(a) $n = 200$, $p = 0{,}01649$



(b) $n = 200$, $p = 0{,}02649$



(c) $n = 200$, $p = 0{,}03649$

Figura 1: Nueve realizaciones de grafos ER con $n = 200$ y tres valores de p alrededor del umbral de conectividad $p \approx \ln()/n$.

muestran tres realizaciones de un grafo SBM, con las matrices:

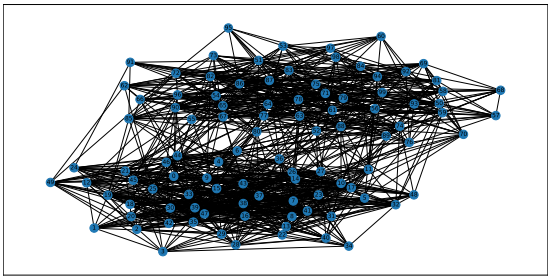
$$Q_1 = \begin{pmatrix} 0{,}4 & 0{,}05 \\ 0{,}05 & 0{,}3 \end{pmatrix} \quad Q_2 = \begin{pmatrix} 0{,}2 & 0{,}5 \\ 0{,}5 & 0{,}1 \end{pmatrix} \quad Q_3 = \begin{pmatrix} 0{,}8 & 0{,}05 \\ 0{,}05 & 0{,}8 \end{pmatrix} \tag{3}$$

Intuitivamente, los valores de la diagonal indican que tan densamente conectada esta una comunidad, mientras que los valores fuera indican que tanto se conectan entre comunidades. Viendo la figura 2b vemos el resultado de la matriz Q_2, donde el valor de intraconexión de una comunidad es similar al valor de interconexión entre comunidades. Esto deriva en un valor propio negativo, y en un grafo con una sola comunidad.

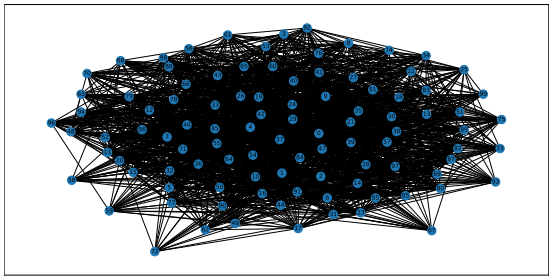
En los otros dos casos, tenemos matrices Q con valores propios positivos. Podemos notar como las comunidades de la figura 2c son muchas densas que en la figura 2a, lo cual es esperable dada la matriz Q asociada. Si observamos los valores propios, vemos que son mayores para Q_3, de lo que podemos concluir que hay una correlación entre la densidad de la comunidad y el valor propio asociado.

En resumen, los valores propios de la matriz Q son indicadores de la cantidad y densidad de las comunidades del grafo generado, donde la cantidad de valores propios positivos es igual a la cantidad de comunidades. Apliquemos este resultado a una matriz Q con $n = [45, 5, 45, 5]$:

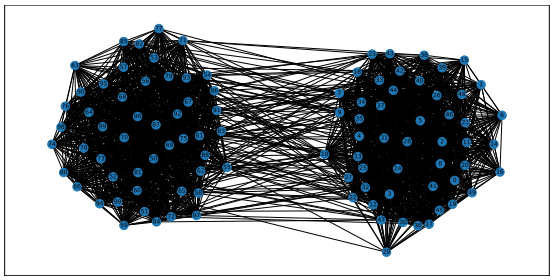
$$Q = \begin{pmatrix} 0{,}05 & 0{,}9 & 0{,}0 & 0{,}0 \\ 0{,}9 & 0{,}8 & 0{,}0 & 0{,}5 \\ 0{,}0 & 0{,}0 & 0{,}05 & 0{,}9 \\ 0{,}0 & 0{,}5 & 0{,}9 & 0{,}9 \end{pmatrix} \tag{4}$$



(a) SBM con valores propios 0,421 y 0,279, asociado a la matriz Q_1 (3).



(b) SBM con valores propios 0,652 y $-0{,}352$, asociado a la matriz Q_2 (3).



(c) SBM con valores propios 0,85 y 0,75, asociado a la matriz Q_3 (3).

Figura 2: Tres realizaciones de un grafo SBM con dos comunidades de igual tamaño, y dos valores distintos de Q.

Antes que nada, intuitivamente podemos ver que hay comunidades que no van a existir, puesto que su valor en la diagonal es menor que los valores fuera de la diagonal. Calculando los valores propios de la matriz Q, obtenemos:

$$\lambda_1 = 1{,}81, \quad \lambda_2 = 1{,}11, \quad \lambda_3 = -0{,}71, \quad \lambda_4 = -0{,}41 \tag{5}$$

Podemos ver que hay dos valores propios positivos, y dos negativos. Por lo tanto, deberíamos obtener un grafo con dos comunidades. La figura 3a muestra el resultado de la generación, donde efectivamente observamos que el grafo resultante tiene dos comunidades. En la figura 3b podemos ver la matriz de adyacencia del grafo resultante, donde se entiende la intuición que planteamos inicialmente: la conexión intracomunidad (valor de la diagonal) tiene que ser suficientemente mayor que la conexión intercomunidad (valor fuera de la diagonal) para que se forme una comunidad.

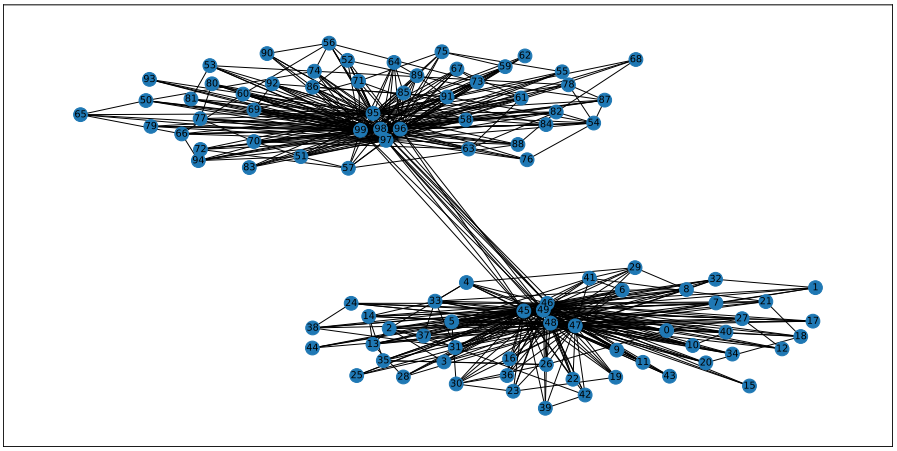
4. Grafos RDPG

4.1. definicion

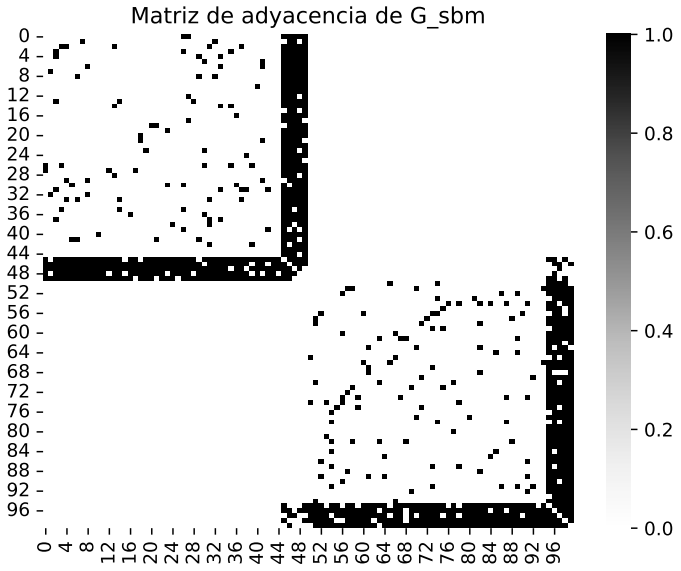
4.2. inferencia

4.3. Clustering

5. Ejemplo real



(a) Realización del grafo SBM.



(b) Matriz de adyacencia del grafo SBM.

Figura 3: SBM con valores propios (5), asociado a la matriz Q (4).

Referencias

[1] Paul Erdős and Alfréd Rényi. On random graphs. i. *Publicationes Mathematicae*, 6(3-4):290–297, 1959.