Maestría en Inteligencia Artificial

Aprendizaje Automático Avanzado
Identificación de datos

Dr. Gaddiel Desirena López

1. Variables numéricas y variables categóricas

Variable categórica (cualitativa): Contienen un número finito de categorías o grupos distintos.

- ¿Importa cuán diferentes son dos valores, o solo que son diferentes?
- Una variable categórica podría representar las principales ciudades del mundo, las cuatro estaciones en un año, petróleo, viajes, tecnología, etc.
- El dominio es un conjunto de valores discretos.
- El número de valores de categoría siempre es finito en un conjunto de datos del mundo real.

Variable numérica (cuantitativa): Los valores son números que suelen representar un control o una medición.

- El dominio es el conjunto de valores numéricos.
- La característica numérica puede ser continua o discreta.
- La característica numérica puede ser escalada: $u = 2v$.

Variable ordinal: El dominio es el conjunto de valores ordenados.

- Puede ser numérica o categórica.

Un ejemplo de estas variables se encuentra en la Tabla 1.

Tabla 1: Datos demográficos

Nombre	Edad	Sexo	Estudios
Fernando	32	Masculino	Maestría
Karen	32	Femenino	Maestría
Rosario	58	Femenino	Secundaria
Fernando	59	Masculino	Preparatoria
Carlos	31	Masculino	Doctorado
Marlene	31	Femenino	Mestría
Martín	25	Masculino	Licenciatura

2. Valores faltantes

Los datos faltantes no son raros en conjuntos de datos reales. De hecho, la probabilidad de que falte al menos un punto de datos aumenta a medida que aumenta el tamaño del conjunto de datos. Los datos faltantes pueden ocurrir de varias formas, algunas de las cuales incluyen las siguientes.

Fusión en la fuente de datos: un ejemplo sencillo suele ocurrir cuando dos conjuntos de datos se combinan mediante un identificador de muestra (ID). Si una ID está presente solo en el primer conjunto de datos, entonces los datos combinados contendrán valores faltantes para esa ID para todos los predictores en el segundo conjunto de datos.

Eventos aleatorios: cualquier proceso de medición es vulnerable a eventos aleatorios que impiden la recopilación de datos. Por ejemplo, si una batería se agota o el dispositivo de recolección está dañado, las mediciones no se pueden recolectar y faltarán en los datos finales.

Fallos de medición: por ejemplo, las mediciones basadas en imágenes requieren que una imagen esté enfocada. Otro ejemplo de falla en la medición ocurre cuando un paciente en un estudio clínico pierde una visita médica programada. Las mediciones que se hubieran tomado para el paciente en esa visita faltarían en los datos finales.

2.1. Tipos de valores faltantes

Deficiencias estructurales en los datos: se define como un componente faltante de un predictor que se omitió de los datos. Este tipo de falta es a menudo el más fácil de resolver una vez que se identifica el componente necesario.

Por un caso específico o suceso no aleatorio: Este tipo de datos faltantes son los más difíciles de manejar.

Sucesos aleatorios: Éste se subdivide en dos categorías:

Datos perdidos completamente al azar: la probabilidad de que falte un resultado es igual para todos los puntos de datos (observados o no observados). En otras palabras, los valores perdidos son independientes de los datos. Esta es la mejor situación.

Datos faltantes al azar: la probabilidad de que falten resultados no es igual para todos los puntos de datos (observados o no observados). En este escenario, la probabilidad de que falte un resultado depende de los datos observados pero no de los datos no observados.

3. Cardinalidad de variables categóricas

La cardinalidad de un conjunto es la medida del "número de elementos en el conjunto". Por ejemplo, el conjunto $A = \{2, 4, 6\}$ contiene 3 elementos, y por tanto A tiene cardinalidad 3. La cardinalidad de un conjunto A usualmente se denota $|A|$ o como $\#A$.

Como una aplicación del uso de la cardinalidad, se enuncian los siguientes ejemplos:

Características de color:

Una manera de encontrar la cardinalidad de color es [1]

- Convertir la imagen de tres matrices (R, G, B) a una sola matriz con los valores concatenados RGB.
- Cuantizar los datos obtenidos escogiendo una distancia adecuada (en este ejemplo, se hace una separación cada 0x10 valores).

Ej:
- color1=0x000000-0x000010
- color2=0x000010-0x000020

- ...

Note tres cosas: 1. La codificación es hexadecimal, 2. Cada dos dígitos se representa un color (azul=0x0000FF, rojo=0xFF0000, verde=0x00FF00), 3. La cardinalidad depende de la logitud entre valores que se elija y no del tamaño de la imagen, ya que se enumeran la cantidad de colores que hay en ésta.

Características de textura:
Una de las formas de modelar estas características es la matriz de co–ocurrencia en escala de grises definida como

$$M_{i,j} = \{\#[I_{i,j}]|I_{i,j} = I_{i+1,j+1}\}.$$

Esta forma de modelado evalúa el número de píxeles que comparten el tono de gris, considerando un tono negro como "más profundidad" y un tono claro como "mayor elevación".

4. Relaciones lineales

- Cuantifica como se vinculan dos variables

$$Y = f(X), \quad f : R^p \to R^q$$
$$Y = WX, \quad W \in R^{q \times p}$$

- Este mapeo es independiente del orden de la variables:

$$\{x_1, x_2, x_3, \dots\} \to \{y_1, y_2, y_3, \dots\}.$$
$$\{x_5, x_1, x_6, \dots\} \to \{y_5, y_1, y_6, \dots\}.$$

- Nos permite reformular problemas no–lineales a lineales

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 / x_2^2 + \dots$$
$$\text{con } x_1/x_2^2 = x_3.$$

Por ejemplo, si el conjunto de características Y representa sensaciones de alimentos (picante/caliente, mentolado/fresco) y el conjunto X representa color (rojo, verde y azul)

$$\begin{aligned} Y &= \{1, 2\} \\ X &= \{1, 2, 3\} \end{aligned}$$

La matriz que lo relaciona es

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.55 & 0.3 \end{bmatrix}.$$

Otras relaciones

Estación del año	$\sim$	Temperatura,
Altura	$\sim$	Presión,
Voltaje	$\sim$	Corriente.

Mas se puede obtener una función exacta que, dado un valor de entrada en X, se obtenga un valor de salida en Y. Por ejemplo: Se desea obtener una función lineal que pase exactamente por los datos relacionados $X = [1.1,\ 2]$ y $Y = [-0.9\ 0]$. Únicamente hay dos pares de puntos, por lo que es posible encontrar una función $y = mx + b$ que pase exactamente por ellos. Se prosigue de la siguiente manera:

- Como ya se tiene la estructura de la ecuación, se sustituyen los pares de datos que se desean relacionar

X		Y	$y = mx + b$
1.1	$\sim$	-0.9	$-0.9 = m(1.1) + b$
2	$\sim$	0	$0 = m(2) + b$

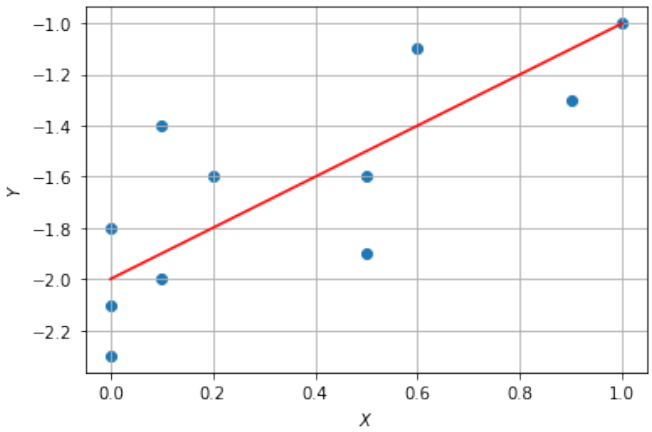

Figura 1: Regresión lineal de los conjuntos X y Y.

- Se tienen entonces dos ecuaciones y dos incógnitas, m y b, se resuelve la siguiente ecuación

$$\begin{bmatrix} 1.1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} -0.9 \\ 0 \end{bmatrix}$$

  resultando $m = 1$ y $b = -2$, por lo que la función es $y = x - 2$.

Lo que significa que, en lugar de usar cualquier valor de y (por ejemplo), se toma el valor correspondiente en x y se le resta 2.

Ahora bien, considere tener dos conjuntos de variables

$$\begin{array}{rcl} X & = & [0, \quad 0.1, \quad 0, \quad 0.2, \quad 0.1, \quad 0, \quad 1, \quad 0.5, \quad 0.6, \quad 0.5, \quad 0.9] \\ Y & = & [-1.8, \quad -2, \quad -2.3, \quad -1.6, \quad -1.4, \quad -2.1, \quad -1, \quad -1.6, \quad -1.1, \quad -1.9, \quad -1.3]. \end{array}$$

Gráficamente lo que se desea es obtener la representación matemática de la línea mostrada en la Figura 1. El objetivo sigue siendo encontrar los valores de m y b en la ecuación $y = mx + b$, pero ahora la línea no pasa exactamente por los puntos, entonces se busca la **mejor** línea que pase entre los puntos. Como criterio, que nos indica si es mejor o peor una función, usaremos el promedio de todos los cuadrados de cada diferencia entre la aproximación de la función propuesta y el valor que se desea relacionar, es decir, para cada $y \in Y$ y $x \in X$, minimizar $\frac{1}{N}\sum^{N}(mx + b - y)^2$, donde N es el número de datos [2].

PYTHON

SCIPY.OPTIMIZE.MINIMIZE(FUN, [M0,B0], ARGS=(X,Y))

LR=SKLEARN.LINEAR_MODEL.LINEARREGRESSION() LR.FIT(X,Y)

5. Distribución de datos

Es la agrupación de datos en diferentes categorías indicando el número de observaciones [3, 4]:

$\{1, 1, 2, 1, 3, 2, 1, 4, 2, 3\}$

Dato	Observaciones
1	4
2	3
3	2
4	1

PYTHON:

PANDAS.DATAFRAME.HIST()

MATPLOTLIB.PYPLOT.HIST(X) -> (CONTEO, LIMITES)

En un conjunto grande de datos, la representación en una tabla resulta poco útil. El objetivo, a partir de un conjunto de datos, es obtener un conjunto pequeño de números que resuman bien a éste a partir de **medidas de posición**, de **dispersión** y de **forma**.

5.1. Medidas de posición o de tendencia central

Estas medidas centralizan la información, también se les conoce como promedios:

- Media aritmética. La media aritmética se ve muy alterada por valores extremos de la variable.

$$x = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

PYTHON:

NUMPY.MEAN(X)

- Media recortada (α-trimmed) es la media aritmética calculada quitando el α por ciento de los datos inferiores y superiores.

PYTHON:

SCIPY.STATS.TRIM_MEAN(X,*PORCENTAJE*)

SCIPY.STATS.TMEAN(X,(MIN,MAX))

- Media ponderada. Es el promedio de cada muestra multiplicando a cada una por un peso

$$x = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_n w_n}{w_1 + w_2 + \cdots + w_n}$$

PYTHON

NUMPY.AVERAGE(X,W)

- Media geométrica. Es la raíz n-ésima del producto de todos los datos

$$x = \sqrt[n]{x_1 x_2 \cdots x_n}$$

PYTHON

SCIPY.STATS.GMEAN(X)

- Media armónica. Es el inverso de la media aritmética de los valores inversos.

$$x^{-1} = \frac{1}{n} \sum_{i=1}^{n} x_i^{-1}$$

Interesante en la media de las velocidades. PYTHON

SCIPY.STATS.HMEAN(X)

- Mediana. El valor central de los datos ordenados. Presenta robustez, no se deja afectar por valores extremos.

PYTHON

NUMPY.MEDIAN(X)

- Moda. El valor que más veces se repite.

PYTHON

SCIPY.STATS.MODE(X)

- Cuantil o percentil. Aquel valor que divide a la variable en dos partes, dejando a su izquierda el $p\,\%$ de los datos y a su derecha el $100 - p\,\%$.

Cuartiles. Si consideramos los percentiles 25, 50 y 75, los datos son divididos en cuatro partes y suelen llamarse **primer cuartil** Q_1, **segundo cuartil** o mediana Q_2 y **tercer cuartil** Q_3. Así mismo, los datos se pueden dividir en diez partes, llamando a cada una de ellas **deciles**, éstos se representan como D_1, D_2, ..., D_9.

PYTHON

NUMPY.PERCENTILE(X,25)

NUMPY.QUANTILE(X,0.25)

PANDAS.DATAFRAME.QUANTILE(0.25)

5.2. Medidas de dispersión

Lo procedente es saber qué fiabilidad nos ofrecen esas pocas cantidades o números, es decir, cuánta variabilidad existe en el conjunto de datos. Si hay poca variabilidad, la información de los valores medios será muy precisa. Si existe mucha variabilidad, la información será menos precisa.

- Varianza y desviación estándar. Son las medidas de dispersión más importantes, estando íntimamente ligadas a la media aritmética como medida de representación de ésta.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$$

PYTHON

NUMPY.VAR(X)

NUMPY.STD(X)

- Rango. Se define como la diferencia entre los valores máximo y mínimo

$$\text{máx}(X) - \text{mín}(X)$$

PYTHON

NUMPY.PTP(X)

SCIPY.STATS.IQR(X,RNG=(0,100)

- Rango intercuartílico o IQR. Se define como la diferencia entre los cuartiles Q_3 y Q_1

$$R_i = Q_3 - Q_1$$

PYTHON

SCIPY.STATS.IQR(X)

- Coeficiente de variación (Pearson). Se define como el cociente entre la desviación estándar y el valor absoluto de la media aritmética.

$$cv = \frac{\sigma}{|x|}$$

Permite comparar la dispersión de varias distribuciones sin importar la unidades de éstas.

PYTHON

SCIPY.STATS.VARIATION(X)

A excepción de este último coeficiente, el resto de las medidas depende de las unidades en las que se expresen los datos (metros, kilómetros, segundos, horas,...), haciendo imposible compara las muestras. Más adelante se retoma este tema para hacer posible la comparación entre diferentes variables que tengan en cuenta el tamaño de las observaciones.

5.3. Medidas de forma

- Simetría: Una distribución o variable es simétrica si, sobre la media o mediana, los datos se distribuyen de igual forma a la izquierda o a la derecha. Si una distribución es simétrica, la media aritmética y la mediana van a coincidir.

  Coeficiente de Fisher (g_1) es la relación entre el momento central estandarizado de orden 3 (m_3) y la desviación estandar al cubo (σ^3).

$$g_1 = \frac{m_3}{\sigma^3},$$

donde $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x - x_i)^3$.

Si $g_1 < 0$ la variable presenta una "cola" alargada hacia la izquierda y se dice que presenta una asimetría izquierda, si $g_1 > 0$ la asimetría es derecha, si $g_1 = 0$ la variable es simétrica.

PYTHON

PANDAS.DATAFRAME.SKEW()

- Apuntamiento o Curtosis: Es la concentración en la zona central. Se calcula como el segundo coeficiente de Fisher:

$$g_2 = \frac{m_4}{\sigma^4}$$

Este valor se compara con la curva gausiana, cuya curtosis es 3. Si $g_2 < 3$ la variable es platicúrtica (menos apuntada que la normal), si $g_2 > 3$ es leptocúrtica (la variable es más apuntada que la normal), si $g_2 = 3$ ésta es mesocúrtica (tiene el mismo apuntamiento que la normal).

PYTHON

PANDAS.DATAFRAME.KURT()

Notita: Este método resta 3 al valor, comparando la normal con 0.

6. Valores atípicos

Los valores que se alejan "demasiado" de donde se acumulan el común de los valores, se les considera atípicos. Un criterio para identificarlos es el diagrama de caja.

6.1. Diagramas de caja

Sirve para visualizar tanto la dispersión como la forma del conjunto de datos.

- A los datos que se encuentren a una distancia de Q_1 por la izquierda, o de Q_3 por la derecha, superior a 1.5 veces el recorrido intercuartílico $R_i = Q_3 - Q_1$, se le llaman **atípicos** de primer nivel.
- Cuando la distancia, por uno de los dos lados, es superior a $3R_i$, el valor atípico se denomina de segundo nivel, o **dato extremo**.

La Figura 2 muestra un ejemplo descriptivo del diagrama de caja.

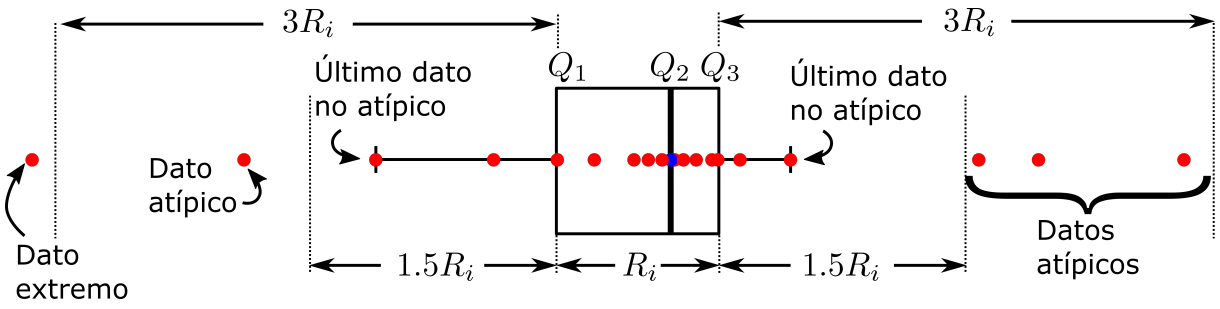

Figura 2: Características de un diagrama de caja

PYTHON

PANDAS.DATAFRAME.BOXPLOT()

Referencias

[1] Chandakkar, P. S., Venkatesan, R., & Li, B. (2018). Feature Extraction and Learning for Visual Data. In Feature Engineering for Machine Learning and Data Analytics (pp. 55-85). CRC Press.

[2] `https://scikit-learn.org/stable/modules/linear_model.html`

[3] https://bookdown.org/aquintela/EBE/

[4] https://www.odiolaestadistica.com/estadistica-python/

[5] https://docs.scipy.org/doc/scipy/reference/stats.html