aepd agencia española protección datos

EDPS

10

MALENTENDIDOS SOBRE EL

MACHINE LEARNING

(APRENDIZAJE AUTOMÁTICO)

La UE ha identificado la inteligencia artificial (IA) como una de las tecnologías más relevantes del siglo XXI, y ha destacado su importancia en la estrategia para la transformación digital de la UE[1]. Al tener una amplia gama de aplicaciones, la IA puede contribuir en áreas tan dispares como el tratamiento de enfermedades crónicas, la lucha contra el cambio climático o la anticipación de amenazas de ciberseguridad.

"Inteligencia artificial", sin embargo, es un término general que engloba a aquellas tecnologías que tienen como objetivo imitar las capacidades de razonamiento humano, que pueden tener aplicaciones y limitaciones muy diferentes. Con frecuencia, los proveedores de tecnología promocionan sus sistemas haciendo referencia a la IA y, sin especificar qué tipo de IA.

El aprendizaje automático (Machine Learning o ML) es una rama específica de la IA, aplicada a la resolución de problemas específicos y limitados, como tareas de clasificación o predicción. A diferencia de otros tipos de IA que intentan emular la experiencia humana (por ejemplo, sistemas expertos[2]); el comportamiento de los sistemas de aprendizaje automático no está definido por un conjunto predeterminado de instrucciones.

Los modelos de ML se entrenan utilizando conjuntos de datos. Durante su entrenamiento, los sistemas de ML se adaptan de forma autónoma a los patrones encontrados en las diferentes variables de un conjunto de datos dado, creando correlaciones. Una vez entrenado, el sistema utilizará los patrones aprendidos para generar un resultado.

A diferencia de otros tipos de sistemas de IA[3], el rendimiento[4] de los modelos de ML depende en gran medida de la precisión y la representatividad de los datos utilizados para su entrenamiento (datos de entrenamiento).

El objetivo de este documento es dilucidar los conceptos erróneos comunes que rodean a los sistemas de ML, al tiempo que subrayar la importancia de implementar estas tecnologías de acuerdo con los valores de la UE, los principios de protección de datos y el total respeto a las personas.

---

1 Comunicación de la Comisión: Inteligencia artificial para Europa, **https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=COM%3A2018%3A237%3AFIN**

2 Los sistemas expertos son programas informáticos diseñados para resolver problemas complejos en áreas específicas. Se basan en una base de conocimiento, que define las reglas para la toma de decisiones, y en un motor de inferencia, que aplica las reglas.

3 En aprendizaje automático, el algoritmo aprende reglas a medida que establece correlaciones entre entradas y salidas. En el razonamiento simbólico, las reglas se crean a través de la intervención humana. Primero, los humanos deben aprender las reglas por las cuales dos fenómenos se relacionan, y luego traducir a código esas relaciones en el sistema de razonamiento simbólico. Por lo tanto, la precisión del sistema simbólico de IA se basa en la calidad de la definición humana de esas relaciones, en lugar de la calidad de los conjuntos de datos de entrada.

4 En pocas palabras, el rendimiento de un sistema de ML define cómo de "buenas" son realmente sus predicciones. Aunque sea un concepto simple, la complejidad tiene que ver con identificar lo que se considera "bueno". Existen varias "métricas de rendimiento", que permiten evaluar modelos de ML: La exactitud es la fracción de predicciones acertadas de un modelo; La precisión es la relación entre el número de resultados correctos y el número de todos los resultados devueltos; La exhaustividad es la relación entre el número de resultados correctos y el número de resultados que deberían haberse devuelto como correctos. Dependiendo del contexto de la aplicación, algunas métricas de rendimiento podrían ser más relevantes que otras.

1 MALENTENDIDO

Correlación implica causalidad.

Realidad: Causalidad es más que solo establecer correlaciones.

La "causalidad" es la relación que existe entre causa y efecto. La "correlación" es la relación que existe entre dos factores que ocurren o evolucionan con alguna sincronización.

Los sistemas de ML son muy eficientes en la búsqueda de correlaciones, pero carecen de la capacidad analítica para ir más allá de eso y establecer una relación causal [5].

Sea un ejemplo, el sistema de ML es alimentado con un conjunto de datos compuesto por las puntuaciones en test de inteligencia y la altura de los participantes (pero no la edad), un modelo de ML podría predecir erróneamente que las personas altas son más inteligentes que las personas bajas, al establecer una correlación entre altura y puntuación. Sin embargo, tal fenómeno podría estar originado por el hecho de que los niños comúnmente obtienen puntuaciones menores que los adultos en las pruebas de coeficiente intelectual.

Como otro ejemplo, también podríamos mencionar que entrenar un sistema de ML para inferir enfermedades es posible utilizando un conjunto de datos con correlaciones síntoma-enfermedad. Sin embargo, ese mismo sistema podría no ser adecuado para explicar qué está causando la enfermedad.

Estos ejemplos reales demuestran que la supervisión humana es necesaria para garantizar que los sistemas de ML identifiquen las variables relevantes (las causas últimas) para una predicción o clasificación.

5 **https://www.wired.com/story/ai-pioneer-algorithms-understand-why**

2 MALENTENDIDO

Al desarrollar sistemas de aprendizaje automático, cuanto más datos y mayor sea la variedad, mejor.

Realidad: Los conjuntos de datos de entrenamiento de ML deben seleccionarse para cumplir umbrales de precisión y representatividad.

El creciente desarrollo de los sistemas de ML ha llevado a una mayor demanda de datos personales y no personales, porque los desarrolladores de ML no tienen suficientes datos para mejorar el rendimiento de sus sistemas.

Por lo general, el entrenamiento de los sistemas de ML requiere grandes cantidades de datos, dependiendo de la complejidad de la tarea a resolver. Sin embargo, utilizar más datos de entrenamiento en el desarrollo de modelos de aprendizaje automático no siempre mejorará el rendimiento del sistema. De hecho, podría crear nuevos problemas o empeorar los existentes.

Por ejemplo, incluir más imágenes masculinas de piel clara a los conjuntos de datos de entrenamiento de reconocimiento facial no ayudará a corregir ningún sesgo de género o étnico existente de los sistemas[6].

El RGPD exige que el tratamiento de los datos personales sea proporcional a su finalidad. Desde el punto de vista de la protección de datos, no es una práctica proporcional aumentar sustancialmente la cantidad de datos personales en el conjunto de datos de entrenamiento para tener únicamente una ligera mejora en el rendimiento de los sistemas.

Más datos no necesariamente mejorarán el rendimiento de los modelos de ML. Por el contrario, más datos podrían implicar una generación de más sesgo en el modelo.

6 Análisis de sesgo de género y étnico en: Gender Shades project **http://gendershades.org**

3 MALENTENDIDO

ML necesita datos de entrenamiento completamente libres de errores.

Realidad: Para conseguir un buen rendimiento los modelos de ML se necesitan datos de entrenamiento con tan solo una calidad superior a un cierto umbral.

El rendimiento de los modelos de ML depende, entre otros factores, de la calidad de los datos de entrenamiento, validación y prueba. Por lo tanto, esos conjuntos de datos deben ser capaces de definir un caso real de forma suficientemente completa y precisa. Sin embargo, la ciencia estadística sugiere que, a pesar de la presencia de errores individuales en los datos [7] de entrada, es posible calcular con precisión el resultado promedio cuando se procesan grandes cantidades de datos.

Los modelos de ML son tolerantes a inexactitudes ocasionales en registros individuales [8], pues descansan en la calidad del conjunto de datos de entrenamiento considerando éste como un todo. Algunos modelos de ML se entrenan utilizando datos sintéticos, es decir, conjuntos de datos de entrenamiento generados artificialmente, que imitan datos reales. Incluso si ningún dato real coincide con precisión con los datos sintéticos, los modelos de ML entrenados con datos sintéticos pueden producir buenos rendimientos.[9]

La Privacidad Diferencial es una técnica que introduce ruido en los conjuntos de datos de entrenamiento para preservar la privacidad de los interesados. A pesar de las imprecisiones introducidas por la privacidad diferencial, los modelos de ML son capaces de lograr buenos rendimientos[10].

7 "When you have plenty of data the law of big numbers tends to make sure the data is evenly distributed." (p. 43) Practical Machine Learning with H20 by Darren Cool tried to k, O'Really Media Inc. "Third, data anomalies were eliminated in the data cleansing process, due to the so-called law of big numbers." **https://link.springer.com/article/10.1186/s40537-019-0216-1**

8 De hecho, debido al gran volumen de variables de entrada en algunos modelos de ML, a menudo es necesario utilizar técnicas que introducen ruido en los datos de entrada, como Análisis de Componentes Principales (PCA), una técnica para agregar variables. Por supuesto, el ruido introducido en los datos de entrada debe estar por debajo del valor de rendimiento aceptable de la aplicación.

9 Algunos ejemplos de modelos de aprendizaje automático entrenados con datos sintéticos: Amazon Alexa **https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases** Google Waymo **https://blog.waymo.com/2019/08/learning-to-drive-beyond-pure-imitation_26.html**

10 Recent research has shown, counterintuitively, that differential privacy can improve generalization in machine learning algorithms - in other words, differential privacy can make the algorithm work better!' **https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy**

11 En el aprendizaje automático, los parámetros son los valores que un algoritmo de aprendizaje puede cambiar de forma independiente a medida que aprende. Estos valores se optimizan a medida que el modelo aprende, perfeccionando así su razonamiento.

4 MALENTENDIDO

El desarrollo de sistemas de ML requiere grandes repositorios de datos o el intercambio de conjuntos de datos de diferentes fuentes.

Realidad: El aprendizaje federado permite el desarrollo de sistemas de aprendizaje automático sin compartir datos de entrenamiento.

La agrupación de datos y el sistema de aprendizaje automático en una infraestructura de computación en la nube controlada por desarrollador del sistema de ML es una solución común para evitar restricciones de rendimiento. Esta es una arquitectura conocida como aprendizaje centralizado. Sin embargo, aunque el aprendizaje centralizado puede resolver problemas de rendimiento, hay algunas consideraciones que se deben tener en cuenta. Una de ellas es que, los datos personales requieren que tanto el responsable como el destinatario de los datos cumplan con los principios RGPD de responsabilidad proactiva, seguridad y limitación de propósito, entre otros. Otra es que los repositorios más grandes de datos personales aumentan el interés de terceros para obtener acceso no autorizado y exacerban el impacto de una brecha de datos personales. El aprendizaje distribuido on-site y el aprendizaje federado son arquitecturas de desarrollo alternativas al aprendizaje automático centralizado. En el aprendizaje distribuido on-site, cada servidor del responsable del tratamiento descarga un modelo de ML genérico o previamente entrenado desde un servidor remoto. Cada servidor local utiliza su propio conjunto de datos para entrenar y mejorar el rendimiento del modelo genérico.

Después de que el servidor remoto haya distribuido el modelo inicial a los dispositivos, no es necesaria ninguna otra comunicación. Implica las mismas técnicas utilizadas en el aprendizaje centralizado, pero en los servidores del responsable. En el aprendizaje federado, cada servidor del responsable entrena un modelo con sus propios datos y envía únicamente los parámetros del modelo a un servidor central[11] para su agregación. Los datos permanecen en los dispositivos y el conocimiento se comparte a través de un modelo agregado con pares.[12] Ninguna arquitectura de aprendizaje se adapta a todas las tareas. Sin embargo, acumular datos en uno o varios servidores no siempre es la mejor ni la más eficiente solución. Además, podría constituir un obstáculo para que las PYMES realicen desarrollos basados en ML.

12 Abdulrahman, Sawsan & Tout, Hanine & Ould-Slimane, Hakima & Mourad, Azzam & Talhi, Chamseddine & Guizani, Mohsen. (2020). A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond. IEEE Internet of Things Journal. PP. 10.1109/JIOT.2020.3030072. **http://dx.doi.org/10.1109/JIOT.2020.3030072**

5 MALENTENDIDO

Los modelos de ML mejoran automáticamente con el tiempo.

Realidad: Una vez implementados, el rendimiento de los modelos de ML puede deteriorarse y no mejorará a menos que reciba entrenamiento adicional.

Durante el entrenamiento de un modelo de ML, el algoritmo se prueba constantemente. Cuando el modelo está maduro (es decir, puede resolver correctamente los problemas para los que fue diseñado), se considera adecuado para ser puesto en producción.

Un modelo que se implementa y ya no se entrena no "aprenderá" más correlaciones de los datos entrantes, sin importar la cantidad de datos que se le proporcionen. Esto significa que, a menos que los modelos de ML continúen siendo entrenados, no se puede esperar que evolucionen. Esto es un riesgo para la precisión del sistema, ya que su obsolescencia hacia la realidad puede poner en peligro su capacidad para hacer juicios ajustados y justos.

La capacidad predictiva de los modelos de ML, una vez en producción, puede deteriorarse con el tiempo por dos motivos diferentes: debido a la deriva de los datos que se emplean como entrada al modelo (cambios sustanciales en los datos de entrada) o debido a la deriva del concepto (cuando nuestra interpretación de los datos cambia mientras que la distribución general de los datos no cambia). [13]

Ya que el contexto del tratamiento en el que un sistema ML está actuando puede evolucionar, es necesario monitorizar el sistema para detectar cualquier deterioro del modelo y actuar sobre el mismo (por ejemplo, volviendo a entrenar el modelo con nuevos datos).

13 A Comprehensive Guide on How to Monitor Your Models in Production **https://neptune.ai/blog/how-to-monitor-your-models-in-production-guide**

6 MALENTENDIDO

Las decisiones automáticas tomadas por los algoritmos de ML no pueden ser explicadas.

Realidad: Un modelo de ML bien diseñado puede producir decisiones comprensibles para todas las partes interesadas relevantes.

Varios enfoques son posibles para proporcionar explicaciones de las decisiones basadas en IA, y la mayoría de ellos también se pueden aplicar a las decisiones del modelo de ML.

Algunos enfoques aclaran el proceso de creación del modelo, especificando qué parámetros [14] e hiperparámetros se han aplicado y cuánta influencia tuvo cada uno en el modelo resultante. Otros explican cómo el modelo interpreta las características [15] de los datos entrantes, lo que permite a los individuos comprender y anticipar cómo se comportará el sistema en una situación particular. Otros enfoques no explican el comportamiento general del modelo, sino que se centran en cómo una entrada concreta influye en conseguir una salida determinada [16].

Pueden ser necesarios diferentes grados de detalle en la explicación del modelo, dependiendo de los individuos y el contexto. El enfoque adecuado será el que pueda describir de forma clara a los interesados la toma de decisiones en los procesos de entrenamiento y creación del modelo de ML.

14 Un hiperparámetro es un parámetro cuyo valor se establece antes de que comience el proceso de aprendizaje automático. Por el contrario, los valores de otros parámetros se derivan a través del entrenamiento.

15 Por ejemplo, el valor de presión sanguínea de un paciente podría ser muy relevante para detectar una determinada enfermedad, mientras que quizás la edad del paciente no sea tan relevante.

16 Arya, V. et al. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques". ArXiv abs/1909.03012 'Tr(2019): **https://arxiv.org/abs/1909.03012v2**

7 MALENTENDIDO

La transparencia en ML viola la propiedad intelectual y no es entendida por el usuario.

Realidad: Es posible proporcionar una transparencia significativa a los usuarios de IA sin dañar la propiedad intelectual.

Las personas deben recibir suficiente información sobre cómo se tratan sus datos personales, y los tratamientos basados en sistemas de ML no son una excepción.

Este tipo de transparencia no implica necesariamente la divulgación de información técnica detallada que, en la mayoría de los casos, no sería significativa para los usuarios.

De la misma manera que un prospecto de medicamentos proporciona información sobre usos correctos e indebidos o efectos secundarios, abstrayendo al usuario de las descripciones químicas detalladas, un sistema de ML debe ofrecer a sus usuarios información significativa que los haga conscientes de la lógica aplicada, así como la importancia y las consecuencias esperadas del procesamiento.

Al procesar datos personales utilizando ML, los responsables del tratamiento deben informar adecuadamente a los interesados sobre los posibles impactos en su vida diaria.

Algunos ejemplos de información significativa podrían ser el describir las limitaciones del sistema, las métricas de rendimiento alcanzadas, los datos personales utilizados como entrada y generados como salida, el impacto de ciertos datos de entrada en los resultados, las comunicaciones a terceros, y los riesgos que supone para los derechos y libertades.

8 MALENTENDIDO

Los sistemas de ML están sujetos a menos sesgos que los propios humanos.

Realidad: Los sistemas de ML están sujetos a diferentes tipos de sesgos y algunos de estos provienen de sesgos humanos.

Los modelos de ML pueden estar libres de sesgo humano o favoritismo hacia un individuo o un grupo en función de sus características inherentes o adquiridas. Sin embargo, los sistemas de ML se seleccionan, diseñan, ajustan y entrenan con datos, que, en la mayoría de los casos, fueron seleccionados por humanos. Los sistemas de ML podrían estar sujetos a más de veinte tipos de sesgos derivados del procesamiento de datos[17].

Algunos de los sesgos que afectan a los sistemas de ML replican los sesgos humanos (por ejemplo, un modelo entrenado con perfiles históricos de CEO estará sesgado hacia candidatos masculinos). Otros posibles sesgos de ML dependen de decisiones humanas, como la forma en que se muestrean los datos de entrenamiento o el modo en el que se presentan los resultados. A veces, los sistemas de aprendizaje automático se utilizan en contextos que no son los mismos para los que se diseñaron los modelos.

En resumen, el objetivo de los sistemas de ML es reflejar la experiencia y el conocimiento proporcionados por sus creadores.

Sin embargo, los sistemas de ML no incorporan la humanidad requerida para manejar situaciones excepcionales: no tienen una visión global del problema y tienen una capacidad limitada para adaptarse a los cambios contextuales y ser flexibles ante circunstancias imprevistas.

17 Ninareh MehrabiFred Morstatter, Nripsuta Saxena, Kristina Lerman, y Aram Galstyan. A Survey on Bias and Fairness in Machine Learning.". (2019) **https://arxiv.org/abs/1908.09635v2**

9

MALENTENDIDO

ML puede predecir con precisión el futuro.

Realidad: Las predicciones de un sistema ML solo son precisas cuando los eventos futuros reproducen tendencias pasadas.

Los sistemas de ML tienen en cuenta los datos presentes en los conjuntos de datos y los utiliza para extraer proyecciones de posibles resultados futuros.

Por tanto, los sistemas de ML no hacen conjeturas sobre el futuro, sino más bien pronósticos, que se basan en eventos pasados y se proporcionan a los sistemas durante el entrenamiento.

Algunos modelos de aprendizaje de ML podrían evolucionar para adaptarse a nuevos datos, como los modelos utilizados en la creación de perfiles en marketing o medios de comunicación online.

Sin embargo, son incapaces de adaptarse a un escenario completamente nuevo o a eventos que cambien rápidamente. Para adaptar sus predicciones a tales cambios, la mayoría de los modelos necesitarán grandes cantidades de nuevos datos.

10

MALENTENDIDO

Los interesados son capaces de anticipar las posibles salidas que los sistemas de ML pueden dar con sus datos.

Realidad: La capacidad de ML para encontrar correlaciones no evidentes en los datos puede terminar por revelar información personal adicional, sin que el interesado sea consciente de ello.

Los sistemas de ML son excelentes para encontrar correlaciones en los datos. Son capaces de identificar patrones en los datos personales que van más allá de los planteados explícitamente en el desarrollo del modelo, y que podrían ser desconocidos incluso para los individuos afectados (por ejemplo, una predisposición a una enfermedad). Este potencial plantea varias preocupaciones desde el punto de vista de la protección de datos.

Por un lado, los interesados pueden verse afectados por decisiones basadas en información que no conocen y no tenían forma de anticipar y/o reaccionar.

Por otro lado, los sujetos de datos pueden recibir información descubierta por un modelo de ML sobre ellos en lugares o situaciones donde podrían producir un mayor impacto en sus vidas debido al contexto específico. Por ejemplo, al recibir cupones de descuento por correo de una tienda, basados en sus hábitos de compra, podría revelar cierto nivel de adicción a los juegos de azar.

Cuando los sistemas de ML procesan datos personales para crear inferencias más allá del propósito declarado del tratamiento, por ejemplo, al realizar algún tipo de elaboración de perfiles (predicciones o clasificaciones) de individuos, el responsable debe cumplir con todos los principios de protección de datos, incluida la licitud, transparencia (art. 5.1.a) del RGPD) y limitación de la finalidad (art. 5.1.b) del RGPD).

Cualquier tipo de procesamiento posterior de datos personales requiere una base legal y un propósito claro.