Lección 1. Tratamiento de datos estructurados.

Al capturar información, no importa el medio, esta tendrá diferentes problemas como pueden ser datos nulos, mal formados, estructuras no válidas, entre otros. En esta lección se aprenderán los conceptos necesarios para que el estudiante esté en capacidad de tomar las decisiones apropiadas en cuanto a qué decisiones tomar cuando un conjunto de datos contiene alguna o todas las condiciones descritas.

Según la forma en la que se construyen los datos y las relaciones que tienen unas variables con otras, se pueden clasificar los datos en dos grupos.

Datos estructurados, que corresponden a aquellos datos que tienen un modelo predefinido, en donde hay claramente una imposición de cómo deben almacenarse los datos, qué tipo de dato debe tener cada valor (numérico, texto, entre otros) y las relaciones que ciertos valores tienen con otros. A esta forma de guardar la información se le conoce como datos estructurados.

También existen datos que no poseen una estructura clara para un ordenador. Tal es el caso de ficheros como grabaciones de audio, imágenes y videos. En estos casos no existe una estructura rígida de almacenamiento de la información como es el caso de las tablas y las bases de datos. Por lo tanto, para un computador será muy complejo determinar qué contienen estos datos.

Dependiendo del tipo de datos que se tenga, el procesamiento cambia. En el caso de los datos no estructurados, se suelen ajustar las longitudes de los archivos de audio, recortar componentes en frecuencia y extraer características de tamaños fijos para darles una estructura reconocible por un procesador. De la misma forma, sucede con las imágenes y los videos. Al preprocesar los datos es necesario crear estructuras por medio de algoritmos de extracción de características o incluso correr modelos de aprendizaje de máquina que tengan la capacidad de determinar (con cierto margen de error) qué contiene un video o una imagen. También se requiere la estandarización de la información en formatos que posean un mismo tamaño de imagen (ancho por alto), un nivel de codificación del color similar, entre otros aspectos a tener en cuenta.

De igual forma, en los datos estructurados es necesario cierto preprocesamiento antes de iniciar con la fase de exploración y análisis de la información. La necesidad surge de que los datos suelen venir con información problemática.

Desde la captura de la información a través de instrumentos como pueden ser sensores, simuladores, encuestas, evaluaciones, observaciones, entre muchas otras formas de adquirir datos; los datos vienen asociados a una incertidumbre inherente al proceso de captura, es decir, ruido y procesos que distorsionan la información por el solo hecho de medirlos.

Dependiendo del tipo de distorsión podemos establecer que proviene de ruido, el cual siempre existirá, o puede provenir de fuentes que sí se pueden mitigar (error sistemático). Por ejemplo, un instrumento descalibrado, una encuesta que tenga un valor por defecto sin requerir que sea llenado para poder guardar, una persona que se descuida al realizar las observaciones y se pasa por alto eventos que no observa o incluso, un programa que tenga un error de cálculo implícito y difícil de descubrir. Todas las fuentes de error sistemático se pueden eliminar, sin embargo, los datos pueden haber sido almacenados con tales imprecisiones.

La información luego de ser capturada se suele transmitir (o almacenar) en medios digitales. Esto también puede inducir al error ya que un valor puede redondearse o aproximarse involuntariamente, un fichero puede transmitirse por una red de datos y el ruido del ambiente puede alterar un uno por un cero (o viceversa) generando un dato no válido (erróneo), un sistema de cómputo puede presentar errores en la manipulación de archivos, entre otras causas que originan datos faltantes o datos no válidos.
Por los motivos descritos se requiere que los datos sean preprocesados, de esta forma se garantizan aspectos como:

Aseguramiento de la calidad de la información:

los datos preprocesados deben venir sin errores o con la menor cantidad de errores posibles. De esta forma, se remueven valores atípicos, valores faltantes y tendencias incorrectas. Supongamos que una persona va a una entidad bancaria a pedir un crédito, el banco quiere emplear un modelo de machine learning para que, dadas ciertas características del cliente (comportamiento de pago, puntaje crediticio y otros aspectos) se pueda clasificar como confiable o no confiable. Si un modelo se ha entrenado mayoritariamente con datos faltantes en alguna columna, es muy posible que los optimizadores encuentren esta tendencia y la integren a sus rutinas en un proceso de machine learning. Eso induciría a modelos que entregan información errónea (por ejemplo, un alto riesgo solo por que la persona no llenó el campo “teléfono fijo”). Los datos que tienen un aseguramiento de la calidad permiten mejores procesos de conocimiento de los datos, así como mejores modelos.

Facilidad para el análisis:

otro aspecto importante de preprocesar los datos es que se hacen fáciles los procesos siguientes, por ejemplo, el análisis exploratorio es mucho más fácil de efectuar si ya se cuentan con datos preprocesados y sin errores.

Ventajas de cara al uso de machine learning:

Los datos preprocesados son esenciales para la creación de modelos de machine learning, es común que un conjunto de datos que no ha sido preprocesado tenga muchos registros con valores no válidos o con inconsistencias. Entrenar cualquier modelo con este tipo de datos puede generar errores y que simplemente el modelo no entrene (lo cual puede ser un buen caso, terminar en errores) o se pueden obtener modelos que están sesgados a una tendencia diferente de la real. Por ejemplo, modelos que son especialmente buenos en detectar una clase de las demás (con valores de precisión del 100%).

Los procesos más habituales en la limpieza de datos son:

Tratamiento de registros incompletos: se trata de dar manejo a aquellas filas que tengan datos faltantes, puede ser uno de ellos o varios, según sea el caso y el conjunto de datos se debe decidir qué hacer.

Análisis de outliers: Los outliers son datos atípicos, es decir que suelen estar 5 o más desviaciones estándar por encima de la media de los datos, cuando esto sucede hay que revisar las causas que pudieron dar origen al outlier. Por ejemplo, un error en la toma del dato, un error en su conversión o transmisión, un dato mal digitado, entre otros errores que pueden ser solucionados. También, dependiendo del conjunto de datos son valores atípicos por cambios. Por ejemplo, si se analiza el tráfico de las redes de datos, se deben excluir fechas como navidad o cuando ocurre un terremoto ya que la gente tiende a hacer muchas más llamadas de lo habitual, saturando la capacidad del sistema.

Tratamiento de datos incoherentes: existen casos en que hay discrepancias de la información. Por ejemplo, una edad negativa, una fecha en el futuro en un registro que ya ocurrió, una letra en un campo numérico, entre otros casos que se considerarían por fuera del rango de una variable aleatoria. Según sea el caso se debe decidir si se conserva el dato (y se completa) o si se elimina.

Balanceo de datos: en conjuntos de datos en donde un evento ocurre muy pocas veces, es fácil sesgar un modelo de machine learning a que aprenda el comportamiento de la clase mayoritaria y solo lo haga bien con esa clase. Por ejemplo, si tengo datos de transacciones bancarias y quiero descubrir un fraude, es probable que por cada 1000 transacciones tenga un fraude. Si se encuentra un modelo que diga las mil veces “no es fraude” se conseguiría una precisión alta. Pero el recall del modelo es bajo. La solución a este problema es re muestrear los datos para poder tener una mejor distribución de los eventos positivos y negativos.

¿Qué hacer con los valores faltantes?

Hay varios motivos por los que los datos pueden estar faltando, algunas decisiones típicas en estos escenarios son:

- Eliminar los registros que tengan datos faltantes.
- Sustituir los valores por un valor ficticio (una categoría diferente o un número como el cero).
- Sustitución por la media: permite rellenar campos numéricos con la media, lo que no afecta la distribución de probabilidad.
- Sustitución frecuente: reemplazar los datos faltantes con la moda.
- Sustitución por medio de regresión: se emplea algún modelo de regresión que permita reemplazar los valores según los otros datos asociados a un registro.

Como segunda etapa del preprocesamiento de los datos, es común encontrar tareas como la normalización de los datos. Es decir, escalar el conjunto de datos para que tenga una media de cero y una varianza de uno, esto reduce el tiempo durante el entrenamiento de los modelos de machine learning y también mitiga problemas como la explosión del gradiente.

Discretización de los datos:

se trata de llevar un valor o un intervalo de valores establecido (discreto).

Como tercera etapa se suelen incluir procesos que transforman los datos a una representación que facilita su análisis y la creación de modelos alrededor de estos datos. Dentro de esos procesos se tienen:

- Reducción de dimensionalidad.
- Extracción de características.
- Etapas de agregación de los datos.

En el caso de la reducción de dimensionalidad se tienen muchos algoritmos como PCA, Relieff, entre otros, que permiten conocer cuáles características (columnas) dentro de un conjunto de datos son las más importantes.

Suele ocurrir que la mayor entropía (cantidad de información para un punto dado) sea aportada por pocas características. Por ejemplo, si se analiza un conjunto de datos de 40 características (columnas) utilizando la medida de la entropía, suele ocurrir que los datos se pueden separar empleando 3 o 4 columnas, según sea el caso. Entrenar un modelo con 3 o 4 características que tengan el 90% de la información o más, significa una ganancia competitiva en contraste con equipos de ciencias de datos que no tengan un dataset con las dimensiones reducidas.

Extracción de características:

no todos los datos vienen bien estructurados, por eso se requieren de algoritmos que permitan describir mejor cada punto de datos empleando un algoritmo que extraiga números, que representan características relevantes y las almacene como puntos dentro de un array. Este tipo de procesos son comúnmente ejecutados con datos no estructurados (imágenes, videos y audios).

Agregación de los datos:

en el caso de la agregación, se trata de tomar datos que pueden ser diferentes o provenir de bases de datos distintas, pero que podrían pertenecer a un mismo grupo de datos y se necesita vincular esos datos para poder almacenarlos correctamente. Por ejemplo, complementar un arreglo de datos teniendo en cuenta dos o más fuentes de información.