Taller 1

Programacion En Lenguages Estadisticos

Integrantes:
Julio Andres Miranda Sierra
Camilo Andres Mendoza Araujo
Prof. Jose Francisco Ruiz Monos
UNAL-SEDE LA PAZ

15 de agosto - 2022

Elementos de datos estructurados

Los datos provienen de muchas fuentes: mediciones de sensores, eventos, texto, imágenes y videos. El Internet de las cosas (IoT) está arrojando flujos de información. Gran parte de estos 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 que no son palabras, a menudo organizados por secciones, subsecciones, etc. Los flujos de clics son secuencias de acciones realizadas por un usuario que interactúa con una aplicación o una página web. De hecho, un gran desafío de la ciencia de datos es convertir este torrente de datos sin procesar en información procesable. Para aplicar los conceptos estadísticos cubiertos en este libro, los datos sin procesar sin estructura deben procesarse y manipularse en una forma estructurada. Una de las formas más comunes de datos estructurados es una tabla con filas y columnas, ya que los datos pueden surgir de una base de datos relacional o recopilarse para un estudio.

Hay dos tipos básicos de datos estructurados: numéricos y categóricos. Los datos numéricos se presentan en dos formas: continuos, como la velocidad del viento o la duración del tiempo, y discretos, como el recuento de la ocurrencia de un evento. Los datos categóricos solo toman un conjunto fijo de valores, como un tipo de pantalla de TV (plasma, LCD, LED, etc.) o el nombre de un estado (Alabama, Alaska, etc.). Los datos binarios son un caso especial importante de datos categóricos que toman 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 esto es una calificación numérica (1, 2, 3, 4 o 5).

¿Por qué nos molestamos con una taxonomía de tipos de datos? Resulta que a los efectos del análisis de datos y el modelado predictivo, el tipo de datos es importante para ayudar a determinar el tipo de visualización, análisis de datos o modelo estadístico. De hecho, el software de ciencia de datos, 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 manejará los cálculos para esa variable.

Términos clave 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, flotante, numérico). Discreto: Datos que solo pueden tomar valores enteros, como recuentos. (Sinónimos: número entero, cuenta). Categórico: Datos que pueden tomar solo un conjunto específico de valores que representan un conjunto de categorías posibles. (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 el análisis. Después de todo, las categorías son simplemente 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 cómo deben comportarse los procedimientos estadísticos, 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, scikit-learn admite datos ordinales con sklearn.preprocessing.OrdinalEncoder.
- El almacenamiento y la indexación se pueden optimizar (como en una base de datos relacional).
- Los valores posibles que puede tomar una variable categórica determinada se imponen en el software (como una enumeración).

El tercer "beneficio" puede dar lugar a un comportamiento no deseado o inesperado: el comportamiento predeterminado 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 que se importaron originalmente, y la asignación de 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 read_csv.

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

Otras lecturas

- La documentación de pandas describe los diferentes tipos de datos y cómo se pueden manipular en Python.
- Los tipos de datos pueden ser confusos, ya que los tipos pueden superponerse y la taxonomía en un software puede diferir de la de otro. El sitio web R Tutorial cubre la taxonomía de R. La documentación de pandas describe los diferentes tipos de datos y cómo se pueden manipular en Python.
- Las bases de datos son más detalladas en su clasificación de tipos de datos, incorporando consideraciones de niveles de precisión, campos de longitud fija o variable y más; consulte la guía de SQL de W3Schools.

Datos rectangulares

El marco de referencia típico para un análisis en ciencia de datos es un objeto de datos rectangular, como una hoja de cálculo o una tabla de base de datos.

Datos rectangulares es el término general para una matriz bidimensional con filas que indican registros (casos) y columnas que indican características (variables); El marco de datos es el formato específico en R y Python. Los datos no siempre comienzan de esta forma: los datos no estructurados (p. ej., texto) deben procesarse y manipularse para que puedan representarse como un conjunto de características en los datos rectangulares (consulte “Elementos de los datos estructurados” en la página 2). Los datos de las bases de datos relacionales deben extraerse y colocarse en una sola tabla para la mayoría de las tareas de modelado y análisis de datos.

1 Medidas de tendencia central y dispersión

1.1 media aritmética:

La media aritmética, o también promedio, es el numero y/o valor que se obtiene al sumar todos los datos que tengamos y luego se dividen por la cantidad de datos que hay. Un ejemplo de esto es: si tenemos una lista con los números 3, 10, 9, 2 y 5, se tienen que sumar y luego se divide por 5 que es la cantidad de números.

La media aritmética se formula de la siguiente manera.

$$x^{-} = \frac{X_1 + X_2 + X_3 \cdots + X_n}{N}$$

1.2 Mediana:

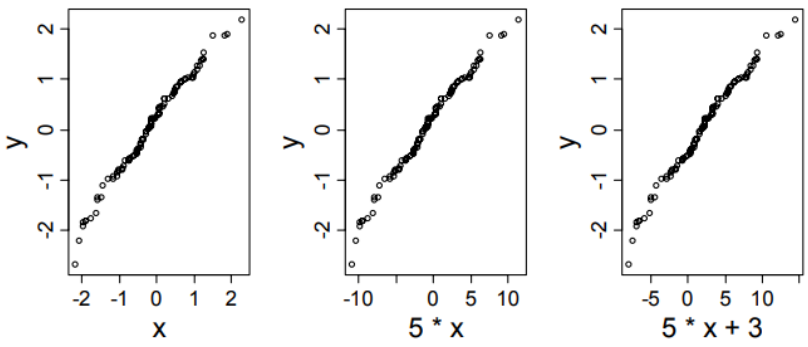
La mediana es el número que se puede encontrar en el medio de una lista de datos. Por ejemplo, en una lista con los números 1, 2, 3, 4, 5, 6, 7, el número "4" es la mediana de la lista de números.

1.3 Cuantiles:

Los cuantiles son puntos que dividen a funciones de distribución de una variable aleatoria, es decir, que los cuantiles son una medida de posición que divide por grupos a una lista de datos.

1.4 Gráficos cuantil - cuantil:

Los gráficos cuantil – cuantil o gráficos Q – Q, tienen como propósito de mostrar gráficamente cómo es la distribución de dos conjuntos de datos y así poder evaluar las semejanzas o diferencias entre ambos grupos de datos.
La función qqplot(x, y, plot=T) grafica las funciones quantile de una muestra vs. la de la otra. Vemos que el Q-Q plot no cambia por una transformación lineal de los datos.



1.5 Moda:

La moda es el número o valor que se repite con mucha frecuencia en un conjunto de datos. Por ejemplo, 4, 3, 7, 4, 9, 10, 4, el número "4" es la moda de ese conjunto de datos debido a que es el valor que más se repite.

1.6 Media geométrica:

La media geométrica es una medida que se obtiene a partir de la multiplicación de un conjunto de números para luego sacar raíz cuadrada.

1.7 Media harmónica:

La media armónica es una media similar a la media aritmética (medida que da la misma ponderación a los valores de un grupo), pero con la diferencia que en la media armónica se suman todos los valores y luego se dividen por el número de observaciones (Número de elementos en el que se calcula la media).
Acontinuacion un claro ejemplo del proceso;

$$H = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}}$$

2 Medidas de dispersion

2.1 Rango:

El rango el es valor que enseña la diferencian entre el valor mínimo y el valor máximo de una muestra.

2.2 Rango Intercuartil:

El rango intercuartil es una agrupación de datos que muestra la diferencia que hay entre el primer cuartil y penúltimo cuartil.

2.3 Desviación absoluta:

La desviación absoluta o desviación absoluta media es una medida estadística que tiene como propósito el cálculo de la variación de la media de un conjunto de datos

2.4 Varianza:

La varianza es la medida estadística que representa la variación que presentan algunos conjuntos de datos con respecto a su media

$$Var(X) = \frac{\sum_1^n (x_i - \bar{X})^2}{n}$$

2.5 Desviación estándar:

La desviación estándar es un índice que tiene como objetivo el mostrar la diferencia que tienen los conjuntos de valores con su media.

2.6 Coeficiente de variación:

El coeficiente de variación o también llamado, coeficiente de Pearson es una medida de dispersión que muestra como es el movimiento relativo de los valores en un conjunto, es decir, que esta mide si una variable se mueve mucho o poco.

3 Diagramas de caja

Los diagramas de caja, también conocidos por "Boxplot", son un método para la representación de grafica de variables cualitativas o cuantitativas. Esto sirve para identificar con mayor rapidez y eficacia los cuantiles de los conjuntos de variables cuantitativas, o ya sea también cualitativas.

4 Medidas de concentracion

4.1 Curva de Lorenz:

La curva de Lorenz es una forma grafica que sirve para representar la distribución relativa que tiene una variable de un conjunto o dominio. Este conjunto puede ser cualitativos o cuantitativos

4.2 Coeficiente Gini:

El coeficiente de Gini o índice de Gini es una medida estadística que se encarga de medir la desigualdad de las variables. Esta es muy usada para medir la desigualdad salarial de las personas

Relacion entre Posit™ y R Studios

Posit™ es el nuevo nombre que tiene RStudio, este cambio se debe a que no solo se quiera actualizar la interfaz de programación de RStudio, sino que también se quiere adquirir a los usuarios de Python y Visual Studios Core. Esta interfaz busca actualizar en entorno de programación en lenguaje R y adicionar el lenguaje Python para volverse una interfaz de programación "Bilingües".

Referencias

- Kelmansky, D. D. (2008). Análisis de datos. Dm. Recuperado 16 de agosto de 2022, de ttps://www.dm.uba.ar/materias/analisis$_d e_d atos$/2008/1/$teoricas/Teor5.pdf$: : $text = Gr$

- "Elements of structured data" (p ags. 2-4) del libro "Bruce, P., Bruce, A., Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media".