Trabajo Especial 2:
Cadenas de Markov y modelo PageRank

FaMAF, UNC

Mayo 2016

1. Conceptos preliminares

Sea $G = (\mathcal{V}, \mathcal{E}, \mathrm{A})$ un grafo dirigido, con $\mathcal{V} = \{1, 2, \ldots, n\}$ un conjunto (contable) de vértices o nodos y $\mathcal{E}$ un conjunto de arcos: conjunto de pares ordenados $(i, j) \in \mathcal{V} \times \mathcal{V}$. Los elementos de la matriz $\mathrm{A} = [a_{ij}]$, $1 \leq i \leq n$, $1 \leq j \leq n$, son tales que $a_{ij} > 0 \iff (i, j) \in \mathcal{E}$ y 0 en caso contrario.

Si $a_{ij} > 0$ se dice que los vértices i y j son *adyacentes*, es decir, existe un arco (dirigido) entre los nodos i y j. Un camino entre los nodos i y $j = i + q$ es una secuencia c_{ij} de pares adyacentes, de la forma $c_{ij} = \{i, i+1, \ldots, i+q | (i, i+1), (i+1, i+2), \ldots, (i+q-1, i+q) \in \mathcal{E}\}$. En este caso, se dice que el camino tiene longitud q.

Sea $n = |\mathcal{V}|$ el número de nodos en G. Dado un $i \in \mathcal{V}$ se define su *grado de salida* (out-degree) como $d_i^+ = \sum_{j=1}^n a_{ij}$ y su *grado de entrada* (in-degree) como $d_i^- = \sum_{j=1}^n a_{ji}$. En general $d_i^+ \neq d_i^-$. Sin embargo, se cumple que $d := \sum_{i=1}^n d_i^+ = \sum_{i=1}^n d_i^-$. Se define el *volumen* del grafo G como $vol(G) = d$.

La *vecindad* del nodo i se define como el conjunto $N_i := N_i^+ = \{j | (i, j) \in \mathcal{E}\}$, es decir, mediante el conjunto de nodos unidos por un arco que tiene origen en el vértice i.

1.1. Caminatas aleatorias en grafos

Considérese la siguiente caminata aleatoria en G: comenzando por un vértice arbitrario, si a tiempo t un caminante se encuentra en el nodo i, el mismo efectúa una transición hacia el nodo $j \in N_i$ con probabilidad $p_{ij} = a_{ij}/d_i^+$ (independiente de t). Cada "paso" del proceso se puede considerar asociado a una variable aleatoria X_t la que toma valores en $\mathcal{V}$. De esta manera, la secuencia $X_1, X_2, \ldots, X_t, \ldots$ es una cadena de Markov definida en el espacio de estados $\mathcal{V}$. Se cumple que:

$$P(X_{t+1} = j | X_t = i, X_{t-1} = i-1, \ldots, X_1 = 1) = P(X_{t+1} = j | X_t = i) = p_{ij}$$

Sea $p_{ij}^{(t)}$ la probabilidad de transición del nodo i al nodo j en exactamente t pasos. La misma corresponde a suma de las probabilidades de todos los caminos posibles de longitud t que comienzan en el nodo i y finalizan en el nodo j. En particular $p_{ij}^{(1)} = p_{ij}$. Sigue que:

$$p_{ij}^{(t+1)} = \sum_k p_{ik} p_{kj}^{(t)}$$

Sea π_i la probabilidad *inicial* asociada al estado i-ésimo de la cadena. La probabilidad de alcanzar dicho estado al cabo de t pasos es por lo tanto

$$\pi_i^{(t)} = \sum_k \pi_k p_{ki}^{(t)}$$

Sea $\boldsymbol{\pi}^{(t)} = \left(\pi_1^{(t)} \cdots \pi_n^{(t)}\right)^T$ el vector asociado a la distribución de probabilidades de estados a tiempo t y $\boldsymbol{\pi}^{(0)}$ la distribución inicial. Para $t > 0$ se puede escribir:

$$\boldsymbol{\pi}^{(t)T} = \boldsymbol{\pi}^{(0)T} \mathrm{P}^t$$

donde $\mathrm{P}^t = [p_{ij}^{(t)}]$ y $\mathrm{P}^{t+1} = \mathrm{P}\mathrm{P}^t$. De la definición de p_{ij} sigue que $\sum_{j=1}^n p_{ij} = 1$, por lo que P es una *matriz estocástica*. Así, una caminata aleatoria en G es una cadena de Markov con estados en $\mathcal{V}$ y matriz de transición $\mathrm{P} = [p_{ij}]$. La probabilidad de un caminante de transitar del vértice i al j, para $(i, j) \in \mathcal{E}$, está dada por:

$$p_{ij} = \frac{a_{ij}}{d_i^+} = \frac{a_{ij}}{\sum_{j=1}^n a_{ij}}$$

Sea $\mathrm{D} = \mathrm{diag}\left(d_1^+, \ldots, d_n^+\right)$ la matriz diagonal de grados de salida. Se cumple que $\mathrm{P} = \mathrm{D}^{-1}\mathrm{A}$. La distribución $\boldsymbol{\Pi} = (\Pi_1 \cdots \Pi_n)^T$ se dice *estacionaria* si se cumple:

$$\boldsymbol{\Pi}^T = \boldsymbol{\Pi}^T \mathrm{P} = \left({}^{\Pi_1}\!/_{d_1^+} \cdots {}^{\Pi_n}\!/_{d_n^+}\right)^T \mathrm{A}. \tag{1}$$

1.2. * Existencia de la distribución estacionaria

Sea $\mathcal{C}$ un conjunto cualquiera de estados y $\mathcal{C}'$ su complemento, tales que $\mathcal{V} = \mathcal{C} \cup \mathcal{C}'$. Se dice que $\mathcal{C}$ es *cerrado* si no se pueden alcanzar estados en $\mathcal{C}'$ a partir de estados de $\mathcal{C}$, es decir: $p_{ij} = 0$, $\forall i \in \mathcal{C}$, $\forall j \in \mathcal{C}'$.

Se dice que el estado k tienen período $T > 1$ si $p_{kk}^{(t)} = 0$ salvo que $t = \nu T$ ($\nu = 1, 2, \ldots$), con T el mayor entero cumpliendo esta propiedad. El estado k es *aperiódico* si no existe tal T. Una cadena se denomina aperiódica si sus estados los son.

Una cadena de Markov es *irreductible* si no existen conjuntos cerrados de estados distintos al conjunto de todos los estados. Dicho de otra manera, una cadena es irreducible si y solo si cada estado puede ser alcanzado desde cualquier otro estado en un número finito de pasos.

Sea $f_{ij}^{(t)}$ la probabilidad de que el primer pasaje por el estado j, comenzando en el estado i, ocurra en el paso t-ésimo. En este caso,

$$f_{ij} = \sum_{t=1}^{\infty} f_{ij}^{(t)}$$

es la probabilidad de que, comenzando en el nodo i, el caminante alcance *alguna vez* el nodo j. Se cumple que $f_{ij} \leq 1$. Si $f_{ij} = 1$, $\{f_{ij}^{(t)}\}$ corresponde a la *distribución de primeros pasajes* para j. En particular, $\{f_{jj}^{(t)}\}$ representa la *distribución de tiempos de recurrencia* para j. Si $f_{jj} = 1$, se puede definir el *tiempo medio de recurrencia* como:

$$\mu_j = \sum_{t=1}^{\infty} t f_{jj}^{(t)} < \infty$$

El estado j se llama *persistente* si $f_{jj} = 1$ y *transitorio* si $f_{jj} < 1$. Un estado persistente y aperiódico j con $\mu_j < \infty$ se denomina *ergódico*. Una cadena ergódica es una cadena compuesta de tales estados.

Se demuestra que *una cadena de Markov irreductible posee una distribución estacionaria (que satisface* (1)*) si y solo si la misma es ergódica. En este caso* $\pi_j > 0$ $\forall j$ *y las probabilidades* $\pi_j^{(t)}$ *tienden a sus valores estacionarios independientemente de* $\pi_j^{(0)}$ (Feller, 1968). Más aún, el límite

$$\Pi_j = \lim_{t \to \infty} p_{ij}^{(t)} = \frac{1}{\mu_j}$$

existe y es independiente de la distribución inicial (Feller (1968), §XV.7). Nótese que $\mathbf{\Pi}$ en (1) corresponde en este caso a la distribución de tiempos medios de recurrencia.

1.3. Cómputo de la distribución estacionaria

La distribución estacionaria $\mathbf{\Pi}$ se puede hallar resolviendo el problema de autovalores (1), sujeto a la normalización $\mathbf{\Pi}^T\mathbf{e} = 1$, donde $\mathbf{e}$ representa el vector n-dimensional cuyos elementos son todos iguales a 1. El método más utilizado para tal fin es el *método de las potencias* el cual permite encontrar el autovector asociado al mayor autovalor de P de manera iterativa. El método consiste en calcular, partiendo de una distribución inicial arbitraria $\mathbf{\Pi}^{(0)}$, el siguiente esquema iterativo:

$$\mathbf{\Pi}^{(k+1)^T} \longleftarrow \frac{\mathbf{\Pi}^{(k)^T}\mathrm{P}}{\|\mathbf{\Pi}^{(k)}\mathrm{P}\|_1}$$

Una de las desventajas de este método es su velocidad de convergencia, por lo que en general se recurre a descomposiciones de la forma (1) a fin de aprovechar la estructura *rala* de la matriz de adyacencias A.

2. Relevancia de nodos en grafos dirigidos: modelo PageRank

El modelo PageRank propuesto por Page et al. (1999) se basa en considerar la estructura de hipervínculos de la web como una cadena de Markov con una matriz de transición $\mathrm{P} = [p_{ij}]$. Las probabilidades de transición p_{ij} asociadas al nodo i se distribuyen uniformemente entre los $j \in N_i^+$. La relevancia (o *ranking*) de una página en particular resulta de la probabilidad estacionaria asociada al estado correspondiente. Sin embargo, la estructura del problema requiere de ciertas consideraciones respecto de la construcción de P a fin de que tal distribución límite exista.

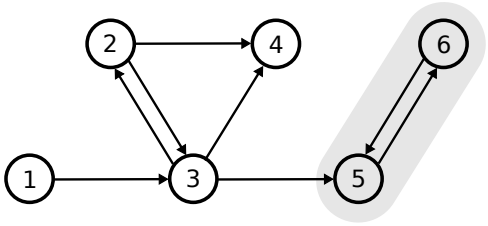
Considérese la estructura de hipervínculos de la web como un grafo dirigido, donde los nodos representan páginas y los arcos hipervínculos entre las mismas. Supóngase que, comenzando de cualquier nodo, un caminante en el nodo i tiene igual probabilidad de transitar hacia cualquier $j \in N_i^+$, es decir, $p_{ij} = 1/d_i^+ = 1/\sum_j a_{ij}$, con $a_{ij} = 1$ si $(i,j) \in \mathcal{E}$. Una particularidad en esta estructura es la existencia de nodos i tales que $N_i^+ = \emptyset$, es decir, que no poseen arcos salientes. El efecto de tales nodos se traduce en que la fila correspondiente de la matriz P contiene únicamente ceros. Al no ser una matriz estocástica, P no es una matriz de transición válida. Los nodos mencionados se denominan "colgados" (*dangling / dead end*) (p. ej., nodo 4 en la figura 1). Una forma de remediar dicha situación es reemplazar la fila correspondiente en P por el vector $\frac{1}{n}\mathbf{e}^T$. Si bien esto resuelve el problema de estocasticidad, la matriz resultante no es irreductible: existen nodos que no pueden ser alcanzados en un número finito de pasos (p. ej., nodo 1 en la figura 1). Por otro lado, existen a su vez bucles *absorbentes*[1] los cuales actúan como "sumideros" (*sinks / spider trap*) (p. ej., nodos 5–6 en la figura 1).

Page et al. (1999) propusieron la siguiente formulación para la matriz de transición:

$$\tilde{\mathrm{P}} = \alpha\mathrm{P} + (1-\alpha)\frac{\mathbf{e}\mathbf{e}^T}{n}$$

con $0 \leq \alpha \leq 1$. En este caso, la combinación convexa de P con la matriz de perturbación $\mathrm{E} = \frac{\mathbf{e}\mathbf{e}^T}{n}$ asegura que $\tilde{\mathrm{P}}$ sea irreductible por definición (Langville and Meyer, 2004). La intuición de dicho modelo es capturar el comportamiento de un "navegante aleatorio" que, con probabilidad $(1-\alpha)$, se aburre y efectúa un salto a un sitio arbitrario. Una extensión a éste modelo –conocida como

[1] Un estado se dice absorbente si forma un conjunto cerrado, en cuyo caso $p_{ii} = 1$. Un conjunto $\mathcal{C} \subseteq \mathcal{V}$ se llamará absorbente si es cerrado y existe al menos un $(i,j) \in \mathcal{E}$ tal que $i \in \mathcal{C}'$ y $j \in \mathcal{C}$.



$$P = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figura 1: Ejemplo de grafo dirigido y su matriz de transición.

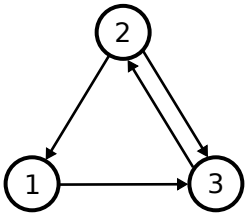
personalización– consiste en reemplazar $\frac{1}{n}\mathbf{e}^T$ por $\mathbf{v}^T$, un vector (distribución sobre estados) que refleja las preferencias de cada usuario en particular (Haveliwala, 2002).

El agregado de la matriz E equivale a crear un arco entre cada uno de los n nodos (método de máxima irreductibilidad). Una alternativa a esta aproximación consiste en agregar un nodo $j = n + 1$ al grafo original tal que $(i, j) \in \mathcal{E}$ y $(j, i) \in \mathcal{E}$, es decir, se encuentre conectado a cada uno de los n nodos y estos a él (método de mínima irreductibilidad).

3. Ejercicios

Para la realización del trabajo se dispone de las siguientes definiciones (archivo `te2_pagerank.py`):

- `randg(N)`: genera un grafo aleatorio de N nodos. Un grafo se representa como una N-tupla cuyos elementos son M_i-tuplas, $i = 1, \ldots, N$, con los índices conectados al nodo i. Por ejemplo, el grafo:



  se representará como `G=((2,), (0, 2), (1))` en el caso en que los nodos 1–3 se identifiquen con los índices 0–2.

- `g2p(g)`: cómputo de la matriz de transición a partir de un grafo `g`
- `g2p_pagerank(g, alpha)`: cómputo de la matriz de transición PageRank a partir de un grafo `g` y un factor de *dumping* `alpha`
- `power_iter_one_step(x, p)`: realiza un paso del método de las potencias utilizando la distribución `x` y matriz de transición `p`.
- Grafos `G1` y `G2` con 10 y 100 nodos, respectivamente.

Se pide:

1. Realizar una simulación que permita estimar la distribución estacionaria sobre le grafo `G1`, mediante la implementación de un caminante aleatorio. El mismo comienza en un nodo elegido al azar, y la probabilidad de que estando en un nodo i salte a un nod adyacente j es P_{ij}. Evaluar para $n = 100$ pasos. Comparar con el método de las potencias para valores crecientes de N, empleando la matriz de transición original y la modificada ($\alpha = 0{,}85$).

2. El *tiempo de cruce* para un nodo es el número de movimientos del caminante entre visitas sucesivas a dicho nodo. Mediante simulación, estime la distribución de tiempos de cruce sobre nodos en `G1` y su valor medio. Compare con los valores de PageRank para cada nodo y formule una hipótesis respecto de la misma. Valide su hipótesis en `G2`.

3. El *tiempo de cubrimiento* es la esperanza sobre el número de pasos que requiere un caminante aleatorio para visitar cada nodo al menos una vez, comenzando de un nodo elegido de forma aleatoria. Estime los tiempos de cubrimiento en `G1` y `G2` para distintos valores de α.

4. Mediante simulación (empleando la función `randg`), analice el comportamiento de los tiempos de cubrimiento al variar el número de nodos en el grafo.

5. Supongamos que se dispone de una web con N páginas (nodos). Un *spammer* agrega una nueva página `S` a la web y desea que la misma tenga el mayor *ranking* posible. Para ello elabora dos estrategias:

   a) Crear K páginas ficticias y hacer que todas ellas apunten a `S`. Después de esta adición, el número total de páginas en la web es $N + S + 1$.

   b) *Hackear* K de las N páginas existentes y agregar enlaces a `S`.

   Evalúe cual de las estrategias es la más efectiva sobre `G2`.

Referencias

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley.

Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proc. Int. World Wide Web Conference (WWW)*.

Langville, A. N. and Meyer, C. D. (2004). Deeper inside PageRank. *Internet Mathematics*, 1(3):335–400.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.