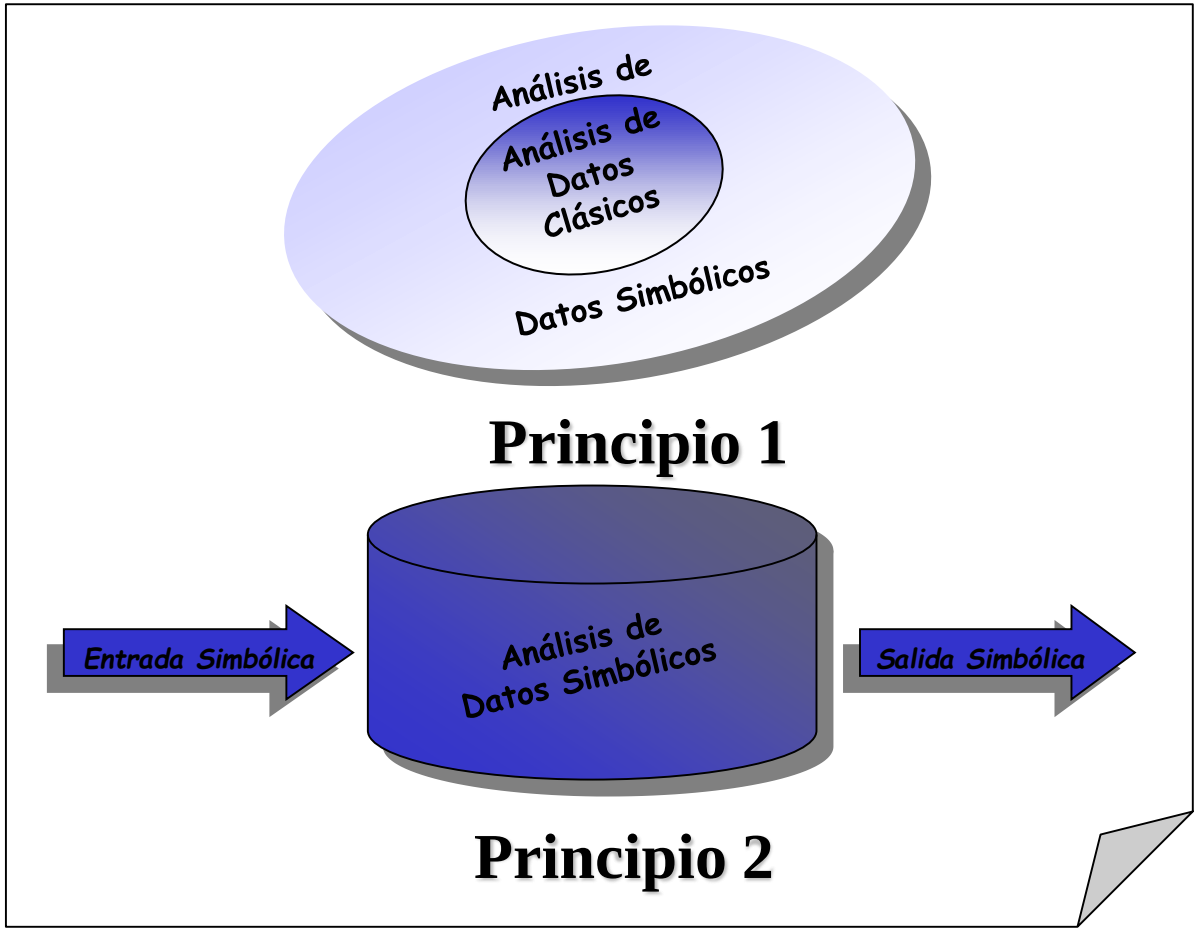
PR0MiDAT
IBEROAMERICANO
Programa Iberoamericano de Formación en Minería de Datos

Estadísticas Básicas sobre Datos Simbólicos

Principios del ADS:



La Media Simbólica

En análisis de datos clásico la media se define como sigue. Sea Y una variable cuantitativa y sea $y_1, y_2, \ldots, y_m$ los m valores observados de esta variable, entonces la media de Y es:

$$\overline{Y} = \frac{1}{m} \sum_{i=1}^{m} y_i.$$

Bertrand y Goupil en (([7, Bock and Diday (2000)])) generalizaron esta definición para datos de tipo intervalo como sigue:

$$\overline{Y} = \frac{1}{m} \sum_{i=1}^{m} \frac{\underline{y}_i + \overline{y}_i}{2},$$

con $Y(i) = [\underline{y}_i, \overline{y}_i]$, $i = 1, 2, \ldots, m$.

La media simbólica:

Esta definición de media dada por Bertrand y Goupil tiene como entrada una variable de tipo intervalo, sin embargo, la salida es un valor numérico, esto tiene en gran inconveniente de que no refleja la variación que podría existir en el resultado. Por ejemplo, si $E = \{1, 2, 3, 4\}$, $Y(E) = \{[1, 2], [-1, 4], [2, 3], [-1, 1]\}$ y $Z(E) = \{[0, 3], [-2, 5], [1, 4], [-2, 2]\}$ entonces ambas variables tienen la misma media ($\overline{Y} = \overline{Z} = 1{,}375$), aún cuando es claro que la variable Z tiene mucho mayor variación que la variable Y.

Por esta razón en [Rodríguez 2000] se define la Media de una variable de tipo intervalo como intervalo que refleja la mínima y la máxima media posible dada la variación posible en los datos.

Definition 1 *Sea Y una variable de tipo intervalo definida en $E = \{1, 2, \ldots, m\}$ por $Y = \{[\underline{y}_1, \overline{y}_1], [\underline{y}_2, \overline{y}_2], \ldots, [\underline{y}_m, \overline{y}_m]\}$ entonces la media simbólica se define por:*

$$\overline{Y} = \left[\frac{1}{m} \sum_{i=1}^{m} \underline{y}_i, \frac{1}{m} \sum_{i=1}^{m} \overline{y}_i \right]. \tag{1.3}$$

Con esta definición, en ejemplo anterior, si $E = \{1, 2, 3, 4\}$, $Y(E) = \{[1, 2], [-1, 4], [2, 3][-1, 1]\}$ y $Z(E) = \{[0, 3], [-2, 5], [1, 4], [-2, 2]\}$ entonces ambas variables tienen diferente media $\overline{Y} = [0{,}25, 2{,}5]$ y $\overline{Z} = [-0{,}75, 3{,}5]$.

La mediana simbólica

En análisis de datos clásico la Mediana es el valor que está en el centro de los datos cuando estos están ordenados, es decir, 50 % de los datos son más grandes que la mediana y 50 % son más pequeños que la mediana.

Formalmente si se tienen m valores $y_1, y_2, \ldots, y_m$ para una variable cuantitativa Y y se supone que estos valores están ordenados, entonces el valor de la mediana depende de si el número m es par o impar, como sigue:

- Si m es impar, entonces la mediana está en la posición $\frac{m+1}{2}$ que es exactamente la posición que separa los datos en dos grupos del mismo tamaño.
- Si m es par, entonces la mediana está entre la posición $\frac{m}{2}$ y la posición $\frac{m}{2}+1$ de tal forma que divide los datos en dos grupos con la misma cantidad de elementos, con $\frac{m}{2}$ elementos cada uno. En este caso la mediana se define como la media entre los datos $y_{\frac{m}{2}}$ y $y_{\frac{m}{2}+1}$, es decir, $\mathrm{Me}(Y)=\frac{y_{\frac{m}{2}}+y_{\frac{m}{2}+1}}{2}$.

La mediana simbólica

Definition 2 *Sea Y una variable de tipo intervalo definida en $E = \{1, 2, \ldots, m\}$ por $Y = \{[\underline{y}_1, \overline{y}_1], [\underline{y}_2, \overline{y}_2], \ldots, [\underline{y}_m, \overline{y}_m]\}$, entonces la mediana simbólica se define como:*

$$Me(Y) = \left[\underline{Me}, \overline{Me}\right], \tag{1.4}$$

donde $\underline{Me}$ la media clásica de $\{\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_m\}$ and $\overline{Me}$ es la media clásica de $\{\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_m\}$.

Ejemplo:

Y Pulse Rate	Y_1 Systolic Pressure	Y_2 Diastolic Pressure
$[44, 68]$	$[90, 100]$	$[50, 70]$
$[62, 72]$	$[90, 130]$	$[70, 90]$
$[56, 90]$	$[140, 180]$	$[90, 100]$
$[70, 112]$	$[110, 142]$	$[80, 108]$
$[54, 72]$	$[90, 100]$	$[50, 70]$
$[70, 100]$	$[130, 160]$	$[80, 110]$
$[63, 75]$	$[60, 100]$	$[140, 150]$
$[72, 100]$	$[130, 160]$	$[76, 90]$
$[76, 98]$	$[110, 190]$	$[70, 110]$
$[86, 96]$	$[138, 188]$	$[90, 110]$
$[86, 100]$	$[110, 150]$	$[78, 100]$

Cuadro 1.1: Ejemplo de once pacientes.

Las medianas simbólicas de estas tres variables son: $\text{Me}(Y) = [70, 97]$, $\text{Me}(Y_1) = [110, 146]$ and $\text{Me}(Y_2) = [77, 100]$.

La varianza y la desviación estándar

En análisis de datos clásicos la medida de dispersión más usada es la desviación estándar.

Sea Y la variable cualitativa y sea $y_1, y_2, \ldots, y_m$ los m los valores observados de esta variable, entonces la desviación estándar de Y se define como:

$$\sigma_Y = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \overline{Y})^2}$$

La varianza de Y se define como:

$$Var(Y) = \sigma_Y^2$$

La varianza y la desviación estándar

Bertrand y Goupil generalizaron la definición anterior para variables de tipo intervalo como sigue:

$$\sigma_Y = \sqrt{\frac{1}{3m}\sum_{i=1}^{m}(\overline{y}_i + \underline{y}_i)^2 - \frac{1}{4m^2}\left(\sum_{i=1}^{m}(\overline{y}_i + \underline{y}_i)\right)^2}$$

donde $Y(i) = [\underline{y}_i, \overline{y}_i]$ para $i = 1, 2, \ldots, m$ (ver [7, Bock and Diday (2000)]).

Sin embargo, esta definición tiene el mismo problema que el mencionado en definiciones anteriores, por ejemplo, si $E = \{1, 2, 3, 4\}$ y $Y(E) = \{[1, 2], [-1, 4], [2, 3], [-1, 1]\}$, $Z(E) = \{[0, 3], [-2, 5], [1, 4], [-2, 2]\}$ entonces ambas variables tienen la misma desviación estándar, es decir, $\sigma_Y = \sigma_Z = 0{,}892$, aún cuando la variable Z tiene mucho mayor variación que la variable Y.

La varianza y la desviación estándar

Para evitar este problema, para variables de intervalo debemos definir la varianza de manera tal que este índice mida cuán distantes son los datos de la media de tipo intervalo, pero de tal manera que esta variación sea también un intervalo que contemple la varianza mínima y máxima posibles.

Definition 3 *Sea Y una variable de tipo intervalo definida en $E = \{1, 2, \ldots, m\}$ por $Y = \{[\underline{y}_1, \overline{y}_1], [\underline{y}_2, \overline{y}_2], \ldots, [\underline{y}_m, \overline{y}_m]\}$ y sea $\overline{Y} = [\alpha, \beta]$ entonces la desviación estándar para variables de tipo intervalo se define como:*

$$\sigma_Y = \left[\sqrt{\frac{1}{m}\sum_{i=1}^{m} \min_{\substack{x\in[\underline{y}_i,\overline{y}_i] \\ y\in[\alpha,\beta]}} (x-y)^2}, \sqrt{\frac{1}{m}\sum_{i=1}^{m} \max_{\substack{x\in[\underline{y}_i,\overline{y}_i] \\ y\in[\alpha,\beta]}} (x-y)^2} \right], \tag{1.5}$$

y la varianza para variables de tipo intervalo se define como:

$$Var(Y) = \left[\frac{1}{m}\sum_{i=1}^{m} \min_{\substack{x\in[\underline{y}_i,\overline{y}_i] \\ y\in[\alpha,\beta]}} (x-y)^2, \frac{1}{m}\sum_{i=1}^{m} \max_{\substack{x\in[\underline{y}_i,\overline{y}_i] \\ y\in[\alpha,\beta]}} (x-y)^2 \right].$$

Ejemplo:

Y Pulse Rate	Y_1 Systolic Pressure	Y_2 Diastolic Pressure
[44, 68]	[90, 100]	[50, 70]
[62, 72]	[90, 130]	[70, 90]
[56, 90]	[140, 180]	[90, 100]
[70, 112]	[110, 142]	[80, 108]
[54, 72]	[90, 100]	[50, 70]
[70, 100]	[130, 160]	[80, 110]
[63, 75]	[60, 100]	[140, 150]
[72, 100]	[130, 160]	[76, 90]
[76, 98]	[110, 190]	[70, 110]
[86, 96]	[138, 188]	[90, 110]
[86, 100]	[110, 150]	[78, 100]

Cuadro 1.1: Ejemplo de once pacientes.

EJEMPLO:

Usando los datos del Cuadro 1.1, se obtiene

$\sigma_Y = [0,35,55]$, $\sigma_{Y_1} = [4{,}68, 49{,}60]$ y $\sigma_{Y_2} = [1{,}52, 32{,}23]$.

La correlación clásica:

La definición clásica de correlación es la siguiente:

Definition 4 Sea $Y = (y_1, y_2, \ldots, y_m)$ y $X = (x_1, x_2, \ldots, x_m)$ dos variables cuantitativas aplicadas a m individuos, donde x_i y y_i son los valores tomados por estas variables X y Y para el individuo i respectivamente. Entonces:

- La variance de Y se define como:

$$\sigma_Y^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \overline{Y}).$$

- La covarianza entre las dos variables variables X y Yse define como:

$$Cov(X,Y) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{X})(y_i - \overline{Y}).$$

- La correlación entre estas dos variables variables X y Y se define como:

$$R(X,Y) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{x_i - \overline{X}}{\rho_X}\right)\left(\frac{y_i - \overline{Y}}{\rho_Y}\right) = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2\sigma_Y^2}}.$$

La correlación simbólica (Billard-Diday):

Emulando la definición anterior, la correlación para variables de tipo intervalo se define como sigue:

Definition 5 Sean $X = ([\underline{x}_1, \overline{x}_1], [\underline{x}_2, \overline{x}_2], \ldots, [\underline{x}_m, \overline{x}_m])$ and $Y = ([\underline{y}_1, \overline{y}_1], [\underline{y}_2, \overline{y}_2], \ldots, [\underline{y}_m, \overline{y}_m])$ dos variables de tipo intervalo.

- La varianza de Y se define como:

$$\sigma_Y^2 = \frac{1}{3m} \sum_{i=1}^{m} (\overline{y}_i + \underline{y}_i)^2 - \frac{1}{4m^2} \left(\sum_{i=1}^{m} (\overline{y}_i + \underline{y}_i) \right)^2. \tag{1.6}$$

- La covarianza entre X y Y se define como:

$$Cov(X,Y) = \sigma_{XY} = \frac{1}{4m} \sum_{i=1}^{m} \frac{(\overline{x}_i^2 - \underline{x}_i^2)(\overline{y}_i^2 - \underline{y}_i^2)}{(\overline{x}_i^2 - \underline{x}_i^2)(\overline{y}_i^2 - \underline{y}_i^2)} - \overline{X} \cdot \overline{Y}. \tag{1.7}$$

- La correlación entre Y_1 and Y_2 se define como [5, Billard and Diday (2000)]:

$$R(X,Y) = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}. \tag{1.8}$$

Nube de puntos tipo intervalo:

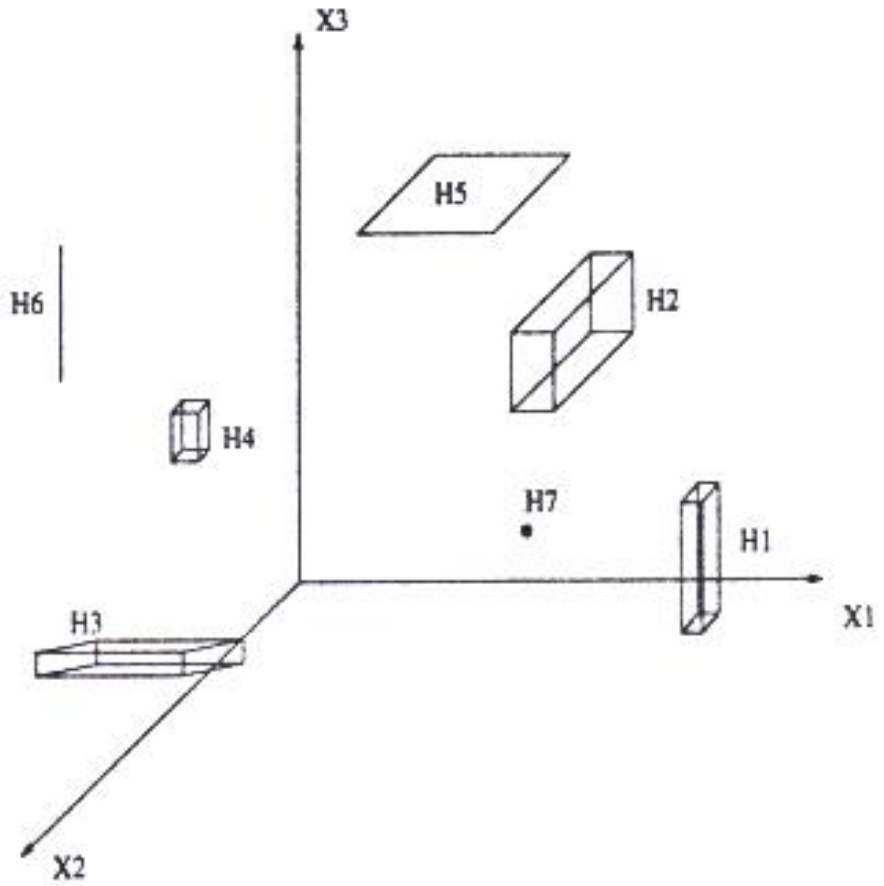


FIG. 1.1: *Nuage d'objets décrits par trois variables de type intervalle*

Nube de puntos tipo intervalo:

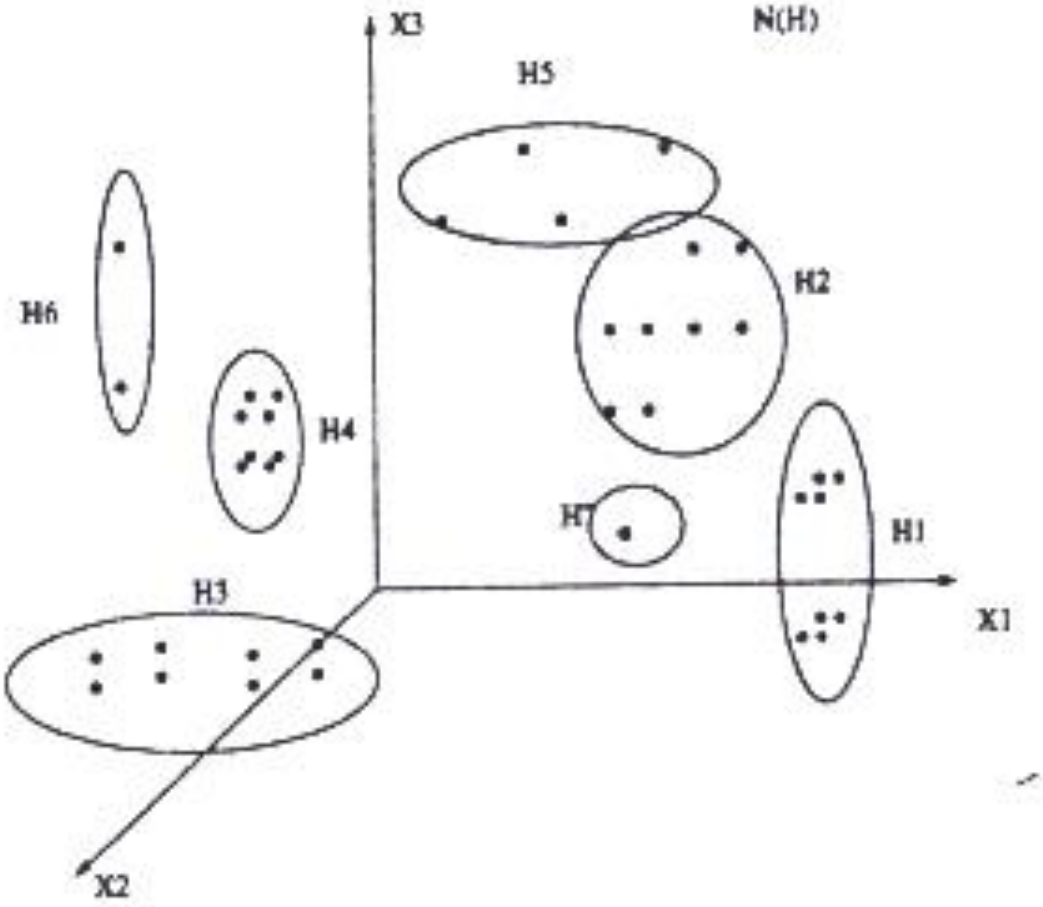


FIG. 1.2: *Nuage N(H) des sommets des hyper-rectangles*

PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

Gracias….