큐잉 이론과 대기행렬

1. 큐잉 이론의 이론적 배경과 수식

1.1 포아송 분포와 지수 분포

- 포아송 분포와 지수 분포는 큐잉 이론에서 가장 기본적인 확률 분포입니다.
- 포아송 분포: 특정 시간 내에 일어나는 사건의 수를 나타내는 확률 분포로, 도착률 λ 를 따르는 고객의 도착 패턴을 설명합니다.
- 공식: $P(X = k) = \frac{(\lambda t)^k e^{(-\lambda t)}}{k!}$

여기서 X는 시간 t 동안의 고객 도착 수이며, λ는 평균 도착률입니다.

- 지수 분포: 두 사건 사이의 시간 간격을 나타내며, 서비스 시간이나 도착 간격을 모델링할 때 사용합니다.
- 공식: $f(t) = \lambda e^{(-\lambda t)}$

여기서 λ 는 평균 도착률 또는 서비스율입니다.

1.2 $M/M/1$ 모델 수식과 이론적 배경

$M/M/1$ 모델은 가장 단순하면서도 중요한 큐잉 모델로, 다음과 같은 가정을 따릅니다:

- 도착 시간과 서비스 시간이 모두 포아송 분포를 따름 (Markovian)
- 하나의 서버가 고객을 처리함
- 대기열 길이는 무한함.

주요 수식

1 시스템 이용률(ρ):

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

여기서 λ 는 도착률, μ 는 서비스율을 나타냅니다. $\rho < 1$ 일 때 시스템이 안정적임을 의미합니다.

2 평균 시스템 내 고객 수 (L):

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

이는 도착률이 서비스율에 가까울수록 시스템 내 고객 수가 증가함을 의미합니다.

3 평균 대기 시간 (W_q):

$$W_q = \frac{\lambda}{(\mu (\mu - \lambda))}$$

고객이 대기열에서 대기하는 평균 시간을 계산합니다.

4 평균 시스템 체류 시간 (W):

$$W = \frac{1}{(\mu - \lambda)}$$

고객이 시스템 전체에서 소비하는 총 시간입니다.

1.3 $M/M/c$ 모델 수식과 이론적 배경

이 모델은 c개의 서버가 있는 상황을 모델링합니다.

1 P_0 (시스템이 비어있는 확률):

$$P_0 = \left[\sum \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} \frac{1}{1-\rho}\right]^{-1}$$

2 평균 대기 고객 수 (L_q):

$$L_q = \frac{(\lambda/\mu)^c \rho}{(c!\,(1-\rho)^2)} P_0$$

이는 $M/M/1$ 모델과는 달리, 여러 서버가 있을 때의 대기 고객 수를 정확하게 계산합니다.

3 평균 시스템 내 고객 수 (L):

$$L = Lq + \frac{\lambda}{\mu}$$

이는 대기열에 있는 고객과 이미 서비스를 받고 있는 고객을 합친 수입니다.

1.4 $M/G/1$ 모델 이론과 Pollaczek-Khinchine 공식

$M/G/1$ 모델은 서비스 시간이 임의의 일반 분포를 따를 때 사용되는 모델입니다.

평균 대기 고객 수 (L_q):

$$L_q = \frac{(\lambda^2 E[S^2])}{\left(2(1-\rho)\right)}$$

여기서 $E[S^2]$는 서비스 시간의 분산을 의미하며, 시스템 변동성의 영향을 나타냅니다.

2. 실제 사례를 통한 적용 예시

2.1 슈퍼마켓 계산대 ($M/M/c$ 모델)

상황 : 슈퍼마켓에 3 개의 계산대가 있으며, 고객은 평균적으로 1 분마다 도착하고, 각 계산대는 평균적으로 3 분마다 고객을 처리합니다.

분석:

- λ = 1 (고객/분), μ = 1/3 (고객/분), c = 3
- 시스템 이용률: $\rho = \frac{\lambda}{c\mu} = \frac{1}{3} = 0.333$
- 안정적인 시스템이며, 공식에 따라 $P_0, L_q, \boldsymbol{L}, W_q, \boldsymbol{W}$를 계산할 수 있습니다.

2.2 콜센터 운영 ($M/G/1$ 모델)

상황 : 콜센터에 상담원 1 명이 있고, 전화는 2 분마다 평균적으로 걸려오며, 상담 시간의 평균은 4 분, 분산은 16 분입니다.

분석:

- $\lambda = 0.5$ (통화/분), $\mu\ =\ 0.25$ (통화/분), 서비스 시간 분산 $[S^2] = 16E$
- $L_q = \frac{\lambda^2 E[S^2]}{2(1-\rho)} = \frac{(0.5)^2 \times 16}{2(1-0.5)} = 4$
- 이를 통해 평균적으로 4 명의 고객이 대기한다는 것을 알 수 있습니다.

3. 큐잉 이론의 실제적 활용과 의의

- **효율적인 자원 배치**: 은행, 병원, 콜센터 등에서 대기 시간을 줄이고 효율적으로 인력을 배치하는 데 사용됩니다.
- **제조 공정 최적화**: 생산 라인의 병목 현상을 줄이고 생산 속도를 최적화합니다.
- **IT 시스템**: 네트워크 대역폭 관리, 서버의 부하 균형을 유지하는 데 활용됩니다.