Presentación

Os pedimos en esta actividad que resolváis el caso de uso propuesto mediante el modelado con Cadenas de Markov. Este caso de uso os permitirá poner en práctica los conceptos trabajados en este reto, entender y coger destreza en su aplicación a un caso de uso concreto utilizando datos reales o realistas. Veréis también la necesidad de utilizar un lenguaje de programación, como por ejemplo `R`, para su resolución y cogeréis destreza en su utilización.

Competencias

En esta PEC se trabajan las siguientes competencias del Grado en Ciencia de Datos Aplicada:

- Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.
- Utilizar de forma combinada los fundamentos matemáticos, estadísticos y de programación para desarrollar soluciones a problemas en el ámbito de la ciencia de datos.
- Uso y aplicación de las TIC en el ámbito académico y profesional.

Objetivos

Los objetivos concretos de esta Práctica son:

- Comprender la utilidad de los conceptos de álgebra lineal que se han trabajado en los retos 1-3 en la aplicación en el ámbito de la ciencia de datos mediante el análisis de componentes principales y la descomposición en valores singulares.
- Ser capaz de resolver un problema utilizando la descomposición en valores singulares en un caso de uso utilizando datos reales o realistas.
- Entender la utilidad de utilizar un lenguaje de programación para el tratamiento de grandes volúmenes de datos.

- Coger destreza en la utilización del lenguaje R para la resolución de problemas con un gran volumen de datos.

Descripción de la Práctica a realizar

Saber modelizar un sistema mediante sistemas dinámicos en tiempo discreto nos puede ser muy útil en nuestra carrera profesional en el ámbito de la ciencia de datos. El ejemplo más famoso de este tipo de modelos es el del algoritmo PageRank que utiliza el buscador Google.

Por un lado, os pedimos que respondáis un **cuestionario** (se puede encontrar en el aula Moodle entrando en el enlace "Cuestionarios" en la parte derecha del aula) en el que vamos a trabajar la parte más instrumental de este reto en una serie de preguntas genéricas.

Os pedimos también que resolváis la práctica descrita en este documento. Estos ejercicios os plantearán escenarios propios de la ciencia de datos y veréis como los conceptos trabajados en este reto tienen relevancia en estos contextos.

Recursos

Recursos Básicos

- Capítulo *Cadenas de Markov de tiempo discreto* del libro *Una introducción amable a la teoría de colas.*
- *Introducción a la teoría de matrices positivas: aplicaciones.*
- Documento Problemas sobre modelos matriciales en la ciencia de datos.

Recursos Complementarios

- Caso de uso y guía de resolución en `R`.

Criterios de valoración

- La práctica se ha de resolver de manera individual.
- Es necesario justificar todos los pasos realizados en la resolución de la Práctica.

Tened en cuenta que las dos actividades que se plantean en este reto (la resolución de la práctica que se plantea en este documento y el cuestionario) serán parte de la nota de prácticas (Pr = (Pr1 + Pr2) / 2). La nota de estas actividades corresponde a la Pr2 (con un peso del 20 % para el cuestionario y un 80 % para la práctica). Para más información sobre el modelo de evaluación de la asignatura, consultad el plan docente.

Formato y fecha de entrega

Para realizar la práctica correctamente, es necesario consultar y contestar la tabla resumen asociada a la práctica con los resultados obtenidos. La encontraréis en el aula Moodle RETO 5 - Tabla resumen de la práctica 2". Tenéis dos intentos para responder a la tabla, el primero de ellos con retroalimentación en las respuestas. Fijaos que en el enunciado de la tabla encontraréis parámetros necesarios para realizar la práctica y que cambian en cada intento.

Es necesario entregar un único documento en PDF que incorpore:

- la resolución de la práctica (memoria técnica detallada). **Es necesario especificar a qué intento de la tabla corresponde**;
- el código `R`; y
- las imágenes o figuras que se os piden.

Este fichero debe entregarse en el espacio del registro de la evaluación continua (RAC) del aula antes de las 23:59 horas del día 07/07/2023 (hora central europea (CET)).

Recordad que la práctica es **individual**. La detección de carencia de originalidad será penalizada de acuerdo con la normativa vigente de la UOC. Además, al realizar la entrega, aseguraos de comprobar que el archivo colgado es el correcto; pues es responsabilidad del alumnado realizar la entrega correctamente.

No se aceptarán entregas fuera de plazo ni en formatos que no sean los especificados.

1. Cadenas de Markov discretas / El camino aleatorio de la persona apasionada por la lectura

Una persona apasionada por la lectura camina alrededor de una isla de casas del Eixample de Barcelona. Su casa (**H**) se encuentra en una de las esquinas de la isla de casas. También hay una biblioteca municipal (**B**) en una de las esquinas contiguas, con una extensísima colección de libros. Este apasionado lector o lectora, cada vez que llega a una esquina (**C1** o **C2**), aleatoriamente, gira a la izquierda o vuelve atrás. La probabilidad de girar a la izquierda (moverse en sentido contrario a las agujas de un reloj) es del $100E\,\%$ (consultar vuestro valor de E en el cuestionario asociado a la práctica). Si llega a casa (**H**) o a la biblioteca (**B**), ya no sale porque se queda leyendo apasionadamente. Para más detalles, ver la Figura 1.

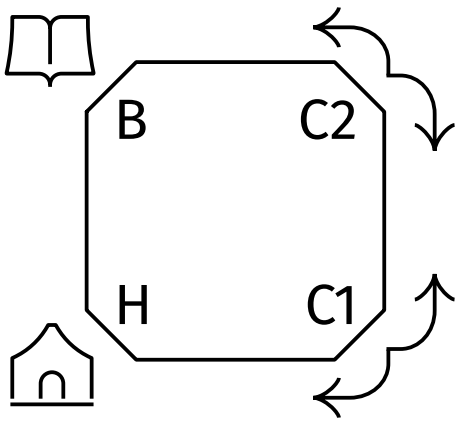

Figura 1: El camino aleatorio de la persona apasionada por la lectura.

1. [10 %] **Representar gráficamente (a mano o con ordenador de forma esquemática) la cadena de Markov** que describe la ubicación de la persona apasionada por la lectura. Tener en cuenta que esta cadena de Markov tiene cuatro estados, que podemos denotar como **H** (casa), **C1** (primera esquina), **C2** (segunda esquina) y **B** (biblioteca). Mostrar la probabilidad de las transiciones entre dos estados. **Completar también la matriz de transición** de la Tabla 1 siguiendo las indicaciones descritas en el enunciado de la práctica y vuestro valor de E.

Tabla 1: Matriz de transición entre la posición inicial de la persona apasionada por la lectura y la posición final.

	Ubicación final			
Ubicación inicial	**H**	**C1**	**C2**	**B**
H				
C1				
C2				
B				

2. [10 %] Considerar la matriz de transición de la Tabla 1 (ya completada) y **definir la matriz de transición** `P` por filas (¡no por columnas!). Si generáis un vector `x` con los 16 valores de la tabla (por filas), podéis generar la matriz `P` de la siguiente manera:

```
labels<-c("H","C1","C2","B")
byRow <- TRUE
P<-matrix(data=x,byrow=byRow,nrow=4,dimnames=list(labels,labels))
```

   Comprobar que la suma de las probabilidades de cada una de las cuatro filas es 1. **Calcular también la suma de las probabilidades de cada una de las columnas**.

3. [10 %] Una cadena de Markov se llama *regular* (también *primitiva* o *ergódica*) si existe alguna potencia positiva de la matriz de transición cuyas entradas sean todas estrictamente mayores que cero. Pero no será necesario examinar las potencias de la matriz de transición $\texttt{mcP}^m$ para todo m hasta el infinito (¡afortunadamente!). Un teorema nos dice que es suficiente examinar las potencias $\texttt{mcP}^m$ para valores naturales de m menores o iguales a $(n-1)^2+1$ donde n es el número de estados (en nuestro caso, $n=4$). **Generar un programa o escribir unas líneas de código** que, dada la cadena de Markov `mcP`, nos devuelva una lista `mcL` de $m=(4-1)^2+1=10$ elementos, donde el elemento i-ésimo de la lista sea la suma del número de ceros de la potencia $\texttt{mcP}^i$. Como ayuda, pensar que el primer elemento de la lista debe ser 10. **¿Es la cadena de Markov regular?**

4. [10 %] **Identificar y justificar qué** estados son absorbentes. Relacionar la existencia de estados absorbentes con la regularidad o no de la cadena de Markov. Es decir:

   - ¿puede una cadena de Markov regular tener estados absorbentes?
   - si una cadena de Markov tiene algún estado absorbente, ¿puede ser regular?

5. [10 %] Cuando se trabaja con cadenas de Markov con estados absorbentes, como es el caso en esta práctica, a menudo es conveniente reorganizar la matriz de manera que las filas y columnas correspondientes a los estados absorbentes se enumeren primero. Esto se llama

forma canónica. **Reescribir la matriz de transición de la Tabla 1 en la forma canónica**. Obtendréis una matriz de transición con esta estructura:

$$\mathtt{Q} = \left(\begin{array}{c|c} \mathbb{I} & \mathbf{0} \\ \hline \mathbf{A} & \mathbf{D} \end{array}\right)$$

donde:

- $n = 4$ es el número de estados de la cadena de Markov;
- κ es el número de estados absorbentes;
- $\mathtt{Q}$ es una matriz cuadrada de n filas y n columnas;
- $\mathbb{I}$ es la matriz identidad de κ filas y κ columnas;
- $\mathbf{0}$ es la matriz nula de κ filas y $(n - \kappa)$ columnas;
- $\mathbf{A}$ es una matriz de $(n - \kappa)$ filas y κ columnas;
- $\mathbf{D}$ es una matriz de $(n - \kappa)$ filas y $(n - \kappa)$ columnas.

6. [15 %] Considerar ahora la cadena de Markov donde la probabilidad de girar a la izquierda es de $100L$ % donde $L \in (0, 1)$. En este caso, la probabilidad de comenzar en la esquina **C1** y terminar en la biblioteca (**B**) viene definida por la serie numérica

$$p_{\mathbf{C1}\to\mathbf{B}} = L^2 \sum_{k=0}^{+\infty} L^k (1 - L)^k,$$

y la probabilidad de comenzar en la esquina **C2** y terminar en la biblioteca (**B**) viene definida por la serie numérica

$$p_{\mathbf{C2}\to\mathbf{B}} = L \sum_{k=0}^{+\infty} L^k (1 - L)^k.$$

Adicionalmente, sabemos que en el límite, todos los estados **H**, **C1**, **C2** y **B**, acaban en uno de los estados absorbentes.

Sabiendo que la suma de una progresión geométrica de razón r es

$$\sum_{k=0}^{+\infty} r^k = \frac{1}{1 - r},$$

¿para qué valor de L la probabilidad de comenzar en la esquina **C1** y terminar en la biblioteca (**B**) **es igual** a la probabilidad de comenzar en la esquina **C1** y terminar en casa (**H**)? ¿Para qué valor de L la probabilidad de comenzar a la esquina **C2** y acabar en la biblioteca (**B**) **es** C **veces superior** a la probabilidad de comenzar a la esquina **C1** y acabar en la biblioteca (**B**)? Consultar vuestro valor de C en el cuestionario asociado a la práctica y dar el resultado exacto, no una aproximación decimal.

7. [15 %] Teniendo en cuenta vuestro valor de E, **¿cuál es la probabilidad de que el lector o lectora acabe en la biblioteca (B) si comienza en la esquina C1? ¿Y la probabilidad de que acabe en casa (H)?** Son compatibles estas probabilidades con el hecho de que la esquina **C1** esté más cerca de casa (**H**)? Usar los resultados de la pregunta 6.

8. [20 %] Para determinar la tendencia a largo plazo, utilizamos la **matriz solución** calculada como

$$\mathbf{S} = \lim_{n\to\infty} \mathtt{Q}^n,$$

donde `Q` es la matriz de transición en forma canónica definida en la pregunta 5. Es decir, elevamos la matriz de transición a potencias elevadas y miramos su convergencia.

En el caso de una cadena de Markov con estados absorbentes, en el límite se absorben todos los estados no absorbentes, es decir, las columnas asociadas a los estados no absorbentes de la matriz $\mathbf{S}$ están formadas por ceros. Se puede demostrar que

$$\mathbf{S} = \lim_{n\to\infty} \mathtt{Q}^n = \left(\begin{array}{c|c} \mathbb{I} & \mathbf{0} \\ \hline \mathbf{S}_r & \mathbf{0} \end{array}\right),$$

donde $\mathbf{S}_r$ es la **matriz de solución reducida** (excluyendo las columnas asociadas a los estados no absorbentes y las filas asociadas a los estados absorbentes) que se puede calcular como

$$\mathbf{S}_r = (\mathbb{I}_{n-\kappa} - \mathbf{D})^{-1}\mathbf{A},$$

donde $\mathbf{A}$ y $\mathbf{D}$ están definidas en la pregunta 5 y $\mathbb{I}_{n-\kappa}$ es la matriz identidad de dimensión $n-\kappa$. La matriz $\mathbf{S}_r$ tiene $(n-\kappa)$ filas (correspondientes a los estados no absorbentes) y κ columnas (correspondientes a los estados absorbentes).

Calcular, con `R`, la matriz $\mathbf{S}_r$ y comprobar que en el límite (escogiendo valores de n suficientemente grandes),

$$\lim_{n\to\infty} \mathtt{Q}^n = \left(\begin{array}{c|c} \mathbb{I} & \mathbf{0} \\ \hline \mathbf{S}_r & \mathbf{0} \end{array}\right).$$

Relacionar los valores de esta matriz con los valores que podéis calcular con las fórmulas de la pregunta 6.