ITESO, Universidad
Jesuita de Guadalajara

Módulo I: Minería de Datos

Dr. Gaddiel Desirena López

Ciencia de Datos e Inteligencia de Negocios Primavera 2021

Contenido

Variables numéricas y variables categóricas

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

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

Tipos de características

Categórica: El dominio es un conjunto de valores discretos.

Ordinal: El dominio es el conjunto de valores ordenados.

Numérica: El dominio es el conjunto de valores numéricos. La característica numérica también es llamada contínua. La característica numérica puede ser escalada: $u = 2v$.

Variables numéricas y variables categóricas

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

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.

Valores faltantes

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.

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.

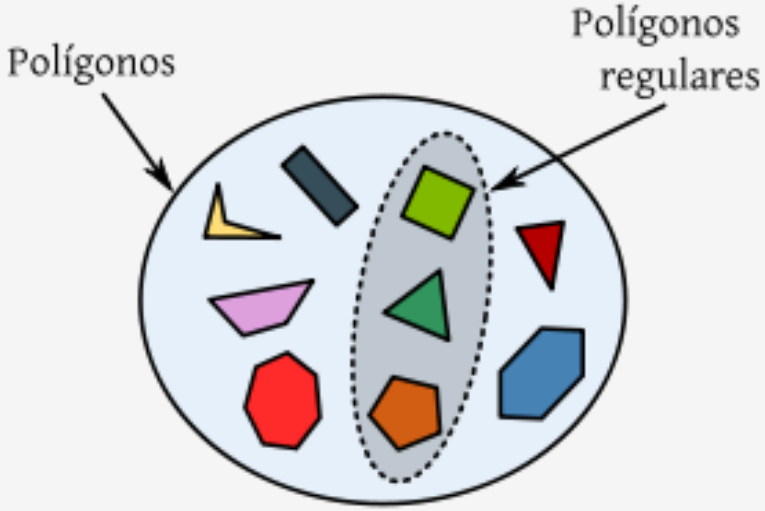
Por un caso específico o suceso no aleatorio: Este tipo de datos faltantes son los más difíciles de manejar.

Valores faltantes

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.

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 $\#A$.

Cardinalidad de variables categóricas

Características de color

Una manera de encontrar la cardinalidad de color es

Características de textura

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

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

Cardinalidad de variables categóricas

Características de color

Una manera de encontrar la cardinalidad de color es

- convertir la imagen (R,G,B) a un solo canal RGB.
- Cuantizar los datos obtenidos.

Características de textura

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

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

Cardinalidad de variables categóricas

Características de color

Una manera de encontrar la cardinalidad de color es

- convertir la imagen (R,G,B) a un solo canal RGB.
- Cuantizar los datos obtenidos.

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

Características de textura

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

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

Características de forma

Para representar este conjunto se pueden detectar los puntos p en el contorno de una imagen. Las características de éste se computan como la cardinalidad de los vectores que conectan p con los demás puntos:

$$|\{p \neq q_i | (p - q_i) \in bin\{k\}\}|$$

Exploratory Data Analysis (EDA)

Eda es la sigla en inglés para Exploratory Data Analysis y consiste en una de las primeras tareas que tiene que desempeñar el Científico de Datos. Es cuando revisamos por primera vez los datos que nos llegan, por ejemplo un archivo CSV que nos entregan y deberemos intentar comprender "¿de qué se trata?", vislumbrar posibles patrones y reconociendo distribuciones estadísticas que puedan ser útiles en el futuro.

Técnicas para el EDA

Dado que cada conjunto de datos suele ser único, el EDA se hace bastante "a mano", pero podemos seguir diversos pasos ordenados para intentar acercarnos a ese objetivo en pocas horas.
A nivel programación y como venimos utilizando Python, encontramos a la conocida librería Pandas, que nos ayudará a manipular datos, leer y transformarlos.

Exploratory Data Analysis (EDA)

Algunas pruebas que se pueden hacer con los datos en un EDA son:

- Si hay datos categóricos, agruparlos, contabilizarlos y ver su relación con las clases de salida
- gráficas de distribución en el tiempo, por ejemplo si tuviéramos ventas, para tener una primera impresión sobre su estacionalidad.
- Rankings del tipo “10 productos más vendidos” ó “10 ítems con más referencias por usuario”.
- Calcular importancia de Features y descartar las menos útiles.

Data Qual ity Report (DQR)

Data Cleaning

Medidas de Similitud