XAI - Análisis Papers

Terminología

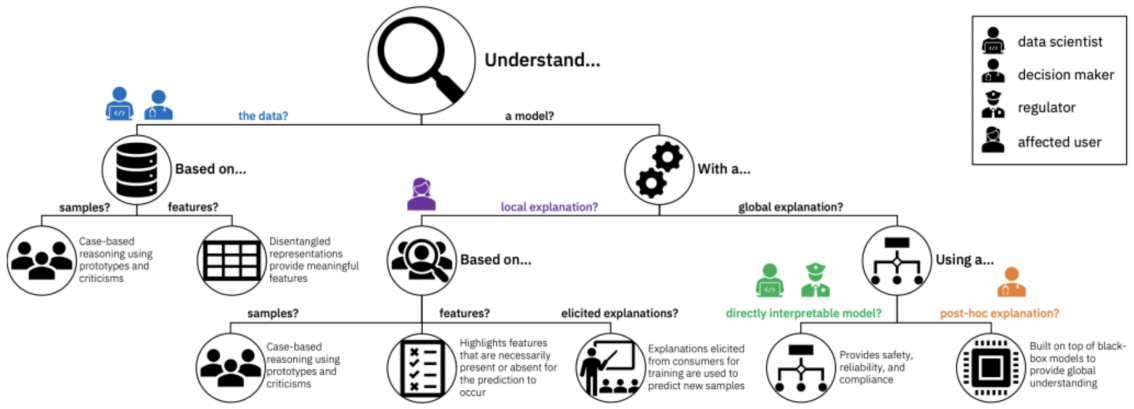
Machine learning: El aprendizaje automático es un conjunto de métodos que utilizan los ordenadores para hacer y mejorar predicciones o comportamientos basados en datos. (Molnar, 2022)
Modelo de caja negra (Black Box Model): es un sistema que no revela sus mecanismos internos. En el aprendizaje automático, la "caja negra" describe modelos que no se puede entender mirando sus parámetros (por ejemplo, una red neuronal)
Machine Learning interpretable: se refiere a los métodos y modelos que hacen que el comportamiento y las predicciones de los sistemas de aprendizaje automático sean comprensibles para los humanos
Inteligencia Artificial Explicable: Definimos el aprendizaje automático interpretable como la extracción de conocimiento relevante de un modelo de aprendizaje automático sobre las relaciones contenidas en los datos o aprendidas por el modelo. (Murdoch et al., 2019)
Para Google, la XAI se define como un conjunto de de herramientas y marco conceptual que puede ser usado para entender, como tus modelos de machine learning toman decisiones.

Que es la inteligencia Artificial Explicable, definiciones:

XAI es un campo de investigación que tiene como objetivo hacer que los resultados de los sistemas de IA sean más comprensibles para los seres humanos. (IBM, 2018)

El término fue acuñado por primera vez en 2004 por Van Lent et al en su trabajo “Un sistema de inteligencia artificial explicable para el comportamiento táctico de unidades pequeñas” (Proc. 16th Conf. Innov. Appl. Artif. Intell., pp. 900-907, 2004), para describir la capacidad de su sistema para explicar el comportamiento de las entidades controladas por IA en la aplicación de juegos de simulación.

Según la definición de IBM “Inteligencia artificial explicable (XAI) es un conjunto de procesos y métodos que permite a los usuarios humanos comprender y confiar en los resultados y la información generados por algoritmos de machine learning. La IA explicable se utiliza para describir un modelo de IA, su impacto esperado y posibles sesgos. Ayuda a caracterizar la precisión del modelo, la imparcialidad, la transparencia y los resultados en la toma de decisiones basada en IA” (IBM, 2018)



Árbol de decisiones 360 de explicabilidad de la IA

Respecto a la distinción del concepto Explicabilidad e interpretabilidad:
Al día de hoy, no hay consenso definitivo respecto a si llamar Inteligencia Artificial Explicable o Interpretable, ya que lo usan indistintamente (Arenas, 2020)

¿Qué es la interpretabilidad?

Definición Interpretabilidad significa explicar o presentar en términos comprensibles.
En el contexto de los sistemas de ML, definimos la Interpretabilidad como la capacidad de explicar o presentar en términos comprensibles para un humano. Una definición formal de explicación sigue siendo difícil de alcanzar; en el campo de la psicología, Lombrozo [2006] afirma que "las explicaciones son... la moneda con la que intercambiamos creencias" y señala que apenas se están empezando a abordar cuestiones como qué constituye una explicación, qué hace que algunas explicaciones sean mejores que otras, cómo se generan las explicaciones y cuándo se buscan las explicaciones.

Interpretabilidad es el término preferido por la comunidad científica para describir las técnicas de IA que son fácilmente comprensibles. Este término se utiliza para describir modelos o algoritmos y no sistemas inteligentes completos (Matthieu Bellucci∗, Nicolas Delestre, Nicolas Malandain, Cecilia Zanni-Merk, 2021)
Adadi et al. el 2018 proponen la siguiente definición: "Un sistema interpretable es un sistema en el que un usuario no sólo puede ver, sino también estudiar y comprender cómo se mapean matemáticamente las entradas a las salidas" (A. Adadi and M. Berrada, 2018)

Podemos concluir que la interpretabilidad se refiere a la capacidad de un objeto para ser entendido y estudiado por un usuario, con un esfuerzo cognitivo razonable. Utilizamos el término objeto en lugar de modelo, porque como mencionó Lipton, la entrada debe ser comprensible al igual que el modelo. Por lo tanto, también podríamos definir un input como

interpretable. La interpretabilidad también engloba diferentes nociones como la descomponibilidad y la simulabilidad.

La literatura muestra evidencia de numerosos estudios sobre la filosofía y las metodologías de la XAI. Sin embargo, hay una evidente escasez de estudios secundarios en relación con los dominios y tareas de la aplicación, por no hablar de revisar los estudios siguiendo las directrices prescritas, que pueden permitir a los investigadores comprender las tendencias actuales en XAI, lo que podría conducir a futuras investigaciones para el desarrollo de métodos específicos de dominios y aplicaciones

Desde Google Cloud, el progreso de la investigación en XAI ha avanzado rápidamente. Ellos proponen la siguiente clasificación:

Clasificación	Métodos	Característica
Input attribution	LIME, Anchors, LOCO, SHAP, DeepLift, Integrated Gradients, XRAI	Te dice, que de tus datos de entrada, cuál es la que mayor influencia ejerce en el resultado
Example influence/matching	MMD Critic, Representer Point Selection, Influence Functions, Attention-Based Prototypical Learning	Se basan en evidencias similares. Esto quiere decir que si a ti te han denegado un crédito, significa que hay otro cliente similar que se le han rechazado, y eso levanta evidencia para rechazar y cumplir con “la regla”
Concept testing/extraction	TCAV, DeepR, Towards Automatic Concept-based Explanations	Simplificar el modelo de deep learning por algo más entendible por un humano (como podrían ser los árboles de decisión o similar), como los modelos sustitutos.
Distillation	Distilling the Knowledge in a Neural Network, Distilling a Neural Network Into a Soft Decision Tree	

Además, seguimos viendo enfoques novedosos de modelos intrínsecamente interpretables y controlables, como las redes reticulares profundas y los modelos bayesianos. (Google Cloud, n.d.)

Modelos Interpretables

Regresión Lineal: La predicción se modela como una suma ponderada de las características. Además, el modelo lineal viene con muchas otras suposiciones. La mala noticia es (bueno, no realmente noticias) que todas esas suposiciones a menudo se violan en la realidad: el resultado, dado que las características, podrían tener una distribución no gaussiana, las características podrían interactuar y la relación entre las características y el resultado podría ser no lineal
Regresión Logística: La regresión logística modela las probabilidades de los problemas de clasificación con dos posibles resultados. Es una extensión del modelo de regresión lineal para problemas de clasificación. Los modelos de regresión lineal y regresión logística fallan en situaciones en las que la relación entre las características y el resultado no es lineal o en las que las características interactúan entre sí
GLM, GAM : tienen problemas por que los supuestos no se cumplen en realidad
Árboles de Decisión: Las predicciones individuales de un árbol de decisiones se pueden explicar descomponiendo la ruta de decisión en un componente por característica. Podemos rastrear una decisión a través del árbol y explicar una predicción por las contribuciones añadidas en cada nodo de decisión
Reglas de Decisión: Una regla de decisión es una simple declaración IF-THEN que consiste en una condición (también llamada antecedente) y una predicción. La utilidad de una regla de decisión suele resumirse en dos números: Soporte y precisión. Al añadir más características a la condición, podemos lograr una mayor precisión, pero perder soporte.

Clasificador Naive Bayes. El clasificador de Naive Bayes utiliza el teorema de probabilidades condicionales de Bayes. Para cada característica, calcula la probabilidad de una clase en función del valor de la entidad. El clasificador Naive Bayes calcula las probabilidades de clase para cada característica de forma independiente, lo que equivale a una suposición fuerte (= ingenua) de independencia condicional de las características. Naive Bayes es un modelo de probabilidad condicional y modela la probabilidad de una clase

KNN: El método k-nearest neighbor se puede utilizar para la regresión y la clasificación y utiliza los vecinos más cercanos de un punto de datos para la predicción. Para la clasificación, el método k-nearest neighbor asigna la clase más común de los vecinos más cercanos de una instancia. Para la regresión, se necesita el promedio del resultado de los vecinos. Las partes difíciles son encontrar la k correcta y decidir cómo medir la distancia entre las instancias, lo que en última instancia define el vecindario. Una forma de reducir tu instancia a las características más importantes, presentar a los vecinos más cercanos puede darte buenas explicaciones (Molnar, 2022)

Tendencias de XAI

“Deja de explicar los modelos de aprendizaje automático de la caja negra para tomar decisiones de alto riesgo y usa modelos interpretables en su lugar” (Rudin, 2019)

- Los modelos de aprendizaje automático de caja negra se están utilizando actualmente para la toma de decisiones de alto riesgo en toda la sociedad, causando problemas en la atención médica, la justicia penal y otros ámbitos. Algunas personas esperan crear métodos para explicar cómo estos modelos de caja negra aliviarían algunos de los problemas, pero tratar de explicar los modelos de caja negra, en lugar de crear modelos que sean interpretables en primer lugar, es probable que perpetúe las malas prácticas y potencialmente pueda causar un gran daño a la sociedad.
- El camino a seguir es diseñar modelos que sean inherentemente interpretables. Esta perspectiva aclara el abismo entre explicar las cajas negras y usar modelos inherentemente interpretables, describe varias razones clave por las que se deben evitar las cajas negras explicables en las decisiones de alto riesgo, identifica los desafíos para el aprendizaje automático interpretable y proporciona varias aplicaciones de ejemplo en las que los modelos interpretables podrían reemplazar potencialmente los modelos de caja negra en la justicia penal, la atención médica y la visión por ordenador

SHAPLEY VALUE

Se podría definir como la contribución marginal promedio de un valor de característica en todas las coaliciones posibles.
El valor de Shapley funciona tanto para la clasificación (si estamos tratando con probabilidades) como para la regresión.

El valor de Shapley se define a través de una función de valor (val) de jugadores en S (coalición).
El valor de Shapley de un valor de característica es su contribución al pago, ponderado y sumado sobre todas las posibles combinaciones de valores de características

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left(val\left(S \cup \{j\}\right) - val(S)\right)$$

donde S es un subconjunto de las características utilizadas en el modelo, x es el vector de valores de características de la instancia a explicar y p el número de características $val_x(S)$ es la predicción de valores de características en el conjunto S que están marginados sobre las características que no están incluidas en el conjunto S

References

A. Adadi and M. Berrada. (2018, Sept 17). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160. 10.1109/ACCESS.2018.2870052

Arenas, J. (2020, December 19). *¿IA Explicable o IA Interpretable?* CEINE. Retrieved April 17, 2022, from https://www.ceine.cl/ia-explicable-o-ia-interpretable/

Doshi, F., & Kim, B. (2017, February 28). *[1702.08608] Towards A Rigorous Science of Interpretable Machine Learning*. arXiv. Retrieved April 17, 2022, from https://arxiv.org/abs/1702.08608

Google Cloud. (n.d.). *AI Explanations Whitepaper*. AI Explanations Whitepaper. Retrieved April 17, 2022, from https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf

IBM. (2018). *IBM XAI*. Retrieved Abril 17, 2022, from https://www.ibm.com/cl-es/watson/explainable-ai

Islam, M.R.; Ahmed, M.U.; Barua, S.; Begum,. (2022, January 22). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences*, *12*(3), 1353. https://doi.org/10.3390/app12031353

Matthieu Bellucci∗, Nicolas Delestre, Nicolas Malandain, Cecilia Zanni-Merk. (2021, Oct). Towards a terminology for a fully contextualized XAI. *Procedia Computer Science*, *2021*(192), 241-250. https://doi.org/10.1016/j.procs.2021.08.025

Molnar, C. (2022). *Interpretable Machine Learning* (Second Edition ed., Vol. 1). Christoph Molnar. https://christophm.github.io/interpretable-ml-book/extend-lm.html

Murdoch, W.J., Singh, S., Kumbier, K., & Abbasi-Asl, R. (2019, Oct 29). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America,*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Rudin, C. (2019, Mayo 13). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *nature machine intelligence*, *1*(206-2015), 206–215. doi.org/10.1038/s42256-019-0048-x