Ciencia de Datos en inteligencia de Negocios
Lectura: Datos y su clasificación.

Datos.

Un dato es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, fenómenos y entidades. También suele ser llamado dato a un valor que recibe una computadora por algún medio, los datos representan la información que el programador manipula en la construcción de una solución o en el desarrollo de un algoritmo [1].

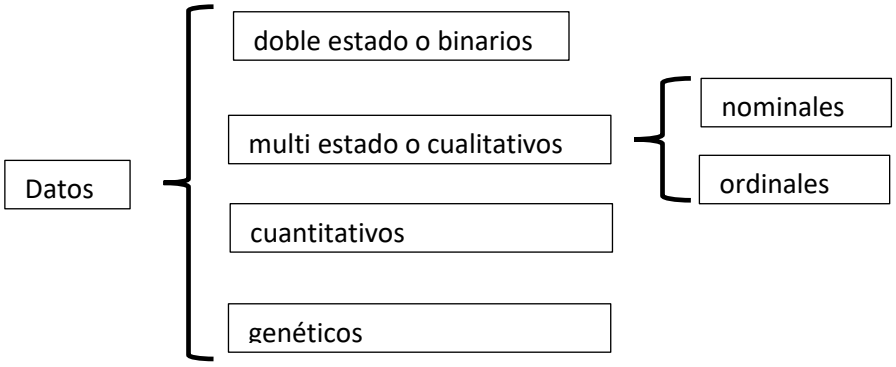
Los datos aisladamente pueden no contener información relevante. Solo cuando un conjunto de datos se examina conjuntamente basado en un enfoque, hipótesis o teoría, se puede apreciar la información contenida en dichos datos. Los datos pueden consistir en números, estadísticas o proposiciones descriptivas. Los datos agrupados, estructurados e interpretados se consideran que son la base de información relevante que se pueden utilizar en la toma de decisiones, la reducción de la incertidumbre, la realización de cálculos y estimaciones [1].

Clasificación de tipos de datos.

Dentro de taxonomía numérica de los datos, se puede encontrar la siguiente clasificación de la naturaleza de los datos para posibles caracterizaciones.

1. **Datos doble estado o binarios.** Son obtenidos cuando solo presentan en la característica evaluada dos posibilidades de respuesta, como ausencia o presencia de alguna característica. Generalmente son codificadas como 0 para las negativas y 1 para las positivas. Presentan una distribución estadística binomial y el cálculo de sus estadísticos básicos deben de realizarse con las fórmulas diseñadas para esta distribución [2].
2. **Datos multi estado o cualitativos.** Presentan más de dos posibilidades de respuesta y pueden ser con secuencia lógica o sin esta, es decir, este tipo de datos engloba a los estadísticos nominales (sin secuencia lógica) y a los ordinales (con secuencia lógica), cada uno de ellos con características propias y métodos diseñados para cada uno [2].
    a. **Datos nominales.** Son aquellos con más de dos posibilidades de respuesta pero que no llevan un orden lógico. Por ejemplo, a los colores se pueden asignar un número a cada uno, es decir, rojo = 1, amarillo = 2, azul = 3, etc; sin embargo, no se puede concluir que rojo es menor que azul, son diferentes pero no en orden o secuencia lógica.
    b. **Datos ordinales.** Pueden ser ordenados con una secuencia lógica. Por ejemplo, al tamaño de hoja se puede asignar un número para designar que tan grande es, es decir, que puede ser chica = 1, mediana = 2, grande = 3. En este caso si se puede decir que grande es más que chica o que la mediana aunque no de una manera cuantitativa. Estos datos cumplen generalmente con las distribuciones polinomiales u otras distribuciones derivadas de variables discretas. No obstante, datos ordinales y nominales son estadísticamente diferentes.

3. **Datos cuantitativos.** Son datos que pueden contarse y que son continuos (presentan cualquier valor real); generalmente con una distribución normal e incluso algunos datos discretos pueden ser utilizados con un buen muestreo, pero mediante estadísticos de tendencia central para obtener normalidad. Ejemplos de este tipo de datos pueden ser la altura de planta, peso de fruto y de semilla, cantidad de materia, número de hojas, de frutos y de flores, entre otras (si el muestreo es adecuado) [2].
4. **Datos genéticos.** Los datos genéticos son aquellos en los cuales se puede conocer características que cumplan con la genética mendeliana y de poblaciones, es decir, donde se conocen el número de *loci* que se está evaluando y el número de *alelos* por locus.



Medidas de distancia y similitud.

Ya que se tiene determinado el tipo de dato a utilizar, se pueden realizar una medida de similitud proponiendo el índice adecuado apropiado. La primera y más importante es evitar la combinación de datos, esto debido a que cada tipo de dato presenta características propias que no comparten con los de otra naturaleza.

Un índice de similitud δ_{ij} es una medida de que tan parecido es un dato i con otro j. Generalmente, las similaridades están acotadas en el rango de cero a uno; un aumento en la similaridad implica un aumento de la semejanza entre datos o variables, y toda similaridad de un dato consigo mismo debería ser igual al máximo valor posible, es decir, uno. Las distancias en cambio disminuyen con un aumento del parecido, no son negativas y la distancia de un elemento consigo mismo es cero. Tanto las matrices de similaridades como las de distancias son simétricas; es decir, la distancia entre el individuo 'a' y el 'b' es la misma que entre el 'b' y el 'a'.

Dependiendo del método elegido para la ordenación, la clasificación, o el cálculo de índices de diversidad, así como de la escala de medición de los rasgos funcionales, la asociación entre los datos se expresará en términos de similaridad o distancia. Sin embargo, las similaridades pueden transformarse en distancias y viceversa.

Para el rango cero-uno, la similaridad δ_{ij} puede ser transformada a distancia d_{ij} de la siguientes formas:

$$d_{ij} = 1 - \delta_{ij}$$

$$d_{ij} = \sqrt{1 - \delta_{ij}}$$

$$d_{ij} = \sqrt{\delta_{ii} - 2\delta_{ij} + \delta_{jj}}$$

$$d_{ij} = -\log(\delta_{ij})$$

El uso de índices de diversidad funcional basados en distancias, así como los métodos de clasificación y/o de ordenación requiere una comprensión de las propiedades de la escala de medición de los rasgos funcionales de las especies, y de las características de las medidas de semejanza asociadas a cada tipo de datos. El estudio de la diversidad funcional incluye rasgos funcionales expresados en diferentes formas: variables binarias (presencia/ausencia), variables cualitativas (nominales y ordinales) y variables cuantitativas (discretas o continuas).

Es importante aclarar que las medidas de distancia no tienen sentido físico, ya que solo son usados como indicadores de proximidad, cercanía o semejanza.

Ciencia de Datos en inteligencia de Negocios
Lectura: Datos y su clasificación.

Similitud en datos de doble estado o binarios

Cuando la matriz de datos X, proviene de la observación de n atributos que toman el valor 0 si la característica está ausente y el valor 1 si está presente, la información del grado de asociación entre cualquier par de individuos y puede representarse como una tabla de contingencia $2x2$.

		Individuo j		
		Presente (1)	Ausente (0)	
Individuo i	Presente (1)	a	b	$a+b$
	Ausente (0)	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d$

Donde a es el número de caracteres presentes comunes, b es el número de caracteres presente en i pero no en j, c es el número de caracteres presentes en j pero no presentes en i y d es número de caracteres no presentes en ambos datos. Para la matriz X de dimensión $m \times n$, es posible crear o construir $\frac{n(n-1)}{2}$ tablas de contingencia. Generalmente, a la tabla de contingencia también se le conoce como matriz de confusión.

Se han propuesto diversas medidas de similaridad que verifican estas propiedades, entre otros, Jaccard (1908), Rusell y Rao (1940), Sorensen (1948) y Sokal y Michener (1958). Sin embargo, existen similitudes que no verifican las propiedades de simetría y rango tales como la de Kulczynski (1970) acotada en el rango [0,") y otros que expresan dependencia estocástica entre x t y x j como son las de Yule (1912) y la de Pearson (1926), acotadas en el rango (-1,1), donde la mayor disimilaridad corresponde a -1, la similaridad total a 1 y el valor 0 se asocia a la independencia estocástica

No existe un criterio universal de cuándo usar una u otra similitud. Los diferentes autores que han abordado el tema coinciden en que la elección de una determinada similitud dependerá del peso que se desea dar a las frecuencias de a, b, c y d, del tipo de datos que se quieran representar y de la situación experimental (Legendre y Legendre 1979, Gower y Legendre 1986).

Tabla 1. Indices de similaridad para datos binarios

Similaridad		Simétrica	Rango
Emparejamiento simple	$\frac{a+d}{a+b+c+d}$	Si	$[0,1]$
Rogers y Tanimoto	$\frac{a+d}{a+2b+2c+d}$	Si	$[0,1]$
Hamman	$\frac{(a+d)-(b+c)}{a+b+c+d}$	Si	$[-1,1]$
Yule	$\frac{ad-bc}{ad+bc}$	Si	$[-1,1]$
Pearson	$\frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$	Si	$[-1,1]$
Jaccard	$\frac{a}{a+b+c}$	No	$[0,1]$
Russel y Rao	$\frac{a}{a+b+c+d}$	No	$[0,1]$

No hay que olvidar que cada índice de similitud tiene propiedades distintas por lo que hay que considerar el objetivo que se quiere obtener para elegir el índice adecuado.

Similitud en datos multi estado nominales

Si las categorías para cada variable son codificadas por ejemplo, como: $0,1,2,3,\ldots,k$, el grado de asociación entre cualquier par de individuos x_i y x_j puede medirse a través de la expansión del emparejamiento simple que se expresará como:

$$\delta_{ij} = \frac{número\ de\ caracteres\ coincidentes}{número\ total\ de\ caracteres}$$

No obstante, cuando el cero representa ausencia del carácter es recomendable ignorar el empate de ceros en forma similar como lo hace Jaccard.

Para el tratamiento de las variables 'indicadoras excluyentes' pueden utilizarse dos estrategias: uso de variables 'auxiliares' (dummy) o desdoblamiento en tantas variables como estados posibles presente el carácter. En el caso de variables 'auxiliares' cada variable estará representada por tantas pseudo variables como número de estados diferentes menos uno. Así cada especie tendrá asociado un perfil con un 1 en el estado en que se encuentre, estando el último estado representado solo por ceros.

En estos casos debería utilizarse una medida de similitud que incluyera la presencia-ausencia (b y c) ya que contribuye a la diferencia entre las especies, sin incluir el componente de ausencia-ausencia, ya que aumentaría artificialmente la similitud.

Cuando se realiza el desdoblamiento de una variable nominal en todos sus posibles estados, se identifica la presencia o ausencia de cada estado del rasgo funcional en estudio, pero como estos estados son excluyentes cada especie tendrá un solo valor de presencia (1) y el resto serán ceros.

Similitud en datos multi estado ordinales

Las variables ordinales pueden considerarse como variables cuantitativas si la asignación del ranking no es caprichosa sino que refleja en cierta forma una diferencia entre los estados de la variable. Por ejemplo, si se considera la resistencia al fuego de un conjunto de especies usando las categorías: muy baja, baja, media, alta y muy alta; puede ser razonable asignarle valores: 0, 1, 2, 3, 4, respectivamente ya que las categorías consecutivas pueden considerarse como equidistantes. De esta manera, la nueva variable numérica podría ser tratada como una variable cuantitativa. Cuando las variables ordinales tienen categorías no equidistantes, no es razonable suponer equidistancias.

Ciencia de Datos en inteligencia de Negocios
Lectura: Datos y su clasificación.

Similitud en datos cuantitativos

Generalizando para m especies y n variables aleatorias cuantitativas (rasgos funcionales), la distancia usual que se observa entre el par de unidades x_i y x_j cuando se representan en el espacio de coordenadas definido por n variables cuantitativas, es conocida como distancia Euclideana:

$$d_{ij} = \sqrt{\sum_{t=1}^{n} \left(x_{it} - x_{jt}\right)^2}$$

donde n son los rasgos funcionales. La distancia Euclídeana es la más conocida, la de mayor uso y es la herramienta fundamental de cálculo de la mayoría de los métodos multivariados basados en distancias. Sin embargo, presenta varios inconvenientes: no está acotada, es sensible a cambios de escalas y considera las n variables estocásticamente independientes.

Se han propuesto varias transformaciones que permiten minimizar y/o eliminar estos inconvenientes, entre otras: se recomienda utilizarla en caso de homogeneidad entre la naturaleza física de las variables, cuando esto no es posible se puede estandarizar cada variable por su rango r_t asegurando que la contribución de cualquier variable estará acotada en el intervalo $(0,1)$. Además puede dividirse por la cantidad de variables obteniendo una distancia media que oscilará en este rango y facilita su conversión a similaridad, la expresión estará definida por:

$$d_{ij} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \frac{\left(x_{it} - x_{jt}\right)^2}{r_t}}$$

A continuación se presenta la formulación y propiedades de las distancias y disimilaridades no negativas más utilizadas en los estudios de diversidad. Las más usadas son las distancias: Euclídea, Manhattan y Mahalanobis.

Tabla 2. Indices de similaridad para datos cuantitativos

Similaridad		Simétrica	Rango
Euclideana	$\sqrt{\sum_{t=1}^{n} \left(x_{it} - x_{jt}\right)^2}$	Si	$[0, \infty]$
Manhattan	$\sum_{t=1}^{n} \lvert x_{it} - x_{jt} \rvert$	Si	$[0, \infty]$
Bray-Curtis	$\frac{\sum_{t=1}^{n} \lvert x_{it} - x_{jt} \rvert}{\sum_{t=1}^{n} \left(x_{it} + x_{jt}\right)}$	Si	$[0, \infty]$
Canberra	$\sum_{t=1}^{n} \frac{\lvert x_{it} - x_{jt} \rvert}{\left(x_{it} + x_{jt}\right)}$	Si	$[0, \infty]$
Minkowski	$\sqrt[p]{\sum_{t=1}^{n} \left(x_{it} - x_{jt}\right)^p}$	Si	$[0, \infty]$
Mihalanobis	$\sqrt{\sum_{t=1}^{n}\sum_{t=1}^{n} \left(x_{it} - x_{jt}\right)\sigma_{lt}{}^{-1}\left(x_{il} - x_{jl}\right)}$	Si	$[0, \infty]$

Bibliografía

[1] Wikipedia, «Dato,» 15 08 2017. [En línea]. Available: https://es.wikipedia.org/wiki/Dato.

[2] C. A. y. E. L. D. Nuñez-Colin, «Uso correcto del análisis de clúster en la caracterización de germoplasma vegetal,» *Agronomía Mesoamericana,* vol. 22, nº 2, pp. 415-427, 2011.