Preparación de los datos de entrada

Clase nro. 6

CURSO 2010

Objetivo

- **Modelado de las características estocásticas de los sistemas.**
- **Variables aleatorias con su distribución de probabilidad.**
- **Por ejemplo:**
  - **cantidad de arribos en un determinado período de tiempo (o tiempos entre arribos),**
  - **duración de una actividad.**

Datos

- **Si se dispone de datos existentes: determinar la mejor manera de utilizarlos; importante conocer cómo fueron recolectados.**
- **Si es posible recolectarlos: definir bien los requerimientos.**
- **Ante ausencia de datos: establecer hipótesis adecuadas.**

Uso de los datos

1. **Reproducir la situación observada (*trace-driven simulation*).**
2. **Definir una distribución empírica de probabilidad (*muestreo de histograma*).**
3. **Ajustar una distribución teórica (*por ejemplo: Poisson, Normal, etc*.).**

Recomendaciones

- **Usar** *t-driven sim* **para validar los modelos.**
- **Usar** *distr. empírica* **cuando no es posible ajustar a alguna distribución teórica.**
- **Siempre tratar de usar** *distr.teórica* **porque:**
  - **Elimina "irregularidades" en los datos,**
  - **Permite generación de valores extremos (fuera del rango de datos observados),**
  - **Representación compacta.**

Distribuciones teóricas

- **Determinar valores de sus parámetros, de forma de ajustar a los datos.**
- **Distribuciones contínuas: 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 nos centramos en 3 y principalmente 4

Distribuciones teóricas

- **Verificación de independencia de los datos:**
  - **<u>Independencia</u>: hipótesis asumida por métodos de estimación de parámetros (*máxima verosimilitud*) y de determinación de la bondad del ajuste (*Chi-Cuadrado*).**
  - **Métodos: gráficos de correlación y de dispersión.**

Distribuciones teóricas

- **Hipótesis acerca de familias de distribuciones:**
  - **Determinar la forma de la curva, sin preocuparse (en esta etapa) por los parámetros específicos.**
  - **Estadísticos (de las distribuciones, a partir de los datos): mínimo, máximo, media, mediana, varianza, coeficiente de variación, lexis ratio, skewness.**
  - **Histogramas (distribuciones continuas) y gráficos de líneas (distribuciones discretas).**

S.E.D. 2010

Distribuciones teóricas
Estimación de parámetros

Observaciones: $X_1,\ldots,X_n$

Distribución de probabilidad (densidad): $f(x)$

Parámetro(s): θ

Distribución paramétrica: $f_\theta(x)$

<u>Objetivo</u>: hallar el valor de θ en función de $X_1,\ldots,X_n$, que mejor ajusta la distribución a las observaciones.

S.E.D. 2010

Distribuciones teóricas
Estimación de parámetros

Estimadores de máxima verosimilitud:

Determinan el valor de θ asumiendo que se observaron los datos $X_1,\ldots,X_n$ porque son los más probables.

Función de verosimilitud $V(\theta)$ se define como:

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

Maximización de $V(\theta)$:

$$dV(\theta)/d\theta = 0$$

S.E.D. 2010

Distribuciones teóricas
Estimación de parámetros

Ejemplo: Estimador de máxima verosimilitud para distribución exponencial negativa

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

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

$dV/d\lambda = 0$ y $d^2V/d\lambda^2 < 0 \Rightarrow$

$$\lambda = \Sigma_{i=1..n} X_i/n$$

S.E.D. 2010

Distribuciones teóricas

Estimación de parámetros

Otras distribuciones:

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

S.E.D. 2010

Distribuciones teóricas

Bondad del ajuste

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

Procedimientos heurísticos: comparación de frecuencias, gráficos de probabilidad.

Tests 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.

S.E.D. 2010

Distribuciones teóricas

Test Chi-Cuadrado

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,
$E_j = nP_j$, donde P_j es la probabilidad teórica de la categoría j
El estadístico

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

tiene una distribución Chi-Cuadrado con k-1 grados de libertad

S.E.D. 2010

Distribuciones teóricas

Test Chi-Cuadrado

$\chi^2_{k\text{-}1,1\text{-}\alpha}$: valor de la tabla de la distribución Chi-Cuadrado

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

α: probabilidad de cometer error de rechazar H_0 cuando es verdadera (error de Tipo I)

S.E.D. 2010

Distribuciones teóricas

Test Chi-Cuadrado

Recomendaciones:

- Construir categorías equiprobables
- Asegurar que $nP_j \geq 5 \quad \forall$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, entonces se utiliza una distribución empírica.

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, $F \leq 1$).
Se muestrean valores de la distribución F, de forma análoga al método de distribución discreta.

Distribuciones empíricas

Algoritmo de muestreo:

u = valor aleatorio uniforme en (0,1);

$x = 1$;

While $u > F(x)$ do

$x = x + 1$;

Caso discreto:

retornar $x - 1$

Caso contínuo:

retornar $x - 1 + (u\text{-}F(x\text{-}1))/(F(x)\text{-}F(x\text{-}1))$

S.E.D. 2010