Cadenas de Markov y Perron-Frobenius

Pablo Lessa

10 de octubre de 2014

1. Cadenas de Markov

En 1996 Larry Page y Sergey Brin, en ese momento en Stanford, inventaron una manera de asignar un "ranking de importancia" a las páginas web de internet. Utilizar este método, llamado PageRank, fue uno de los elementos que permitió que su motor de búsqueda web, Google, se destacara por encima de las alternativas que existían en ese momento.

La idea detrás de PageRank es simple. Se considera un personaje ficticio que navega por internet siguiendo las siguientes reglas

1. Con un 85 % de chance elije un enlace al azar de la página en la que se encuentra y lo sigue.

2. Con un 15 % de chance elije una página web al azar de todo internet y va navega a ella directamente (esto modela el hecho de que una vez cada tanto escribe directamente una dirección web en el navegador en lugar de seguir un enlace).

El PageRank de una pagina web es la frecuencia con la cual este navegador ficticio visita a dicha página después de navegar durante muchísimo tiempo.

Vamos a formalizar este modelo en esta sección respondiendo también a las siguientes preguntas: ¿Porqué existe esa frecuencia límite (porqué no oscila con n la fracción de veces que el navegador pasa en cada página? ¿El PageRank de una página depende de la página inicial en la que empieza el navegador? ¿Cómo podría calcularse el PageRank y porqué dicho procedimiento está justificado?

1.1. Medidas de Markov

Supongamos que X es un conjunto finito cuyos elementos representan las páginas web de internet (según una búsqueda web que recién hice hoy en día se estima que hay más de mil millones de páginas web... en todo caso es un número finito). Las posibles "trayectorias" del navegante al azar son puntos del espacio métrico separable y completo $X^{\mathbb{N}}$ (el espacio de sucesiones de elementos en P con la distancia habitual suma de 2^{-n} en los n en los cuales las dos sucesiones difieren).

Para modelar el comportamiento del navegador al azar tenemos que construir una probabilidad μ en X.

El modelo es el siguiente: Fijamos una matriz $(q(x,y))_{x,y\in X}$ donde $q(x,y)$ es la probabilidad de que el navegador al azar vaya a y en un paso si se encuentra en la página web x. Según lo dicho anteriormente tenemos

$$q(x,y) = \frac{0{,}15}{|X|} + \frac{\text{número de enlaces de } x \text{ que apuntan a } y}{\text{número total de enlaces en la página web } x}.$$

Luego falta definir en qué página comienza el navegante. Vamos a suponer que comienza en una página al azar de X con distribución p. Es decir p es una probabilidad en X. El caso en que el navegante siempre comienza en cierta página x_0 se corresponde a la probabilidad p que da masa 1 a x_0 y 0 al resto de las páginas.

Entonces nuestro modelo, la probabilidad μ en $X^{\mathbb{N}}$, es la única medida tal que para todo cilindro $[x_1,\dots,x_n] \subset P^{\mathbb{N}}$ se cumple

$$\mu_{p,q}([x_1,\dots,x_n]) = p(x_1)q(x_1,x_2)q(x_2,x_3)\cdots q(x_{n-1},x_n).$$

Ejercicio 1. *Demostrar que para todo conjunto finito P, toda matriz $(q(x,y))_{x,y\in X}$ que cumple $\sum_y q(x,y) = 1$ para todo x, y toda probabilidad p en X existe una única probabilidad $\mu_{p,q}$ en $P^{\mathbb{N}}$ tal queremos*

$$\mu_{p,q}([x_1,\dots,x_n]) = p(x_1)q(x_1,x_2)q(x_2,x_3)\cdots q(x_{n-1},x_n)$$

para todo cilindro $[x_1,\dots,x_n]$.

Notemos que la medida $\mu_{p,q}$ no tiene porqué ser invariante para el shift $\sigma : X^{\mathbb{N}} \to X^{\mathbb{N}}$ definido por

$$\sigma(x_1,x_2,\dots) = (x_2,x_3,\dots).$$

El resultado más básico sobre este tipo de medidas es completamente elemental, dependiendo solamente de álgebra lineal en $\mathbb{R}^n$. Sin embargo es un resultado muy importante con una gran variedad de generalizaciones y aplicaciones en diferentes contextos.

Ejercicio 2 (Teorema de Perron-Frobenius)**.**

1. *Demostrar que toda matriz de $n \times n$ con entradas no negativas A existe un vector v con entradas no negativas que es vector propio de A.*
2. *Si además A cumple la propiedad de que la suma de las coordenadas de Av es igual a la de v para todo $v \in \mathbb{R}^n$, deducir que existe un vector propio con entradas no negativas y tal que su suma de coordenadas es* 1*.*
3. *Si además todas las coordenadas de A son estrictamente positivas demostrar que el v de la parte anterior es único.*
4. *Demostrar que para todo conjunto finito X y toda matriz de transición $(q(x,y))_{x,y\in X}$ satisfaciendo la condición del ejercicio anterior, existe una probabilidad p en X tal que $\mu_{p,q}$ es shift invariante en $X^{\mathbb{N}}$.*

5. *En las condiciones de la parte anterior, si además existe n tal que para todo $x, y \in X$ existen $x_1, \ldots, x_{n-1}$ tales que $q(x, x_1)q(x_2, x_3) \cdots q(x_{n-1}, y) > 0$, deducir que p es única y que la medida $\mu_{p,q}$ es ergódica para el shift.*

En las condiciones del último ítem del ejercicio anterior se dice que la cadena de Markov (o medida de Markov) es irreducible y aperiódica. Irreducible porque empezando en cualquier x hay probabilidad positiva de llegar a cualquier otra página y después de algunos pasos. Aperiódica porque el mayor divisor común de los tiempos en los cuales es posible llegar de x a y es 1. Notemos que el modelo de interés para PageRank es una cadena irreducible y aperiódica.

Ejercicio 3. *Si $X = \{x, y\}$ y $q(x, x) = q(y, y) = 0, q(x, y) = q(y, x) = 1$, ¿Cuales son las probabilidades p en X que hacen que $\mu_{p,q}$ sea shift invariante? ¿Cuáles hacen que $\mu_{p,q}$ sea ergódica? Repetir el ejercicio con $X = \{0, 1, 2, 3\}$ donde $q(a, b) = 1/2$ sí y sólo si $a - b = \pm 1 (mod\ ) 4$ y 0 en caso contrario (esto modela caminar en un cuadrado eligiendo ir horario o antihorario en cada paso con probabilidad $1/2$).*

1.2. ¿Porqué existe el PageRank? y ¿Cómo se calcula?

Hemos fijado nuestro modelo para el navegador al azar, i.e. la probabilidad $\mu_{p,q}$ en $X^{\mathbb{N}}$. El PageRank de una página x se define como

$$\mathrm{PR}(x) = \lim_{n \to +\infty} \frac{1_x(x_1) + \cdots + 1_x(x_n)}{n}$$

donde $(x_1, \ldots, x_n, \ldots)$ es una sucesión típica para la probabilidad $\mu_{p,q}$.

Es decir se podría aproximar el valor de PR numéricamente usando una computadora para simular una trayectoria muy larga del navegador al azar y calculando la fracción de tiempo que pasó en cada página. Esto parece un procedimiento muy ineficiente (dado que hay más de mil millones de páginas).

El procedimiento basado en simulación está justificado por el siguiente resultado.

Corolario 1 (Existencia de PageRank)**.** *En el modelo fijado para el navegador al azar, existe un conjunto $A \subset X^{\mathbb{N}}$ con $\mu_{p,q}(A) = 1$ probabilidad total tal que el límite*

$$PR(x) = \lim_{n \to +\infty} \frac{1_x(x_1) + \cdots + 1_x(x_n)}{n}$$

existe y toma el mismo valor para toda trayectoria $(x_1, x_2, \ldots)$

Ahora enunciaremos el resultado general del cual lo anterior es un corolario.

Teorema 1. *Sea X un conjunto finito y q una matriz de transición que determina una cadena irreducible y aperiódica. Sea p_0 la única probabilidad en X tal que $\mu_{p_0,q}$ es shift invariante. Entonces para toda probabilidad p en X se cumple:*

$$\mu_{p,q}\left(\left\{(x_1, \ldots) \in X^{\mathbb{N}} : \lim_{n \to +\infty} \frac{1_x(x_1) + \cdots + 1_x(x_n)}{n} = p_0(x)\right\}\right) = 1,$$

para todo $x \in X$.

Demostración. Empecemos poniéndole el nombre A al conjunto de trayectorias en $X^{\mathbb{N}}$ que queremos mostrar que tiene medida 1 para toda probabilidad $\mu_{p,q}$, es decir

$$A = \left\{ (x_1, \ldots) \in X^{\mathbb{N}} : \lim_{n \to +\infty} \frac{1_x(x_1) + \cdots + 1_x(x_n)}{n} = \mathbb{E}_{\mu_{p_0,q}}(1_x) \right\}.$$

Como $\mu_{p_0,q}$ es ergódica para el shift se tiene $\mu_{p_0,q}(A) = 1$.

Dada una probabilidad cualquiera p en X mostraremos que $\mu_{p,q}$ es absolutamente continua respecto a $\mu_{p_0,q}$ (i.e. todo conjunto de medida nula para $\mu_{p_0,q}$ tiene medida nula para $\mu_{p,q}$). De esto se deduce el resultado inmediatamente.

Para mostrar que $\mu_{p,q}$ es absolutamente continua respecto a $\mu_{p_0,q}$ notemos que, como q es irreducible y aperiódica, existe n tal que para todo $x, y \in X$ tenemos existen $x_1, \ldots, x_{n-1}$ tales que $q(x, x_1) \cdots q(x_{n-1}, y) > 0$. Entonces para este valor de n se tiene

$$\mu_{p_0,q}(\{((x_1, x_2, \ldots) \in X^{\mathbb{N}} : x_n = x\}) > 0$$

para todo $x \in X$.

Como $\mu_{p_0,q}$ es shift invariante esto implica que

$$p_0(x) = \mu_{p_0,q}(\{((x_1, x_2, \ldots) \in X^{\mathbb{N}} : x_1 = x\}) > 0$$

para todo X.

Como consecuencia de lo anterior dada cualquier otra probabilidad p existe C tal que $p(x) \leq C p_0(x)$ para todo $x \in X$. De esto se deduce que

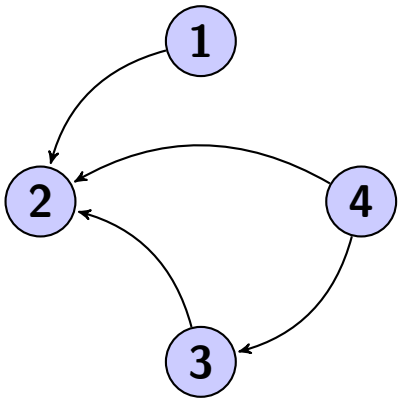
$$\mu_{p,q}([x_1, \ldots, x_m]) \leq C \mu_{p_0,q}([x_1, \ldots, x_m])$$

para todo cilindro $[x_1, \ldots, x_n] \subset X^{\mathbb{N}}$. Esto implica que cualquier conjunto de medida nula para $\mu_{p_0,q}$ también tiene medida nula para $\mu_{p,q}$, lo cual concluye la demostración. □

El teorema anterior implica que el PageRank de una página web x es simplemente $p_0(x)$ donde p_0 es la única probabilidad en X tal que $\mu_{p_0,q}$ es shift invariante. Esto reduce el problema de calcular el PageRank al de calcular un vector propio de una matriz con entradas positivas. Para esto esto existen buenos métodos numéricos uno de los más simples es simplemente calcular un potencia grande de la matriz.

Ejercicio 4. *Sea A una matriz $n \times n$ con entradas estrictamente positivas y tal que la suma de coordenadas de Av es igual a la de v para todo $v \in \mathbb{R}^n$. Mostrar que dado cualquier v con entradas no negativas y con suma de coordenadas igual a 1, $A^n v \to v_0$ donde v_0 es el único vector propio de A con entradas positivas cuyas coordenadas suman 1.*

Ejercicio 5. *En el siguiente grafo (se supone que cada vértices es una página web y las flechas son los enlaces) ¿Cuál te parece que sería el orden de "importancia" de los nodos? Calculá el PageRank exactamente, o al menos con suficiente exactitud para ver si tu intuición coincide con el resultado.*


1
2
4
3