Práctica 1. El algoritmo PageRank

En esta práctica, repasaremos algunos comandos de Matlab, experimentaremos con distintos tipos de matrices de transición y veremos un ejemplo de aplicación de las cadenas de Markov: *PageRank*, el algoritmo de clasificación de páginas web que dio origen al popular buscador Google.

1. Vectores, matrices y ecuaciones lineales en Matlab

En este apartado se repasan todos los comandos de Matlab útiles en esta práctica. En el fichero **comandos_matlab_1.m** encontrará ejemplos de uso de cada comando. Ejecute aquellos que no conozca y compruebe su funcionamiento.

2. Probabilidades en n saltos. Ecuación de Chapman-Kolmogorov

Ejecute el fichero **matices_X.m**, donde X = 1 si su turno de prácticas es el primero y X = 2 si su turno es el segundo. Este fichero crea tres matrices: Pa, Pb y Pc. Evalúe a qué valores convergen las probabilidades de transición en n etapas, y a partir de los resultados, clasifique las clases contenidas en cada matriz.

	Pa	Pb	Pc
Número de estados transitorios			
Número de clases recurrentes aperiódicas			
Número de clases recurrentes periódicas			
Es ergódica (si/no)			

3. Solución de las ecuaciones de balance

Programe la función **v = SolveErgodicDTMC(P).** El argumento (P) es una matriz de transición con una sola clase recurrente y aperiódica y el resultado (v) es el vector de estado estacionario de P.

4. Algoritmo PageRank

Programe la función **[r i] = PageRank(A, $\boldsymbol{\alpha}$)**, que devuelve el vector **r** con los *Ranks* y el vector **i** con los identificadores de las páginas, ordenados de mayor a menor *Rank*. El argumento $\boldsymbol{\alpha} \leq 1$ es un escalar positivo y **A** es la matriz de incidencias definida así:

$$\boldsymbol{A}(i,j) = \begin{cases} 1 & \text{si } i \text{ contiene un hipervínculo a } j \\ 0 & \text{en caso contrario} \end{cases}$$

Anexo: Fundametos de PageRank

El objetivo de *PageRank* es ordenar por importancia las direcciones web. La importancia de una página j se basa en el número de páginas que la referencian (páginas que contienen hipervínculos a j) así como la importancia de las páginas que la referencian. Supongamos que una página i referencia a j. La contribución de i en la importancia de j es proporcional a la importancia de i e inversamente proporcional al número total de hipervínculos de i.

Denominemos $H(j)$ al conjunto de páginas que contienen hipervínculos a j. La importancia (*Rank*) de j (r_j) se define inicialmente como:

$$r_j = \sum_{k \in H(j)} r_k \frac{1}{n_k}$$

Donde n_k es el número de hipervínculos que contiene la página k.

Si definimos la matriz de hipervínculos $\boldsymbol{Q}$ de la siguiente forma:

$$\boldsymbol{Q}(i,j) = \begin{cases} \frac{1}{n_i} & \text{si } i \text{ contiene un hipervínculo a } j \\ 0 & \text{en caso contrario} \end{cases}$$

Resulta que $\boldsymbol{Q}$ es una matriz de transición y el ranking de las páginas se puede obtener resolviendo las ecuaciones de balance:

$$\mathbf{r} = \mathbf{rQ} \text{ con } \mathbf{re} = 1, \text{ donde } \mathbf{e} = [1,1,\dots,1]^T.$$

Sin embargo nada nos garantiza que $\boldsymbol{Q}$ sea ergódica. De hecho en la realidad no lo es. Por tanto, para resolver el problema del *PageRank* se redefine el *Rank* de la siguiente forma:

$$r_j = \alpha \sum_{k \in H(j)} r_k \frac{1}{n_k} + \frac{(1-\alpha)}{N}$$

Donde N es el número total de páginas y α es un factor de ponderación ajustado a 0.85 generalmente. Es decir el *Rank* de j (r_j) es una media ponderada entre una importancia homogénea entre todas las páginas y el *Rank* definido anteriormente.

Para obtener la expresión matricial de *PageRank* definimos la matriz $\mathbf{M}$:

$$\mathbf{M} = \alpha\mathbf{Q} + \frac{(1-\alpha)}{N}\mathbf{1}_{N\times N}$$

Donde $\mathbf{1}_{N\times N}$ es una matriz de N×N con todos sus elementos iguales a 1. Las ecuaciones quedan así:

$$\mathbf{r} = \mathbf{rM} \text{ con } \mathbf{re} = 1$$