Simulación a Eventos Discretos

Tema 4: Modelado estocástico (parte 3)

Objetivo

- Nos interesa modelar las características estocásticas de los sistemas.
- Utilizamos variables aleatorias con su correspondiente distribución de probabilidad.
- Por ejemplo: la cantidad de arribos en un determinado período de tiempo (o el tiempo entre arribos), la duración de una actividad.

Datos

Necesitamos datos para construir las distribuciones adecuadas.

- Si se dispone de datos existentes, debemos determinar la mejor forma de utilizarlos; es importante conocer cómo fueron recolectados.

- Si es posible recolectarlos, es imprescindible definir claramente los requerimientos.

- Ante ausencia de datos, debemos establecer hipótesis adecuadas.

Datos

¿Para qué se usan?

- Reproducir en la simulación la situación observada (*trace-driven simulation*).
- Definir una *distribución empírica de probabilidad*.
- Ajustar una *distribución teórica paramétrica* (Poisson, Normal, etc.).

Recomendaciones

- Usar trace-driven simulation para validar los modelos.
- Usar una distribución empírica cuando no es posible ajustar a alguna distribución teórica.
- Siempre tratar de usar una distribución teórica, porque: elimina irregularidades en los datos, permite generación de valores extremos (fuera del rango de los datos observados) y tiene una representación compacta.

Distribuciones teóricas

Debemos determinar los valores de sus parámetros de forma de ajustar a los datos.

Distribuciones continuas: tienen parámetros de ubicación, escala y forma (Exponencial negativa, Normal, Weibull, etc.).

Distribuciones discretas: Poisson, Binomial, etc.

Distribuciones teóricas

Proceso de ajuste:

1. Verificación de independencia de los datos.
2. Hipótesis acerca de familias de distribuciones.
3. Estimación de parámetros.
4. Determinación de la bondad del ajuste.

En este curso asumimos 1 y 2 dados; realizamos 3 y 4.

Verificación de la independencia de los datos

- Es un aspecto importante, dado que la independencia de los datos es una hipótesis asumida por métodos de estimación de parámetros ($máxima\ verosimilitud$) y de determinación de la bondad del ajuste ($Chi\text{-}cuadrado$).

- Métodos: gráficos de correlación y de dispersión (ver Law, 2015).

Hipótesis acerca de familias de distribuciones

- Implica determinar la forma de la curva, sin preocuparse (en esta etapa) por los parámetros específicos.

- Se utilizan diferentes estadísticos de las distribuciones, calculados a partir de los datos: mínimo, máximo, media, mediana, varianza, coeficiente de variación (ver Law, 2015).

Hipótesis acerca de familias de distribuciones

Gráficamente se utilizan histogramas (para distribuciones continuas) y gráficos de líneas (para distribuciones discretas). Por ejemplo:

Familias de distribuciones: algunas aplicaciones

- Binomial: Número de piezas defectuosas en un lote de tamaño dado.
- Poisson: Cantidad de arribos de clientes en un intervalo de tiempo dado.
- Normal: Proceso resultante de la suma de varios otros.
- Exponencial: Tiempo entre eventos, procesos sin memoria.
- Uniforme: Incertidumbre completa, todos los valores tienen la misma probabilidad.
- Triangular: Solo se conocen los valores mínimo, máximo y más probable.

Estimación de parámetros

Observaciones: $X_1, \ldots X_n$

Función de probabilidad: $f(x)$

Parámetro(s): θ

Función de probabilidad paramétrica: $f_\theta(x)$

Objetivo: encontrar el valor de θ en función de $X_1, \ldots X_n$ que mejor ajusta la distribución a las observaciones.

Estimación de parámetros

Estimadores de *máxima verosimilitud* (MLE, siglas en inglés): determinan el valor de θ asumiendo que se observaron los datos $X_1, \ldots X_n$ porque son los más probables.

La función de verosimilitud se define como:

$$V(\theta) = f_\theta(X_1) f_\theta(X_2) \ldots f_\theta(X_n)$$

Queremos maximizar $V(\theta) \Rightarrow dV(\theta)/d\theta = 0$

Estimación de parámetros

Ejemplo: el estimador de máxima verosimilitud para la distribución exponencial negativa se obtiene se la siguiente forma:

$V(\lambda) = 1/\lambda \exp(-X_1/\lambda) \ldots 1/\lambda \exp(-X_n/\lambda)$

$\ln(V) = -n\ln(\lambda) - (1/\lambda) \sum_{i=1..n} X_i$

$dV/d\lambda = 0$ y $d^2V/d\lambda^2 < 0 \Rightarrow \lambda = \sum_{i=1..n} X_i/n$

Otras distribuciones requieren la aplicación de métodos numéricos para la maximización de la función de verosimilitud $V(\theta)$.

Estimación de parámetros

Estimadores de máxima verosimilitud para otras distribuciones:

Distribución	Función de probabilidad	MLE $\hat{\mu}$	MLE $\hat{\sigma}$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x-\mu)^2/2\sigma^2\right)$	$\frac{\sum_{i=1}^{n} X_i}{n}$	$\sqrt{\frac{\sum_{i=1}^{n}(X_i-\hat{\mu})^2}{n}}$
Lognormal	$\frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-(lnx-\mu)^2/2\sigma^2\right)$	$\frac{\sum_{i=1}^{n} lnX_i}{n}$	$\sqrt{\frac{\sum_{i=1}^{n}(lnX_i-\hat{\mu})^2}{n}}$

Estimación de parámetros

Objetivo: determinar la "calidad" del ajuste de una distribución a los datos.

Procedimientos heurísticos: comparación de frecuencias, gráficos de probabilidad (ver Law, 2015).

Test de bondad de ajuste: test de hipótesis donde

$H_0 = \{$las X_i son muestras IID de una distribución $F\}$

Por ejemplo: Chi-cuadrado, Kolmogorov-Smirnov.

Test Chi-cuadrado

Los datos $X_1, \ldots X_n$ se dividen en k categorías.

O_j: cantidad de datos $observados$ en la categoría j.
E_j: cantidad de datos $esperados$ en la categoría j, según la distribución considerada. $E_j = nP_j$ donde P_j es la probabilidad teórica de la categoría j.

El estadístico

$$\chi^2 = \sum_{j=1..k} (O_j - E_j)^2 / E_j$$

tiene una distribución Chi-cuadrado con $k - 1$ grados de libertad.

Test Chi-cuadrado

Siendo $\chi^2_{k-1,1-\alpha}$ el valor de la tabla de la distribución Chi-cuadrado.

- Si $\chi^2 > \chi^2_{k-1,1-\alpha}$: rechazar H_0
- Si $\chi^2 \leq \chi^2_{k-1,1-\alpha}$: no rechazar H_0

donde α es la probabilidad de cometer el error de rechazar H_0 cuando es verdadera (error de Tipo I).

Test Chi-cuadrado

Recomendaciones:

- Construir categorías equiprobables.
- Asegurar que $nP_j \geq 5 \ \forall j$ en $1..k$.
- No olvidar las "colas" de las distribuciones.

Observaciones:

- El test tiende a no rechazar H_0 cuando hay pocos datos (n pequeño) y a rechazar H_0 cuando hay muchos datos (n grande).

Distribuciones empíricas

Cuando no es posible encontrar un patrón según una distribución teórica conocida (pero hay datos), entonces se puede utilizar una distribución empírica.

Para eso se construye una tabla de k pares de valores $(x_j, F(x_j))$, donde F es la distribución (acumulada) de observaciones (F monótona creciente con j, $F \leq 1$).

Luego se muestrean valores de la distribución F utilizando el método de la transformación inversa (ver clase anterior).

Ausencia de datos

Cuando no hay datos (por ejemplo, si el modelo está en una etapa de demostración preliminar de un proyecto), la entrada de la simulación debe configurarse con otro enfoque.

Fuentes alternativas: especificación de productos, juicio de expertos, limitaciones físicas, naturaleza del proceso.

Distribuciones más usadas cuando no hay datos o son muy escasos: uniforme, triangular, beta.