El paradigma Big Data y el Aprendizaje Automático

JOSÉ A. ÁLVAREZ-JAREÑO Y JOSÉ M. PAVÍA

Actualmente, cada individuo es una fuente inagotable de datos. La mayoría de ellos los generamos de manera inconsciente, y por lo tanto, los cedemos a las empresas o instituciones de la misma forma. El uso de una tarjeta de crédito en un supermercado genera muchos datos para la empresa que emitió la tarjeta, para la entidad que nos la vendió y para la tienda en la que compramos. Sin embargo, se generan muchos más datos por el simple hecho de tener encendido el teléfono móvil. Los datos son la base de este nuevo paradigma, y como indican Mayer-Schönberger y Cukier (2013) son un recurso y una herramienta.

La cantidad de datos que se almacenan hoy día es inmensamente mayor que en cualquier momento anterior, y la tendencia es que siga creciendo. El volumen de información disponible crece a un ritmo acelerado, y como se puede leer en The Economist: "La cantidad de información digital se multiplica por 10 cada 5 años, mientras que la Ley de Moore indica que la capacidad de procesamiento se duplica cada 18 meses". Los datos se incrementan mucho más deprisa que la capacidad para procesarlos.

LOS DATOS

Los datos se pueden clasificar en función de sus características en tres tipos: estructurados, semi-estructurados y sin estructurar. Los datos estructurados son la base de la información de las empresas, forman su contabilidad y los informes con los que se toman las decisiones. Tendrán un formato o esquema prefijado con anterioridad. La fecha, el nombre, el DNI o los dígitos de las cuentas bancarias conforman este primer grupo.

Los datos semi-estructurados surgieron con el auge de las páginas webs. Estos datos no tienen una estructura, pero sí que tienen un flujo lógico, con etiquetas y marcadores que permiten identificarlos. Los más habituales son los registros de Web logs de las conexiones a Internet. Los Web logs son los registros de la actividad en Internet de un ordenador y pueden ser recopilados y analizados con herramientas o técnicas especiales.

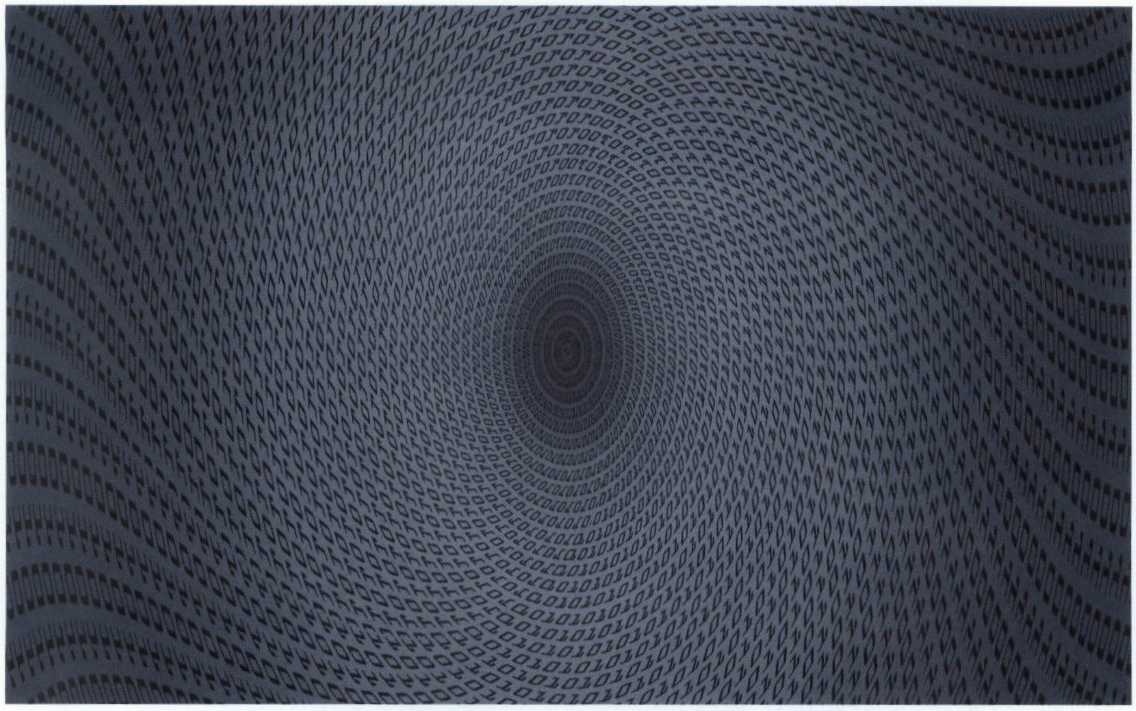
Los datos sin estructura podrían ser cualquier cosa, y se almacenarán como "documentos" u "objetos". La mayor parte de ellos serán ficheros de audio, video, fotografías, posts de Twitter, documentos impresos, correos electrónicos, mensajes, libros electrónicos, etc. Este tipo de datos puede proporcionar una importantísima cantidad de información, siempre y cuando, se disponga de las técnicas y el conocimiento adecuado para hacerlo.

Los datos también se pueden clasificar en función de su origen:

- La interacción entre humanos a través de un sistema digital que registra la actividad de la interacción. Los ejemplos de estos datos son los correos electrónicos, los foros de Internet o las redes sociales, en los que los datos son generados por personas y almacenados y procesados por ordenadores.
- La interacción entre un humano y una máquina. Al navegar o surfear por Internet las personas interactúan con ordenadores que registran la actividad que se produce en sus páginas web. De igual forma cuando se utiliza la banca on-line, el comercio electrónico, o los dispositivos GPS de navegación.
- La interacción entre maquinas (M2M). Este tipo de interacción es la que más ha crecido en los últimos años, ya que cada vez son más los sensores que se instalan para la recogida de datos que posteriormente se procesan o para el funcionamiento de multitud de servicios. Nuestro teléfono está compartiendo información con las antenas que prestan el servicio de telefonía móvil, o con los satélites del programa GPS que orbitan a 20.200 km. de altura.

BIG DATA

El término Big Data se utilizó por primera vez en 1997 por Michael Cox y David Ellsworth para referirse a las grandes cantidades de datos que son difíciles de gestionar mediante bases de datos relacionales. Pero no será



hasta 2001 cuando Doug Laney propondrá las tres Vs como principales características del Big Data: Volumen, Velocidad y Variedad. Con posterioridad otros autores han ido añadiendo nuevas características que también empiezan por V; y según al autor serán 5, 7 o hasta 9 Vs.

Volumen se refiere a la gran cantidad de datos que se precisa almacenar y procesar actualmente. No solo se pueden explotar los datos de los clientes de la base de datos interna de la compañía, ahora también se pueden utilizar todos los datos del tráfico de la página web para tener un mejor conocimiento de nuestros posibles clientes.

Dependiendo del tipo de datos y de negocio, los datos se deberán procesar rápidamente para actuar consecuentemente. La Velocidad de procesamiento de los datos será la segunda característica. No todas las acciones precisarán de una respuesta rápida, pero si se detecta el uso fraudulento de una tarjeta de crédito, la respuesta del sistema debe ser lo más rápida posible para minimizar los daños. Si las transacciones no se están incorporando en la base de datos, cuando se detecte el fraude puede ser muy tarde.

La tercera V hace referencia a la Variedad de los datos. La necesidad de integrar en una base datos procedentes de distintas fuentes, con diferentes estructuras y composición, es un nuevo reto. Las bases de datos relacionales están pensadas y organizadas para albergar datos estructurados, sin embargo, la mayor parte de los datos que se generan son desestructurados (archivos de audio, video, fotografías, documentos, libros, etc.).

Dos nuevas Vs son incluidas por casi todos los autores, Joyanes (2014): Veracidad y Valor. Uno de los principales objetivos del Big Data es servir de herramienta a la toma de decisiones de las empresas y profesionales, por lo que será fundamental que los datos que se utilicen en este procesos sean veraces. Las técnicas estadísticas sobre las que se basan los análisis de los datos están contrastadas y producirán resultados correctos siempre que los datos que se utilicen sean de calidad. Es el principio de GIGO (garbage in, garbage out), tal como expone Silver (2014) si entra basura sale basura, y las decisiones serán equivocadas.

Finalmente la V de Valor hace referencia a que los datos tienen que servir para algo. Los datos por sí solos no aportan nada. Una vez organizados y clasificados serán información, y posteriormente si son analizados podrán generar conocimiento. La obtención de este conocimiento tendrá un valor y este se deberá obtener de una manera rentable y eficiente.

El director de Investigación y Ciencia Computacional del CERN, Sergio Bertolucci, informaba en El Mundo que en sus instalaciones se genera un gran volumen de datos (30 petabytes de datos al año), pero que *«Ahora el juego está en ver cuál es el valor de estos datos»*.

APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)

Tal como afirman Caballero y Martín (2015) "El fin último del Big Data no es acumular datos, sino extraer información útil a partir de los datos". Para extraer información y convertirla en conocimiento se precisará de

diferentes técnicas estadísticas que se integrarán dentro de lo que se conoce como aprendizaje automático.

El término aprendizaje automático es genérico y englobará una serie de resultados de investigación, técnicas y herramientas con el objetivo de extraer información útil de los grandes volúmenes de datos. Habitualmente, las técnicas utilizadas suelen dividirse en métodos de Aprendizaje Supervisado y No Supervisado.

La diferencia de ambos métodos es que en el aprendizaje no supervisado todas las variables son de la misma naturaleza y no se precisa de una variable objetivo o dependiente. Sin embargo, en el aprendizaje supervisado existirá una variable respuesta que se modelizará en base a una serie de covariables, predictores o variables exógenas. La variable respuesta, en algunas ocasiones, habrá sido etiquetada en base al conocimiento de personas expertas en la materia.

Aunque cada procedimiento de extracción de información y conocimiento es diferente, y que no existe un modelo que sirva para todos los casos, sí que es conveniente llevar a cabo una serie de acciones para comprender los datos que se están analizando.

Objetivo del análisis

La primera cuestión fundamental es establecer lo más nítidamente posible qué es lo que se va a buscar en los datos. La mayoría de las veces los datos no tienen un valor evidente hasta que no se realizan las preguntas adecuadas. Realizar preguntas de carácter general sólo conduce a un callejón sin salida. Será necesario acotar el problema que se desea estudiar para iniciar una reflexión sobre los datos disponibles y su idoneidad para atacarlo. Tan importante como los datos serán las preguntas que nos planteamos sobre los mismos.

Una vez identificado el problema a tratar, a continuación se deberá elegir el método o técnica estadística apropiada para la resolución del mismo. Para dar una respuesta a un problema se pueden utilizar diferentes técnicas, algunas de ellas complementarias y otras sustitutivas.

La primera cuestión fundamental es establecer lo más nítidamente posible qué es lo que se va a buscar en los datos. La mayoría de las veces los datos no tienen un valor evidente hasta que no se realizan las preguntas adecuadas

Visualización de los datos a analizar

Es bien conocida la expresión que "más vale una imagen que mil palabras". Normalmente, es más fácil mostrar una imagen que intentar hacer una extensa explicación. Algo parecido ocurre con los datos, "más vale un gráfico que todos los datos". Los humanos no son muy buenos procesando datos, sin embargo, son muy buenos realizando otras tareas visuales, y es capaz de que un ojo entrenado pueda ver y deducir patrones en una representación gráfica de los datos que tanto le cuesta comprender.

Este paso no es obligatorio efectuarlo para poder realizar un correcto análisis de los datos, pero sí que puede ayudar al científico a tener una visión global de los mismos. Esta primera representación gráfica puede aportar ideas sobre la estructura de los datos, la distribución, e incluso sobre la solución del problema. Con la llegada de los potentes procesadores que pueden realizar representaciones gráficas de grandes conjuntos de datos, se ha desarrollado toda una rama de la analítica que se conoce como "visual analytics".

Comprobación de los supuestos del modelo

Todas las técnicas y/o modelos estadísticos tendrán unos supuestos para su aplicación, y la siguiente cuestión será comprobar que los datos que se van a analizar cumplen con esas premisas. Si la técnica exige que los datos provengan de una población normal, o si una serie temporal debe ser estacionaria, se deberá comprobar que efectivamente se cumplen con los supuestos de partida. Ignorar esta fase, puede llevar al científico a realizar un correcto análisis de los datos con una técnica adecuada, y sin embargo, los resultados no ser reflejo de la realidad. Se sabe que una metodología estadística funciona bien bajo determinados principios, incumplirlos implicará modificar el funcionamiento y los resultados podrán ser irrelevantes.

A la hora de realizar una encuesta es fundamental para que los resultados sean válidos que la selección de los encuestados sea adecuada y se incorpore en el proceso de inferencia toda la información relevante. Es mucho más importante la forma en la que se escogen a los encuestados que el número de encuestas realizadas. De nada sirve tener un gran numero de encuestas si los encuestados son tratados como una muestra representativa de la población que se pretende analizar y en realidad no lo son. Las técnicas estadísticas utilizadas para los sondeos electorales son las apropiadas siempre y cuando los datos recogidos cumplan con unos criterios mínimos que no se pueden obviar y se emplean las técnicas apropiadas para compensar por las debilidades que presenten los datos.

Selección de los datos

El siguiente paso sería seleccionar los datos con los que se va a realizar el análisis. En función del objetivo se utilizará diferentes técnicas de selección y tratamiento de los datos. De acuerdo con Siegel (2014) tres son los niveles de análisis de datos:

- Análisis descriptivo: el objetivo es saber qué ha pasado o qué está pasando en una determinada situación. Únicamente se pretende describir que ha ocurrido en una empresa o en un país durante un período de tiempo. Permitirá realizar afirmaciones como: "el consumo eléctrico en el último trimestre creció un 10% respecto al mismo periodo del año anterior". Hace una descripción de la situación sin cuestionar que ocurrirá en el futuro.
- Análisis predictivo: al realizar este análisis el propósito final es poder predecir qué ocurrirá en el futuro. Ahora se esta interesado en saber cúal será el consumo eléctrico del próximo trimestre para poder tomar decisiones sobre la capacidad de suministro de las empresas eléctricas. ¿Se podrá abastecer todo el consumo? ¿Será necesario importar electricidad de otros países?
- Análisis prescriptivo: es el más complejo de todos, ya que su objetivo es identificar qué se debe hacer para obtener un resultado. Las preguntas que busca responder este tipo de análisis es ¿qué hacer para que mis clientes no se marchen a la competencia? Para poder llevar a cabo este análisis se precisa de los dos anteriores, ya que será necesario conocer la situación (histórica) y haber realizado predicciones que permitan comprobar cómo han reaccionado los sujetos a determinadas acciones.

Un ejemplo que se utiliza habitualmente para ilustrar estos tres niveles del análisis de datos:

- Ayer llovió en mi ciudad. Análisis descriptivo, hecho cierto y conocido que describe el tiempo atmosférico en un día concreto en un lugar exacto.
- Es probable que hoy vuelva a llover. Análisis predictivo, de la situación del tiempo en los pasados días se infiere que hoy hay mucha probabilidad que vuelva a llover.
- Sería conveniente salir de casa con paraguas. Si el individuo no pretende volver a casa como una sopa, debería llevarse un paraguas o al menos un chubasquero. Si la probabilidad de lluvia en la ciudad es muy elevada, lo más lógico sería tomar medidas.

Según un estudio realizado por Gartner en 2013, sólo el 3% de las compañías analizadas emplean soluciones de analítica prescriptiva, limitándose el 84% de las mismas a realizar análisis descriptivos. Para un análisis descriptivo únicamente se precisará un conjunto de datos para realizar el análisis.

Si se pretende dar un segundo paso y efectuar una predicción para el futuro, los datos deberán ser también representativos de la población sobre la que se desean realizar inferencias, pero, además, se deberán particionar en dos subconjuntos. El primero de los conjuntos será el conjunto de entrenamiento con el que se modelizarán los datos para, posteriormente, realizar la predicción, y el segundo será el conjunto de comprobación con el que se determinará la capacidad predictiva del modelo propuesto.

La forma en la que se realizará está división de los datos deberá ser también aleatoria o respetar la secuencia temporal de acuerdo con la cual han sido obtenidos. Los dos conjuntos que salgan de la muestra inicial de datos deben disponer de la misma estructura, y para ello el muestreo aleatorio será el más indicado en el primer caso.

Las aseguradoras y los actuarios están acostumbrados a trabajar con datos. Es la base de su negocio. Para sacar partido a la nueva situación sólo necesitan incrementar su arsenal de técnicas y ampliar su visión de lo que significa un dato para continuar extrayéndoles valor.

Si se dispone de una muestra grande, el científico optará por el tamaño de ambos conjuntos. Las divisiones más habituales son 80% para entrenamiento y 20% para comprobación, o bien, dos tercios y un tercio. El conjunto de entrenamiento suele ser mayor que el de comprobación porque el objetivo es extraer la mayor información posible de los datos de entrenamiento. Aunque obviamente es posible encontrar excepciones.

Para realizar la analítica prescriptiva, se dispondrá de diferentes técnicas estadísticas que se podrán aplicar en función de las necesidades de la empresa. Será necesario la realización de experimentos para poder inferir que opción es mejor. Las pruebas A/B o los modelos de respuesta incremental (más conocidos por su expresión en inglés, uplift models) serán algunas de las soluciones más habituales. Para poder realizarlas será necesario establecer a priori un grupo de control. De esta forma se podrá observar que efecto causa la medida aplicada en comparación con el grupo de control, tal como indica Siegel (2014).

Aplicación de la técnica estadística

El siguiente paso consistirá en aplicar con ayuda de la herramienta informática adecuada la técnica o técnicas estadísticas que permitan resolver el problema. Si se ha optado por una técnica de regresión se deberán comprobar los diferentes parámetros del modelo para obtener los resultados acordes a los datos que se están analizando. La técnica a aplicar no tiene porqué ser única, y se pueden conformar modelos mediante procedimientos de ensamble. La utilización conjunta de varias técnicas estadísticas generaría una mejor capacidad predictiva que cada una de ellas por separado. En palabras de Silver (2014) "la predicción conjunta supera a menudo incluso la mejor predicción individual".

Análisis de resultados

El último paso será interpretar los resultados y obtener las conclusiones. La experiencia del científico será determinante para realizar una interpretación acertada de los resultados obtenidos. El mismo resultado puede ser interpretado de diferentes maneras por diferentes personas. No olvidar, que los resultados electorales son los mismos para todos los partidos, sin embargo, cuando hacen las valoraciones todos parecen haber ganado las elecciones.

Ser objetivo en la interpretación de los resultados puede ser determinante en la solución que se aplique. Con un 50% de probabilidad de lluvia, se puede optar por coger el paraguas o no hacerlo, pero con un 80% parece arriesgado no cogerlo. La experiencia del analista será muy importante en la gestión que se haga de los resultados.

De igual forma, la capacidad de comunicar los resultados a los directivos de la compañía puede ser lo más importante de todo el proceso. Si el científico no es quien debe tomar las decisiones finales y los resultados se deben trasladar a un equipo directivo, las ideas y como se expresen puede cambiar la toma de decisiones de la dirección.

CONCLUSIÓN

Los datos están creciendo a gran velocidad. Las oportunidades están disponibles para todas las empresas de todos los sectores. Aquellas empresas que tomen la iniciativa comenzarán a ganar el futuro. El sector asegurador, además, es un candidato ideal para adoptar estas nuevas tecnologías. Las aseguradoras y los actuarios están acostumbrados a trabajar con datos. Es la base de su negocio. Para sacar partido a la nueva situación sólo necesitan incrementar su arsenal de técnicas y ampliar su visión de lo que significa un dato para continuar extrayéndoles valor. El futuro está aquí y no te va a esperar. Toma la iniciativa.

REFERENCIAS

- Caballero, R. y E. Martín (2015) **Las Bases del Big Data**. Catarata, Madrid.
- Joyanes, L. (2014) **Big Data. Análisis de Grandes Volúmenes de Datos en Organizaciones**. Marcombo Ediciones Técnicas, Barcelona.
- Mayer-Schönberger, V. y K. Cukier (2013) **Big Data. La revolución de los datos masivos**. Turner Publicaciones, Madrid.
- Siegel, E (2014) **Analítica Predictiva**. Ediciones Anaya, Madrid.
- Silver, N. (2014) **La Señal y el Ruido**. Ediciones Península, Barcelona.