Matemática para Ciencia de Datos

Cadenas de Markov

Dr. Daniel Alexis Gutierrez Pachas (dgutierrezp@ucsp.edu.pe)

26 de noviembre de 2023

Departamento de Ciencia de la Computación, Universidad Católica San Pablo, Arequipa, Perú.

Cadenas de Markov

Una cadena de Markov es un proceso aleatorio compuesto de una secuencia de variables aleatorias:

$$\Theta : \{\theta(1), \theta(2), \theta(3) \ldots\}.$$

La propiedad indica que si se conoce la historia del sistema hasta su instante actual, su estado presente resume toda la información relevante para describir en probabilidad su estado future.

Asumimos $\mathbb{I} = \{1, 2, \ldots, N\}$ y asumiendo una secuencia de índices $i_1, i_2, \ldots, i_k$ del conjunto $\mathbb{I}$ tenemos que:

$$\begin{aligned} &P(\theta(k+1) = i_{k+1} \,|\, \theta(k) = i_k, \theta(k-1) = i_{k-1}, \ldots, \theta(2) = i_2, \theta(1) = i_1) \\ =&P(\theta(k+1) = i_{k+1} \,|\, \theta(k) = i_k. \end{aligned}$$

Propiedades

Sea $p_{ji} = P(\theta(k+1) = i \,|\, \theta(k) = j)$, y definimos el vector distribución $\pi(k) = (\pi_1(k), \ldots, \pi_N(k))$. Con la formula anterior obtenemos

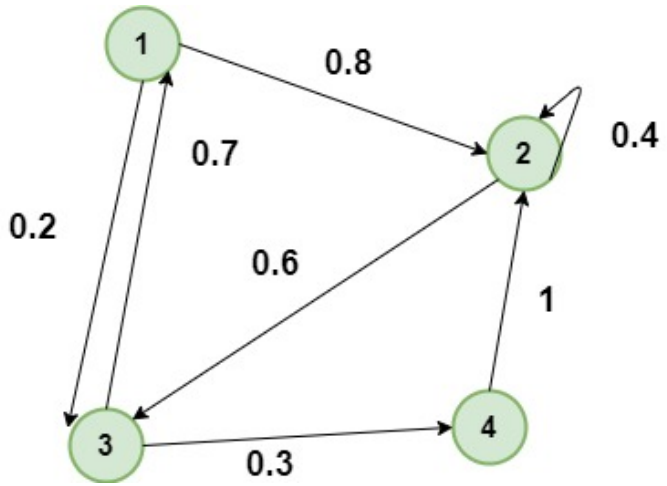
$$\begin{aligned}\pi_i(k+1) &= \sum_{j=1}^{N} P(\theta(k+1) = i \,|\, \theta(k) = j) \cdot P(\theta(k) = j) \\ &= \sum_{j=1}^{N} p_{ji} \cdot \pi_j(k)\end{aligned}$$

En términos matriciales $\pi(k+1) = \pi(k)\mathcal{P}$, donde

$$\mathcal{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix}, \text{ donde } \sum_{i=1}^{N} p_{ji} = 1, \forall\, j = 1, \ldots, N.$$

Ejemplo 1

Sea el grafo.

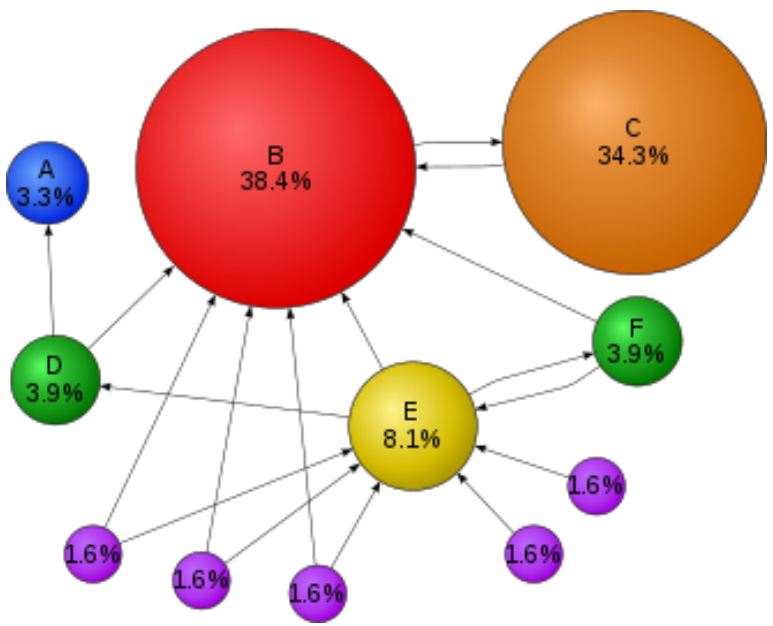


Tenemos que

$$\mathcal{P} = \begin{bmatrix} 0 & 0{,}8 & 0{,}2 & 0 \\ 0 & 0{,}4 & 0{,}6 & 0 \\ 0{,}7 & 0 & 0 & 0{,}3 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Page Rank

PageRank es una familia de algoritmos creada y desarrollada por Google para optimizar las búsquedas de páginas web. PageRank interpreta un enlace de una página A a una página B como un voto, de la página A, para la página B. Pero considera más allá del volumen de votos; también analiza la página que emite el voto. Los votos emitidos por las páginas consideradas importantes, valen más, y ayudan a hacer a otras páginas importantes.



Page Rank

El algoritmo inicial del PageRank lo podemos encontrar en el documento original donde sus creadores presentaron el prototipo de Google: “The Anatomy of a Large-Scale Hypertextual Web Search Engine”

$$\mathrm{PR}(A) = (1-d) + d \sum_{i=1}^{N} \frac{\mathrm{PR}(i)}{C(i)},$$

donde $\mathrm{PR}(A)$ es el PageRank de la página A, d es un factor de amortiguación que tiene un valor entre 0 y 1, $\mathrm{PR}(i)$ son los valores de PageRank que tienen cada una de las páginas i que enlazan a A y $C(i)$ es el número total de enlaces salientes de la página i (sean o no hacia A).

Generalmente se usa $d = 0{,}85$ y representa la probabilidad de que un navegante continúe pulsando links. Entonces, la probabilidad de que el usuario deje de pulsar links y navegue directamente a otra web aleatoria es $1 - d$.

Calculando el Page Rank

Los pasos a seguir son:

(a) Dado un grafo, calcular su matriz de adyacencia.

(b) Calcular la matriz de transición asociada.

(c) Calcular la matriz de Google.

(d) Calcular la distribución estacionária a la matriz de Google.