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.

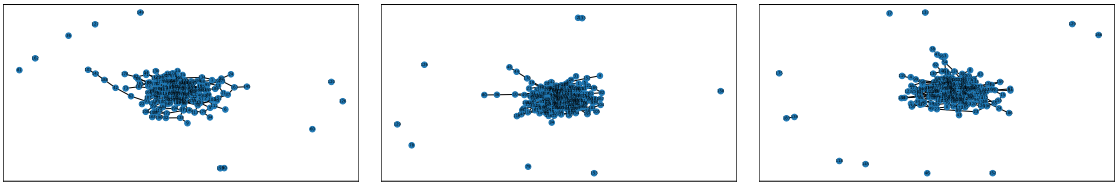
La primera observación es que para valores de p menores que el umbral de conectividad, el grafo es siempre disconexo, mientras que para valores de p mayores que el umbral de conectividad, el grafo es siempre conexo. Cabe recalcar que las condiciones de las cotas (1) y (2) son probabilísticas, por lo 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.

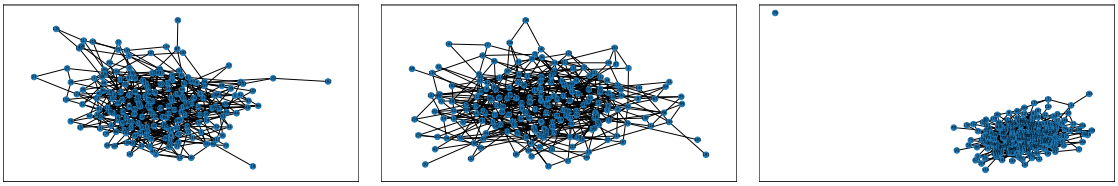
3. Grafos SBM

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.

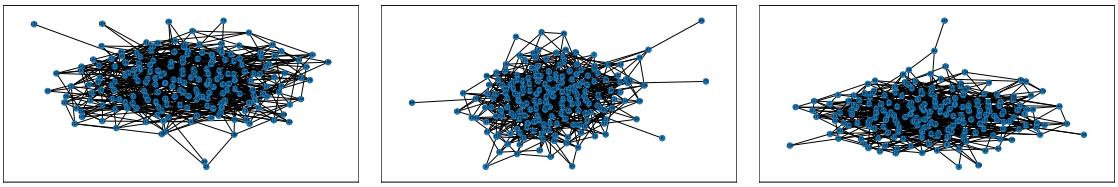
Analizemos el impacto de los vectores propios de la matriz Q en el grafo generado. Por simplicidad, trabajaremos con dos comunidades de igual tamaño $n_1 = n_2 = 50$. En la figura 2 se



(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:

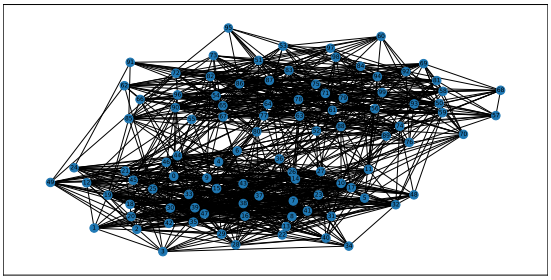
$$Q_1 = \begin{pmatrix} 0{,}4 & 0{,}05 \\ 0{,}05 & 0{,}3 \end{pmatrix} \quad Q_2 = \begin{pmatrix} 0{,}2 & 0{,}5 \\ 0{,}5 & 0{,}1 \end{pmatrix} \quad Q_3 = \begin{pmatrix} 0{,}8 & 0{,}05 \\ 0{,}05 & 0{,}8 \end{pmatrix} \tag{3}$$

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.

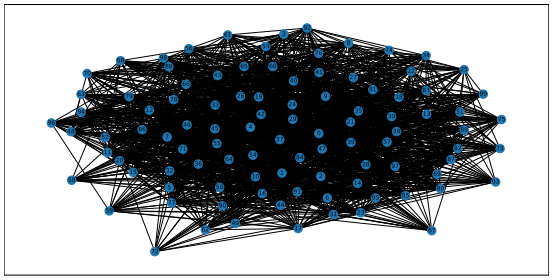
En los otros dos casos, tenemos matrices Q con valores propios positivos. Podemos notar como las comunidades de la figura 2c son muchas densas que en la figura 2a, lo cual es esperable dada la matriz Q asociada. Si observamos los valores propios, vemos que son mayores para Q_3, de lo que podemos concluir 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 $n = [45, 5, 45, 5]$:

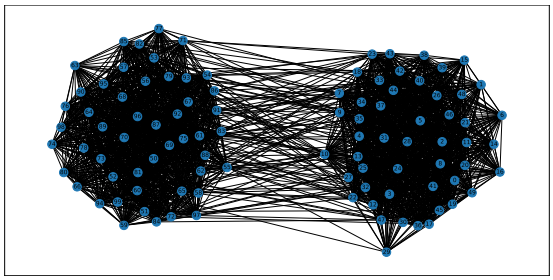
$$Q = \begin{pmatrix} 0{,}05 & 0{,}9 & 0{,}0 & 0{,}0 \\ 0{,}9 & 0{,}8 & 0{,}0 & 0{,}5 \\ 0{,}0 & 0{,}0 & 0{,}05 & 0{,}9 \\ 0{,}0 & 0{,}5 & 0{,}9 & 0{,}9 \end{pmatrix} \tag{4}$$



(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 incialmente: 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 Random Dot Product Graph (RDPG)

Los Random Dot Product Graphs (RDPG) representan una generalización tanto de los grafos Erdös-Rényi como de los Stochastic Block Models, introduciendo el concepto de *posiciones latentes*. En este modelo, cada nodo $i = 1, \ldots, n$ se asocia con un vector $\mathbf{x}_i \in \mathbb{R}^d$ en un espacio latente de dimensión d, donde la probabilidad de conexión entre dos nodos i y j está determinada por el producto interno de sus respectivas posiciones latentes: $P(A_{ij} = 1) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$.

4.1. Relación con modelos previos

Para verificar la generalidad del modelo RDPG, analizamos cómo los modelos ER y SBM pueden ser expresados como casos particulares. En el caso de un grafo Erdös-Rényi con probabilidad de conexión p, la matriz de probabilidades es $\mathbf{P} = p\mathbf{1}_{n\times n}$. Para representar esto en el modelo RDPG, consideramos $d = 1$ y tomamos $\mathbf{x}_i = \sqrt{p}$ para todos los nodos, de forma que $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = p$ para cualquier par de nodos.

Para el caso SBM, trabajamos con la matriz de probabilidades entre comunidades $\mathbf{Q} \in \mathbb{R}^{C\times C}$ y una matriz $\mathbf{Z} \in \mathbb{R}^{N\times C}$, donde $Z_{ij} = 1$ si el nodo i pertenece a la comunidad j, y $Z_{ij} = 0$ en caso contrario. La matriz de probabilidades entre nodos se expresa como $\mathbf{P} = \mathbf{ZQZ}^T$. Para representar esto como RDPG, necesitamos encontrar una descomposición $\mathbf{XX}^T = \mathbf{ZQZ}^T$.

Sin embargo, no todos los grafos SBM pueden ser representados exactamente por un RDPG. Esto se debe a que la matriz $\mathbf{P} = \mathbf{XX}^T$ debe ser semidefinida positiva por construcción (al ser una matriz de Gram), lo que impone restricciones sobre los valores propios de $\mathbf{Q}$. Cuando $\mathbf{Q}$ tiene valores propios negativos no es posible la descomposición exacta bajo el modelo RDPG estándar.

4.2. Inferencia de posiciones latentes

Dado un grafo G con matriz de adyacencia $\mathbf{A}$, el objetivo es estimar la matriz de posiciones latentes $\mathbf{X}$. Dado que $\mathbb{E}[\mathbf{A}] = \mathbf{P} = \mathbf{XX}^T$, buscamos la mejor aproximación de rango d de la matriz $\mathbf{A}$. La estimación se obtiene resolviendo:

$$\hat{\mathbf{X}} = \underset{\mathbf{X}:\text{rango}(\mathbf{XX}^T)=d}{\operatorname{argmin}} ||\mathbf{A} - \mathbf{XX}^T||_F^2 \tag{6}$$

La solución utiliza la descomposición espectral de $\mathbf{A} = \mathbf{Q\Lambda Q}^T$. Ordenando los valores propios por magnitud y seleccionando los d mayores, obtenemos $\hat{\mathbf{X}} = \hat{\mathbf{Q}}\sqrt{\hat{\mathbf{\Lambda}}}$, donde $\hat{\mathbf{Q}}$ y $\hat{\mathbf{\Lambda}}$ corresponden al "recorte" apropiado.

Para la implementación utilizamos la biblioteca *graspologic*, específicamente la clase `AdjacencySpectralEmbed`. En los experimentos con grafos ER, realizamos la gráfica de la figura 4, que muestra la varianza de las estimaciones $\mathbf{x}_i$ y de $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ para distintos valores de n y p. En ambos casos, al aumentar n, la varianza disminuye, reflejando una mayor estabilidad en las estimaciones. Sin embargo, al aumentar p, la varianza de $\mathbf{x}_i$ disminuye, mientras que la varianza de $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ se mantiene prácticamente constante. Esto sugiere que el producto $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ es más robusto frente a variaciones en p, probablemente debido al efecto de promediado en el producto matricial.

La selección de la dimensión d se realiza mediante el análisis de los valores propios de la matriz de adyacencia. Al graficar los valores propios ordenados por magnitud, buscamos un "codo" en la curva que indique una separación clara entre los valores propios señal y ruido. La biblioteca *graspologic* automatiza este proceso mediante el parámetro `n_elbows`. Se muestra en la figura 5 como el codo se hace presente en $d = 1$ tal como se determinó en el código igualando `n_elbows` a 1.

4.3. Análisis de embeddings en grafos SBM

Para grafos SBM, analizamos la interpretación geométrica de las posiciones latentes estimadas. En experimentos con grafos de dos comunidades, representamos los embeddings en coordenadas polares, donde observamos que el ángulo entre vectores $\mathbf{x}_i$ y $\mathbf{x}_j$ está relacionado con la probabilidad de conexión: ángulos pequeños corresponden a alta probabilidad (nodos de la misma comunidad), mientras que ángulos grandes indican baja probabilidad de conexión (nodos de comunidades diferentes).

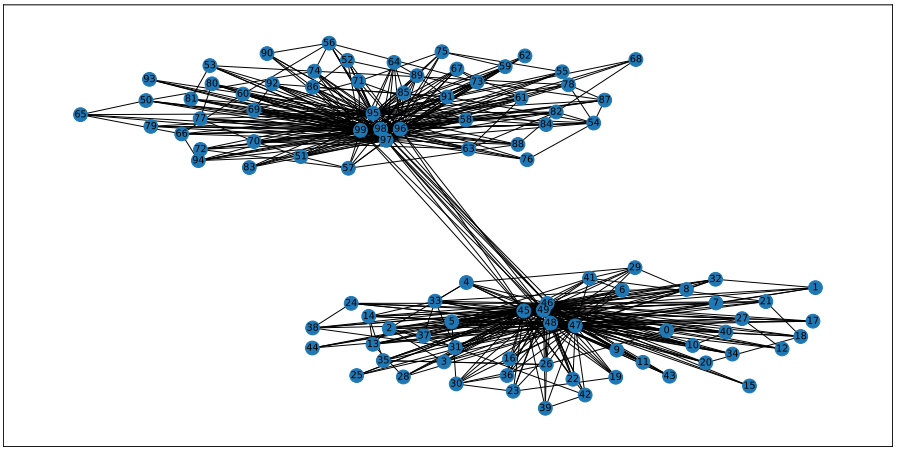
La magnitud $||\mathbf{x}_i||$ está relacionada con la probabilidad de conexión intra-comunidad, aproximadamente $\sqrt{p_c}$ donde p_c es la probabilidad de conexión dentro de la comunidad correspondiente. Esta interpretación geométrica facilita la comprensión de cómo el modelo RDPG captura la estructura de comunidades. Se observa en la figura 6 la distinción clara de dos comunidades, donde una se localiza sobre el eje horizontal y otra sobre el eje vertical. Si medimos los ángulos entre las componentes de una misma comunidad, obtenemos valores pequeños que tienden a cero, mientras que entre comunidades los ángulos son prácticamente ángulos rectos.

4.4. Detección de comunidades mediante clustering

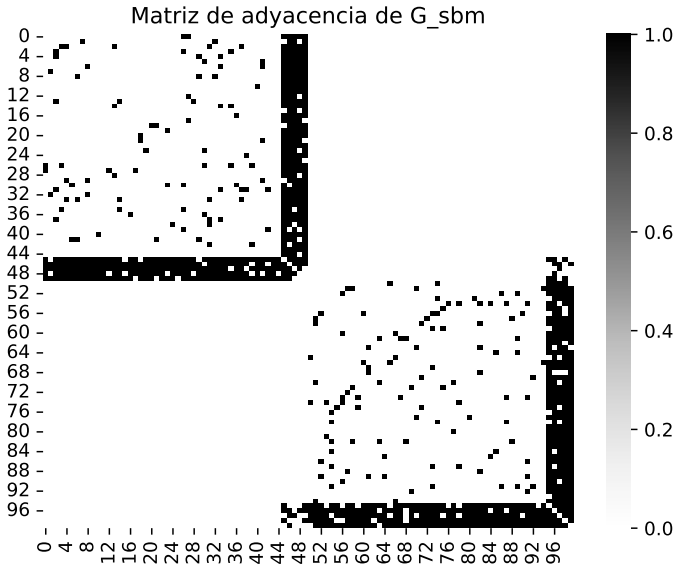
Una vez obtenidas las posiciones latentes, aplicamos algoritmos de clustering sobre los vectores $\mathbf{x}_i$ para detectar comunidades. Si bien lo clásico es utilizar k-means, los clusters observados muestran que utilizar Gaussian Mixture Models (GMM) puede ser una mejor opción.

Para la detección de comunidades a partir de los embeddings latentes, empleamos la clase `AutoGMMCluster` de la biblioteca *graspologic*, que permite ajustar automáticamente los hiperparámetros del modelo de mezclas gaussianas (GMM). En la figura 7 se ilustran los resultados obtenidos para diferentes valores de Q y n. Se observa que la tarea de identificar comunidades se vuelve más desafiante cuando la matriz Q presenta valores similares en la diagonal y la antidiagonal, ya que en estos casos la probabilidad de pertenencia a una u otra comunidad es comparable, dificultando la separación clara entre grupos.

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).

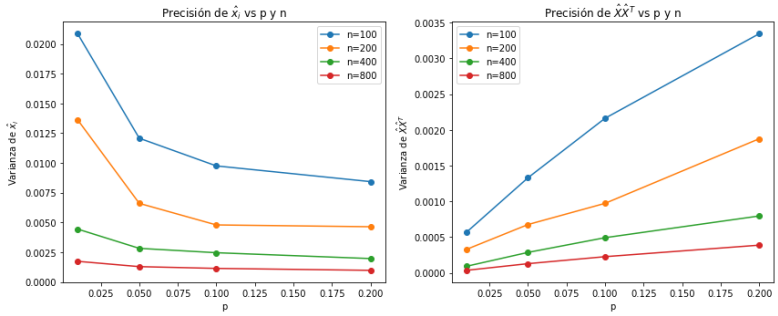


Figura 4: Varianza de las estimaciones $\mathbf{x}_i$ y $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ para distintos valores de n y p en grafos ER.

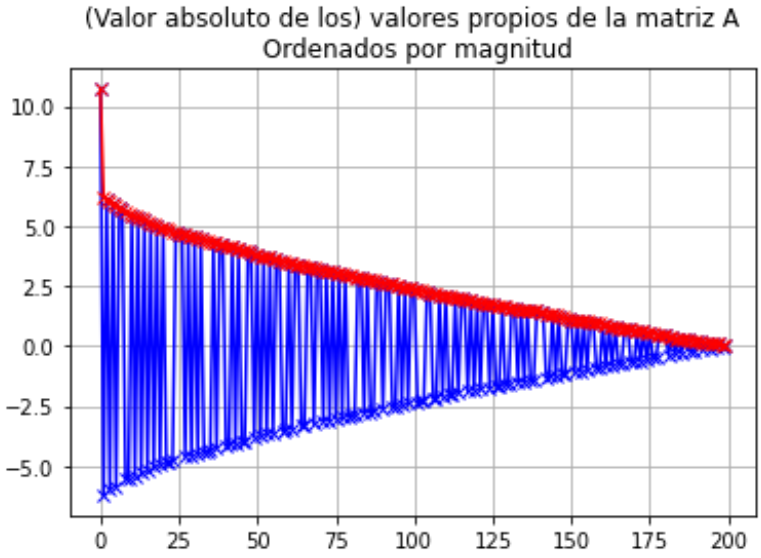


Figura 5: Valor absoluto de los valores propios de $\mathbf{A}$ ordenados por magnitud.

Figura 6: Representación en coordenadas polares de los embeddings latentes estimados para un grafo SBM con dos comunidades.

Figura 7: Resultados de la detección de comunidades mediante clustering de los embeddings latentes en grafos SBM para distintos valores de Q y n.

Referencias

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