

Procesos estocásticos

Sesión 10. Teoría de colas

Enrique Miranda

Universidad of Oviedo

Máster Universitario en Análisis de Datos
para la Inteligencia de Negocios

Contenidos

1. Elementos de un modelo de colas.
2. Parámetros de interés.
3. Ejemplos de modelos de colas.
4. Redes de colas.

Modelos de colas

Los procesos de nacimiento y muerte constituyen un caso particular de **modelos de colas**.

En un modelo de colas, tenemos individuos (clientes) que provienen de una población y que llegan a un servicio con un cierto número de servidores o recursos. Si al llegar todos los servidores están ocupados, deberá esperar en cola para obtener servicio.



El primer trabajo en teoría de colas fue publicado por **Erlang** en 1909. Se ocupó del problema del dimensionamiento de líneas y centrales para el servicio de telefonía.

Otras contribuciones importantes fueron realizadas por **Engset**, **Palm**, **Pollaczek** y **Kendall**, entre otros.

Elementos de un modelo de colas

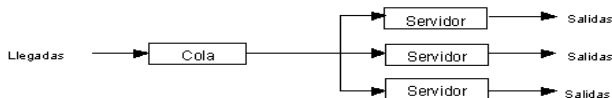
- ▶ La población o fuente de clientes.
- ▶ El modelo de llegadas.
- ▶ El modelo de servicio de cada servidor.
- ▶ El número de servidores o canales.
- ▶ El número de etapas del servicio.
- ▶ La capacidad del sistema.
- ▶ La disciplina de cola.

Ejemplos

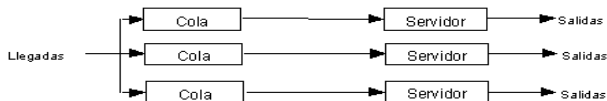
Una Cola , Un Servidor



una cola, múltiples servidores



Varias colas, múltiples servidores



Una cola, Servidores Secuenciales



Tipos de modelos de llegada

Se consideran varias distribuciones posibles para los tiempos entre dos llegadas consecutivas, que suelen resumirse con una letra:

- ▶ M: tiempos entre llegadas exponenciales.
- ▶ D: tiempos con patrón determinista o constante.
- ▶ E_k : tiempos con distribución de Erlang con k etapas.
- ▶ H: tiempos con distribución hiperexponencial.
- ▶ G: tiempos con una distribución arbitraria.

Si el patrón de llegadas no cambia con el tiempo, el modelo se dice **estacionario**. También existe la posibilidad de que se produzcan **llegadas en masa**.

Modelos de servicio

Se describe mediante la distribución del número de clientes servidos por unidad de tiempo. Se suele considerar independiente de los tiempos de llegada, y está condicionado a que el servidor esté ocupado, esto es, a que siempre haya clientes en el sistema.

Se utiliza la misma notación M, D, E_k, H, G para denotar las distribuciones. También existen modelos de **servicios en masa**.

Número de servidores/etapas

Existen modelos para uno o varios servidores, y para una o varias etapas. Por otro lado, en el caso de varias etapas también puede producirse un modelo de retroalimentación, típico en modelos de control de calidad.

Por otro lado, en otros modelos existe una limitación en el número máximo de clientes que puede haber en el sistema.

Disciplina de cola

- ▶ FIFO: El primero en llegar es el primero en ser servido.
- ▶ LIFO: El último en llegar es el primero en ser servido.
- ▶ SIRO: El servicio se produce en orden aleatorio.

Por defecto se considera el sistema FIFO. También se pueden incluir clases de prioridades sobre los clientes.

Se suele usar la notación de **Kendall** para un modelo de colas: $A/B/c/K/m/Z$, donde A denota el tiempo entre llegadas, B el tiempo de servicio, c el número de servidores, K la capacidad del sistema, m el tamaño de la población y Z la disciplina de cola. Si $K/m/Z = \infty/\infty/FIFO$ se omite la segunda parte.

Parámetros de interés en un modelo de colas

Para medir el rendimiento de un modelo de colas se suelen mirar los siguientes parámetros:

- ▶ Probabilidad de que haya n clientes en el sistema.
- ▶ Trabajo: **tiempo medio de servicio**, $\frac{1}{\mu}$.
- ▶ **Intensidad de tráfico**: $r = \frac{\lambda}{\mu}$, siendo λ el número medio de llegadas. Determina el número mínimo de servidores que se necesitan para que el sistema sea estable.
- ▶ **Uso del servidor**: $\rho = \frac{\lambda}{s\mu}$, siendo s el número de servidores.
- ▶ **Tiempo de permanencia** en el sistema, W y en la cola, W_q .
- ▶ **Número de clientes** en el sistema, N , y en la cola, N_q .

Fórmulas de Little

Cuando las tasas de llegada al sistema son constantes en λ , se verifican las igualdades

$$L = E(N) = \lambda W$$

y

$$L_q = E(N_q) = \lambda W_q.$$

La idea de las fórmulas de **Little** es que el número medio L de clientes en el sistema es igual a la tasa media de llegadas al sistema por el tiempo medio de permanencia en el sistema, y análogamente en lo que se refiere a la cola.

Por otro lado, también se cumple $W = W_q + \frac{1}{\mu}$: el tiempo medio en el sistema es igual a la suma de los tiempos medios de servicio y en la cola.

El modelo M/M/1

Se dispone de un sólo canal de servicio, las llegadas siguen una Poisson $\mathcal{P}(\lambda)$ y el tiempo de servicio es exponencial $\exp(\mu)$. Es un ejemplo de proceso de nacimiento y muerte. Sus parámetros son:

- ▶ Número medio de clientes en el sistema:

$$L = E(N) = \frac{\lambda}{\mu - \lambda}.$$

- ▶ Número medio de clientes en la cola:

$$L_q = L - (1 - \pi_0) = \frac{\lambda^2}{\mu(\mu - \lambda)}.$$

- ▶ Tamaño esperado de la cola: $\frac{\lambda}{\mu(\mu - \lambda)}.$

- ▶ La distribución del tiempo en el sistema es $\exp(\mu - \lambda)$, y por lo tanto $W = \frac{1}{\mu - \lambda}.$

El modelo M/M/1/K

A diferencia del anterior, en éste sólo se admite a un máximo de K clientes en el sistema. Obtenemos las siguientes medidas de rendimiento:

$$L = \begin{cases} \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} & \text{si } \rho \neq 1 \\ \frac{K}{2} & \text{si } \rho = 1. \end{cases}$$

- ▶ Tasa media de entradas al sistema: $\lambda_e = \lambda(1 - \pi_K)$, siendo $\pi_K = \frac{\rho^K}{\rho^{K+1}-1}$ la probabilidad de que el sistema esté al máximo de su capacidad.
- ▶ Tiempo medio en el sistema: $W = \frac{L}{\lambda_e}$.

El modelo M/M/s

Tenemos ahora s servidores en paralelo. Se obtiene

$$\blacktriangleright L_q = p_0 \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2}.$$

$$\blacktriangleright W_q = \frac{L_q}{\lambda}.$$

$$\blacktriangleright W = \frac{1}{\mu} + W_q.$$

$$\blacktriangleright L = \lambda W,$$

siendo

$$p_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!(1-\rho)}}$$

la probabilidad de que no haya clientes en el sistema.

Cuando hacemos $s \rightarrow \infty$, se ve cómo la variable N tiende a una distribución de Poisson con parámetro $\frac{\lambda}{\mu}$.

El modelo M/M/1/K/K

Consideramos ahora un modelo con una población finita de tamaño K . Se conoce también como **modelo de interferencia de máquinas**: podemos considerar K máquinas independientes, cada una con tasa λ de fallos y un sistema de reparación con tasa μ de servicio.

Definimos $B(K, z) = \frac{z^K/K!}{\sum_{n=0}^K z^n/n!}$. Entonces:

- ▶ $\rho = 1 - B(K, \frac{1}{r})$.
- ▶ $\lambda_e = \mu\rho$.
- ▶ $W = \frac{K}{\lambda_e} - \frac{1}{\lambda}$.
- ▶ $W_q = \frac{K}{\lambda_e} - \frac{1}{\lambda} - \frac{1}{\mu}$.

Otros modelos

Existen muchos otros modelos de colas que pueden resultar interesantes en la práctica, como son el $M/M/\infty$, el $M/M/s/K$ o el $M/M/1/\infty/H$, entre otros. No obstante, los modelos de colas también sufren de algunas limitaciones:

- ▶ Los cálculos de los parámetros de interés son complejos para muchos de estos modelos, y en algunos casos no existe solución analítica; debemos entonces recurrir a la simulación.
- ▶ Las distribuciones matemáticas consideradas suelen ser una aproximación de la realidad, sobre todo el tiempo de servicio exponencial.
- ▶ En particular, cuando el sistema de servicio no es FIFO los cálculos son bastante más complicados.

Redes de colas

Hasta ahora suponíamos que en el modelo de colas teníamos una sola etapa de servicio. Sin embargo, en la práctica podemos tener varias etapas de servicio, y los clientes pueden tener un comportamiento variable:

- ▶ Pueden llegar al sistema por distintas etapas.
- ▶ Al finalizar un servicio acudir a otro distinto o abandonar el sistema.
- ▶ Pueden volver a etapas visitadas previamente, e incluso abandonar el sistema definitivamente.

Hablamos entonces de una **red de colas**.

Redes abiertas y cerradas

Una **red abierta** es aquélla en la que los clientes pueden entrar o salir del sistema.

Una **red cerrada** es aquélla en la que ni entran ni salen clientes, sino que el número es constante en el tiempo. Es el caso de los sistemas de reparación de máquinas.

Redes abiertas (I): colas en serie

Consideramos por ejemplo un modelo en que los clientes llegan según una $\mathcal{P}(\lambda)$, y pasan sucesivamente por dos colas en serie, con tasas de servicio μ_1, μ_2 , respectivamente. Entonces:

- ▶ El número de clientes en cada uno de los servidores es independiente del otro.
- ▶ Los tiempos de espera de un cliente en cada cola no son independientes.
- ▶ En cambio, los tiempos totales (añadiendo el tiempo de servicio) sí lo son.

Redes abiertas (II): redes de Jackson

Son una red de colas con R nodos con las siguientes características:

- ▶ Cada nodo i consiste en s_i servidores idénticos, con tasa de servicio μ_i .
- ▶ La llegada al nodo i desde fuera del sistema sigue una $Poisson(\lambda_i)$.
- ▶ Una vez servido en el nodo i , un cliente pasa al nodo j con probabilidad p_{ij} , o sale del sistema con probabilidad $1 - \sum_j p_{ij}$.

Entonces,

- ▶ La tasa media de llegadas a totales al nodo i es $\Lambda_i = \lambda_i + \sum_j \Lambda_j p_{ji}$.
- ▶ Si $\Lambda_i < s_i \mu_i \forall i$, entonces cada servidor se comporta como un modelo $M/M/s_i$ con llegadas Λ_i y servicio μ_i de forma independiente.

Redes de Jackson cerradas

Consideramos las hipótesis anteriores junto con $\sum_j p_{ij} = 1$ para todo i (para que sea cerrada).

Existen fórmulas recursivas para la permanencia en cada servidor en función del número total de clientes.

Jackson también estudió situaciones más generales:

- ▶ Las llegadas desde fuera del sistema pueden depender del total de clientes en el sistema.
- ▶ La tasa de servicio en un nodo puede depender del número de clientes en el nodo.
- ▶ Se puede incluir el tiempo de viaje entre los nodos de la red.

Otras redes de colas

Otros tipos de redes de colas son:

- ▶ Modelos de **colas en serie**, en los que la entrada de clientes sólo es posible en la primera cola, y la salida por la última. Son una red de Jackson abierta.
- ▶ **Colas cíclicas**, parecidas a las anteriores pero conectadas de forma circular. Son una red de Jackson cerrada.
- ▶ **Colas con bloqueo**, en los que algunos de los modelos de colas tienen capacidad limitada.