Untitled

YAN

2023-05-18

Temas a tratar

- Introducción al análisis estadístico multivariado
- Distancias en el análisis estadístico multivariado
- Distancias con variables cualitativas
- Distancias con variables cuantitativas
- Conclusiones Generales

Analisis estadistico multivariado

El análisis estadístico multivariado es una técnica que permite analizar conjuntos de datos que involucran múltiples variables, estudiando cómo se relacionan entre sí y cómo afectan conjuntamente a un resultado o variable de interés mediante el uso de diversas técnicas estadísticas.

Distancias en el análisis estadístico multivariado

- En el análisis estadístico multivariado, se trabaja con múltiples variables.
- Las distancias son una medida de la similitud o diferencia entre los objetos (individuos, variables, etc.) en función de estas variables.
- Son ampliamente utilizadas en análisis de datos, clusterización, clasificación, entre otros.
- En esta presentación, nos enfocaremos en algunas de las distancias más utilizadas en el análisis estadístico multivariado.

Selección de la medida de distancia en análisis estadístico multivariado

- La elección de la medida de distancia adecuada depende del objetivo del análisis.
- Depende del tipo de datos y de la escala de medida de las variables.
- La elección también puede depender de la estructura de los datos (por ejemplo, si hay datos faltantes o valores extremos).
- Es importante tener en cuenta las propiedades de las medidas de distancia, como la simetría, la triangularidad y la identidad de los indiscernibles.

Selección de la medida de distancia en análisis estadístico multivariado

- Si el objetivo es encontrar grupos de objetos similares, se pueden utilizar medidas de distancia que enfaticen la similitud, como la distancia euclidiana.
- Si el objetivo es clasificar objetos en diferentes categorías, se pueden utilizar medidas de distancia que minimicen la variabilidad dentro de cada categoría, como la distancia de Mahalanobis.
- Si el objetivo es determinar la estructura subyacente de los datos, se pueden utilizar medidas de distancia que revelen patrones de covariación entre las variables, como la distancia de correlación.
- Si el objetivo es identificar objetos anómalos o extremos, se pueden utilizar medidas de distancia robustas, como la distancia de Minkowski con un valor de p mayor que 1.

Medidas de distancia entre individuos

- La elección de la medida de distancia entre individuos puede depender de la escala de medida de las variables y del tipo de variables.
- Si las variables están en diferentes escalas, la distancia euclidiana no será adecuada ya que una variable con una escala más grande tendrá una mayor influencia en la medida de distancia.
- Si las variables son de tipo categórico o nominal, la distancia euclidiana no se puede utilizar y se deben usar medidas de distancia apropiadas para variables categóricas, como la distancia de Gower.
- Si las variables son de tipo ordinal, la distancia euclidiana no es la mejor medida y se pueden utilizar medidas de distancia apropiadas para variables ordinales, como la distancia de Spearman.
- Si las variables son de tipo binario, se puede utilizar la distancia de Hamming.
- Si las variables son mixtas (numéricas y categóricas), se pueden utilizar medidas de distancia apropiadas para datos mixtos, como la distancia de Gower.
- Si las variables están en diferentes escalas, se pueden utilizar medidas de distancia que tengan en cuenta la variabilidad y la escala de cada variable, como la distancia de Mahalanobis.

Medidas de distancia entre variables

- La elección de la medida de distancia entre variables también puede depender de la escala de medida y del tipo de variables.

Distancias con variables cualitativas

En el análisis multivariado de variables cualitativas, la distancia se refiere a la medida de la diferencia entre dos observaciones o individuos en función de sus características o variables cualitativas. Existen diferentes medidas de distancia que se pueden utilizar en el análisis multivariado de variables cualitativas. Algunas de las más comunes son:

- Distancia Hamming
- Distancia Jaccard

Distancia Hamming

La distancia de Hamming es una medida de la distancia entre dos cadenas de igual longitud. La fórmula para calcular la distancia de Hamming es la siguiente:

$$D_H(x, y) = \sum_{i=1}^{n} \mathbb{I}(x_i \neq y_i)$$

Donde x y y son las cadenas que se van a comparar, n es la longitud de las cadenas y $\mathbb{I}(x_i \neq y_i)$ es una función indicadora que devuelve 1 si los caracteres en las posiciones i de x y y son diferentes y 0 en caso contrario.

Ejemplo

Supongamos que tenemos dos cadenas binarias de la misma longitud, x = "0110101" e y = "1100110". Queremos calcular la distancia de Hamming entre estas dos cadenas.

Para ello, podemos utilizar la fórmula anterior,

$$D_H(x, y) = \sum_{i=1}^{7} \mathbb{I}(x_i \neq y_i)$$

$$D_H(x, y) = \mathbb{I}(0 \neq 1) + \mathbb{I}(1 \neq 1) + \mathbb{I}(1 \neq 0) + \mathbb{I}(0 \neq 0) + \mathbb{I}(1 \neq 0) + \mathbb{I}(0 \neq 1) + \mathbb{I}(1 \neq 0)$$

$$D_H(x, y) = 1 + 0 + 1 + 0 + 1 + 1 + 1$$

Por lo tanto, la distancia de Hamming entre las cadenas binarias x e y es de 5. Esto significa que hay 5 posiciones en las que las cadenas difieren.

Distancia Jaccard

Se utiliza para medir la similitud entre dos conjuntos de variables cualitativas.

La distancia de Jaccard entre dos conjuntos A y B se define como:

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Donde $|A|$ representa el tamaño del conjunto A y $A \cap B$ y $A \cup B$ representan la intersección y la unión de los conjuntos A y B, respectivamente. Esta fórmula mide la distancia entre dos conjuntos basándose en la similitud entre ellos.

Ejemplo

A continuación, se presenta una tabla que contiene información acerca de los clientes y los productos adquiridos. En dicha tabla, se representa con el número 0 cuando el cliente no ha comprado el producto y con el número 1 cuando sí lo ha adquirido.

Ejemplo

Con base en los datos recolectados previamente, se realiza el cálculo de la distancia Jaccard entre cada cliente, con el fin de identificar patrones y tendencias en su comportamiento y asi tener una tomar decisiones más precisas e informadas. El resultado de este cálculo se producirá en una matriz de distancias, que sintetizará de manera efectiva los resultados obtenidos.

Conclusiones

- Los clientes cuyas distancias son más cercanas a 0 tienen un alto grado de similitud en sus compras. tal es el caso de los clientes A y B, C y D que presentan una distancia de 0,33 lo que indica que tienen patrones de compra similares.
- Los clientes que tienen una distancia de 0,50 como los clientes B y D, tienen una afinidad del 50% en sus compras. Esto significa que dividen la mitad de los productos que compran.
- Los clientes cuyas distancias son más cercanas a 1 tienen un alto grado de disimilitud en sus compras, como lo son, los clientes A y D, asi como los clientes B y C que presentan una distancia de 0,75 Esta medida indica que estos clientes presentan la mayor disimilitud entre todos los datos analizados en cuanto a sus patrones de compra.

Distancias con variables cuantitativas

En análisis multivariado, se utilizan diferentes métodos para medir la distancia entre observaciones con variables cuantitativas. Estas medidas de distancia son utilizadas en técnicas como análisis de componentes principales, análisis de correspondencias, análisis de conglomerados, entre otras.

A continuación, se describen algunas de las medidas de distancia más comunes en análisis multivariado con variables cuantitativas:

- Distancia Euclidiana
- Distancia Manhattan
- Distancia Mahalanobis

Distancia Euclidiana

La distancia euclidiana es una medida de la distancia entre dos puntos en un espacio euclidiano de dos o más dimensiones.

La distancia euclidiana entre dos puntos p y q en un espacio euclidiano de n dimensiones se define como:

$$d_E(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Donde p_i y q_i son las coordenadas del punto p y el punto q en la i-ésima dimensión, respectivamente.

Ejemplo

Supongamos que tenemos dos vectores P y Q:

$p = (1, 2, 3)$ y $q = (4, 5, 6)$, entonces la distancia euclidiana entre p y q es:

$$d_E(p, q) = \sqrt{(1-4)^2 + (2-5)^2 + (3-6)^2} = \sqrt{27} \approx 5.196$$

Distancia Manhattan

La distancia de Manhattan es una medida de distancia entre dos puntos en un espacio euclidiano de n dimensiones, mide la distancia que hay que recorrer para ir de un punto a otro si sólo se permiten movimientos en línea recta horizontal o vertical.La fórmula para calcular la distancia de Manhattan es la siguiente:

$$D_M(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Donde x y y son los vectores que se van a comparar, n es el número de dimensiones y $|x_i - y_i|$ representa la diferencia absoluta entre la coordenada i de x y la coordenada i de y.

Ejemplo

Supongamos que tenemos dos vectores X y Y:

$x = (1, 2, 3)$ y $y = (4, 5, 6)$, entonces la distancia Manhattan entre x y y es:

$$D_M(x) = |1-4| + |2-5| + |3-6| = 9$$

Distancia Mahalanobis

La distancia de Mahalanobis es una medida de la distancia entre un punto y un conjunto de puntos, teniendo en cuenta la covarianza entre las variables. La fórmula para calcular la distancia de Mahalanobis es la siguiente:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Donde x es el vector de variables, μ es el vector de medias y Σ es la matriz de covarianza.

Donde p_i y q_i son las coordenadas del punto p y el punto q en la i-ésima dimensión, respectivamente.

Ejemplo

En el siguiente ejemplo se utiliza la distancia de Mahalanobis para detectar valores atípicos en un conjunto de datos simulado de características de televisores. lo que puede ayudar a detectar problemas en la producción o a tomar decisiones de marketing y precios más informadas.

```
##   resolucion_pantalla consumo_energia precio distancia_maha
## 1                1593              57    700       3.068845
## 2                1110             197   1500       2.806825
## 3                1846             217   1000       2.725497
## 4                1139             170   3000       2.543593
## 5                1856              80    500       1.464287
## 6                1529             219   1200       1.401068
```

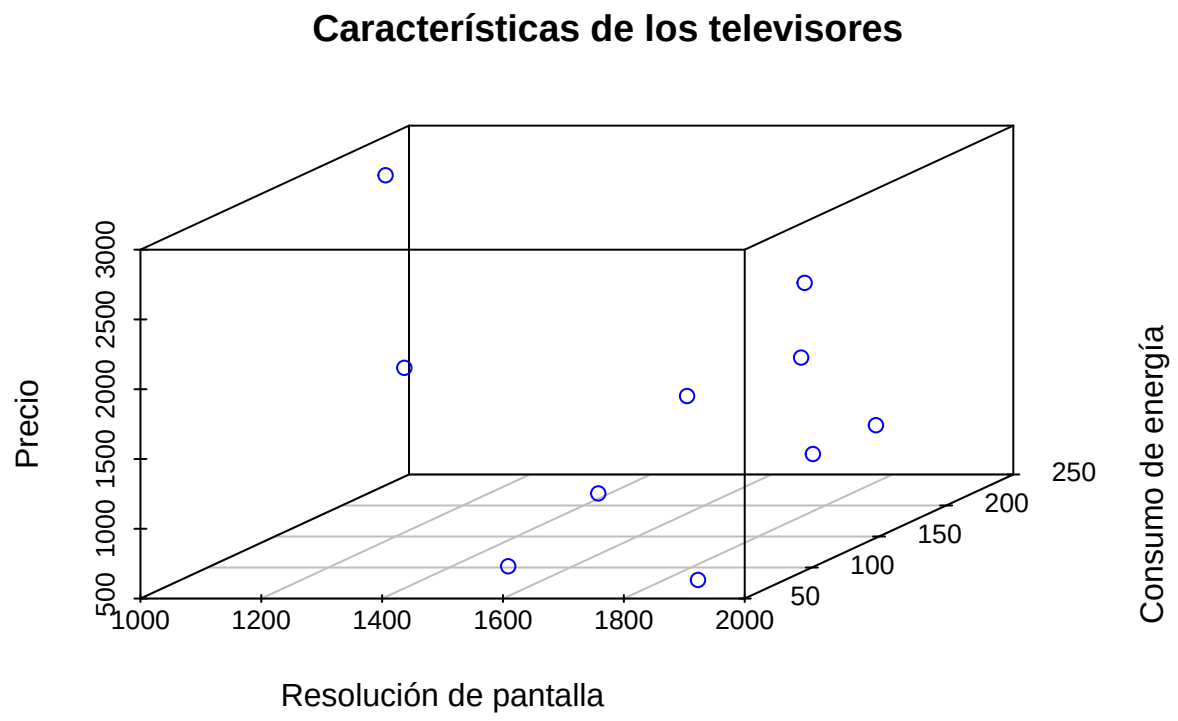
Conclusiones

Al analizar detalladamente la tabla anterior, podemos concluir que aquellos dispositivos tecnológicos que se encuentran significativamente alejados de la media podrían estar experimentando problemas relacionados con sus características. Por el contrario, aquellos dispositivos que se encuentran en posiciones más cercanas a la media sugieren una mayor estabilidad en sus características.

Por consiguiente, los fabricantes de dispositivos electrónicos deben considerar cuidadosamente estos resultados al diseñar y promocionar sus productos, con el fin de optimizar su posicionamiento en el mercado y ofrecer el mejor valor a sus clientes.

Diagrama de dispersion

A continuación, se presenta un gráfico que permite una mejor visualización de la relación entre las variables, lo que facilita la identificación de patrones y tendencias en los datos.



Conclusiones

- Resolución de pantalla y precio: se puede observar una relación positiva entre la resolución de pantalla y el precio de los televisores. Esto sugiere que los televisores con una resolución de pantalla más alta tienden a tener un precio más alto.
- Consumo de energía y precio: se puede observar una relación negativa entre el consumo de energía y el precio de los televisores. Esto sugiere que los televisores que consumen menos energía tienden a tener un precio más alto.
- Distancia entre los puntos: la distancia entre los puntos en la gráfica indica la similitud en las características de los televisores. Por ejemplo, los televisores con una resolución de pantalla similar y un consumo de energía similar tienden a estar más cerca de unos de otros.
- Dos grupos claros: se puede observar que hay dos grupos claros de televisores en la gráfica: uno con una resolución de pantalla alta y un consumo de energía bajo, y otro con una resolución de pantalla más baja y un consumo de energía más alto.