Cadenas de Markov de tiempo continuo

Introducción

Fabián Abarca Calderón

IE0405 – Modelos Probabilísticos de Señales y Sistemas

18

Tema V

UNIVERSIDAD DE COSTA RICA

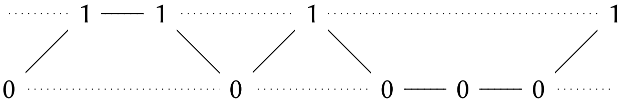
EIE Escuela de Ingeniería Eléctrica

Muchos fenómenos del mundo real pueden ser modelados como una secuencia de **transiciones** de un **estado** a otro, donde cada transición tiene una incertidumbre asociada.
Las cadenas de Markov proveen un modelo para información secuencial que permite que resultados futuros dependan de otros resultados previos.

Definición de cadenas de Markov

Ejemplos de “cadenas de Markov” I

- Un modelo simplista del tiempo atmosférico en un lugar está dado por la secuencia temporal $\{X_1, X_2, X_3, \ldots\}$, donde X_i tiene los dos únicos valores 0 (“seco”) y 1 (“húmedo”).



- Los precios de las acciones de una empresa tienen una incertidumbre respecto a si el precio va a hacer una *transición* hacia un valor mayor o menor.

En ambos casos, las **transiciones** pueden modelarse de manera probabilística y es crucial el conocimiento del estado actual. Las cadenas o procesos de Markov son un modelo de *dependencia*.

Otros ejemplos de aplicaciones

- Predicción de cambios en la demanda de electricidad
- Modelado del movimiento de insectos
- Monitoreo del uso de camillas en hospitales
- Personas que cambian de servicio celular
- Monitoreo de patrones de navegación web para proveer anuncios personalizados

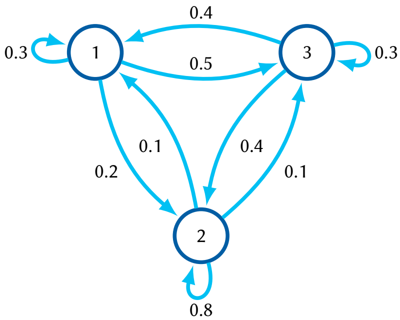
- Las cadenas de Markov son **procesos estocásticos** que cumplen la *propiedad de Markov*.
- Informalmente, esta propiedad establece que los estados futuros dependen únicamente del estado actual, y no de la secuencia de estados para llegar ahí. De algún modo, "trunca la memoria".

Propiedad de Markov o de la falta de memoria

Formalmente, es una *probabilidad condicional* de la forma

$$P(X_n = x \mid X_0, \ldots, X_{n-1}) = P(X_n = x \mid X_{n-1}), \quad \forall n, \forall x \tag{1}$$

Una ciudad tiene tres diferentes “zonas”, numeradas 1, 2 y 3, y los taxis operan en todas las zonas.



La probabilidad de que el próximo pasajero va hacia una zona particular depende de dónde abordó el taxi.

Deducción a partir de la distribución exponencial

Un par de casos especiales de la distribución exponencial son útiles para determinar los parámetros de interés en las cadenas de Markov.

La distribución exponencial como "tiempo de vida"

> La distribución exponencial es la típica caracterización de la probabilidad de los tiempos de espera (o "de vida").

Si T es el tiempo de vida de un componente que está exponencialmente distribuido con parámetro α, entonces T tiene densidad

$$f_T(t) = \begin{cases} 0 & t < 0 \\ \alpha e^{-\alpha t} & t > 0 \end{cases} \tag{2}$$

La media de T es el recíproco del parámetro α, $E[T] = \frac{1}{\alpha}$.

- La variable aleatoria T tiene la *propiedad de falta de memoria*
- "No importa la antigüedad del componente, este opera como si fuera nuevo".

Propiedad de falta de memoria de la distribución exponencial

La propiedad de falta de memoria, matemáticamente, se escribe[1] como

$$P(T > t + s | T > t) = P(T > s) \tag{3}$$

para tiempos $t, s \geq 0$. También se tiene que

$$P(T > t) = e^{-\alpha t} \tag{4}$$

para $t \geq 0$ (porque $F_T(t) = 1 - e^{-\alpha t}$).

A partir de estas definiciones, se determinan dos casos específicos.

[1] Por ejemplo, T podría ser "el tiempo de espera de un autobús", y $s = 5$ min y $t = 30$ min.

- Supóngase que $T_1, T_2, \ldots, T_N$ son variables aleatorias independientes, cada una distribuida exponencialmente. Supóngase que T_i tiene parámetro α_i.
- Suponga N componentes que se "conectan" (o "inician su operación conjunta") al tiempo $t = 0$.
- T_i es el tiempo de vida del i-ésimo componente.
- Sea M el mínimo de todos los tiempos T_i's de los componentes. M es el tiempo en que el primer componente falla.
- M es una variable aleatoria.

- Sea $t \geq 0$. Entonces $M = \min\{T_1, \dots, T_N\}$ es más grande que t si y solo si **todo** $T_i > t$.

$$\begin{aligned} P(M > t) &= P(\min\{T_1, T_2, \dots, T_N\} > t) \\ &= P(T_1 > t, T_2 > t, \dots, T_N > t) \\ &= e^{-\alpha_1 t} \cdot e^{-\alpha_2 t} \cdots e^{-\alpha_N t} \\ &= e^{-(\alpha_1 + \alpha_2 + \ldots + \alpha_N)t} \end{aligned} \tag{5}$$

M está exponencialmente distribuida con parámetro $\alpha_1 + \alpha_2 + \ldots + \alpha_N$ y valor medio $1/(\alpha_1 + \alpha_2 + \ldots + \alpha_N)$.

Ante la pregunta,

> *¿Cuál es la probabilidad de que, cuando el primer fallo suceda, sea el componente j-ésimo?*

La probabilidad de que entre las N variables aleatorias T_i el mínimo sea T_j está dada por la expresión

$$P(T_j = \min\{T_1, T_2, \ldots, T_N\}) = \frac{\alpha_j}{\alpha_1 + \alpha_2 + \cdots + \alpha_N} \tag{6}$$

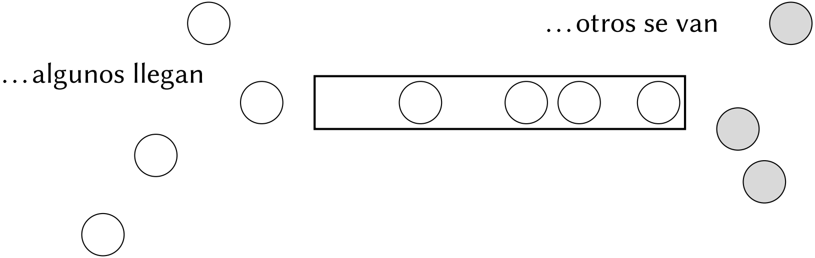
Prueba para dos variables aleatorias

Si T_1 y T_2 son independientes y distribuidos exponencialmente con parámetros respectivos α y β: $f_{T_1,T_2}(t_1, t_2) = \alpha e^{-\alpha t_1} \beta e^{-\beta t_2}$, para $t_1, t_2 > 0$. Con t_1 y t_2 correspondientes a T_1 y T_2:

$$
\begin{aligned}
P(T_1 = \min\{T_1, T_2\}) &= P(T_1 < T_2) \\
&= \iint_{T_1 < T_2} f_{T_1,T_2}(t_1, t_2)\, \mathrm{d}t_1\, \mathrm{d}t_2 \\
&= \int_0^\infty \int_0^{t_2} \alpha e^{-\alpha t_1} \beta e^{-\beta t_2}\, \mathrm{d}t_1\, \mathrm{d}t_2 \\
&= \int_0^\infty (1 - e^{-\alpha t_2}) \beta e^{-\beta t_2}\, \mathrm{d}t_2 \\
&= 1 - \frac{\beta}{\alpha + \beta} = \frac{\alpha}{\alpha + \beta}
\end{aligned}
$$

El proceso de nacimiento y muerte en tiempo continuo

"Nacimiento" y "muerte" es una analogía tétrica que puede interpretarse también como "aparición" y "desaparición" o "llegada" y "salida" o "unión" y "separación" o "haciendo fila" y "listo el trámite"...



El proceso de nacimiento y muerte en tiempo continuo II

- Considérese una "máquina" que puede estar en cualquiera de varios estados en cada instante de tiempo $t \geq 0$.
- El conjunto de estados posibles, el espacio de estados S, será siempre discreto:

$$S = \{0, 1, 2, \ldots, N\} \qquad \text{o} \qquad S = \{0, 1, 2, \ldots\}$$

- Al tiempo t, el estado de la máquina es denotado por X_t.

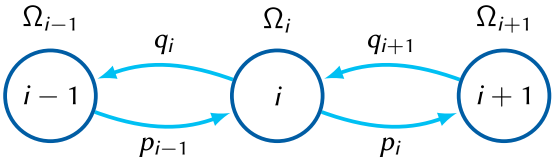
Por ejemplo, X_t podría denotar el número de animales en una poza para beber. El estado de la máquina es el número X_t de animales al tiempo t.

1. Si al tiempo t la máquina está en el estado i, permanece en ese estado por un tiempo aleatorio que es exponencialmente distribuido con parámetro Ω_i.
    - El tiempo de espera promedio en el estado i es el recíproco $1/\Omega_i$
    - Ω_i depende del estado i, pero no depende de otros estados anteriores
    - El estado i pudiera ser **absorbente**: una vez que la máquina entra al estado i, permanecerá siempre ahí.
    - En este caso, $\Omega_i = 0$ y el tiempo de espera promedio es $1/\Omega_i = \infty$.

2. Cuando la máquina sale del estado i, cambia al estado $i+1$ o al estado $i-1$ (habían 9 vacas, luego hay 8 si se va una, o 10 si llega otra). Sea

$$\begin{aligned} p_i &= P(\{\text{próximo estado es } i+1 \mid \text{último estado es } i\}) \\ q_i &= 1 - p_i \\ &= P(\{\text{próximo estado es } i-1 \mid \text{último estado es } i\}) \end{aligned} \tag{7}$$

Entonces p_i y q_i dependen solamente del estado i y no de otros detalles del proceso (tales como el tiempo t, la duración en i o el estado antes de i).



El proceso estocástico $\{X_t\}_{t=0}^{\infty}$ es un récord completo de los estados ocupados por la máquina para todos los tiempos $t \geq 0$.

Dado el presente estado X_t del sistema al tiempo t, los estados futuros de la máquina no dependen de los estados pasados.

- En particular, si el estado al tiempo t es $X_t = i$, entonces es completamente irrelevante si ha estado en el estado i por varios años o si acaba de cambiar al estado i, para predecir cuando se mudará del estado i.

- Dado que la distribución exponencial sigue la propiedad de la falta de memoria, la máquina se comporta como si acabara de moverse al estado i a pesar de qué tan largo hubiera realmente ocupado el estado i.
- La distribución exponencial es la única distribución continua concentrada en $[0, \infty[$ para los tiempos de espera que tiene esta propiedad.

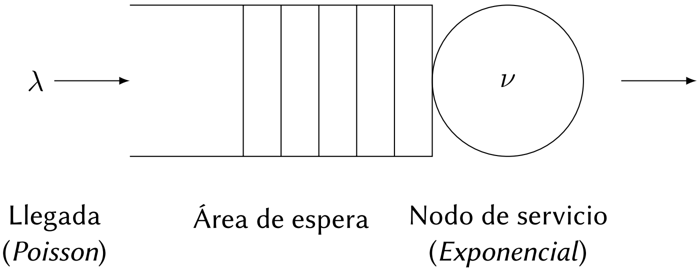
Nótese que si $\Omega_i = 0$ para el estado i, entonces los valores de p_i, q_i son innecesarios de especificar dado que la máquina no puede cambiar del estado i una vez en él.

Consiste de una máquina que puede cambiar entre estados en un espacio de estados S.

- X_t denota el estado ocupado al tiempo t para $t \geq 0$.
- La máquina permanece en el estado i por un periodo de tiempo indeterminado (llamado tiempo de espera o permanencia) que es exponencialmente distribuido con parámetro Ω_i (tiempo de espera promedio $1/\Omega_i$).
- Cuando la máquina cambia, cambia a los estados $i+1$, $i-1$ con probabilidades respectivas p_i, $q_i = 1 - p_i$.

Teoría de colas

La teoría de colas (*queueing theory* en inglés) es un estudio matemático de situaciones muy comunes, como carros que llegan a casetillas de peaje, clientes en una caja de supermercado, aviones que sobrevuelan un aeropuerto esperando aterrizar...



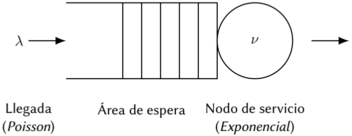
En teoría de colas, la "máquina" consiste de clientes y servidores.

- Es similar a un tanque de agua con un flujo de entrada y un flujo de salida. La relación entre estos flujos determina qué tan lleno o vacío está el tanque.
- Se determinan dos flujos:
  - arribo de los "clientes"
  - tiempo de "servicio".

- Los clientes arriban de acuerdo a una *corriente Poisson* con parámetro λ.

$$P(X = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

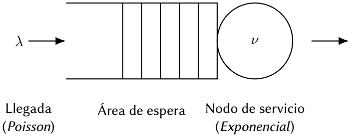
- En promedio, arriban λ clientes por unidad de tiempo.



- Puede haber uno o más servidores.
- Los tiempos de servicio son aleatorios, pero se supone que están exponencialmente distribuidos con parámetro ν (nu).

$$f_T(t) = \begin{cases} 0 & t < 0 \\ \nu e^{-\nu t} & t > 0 \end{cases}$$

- En promedio, el tiempo de servicio es $1/\nu$.



- X_t es la longitud de la cola en el tiempo t, el número de clientes incluyendo a aquel (o aquellos si hay más de un servidor) que son servidos en el tiempo t.
- Entonces el proceso estocástico $\{X_t\}_{t=0}^{\infty}$ es llamado "proceso de nacimiento y muerte".

Carencia de memoria

- Tanto la corriente de llegada de clientes así como la corriente de salida de clientes que han sido servidos obedecen la propiedad de la carencia de memoria.
- Si algún otro cliente se une a la cola, es un hecho que es independiente del número de clientes en fila u otros detalles pasados acerca de la cola.
- Dado que los tiempos de servicio son exponencialmente distribuidos, el hecho de que un cliente termine de ser servido en cierto periodo de tiempo es independiente de detalles de la historia de la cola, incluyendo lo largo que el cliente ha recibido servicio.

Supóngase que hay un cajero en el supermercado y que los clientes que arriban forman una corriente de Poisson con parámetro λ y el tiempo de servicio está exponencialmente distribuido con parámetro ν. El estado X_t en el tiempo t es la longitud de la cola. Así, $S = \{0, 1, 2, \ldots\}$.

Un movimiento de 0 a 1 cliente ocurre en un tiempo aleatorio exponencialmente distribuido con parámetro λ. De esta forma,

$$\Omega_0 = \lambda \qquad p_0 = 1$$

Si el estado al tiempo t es $X_t = i \geq 1$, un movimiento a un estado $i-1$ o $i+1$ puede ocurrir. Sea T el tiempo de arribo del próximo cliente y S sea el tiempo de servicio del cliente que es servido en el instante presente.

T y S son independientes y cada uno está exponencialmente distribuido con los respectivos parámetros λ y ν. Un movimiento ocurre fuera del estado i en un tiempo igual a $\min\{T, S\}$.

- La variable aleatoria $\min\{T, S\}$ está exponencialmente distribuida con parámetro $\lambda + \nu$.
- Un movimiento ocurre al estado $i + 1$ si $T = \min\{T, S\}$; esto ocurre con una probabilidad igual a $\lambda/(\lambda + \nu)$.
- Similarmente, el movimiento hacia $(i - 1)$ clientes ocurre si $S = \min\{T, S\}$, lo que se da con una probabilidad $\nu/(\lambda + \nu)$.

Para resumir, para la cola de un servidor y para $i \geq 1$,

$$\begin{array}{lll} \Omega_0 = \lambda & p_0 = 1 & \\ \Omega_i = \lambda + \nu & p_i = \frac{\lambda}{\lambda+\nu} & q_i = \frac{\nu}{\lambda+\nu} \end{array}$$

La cola con infinito número de servidores I

Ejemplo 2

- Supóngase ahora que los clientes llegan de acuerdo a una corriente de Poisson con parámetro λ
- Esta vez cada cliente recibe servicio instantáneo
- El tiempo de servicio es aún exponencialmente distribuido con parámetro ν
- Hay un infinito número de cajeros en el supermercado. De forma más realista, los clientes se sirven solos. Mejor aún: estamos en La Sabana y todos quienes llegan reciben "el servicio" de estar en el parque.
- Como en la cola de un servidor, $\Omega_0 = \lambda$. Si hay i clientes en el tiempo t, cada uno está recibiendo servicio.
- Sea sus respectivos tiempos de servicio $S_1, S_2, \ldots, S_i$. (Puede pensarse en esta también como "‘permanencia").

- Sea T el tiempo de arribo del próximo cliente. Entonces un movimiento fuera del estado i ocurre en un tiempo igual al $\min\{S_1, S_2, \dots, S_i, T\}$. Dado que cada S_j está exponencialmente distribuido con parámetro ν y T está exponencialmente distribuido con parámetro λ, el tiempo para moverse del estado i está exponencialmente distribuido con parámetro igual a la suma $\nu + \cdots + \nu + \lambda = i\nu + \lambda$.
- El movimiento será a $i + 1$ clientes si $T = \min\{S_1, S_2, \dots, S_i, T\}$. La probabilidad de esto es $\lambda/(i\nu + \lambda)$.
- El movimiento será para $i - 1$ clientes si cualquiera de los i clientes termina de ser servido. La probabilidad de que el cliente j-ésimo termine antes de los otros y también antes de un arribo es $\nu/(i\nu + \lambda)$.

- Cualquiera de los i-clientes podría finalizar para que el movimiento fuera hacia $(i-1)$ clientes. Así, la probabilidad de un movimiento a $(i-1)$ es $i\nu/(i\nu+\lambda)$ (esto último también puede obtenerse como $1-\lambda/(i\nu+\lambda)$).
- Para la cola con infinito número de servidores y para $i \geq 1$

$$\begin{array}{lll} \Omega_0 = \lambda & p_0 = 1 & \\ \Omega_i = i\nu + \lambda & p_i = \frac{\lambda}{i\nu+\lambda} & q_i = \frac{i\nu}{i\nu+\lambda} \end{array}$$

Una oficina de negocios tiene un teléfono con un botón de retención.

- Suponga que las llamadas entrantes forman una corriente de Poisson con parámetro λ.
- Cada llamada toma un tiempo exponencialmente distribuido con un promedio de $1/\nu$ minutos.
- Si una llamada llega durante un tiempo en que el teléfono está ocupado, es colocada en retención. Si otra llamada llega, recibe un tono de ocupado y debe colgar.
- Sea el estado del sistema el número de llamadas que reciben servicio o están retenidas.

¿Cuáles son los parámetros Ω_i y las probabilidades de transición p_i, q_i?

- El espacio de estados es $S = \{0, 1, 2\}$. Cambio del estado 0 al estado 1 ocurre con la llegada de una llamada. De esta forma,

$$\Omega_0 = \lambda \qquad p_0 = 1$$

- Si el estado es 2, un movimiento ocurre al estado 1 tan pronto como la llamada que es servida termina. De esta forma,

$$\Omega_2 = \nu \qquad q_2 = 1$$

- Si el estado es 1, un movimiento ocurre si, ya sea una llamada llega (exponencialmente distribuida con parámetro λ, o si la llamada que es servida termina (exponencialmente distribuida con parámetro ν).

$$\Omega_1 = \lambda + \nu \qquad p_1 = \frac{\lambda}{\lambda + \nu} \qquad q_1 = \frac{\nu}{\lambda + \nu}$$

Es posible definir un procedimiento general para los problemas de colas.

Primero, represéntese el tiempo de espera en el estado i como $\min\{S, T, \ldots\}$ donde $S, T, U, \ldots$ son independientes, cada uno distribuido exponencialmente (tales variables son ya sea tiempos de arribo o tiempos de servicio). Entonces:

Ω_i	p_i	q_i
suma de todos los parámetros	suma de los parámetros de **tiempo de arribo** divididos por Ω_i	suma de los parámetros de **tiempo de servicio** divididos por Ω_i

Los arribos forman una corriente Poisson con parámetro λ. Hay una cola formada, pero dos servidores. Cada tiempo de servicio está exponencialmente distribuido con parámetro ν. Cuando un cliente completa el servicio, un cliente de la fila empieza a ser servido y la cola decrece de tamaño por una unidad.
¿Cuáles son S, Ω_i, p_i, q_i?

- S es el número de clientes en fila o recibiendo servicio. De esta forma, $S = \{0, 1, 2, \ldots\}$. Sea $i \geq 2$.
- Ambos servidores están ocupados y el tiempo de espera en el estado i es $\min\{S_1, S_2, T\}$, donde S_1, S_2 son los respectivos tiempos de servicio para los servidores 1, 2, y T es el tiempo de arribo para un nuevo cliente.

- S_1, S_2, T están exponencialmente distribuidos con parámetros ν, ν, λ. Por el procedimiento mencionado arriba y para $i \geq 2$,

$$\Omega_i = 2\nu + \lambda \qquad p_i = \frac{\lambda}{2\nu + \lambda} \qquad q_i = \frac{2\nu}{2\nu + \lambda}$$

- Si $i = 1$, el tiempo de espera en el estado $i = 1$ es $\min\{S, T\}$ donde S es el tiempo de servicio del cliente que es servido. Así,

$$\Omega_1 = \nu + \lambda \qquad p_1 = \frac{\lambda}{\nu + \lambda} \qquad q_1 = \frac{\nu}{\nu + \lambda}$$

- Finalmente, el tiempo de espera en el estado 0 es el tiempo T de arribo; de esta forma,

$$\Omega_0 = \lambda \qquad p_0 = 1$$

Nota sobre los cambios de estado

- En una cola solamente un cambio ocurre a la vez.
- Por ejemplo, en la cola con infinito número de servidores, si hay 100 clientes al presente recibiendo servicio, un desplazamiento a 101 clientes o a 99 clientes solamente ocurre uno a la vez.
- Para visualizar esto, sea $S_1, \ldots, S_{100}$ los tiempos de servicio respectivos de los clientes y sea T el tiempo de arribo del próximo cliente. Dado que cada S_j y T están exponencialmente distribuidos, ellos son variables aleatorias continuas. La probabilidad de que cualesquiera dos de ellas tengan el mismo valor es cero.
- Pero esto debería ser el caso para que ocurra más de un movimiento al mismo tiempo. *En una cola, solamente un movimiento o cambio de estado ocurre a la vez.*

- **¿Qué es una cadena de Markov?**
  Luis Rincón, https://youtu.be/Trf9P7DnOHQ
- **Origin of Markov chains | Journey into information theory | Computer Science**
  Khan Academy Labs, https://youtu.be/Ws63I3F7Moc