Santander X Innovation Xperts

La interpretabilidad desde el diseño: oportunidades futuras

Octubre 2023

By Santander

Autora

Dra. Irene Unceta

Irene Unceta estudió Física en la Universidad de Barcelona y obtuvo un Máster en Informática (MSc in Computational Science) por la Universidad de Ámsterdam. Realizó sus estudios de doctorado en la Universidad de Barcelona y en BBVA Data & Analytics, donde obtuvo un doctorado industrial en interpretabilidad de modelos de aprendizaje automático aplicados al riesgo financiero. Científica social y de datos a partes iguales, su carrera profesional ha sido un recorrido por la intersección de ambos mundos, con un enfoque principal en el impacto social y la rendición de cuentas del aprendizaje automático.

Índice

01

Introducción

El aprendizaje automático augura un procesamiento de la información más preciso, eficiente y coherente en una amplia gama de campos[1]**.** Desde las finanzas y los seguros hasta la sanidad, la publicidad o el sistema de justicia penal, esta tecnología podría mejorar los procesos de toma de decisiones y suponer importantes beneficios para los usuarios, las empresas y la sociedad en general. Sin embargo, para cumplir esta promesa, los sistemas basados en el aprendizaje automático deben superar ciertas barreras que obstaculizan su uso y adopción.

Las arquitecturas tradicionales de aprendizaje automático, como los árboles de decisión y los sistemas basados en reglas, se adaptan bien a los procesos de representación del conocimiento humano y pueden descomponerse en reglas sencillas e identificables. Tienen, además, un tamaño acotado y su número de variables es limitado. Por el contrario, las tendencias actuales en el diseño y la aplicación de modelos favorecen las arquitecturas grandes de ajuste preciso y relativamente generales.

Los recientes avances en visión artificial y procesamiento del lenguaje natural han introducido mecanismos de autoatención y autosupervisión como, por ejemplo, los de los transformadores. Los modelos híbridos también han ganado popularidad gracias a su capacidad de combinar diferentes estructuras de aprendizaje profundo y medidas probabilísticas para modelar la incertidumbre. Estas arquitecturas son capaces de explotar las nolinealidades de los datos gracias a capas de abstracción muy recursivas. Los artefactos matemáticos necesarios para describir las relaciones entre estas capas son cada vez más complejos. En consecuencia, los resultados son más difíciles de entender. La comprensión suele ser incompatible con el tamaño y la complejidad. La tendencia a aumentar el número y el tamaño de las capas, la complejidad y la recursividad de las transformaciones no lineales, o el número de aprendices, derivará naturalmente en una mayor opacidad.

En este artículo se definen y evalúan los problemas resultantes de la creciente opacidad que, en el futuro inmediato, caracterizará las tendencias del aprendizaje automático. En el **primer apartado**, se abordará la cuestión del rendimiento. A continuación, en el **segundo**, se evaluarán las soluciones existentes, entre ellas la explicabilidad. En el **tercer apartado**, se presenta la interpretabilidad como una posible alternativa, y se defiende como la vía más prometedora para avanzar. Además, se identifican las principales oportunidades que ofrecen los modelos interpretables. Luego se analizan los retos para conseguir deliberadamente la interpretabilidad en la práctica y se indican los ámbitos y los sectores clave en los que estos modelos pueden suponer una ventaja.

1. Los modelos de aprendizaje automático entran en la categoría de métodos actuariales de toma de decisiones, que se establecen a partir de relaciones establecidas empíricamente y codificadas en fórmulas matemáticas que producen resultados automatizados. Como tales, están pensados para evitar los sesgos, las distorsiones y las creencias relacionados con el método clínico, que se basa por completo en el procesamiento humano de la información [Dawes *et al.*, 1989].

02

El reto que plantean la complejidad y la opacidad

Los modelos de aprendizaje automático pueden resultar difíciles de comprender para los humanos, con independencia de cuál sea su arquitectura concreta. Los usuarios no suelen entender bien la lógica que permite a los modelos producir un resultado a partir de la introducción de datos, y los consideran opacos. Esta *opacidad* puede ser intencionada[2]. Los modelos industriales de aprendizaje automático pueden estar sujetos al secreto corporativo o estatal debido al interés legítimo de las empresas, que así consiguen y mantienen una ventaja competitiva[3]. La opacidad también puede deberse al analfabetismo técnico. La población carece, en general, de las competencias necesarias para leer código[4] o seguir las complejas derivaciones matemáticas que hay detrás de la mayoría de los algoritmos. En muchos casos, sin embargo, la **opacidad es inherente al aprendizaje automático**. Muchas veces, los modelos, para extraer conocimiento a partir de grandes volúmenes de datos, entrañan un grado de complejidad que no es compatible con las exigencias del razonamiento a escala humana[5]. Ocurre así sobre todo en el caso de los modelos basados en el aprendizaje profundo y, en menor medida, también en los métodos de conjunto. Las razones que hay tras esto son múltiples.

En primer lugar, los modelos de aprendizaje profundo son grandes y constan de múltiples capas de neuronas ocultas. Esto supone un problema para interpretar su funcionamiento. En el mejor de los casos, las personas pueden entender cómo funcionan las distintas capas por separado. Sin embargo, su capacidad para retener información en la memoria es limitada. Por eso, cuando hay un número creciente de capas, la intuición humana falla en dimensiones considerables. Además, la comprensión de cada uno de los componentes o capas de un modelo puede no ser suficiente para entenderlo en su conjunto. A este efecto se le suele llamar "más es diferente"[6]: el conjunto tiene características adicionales y nuevas que no pueden entenderse mediante la simple observación de las partes por separado. Por lo tanto, la estructura agregada, el modelo, sigue siendo opaca.

2. Para un análisis más profundo de las diferentes formas de opacidad en el aprendizaje automático, véase [Burrel, 2016].
3. Además de la competitividad, los aspectos internos de los modelos pueden mantenerse en secreto en aras de la seguridad, como se expone en [Sandvig *et al.*, 2014].
4. Incluso cuando se cuenta con las capacidades adecuadas, leer código, ya sea comercial o de otro tipo, puede resultar difícil si no existen unas prácticas y unos estándares bien definidos [Matteas *et al.*, 2005].
5. Cabe destacar que se trata de una forma de opacidad que no está relacionada con la capacidad técnica y que afecta indistintamente a diseñadores, programadores y usuarios [Seaver, 2013].
6. Este es un efecto bien conocido en los estudios de sistemas complejos. Fue descrito por primera vez en [Anderson, 1974].

En segundo lugar, una característica fundamental de los modelos de aprendizaje profundo es su alta recursividad: las capas subsiguientes se alimentan y conectan entre sí. Se trata de un mecanismo que añade otra capa de complejidad. Esto ocurre sobre todo en el caso de los *transformadores*, que permiten asignar memoria a las neuronas y hacen que las capas individuales dejen de ser simples. Otra consecuencia de la recursividad es que el nivel de abstracción de las representaciones del conocimiento aprendido deja de ser diacrónico. Los modelos tradicionales proyectaban los datos en estados que se volvían más abstractos con la profundidad. En las capas altamente recursivas, en las que la información va de un lado a otro a través de las distintas capas, eso ya no es así.

En tercer lugar, con la llegada de los modelos de conjunto híbridos y profundos que combinan diferentes arquitecturas surge la cuestión adicional de cómo se combinan estas arquitecturas en un sistema único. Los modelos de conjunto tradicionales combinaban los resultados de aprendices de árbol múltiples o potenciados por gradiente. Entenderlos requería, en primer lugar, inspeccionar los aprendices individuales y luego centrarse en la manera en que se agregaban sus resultados. Los modelos híbridos actuales combinan diferentes arquitecturas de aprendizaje profundo, entre ellas las redes neuronales recurrentes (RNN, por sus siglas en inglés), las redes de memoria a corto-largo plazo (LSTM), las redes neuronales convolucionales (CNN) o las redes generativas antagónicas (GAN) con otros tipos de modelos, como las máquinas de vectores de soporte (SVM) o los modelos bayesianos. La agregación de múltiples aprendices de árbol supone un reto para la comprensión. Entender los sistemas que combinan estas arquitecturas enseguida se vuelve inviable.

Por último, incluso en caso de que los modelos sean sencillos, los sistemas predictivos resultantes pueden seguir siendo opacos. Como se tratará más adelante con mayor detalle, el entrenamiento y la aplicación del aprendizaje automático son un proceso complicado con varias fases. Durante la fase de preprocesamiento de datos, con frecuencia se combinan variables brutas para obtener un conjunto reducido de atributos altamente predictivos que capturen las nolinealidades de los datos. Este proceso ofusca los atributos originales y da lugar a características difícilmente comprensibles para el ciudadano normal. Por eso los modelos basados en estas características, aunque sean sencillos por naturaleza, deben seguir considerándose opacos.

La opacidad, pues, puede emanar del tamaño de los modelos, que pueden tener múltiples capas, o de los patrones que surgen de las intrincadas relaciones entre estas capas, así como de la agregación de varias estructuras multicapa.

Sea cual sea su origen, **la opacidad, y en consecuencia la complejidad, representan una barrera para la adopción del aprendizaje automático a gran escala**. Esta barrera no solo se define como los problemas derivados de la propia opacidad, sino también (y, quizá, de manera más importante) como un interrogante sobre si la complejidad aporta mejoras significativas al rendimiento, la eficiencia y la sostenibilidad.

Figura 1: Problemas derivados de los modelos complejos y opacos

Relacionados con el rendimiento	Derivados de la complejidad	Derivados de la opacidad
→ No hay suficientes pruebas que respalden una correlación entre complejidad y opacidad	→ Ineficiencias en el entrenamiento y la configuración → Consumo de recursos insostenible	→ Modelos empresariales expuestos a puntos ciegos → Usuarios/consumidores, sobre todo en tomas de decisiones de alto riesgo → Incumplimiento de la normativa

Fuente: creado por el autor.

El rendimiento

Buena parte del éxito actual de los modelos complejos se debe a **la creencia de que existe un *trade-off* entre precisión e interpretabilidad**. Cuanto más complejo es un modelo, se supone, mejor es su rendimiento. Sin embargo, esta aparente relación no se basa en pruebas experimentales sólidas y sigue siendo, a día de hoy, una mera percepción[7]. Por el contrario, en los escenarios reales el aparente *trade-off* entre precisión e interpretabilidad suele invertirse[8].

7. Uno de los primeros artículos que denunció la falta de pruebas que respalden esta afirmación fue [Rudin, 2019].
8. Cada vez son más las voces que advierten de que apenas existen pruebas experimentales —si es que las hay— del *trade-off* entre precisión e interpretabilidad desde la perspectiva del usuario final. Como se expone en [Herm *et al.*, 2021], este *trade-off* es muy circunstancial y depende, entre otras cosas, de la aplicación considerada y los datos de entrenamiento

En muchas aplicaciones relevantes no se observan diferencias de rendimiento significativas entre los modelos complejos y otros mucho más sencillos[9]. Por ejemplo, las evidencias demuestran que, cuando se evalúan sobre el terreno en ámbitos como la sanidad, la justicia penal o la visión artificial, los modelos simples no son menos precisos que las soluciones complejas de caja negra[10]. A fin de cuentas, los modelos complejos, al parecer, no son siempre la mejor opción cuando se trata de desarrollar herramientas automatizadas de ayuda al proceso de toma de decisiones en contextos de alto riesgo.

9. Aun así, se trata de una creencia muy arraigada en la comunidad del aprendizaje automático, que ha determinado el programa de investigación y desarrollo en este campo durante las últimas décadas. Un ejemplo ilustrativo es que cada semana se publican cientos de artículos que mejoran unos pocos decimales los estándares de rendimiento del aprendizaje profundo, introduciendo pequeños ajustes en los métodos existentes. Por el contrario, la mayoría de los modelos de árboles de decisión que se entrenan ahora, tanto en el mundo académico como en la industria, se basan en CART, un algoritmo que se remonta a 1984. Desde entonces se han propuesto otros algoritmos, entre ellos ID3, MARS o CHAID. Sin embargo, CART sigue siendo el estándar general cuando se trata de entrenar árboles de decisión. Véase la publicación original en [Breiman *et al.*, 1984] Suponiendo que exista una diferencia de rendimiento entre los modelos más complejos y los más sencillos, no se ha hecho demasiado por reducirla.

10. [Angelino *et al.*, 2018] demuestra que sus modelos sencillos, basados en reglas, son tan eficaces como el modelo de caja negra COMPAS a la hora de predecir detenciones de reincidentes en el sistema judicial estadounidense. [Caruana *et al.*, 2015] analiza dos estudios de caso en los que modelos inteligibles proporcionan una precisión puntera para la predicción del riesgo de neumonía en readmisiones hospitalarias a 30 días.

En síntesis

Razones adicionales

La percepción sesgada del rendimiento de los modelos complejos puede atribuirse a las prácticas actuales en la investigación y el desarrollo del aprendizaje automático.

- **Las mejoras atribuidas a los modelos complejos suelen basarse en comparaciones de datos estáticos.** Esta práctica ignora aspectos relevantes de algunos problemas reales. El proceso de extraer conocimiento a partir de datos puede requerir varias iteraciones. Rara vez se basa en una evaluación única y estática. Además, los experimentos no suelen indicar el rendimiento dentro del contexto ampliado en el que operarán los modelos en la vida real, olvidando cuestiones fundamentales relacionadas con la manera en que se utiliza el aprendizaje automático en la mayoría de las empresas. En general, el *trade-off* óptimo entre coste y beneficio de un problema suele estar sujeto a cambios y, por lo tanto, no debería asumirse que es estático. Así pues, es posible que hayamos sobrestimado colectivamente la ventaja competitiva que ofrecen los modelos complejos en las aplicaciones reales.

- **El rendimiento del modelo suele evaluarse asumiendo que los costes son iguales**. En muchos casos, los modelos se evalúan calculando directamente la tasa de error de todas las muestras sin establecer ninguna distinción. Al adoptar este enfoque, podemos estar comparando modelos en escenarios que son sustancialmente diferentes de aquellos en los que serán productivizados. En los entornos de identificación de fraudes es bien sabido que no se puede suponer que el coste de los falsos positivos es igual al de los falsos negativos. Asimismo, en los de calificación crediticia el coste de estimar incorrectamente el riesgo de impago puede ser muy diferente en función del importe de los préstamos considerados.

→ **Esta tendencia tiene un profundo efecto en la manera de concebir y llevar a cabo la investigación**. Por último, no se debe subestimar la influencia de esta tendencia actual, la del aprendizaje profundo, en las publicaciones científicas. Cada semana se publican en las principales revistas cientos de artículos sobre aprendizaje automático que proponen nuevas arquitecturas cada vez más complejas que ofrecen una ventaja predictiva sobre los métodos existentes. Esta ventaja se demuestra comparando las arquitecturas propuestas con otras más sencillas. Curiosamente, estas últimas siempre obtienen peores resultados. Queda por ver si los esfuerzos dedicados a entrenar esas soluciones más sencillas son comparables a los invertidos en desarrollar las alternativas propuestas. De no ser así, puede que los avances teóricos no se traduzcan en un progreso real que respalde la utilización de modelos complejos en aplicaciones reales.

La complejidad

La sostenibilidad. Los modelos complejos hacen un uso intensivo de los recursos informáticos, lo que puede dar lugar a rendimientos decrecientes. Con el auge del aprendizaje profundo, los modelos han experimentado un aumento espectacular del número de parámetros. Las arquitecturas que constan de miles e incluso millones de parámetros son habituales en campos como el procesamiento del lenguaje natural o la visión artificial[11]. El entrenamiento de estas arquitecturas puede llevar horas e incluso semanas. A veces más, porque el entrenamiento, por sí solo, no es garantía de rendimiento. Los modelos de aprendizaje automático se entrenan mediante un proceso de prueba y error. Por lo tanto, una sola iteración de entrenamiento puede no ser suficiente para obtener el rendimiento deseado. Encontrar la configuración óptima de los parámetros puede implicar varios ciclos e incluso cuando el proceso ha terminado, la solución obtenida puede ser incorrecta. El entrenamiento y la resolución de problemas en los modelos complejos de aprendizaje automático puede ser una tarea larga, tediosa y que conlleve muchos cálculos. Una tarea que, además, puede tener importantes costes medioambientales. En los próximos años, cada vez más, la huella de carbono de los modelos de caja negra formará parte del debate público.

11. Los modelos BERT-large y T5-11B de Google tienen alrededor de 350 millones y 1.000 millones de parámetros, respectivamente. El revolucionario modelo de lenguaje autorregresivo GPT-3, de OpenAI, contiene 175.000 millones de parámetros. Megatron-LM, de NVIDIA, incluye 8.000 millones de parámetros. En [Schwarz *et al.*, 2018]. puede encontrarse un resumen más completo del coste computacional de la mayoría de los modelos comerciales de aprendizaje automático.

La eficiencia. El entrenamiento de transformadores, modelos híbridos y otras redes grandes y complejas puede suponer un coste elevado para las empresas[12]. Pero este no se limita al entrenamiento. Dado el tamaño de los modelos, realizar inferencias sobre muestras individuales puede requerir muchos cálculos adicionales, un coste que se suma al del entrenamiento. En algunos casos, estos costes pueden volverse excesivos.

Un ejemplo

AlphaGo

El problema de la eficiencia no es nuevo. En el año 2016, Deep Mind lanzó AlphaGo, un *software* para jugar al juego Go basado en el aprendizaje por refuerzo. Para jugar una partida, el experimento necesitó 1.920 CPU y 280 GPU, y su coste estimado fue de 35 millones de dólares[13]. Entrenar modelos cada vez más complejos es cada vez más insostenible y costoso. Más aún si se tienen en cuenta las recientes subidas del precio de la electricidad. Este tipo de modelos, que en la mayoría de los sectores ya se consideran un bien básico, podrían quedar rápidamente fuera del alcance de muchas empresas. Si ocurriera así, cabría plantearse si la promesa de un mayor rendimiento universal es razón suficiente para seguir invirtiendo únicamente en estos modelos.

12. Una sola iteración de entrenamiento de BERT-large requiere el uso de 64 chips TPU durante cuatro días y tiene un coste estimado de 7.000 dólares.
13. Últimamente, este modelo ha sido objeto de mucha controversia. La publicación original se puede consultar en [Silver *et al.*, 2016]. En respuesta a la indignación pública que despierta el impacto del entrenamiento de este tipo de modelos, Google ha publicado hace poco una serie de buenas prácticas destinadas a reducir la huella de carbono de la inteligencia artificial. El informe completo sigue en fase de revisión. Se puede acceder a una prepublicación en https://arxiv.org/abs/2204.05149. Obsérvese que el artículo no plantea la que quizá sea la pregunta más relevante: ¿son realmente necesarios estos modelos o existen otras alternativas viables?

La opacidad

Los modelos de negocio. La opacidad supone, por sí sola, un riesgo para las empresas, que sin darse cuenta pueden implementar soluciones defectuosas[14]. Por ejemplo, los modelos pueden estar sesgados y desfavorecer a ciertos colectivos o minorías, basarse en suposiciones erróneas o llevar a predicciones inexactas. La opacidad puede impedir que las empresas identifiquen y resuelvan estos problemas antes de poner sus modelos en producción. Lo cual puede causar importantes perjuicios financieros y/o de reputación[15].

Los usuarios/clientes. La opacidad también supone un riesgo evidente para los usuarios, que tal vez tengan que enfrentarse a decisiones que no entienden o no son capaces de analizar[16]. En el peor de los casos, esto puede impedirles hacer valer sus derechos frente a una toma de decisiones automatizada.

El cumplimiento de la normativa. La cuestión anterior ha llevado a muchos Gobiernos a apoyar la regulación de las herramientas artificiales de apoyo a la toma de decisiones. Desde el 25 de mayo de 2018, la toma de decisiones exclusivamente automatizada está prohibida de manera estricta en todos los Estados miembros de la Unión Europea[17]. En los casos en que los ciudadanos se hallan sujetos a decisiones de alto riesgo basadas, en parte, en sistemas de ayuda de aprendizaje automático, el Reglamento General de Protección de Datos (RGPD) establece que tienen derecho a recibir información comprensible sobre la lógica que subyace en esas decisiones. Entre los campos afectados están, entre otros, la calificación crediticia, la elaboración de perfiles de candidatos o la identificación de fraudes, cuyo impacto, en todos los casos, es significativo en la vida de las personas. El incumplimiento de este reglamento puede causar importantes pérdidas económicas a las empresas que implementen estos sistemas en Europa.

---

14. Los modelos de aprendizaje automático son tan buenos como los datos con los que han sido entrenados [Crawford, 2013]. Cuando estos datos son incorrectos, la opacidad impide que las empresas sean capaces de identificar y corregir los errores.
15. En los últimos años, muchas voces han denunciado las posibles consecuencias negativas del hecho de que las empresas y las instituciones públicas deleguen la toma de decisiones en modelos cuyo funcionamiento interno no se comprende del todo. Véanse, por ejemplo, [Eubanks, 2018] y [O'Neil, 2016].
16. El aprendizaje automático se utiliza cada vez más para fundamentar decisiones de alto riesgo, por lo que se ha dicho y escrito mucho sobre cómo estas prácticas pueden afectar al derecho de las personas a obtener información comprensible sobre el mecanismo que hay detrás de la toma de decisiones automatizada [Selbst, 2017].
17. Si este reglamento reconoce de verdad un derecho a la explicación queda fuera del ámbito de este trabajo. En esto, me alinearé con [Selbst & Powles, 2017] para afirmar que puede interpretarse que el reglamento reconoce al menos el derecho de los usuarios a recibir la información mínima necesaria que les permita impugnar las decisiones a las que están sometidos y, en última instancia, reivindicar sus derechos si se toman decisiones erróneas o injustas.

En síntesis

De la toma de decisiones de bajo riesgo a la de alto riesgo

La utilización del aprendizaje automático se popularizó primero para aplicaciones como la publicidad, la recomendación de productos o la búsqueda en la web. Las decisiones tomadas en estos ámbitos se denominan de *bajo riesgo*. Pueden ser una fuente de ingresos para las empresas, pero no afectan de manera importante a la vida de las personas. En cambio, las que se adoptan en ámbitos como la detección de fraudes, la calificación crediticia o la predicción del resultado de las audiencias de libertad condicional sí afectan de una manera directa. Estas se denominan de *alto riesgo*. El aprendizaje automático surgió como una herramienta de ayuda en el proceso de toma de decisiones para aplicaciones de bajo riesgo y ha ido aumentado poco a poco su presencia en contextos de alto riesgo, donde el aprendizaje profundo se ha convertido en algo habitual en muchos sectores. Sin embargo, los contextos de bajo riesgo y de alto riesgo tienen necesidades diferentes y requieren prácticas distintas. Al no proporcionar información sobre cómo se combinan los atributos de entrada para realizar predicciones, los modelos opacos impiden que las personas entiendan cómo se realizan las predicciones individuales. Esta falta de comprensión, que puede resultar aceptable en las decisiones de bajo riesgo, puede tener graves consecuencias en la toma de decisiones de alto riesgo, en la medida en que disuade a los usuarios de impugnar decisiones que tienen un efecto directo en su vida cotidiana.

03

¿Es la explicabilidad una solución?

La última década ha sido testigo del auge de la investigación destinada a obtener explicaciones *post hoc* que puedan proporcionar esa información significativa abordando, al menos en parte, la lógica aprendida por un modelo entrenado. Este planteamiento toma como punto de partida la existencia de un modelo complejo que necesita ser explicado. Los intentos de aclarar cómo funciona suelen consistir en replicar su comportamiento, bien a escala local o global, utilizando otro modelo distinto que sea más fácil de entender[18]. Este enfoque permite desvelar en parte los mecanismos internos de los modelos complejos. Sin embargo, solo ofrece una solución parcial e incompleta que no puede considerarse definitiva.

La explicabilidad no mejorará el rendimiento. Las técnicas orientadas a explicar *post hoc* los modelos complejos pueden ayudar a desvelar la lógica aprendida pero no pueden aumentar su rendimiento. Estos métodos pueden describir un modelo, pero no modificarlo. En este sentido, la explicabilidad no soluciona la falta de relación entre complejidad y rendimiento descrita antes.

La imprecisión. Estas explicaciones son aproximaciones a lo que pretenden describir. Como tales, incurren en alguna pérdida de información. Las explicaciones obtenidas de modelos complejos siempre están en cierta medida equivocadas. Si estas pudieran reproducir el funcionamiento de un modelo con una fidelidad perfecta, este dejaría de ser necesario. En el mejor de los casos, las explicaciones pueden aspirar a recuperar la mayor parte de la lógica del modelo. Pero incluso en los casos en que lo consiguen, no pueden ser completas: un método explicativo con un 95% de fidelidad se equivoca un 5% de las veces. Por lo tanto, las explicaciones p*ost hoc* nunca pueden ser del todo fieles a los modelos que pretenden explicar[19].

La relevancia. Además, está la cuestión de hasta qué punto las explicaciones *post hoc* son relevantes. Un método que pretende aclarar cómo funciona un sistema determinado debería identificar los factores que este tiene en cuenta a la hora de tomar decisiones. También debería describir cómo se utilizan esos factores y cuál es su contribución individual al resultado. Las actuales formas de explicación no suelen ofrecer esa descripción.

---

18. Algunos ejemplos de estas técnicas son LIME [Ribeiro *et al.*, 2016] y su versión más reciente, Anchors [Ribeiro *et al.*, 2018], así como SHAP [Lundberg *et al.*, 2017].
19. Las explicaciones pueden ser imprecisas o válidas solo para determinadas partes del espacio de características. Remito al lector a [Mittelstadt *et al.*, 2019] para una descripción más detallada de esta cuestión y sus consecuencias.

La incompletitud. Para estar completas, las explicaciones tendrían que describir los distintos elementos que intervienen en el diseño de una solución de aprendizaje automático. Cuando se implementan en escenarios reales, los modelos de aprendizaje automático casi nunca existen de forma aislada. Constituyen, más bien, una pequeña parte de una estructura mayor que implica un sistema de inteligencia artificial[20]. Este sistema lo forman el propio modelo y también los datos de entrenamiento, la estrategia de preprocesamiento de datos, la infraestructura de producción o cualquier dependencia de *software* de terceros. Explicar un sistema de aprendizaje automático exige comprender cómo interactúan entre sí estos distintos elementos. Por ejemplo, los modelos suelen combinar conocimientos externos con los datos de entrenamiento. Este conocimiento puede estar relacionado con la estrategia de negocio de una empresa: esta puede optar por excluir ciertos casos o asociar un mayor valor a otros. Entender cómo el modelo utiliza este conocimiento es relevante para explicar sus predicciones. Sin embargo, los métodos de aprendizaje profundo nos impiden acceder a esta información. Por lo tanto, las explicaciones que se obtengan de ellos no podrán dar una visión completa.

Ejemplo

Los mapas de saliencia

Los mapas de saliencia suelen utilizarse para explicar modelos complejos de aprendizaje profundo basados en datos de imagen o de vídeo. Destacan las regiones de la imagen en las que se centra el modelo. Esto puede utilizarse para determinar el grado de importancia de los diferentes píxeles. Sin embargo, los mapas de saliencia no nos dicen nada acerca de cómo utiliza el modelo estos píxeles. Pueden identificar en qué se fija el modelo, pero no aportan ninguna información sobre cómo el modelo procesa lo que ve. Si no se cuenta con esta información, las explicaciones basadas en técnicas como los mapas de saliencia difícilmente pueden considerarse relevantes o completas.

20. Algunos actores importantes de la comunidad del aprendizaje automático han impulsado la utilización del sistema más general cuando se refieren al aprendizaje automático, con el objetivo de destacar los muchos elementos que intervienen en la concepción de una herramienta automatizada de toma de decisiones. Véase, por ejemplo, la Declaración de Montreal para un desarrollo responsable de la inteligencia artificial en: https://www.montrealdeclaration-responsibleai.com/_files/ugd/ebc3a3_5c89e007e0de440097cef36dcd69c7b0.pdf

04

Las soluciones interpretables: un enfoque alternativo

La creación de modelos que sean *interpretables* desde el principio es otro planteamiento que ha recibido menos atención. Las llamadas *cajas blancas o transparentes* son modelos inherentemente comprensibles, que se han entrenado teniendo en cuenta la interpretabilidad. Entre ellos están los árboles de decisión y los sistemas basados en reglas, como ya se ha explicado. Pero también otros métodos de representación del conocimiento que se adaptan mejor a la comprensión humana. Como se aborda a continuación, la utilización de estos modelos interpretables puede tener varias ventajas respecto a la explicabilidad *post hoc*.

La rentabilidad a largo plazo. Desarrollar modelos interpretables puede resultar costoso. Pero en la mayoría de los casos, el coste de un error es mucho mayor que el de dedicar tiempo y recursos a desarrollar una solución interpretable. Las ventajas a este respecto son múltiples. Los modelos interpretables permiten inspeccionar su interior. Lo cual hace que el proceso de localizar y corregir errores a tiempo sea mucho más ágil.

La flexibilidad. Los modelos interpretables también son más adecuados para trabajar en distintos entornos. Al entender cómo funcionan, pueden adaptarse fácilmente a nuevos contextos, modificando determinadas piezas. Por el contrario, los modelos opacos pueden resultar muy frágiles. Cuando se enfrentan a escenarios distintos de aquellos para los que fueron diseñados pueden generar pérdidas de rendimiento significativas, aunque las diferencias entre las configuraciones no sean importantes. Esto hace que la construcción de sistemas de predicción sólidos basados en cajas negras sea difícil. Dado el elevado coste reputacional y económico que tienen las predicciones incorrectas o deficientes, a largo plazo los modelos interpretables son más rentables. Más aún si se tiene en cuenta que, al comprender sus mecanismos, los elementos de los modelos interpretables se pueden reutilizar, evitando así los elevados costes del reentrenamiento.

Sin embargo, este cambio en la manera de concebir y utilizar los modelos de aprendizaje automático en contextos de alto riesgo como los descritos no solo estará motivado por la regulación. A medida que se extienda el uso de esta tecnología, los usuarios exigirán una mayor transparencia para confiar en ella. Investigaciones recientes sugieren que la interpretabilidad desempeña un papel clave en la aceptación de los productos y servicios basados en el aprendizaje automático[21]. De ser así, es probable que la **adopción de modelos interpretables vaya más allá del rendimiento y los costes**.

21. El papel de la interpretabilidad en la percepción del consumidor se estudia en [Shin, 2021] y [Wanner *et al.*, 2020]

En síntesis

Anticiparse a los riesgos normativos y evitarlos

Invertir en modelos de aprendizaje automático cuyo diseño los haga interpretables también debería ayudar a las empresas a adelantarse a las posibles tendencias y anticiparse a futuros problemas. En abril de 2021, la Comisión Europea publicó la propuesta de Ley de IA[22], que fue adoptada por el Parlamento Europeo en junio de 2023, iniciándose las conversaciones con los países miembros sobre la forma final de la ley. La Ley de IA se basa en el RGPD y será el primer intento de legislar sobre IA en todo el mundo. Clasifica la IA en tres categorías de riesgo: riesgo inaceptable, riesgo alto y riesgo bajo o mínimo. Se entenderá que plantean un riesgo inaceptable los usos que supongan una amenaza grave para el bienestar o la privacidad de los usuarios. Aplicaciones como el reconocimiento facial en espacios públicos, los sistemas de crédito social o las técnicas subliminales, que entran en esta categoría, estarán prohibidas.

Las aplicaciones de aprendizaje automático que tengan un efecto considerable en la vida de las personas se considerarán de alto riesgo. Esto podría incluir la identificación y la categorización biométricas; la gestión del empleo; la aplicación de la ley; el desarrollo de componentes de seguridad; el acceso y disfrute de servicios privados esenciales y de servicios y prestaciones públicas; el acceso a la educación y la formación profesional, así como sus tareas y evaluación; el asilo de migrantes y la gestión del control de fronteras; y la administración de justicia y los procesos democráticos, así como otros muchos ámbitos de alto riesgo. En todos los casos, las aplicaciones incluidas en la categoría de alto riesgo estarán sujetas a leyes y prohibiciones estrictas. A los sistemas de IA de alto riesgo se les exigirá, entre otras cosas, que sean transparentes y permitan a los usuarios interpretar sus resultados y utilizarlos de manera adecuada. Las empresas tendrán que adaptar su tecnología para cumplir estas nuevas normas si quieren seguir siendo competitivas en el mercado europeo de la IA en el futuro próximo.

22. La propuesta completa puede consultarse en: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206

05

Cómo buscar la interpretabilidad en la práctica

En la práctica, conseguir la interpretabilidad de manera deliberada exigirá avances en el entrenamiento y el diseño de modelos como los árboles de decisión o los sistemas basados en reglas. Estos avances deben tener en cuenta la investigación sobre los criterios de optimización y las estrategias de poda para los árboles, así como el desarrollo de nuevas técnicas para el diseño de reglas binarias y multinarias. Otras alternativas, como los modelos lineales, y entre ellos la regresión logística, también podrían ser de interés. Con independencia de su estructura concreta, los modelos interpretables deben tener un tamaño limitado o, cuando no sea así, proporcionar una segmentación clara de los casos. Deben basarse en atributos brutos, que sigan siendo inteligibles para un público amplio. A este respecto, un reto importante será diseñar técnicas para captar los efectos no lineales sin obstaculizar la comprensibilidad general.

Esto no significa que deban dejar de utilizarse los modelos de caja negra. Al contrario, existen múltiples ámbitos de aplicación, por ejemplo los relacionados con el procesamiento de textos e imágenes, en los que suponen una ventaja cualitativa. En aquellos en los que presumiblemente no es así, la exploración de nuevos métodos y técnicas para entrenar modelos inherentemente interpretables puede ofrecer oportunidades sustanciales.

Probablemente **sea en los ámbitos calificados de alto riesgo**, como el empleo, el acceso a servicios esenciales privados y públicos, incluidos el crédito y los seguros, la educación, la física nuclear o el sistema judicial, en los que los **modelos interpretables aportarán mayores beneficios**, al garantizar el cumplimiento de la futura regulación. Estos beneficios también se harán extensivos a las decisiones de alto riesgo en otros sectores, sobre todo en aquellos casos en que los usuarios puedan interactuar directamente con los resultados de los modelos.

En los entornos de bajo riesgo, las ventajas de los modelos complejos posiblemente superen a las desventajas. Por ejemplo, en la optimización de los procesos de producción, la publicidad *online* o la recomendación de productos, donde las decisiones adoptadas por los sistemas de aprendizaje automático no tienen un impacto directo en la vida de las personas. En estos casos, la comprensibilidad del modelo puede no ser un imperativo, y las explicaciones incompletas pueden satisfacer las necesidades de transparencia.

En este sentido, **hay que distinguir entre la perspectiva de los usuarios y de las empresas respecto a los entornos de alto y bajo riesgo**. Las decisiones que se toman para publicitar un producto en lugar de otro, o para recomendar la inversión en un servicio y no en otro, pueden ser cruciales para una empresa, aunque no sean de alto riesgo para el usuario. En casos como este, el marco regulador no influirá en la elección del enfoque de modelización adecuado. Sin embargo, es posible que las empresas deseen entender bien por qué se están tomando esas decisiones.

06

Conclusiones

El auge de unos modelos de aprendizaje profundo cada vez más complejos plantea un reto en términos de complejidad y opacidad, en muchas aplicaciones. Un reto que aún no se ha resuelto del todo. Solucionar este problema optando ahora por un rumbo claro ayudará, al mismo tiempo, a crear valor y a evitar costes imprevistos a medio plazo.

La utilización de modelos complejos surge de una creencia muy arraigada según la cual la mayoría de los problemas son inherentemente complicados y, por lo tanto, se necesitan modelos intrincados para resolverlos. Muchos de estos problemas, probablemente la mayoría, sí son complejos. Aun así, los **problemas complicados no necesitan por fuerza soluciones complicadas**[23]. Con frecuencia, la complejidad reside en la búsqueda de una solución sencilla.

Siguiendo con los argumentos ya expuestos, exigir que **la elección de un modelo de caja negra frente a una solución interpretable se base en la oportunidad y no en ideas preconcebidas parece que conduce a un rendimiento y una eficacia mayores**. Los modelos que por su diseño son interpretables pueden resultar insuficientes para alcanzar determinado rendimiento o satisfacer los requisitos de uso de ciertas aplicaciones. En casos así, los modelos de caja negra pueden ofrecer una ventaja competitiva, siempre que las afirmaciones que se hagan al respecto estén suficientemente respaldadas por pruebas empíricas sólidas. En los demás casos, debe considerarse en primer lugar la inversión en soluciones sencillas.

Avanzar hacia un escenario en el que los modelos interpretables ocupen un lugar más relevante puede tener un **coste de transición**. Sin embargo, hacerlo a tiempo brindará a las partes interesadas del sector privado una **oportunidad única de prepararse para un escenario futuro** en el que es probable que los costes de entrenamiento y de evaluación aumenten, en el que los nuevos avances en la regulación exigirán que las decisiones de alto riesgo sean transparentes y en el que los usuarios no confiarán en productos y servicios que no puedan comprender suficientemente bien. Dejar atrás a corto plazo la situación actual contribuirá a una adopción masiva del aprendizaje automático y, por lo tanto, tendrá un impacto positivo a medio y largo plazo.

23. Se podría hacer un argumento similar al de la navaja de Ockham para afirmar que, en muchos ámbitos, existen soluciones sencillas para resolver problemas complejos. Algunos investigadores han presentado argumentos parecidos. Véase, por ejemplo [Rudin, 2019], que introduce el argumento del efecto *Rashomon*, o [Hand, 2006] que expone el efecto *flat minima* para demostrar que en la mayoría de los casos deben existir soluciones simples y precisas

a1 Apéndice | Los retos futuros de los modelos de caja transparente y cómo afrontarlos

Figura 2. Retos de los modelos de caja transparente y cómo abordarlos

Reto	Enfoque
Falta de capacidades y herramientas para desarrollar soluciones más sencillas	Reorientación, pasar de la competencia técnica al pensamiento analítico + Inversión en ciencia básica + Inversión en ciencia aplicada + Adaptación de las estrategias de selección de candidatos al nuevo objetivo
No se puede garantizar el secreto en las soluciones patentadas	Énfasis en la ventaja competitiva de las empresas a medio plazo, gracias a → Preferencia de los usuarios por la interpretabilidad → Mejoras en el rendimiento y la eficiencia → Minimización de los errores

El cambio de paradigma hacia modelos más interpretables plantea problemas específicos. Los modelos interpretables pueden requerir un esfuerzo considerable en cuanto a conocimiento del dominio. En la actualidad, la educación y la formación en IA se centran sobre todo en los métodos. La comprensión teórica de los problemas y sus implicaciones suele dejarse de lado en favor de la acción. Así, son escasas las competencias orientadas al desarrollo de soluciones más sencillas que planteen un reto teórico. Las herramientas disponibles para desarrollar tales soluciones también son limitadas, y a menudo se han quedado obsoletas. La adopción de estas herramientas pasa por dirigir los esfuerzos al **desarrollo de *software* que sea muy accesible, para utilizarlas** y devolverlas al centro del debate. Como ocurre en tantas ramas de la ciencia, el conocimiento sobre los sistemas interpretables se ampliará a medida que lo haga la comunidad dedicada a su estudio y aplicación práctica.

Para que esto ocurra, **la inversión debe dirigirse a la ciencia básica**. Entender mejor lo que implica el proceso de aprendizaje revertirá positivamente en cómo se diseñan los modelos. Por ejemplo, una creencia habitual es que los modelos de caja negra pueden identificar mejor los patrones relevantes en los datos. Sin embargo, si esos patrones son de verdad relevantes, puede que, con las técnicas adecuadas, un modelo interpretable también sea capaz de encontrarlos.

La inversión también debe dirigirse a la ciencia aplicada. Para que los investigadores diseñen modelos interpretables, debe existir la tecnología que permita hacerlo. Esto puede suponer un reto en cuanto a contratación. Las estrategias de selección de candidatos quizá tengan que adaptarse al nuevo conjunto de competencias que se necesitan. La mayoría de los profesionales que trabajan ahora en inteligencia artificial conocen bien programas como Tensorflow o Pytorch, que permiten entrenar arquitecturas de aprendizaje profundo. Como se ha mencionado antes, tratar con modelos interpretables puede requerir un **conjunto de capacidades completamente diferente**, menos centrado en la competencia técnica y más **orientado al pensamiento analítico**.

Además, está la cuestión de **cómo garantizar el secreto y los derechos de las soluciones patentadas si se quiere que estas sean interpretables**. En la actualidad, muchas empresas obtienen beneficios de la propiedad intelectual conferida a los modelos de caja negra. Si se optara por los modelos interpretables, esos beneficios desaparecerían y las empresas tendrían que adaptar sus modelos de negocio al nuevo escenario[24]. Lo cual puede ser un problema para aquellas que aspiran a establecer su propia cuota de mercado en el sector del aprendizaje automático. Si los elementos internos de los modelos utilizados para fundamentar la toma de decisiones de alto riesgo se hacen públicos y son accesibles, esto puede impedir que las empresas obtengan beneficios legítimos de ellos. Por lo tanto, hay que insistir en que las **empresas seguirán obteniendo ventajas competitivas** con el desarrollo de modelos interpretables que los usuarios entiendan mejor, o cuyos resultados sean mejores que los de los modelos existentes, y que esto tendrá presumiblemente un impacto positivo en su balance.

24. En la actualidad, la expresión de un algoritmo en un archivo de código fuente o un script de programación puede tener derechos de autor, aunque los algoritmos en sí no los tienen. Las empresas pueden impedir la reproducción no autorizada de su código fuente o proteger productos y servicios específicos basados en ese código, pero los algoritmos que hay detrás de ese código no pueden patentarse.

Referencias

Anderson, W.P. (1972). *More is Different*. Science, 177, 4047, 393-396

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M. & Rubin, C. (2018). *Learning Certifiably Optimal Rule Lists for Categorical Data*. Journal of Machine Learning Research, 18, 1-78.

Grupo de Trabajo del artículo 29 (2017). *Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679*. wp251, disponible en https://ec.europa.eu/newsroom/article29/items/622227/en [3.10.2022].

Breiman, L., Friedman, H. J., Olshen, A. R. & Stone, J. C. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, *Monterey*, California.

Burrel, J. (2016). *How the machine 'thinks': Understanding opacity in machine learning algorithms*. Big Data & Society, 3, 1-12.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. (2015). *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission*. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 14-16 August, Sydney, Australia.

Crain, M. (2018). *The limits of transparency: data brokers and commodification*. New Media & Society, 20, 1, 88-104.

Crawford, K. (2013). The hidden biases in big data. Harvard Business Review. Available at https://hbr.org/2013/04/the-hidden-biases-in-big-data, accessed 7 July 2022.

Dawes, M. R., Faust, D. & Meehl, E. P. (1989). *Clinical versus actuarial judgment*. Science, 243, 4899, 1668-1674.

Domingos, P. (2012). *A few useful things to know about machine learning*. Communications of the ACM, 55, 10, 78-87.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police and punish the poor. St. Martin's Press, Nueva York.

Unión Europea, Comisión Europea. (2021). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing an EU common list of safe countries of origin for the purposes of Directive 2013/32/EU of the European Parliament and of the Council on common procedures for granting and withdrawing international protection, and amending Directive 2013/32/EU,* COM/2021/206 final. Disponible en https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206 [10.08.2022]

Hand, J. D. (2006). *Classifier technology and the illusion of progress*. Statistical Science, 21, 1, 1-14.

Herm, L., Heinrich, K., Wanner, J. & Janiesch, C. (2021). *Stop ordering machine learning algorithms by their explainability! A user-cantered investigation of performance and explainability*. In proceedings of the 20th Conference on e-Business, e-Services and e-Society (I3E), 1-4 September, Galway, Ireland.

References

Lundberg, M. S. & Lee, S. (2017). *A unified approach to interpreting model predictions*. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 4-9 December, Long Beach, California, USA.

Matteas, M. & Montfort, N. (2005). *A box, darkly: Obfuscation, weird languages, and code aesthetics*. In Proceedings of the 6th Annual Digital Arts and Culture Conference, 1-3 December, Copenhagen, Denmark.

Miller, A. G. (1956). *The magical number seven, plus or minus two: some limits on our capacity for processing information*. Psychological review, 63, 2, 81-97.

Mittelsadt, B., Russell, C. & Wachter, S. (2019). *Explaining explanations in AI*. In Proceedings of the Conference on Fairness, Accountability and Transparency, 29-31 January, Atlanta, Georgia, USA.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*, Broadway Books, Nueva York.

Diario Oficial de la Unión Europea, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Patterson, D., et al. (2020). *The carbon footprint of machine learning training will plateau, then shrink*. TechRxiv, Preprint, Available https://doi.org/10.36227/techrxiv.19139645.v4, accessed 9 October 2022.

Provost, F. & Fawcett, T. (2001). *Robust classification for imprecise environments*. Machine Learning, 42, 203-231.

Ribeiro, T. M., Singh, S. & Guestrin, C. (2016). *Why should I trust you? Explaining the predictions of any classifier*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 August, San Francisco, California, USA.

Ribeiro, T. M., Singh, S. & Guestrin, C. (2018). Anchors: *High-Precision Model-Agnostic Explanations*. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2-7 February, New Orleans, Louisiana, USA.

Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Machine Intelligence, 1, 206-215.

Rudin, C. & Radin, J. (2019). *Why are we using Black box models when we don't need to? A lesson from an explainable AI competition*. Harvard Data Science Review, 1, 2. Available at https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8. [09.09.2022]

References

Sandvig, C. (2014). Seeing the sort: *The aesthetic and industrial defence of the algorithm*. Journal of the New Media Caucus, 3, 3, 1-21.

Seaver, N. (2014). *Knowing Algorithms*. Presented at Media in Transition 8, 3-5 May, Cambridge, Massachusetts, USA. Available at https://digitalsts.net/wp-content/uploads/2019/03/26_Knowing-Algorithms.pdf [07.07.2022]

Schwarz, R., Dodge, J., Smith, A. N. & Etzioni, O. (2020). *Green AI*. Communications of the ACM, 63, 13, 54-63.

Selbst, D. A. & Powles, J. (2017). *Meaningful Information and the Right to Explanation*. International Data Privacy Law, 7, 4, 233–242.

Shin, D. (2021). *The effects of explainability and causability on perception, trust and acceptance: Implications for explainable AI*. International Journal of Human-Computer Studies, 146, 102551.

Silver, D., et al. (2016). *Mastering the game of Go with deep neural networks and tree search*. Nature, 529, 7587, 484-.

Universidad de Montreal. (2018). *Montréal declaration for a responsible development of artificial intelligence* https://www.montrealdeclaration-responsibleai.com/_files/ugd/ebc3a3_5c89e007e0de440097cef36dcd69c7b0.pdf [23.10.2018]

Wanner, J., Herm, L., Heinreich, K. & Janiesch, C. (2020). *A social evaluation of perceived goodness of explainability in machine learning*. Journal of Business Analytics, 5, 1, 29-50.

esade

Santander X Innovation Xperts

www.santander.com/santander-x-innovation-xperts-es

By Santander