Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Septiembre, 2022

Representaciones de conjuntos y matrices (1)

- Para analizar datos nominales y ordinales, podemos definir relaciones entre pares de dichos datos, que pueden analizarse utilizando métodos relacionales específicos.
  - Por tanto, denotamos datos de características numéricas como el conjunto:

$$X = \{x_1, \dots x_n\} \subset \mathbb{R}^p$$

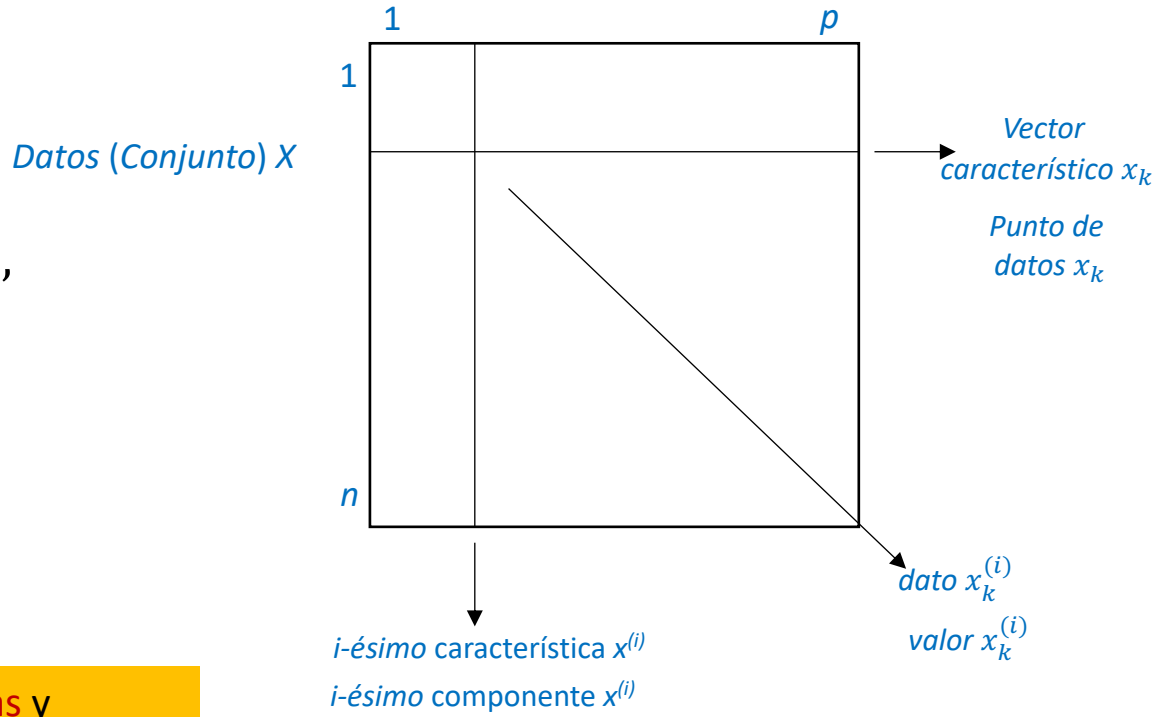
  - Con *n* elementos, donde cada elemento es un vector de características de valor real con *p*-dimensiones, donde *n* y *p* son números enteros positivos. Para *p* = 1, llamamos a *X* un conjunto de datos escalares. Como alternativa a la representación del conjunto, los datos de características numéricas también se representan a menudo como una matriz.

$$X = \begin{pmatrix} x_1^{(1)} & \cdots & x_1^{(p)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \cdots & x_n^{(p)} \end{pmatrix}$$

  - Así, los vectores $x_1, \dots x_n$ son vectores de fila, aunque matemáticamente los conjuntos de datos y las matrices de datos se usan comúnmente como representaciones de datos equivalentes.

Representaciones de conjuntos y matrices (2)

- En la siguiente figura se muestran los términos y notaciones comunes en la representación de una matriz de datos.
  - Cada fila de la matriz de datos corresponde a un elemento del conjunto de datos.
  - Se llama vector de características o punto de datos x_k, $k = 1, \dots, n$.
  - Cada columna de la matriz de datos corresponde a un componente de todos los elementos del conjunto de datos.
  - Se denomina *i*-ésima característica o *i*-ésima componente $x^{(i)}, i = 1, \dots, p$



Las filas y columnas se identifican usando: subíndices para filas y superíndices entre paréntesis para columnas. Las notaciones alternativas en la literatura son: $x(k,.)$ y $x(.,i)$. Un elemento único de la matriz se denomina componente de un elemento del conjunto de datos. Se llama dato o valor $x_k^{(i)}, k = 1, \dots, n; \quad i = 1, \dots, p$

Representaciones de conjuntos y matrices (3)

- El conjunto de datos de Iris se puede escribir como una matriz de datos con 150 filas y 4 columnas,
  - donde cada fila representa un objeto (flor) y cada columna representa una característica (dimensión).
  - La matriz de datos de Iris se puede obtener por concatenación vertical de las tres porciones que se muestran en la Tabla.
  - La información de clase (Setosa, Versicolor, Virginica) puede interpretarse como una quinta característica, en una escala nominal.

Setosa				Versicolor				Virginica			
Sepal		Petal		Sepal		Petal		Sepal		Petal	
Length	Width	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

Relaciones (1)

- Considérese un conjunto de elementos (abstractos), sin referirse a vectores de características numéricas.

$$O = \{o_1, \dots o_n\}$$

  - Algunas veces no hay una representación vectorial disponible para los objetos $o_k, k = 1, \dots, n$; por lo que los métodos convencionales de análisis de datos basados en características no son aplicables. En cambio, la relación de todos los pares de objetos a menudo se puede cuantificar y escribir como una matriz cuadrada.

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

  - Cada valor de relación $r_{ij},\ i, j = 1, \dots, n$, puede referirse a un grado de similitud, disimilitud, compatibilidad, incompatibilidad, proximidad o distancia entre el par de objetos o_i y o_j.
  - R puede ser simétrica, entonces $r_{ij} = r_{ji}$ para toda $i, j = 1, \dots, n$. R puede definirse manualmente o calcularse a partir de características.

Relaciones (2)

- Si las características numéricas X están disponibles, entonces R puede calcularse a partir de X usando una función apropiada $f: \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$.
  - Por ejemplo, un analista puede definir manualmente una matriz relacional para el conjunto de datos Iris, que compara visualmente y luego califica numéricamente alguna relación entre pares de flores, o R puede calcularse a partir de las longitudes y anchuras de sépalos y pétalos.
  - A continuación se presentan dos clases importantes de relaciones, diferencias y similitudes.

Medidas de disimilitud (1)

- Una función d se llama de disimilitud o medida de distancia si para toda $x, y \in \mathbb{R}^p$. Entonces tenemos los siguientes axiomas:
  - $d(x, y) = d(y, x)$
  - $d(x, y) = 0 \iff x = y$
  - $d(x, z) \leq d(x, y) + d(y, z)$
- De estos axiomas se tiene ahora que:
  - $d(x, y) \geq 0$
- Una clase de medidas de disimilitud se define usando una norma $\|\cdot\|$ de $x - y$, entonces:
  - $d(x, y) = \| x - y \|$

Medidas de disimilitud (2)

- Una función $\|\cdot\|: \mathbb{R}^p \rightarrow \mathbb{R}^+$ es una norma sí y solo sí:
  - $\| x \| = 0 \iff x = (0,0,\dots,0)$ (1)
  - $\| \alpha \cdot x \| = |\alpha| \cdot \| x \| \quad \forall \alpha \in \mathbb{R}, x \in \mathbb{R}^p$ (2)
  - $\| x + y \| \leq \| x \| + \| y \| \quad \forall x, y \in \mathbb{R}^p$ (3)

- Por ejemplo, la llamada norma hiperbólica de uso frecuente es:
  - $\| x \|_h = \prod_{i=1}^{p} x^{(i)}$

    - No es una norma según la definición anterior, ya que la condición (1) es violada por $x = (0,1) \neq (0,0)$ con $\| x \|_h = \| (0,1) \|_h = 0$, o la condición (2) se viola por $x = (1,1)$ y $\alpha = 2$, donde $\| \alpha \cdot x \|_h = \| 2 \cdot (1,1) \|_h = \| (2,2) \|_h = 4 \neq |\alpha| \cdot \| x \|_h = |2| \cdot \| (1,1) \|_h = 2$.

Medidas de disimilitud (3)

- Las clases de normas utilizadas con frecuencia son las:
  - Normas matriciales
  - Normas de Lebesgue
  - Norma de Minkowski
- La norma matricial se define como: $\| x \|_A = \sqrt{xAx^T}$
- Con una matriz $A \in \mathbb{R}^{n \times n}$. Casos especiales importantes de la norma matricial son la norma Euclídea.

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- Para $x \in \mathbb{R}^n$ se define su norma Euclídea como:

$$\| x \| = \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + x_2^2 +, \ldots, x_n^2} = \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2}$$

Medidas de disimilitud (4)

- Demostración de la norma Euclídea:
  - Aplicando las propiedades se tiene que para cada $x \in \mathbb{R}^n$ y $\alpha \in \mathbb{R}$, se cumple que:

I. $\| x \| > 0$

$$\sqrt{\langle x, x \rangle} = \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2} > 0$$

II. $\| x \| = 0 \iff x = (0,0, \dots, 0)$

$$\| x \| = 0 \Leftrightarrow \sqrt{\langle x, x \rangle} = 0 \Leftrightarrow \langle x, x \rangle = 0 \Leftrightarrow x = (0,0, \dots, 0)$$

III. $\| \alpha \cdot x \| = |\alpha| \cdot \| x \|$

$$\| \alpha \cdot x \| = \left(\sum_{j=1}^{n} \alpha x_j^2 \right)^{1/2} = \left(\sum_{j=1}^{n} (\alpha)^2 x_j^2 \right)^{1/2} = (\alpha^2)^{1/2} \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2} = |\alpha| \cdot \| x \|$$

Medidas de disimilitud (5)

- Demostración de la norma Euclídea:

IV. $\| x + y \| \leq \| x \| + \| y \|$ para esta propiedad se tiene:

$$\| x + y \|^2 \leq \langle x + y, x + y \rangle =$$

$$= \langle x, x + y \rangle + \langle y, x + y \rangle =$$

$$= \langle x + y, x \rangle + \langle x + y, y \rangle =$$

$$= \langle x + x \rangle + \langle y + x \rangle + \langle x + y \rangle + \langle y + y \rangle =$$

$$= \| x \|^2 + 2\langle x, y \rangle + \| y \|^2 \leq$$

$$\leq \| x \|^2 + 2 \| x \| \| y \| + \| y \|^2 =$$

$$= (\| x \| + \| y \|)^2;$$

Por lo tanto:

$$\| x + y \| \leq \| x \| + \| y \|$$

Cálculo de la norma de una matriz, distancia entre matrices y producto interno entre matrices (1)

- Vamos a calcular la $\| A \|$. Para las matrices: $A = \begin{bmatrix} -1 & 4 \\ 5 & -2 \end{bmatrix}$ $B = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}$

$$\| x \|_2 = \sqrt{\langle x, x \rangle} = \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2}$$

$$\| x \|_1 = \sum_{j=1}^{n} |x_j|$$

$$\| x \|_\infty = \max |x_j|$$

$$\| A \|_2 = < A, A > \qquad < A, A > = a_{11}a_{11} + a_{21}a_{21} + a_{12}a_{12} + a_{22}a_{22}$$

$$< A, A > = a_{11}^2 + a_{21}^2 + a_{12}^2 + a_{22}^2$$

$$< A, A > = a_{11}^2 + a_{21}^2 + a_{12}^2 + a_{22}^2$$

$$\| A \|_2 = (-1)^2 + 4^2 + 5^2 + (-2)^2$$

$$\| A \|_2 = 46$$

$$\sqrt{\| A \|^2} = \sqrt{46}$$

$$\| A \| = \sqrt{46}$$

Cálculo de la norma de una matriz, distancia entre matrices y producto interno entre matrices (2)

- Ahora vamos a calcular la $d(A,B)$ con las matrices: $A=\begin{bmatrix}-1 & 4\\ 5 & -2\end{bmatrix}$ $B=\begin{bmatrix}0 & -1\\ 2 & 1\end{bmatrix}$

$d(A,B) = \| A-B \|$ Primeramente calcular la operación $A-B=?$

$$A-B=\begin{bmatrix}-1-0 & 4-(-1)\\ 5-2 & -2-1\end{bmatrix}=\begin{bmatrix}-1 & 5\\ 3 & -3\end{bmatrix}$$

$$A-B=\begin{bmatrix}-1 & 5\\ 3 & -3\end{bmatrix}$$

Ahora se procede a calcular la norma de $\| A-B \|^2$

$\| A-B \|^2 =< A-B, A-B >$

$\| A-B \|^2 = (-1)^2+5^2+3^2+(-3)^2$

$\| A-B \|^2 = 44$

$\sqrt{\| A-B \|^2} = \sqrt{44}$

$\| A-B \| = \sqrt{44}$

$d(A,B) = \sqrt{44}$

$d(A,B) = 2\sqrt{11}$

Cálculo de la norma de una matriz, distancia entre matrices y producto interno entre matrices (2)

- Ahora vamos a calcular el producto interno de $< A, B >$ con las matrices:

$$< A, B > = a_{11}b_{11} + a_{21}b_{21} + a_{12}b_{12} + a_{22}b_{22}$$

$$A = \begin{bmatrix} -1 & 4 \\ 5 & -2 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}$$

$$< A, B > = -1(0) + 4(-1) + 5(2) - 2(1)$$

$$< A, B > = 4$$

Proyecto 2 (1)

Proyecto 2. Programa en R que permita calcular la disimilitud con base en la Norma Euclídea sobre el conjunto de datos Iris, para las 3 clases de flores que están descritas en el conjunto de datos.

1. **Calcular el valor de disimilitud con la norma Euclídea entre las flores: Setosa, Versicolor y Virginica, tomando como base la longitud y anchura (área) del sépalo y pétalo de cada flor.**
2. **Establecer el umbral de disimilitud entre las 3 clases de flores, con base en los valores de área de cada flor.**
3. **Calcular la distancia entre los elementos de la clase Setosa-Versicolor; Setosa-Virginica; Versicolor-Virginica; Versicolor-Setosa; Virginica-Setosa; Virginica-Versicolor.**
4. **Calcular el producto interno entre cada clase de flores Setosa-Versicolor; Setosa-Virginica; Versicolor-Virginica; Versicolor-Setosa; Virginica-Setosa; Virginica-Versicolor**
5. **Calcular el producto interno entre: Setosa-Versicolor; Setosa-Virginica; Versicolor-Virginica; Versicolor-Setosa; Virginica-Setosa; Virginica-Versicolor**