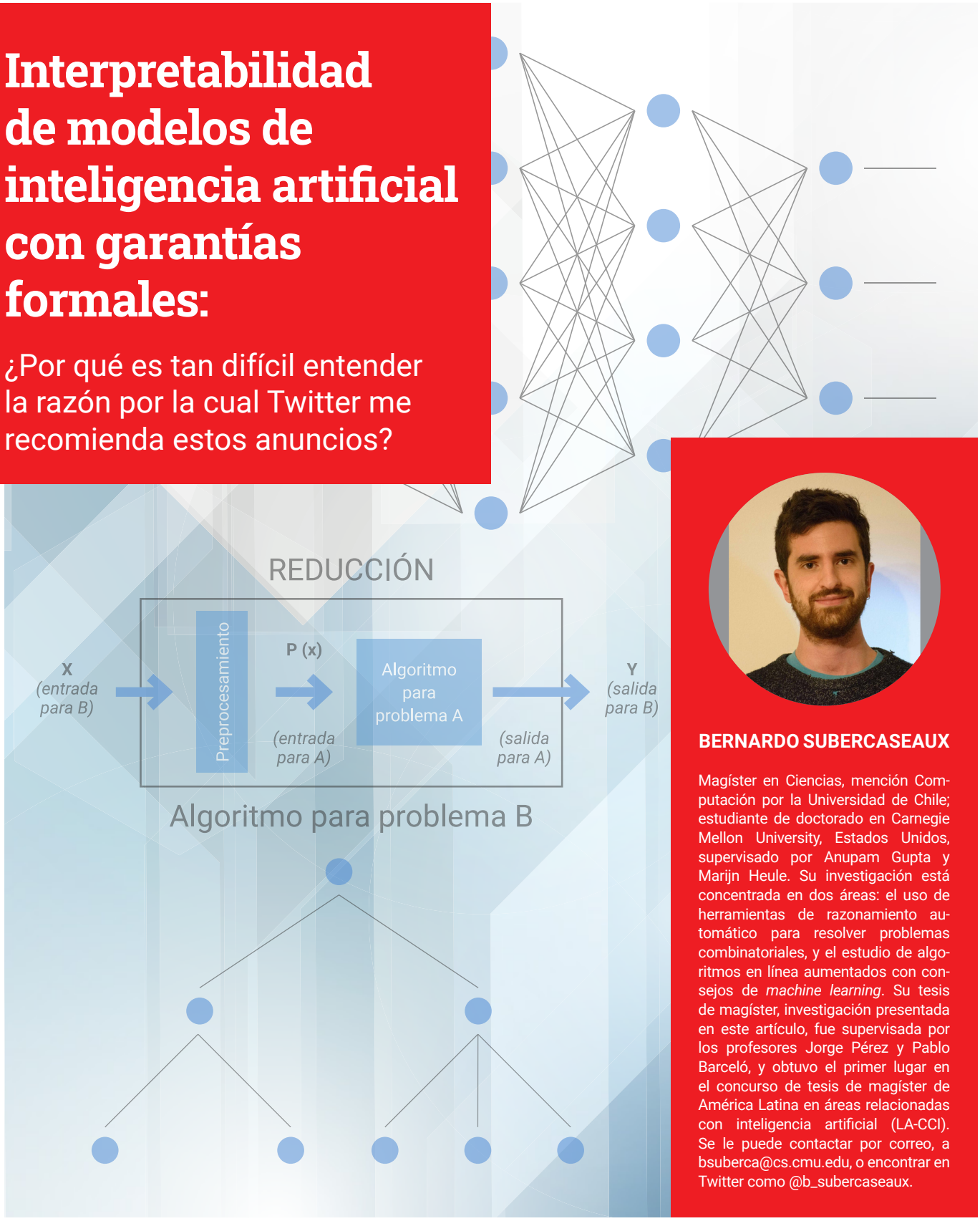
Interpretabilidad de modelos de inteligencia artificial con garantías formales:

¿Por qué es tan difícil entender la razón por la cual Twitter me recomienda estos anuncios?



BERNARDO SUBERCASEAUX

Magíster en Ciencias, mención Computación por la Universidad de Chile; estudiante de doctorado en Carnegie Mellon University, Estados Unidos, supervisado por Anupam Gupta y Marijn Heule. Su investigación está concentrada en dos áreas: el uso de herramientas de razonamiento automático para resolver problemas combinatoriales, y el estudio de algoritmos en línea aumentados con consejos de *machine learning*. Su tesis de magíster, investigación presentada en este artículo, fue supervisada por los profesores Jorge Pérez y Pablo Barceló, y obtuvo el primer lugar en el concurso de tesis de magíster de América Latina en áreas relacionadas con inteligencia artificial (LA-CCI). Se le puede contactar por correo, a bsuberca@cs.cmu.edu, o encontrar en Twitter como @b_subercaseaux.

RESUMEN. Considerando el alto impacto que tienen las decisiones que toman algoritmos de inteligencia artificial en nuestras vidas, se vuelve fundamental que como ciudadanas y ciudadanos podamos confiar en tales algoritmos, y que podamos interpretar o explicar sus decisiones. Sin embargo, un problema importante del área de interpretabilidad en inteligencia artificial, que apunta a resolver estos problemas, es su falta de definiciones precisas sobre conceptos como *interpretabilidad*, *explicación*, etc. En particular, a pesar de aparecer frecuentemente en la literatura, afirmaciones como *"los árboles de decisión son más interpretables que las redes neuronales"* no tienen un significado formal concreto y, por tanto, es difícil construir una ciencia precisa en base a ellas. Mi trabajo de magíster en el DCC justamente consistió en intentar matizar y formalizar ese tipo de afirmaciones, construyendo una noción matemáticamente precisa de interpretabilidad que permite comparar diferentes tipos de modelos de inteligencia artificial en torno a distintos tipos de explicaciones para sus decisiones. Usando técnicas de algoritmos y complejidad computacional, nuestros resultados consisten en una variedad de teoremas que apoyan y fundamentan la sabiduría popular del área.

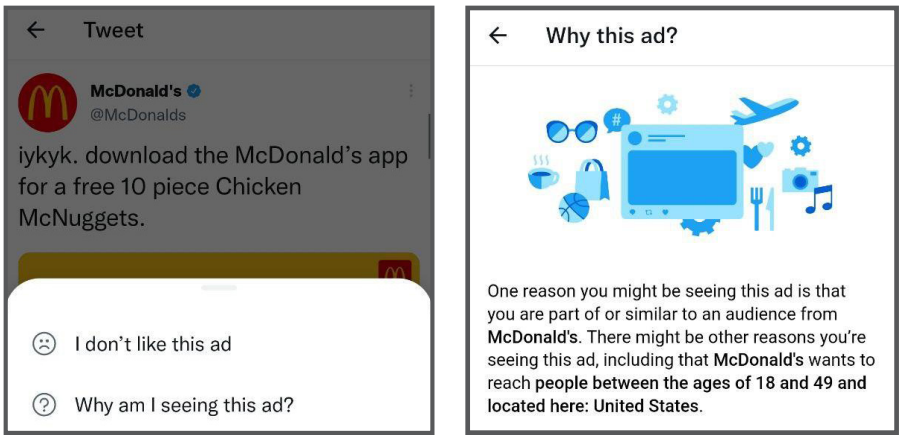


Figura 1. Explicación entregada por Twitter sobre los anuncios mostrados a los usuarios.

¿Qué es la interpretabilidad y por qué importa?

En el mundo moderno prácticamente todas y todos estamos sometidos a decisiones tomadas por modelos de inteligencia artificial: sistemas de recomendación que deciden cuáles anuncios sugerirnos, qué película recomendarnos, o peor aún, cuál de dos noticias contradictorias mostrarnos primero al buscar un asunto político en Internet. Otras aplicaciones tienen influencias aún más directas en nuestra vida práctica: aseguradoras que usan nuestros datos para predecir nuestro riesgo y así ajustar sus tarifas, bancos que utilizan modelos de inteligencia artificial para decidir si otorgar o no un préstamo, y el escalofriante caso en Estados Unidos del sistema penal utilizando un modelo para predecir reincidencia criminal y con ello influir en sentencias.

Debido a su alto impacto en la ciudadanía, el uso actual de modelos de inteligencia artificial levanta una variedad de preocupaciones éticas e ingenieriles. Un concepto de particular importancia al desplegar tecnologías que hacen uso del aprendizaje de máquinas es el de "confianza"; si la ciudadanía va a ser objeto de decisiones tomadas por algoritmos, es fundamental que pueda confiar en ellos y la calidad de sus decisiones. Más aún, esa confianza no tiene por qué ser a ciegas, y en esa línea hay movimientos que abogan por el "derecho a explicaciones" [1], visión en la cual se entiende como un derecho de las personas el poder inspeccionar y recibir explicaciones sobre las decisiones a las cuales están siendo sometidas.

Por ejemplo, al recibir un anuncio en Twitter, es posible preguntar *por qué uno ha recibido tal anuncio*. La respuesta usualmente consiste en algunas características de nuestro perfil a las cuales los anunciantes apuntan como público objetivo. Como puede apreciarse en la Figura 1, la explicación es bastante incompleta e imprecisa; no a toda persona entre 18 y 49 años en Estados Unidos se le presentan anuncios de McDonald's, lo que implica que otros factores están siendo considerados para decidir si presentarnos tales anuncios.

El área de *interpretabilidad* o *explicabilidad* en inteligencia artificial (XAI)[1] tiene por objetivo el diseño de modelos

1 La diferencia entre los conceptos de *interpretabilidad y explicabilidad* es materia de discusión y, más en general, aún no hay consenso sobre sus definiciones [2,3].

Figura 2. Los mapas de saliencia para predicciones contradictorias pueden ser extremadamente similares. Imagen obtenida de [9], quien rinde a su vez cortesía a Chaofan Chen y Checkermallow.

de inteligencia artificial en los cuáles los humanos podamos confiar y utilizar eficazmente [4]. Una variedad de definiciones de interpretabilidad puede reunirse alrededor de la siguiente idea: "*La interpretabilidad tiene que ver con el grado en el cual los humanos podemos entender las causas de una decisión*" [5–7]. Notemos que se trata de una idea bastante vaga; la semántica de las palabras "entender" o "causa" es bastante ambigua. A pesar de su creciente relevancia [8], en particular entendiendo el creciente número de aplicaciones de inteligencia artificial que afectan a la ciudadanía, el área de interpretabilidad o explicabilidad sufre de un fuerte problema definicional: no hay común acuerdo en lo que interpretabilidad o explicabilidad quieren decir exactamente, o cómo deben medirse. Es en este contexto que se enmarcó mi trabajo de magíster, con el objetivo de estudiar a través de formalismos matemáticos la interpretabilidad de diferentes modelos de inteligencia artificial.

Hacia una noción formal de interpretabilidad

Antes de presentar nuestro trabajo, es relevante entender por qué tiene sentido esforzarse en entender la interpretabilidad desde un punto de vista matemático y formal. El famoso artículo *The Mythos of Model Interpretability* [3] ilustra cómo la falta de definiciones claras para conceptos como interpretabilidad es un problema importante en el área; a falta de definiciones claras y métricas consistentes no es posible evaluar adecuadamente el progreso en el área, o comparar los diferentes métodos. Más aún, el artículo *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead* [9], destaca una serie de problemas prácticos con la forma en que los métodos actuales buscan explicar las decisiones que toman modelos opacos como las redes neuronales profundas. Expondré un ejemplo de Rudin. Una tarea clásica en visión computacional es la clasificación de imágenes; detectar, por ejemplo, si una imagen contiene o no un perro siberiano. Imaginemos por un momento que tenemos una red neuronal profunda que ha aprendido a resolver esta tarea con gran precisión. ¿Cómo podemos estar convencidos de que realmente ha aprendido a reconocer perros siberianos? Una hipótesis alternativa, por ejemplo, es que la mayoría de las fotos de perros siberianos con que ha sido entrenada la red ocurren en un paisaje con nieve, y en realidad la red no ha aprendido nada sobre siberianos, sino simplemente que sobre una cierta cantidad de pixeles blancos es conveniente predecir que hay un siberiano en la imagen. Para descartar tales hipótesis, una idea bastante natural es intentar identificar cuáles *pixeles,* o características más en general, están siendo relevantes en una decisión. El problema empieza al definir lo que "relevantes" quiere decir. Varias técnicas de interpretabilidad sobre visión computacional, bajo el nombre de *mapas de saliencia* [8], se basan en computar (o aproximar) el gradiente de la predicción con respecto a las características (i.e., por cada característica, estudiar el efecto que tiene en la predicción el perturbarla según una pequeña cantidad). Es decir, implícitamente definen las características más *relevantes* como aquellas que más efecto tienen localmente en la función de predicción. La Figura 2 ilustra cómo esta idea puede carecer completamente de poder explicativo. Fundamentalmente, el problema es que decir que un conjunto S de características tiene un gradiente de gran magnitud en una localidad del espacio no tiene por qué explicar nada. En otras palabras, si entendemos los mapas de saliencia como una explicación, pareciera que estos requieren a su vez otra explicación para entender

Si la ciudadanía va a ser objeto de decisiones tomadas por algoritmos, es fundamental que pueda confiar en ellos y en la calidad de sus decisiones.

lo que implican. Este problema, de "explicaciones que requieren explicación" se ha comentado repetidas veces en la literatura.

En este contexto de falta de definiciones, y abundancia de técnicas de explicabilidad cuya semántica no es del todo clara, nuestro trabajo busca avanzar la discusión sobre interpretabilidad en una dirección más formal y basada en principios. En particular, es muy común encontrar en la literatura afirmaciones como *"Los árboles de decisión son modelos claramente interpretables, mientras que las redes neuronales profundas son cajas negras no interpretables"*, sin embargo, precisamente por la ausencia de definiciones concretas, no es fácil verificar tales afirmaciones.

Una alternativa posible es llevar a cabo estudios experimentales, en los cuales científicas o científicos de datos evalúan manualmente qué tan interpretables son diferentes modelos en particular, entrenados sobre un conjunto de datos particular. Esta alternativa, si bien útil en la práctica, posee algunos problemas metodológicos. En primer lugar, en ausencia de una definición clara de interpretabilidad, los experimentos suelen definir criterios de interpretabilidad particulares a la situación, los cuales pueden ser cuestionables. Por ejemplo, un experimento puede consistir en presentar a los participantes dos modelos, sus decisiones sobre un conjunto de datos, y luego frente a datos nuevos hacer que los participantes predigan lo que cada modelo decidirá. La idea sería entonces que, si un modelo es más interpretable, será más fácil para el participante humano predecir su comportamiento frente a nuevos datos. De manera implícita, un estudio en esta línea ha definido la interpretabilidad en base a nuestra capacidad de predecir futuras predicciones de un modelo. Por otra parte, una comparación experimental en un momento particular, con un grupo particular de participantes no necesariamente reflejará una verdad subyacente; de la misma forma, no puede decirse de manera definitiva que un algoritmo *A* es más eficiente que un algoritmo *B* solamente en base a un experimento particular, ejecutado en una máquina particular, en base a unas implementaciones particulares.

El aporte principal de nuestro trabajo [10] es presentar una definición formal para establecer cuándo un modelo *A* es más interpretable que un modelo *B*, lo que nos permite establecer afirmaciones como *"para este determinado tipo de explicación, los árboles de decisión son más interpretables que las redes neuronales"* no ya apelando a la intuición humana, sino a través de un teorema matemático. Nuestro punto de partida es flexibilizar la noción de *"A es más interpretable que B"* al parametrizarla según diferentes tipos de explicaciones (que detallaré más adelante); en lugar de decir *"A es más interpretable que B"*, diremos *"A es más interpretable que B con respecto al tipo de explicación E"*. Una ventaja de esta decisión es su flexibilidad frente a los distintos tipos de explicaciones posibles, al no forzar la noción de interpretabilidad a referir a ninguna forma de explicación en particular. Ahora, dado un tipo de explicación *E*, diremos que *"A es más interpretable que B con respecto al tipo de explicación E"* si computar una explicación de tipo *E* sobre el modelo *A* tiene menor complejidad computacional que hacerlo sobre el modelo *B*.

Una brevísima introducción a la complejidad computacional

La complejidad computacional es un formalismo matemático, cuyo desarrollo empezó en los años sesenta, que permite comparar formalmente la dificultad de problemas computacionales. Usualmente, la complejidad computacional de un problema se define en términos del mínimo número de pasos que un computador (i.e., una máquina de Turing) debe realizar para resolver el problema. Determinar tal mínimo número de pasos, incluso asintóticamente, es extremadamente difícil en la práctica; la comunidad matemática no ha podido determinar tal número para prácticamente ningún problema. Por ejemplo, no se sabe cuál es el mínimo número de pasos que se requieren para un problema tan simple como: *dado un arreglo de números enteros, potencialmente negativos, determinar si existen 3 de ellos que suman 0*. No obstante, el área de complejidad computacional es capaz de determinar que un problema *A* es más difícil que otro problema *B*, sin necesariamente saber la complejidad exacta de *A* o de *B*. ¡Esto suena muy útil para nuestro problema de interpretabilidad, porque podremos comparar la interpretabilidad de distintos modelos sin estar forzados a definirla para un modelo aislado! ¿Pero cómo es posible decir que un problema *A* es más difícil que un problema *B* sin saber qué tan difícil es cada uno? La técnica usual se llama *reducción* y consiste en mostrar que, si uno fuese capaz de resolver el problema *A* eficientemente, entonces eso permitiría resolver el problema *B* eficientemente. Por ejemplo, si consideramos *A = dados dos números x e y, computar x+y*; *B = dado un número x, computar 2x,* podremos decir que *A* es más difícil que *B* ya que, de saber resolver *A* eficientemente, puede computarse *B*(x)

El área de interpretabilidad sufre de un fuerte problema definicional; no hay común acuerdo en lo que interpretabilidad quiere decir exactamente, o cómo debe medirse.

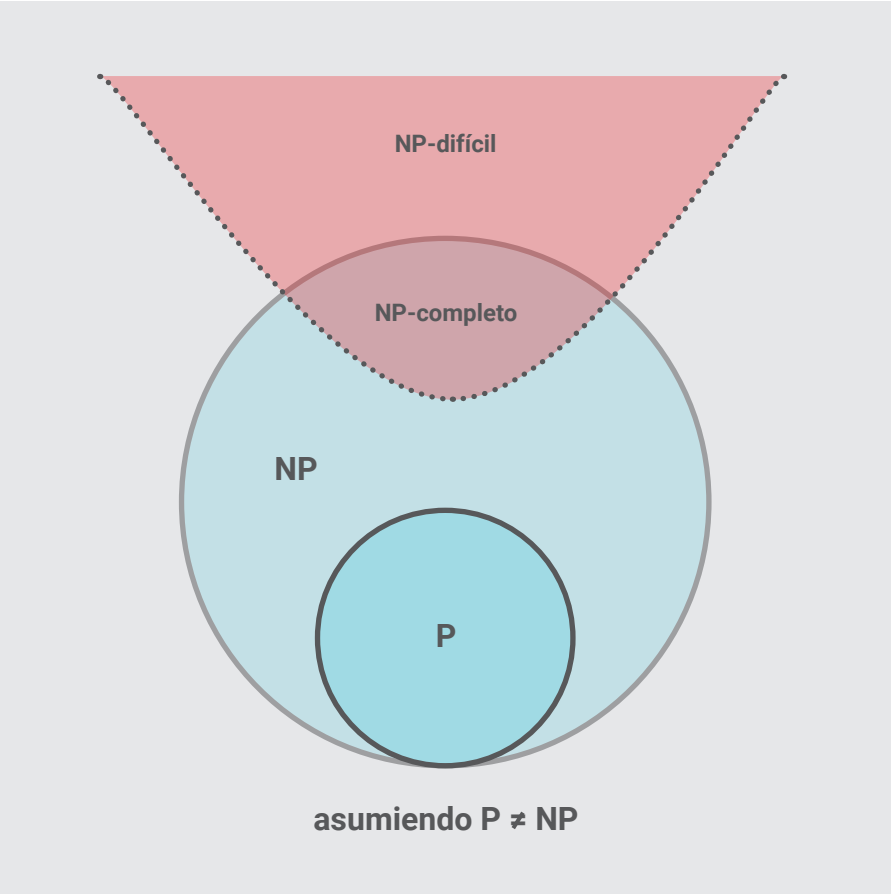


Figura 3. Diagrama de las clases de complejidad P y NP.

simplemente como *A(x, x)*. Por el contrario, no es evidente cómo una máquina que resuelve *B*, es decir, que duplica números, puede usarse para resolver el problema *A*. Para organizar estas comparaciones de dificultad entre problemas, el área de complejidad computacional ha definido una serie de *clases*, las cuáles contienen todos los problemas que no son más difíciles que uno dado (ver Figura 3). Por ejemplo, la clase *NP* puede definirse como la clase de problemas que no son más difíciles que el problema de *satisfacibilidad booleana*, es decir, determinar si una fórmula de la lógica proposicional puede hacerse verdadera bajo alguna asignación de sus variables. Por otra parte, la clase *P* puede definirse como los problemas que no son más difíciles que una restricción particular del problema de satisfacibilidad booleana [11]. Cuando un problema *A* no es más difícil que el problema de satisfacibilidad booleana, decimos que *A pertenece a NP*. Por el contrario, si es igual o más difícil decimos que es *NP-difícil* (i.e., difícil con respecto a *NP*). Finalmente si, un problema es difícil con respecto a una clase a la cual pertenece, se dice que es *completo* para esa clase.

Probablemente el problema abierto más importante de la computación teórica es si acaso las clases *P* y *NP* son iguales. Mientras tanto, se cree fuertemente que *NP* contiene problemas más difíciles que *P*, y por tanto al demostrar que un cierto problema *A* es *NP-difícil*, y que en cambio otro problema *B* está en *P*, se suele considerar que esto es evidencia contundente para decir que *A es más difícil que B*. ¡Usaremos estas ideas para demostrar que interpretar modelos como una red neuronal profunda es más difícil que interpretar un árbol de decisión, bajo distintos tipos de explicaciones!

Tipos de explicaciones

Se presentan tres tipos de explicaciones distintas a través de un ejemplo. Consideremos un banco no particularmente ético que utiliza un modelo de inteligencia artificial para decidir si un postulante debiese o no recibir un determinado préstamo.

Habiendo sido rechazada la postulación presentada en la Figura 4, consideremos los siguientes tipos de explicaciones:

1. El banco no da préstamos a personas con antecedentes penales.
2. Si el postulante hubiese tenido una casa propia, se le habría dado el préstamo.
3. La mayoría de los postulantes casados y sin trabajo estable son rechazados.

A una explicación del primer tipo le llamamos *mínima razón suficiente*, y se formaliza como un subconjunto minimal de las características tal que

El aporte principal de nuestro trabajo es presentar una definición formal para establecer cuándo un modelo A es más interpretable que un modelo B.

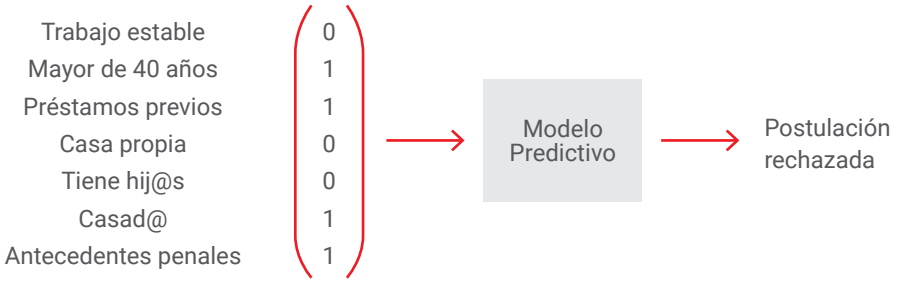


Figura 4. Ilustración simplificada de un modelo predictivo siendo usado por un banco para decidir si otorgar un préstamo a un postulante.

	Á. de Decisión	Perceptrón	MLPs
Mínimo cambio necesario	P	P	NP-completo
Mínima razón suficiente	NP-completo	P	Σ_2^p
Conteo de compleciones	P	#P-completo	#P-completo

Tabla 1. Resumen de resultados de complejidad para distintos tipos de explicaciones sobre distintos tipos de modelos.

cualquier entrada que respete tales características obtendrá el mismo veredicto. En el ejemplo presentado, cualquier instancia posible que tenga un 1 en la componente correspondiente a *antecedentes penales* será rechazada. A una explicación del segundo tipo le llamamos *mínimo cambio necesario*, en cuanto se basa en el mínimo número de características que se deberían cambiar en una instancia para cambiar el veredicto que un modelo en cuestión le da. El tercer tipo, *conteo de compleciones*, es un poco más complejo, y se basa en computar la proporción de instancias que son aceptadas o rechazadas condicionando en algún conjunto de características, como por ejemplo estar casado y no tener trabajo estable. Más aún, en este contexto es posible formalizar la noción de *decisión sesgada*, como aquella en que el veredicto puede cambiar al modificar una característica protegida (edad, género, raza, etc.). Un teorema sencillo de demostrar es que una decisión es sesgada si y solamente si es posible explicarla a través de una mínima razón suficiente que contiene una característica protegida.

Nuestro trabajo considera también otros tipos de explicaciones [12] y está fuertemente inspirado por el trabajo de [13]. Más aún, hemos extendido nuestro análisis a tipos arbitrarios de explicaciones expresables en un lenguaje lógico de primer orden [14].

Resultados

Nuestro estudio se centra en tres clases de modelos: *perceptrones* (o modelos lineales), *árboles de decisión*, y *perceptrones multicapa* (redes neuronales). La Tabla 1 resume nuestros resultados, determinando la complejidad de los diferentes tipos de explicaciones, sobre los diferentes modelos estudiados. Las clases de complejidad *#P* y Σ_2^p tienen definiciones ligeramente más complicadas que aquí omitimos, pero lo relevante es que son clases que se creen incluso más difíciles que *NP*. Es decir, nuestros resultados se pueden resumir diciendo que, para los diferentes tipos de explicaciones considerados, el computarlas sobre perceptrones multicapa es más difícil que sobre árboles de decisión o perceptrones, pero que la comparación entre estos dos últimos depende crucialmente del tipo de explicación considerada.

A continuación presentaré una versión simplificada de las demostraciones más sencillas. Consideremos el problema del *mínimo cambio requerido*. Para demostrar que es *NP*-completo para redes neuronales, usaremos primero un resultado conocido que dice que dada una fórmula booleana F, es posible construir eficientemente una red neuronal M_F que la simula. Esta propiedad no la tienen ni los árboles de decisión ni los modelos lineales. Asumamos ahora que queremos decidir si F es satisfacible, y recordemos que la definición de *NP* nos dice que si demostramos que resolver el problema de *mínimo cambio requerido* sobre redes neuronales permite determinar si F es satisfacible, entonces habremos demostrado que nuestro problema es *NP*-díficil. Asumamos además que F no es satisfecha al asignar todas sus variables a 0, pues de lo contrario la satisfacibilidad de F es trivial. Construyamos ahora la red M_F, y observemos que dada su equivalencia a

Usaremos esta idea [de reducción del área de complejidad computacional] para demostrar que interpretar modelos como una red neuronal profunda es más difícil que interpretar un árbol de decisión, bajo distintos tipos de explicaciones.

F, debe cumplir que $M_F(\bar{0}) = 0$, donde $\bar{0}$ representa el vector de 0s. Ahora, sí el *mínimo cambio requerido* a $\bar{0}$ para cambiar su veredicto en M_F es finito, entonces eso quiere decir que existe una instancia $\bar{x}$ tal que $M_F(\bar{x}) = 1$, y por tanto una asignación de variables asociada a $\bar{x}$ que satisface a *F*. Esto muestra que resolver el problema de *mínimo cambio requerido* permite determinar la satisfacibilidad de *F* y, por tanto es un problema *NP*-difícil. Por el contrario, para los árboles de decisión y los modelos lineales, el problema puede resolverse en tiempo polinomial, mediante programación dinámica y un algoritmo *glotón*, respectivamente. Curiosamente, si a las diferentes características se les asocia un costo de cambio, por ejemplo diciendo que cambiar de "no tener casa propia" a "tener casa propia" cuesta más que de "no tener auto" a "tener auto", entonces el problema se vuelve *NP*-completo para modelos lineales, pero sigue pudiendo resolverse eficientemente sobre árboles de decisión mediante programación dinámica. ■

REFERENCIAS

[1] Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". AI Magazine, 38(3):50–57, 2017.

[2] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018.

[3] Zachary C Lipton. The mythos of model interpretability. Queue, 16(3):31–57, 2018.

[4] David Gunning and David Aha. DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2):44–58, 2019.

[5] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267:1–38, 2019.

[6] Or Biran and Courtenay V. Cotton. Explanation and Justification in Machine Learning: A Survey. 2017.

[7] Finale Doshi-Velez and Been Kim. A Roadmap for a Rigorous Science of Interpretability. CoRR, abs/1702.08608, 2017.

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58:82–115, 2020.

[9] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5):206–215, 2019.

[10] P. Barceló, M. Monet, J. Pérez, and B. Subercaseaux. Model interpretability through the lens of computational complexity. In 34th Conference on Neural Information Processing Systems (NeurIPS), 2020.

[11] Sanjeev Arora and Boaz Barak. Computational complexity: a modern approach. Cambridge University Press, 2009.

[12] B. Subercaseaux. Model Interpretability through the Lens of Computational Complexity. Tesis de Magíster. Universidad de Chile, 2020.

[13] A. Darwiche and A. Hirth. On the reasons behind decisions. In ECAI, pages 712–720, 2020.

[14] M. Arenas, D. Báez, P. Barceló, J. Pérez and B. Subercaseaux. Foundations of Symbolic Languages for Model Interpretability. In 35th Conference on Neural Information Processing Systems (NeurIPS), 2021.