Diplomado Ciencia de Datos
Módulo I:
Introducción, Manipulación, Exploración y Visualización de Datos

Carla Paola Malerva Reséndiz

19 de noviembre de 2020

Índice

1. Calidad de datos

La calidad de los datos le permite preparar y gestionar los datos, al tiempo que los pone a disposición de toda su organización. Los datos de alta calidad permiten a los sistemas estratégicos integrar todos los datos relacionados para proporcionar una visión completa de la organización y las interrelaciones dentro de la misma.
La calidad de los datos es una característica esencial que determina la confiabilidad de la toma de decisiones.
No existen estandarizaciones, ni una talla única en lo que se refiere a data quality, es más una percepción.

Las seis dimensiones de la calidad de datos

- Completitud:
  Se refiere a cuando todos los campos y registros están dentro del conjunto de datos, no existen espacios en blanco.
- Conformidad:
  Los datos deben estar en un formato estándar y legible.
- Consistencia:
  No debe existir información contradictoria.
- Duplicación:
  De acuerdo a la unidad muestral, cada registro debe ser distinto, no podemos contar con la misma información más de una vez.
- Integridad:
  ¿Toda la información relevante de un registro está presente de forma que se pueda utilizar?
- Precisión:
  Si los datos no son precisos, estos no pueden ser utilizados. Para detectar si estos son preciso, se compara el dato con una fuente de referencia

Llevar a cabo una gestión de la calidad de datos adecuada depende de saber cómo medirla

Beneficios:

- Fácil procesamiento
- Fácil análisis
- Ayuda a la organización a tomar mejores decisiones
- Mejor eficiencia operativa

2. Conversión de estructuras OLAP y OLTP a TAD, ingeniería de características

- OLTP
  El procesamiento de transacciones en línea, conocido en breve como OLTP, admite aplicaciones orientadas a transacciones en una arquitectura de 3 niveles. OLTP administra las transacciones diarias de una organización. El objetivo principal es el procesamiento de datos y no el análisis de datos.

  Un ejemplo de sistema OLTP : Suponga que una pareja tiene una cuenta conjunta en un banco. Un día, ambos llegan simultáneamente a diferentes cajeros automáticos exactamente a la misma hora y quieren retirar el monto total presente en su cuenta bancaria.
  Sin embargo, la persona que complete el proceso de autenticación primero podrá obtener dinero. En este caso, el sistema OLTP se asegura de que la cantidad retirada nunca supere la cantidad presente en el banco. La clave a tener en cuenta aquí es que los sistemas OLTP están optimizados para la superioridad transaccional en lugar del análisis de datos.

- OLAP
  Procesamiento analítico en línea, una categoría de herramientas de software que proporcionan análisis de datos para decisiones comerciales. Los sistemas OLAP permiten a los usuarios analizar información de bases de datos de múltiples sistemas de bases de datos a la vez.
  El objetivo principal es el análisis de datos y no el procesamiento de

datos.

Un ejemplo de sistema OLAP puede ser :Amazon analiza las compras de sus clientes para crear una página de inicio personalizada con productos que probablemente interesen a sus clientes.

- TAD (Tabla análitica de datos)
  Es una tabla plana que se utiliza para construir modelos analíticos de aprendizaje máquina y realizar modelación supervisada o no supervisada. Un registro único en esta tabla representa el tema del análisis (por ejemplo, un cliente) y almacena todos los datos (variables) que describen este tema. TAD puede desarrollarse como una instancia más general aplicable para resolver problemas comerciales generales, pero con mayor frecuencia se desarrolla para resolver problemas comerciales muy específicos.

2.1. Ingeniería de características

- Codificación a nivel nominal
  El método principal que tenemos es transformar nuestro datos categóricos en variables "dummy". Las variables "dummy"toman el valor cero o uno para indicar la ausencia o presencia de un categoría. Son suplentes numéricos, para datos cualitativos.
  Ejemplo: Género : "hombre","mujer", en este caso usamos un código ficticio cuando la persona es mujer = 1, y mujer = 0 cuando la persona es hombre.
  Al trabajar con variables "dummy", es importante evitar información duplicada.

- Codificación a nivel ordinal
  En estas variables todavía hay información útil, sin embargo, necesitamos transformar las cadenas en datos numéricos. En el nivel ordinal, dado que hay un significado en los datos que tienen un orden específico, no tiene sentido usar dummies. Para mantener el orden, usaremos un codificador de etiquetas.
  Por codificador de etiquetas, queremos decir que cada etiqueta en nuestros datos ordinales tendrá un valor numérico asociado. En nuestro ejemplo, esto significa que los valores de la columna ordinal (no me gusta, intermedio, me gusta) se representarán como 0, 1 y 2.

- Ingeniería a nivel continuo
  Para las variables de tipo continuo existen muchas opciones para la generación de nuevas variables a partir de la información existente, algunas de las opciones son: generación de ventanas de tiempo, añadir información de periodos anteriores, creación de características que consta de todas las combinaciones polinomiales de las características con grado menor o igual al grado especificado.

- Ingeniería para texto
  Hasta este punto, hemos estado trabajando con datos categóricos y numéricos. Si bien nuestros datos categóricos se han presentado en forma de una cadena, el texto ha sido parte de una sola categoría. Ahora profundizaremos en datos de texto de formato más largo. Esta forma de datos de texto es mucho más compleja que el texto de una sola categoría, porque ahora tenemos una serie de categorías o tokens.
  Antes de seguir trabajando con datos de texto, asegurémonos de comprender bien lo que queremos decir cuando nos referimos a datos de texto. Considere un servicio como Yelp, donde los usuarios escriben reseñas de restaurantes y negocios para compartir sus pensamientos sobre su experiencia. Estas revisiones, todas escritas en formato de texto, contienen una gran cantidad de información que sería útil para fines de aprendizaje automático, por ejemplo, para predecir el mejor restaurante para visitar.

  Este tipo de trabajo puede denominarse procesamiento del lenguaje natural (NLP), es importante tener en cuenta que todos los modelos de aprendizaje automático requieren entradas numéricas, por lo que debemos ser creativos y pensar estratégicamente cuando trabajamos con texto y convertimos esos datos en características numéricas. Hay varias opciones para hacer.

  - CountVectorizer
    Es el método más utilizado para convertir datos de texto en sus representaciones vectoriales. Es similar a las variables "dummy", en el sentido de que CountVectorizer convierte columnas de texto en matrices donde las columnas son tokens y los valores de celda son recuentos de apariciones de cada token en cada documento. La matriz resultante se conoce como matriz documento-término

porque cada fila representará un documento.

- Tf-idf vectorizer
  Un Tf-idfVectorizer se puede dividir en dos componentes.
  Primero, la parte tf, que representa la frecuencia del término, y la parte idf, que significa frecuencia inversa del documento. Es un método de ponderación de términos que tiene aplicaciones en la recuperación y agrupación de información.
  Se da un peso para evaluar qué tan importante es una palabra para un documento en un corpus. Veamos cada parte un poco más:

  - tf:
    frecuencia del término: mide la frecuencia con la que aparece un término en un documento. Dado que los documentos pueden tener una longitud diferente, es posible que un término aparezca muchas más veces en documentos más largos que en documentos más cortos. Por lo tanto, la frecuencia de los términos a menudo se divide por la longitud del documento, o el número total de términos en el documento, como una forma de normalización.
  - idf:
    frecuencia inversa del documento: mide la importancia de un término. Al calcular la frecuencia de los términos, todos los términos se consideran igualmente importantes. Sin embargo, ciertos términos, como "la", "deƶ ƶ", pueden aparecer muchas veces pero tienen poca importancia. Por lo tanto, debemos ponderar menos los términos frecuentes, mientras que ampliamos los raros.

  Para volver a enfatizar, un TfidfVectorizer es lo mismo que CountVectorizer, en el sentido de que construye características a partir de tokens, pero va un paso más allá y normaliza los conteos a la frecuencia de ocurrencias en un corpus.

3. Visualización de datos

La visualización de datos es la práctica de traducir información en un contexto visual, como un mapa o gráfico, para facilitar que el cerebro hu-

mano comprenda y extraiga información útil. El objetivo principal de la visualización de datos es facilitar la identificación de patrones, tendencias y valores atípicos en grandes conjuntos de datos. El término a menudo se usa indistintamente con otros, incluidos gráficos de información, visualización de información y gráficos estadísticos.

La visualización de datos es uno de los pasos del proceso de ciencia de datos, que establece que una vez que los datos se han recopilado, procesado y modelado, deben visualizarse para obtener conclusiones.
La visualización de datos proporciona una forma rápida y efectiva de comunicar información de manera universal utilizando información visual.

Beneficios de la visualización de datos:

- Capacidad de absorber información rápidamente, mejorar los conocimientos y tomar decisiones más rápidas
- Mayor comprensión de los próximos pasos que deben tomarse para mejorar la organización
- Capacidad mejorada para mantener el interés de la audiencia con información que puedan entender
- Distribución fácil de información que aumenta la oportunidad de compartir ideas con todos los involucrados

4. Análisis exploratorio univariado de variables continuas y discretas.

El análisis exploratorio de datos es un enfoque para analizar conjuntos de datos para resumir sus principales características, a menudo con métodos visuales,también se refiere al proceso crítico de realizar investigaciones iniciales sobre datos para descubrir patrones, detectar anomalías, probar hipótesis y verificar suposiciones con la ayuda de estadísticas resumidas y representaciones gráficas.
De acuerdo al nivel de dato se pueden obtener distinta información o una representación visual adecuada.

- Nivel Nominal
  Consiste en datos que se describen puramente por su nombre. Los ejemplos básicos incluyen el tipo de sangre (A, O, AB), especies de animales o nombres de personas. Todos estos tipos de datos son cualitativos.
  En este nivel, no podemos realizar ninguna operación matemática cuantitativa, como suma o división. No podemos encontrar un valor medio a nivel nominal. Sin embargo, podemos hacer recuentos básicos.
  Las representaciones usuales para este tipo de nivel de datos son gráficas de pastel o gráfica de barras.

- Nivel Ordinal
  La escala ordinal hereda todas las propiedades del nivel nominal, pero tiene importantes propiedades adicionales:
  Los datos en el nivel ordinal pueden ordenarse naturalmente, esto implica que algunos valores de los datos en la columna pueden considerarse mejores o mayores que otros, al igual que con el nivel nominal, los datos en el nivel ordinal siguen siendo de naturaleza categórica.
  En el nivel ordinal, todavía podemos hacer recuentos básicos como lo hicimos en el nivel nominal, pero también podemos introducir comparaciones y ordenaciones en la mezcla. Por esta razón, podemos utilizar nuevos gráficos en este nivel. Podemos usar gráficos de barras y circulares como lo hicimos en el nivel nominal, pero debido a que ahora tenemos ordenamiento y comparaciones, podemos calcular medianas y percentiles. Con medianas y percentiles, son posibles los diagramas de caja.

- Nivel de continuo
  Con la capacidad de sumar valores, podemos introducir dos conceptos familiares, la media aritmética (denominada simplemente media), la desviación estándar, mediana, percentiles, mínimo, máximo y conteos.
  Para este nivel se puede hacer uso de histogramas, visualizaciones de dispersión, diagramas de caja y bigotes, mapas de calor, entre otros más.

5. Valores Atípicos

Un valor atípico es algo que es extraño o diferente de la multitud. Algunos estadísticos definen los valores atípicos como "tener un comportamiento

subyacente diferente al del resto de los datos". Alternativamente, un valor atípico es un punto de datos que está distante de otros puntos.
El valor atípico puede deberse sólo a una variabilidad en la medición o puede indicar errores experimentales.

Los valores atípicos se introducen por primera vez en la población al recopilar los datos o se refiere a información errónea. Si es posible, los valores atípicos deben excluirse del conjunto de datos. Sin embargo, la detección de instancias anómalas puede resultar difícil y no siempre es posible.

Existen dos tipos de valores atípicos:

- Puntuales o Globales
  Son aquellos valores donde su valor está muy por fuera de la totalidad del conjunto de datos en el que se encuentra. Por ejemplo : En una clase, la edad de todos los estudiantes será de aprox. similar, pero si ve un registro de un estudiante con una edad de 500 años ese valor se considerará como un valor atípico global.

- Contextuales
  Observaciones consideradas anómalas dado un contexto específico. Un punto de datos se considera un valor atípico contextual si su valor se desvía significativamente del resto de los puntos de datos en el mismo contexto. Por ejemplo:
  El mercado de valores se desploma debido a la estafa en 1992 o en 2020 debido a COVID-19. Los puntos de datos habituales están cerca unos de otros, mientras que los puntos de datos durante el período específico subirán o bajarán mucho. Esto no se debe a errores, son punto de datos de observación real. Estos datos son valores atípicos contextuales.

Es importante identificar y eliminar los valores atípicos ya que los algoritmos de aprendizaje automático son sensibles al rango y distribución de valores de atributos. Los datos atípicos pueden estropear y engañar el proceso de entrenamiento, lo que da como resultado tiempos de entrenamiento más prolongados, modelos menos precisos.

Existen dos alternativas para identificar si existen valores atípicos:

- Visualización
  Consiste en crear gráficos de caja y bigotes, histogramas o gráfica de dispersión para identificar si existen valores que tienen un comportamiento diferente al resto de los demás.

- Funciones matemáticas:

  - Análisis de percentiles.
    La técnica consiste en definir un corte inferior y uno superior basado en los percentiles ,por ejemplo tomar como punto inferior al percentil 5 y como punto superior al percentil 95 , de tal forma que los valores que estén por debajo del percentil 5 se considerarán como valores atípicos y de igual forma con el percentil 95, los valores por encima de dicho percentil se considerarán como valores atípicos.
  - IQR Score
    Medida de dispersión estadística que es igual a la diferencia entre el percentil 75 y el 25. Una regla de uso común dice que un punto de datos es un valor atípico si es más de 1.5IQR por encima del tercer cuartil o por debajo del primer cuartil. Dicho de otra manera, los valores atípicos bajos están debajo Q1 -1.5IQR y los valores atípicos están por encima de Q3+1.5IQR. Cuando la escala se toma como 1,5 de acuerdo con el método IQR, cualquier dato que se encuentre más allá de 2,7 desviaciones de la media, en cualquier lado, se considerará un valor atípico. Y este rango de decisión es el más cercano a lo que nos dice la Distribución Gaussiana, es decir, 3 desviaciones. En otras palabras, esto hace que la regla de decisión se acerque más a lo que la distribución gaussiana considera para la detección de valores atípicos.
  - Z-score
    El puntaje estándar o puntaje z es el número con signo de desviaciones estándar por el cual el valor de una observación o punto de datos está por encima del valor medio de lo que se está observando o midiendo.Se buscan los datos que están demasiado lejos

de cero (outliers) , el umbral usual es de -3 a 3 desviaciones (Una limitación de este método es que solo se puede utilizar cuando los datos no están muy sesgados. Requiere que los datos estén cerca de lo normal.)

6. Técnicas de imputación de valores ausentes

Los datos faltantes se definen como valores no disponibles que serían útiles o significativos para el análisis de los resultados.

Hay muchos tipos de datos faltantes y muchas razones por las cuales pueden ocurrir. En los conjuntos de datos, los valores faltantes usualmente se pueden representar como '?', 'Nan', 'N / A', celda en blanco o, a veces, '-999', 'inf', '-inf'.

Existen dos tipos de valores faltantes:

- Errores visibles:
  Son valores como celdas en blanco, NA(no disponible), NaN (no es un número)
- Errores ocultos:
  Valores los cuales están fuera de la naturaleza de la variable, por ejemplo si se está trabajando con la variable salario y encontramos registros donde el salario es negativo , estos valores se considerarán como valores faltantes.

Los valores faltantes presentan un obstáculo a la hora de crear modelos predictivos, análisis de clusters. Ya que muchos modelos no lidian con ellos y si se opta por eliminarlos se estará perdiendo información, a menos que el porcentaje de valores faltantes sea alto. Dado esto existen distintas técnicas de imputación de valores ausentes.

De acuerdo a la naturaleza de cada variable es que se utiliza un método u otro.

- Media
  Consiste en imputar los valores ausentes con el promedio de los valores existentes. Si los valores aún presentan valores extremos la media podría verse afectada y en consecuencia los valores imputados, es importante tener presente lo mencionado anteriormente, por otro lado si se realizó un conjunto de entrenamiento y uno de prueba , la forma correcta de imputación es utilizando en este caso la media de los datos de entrenamiento para imputar los registros de entrenamiento y prueba.

- Mediana
  Consiste en obtener la mediana , es decir, el percentil 50 de los datos existentes y posteriormente imputar los valores ausentes con dicho valor.Una ventaja de este método es que si existen valores extremos en los datos , la mediana no se verá tan afectada.

- Moda
  En comparación a los demás método es uno de los más sofisticados, de la misma forma resulta ser el más costoso de implementar , la idea detrás de este método es entrenar un modelo haciendo uso del resto de características y definiendo como variable objetivo aquella variable que posee valores ausentes ,de esta forma a partir de las demás características se podría predecir los valores faltantes.

- Creación de categoría missing
  Para algunas soluciones que son particulares se crea una nueva categoría que lleva por nombre "missings", esa es la forma de lidiar con los valores ausentes, creando una nueva categoría.

- Interpolación lineal
  La interpolación lineal es un procedimiento muy utilizado para estimar los valores que toma una función en un intervalo del cual conocemos sus valores en los extremos (x1, f(x1)) y (x2,f(x2)). Para estimar este valor utilizamos la aproximación a la función f(x) por medio de una recta r(x). Este método es útil cuando se tienen series de tiempo, de esta forma se usarán los valores extremos al valor ausente para estimar el valor faltante. Además de interpolación lineal se puede ajustar un spline para estimar aquellos valores faltantes.

- Siguiente valor

Consiste en tomar el siguiente valor no nulo o e imputarlo en los valores anteriores. Este método es util en variables que dependen del tiempo.

- Anterior valor
  Consiste en tomar el último valor no nulo o e imputarlo en los valores posteriores. Este método es util en variables que dependen del tiempo.

Para poder seleccionar el mejor método se pueden realizar pruebas de bondad de ajuste como lo es la prueba de Kolmogorov-Smirnov de la variable original con la variable imputada y de esa forma elegir aquel que obtenga mejores resultados en la prueba.

7. Reducción de dimensionalidad.

En el aprendizaje automático la reducción de dimensionalidad es el proceso de reducción del número de variables iniciales. A medida que aumenta la generación y recopilación de datos se vuelve más desafiante y difícil visualizar o hacer inferencias sobre los mismos, dado esto es de gran utilidad hacer una selección de características relevantes o proyectar los datos a un subespacio de menor dimensión de tal forma que se capte la esencia de los datos.

Es importante realizar reducción de dimensiones ya que se podrían obtener distinto beneficios al hacerlo, por ejemplo:

- Menos dimensiones en el conjunto de datos podría conducir a un menor tiempo de entrenamiento en un modelo de aprendizaje máquina.
- Con algunas técnicas se puede evitar el la multicolinealidad , esto se podría conseguir al eliminar variables redundantes
- Ayuda a la visualización de datos, si se cuenta con un número de variables muy amplio es difícil realizar una representación gráfica de los datos , sin embargo con algunas técnicas se puede reducir a un espacio menor para permitirnos trazar y observar patrones con mayor claridad.

La reducción de dimensiones se puede realizar de dos formas diferentes:

- Selección de características Se refiere a mantener solo las variables más relevantes del conjunto de datos original.

- Reducción de dimensionalidad Generar un conjunto más pequeño de nuevas variables , donde cada una de las cuales es una combinación de las variables originales.

7.1. Relación de valor perdido

Consiste en eliminar aquellas variables que no contienen mucha información , es decir , se cuenta con un gran número de valores faltantes.

7.2. Baja Varianza

Se pueden omitir aquellas variables donde todas las observaciones tienen el mismo valor, a las cuales se les denomina variables unitarias , además se pueden omitir aquellas que tengan una varianza muy baja, este tipo de variable no aportan mucha información.

7.3. Alta correlación entre características

La alta correlación entre dos variables nos indica que tienen tendencias similares y es probable que contengan información similar. Si el coeficiente de correlación cruza un cierto valor de umbral se podría hacer la selección de una de las variables altamente correlacionadas.

7.4. Correlación con el objetivo

Identificar aquellas variables que tengan una correlación muy baja con el objetivo. Si una variable tiene una corrección muy baja con el objetivo, existe la posibilidad de que no sea muy relevante o no aporte mucha información al modelo(predicción).

7.5. Análisis de multicolinealidad

La multicolinealidad en regresión es una condición que ocurre cuando algunas variables predictoras incluidas en el modelo están correlacionadas con otras variables predictoras.
“La multicolinealidad ocurre cuando dos o más variables independientes están altamente correlacionadas entre sí en un modelo de regresión.” Esto significa que se puede predecir una variable independiente a partir de otra variable

independiente en un modelo de regresión.

La multicolinealidad puede incrementar la varianza de los coeficientes de regresión, haciéndolos inestables.
Las siguientes son algunas de las consecuencias de los coeficientes inestables:

- Los coeficientes pueden parecer insignificantes incluso cuando exista una relación significativa entre el predictor y la respuesta.
- Los coeficientes de los predictores muy correlacionados varían ampliamente de una muestra a otra.
- La eliminación de cualquier término muy correlacionado del modelo afectará considerablemente los coeficientes estimados de los demás términos muy correlacionados. Los coeficientes de los términos muy correlacionados incluso pueden tener el signo equivocado.

Existen distintas causas por la cuales se puede presentar multicolinealidad:

- Experimentos mal diseñados
- Creación de nuevas variables
- Incluyendo variables idénticas en el conjunto de datos
- Uso inexacto de variables dummy

Para medir la multicolinealidad se puede examinar los factores de inflación de la varianza (VIF).La puntuación VIF de una variable independiente representa qué tan bien la variable se explica por otras variables independientes.

Si el VIF tiene valores por debajo de 5 indica que no existe multicolinealidad, por otro lado si superan el umbral de 5 o 10 , nos indicaría que existe multicolinealidad en nuestros datos.

La multicolinealidad no afecta la bondad de ajuste ni la bondad de predicción. Los coeficientes (función discriminante lineal) no pueden interpretarse de forma fiable, pero los valores (clasificados) ajustados no se ven afectados.

7.6. Análisis de componentes principales

Método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.

Cada componente principal (Zi) se obtiene por combinación lineal de las variables originales. Se pueden entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales.

Una vez calculada la primera componente (Z1) se calcula la segunda (Z2) repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no puede estar correlacionada con la primera componente.

Ayuda a eliminar la multicolinealidad, pero la explicabilidad se ve comprometida.

El escalado de variables es muy importante ya que el proceso de PCA identifica aquellas direcciones en las que la varianza es mayor. Como la varianza de una variable se mide en su misma escala elevada al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media 0 y desviación estándar 1, aquellas variables cuya escala sea mayor dominarán al resto. De ahí que sea recomendable estandarizar siempre los datos.

7.7. Análisis de varianza explicada

La varianza explicada (también llamada variación explicada) es usada para medir la discrepancia entre un modelo y los datos actuales. En otras palabras, de un conjunto de datos original, cuánta varianza explica un conjunto de sus variables o componentes principales, esto con el objetivo de mantener la mayor varianza posible y así reducir la dimensión del problema. En análisis de componentes principales nos interesa conocer la proporción de varianza explicada por cada uno de los componentes principales, o dicho de otra manera, cuanta información presente en los datos se pierde por la proyección de las observaciones sobre los primeros componentes principales.

7.8. Importancia de variables

Se refiere a que tan importantes o influyentes son las variables para un modelo, en este caso, que tan se puede prescindir de algunas de ellas o bien, cuales son esenciales mantener. Para conocer la importancia de variables , los modelos basados en árboles son los más utilizados para la selección de características dada su importancia.Esto nos ayuda a seleccionar un subconjunto más pequeño de características

7.9. Poder predictivo de características.

7.9.1. Transformación entrópica