¿CUÁNDO USAR MACHINE LEARNING?

Introducción.

Imaginen que trabajan para una empresa donde les interesa analizar el comportamiento histórico de las ventas, y para eso cuentan con un dataset que contiene el monto total facturado cada día, para un año en particular.

La idea es lograr predecir el nivel de ventas para un mes determinado y para años posteriores. Pero, ¿se podría usar el Machine Learning en este caso? y ¿cómo sabemos si es realmente la alternativa más adecuada?

Pues este sencillo ejemplo es un caso típico de lo que ocurre cuando nos enfrentamos a una situación en donde debemos usar los datos para resolver un problema en particular.

Y como el Machine Learning es un área que en los últimos años ha alcanzado unos logros impresionantes, a veces la tendencia es usarlo para resolver todo tipo de problemas en donde estén involucrados datos.

Pero esto no quiere decir que lo podamos usar prácticamente en cualquier situación. De hecho, si lo usamos cuando realmente no es necesario, la empresa para la que trabajamos puede perder tiempo y dinero, y probablemente tendremos un modelo que no tiene un buen desempeño o que no resulta útil.

Es decir, hay casos en donde el Machine Learning simplemente NO es necesario… ¿Pero cómo saber cuándo usarlo y cuándo no?

Para esto debemos partir de esta definición muy importante: "***El Machine Learning es un enfoque que permite aprender patrones complejos a partir de datos existentes, y usar estos patrones para realizar predicciones sobre datos nunca antes analizados***"

Trata de recordar los términos **aprender patrones complejos** y **realizar predicciones**, porque son esenciales para lo que vamos a hablar ahora.

Lo esencial antes de hacer cualquier cosa: los datos.

La primera condición que debemos verificar está relacionada con los datos, que son el punto de partida de cualquier proyecto de Machine Learning. En este caso debemos garantizar que se cumplen al menos dos condiciones: que son **accesibles** y que son **asequibles**.

Que sean accesibles quiere decir que los podemos recolectar fácilmente. Por ejemplo, podemos intentar crear un modelo para predecir el volumen de ventas de una empresa, pero es imposible construirlo a menos que tengamos acceso a datos de varios años atrás.

Y que los datos sean asequibles significa que en muchos casos no serán de uso libre y probablemente será necesario comprarlos. Por ejemplo, imaginemos que queremos desarrollar un sistema para prevenir fraudes con tarjetas de crédito. Para poderlo construir necesitamos un dataset con datos reales, que muy probablemente pertenecerán a una entidad bancaria. Lo más probable es que este dataset no sea de acceso libre, y si queremos utilizar esos datos puede resultar necesario adquirirlos.

Así que en resumen, el paso cero es tener certeza de que contamos con los datos que necesitamos.

El aspecto ético: un dilema que se debe resolver desde el comienzo.

Bien, si los datos son accesibles y asequibles, debemos analizar ahora otra situación que de hecho es menos de tipo técnico y más de tipo humano: el tema ético.

Y este es un aspecto súper importante porque, independientemente del problema que queramos resolver, debemos asegurarnos de que a los datos y al modelo se les dará un uso ético.

Imaginemos por ejemplo el caso de los vehículos autónomos: ¿qué pasa si uno de estos vehículos atropella a una persona? ¿A quién podemos culpar de este accidente? Es un dilema ético de fondo, que realmente ha generado una discusión en torno a este tema y ha creado una barrera que ha dificultado el uso masificado de estos vehículos.

Y dilemas como este se vienen dando en diferentes ámbitos de nuestra vida diaria, desde el mismo uso de las redes sociales, pasando por la medicina y llegando incluso a temas militares.

Así que no podemos confiar ciegamente ni en los datos ni en los algoritmos ni en el Machine Learning. Así como resulta benéfico en muchas situaciones, también existen casos, como el de los vehículos autónomos, en donde su uso puede llevar a riesgos con consecuencias catastróficas.

Una sugerencia: ante cualquier dilema ético que pueda tener consecuencias negativas sobre cualquier actor de nuestro entorno, es mejor no usar ni el Machine Learning, y de hecho ningún otro enfoque.

¿Determinístico o estocástico?

Resuelto el tema ético, el siguiente paso es tener claridad sobre el proceso que está detrás de la generación de los datos, que puede ser **determinístico** o **estocástico**.

Un ***proceso determinístico*** quiere decir que puedo predecir un evento futuro con una certeza del 100%, es decir es un proceso donde no habrá ningún tipo de aleatoriedad. Por ejemplo, para calcular el saldo de mi cuenta de ahorros a final de mes simplemente debo tomar los abonos hechos y sumarle lo que ha ganado por cuenta de la tasa de interés.

Por otro lado, en un ***proceso estocástico*** no tendremos una certeza del 100%, y en su lugar tendremos una estimación o una probabilidad de que el evento ocurra. Es decir que en estos procesos tendremos un cierto grado de **aleatoriedad**. En el caso anterior, un ejemplo de un proceso estocástico sería el comportamiento de la tasa de interés, pues esta depende de las condiciones del mercado, de dónde invierta el dinero el banco, de la volatilidad de las acciones, etc., y por tanto tendrá ese grado de aleatoriedad.

Pero un proceso estocástico es diferente de un proceso aleatorio, como lanzar un dado. En este último caso es imposible predecir, en cada lanzamiento, en qué número caerá, mientras que la tasa de interés tiene un mayor grado de predictibilidad. Por ejemplo, con un alto grado de confianza puedo decir que prácticamente nunca tendremos una tasa de interés del 1.000%, aunque sería fantástico!

Así que si el proceso es determinístico no hace falta usar el Machine Learning, podemos simplemente recurrir al modelo o las ecuaciones que tengamos para resolver el problema. Pero el Machine Learning resulta viable si el proceso es estocástico.

¿Análisis histórico o predicciones?

Si los datos se generan a través de un proceso estocástico, el siguiente paso es definir lo que queremos hacer con esos datos: es decir, ¿se trata de un análisis de **comportamiento histórico** o estamos intentando hacer una **predicción**?

Volvamos al ejemplo inicial, el del volumen de ventas de la empresa. Si simplemente queremos saber en qué mes o día se tuvieron más ventas, basta con consultar el registro histórico que tenemos. Acá realmente no necesitamos Machine Learning.

Pero si, por el contrario, lo que queremos es tomar estos datos históricos y predecir lo que podría ocurrir en años posteriores, en este caso el Machine Learning podría funcionar.

Decimos “podría”, porque realizar predicciones no es la única condición (recordemos que podemos predecir el saldo de nuestra cuenta usando una simple ecuación).

Y acá vale la pena retomar la definición que les mencionaba al comienzo: “El Machine Learning es un enfoque que permite aprender patrones complejos a partir de datos existentes, y usar estos patrones para realizar predicciones sobre datos nunca antes analizados”.

Entonces es evidente que una de las tareas del Machine Learning es **realizar predicciones**, en tareas como la clasificación o la regresión.

Pero además debe haber algún tipo de **patrón** en los datos, y la idea es que con el Machine Learning el modelo pueda aprender a identificar ese patrón. Por ejemplo, aprender a predecir el número ganador de la lotería resulta imposible, porque es un proceso totalmente aleatorio, no hay patrones, pero aprender a predecir con cierta precisión la variación de la tasa de interés sí sería posible con el Machine Learning.

Pero además estos patrones deben ser **complejos**. Predecir cuándo habrá un eclipse de Sol requiere recurrir a la física para predecir las trayectorias de la Tierra y el Sol. Pero predecir el valor de un inmueble a partir de su área, del número de habitaciones, del año en que fue construido, de si hay o no escuelas o supermercados en el vecindario es posible con el Machine Learning, porque allí tenemos un patrón más complejo.

La Interpretabilidad.

Bien, estando seguros de que queremos hacer una predicción, la siguiente condición a analizar sería la **interpretabilidad**, que es uno de los grandes signos de interrogación del Machine Learning.

La interpretabilidad es ***el grado con el que un humano puede comprender las razones de una decisión tomada por el modelo de Machine Learning.***

Por ejemplo, si trabajamos para una entidad bancaria y estamos desarrollando un sistema de detección de fraudes, allí resulta fundamental poder explicar los motivos por los cuales una transacción fue clasificada como fraudulenta; de esta forma la entidad puede tomar los correctivos necesarios para evitar este tipo de fraudes a futuro.

Desafortunadamente la mayor parte de los modelos de Machine Learning no son fácilmente interpretables, exceptuando posiblemente los árboles de regresión, los árboles de clasificación y los bosques aleatorios en algunos casos particulares.

Estos modelos generalmente funcionan como una caja negra, que logra altos niveles de precisión pero donde lo que ocurre en el interior de esa caja tiene todavía un cierto halo de misterio para nosotros los humanos.

Así que, en resumen, si requerimos interpretabilidad realmente la mayoría de las veces el Machine Learning NO es la vía recomendada.

De nuevo los datos: ¿Tenemos suficientes?

Pero si la interpretabilidad no es un inconveniente, el Machine Learning podría ser una alternativa correcta, así que poco a poco vamos llegando al punto que nos interesa: es decir, ¡saber en qué casos se puede usar el Machine Learning!

Para llegar a este punto debemos determinar si la cantidad de datos que hemos recolectado es suficiente, y esto es muy importante porque ***para lograr que el modelo de Machine Learning aprenda a identificar patrones se requiere generalmente una gran cantidad de datos***.

Aunque el número exacto depende del problema en particular que estemos resolviendo, podemos usar tres reglas básicas para tener una primera estimación de la cantidad de datos que se requeriría:

1. Si es un problema de clasificación podemos definir un factor dependiendo del número de clases. Por ejemplo, podemos definir que por cada categoría se requieren al menos 100, 1000 o 10.000 ejemplos de entrenamiento
2. También se puede definir un factor dependiendo del número de características de cada dato. Es decir, si cada ejemplo de entrenamiento está representado como un vector de, por ejemplo, 9 características, entonces el número de ejemplos puede ser de al menos 100, 1.000, 10.000 veces ese número
3. Por último se puede definir un factor dependiendo del número de parámetros del modelo. Así, si por ejemplo estoy entrenando una red neuronal con 100.000 parámetros, podríamos definir una cantidad de ejemplos de entrenamiento que puede ser 5 o 50 veces ese número.

Pero esto siempre será una aproximación, y es necesario entrenar el modelo para verificar si se requieren más datos o no.

Desafortunadamente ***es imposible calcular de forma precisa y con antelación un valor exacto para el tamaño del dataset, pues en el Machine Learning este tamaño depende de muchas variables que son propias del problema a resolver***, como el tipo de datos o el tipo de modelo a implementar. Por ejemplo, una Máquina de Soporte Vectorial es un modelo menos complejo que una Red Neuronal, una Red Convolucional o una Red Transformer, y por tanto la primera requerirá muy probablemente menos ejemplos de entrenamiento.

ADAPTACION del blog: ¿Cuándo usar el Machine Learning?, mayo2021, Miguel Sotaquirá, CODIFICANDOBITS.