Universidad Nacional de Colombia

Sede La Paz

CURSO - Programación en lenguajes estadísticos

Módulo 1 - Quinto semestre 2022

TALLER I

Estudiantes	Carrera
Erick Enrique Bastidas Santana	Ingeniería Mecatrónica
Raul Ricardo Reales Cohen	Ingeniería Mecatrónica

1. **Traducción de la sección "Elements of structured data" (págs. 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".**

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, mucho de esto los datos no están estructurados. Las imágenes son una colección de píxeles, y cada píxel contiene RVA (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 clics son secuenciales de acciones de un usuario que interactúa con una aplicación o una página web. De hecho, un importante 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 recopilados para un estudio.

Hay dos tipos básicos de datos estructurados: numéricos y categóricos. Los 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 (Ala- bama, 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.

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, 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 el noción de datos categóricos y ordinales para análisis. Después de todo, las categorías son meramente un colección de valores de texto (o numéricos), y la base de datos subyacente maneja 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 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 el 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: no estructurados los datos (p. ej., texto) deben procesarse y manipularse para que puedan representarse como un conjunto de características en los datos rectangulares (consulte “Elementos de 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.

2. **Definiciones de “Medidas de tendencia central y dispersión”: 1. Medidas de tendencia central (media aritmética, mediana y cuantiles, gráficos cuantil-cuantil, moda, media geométrica y media armónica). 2. Medidas de dispersión (rango y rango intercuartil, desviación absoluta, varianza y desviación estándar, y coeficiente de variación). 3. Diagramas de caja. 4. Medidas de concentración (curva de Lorenz y coeficiente Gini).**

Definiciones

Medidas de tendencia central: Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda.

- **Media aritmética:** La medida de tendencia central más conocida y utilizada es la media aritmética o promedio aritmético. Se representa por la letra griega μ cuando se trata del promedio del universo o población y por $\bar{Y}$ (léase Y barra) cuando se trata del promedio de la muestra. Es importante destacar que μ es una cantidad fija mientras que el promedio de la muestra es variable puesto que diferentes muestras extraídas de la misma población tienden a tener diferentes medias. La media se expresa en la misma unidad que los datos originales: centímetros, horas, gramos, etc.
- **Mediana:** Otra medida de tendencia central es la mediana. La mediana es el valor de la variable que ocupa la posición central, cuando los datos se disponen en orden de magnitud. Es decir, el 50% de las observaciones tiene valores iguales o inferiores a la mediana y el otro 50% tiene valores iguales o superiores a la mediana. Si el número de observaciones es par, la mediana corresponde al promedio de los dos valores centrales. Por ejemplo, en la muestra 3, 9, 11, 15, la mediana es (9+11)/2=10.
- **Moda:** La moda de una distribución se define como el valor de la variable que más se repite. En un polígono de frecuencia la moda corresponde al valor de la variable que está bajo el punto más alto del gráfico. Una muestra puede tener más de una moda.
- **Cuantil:** Un cuantil es aquel punto que divide la función de distribución de una variable aleatoria en intervalos regulares.
- **Gráficos cuantil-cuantil:** Los diagramas cuantil-cuantil son una herramienta de exploración utilizada para evaluar las similitudes entre la distribución de una variable numérica y una distribución normal, o entre las distribuciones de dos variables numéricas.
- **Media geométrica:** Para un conjunto de n números, la media geométrica es la raíz enésima del producto de esos números. Por ejemplo, la media geométrica de los números 2, 3 y 14 es igual a $(2 * 3 * 14)^{1/3} = (84)^{1/3} = 4,37952$.
- **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.

Medidas de dispersión: Las medidas de dispersión tratan, a través del cálculo de diferentes fórmulas, de arrojar un valor numérico que ofrezca información sobre el grado de variabilidad de una variable.

- **Rango:** Puede ser encontrado en otras bibliografías como recorrido, es la diferencia entre la puntuación mayor y la puntuación menor de un conjunto de datos. Nos deja ver que tan grande puede ser una variación o un cambio.
- **Rango intercuartílico:** El rango intercuartílico IQR (o rango intercuartil) es una estimación estadística de la dispersión de una distribución de datos. Consiste en la diferencia entre el tercer y el primer cuartil. Mediante esta medida se eliminan los valores extremadamente alejados. El rango intercuartílico es altamente recomendable cuando la medida de tendencia central utilizada es la mediana (ya que este estadístico es insensible a posibles irregularidades en los extremos).
- **Desviación absoluta:** Esta desviación muestra la variación que tiene cada uno de los datos de un grupo con respecto a su media aritmética, lo que nos permitirá determinar que tan homogéneo es el grupo de datos.

- **Varianza:** Es la medida de dispersión que resulta de obtener el promedio de cada una de las mediciones de un grupo de datos respecto a la media. No usa la misma unidad que los datos, sino su cuadrado; por ejemplo si los datos obtenidos se refieren al peso de un grupo de personas (kg) la varianza lo expresará en kg2.
- **Desviación estándar:** también llamada desviación típica, se define como la raíz cuadrada de la varianza. Pretende regresar la medida de variabilidad a las mismas unidades que presentaban los datos originalmente y es preferida para fines descriptivos.

  Tanto varianza como desviación estándar no pueden arrojar números negativos, suelen ser utilizadas para el análisis de poblaciones y muestras de tal forma que hay dos maneras de representarlas: cuando se hable de varianza muestral se simbolizara como s^2 cuando se hable de varianza poblacional se representara con sigma minúscula cuadrada σ^2 si de desviación muestral estándar hablamos se simbolizara con la letra s, de igual forma con la desviación poblacional estándar se hará uso del sigma σ
- **Coeficiente de variación:** El coeficiente de variación de Pearson, ha resultado ser una medida dedispersión de amplia utilización,cuando se pretende comparar la dispersión entre varias poblaciones, de una o diferentes variables medidas en la misma o diferentes escalas. Este estadígrafo es definido como la relación por cociente que se establece entre la desviación estándar y la media aritméticade la variable; dicho de otro modo, desviación estándar expresada como porcentaje de la media aritmética; Esto lo hace un coeficiente adimensional, invariante por la escala de medición de la variable analizada. Todo lo anterior justifica la preferencia de utilizarlocuando el propósito en la investigación sea la comparación de poblaciones desde el punto de vista de su variabilidad.

Diagramas de cajas: La Estadística se apoya en las representaciones gráficas para mostrar cómo se distribuyen los datos. Entre ellas se encuentran el diagrama de caja y bigotes, el cual es la representación gráfica de una distribución de datos, diseñada para tomar decisiones y razonar acerca de esas distribuciones. Esta representación consta de cinco elementos: el valor mínimo, el primer cuartil (Q1), el segundo cuartil (Q2), el tercer cuartil (Q3) y el valormáximo; como puede notarse ellosdividena los datos en cuatro grupos. Esta representación presenta al mismo tiempo una medida de tendencia central (mediana), dosmedidasde dispersión (rango y rango intercuartil) e indica la simetría o asimetría de la distribución.

Medidas de concentración: Las medidas de concentración nos informan de la concentración de la distribución , entendida en un sentido distinto al de la antinomia "dispersión/ concentración": miden lo que podríamos llamar la concentración en sentido "económico":miden el mayor o menor "grado de igualdad en el reparto de la totalidad de los valores de la variable.

De esta manera si una pequeña parte de la población (unos pocos individuos) tiene una gran parte del total de la variable (renta, salario, capital, etc.), la variable estará muy concentrada (en pocas manos).Sin embargo, si se guardan las proporciones entre individuos y parte del total que se reparten la distribución será igualitaria, homogénea, poco o nada concentrada.

- **Curva de Lorenz:** La Curva de Lorenz grafica la fracción acumulada de una variable aleatoria versus la fracción acumulada de población receptora de esa variable repartida, asevera Chaves, (2009). En términos generales, afirma Medina, (2001), la curva de Lorenz representa el porcentaje acumulado de ingreso recibido por un determinado grupo de población ordenado en forma ascendente de acuerdo a la cuantía de su ingreso. Cuanto más alejada se encuentre la curva de Lorenz de la línea de igualdad

perfecta mayor es la desigualdad que se presenta, caso contrario, entre más cerca se encuentre menor será la desigualdad y al ser igual no existe desigualdad.

- **Coeficiente de Gini:** Es un indicador, que se clasifica entre las medidas estadísticas para el análisis de la distribución del ingreso, no utiliza como parámetro de referencia el ingreso medio de la distribución -a diferencia de la desviación media, la varianza y el coeficiente de variación-, dado que su construcción se deriva a partir de la curva de Lorenz.

3. **¿Qué es Posit™ y qué relación tiene con R Studio?**

   Posit™ es un nuevo nombre que los desarrolladores del software de R, le han dado a Rstudio, mas que grandes cambios este nuevo nombre lo que ofrece, es que la gente entienda que para los desarrolladores es muy importante continuar con un código abierto accesible para todas las personas. No es que no exista una relación entre Rstudio y Posit, simplemente son lo mismo. Con excepción que Posit es más compatible con el Software Python y Visual Studio Code y emplea mas lenguas(Idiomas) los cuales amplían en un buen margen el uso de esta aplicación debido que será utilizado por muchas mas personas en distintas partes del mundo.

Bibliografía

- Enrique Rus Arias, 07 de enero, 2021 Cuantil. Economipedia.com.
- Flores, J., Flores, R. (2018). La enseñanza del diagrama de caja y bigotes para mejorar su interpretación. Revista Bases de la Ciencia. e-ISSN 2588-0764, 3(1), 69-75.
- González, H. A. B. (2020). La curva de Lorenz y el coeficiente de Gini como medidas de la desigualdad de los ingresos. REICE: Revista Electrónica de Investigación en Ciencias Económicas, 8(15), 104-125.
- Mayorga-Ponce , R. B., Reyes-Torres, S. B., Baltazar-Téllez , R. M., Martínez-Alamilla , A. (2021). Medidas de Dispersión. Educación Y Salud Boletín Científico Instituto De Ciencias De La Salud Universidad Autónoma Del Estado De Hidalgo, 9(18), 77-79. https://doi.org/10.29057/icsa.v9i18.7115
- Quevedo, F. (2011). Medidas de tendencia central y dispersión. Medwave, 11(03).
- Ramiro Vásquez, E., Caballero, A. (2011). Inconsistencia del Coeficiente de Variación para expresar la variabilidad de un experimento en un modelo de Análisis de Varianza. Cultivos Tropicales, 32(3), 42-45.