Maestría en Ciencia de Datos del ITAM

Teoría de Grafos para Análisis de Datos

Ricardo Mansilla

10 de enero de 2016

Índice

Introducción

Con la rápida expansión de internet y aparición reciente de nuevas tecnologías que engendrán un volumen masivo de datos, tanto a nivel personal como colectivo, la industria ha reconocido la necesidad de dedicar cada vez mas recursos a la creación de tecnologías y técnicas de análisis de datos para procesar estos grandes volúmenes con baja latencia.

La aplicación de teorías y métodos de uso limitado al estudio de problemas puramente académicos en casos de aplicación real son cada vez mas frecuentes. Parte importante de estos problemas consiste en capturar en estructuras tan compactas y eficientes como sea posible, la interacción y correlación de los elementos (datos) que forman parte de nuestro sistema a estudiar. Las gráficas han demostrado ser en más de un campo de la matemática abstracta una de dichas estucturas extremadamente eficiente. Es por eso que algunos profesionales de la Ciencia de Datos han reconocido el poder que implica su uso y cada vez con más frecuencia las involucran en sus modelos. Problemas importantes del *machine learning* y de la ciencia de datos en general se basan en el uso de estructuras de grafos.

En este curso pretende establecer un camino posible (puesto que en la literatura no existe) para definir y establecer la teoría necesaria en el análisis de datos y el machine learning usando estructuras gráficas.

1. Teoría de Grafos

En muchos problemas que aparecen como casos de estudios en el mundo real, es necesario modelar la interacción entre los entes que forman parte del sistema a estudiar. Por ejemplo, algunos de estos casos podrían consistir en como se relacionan usuarios de un servicio *online* cualquiera. La manera en que se conectan los autores de artículos académicos a través de las citas mutuas que hacen en los mismos. O en un caso mas general, la proximidad bajo alguna medida de distancia que pueden tener dos puntos en algún espacio métrico. En todos ellos, tenemos una colección de entes, que llamamos vértices o nodos, y realciones entre ellos, que pueden verse como una colección de segmentos abstractos que unen a estos entes y modelan su relación, estas son llamadas aristas o flechas. El estudio de las estructuras abstractas de este tipo es lo que llamamos Teoría de Grafos.

Las relaciones entre nuestros entes pueden ser simétricas o no. Es decir, dado un conjunto V y una relación

$$R = \{(a,b) \mid a,b \in V\}$$

se dice que esta es simétrica si para todo $(a,b) \in R \Rightarrow (b,a) \in R$. En otras palabras si la relación es mutua. Por ejemplo, la relación "*padre de*" entre un conjunto de familiares no es simétrica. Puesto que si *A es padre de B*, entonces *B no es padre de A*. Sin embargo la relación "*amigo de*" obviamente lo es. La estructura de grafos que modela una relación no simétrica es una **gráfica dirigida**.

1.1. Grafos y subgrafos

1.1.1. Grafos

De manera simple y suficientemente formal podemos definir una grafo como

Definición 1.1. *Sea un conjunto de elementos V y un conjunto de pares E de elementos de V, entonces definimos una grafo como el par ordenado $G = (V,E)$.*

Dicho de otra forma, sea el conjunto V y el conjunto $E = \{x,y \mid x,y \in V\}$. Entonces definimos $G = (V,E)$.

Al conjunto V se le llama conjunto de vértices y a E el conjunto de aristas.

En general tomaremos el conjunto V como un conjunto finito de elementos a menos que se indique lo contrario.

Los elemento de cada par ordenado que forma una arista se llaman los **extremos** de la arista. Notemos entonces que cada arista está definida por sus extremos unívocamente. Y que además, no podemos tener aristas que tengan extremos fuera del conjunto V.

En la literatura se hace referencia a que la palabra "*grafo*" proviene del hecho de que la estructura anterior puede ser representada de forma gráfica dibujando un punto por cada vértice en V y uniéndolos a través de una linea por cada arista que existe en E. Es posible además encontrar referencias a los términos **gráfica** o **red**.

1.1.2. Subgrafos

Usando la terminología anterior podemos definir lo que es un subgrafo.

Definición 1.2. *Sea $G = (V, E)$, una gráfica (lo que implica que V es un conjunto de vértices y E de aristas). Si tomamos $V_s \subset V$ y $E_s \subset E$ de manera que*

$$\forall e \in E_s, e = (x, y) \Rightarrow x, y \in V_s$$

entonces $G_s = (V_s, E_s)$ es un subgrafo de G.

Dicho de otra manera, si tomamos un subconjunto V_s del conjunto de vértices y un subconjunto E_s de aristas de forma que cada arista en E_s contenga sus extremos en V_s, entonces el grafo formado por $G_s = (V_s, E_s)$ es un subgrafo de G.

Es bastante claro que el requerimiento de que los extremos de cada elemento en E_s estén contenidos en V_s puesto que esta definición necesita ser compatible con la aterior.

Por otro lado, es importante notar, que para que un grafo S sea subgrafo de G (denotado como $S \subseteq G$), es necesario que $\forall e \in S \Rightarrow e \in G$. Es decir, un subgrafo no puede tener aristas que no existan en el grafo original.

1.1.3. Matrices de adyacencia e inicidencia

Existen otras representaciones de los grafos que son bastante comunes en la literatura. Estas representaciones son menos intuitivas pero bastante mas eficiente cuando uno esta haciendo cómputo con grafos. Ambas son representaciones matriciales. La primera de ellas es conocida como la **matriz de adyacencia**. La matriz de adyacencia es básicamente una matriz cuadrada, de entradas binarias, con tantas filas como nodos. Las entradaas de esta matriz son **1**'s si los vértices correspondientes estan conectados y **0**'s en caso contrario.

Dicho de manera mas formal

Definición 1.3. *Sea $G = (V, E)$ un grafo. Entonces puesto que V es finito podemos ordenar sus elementos en una secuencia $v_1, v_2, \ldots, v_n$, lo que nos permite construir la matriz cuadrada M_G de dimensión $n \times n$. En dicha matriz tenemos que*

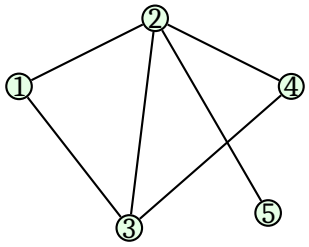
$$M_G(v_i v_j) = 1 \Leftrightarrow (v_i, v_j) \in E$$

y $M_G(v_i v_j) = 0$ en caso contrario. De esta forma tenemos una definición biunívoca de una matriz "equivalente" al grafo G.

La palabra equivalencia debe usarse con cuidado. En general lo anterior es cierto, pero hay que tener en cuenta el límite del significado de equivalencia en la teoría que se está trabajando.

Es importante notar que cada fila de esta matriz tiene tantos **1**'s como aristas inciden en el vértice, es decir, la fila i tiene tantos **1**'s como elementos de E tengan a v_i como extremo. Esta propiedad es fundamental para el desarrollo posterior de nuestra teoría.

Supongamos que tenemos el siguiente grafo



Entonces tenemos una matriz de adyacencia equivalente

$$\mathrm{M}_G = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

La matriz que se muestra arriba tiene todos los elementos de su diagonal iguales a **1**. Esto es un convenio que se establece en la teoría de grafos usual, significando que cada vértice esta unido consigo mismo. A veces sin embargo es conveniente establecer este elemento en la diagonal como **0**. Es decir, escribiríamos

$$\mathrm{M}_G = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Nótese también que estas matrices son simétricas ($M_G = M_G^t$), ya que la relación codificada en el grafo también lo es, es decir el grafo no es dirigida.

Existe otra matriz de gran importancia y es la **matriz de incidencia**. Esta matriz nos dice como se relacionan los vértices y las aristas, es decir, nos permite codificar en una estructura algebráica única las etiquetas de ambos conjuntos. Usando el mismo ejemplo de la gráfica anterior y suponiendo que tenemos sus aristas ordenadas, la matriz correspondiente sería

$$\mathrm{I}_G = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} e1 & e2 & e3 & e4 & e5 & e6 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Es importante ver que esta matriz debe tener en cada columna dos y solamente dos **1**'s puesto que cada arista tiene solo dos extremos. De nuevo, el numero de **1**'s en cada fila nos dice la cantidad de aristas que tienen a dicho vértice como extremo.

Ambas matrices son extremadamente útiles para codificar una estructura de grafo en una unidad de cómputo. La equivalencia entre cada grafo y sus matrices correspondientes es de gran utilidad pues muchos de los resultados mas conocidos y usados han sido probados a través de las matrices de adyacencia correspondientes. El área que estudia esto es conocida como Teoría Algebraica de Grafos.

1.1.4. Caminos y ciclos

Los grafos son buenos entre otras cosas para modelar la dinámica de algunos sistemas. Muchas de estas técnicas exigen la existencia de definiciones formales que permitan establecer a los nodos como "estados" y a las aristas como la posibilidad de "cambiar de estado" lo que se encarga de generar la dinámica buscada.

Definamos primero una simple función que nos será de utilidad mas adelante

Definición 1.4. *Sea $\phi_G : E \longrightarrow V$, de manera que dada $e = (v_1, v_2) \in E$, $\phi_G(e) = \{v_1, v_2\} \subset V$. Es decir la función que envía cada aristas en sus extremos. A esta la llamaremos* **aplicación de extremos** *en G.*

Definición 1.5. *Sea $G = (V, E)$ un grafo, y $\alpha = e_1 e_2 \dots e_k$ una secuencia ordenada de aristas de G tales que $\forall e_i$,*

$$\phi_G(e_i) \cap \phi_G(e_{i-1}) \neq \emptyset,$$
$$\phi_G(e_i) \cap \phi_G(e_{i+1}) \neq \emptyset$$

y

$$\phi_G(e_i) \cap \phi_G(e_{i-1}) \neq \phi_G(e_i) \cap \phi_G(e_{i+1}).$$

Esto es lo mismo que decir que en la secuencia α cada arista comparte exactamente un vértice con la próxima en la secuencia, distinto del que comparte con la anterior (si existe). A esto lo llamamos un **camino** *en G.*

La longitud de un camino es la antidad de aristas que forman parte de él, y se denota como $l(\alpha)$

Definición 1.6. *Si en la definición anterior todas las e_i's son distintas, entonces se dice que α es un* **recorrido** *o* **circuito**.

En general no haremos diferencia entre caminos y circuitos teniendo en cuenta que para fines prácticos solo nos interesan los circuitos (caminos sin repeticiones), que nosotros llamamos caminos.

Definición 1.7. *Sea $G = (V, E)$ un grafo y $\alpha = e_1 e_2 \dots e_k$ un camino en G, tomemos*

$$\{o_\alpha\} = \phi(e_1) - \phi(e_2)$$

y

$$\{f_\alpha\} = \phi(e_k) - \phi(e_{k-1})$$

*entonces se dice que el camino α "***conecta** o_α **con** l_α *" o que "***va desde** o_α **hasta** l_α *". Lo anterior podemos denotarlo como*

$$\alpha : e_1 \sim e_k$$

Nos podemos referir de muchas formas a lo anterior, pero la idea básica es que cada camino empieza en un vértice y termina en otro. Es claro que puede existir más de un camino que conecten dos vértices. Por otro lado, hay un hecho importantísimo que tiene que ver con el conjunto de dichos caminos

Definición 1.8. *Sean $G = (V, E)$ un grafo, $v_1, v_2 \in E$ y $C = \{\alpha \mid \alpha : v_1 \sim v_2\}$. Entonces existe α_0, tal que $\forall \alpha \in C$, $l(\alpha_0) \leq \alpha$. Este camino α_0 es llamado* **camino mínimo** *de v_1 a v_2.*

Más adelante veremos un algoritmo para calcular el camino mínimo, por ahora pensemos en la posibilidad de que un camino regrese en algún momento a su origen.

Definición 1.9. *Sea $G = (V, E)$ un grafo y $\alpha = e_1 e_2 \dots e_k$ un camino en G, entonces si $o_\alpha = f_\alpha$ se dice que α es un* **ciclo**.

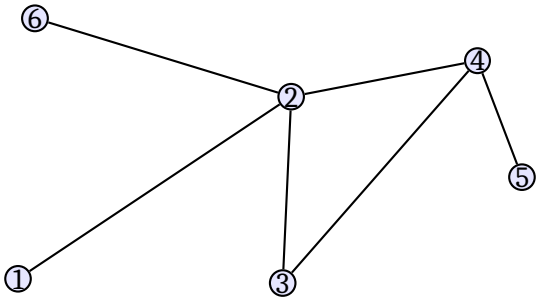
Por último daremos una definición estructural de suma importancia aprovechando la de ciclo

Definición 1.10. *Sea G un grafo. Si G no tiene ciclos entonces se dice que G es un* **árbol**.

Para enunciar algunas de las propiedades básicas de un grafo necesitamos definir un concepto de suma importancia conocido como el **grado de un vértice**.

Definición 1.11. *Sea $G = (V, E)$ un grafo y $v \in V$. Sea además $k = |\{e \in E \; v \in \phi(e)\}|$, es decir el número de aristas que tienen a v como extremo. Se dice entonces que k es el* **grado** *de v, y se denota como $\sigma(v) = k$.*

Examinemos el siguiente grafo y veámos algunas de las propiedades que hemos definido



El vértice **2** tiene grado 4, mientras que el **5** tiene solo grado 1. Es interesante notar también, por ejemplo que existen solo dos caminos que comiencen en el vértice **1** y terminen en **5**. El camino $\alpha = \{(2,3),(3,4),(4,2)\}$ es un ciclo. Por último, si quitáramos la arista $(3,4)$ del grafo tendríamos un árbol.

1.1.5. Propiedades básicas

Comencemos esta sección con algunas propiedades estructurales de los grafos.

Definición 1.12. *Sea $G = (V, E)$ un grafo, se define la* **distancia** *entre dos vértices $v_1, v_2 \in V$ como la longitud del camino mínimo(1.8) entre v_1 y v_2. Esta distancia se denota como $d(v_1, v_2)$.*

La distancia es de hecho una métrica en el grafo puesto que el camino mínimo es invariante ya que solo recorremos a cada camino en el conjunto posible en dirección contraria, la distancia de cualquier vértice a si mismo es nula a través del camino trivial y la desigualdad del triángulo es consecuencia de la minimalidad de la longitud del camino que da la distancia.

Definición 1.13. *Sea $G = (V, E)$ un grafo y $v \in V$. Se define la* **excentricidad** *de v como*

$$max_w\{d(v, w)\}$$

donde $w \in V - \{v\}$. Es decir, la distancia máxima de v al resto de los vértices. Esto se denota como $\epsilon(v)$.

Intuitivamente uno podría decir que si un vértice tiene la excentricidad mas chica está cerca del "centro" del grafo.

Definición 1.14. *La menor de las excentricidades es el* **radio** *del grafo. Esto es*

$$\rho(G) = min_v\{\epsilon(v)\}$$

con $v \in V$. Por tanto el **centro** *del grafo es el vértice donde se alcanza el* **radio**, *en otras palabras, si*

$$\epsilon(v_0) = \rho(G)$$

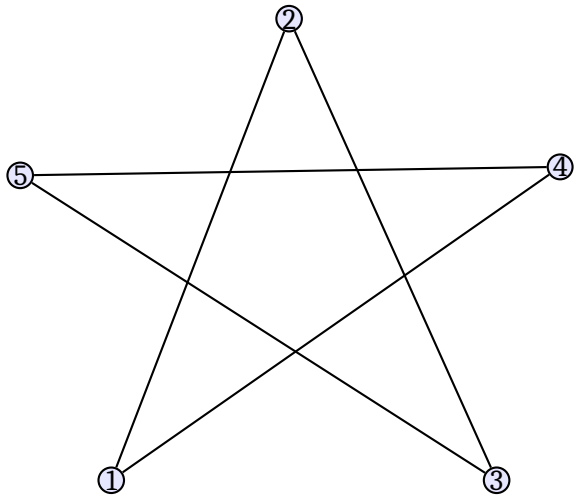
entonces a v_0 se le llama **centro**.

Definición 1.15. *Sea $G = (V, E)$ un grafo, entonces*

$$\delta(G) = max_v\{\epsilon(v)\}$$

con $v \in V$ es conocido como el **diámetro** *del grafo. La* **circunferencia** *es la longitud del ciclo más largo en G.*

Observemos el grafo siguiente



1.1.6. Aplicaciones

1.2. Conectividad

1.2.1. Componentes

1.2.2. Nodos y aristas de corte

1.2.3. Bloques

1.2.4. K-Cores

1.2.5. Otras propiedades

1.2.6. Aplicaciones

1.3. Grafos dirigidos

1.3.1. Matrices de adyacencia e inicidencia

1.3.2. Caminos dirigidos

1.3.3. Ciclos dirigidos

1.3.4. Conectividad

1.3.5. Aplicaciones

1.4. Redes y flujos

1.4.1. Redes

1.4.2. Flujos

1.4.3. Cortes

1.4.4. Aplicaciones

2. Teoría Aleatoria de Grafos

2.1. El modelo de grafos aleatorios

2.1.1. Número de links

2.1.2. Distribución de grados

2.1.3. Aplicaciones

2.2. Propiedades de los grafos aleatorios

2.2.1. Mundo pequeño

2.2.2. Clustering

2.2.3. Redes aleatorias reales

2.2.4. Aplicaciones

3. Teoría Algebraica de Grafos

3.1. Teoría espectral de grafos

3.1.1. Valores propios

3.1.2. Polinomio característico

3.1.3. Aplicaciones

3.2. Grafos regulares

3.2.1. Teoría

3.2.2. Aplicaciones

3.3. Grafos de distancia transitiva

3.3.1. Teoría

3.3.2. Aplicaciones

4. Teoría topológica de Grafos

4.1. Grafos embedidos

4.1.1. Triangulaciones

4.1.2. Simplejos

4.1.3. Aplicaciones: Ejemplo

4.2. Reconstruccion de variedades

4.2.1. Teoría

4.2.2. Aplicaciones: Ejemplo