Taller 1 Programación en Lenguajes Estadísticos

Jose Urquijo, Ramiro Payares, Miler Blanco, Luis Oñate

28/06/2022

Elementos de una Base de Datos

Jhon Tukey, el eminente estadístico cuyas ideas se desarrollaron hace más de 50 años formando la base de la ciencia de datos.

Los datos provienen de muchas fuentes: mediciones de sensores, eventos, texto, imágenes y vídeos. El Internet de las cosas (IoT) está arrojando flujos de información. mucho de esto los datos no están estructurados: las imágenes son una colección de píxeles, y cada píxel contiene RGB (rojo, verde, azul) información de color. Los textos son secuencias de palabras y caracteres no verbales. actores, a menudo organizados por secciones, subsecciones, etc. Los flujos de clic son secuencias de acciones realizadas por un usuario que interactúa con una aplicación o una página web. De hecho, un importante. El desafío de la ciencia de datos es aprovechar este torrente de datos sin procesar para convertirlos en información procesable. Para aplicar los conceptos estadísticos que se tratan en este libro, los datos brutos no estructurados debe ser procesado y manipulado en una forma estructurada. Uno de los más comunes formas de datos estructurados es una tabla con filas y columnas, ya que los datos pueden surgir de una base de datos relacional o recopilar se para un estudio.

Hay dos tipos básicos de datos estructurados: numéricos y categóricos. Datos numéricos viene en dos formas: continua, como la velocidad del viento o la duración del tiempo, y discreta, como el recuento de la ocurrencia de un evento. Los datos categóricos toman solo un conjunto fijo de valores, como un tipo de pantalla de TV (plasma, LCD, LED, etc.) o un nombre de estado (Alabama, Alaska, etc.). Los datos binarios son un caso especial importante de datos categóricos que toma solo uno de dos valores, como 0/1, sí/no o verdadero/falso. Otro tipo útil de datos categóricos son datos

ordinales en los que se ordenan las categorías; un ejemplo de esta es una calificación numérica (1, 2, 3, 4 o 5).

¿Por qué nos molestamos con una taxonomía de tipos de datos? Resulta que para los fines de análisis de datos y modelado predictivo, el tipo de datos es importante para ayudar a determinar el tipo de presentación visual, análisis de datos o modelo estadístico. De hecho, la ciencia de datos El software, como R y Python, utiliza estos tipos de datos para mejorar el rendimiento computacional. Más importante aún, el tipo de datos para una variable determina cómo el software manejar los cálculos para esa variable.

Terminos Claves para Tipos de Datos

- Numérico: Datos que se expresan en una escala numérica.
- Continuo: Datos que pueden tomar cualquier valor en un intervalo. (Sinónimos: intervalo, flotar,numérico)
- Discreto: Datos que solo pueden tomar valores enteros, como recuentos. (Sinónimos: entero, contar)
- Categórico: Datos que pueden tomar solo un conjunto específico de valores que representan un conjunto de posibles categorías. (Sinónimos: enumeraciones, enumerado, factores, nominal)
- Binario: Un caso especial de datos categóricos con solo dos categorías de valores, por ejemplo, 0/1, verdadero/Falso. (Sinónimos: dicotómico, lógico, indicador, booleano)
- Ordinal: Datos categóricos que tienen un ordenamiento explícito. (Sinónimo: factor ordenado)

Los ingenieros de software y los programadores de bases de datos pueden preguntarse por qué necesitamos la noción de datos categóricos y ordinales para análisis. Después de todo, las categorías son meramente una colección de valores de texto (o numéricos), y la base de datos subyacente maneja automáticamente la representación interna. Sin embargo, la identificación explícita de los datos como categóricos, a diferencia del texto, ofrece algunas ventajas:

- Saber que los datos son categóricos puede actuar como una señal que le dice al software qué tan estadístico deben comportarse los procedimientos, como producir un gráfico o ajustar

un modelo. En particular, los datos ordinales se pueden representar como un factor ordenado en R, conservando un orden especificado por el usuario en gráficos, tablas y modelos. En Python, scikitlearn admite datos ordinales con sklearn.preprocessing.Ordinal Encoder El almacenamiento y la indexación se pueden optimizar (como en una base de datos relacional).

- he possible values a given categorical variable can take are enforced in the software (like an enum).
- El tercer "beneficio" puede conducir a un comportamiento no deseado o inesperado: el valor predeterminado El comportamiento de las funciones de importación de datos en R (por ejemplo, read.csv) es convertir automáticamente una columna de texto en un factor. Las operaciones subsiguientes en esa columna supondrán que los únicos valores permitidos para esa columna son los importados originalmente, y asignar un nuevo valor de texto introducirá una advertencia y producirá un NA (valor faltante). El paquete pandas en Python no realizará dicha conversión automáticamente. Sin embargo, puede especificar una columna como categórica explícitamente en la función readcsv.

Ideas Claves

- los datos se clasifican típicamente en el software por tipo.
- Los tipos de datos incluyen numéricos (continuos, discretos) y categóricos (binarios,ordinal).
- La tipificación de datos en el software actúa como una señal para el software sobre cómo procesar la información, datos.

Definiciones

"Medidas de tendencia Central y de Dispersión":

Las medidas de tendencia central son medidas estadísticas con las que intentamos simplificar en un solo valor un conjunto de valores que analizan el comportamiento de una población a partir de su muestra. Sin embargo para hacer análisis más precisos sobre el comportamiento de los datos cercanos a la media es necesario utilizar las medidas de dispersión que son Rango de variación, Varianza, Desviación estándar, y Coeficiente de variación que se encargan de medir el grado de

dispersión de los valores de la variable. Es decir, debemos considerar en que medida o grado los datos difieren entre sí.

1. Medidas de tendencia Central(Media Aritmetica, mediana y cuantiles, graficos cuantil-cuantil, Moda, Media geometrica y Media armonica).

MEDIA: Media aritmética, es la que se obtiene sumando los datos y dividiéndolos por el número de ellos.

MEDIANA: Corresponde al percentil 50 %. Es decir, la mediana divide a la población exactamente en dos.

MODA: Valor o (valores) que aparece(n) con mayor frecuencia. Una distribución unimodal tiene una sola moda y una distribución bimodal tiene dos. Útil como medida resumen para las variables nominales.

MEDIA GEOMETRICA: La media geometrica de un conjunto de n numeros positivos se define como la raiz n-esima del producto de los n numeros.

$sqrt(X1)(X2)...(Xn)$

MEDIA ARMÓNICA: La media armónica es igual al número de elementos de un grupo de cifras entre la suma de los inversos de cada una de estas cifras. En otras palabras, la media armónica es una medida estadística recíproca a la media aritmética, que es la suma de un conjunto de valores entre el número de observaciones.

CUANTILES: Un cuantil es aquel punto que divide la función de distribución de una variable aleatoria en intervalos regulares. Por tanto, no es más que una técnica estadística para separar los datos de una distribución.

2. Medidas de Dispersión (Rango y Rango intercuartil, Desviación absolutad, Varianza, Desviación Estandar, y Coeficiente de Variación).

DESVIACIÓN ESTÁNDAR: Llamada también desviación típica; es una medida que informa sobre la media de distancias que tienen los datos respecto de su media aritmética, expresada en las mismas unidades que la variable.

VARIANZA: Es el valor de la desviación estándar al cuadrado; su utilidad radica en que su valor es requerido para todos los procedimientos estadístico.

ERROR TÍPICO: Llamado también error estándar de la media. Se refiere a una medida d variabilidad de la media; sirve para calcular cuan dispersa estaría la media de realizar un nuevo calculo.

COEFICIENTE DE VARIACIÓN: El Coeficiente de Variación es una medida de dispersión que permite el análisis de las desviaciones de los datos con respecto a la media y al mismo tiempo las dispersiones que tienen los datos dispersos entre sí.

DESVIACIÓN ABSOLUTA: La desviación absoluta media se usa en lugar de la desviación media cuando es necesario que los valores extremos afecten menos al valor de la desviación.

RANGO: El rango, o R, es la diferencia entre los valores más alto y más bajo incluidos en un conjunto de datos.

3.Diagrama de Caja

Un diagrama de caja es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, se muestran a simple vista la mediana y los cuartiles de los datos, y también pueden representarse sus valores atípicos.

4.Medidas de Concentración(Curva de Lorenz y Coeficiente Gini)

El coeficiente de Gini es una medida de la desigualdad ideada por el estadístico italiano Corrado Gini. Normalmente se utiliza para medir la desigualdad en los ingresos, dentro de un país, pero puede utilizarse para medir cualquier forma de distribución desigual.

La curva de Lorenz es una representación gráfica utilizada frecuentemente para plasmar la distribución relativa de una variable en un dominio determinado.

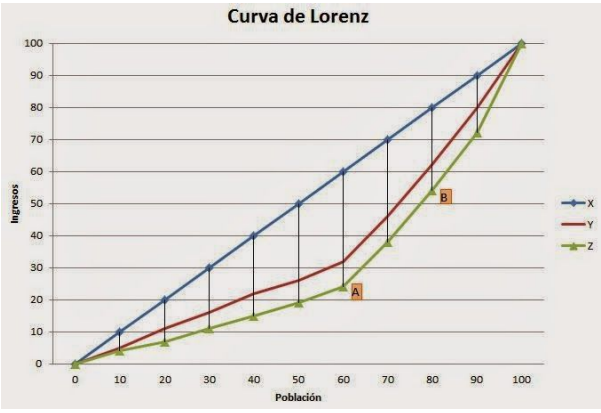

Figura 1: Curva_Lorenz

Posit-TM y que relación tiene con Rstudio

Potsi es una modificación de marca o por ende el nuevo nombre que recibe la plataforma anterior que es conocida como RStudios, su misión principal es crear software de código abierto. Habrá un cambio de herramientas y productos comerciales: RStudio Connect = Posit Connect, Banco de trabajo RStudio = Banco de trabajo Posit, Administrador de paquetes de RStudio = Administrador de paquetes de Posit. En general posit y RStudio son lo mismo.

bibliography:

http://diposit.ub.edu/dspace/bitstream/2445/121804/1/Alca%C3%B1iz%2C%20P%C3%A9rez%2C%20Mar%C3%ADn%20-%20Concentraci%C3%B3n.pdf

https://drive.google.com/file/d/1q3v6K0iaCJPFsuoFM25JMhzEIDZwzHEk/view

https://issuu.com/eduardovielma/docs/media_geom___trica___arm___nica_y_cuad_67b9f66f534aa4

https://www.uaeh.edu.mx/division_academica/educacion-media/repositorio/2010/6-semestre/estadistica/coeficiente-de-variacion.pdf