Ética del aprendizaje automático

Martín Rocamora
rocamora@fing.edu.uy

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Fundamentos de Aprendizaje Automático

Ética del aprendizaje automático

- La utilización del apendizaje automático pone de manifiesto retos éticos y sociales sobre los que es importante reflexionar.*

*UNESCO, "Pensar la inteligencia artificial responsable: una guía de deliberación," 2020

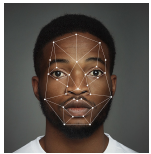
Ética del aprendizaje automático

- La utilización del apendizaje automático pone de manifiesto retos éticos y sociales sobre los que es importante reflexionar.*
- La ética del aprendizaje automático es el conjunto de normas y valores aplicados al desarrollo y a la utilización de estas técnicas.

*UNESCO, "Pensar la inteligencia artificial responsable: una guía de deliberación," 2020

Ética del aprendizaje automático

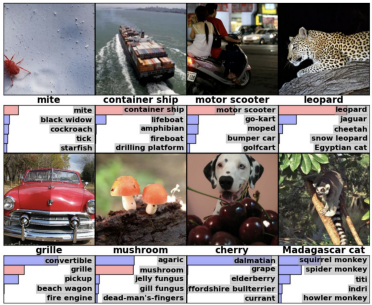
- La utilización del apendizaje automático pone de manifiesto retos éticos y sociales sobre los que es importante reflexionar.*
- La ética del aprendizaje automático es el conjunto de normas y valores aplicados al desarrollo y a la utilización de estas técnicas.
- Actualmente están planteados varios riesgos o desafíos éticos, como por ejemplo la falta de explicabilidad, o el riesgo de discriminación.



*UNESCO, "Pensar la inteligencia artificial responsable: una guía de deliberación," 2020

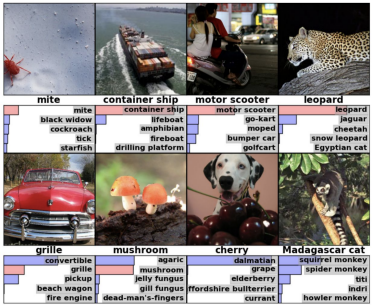
Aprendizaje profundo

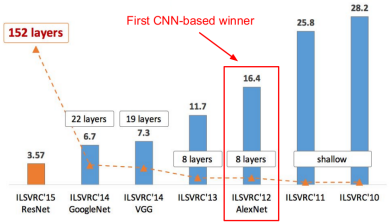
- El aprendizaje profundo produjo avances muy significativos en procesamiento del lenguaje natural, reconocimiento de voz, y visión por computadora (e.g. ImageNet challenge).



Aprendizaje profundo

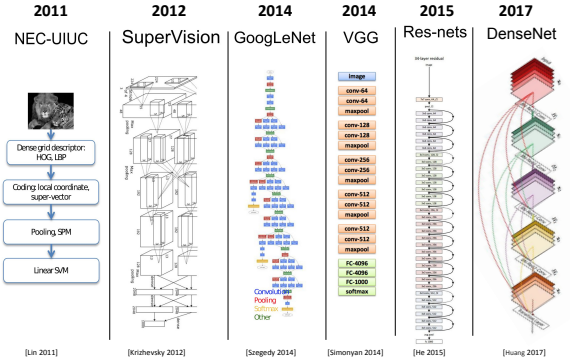
- El aprendizaje profundo produjo avances muy significativos en procesamiento del lenguaje natural, reconocimiento de voz, y visión por computadora (e.g. ImageNet challenge).

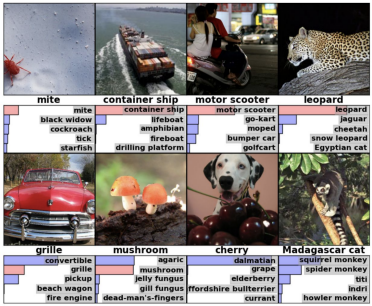


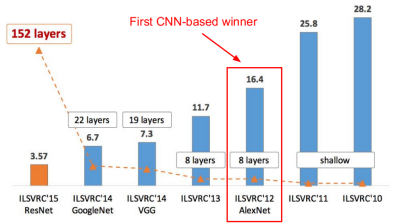


Aprendizaje profundo

- El aprendizaje profundo produjo avances muy significativos en procesamiento del lenguaje natural, reconocimiento de voz, y visión por computadora (e.g. ImageNet challenge).





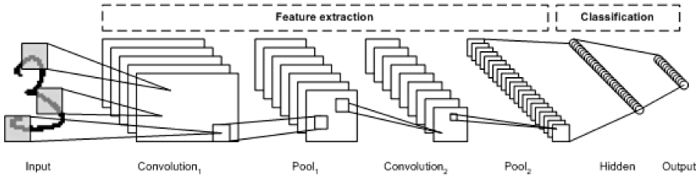


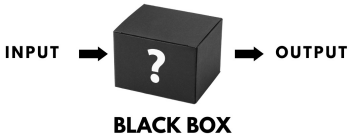
Aprendizaje profundo

- Es difícil entender el proceso de toma de decisiones de las redes neuronales profundas (DNN).

Aprendizaje profundo

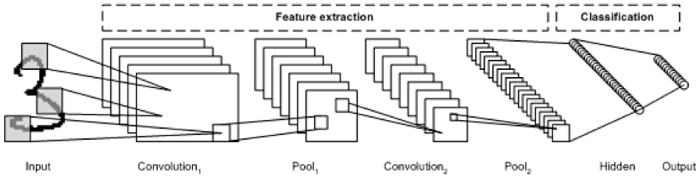
- Es difícil entender el proceso de toma de decisiones de las redes neuronales profundas (DNN).
- Su estructura repetitiva y profunda, con operaciones no lineales, aprende representaciones útiles de los datos, pero también dificulta seguir qué aspectos de la entrada determinan su salida.

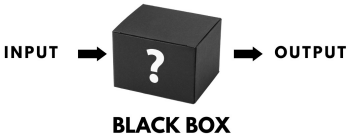




Aprendizaje profundo

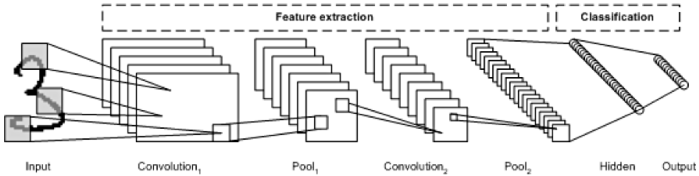
- Es difícil entender el proceso de toma de decisiones de las redes neuronales profundas (DNN).
- Su estructura repetitiva y profunda, con operaciones no lineales, aprende representaciones útiles de los datos, pero también dificulta seguir qué aspectos de la entrada determinan su salida.
- Y también dificulta la extracción del conocimiento capturado por el modelo sobre el problema de una manera que los humanos puedan entender.

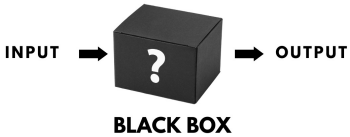




Aprendizaje profundo

- Es difícil entender el proceso de toma de decisiones de las redes neuronales profundas (DNN).
- Su estructura repetitiva y profunda, con operaciones no lineales, aprende representaciones útiles de los datos, pero también dificulta seguir qué aspectos de la entrada determinan su salida.
- Y también dificulta la extracción del conocimiento capturado por el modelo sobre el problema de una manera que los humanos puedan entender.
- Particularmente, en aplicaciones que impactan vidas humanas, como en el cuidado de la salud, la falta de transparencia y de rendición de cuentas puede tener graves consecuencias.





Vulnerabilidad a ataques

- Se puede argumentar que la dificultad para comprender estos modelos no es problemática en muchas aplicaciones.

*C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014

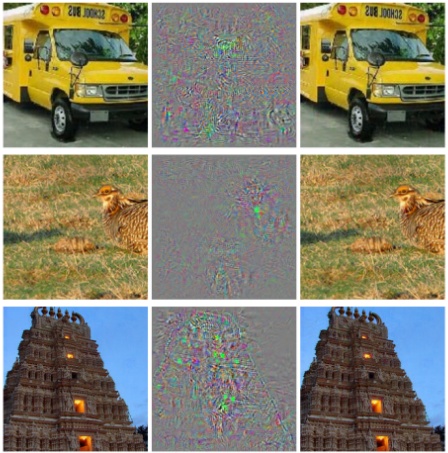
Vulnerabilidad a ataques

- Se puede argumentar que la dificultad para comprender estos modelos no es problemática en muchas aplicaciones.
- Sin embargo, la predicción con alta probabilidad no garantiza que el modelo se comporte siempre como se espera.

*C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014

Vulnerabilidad a ataques

- Se puede argumentar que la dificultad para comprender estos modelos no es problemática en muchas aplicaciones.
- Sin embargo, la predicción con alta probabilidad no garantiza que el modelo se comporte siempre como se espera.
- Se ha demostrado que las DNN pueden ser engañadas con cierta facilidad en tareas de clasificación mediante ataques adversarios.*



*C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014

Sesgo algorítmico

- Errores sistemáticos y repetibles en un sistema informático que crean resultados injustos, como privilegiar una categoría sobre otra de maneras diferentes a la función prevista.

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

Sesgo algorítmico

- Errores sistemáticos y repetibles en un sistema informático que crean resultados injustos, como privilegiar una categoría sobre otra de maneras diferentes a la función prevista.
- El sesgo puede surgir de muchos factores provenientes del diseño del algoritmo o de la forma en que los datos se recopilan, seleccionan, codifican o utilizan para el entrenamiento.

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

Sesgo algorítmico

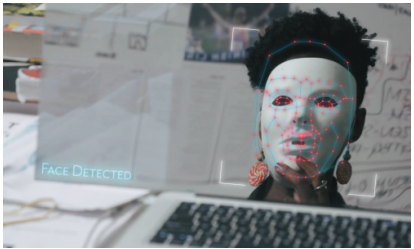
- Errores sistemáticos y repetibles en un sistema informático que crean resultados injustos, como privilegiar una categoría sobre otra de maneras diferentes a la función prevista.
- El sesgo puede surgir de muchos factores provenientes del diseño del algoritmo o de la forma en que los datos se recopilan, seleccionan, codifican o utilizan para el entrenamiento.
- Riesgo de perpetuar los sesgos existentes en la sociedad, reforzando las desigualdades y el tratamiento discriminatorio.*

`Coded biases` - Joy Buolamwini, MIT researcher.

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

Sesgo algorítmico

- Errores sistemáticos y repetibles en un sistema informático que crean resultados injustos, como privilegiar una categoría sobre otra de maneras diferentes a la función prevista.
- El sesgo puede surgir de muchos factores provenientes del diseño del algoritmo o de la forma en que los datos se recopilan, seleccionan, codifican o utilizan para el entrenamiento.
- Riesgo de perpetuar los sesgos existentes en la sociedad, reforzando las desigualdades y el tratamiento discriminatorio.*



`Coded biases` - Joy Buolamwini, MIT researcher.

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

Sesgo algorítmico

Sesgo en modelos generativos texto-a-imagen

confident bus driver

confident network administrator

https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

Sesgo algorítmico

Sesgo en modelos generativos texto-a-imagen

confident bus driver

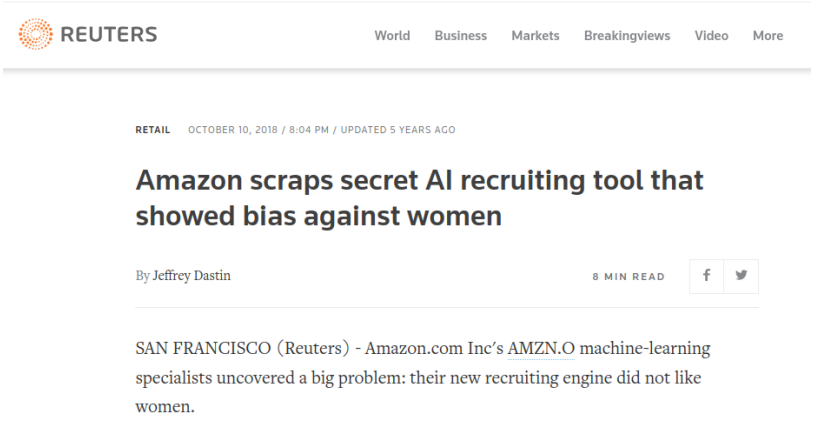
confident network administrator

commited childcare worker

commited nutritionist

https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

Sesgo algorítmico



REUTERS

World Business Markets Breakingviews Video More

RETAIL OCTOBER 10, 2018 / 8:04 PM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

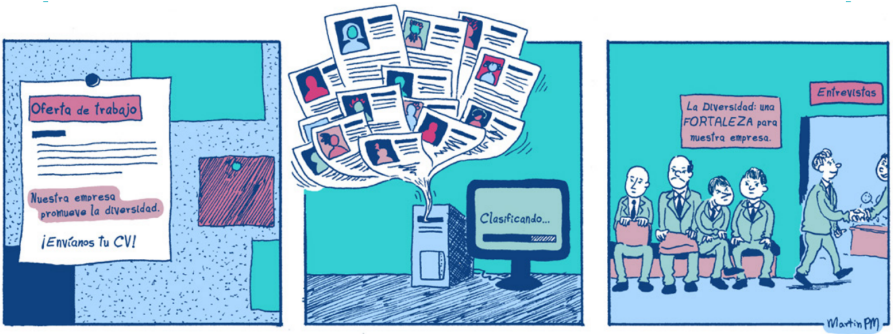
SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Amazon elimina herramienta secreta de reclutamiento de IA que mostraba sesgo contra las mujeres.

Sesgo algorítmico

Actividad: deliberación sobre un caso de uso.

La equidad, diversidad e inclusión en el proceso de contratación.



Buscamos identificar ventajas y desventajas de usar sistemas automáticos en el proceso de contratación.

Sesgo algorítmico

- ¿Una decisión algorítmica de contratación es mejor que una decisión humana, dado que carece de sesgo humano?
- ¿Puede un sistema automático evitar la discriminación?
- En caso de impugnación, ¿cómo debería un(a) empleador(a) justificar una decisión algorítmica de contratación?
- ¿Quién puede considerarse responsable de discriminación en el caso de una contratación automática?
- ¿Por qué delegar las tareas de selección de candidaturas a un sistema de inteligencia artificial de contratación cuando los seres humanos pueden realizarlas correctamente?

Sesgo algorítmico

Se pueden identificar varios aspectos positivos:

- posibilidad de ahorrar costos
- posibilidad de limitar el parcialismo humano
- eficiencia para manejar grandes cantidades de datos

Sesgo algorítmico

Se pueden identificar varios aspectos positivos:

- posibilidad de ahorrar costos
- posibilidad de limitar el parcialismo humano
- eficiencia para manejar grandes cantidades de datos

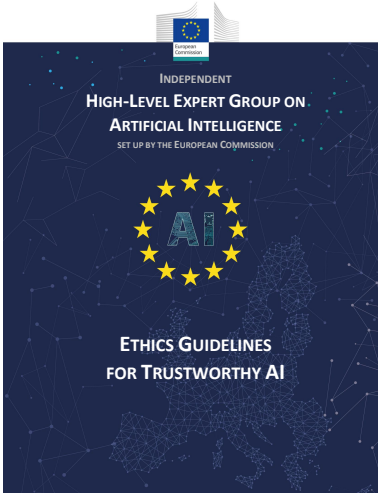
Sin embargo, hay cuestiones importantes a tener en cuenta:

- posibilidad de sesgo algorítmico (e.g. sesgo en los datos)
- justificación de decisiones algorítmicas (explicabilidad)
- sustitución de los recursos humanos

Regulación y recomendaciones

- la noción legal de un derecho a la explicación en la regulación general de protección de datos (GDPR) de la Unión Europea.*
- se distinguen diferentes escenarios de riesgo
- aspectos éticos, legales, sociales, económicos y culturales (ELSEC) asociados al desarrollo de una IA confiable (Trustworthy AI)

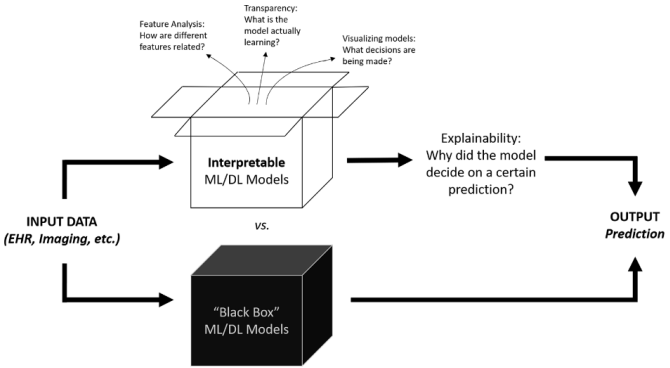
European Commission President, Ursula von der Leyen, Apr. 2021



*The European Parliament and The Council of the European Union, "Regulation (EU) 2016/679. General Data Protection Regulation," 2016

IA explicable o interpretable

- Modelos de aprendizaje automático que brindan explicaciones de sus decisiones *
- Los modelos que se pueden interpretar pueden depurarse y auditarse mejor, ya sea para diseñar métodos de defensa, prever fallos de funcionamiento, descubrir casos extremos o detectar sesgos.



*C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2020

IA explicable o interpretable

- Se utilizan varios términos, como interpretabilidad, explicabilidad, inteligibilidad, comprensibilidad y transparencia, para referirse a la capacidad de comprender cómo funciona el modelo.
- Es una noción de que depende del dominio específico y, por lo tanto, no existe una definición única de qué atributos hacen que los modelos sean interpretables.
- La interpretabilidad a menudo se materializa en la práctica en un conjunto de restricciones del modelo, dictadas por el conocimiento del dominio, como causalidad, monotonía o esparcidad.
- La interpretabilidad se puede lograr en diferentes grados, desde un modelo completamente transparente a uno levemente restringido.

IA explicable o interpretable

- Métodos que resaltan las características de la entrada que más influyen en la salida (post-hoc).

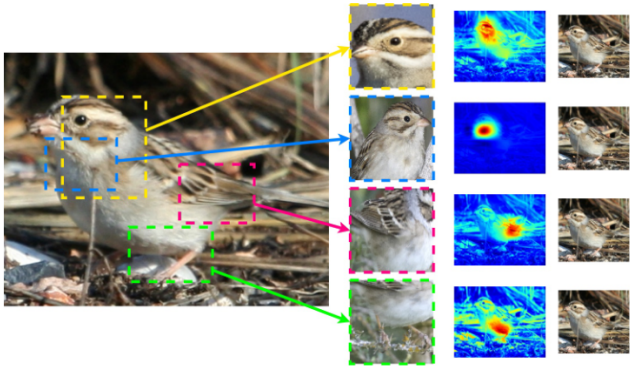
IA explicable o interpretable

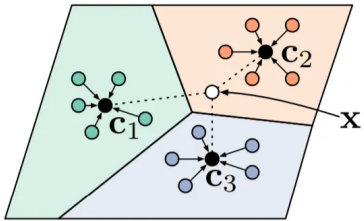
- Métodos que resaltan las características de la entrada que más influyen en la salida (post-hoc).

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

IA explicable o interpretable

- Arquitecturas de redes neuronales interpretables a través de prototipos (interpretable-by-design).*
- Los prototipos se aprenden durante el entrenamiento y las predicciones de la red se basan en la similitud de la entrada con ellos (las explicaciones son los prototipos y distancias correspondientes).





*C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019

References

[1] UNESCO, "Pensar la inteligencia artificial responsable: una guía de deliberación," 2020.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[3] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[4] The European Parliament and The Council of the European Union, "Regulation (EU) 2016/679. General Data Protection Regulation," 2016.

[5] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2020.

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.