SIMULAZIONE DI SISTEMI CASUALI – 2 parte

Processi stocastici e teoria delle code

Processi stocastici

Generalità

La distribuzione di Poisson (degli eventi rari) è caratterizzata dall'avere una funzione di probabilità che dipende da due variabili: la variabile k e la variabile tempo.

Un fenomeno descritto dalla distribuzione di Poisson è un processo stocastico.

"Un processo stocastico è una famiglia di variabili casuali che dipendono da un parametro t"

Un processo stocastico è indicato con la notazione:

$$\{X_t, t \in T\}$$

dove t è un parametro e T è l'insieme dei possibili valori di t.

Di solito con t si indica il tempo, e quindi un processo stocastico è una famiglia di variabili casuali dipendenti dal tempo. Infatti, l'utilizzo dei processi stocastici deriva dall'esigenza di descrivere un fenomeno aleatorio in evoluzione nel tempo.

Le variabili casuali X_t sono definite sull'insieme X, detto *spazio degli stati*, che può essere un insieme continuo, e in tale caso si parla di *processo stocastico continuo*,oppure un insieme discreto, e in tale caso si parla di *processo stocastico discreto*.

Le catene di Markov

Un tipo particolare di processi stocastici è costituito dai processi di Markov (A.A. Markov, matematico russo, 1856-1922), definiti come processi nei quali:

"*il futuro, dato il presente, è indipendente dal passato*".

Un processo di Markov è uno studio generalizzato del problema delle prove ripetute nel quale il tempo assume un'importanza fondamentale. Riprendiamo il concetto generale di processo stocastico, come una famiglia di variabili casuali:

$$\{X_t, t \in T\}$$

definita sullo spazio X comprendente tutti i possibili valori che le variabili casuali possono assumere.
Lo spazio X è definito *spazio degli stati* del processo e i suoi elementi xiÎ X, chiamati *stati del sistema*, rappresentano i possibili risultati di un esperimento.
Un processo stocastico è detto *markoviano* se le *probabilità di transizione* (ossia le probabilità che regolano il passaggio da uno stato ad un altro stato) dipendono *unicamente* dallo stato assunto dal sistema nell'istante precedente a quello considerato. In altre parole, lo stato presente del sistema permette di conoscere il suo comportamento futuro e la storia precedente non ha influenza; per questo motivo i processi markoviani sono detti "senza memoria".

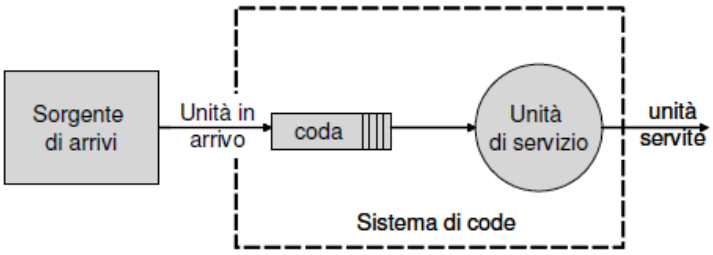
Scopo dello studio delle catene di Markov è la previsione delle probabili sequenze operative di un processo. Ciò implica la conoscenza di tutte le probabilità di transizione da un generico stato di partenza ***i*** al passo ***m*** (cioè all'istante discreto t_m) fino allo stato di arrivo ***j*** al passo ***n*** (istante discreto t_n).

Teoria delle code

Generalità
La teoria delle code ha come oggetto di studio i processi stocastici di formazione delle file di attesa da parte di utenti che si dispongono ordinatamente in attesa di ricevere un servizio erogato dal centro di servizio quando, per motivi aleatori, questo non può evadere immediatamente le richieste di servizio degli utenti.

Se pensiamo alle code alle casse di un supermercato, o agli sportelli di un ufficio postale o ancora ai caselli autostradali, non è sempre possibile prevedere con precisione quando si dovrà provvedere al servizio e/o quanto tempo sarà necessario per effettuarlo. La teoria delle code non risolve direttamente, da sola, questo problema ma fornisce importanti dati necessari per una simile decisione prevedendo varie caratteristiche della linea di attesa come il tempo medio di attesa.

Il processo tipico della coda
Il processo fondamentale usato dalla maggior parte dei modelli di code è il seguente. Le unità che hanno bisogno di servizio generalmente appartengono ad una popolazione di probabili utenti.
Queste unità entrano nel sistema e formano una *coda*. Ad un certo punto, un componente della coda viene scelto per usufruire del servizio secondo certe regole conosciute come la *disciplina del servizio*. Il servizio richiesto da un'unità è eseguito dalla *stazione di servizio*, dopo che l'unità ha lasciato il sistema delle code.
La rappresentazione grafica del processo è la seguente:



I sistemi di file d'attesa (sistemi di congestione) sono caratterizzati da un arrivo casuale di clienti, ciascun richiedente un'operazione (servizio) ad un'apposita unità/stazione di servizio. Anche quando l'afflusso di clienti non sia in media superiore alla massima capacità di smaltimento dell'unità di servizio, a causa dell'aleatorietà dei fenomeni coinvolti, si avrà la formazione di una fila d'attesa, in cui si disporranno i clienti non ancora serviti.

Gli elementi che permettono di definire completamente il fenomeno d'attesa sono:
- la popolazione dei clienti
- il processo d'arrivo
- la coda (in senso stretto)
- i servitori
- il processo di servizio
- la disciplina di servizio.

La **popolazione** è l'insieme dei potenziali clienti, ovvero l'insieme da cui arrivano i clienti e a cui tornano dopo essere stati serviti. Essa può essere finita o infinita. Nel primo caso le modalità di arrivo dei clienti dipendono dal numero di loro correntemente nel sistema.

Il **processo d'arrivo,** che descrive il modo secondo cui i clienti si presentano, è in generale un processo stocastico. Esso è definito in termini della distribuzione dell'**intertempo d'arrivo**, cioè dell'intervallo di tempo che intercorre tra l'arrivo di due clienti successivi.
Per ottenere modelli analiticamente trattabili di solito si assume che sia il processo di arrivo che quello di servizio siano **stazionari**, ovvero che le loro proprietà statistiche non varino nel tempo.

La **coda** (in senso stretto) è formata dai clienti presenti nel buffer in attesa di essere serviti.
La capacità del buffer può essere infinita o finita. Nel secondo caso essa limita di conseguenza la **capacità del sistema,** cioè il numero dei clienti in attesa nel buffer più quelli che correntemente sono serviti. I clienti che arrivano dopo che sia saturata quest'ultima capacità sono respinti. Ad esempio ha capacità di sistema limitata un centralino telefonico che può tenere in attesa solo un numero finito di chiamate. In assenza di centralino la dimensione della coda è addirittura zero, di conseguenza una chiamata o è servita immediatamente o è rifiutata.

I **servitori** sono in numero noto e costante fissato a livello di progetto. Usualmente essi hanno caratteristiche identiche, possono sempre lavorare in parallelo, viceversa non possono mai rimanere inattivi in presenza di clienti in coda. Anche se vi sono di più servitori in una coda in generale si assume l'esistenza di un unico buffer comune, quando infatti ogni servitore ha il suo buffer separato si preferisce pensare ad un insieme di code. Può però essere comodo introdurre, almeno logicamente, più buffer in presenza di clienti provenienti da popolazioni diverse.

Il **processo dei servizi** descrive il modo secondo cui ciascun servitore eroga il servizio, in particolare definisce la durata dello stesso ed è di solito un processo stocastico. Esso è definito in termini delle distribuzioni dei **tempi di servizio** dei diversi servitori**.** Il processo dei servizi è alimentato dal processo d'arrivo. Conseguentemente il processo d'arrivo è indipendente e condiziona il processo dei servizi. Un cliente, infatti, può essere servito solo se è già arrivato.
Quando non c'è nessuno, il servitore è inattivo e quindi non può avvantaggiarsi in vista d'impegni futuri. In altre parole un servitore non può servire in anticipo clienti non ancora arrivati. Non può esistere una coda negativa.

La **disciplina di servizio** specifica quale sarà il prossimo cliente servito fra quelli in attesa al momento in cui si libera un servitore. Le discipline di servizio usualmente considerate, poiché sia molto comuni nella realtà che matematicamente trattabili, sono:

- servizio in ordine di arrivo **FCFS** (first-come first-served) o **FIFO** (first-in first-out);
- servizio in ordine inverso di arrivo **LCFS** (lastcome first-served) o **LIFO** (last-in first-out);
- servizio in ordine casuale **SIRO** (service in random order);
- servizio basato su classi di **priorità** (vedi centri di emergenza quali il pronto soccorso).

Per determinare la capacità di smaltimento dell'unità di servizio si deve tenere conto dei parametri che caratterizzano una fila di attesa, che sono:

- ***ta*** intervallo di tempo tra gli arrivi
- ***ts*** tempo di servizio
- ***tq*** tempo complessivo speso dal generico cliente nella coda prima di essere servito
- ***tw*** tempo complessivo speso dal generico cliente nel sistema
- ***s*** numero di serventi
- ***n*** numero di clienti nel sistema (stato del sistema)

ma anche:

- la ***disciplina di servizio*** legge secondo la quale i clienti in fila d'attesa vengono serviti
- la ***dimensione popolazione di utenti***
- il ***comportamento del cliente dopo il servizio***
- ***l*** lunghezza massima della coda, cioè massimo numeri di utenti in attesa.

Quindi un sistema di file d'attesa è costituito dalla combinazione di due processi stocastici: uno di arrivi, caratterizzato da ***ta***, ed uno di servizio caratterizzato da ***ts***.
Il processo degli arrivi/uscite può dipendere dal numero ***n*** di clienti presenti nel sistema.

Sia ***s*** il numero dei serventi, la lunghezza ***l*** della coda è evidentemente data dallo stato del sistema meno il numero di clienti in fase di servizio. Si ha cioè:

$$\begin{cases} 0 & \text{se } n \le s \\ n\text{-}s & \text{se } n \ge s \end{cases}$$

Inoltre, si può notare che il tempo complessivo speso nel sistema da un cliente è pari a ***tw = tq - ts***

Un'importanza particolare assumono spesso i valori attesi dell'intervallo di tempo tra gli arrivi **E{*ta*}** e del tempo di servizio **E{*ts*}**.
Di solito si usa introdurre una *frequenza media di arrivi* λ ed una *velocità di servizio* ν :

$$\lambda = \frac{1}{E\{t_a\}} \qquad \nu = \frac{1}{E\{t_s\}}$$

Se ***s* = 1** si definisce inoltre *fattore di utilizzazione* **ρ** il rapporto

$$\rho = \frac{E\{t_s\}}{E\{t_a\}} = \frac{\lambda}{\mu}$$

dove **μ** rappresenta la frequenza medie delle partenze.
Tale rapporto può essere interpretato come la frazione di tempo in cui il servente è occupato.

Posto in questi termini **ρ** fornisce una chiara indicazione del grado di congestione, infatti:

- se ρ =1 gli utenti entranti e quelli serviti si equivalgono;
- se ρ >1 il flusso degli arrivi è maggiore del flusso che il servente (o i serventi) riesce a smaltire e si ha pertanto una coda crescente;
- se ρ <1 il servente è in grado di smaltire tutte le richieste e gode di periodi di inattività; non si ha formazione di code.

Per casi come quest'ultimo è stato coniato un indice chiamato coefficiente di utilizzo dei serventi ρ_u che esprime la frazione temporale durante la quale il servente è attivo:

$$P_u = \frac{\lambda}{\mu}$$

Nel caso in cui ***s* > 1**, evidentemente, per poter definire il fattore di utilizzazione del generico servente occorre porre attenzione al fatto che la frequenza degli arrivi al singolo servente certamente non coincide con la frequenza degli arrivi all'unità di servizio, ma sarà in generale un'opportuna frazione di questa.

Consideriamo il casello autostradale e supponiamo che il servente sia un grado di smaltire $\mu = 10$ automobili al minuto e giungano al casello $\lambda = 5$ automobili al minuto. Il servente è perciò occupato solo metà del tempo e infatti risulta:

$$P_u = \frac{(1/2)\,\mu}{\mu} = \frac{1}{2}$$

Se poi sono presenti s casellanti, il tempo di inattività si moltiplica, ovvero:

$$P_u = \frac{\lambda}{S * \mu}$$

La situazione opposta si presenta quando $\lambda > s*\mu$, cioè quando gli arrivi sono più frequenti delle partenze.
In questo caso il valore assunto da ρ_u sarebbe maggiore di 1 ma ciò non è ammissibile in quanto i serventi non possono essere impegnati più del 100%.
In tal caso la formula corretta è:

$$P_u = \min \left\{ \frac{\lambda}{S * \mu} , 1 \right\}$$

Un altro indice è il ***throughput* TH** che indica il numero medio degli utenti serviti nell'unità di tempo.

Se $\lambda \leq s * \mu$ allora tutti gli arrivi sono serviti, quindi **TH** = λ;
Se $\lambda > s * \mu$ allora vengono accolti solo μ arrivi da ciascun servente e quindi **TH = s * μ**

Quindi: **TH = min { λ, s * μ }**