Sesgo algorítmico

PID_00278566

Marc Juárez Miró

Universitat Oberta de Catalunya

Marc Juárez Miró

Investigador posdoctoral en la Universidad del Sur de California, donde estudia problemas de justicia algorítmica. Obtuvo su doctorado en 2019 por la Universidad de Lovaina, con una tesis sobre privacidad y técnicas de análisis del tráfico de redes.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Cristina Pérez Solà

Primera edición: febrero 2021

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Marc Juárez Miró
Producción: FUOC

Índice

Introducción

El paradigma del *big data* ha tenido un efecto transversal en la economía de las sociedades de la era de la información. En prácticamente todos los sectores se han adoptado técnicas para extraer información de los grandes volúmenes de datos que se generan, con el objetivo de optimizar procesos y tomar mejores decisiones. Este paradigma se ha visto reforzado por los adelantos en el campo del aprendizaje automático, al que ha proveído de herramientas para automatizar la extracción de conocimiento de los datos. Mientras que este paradigma ha sido un motor de progreso para resolver algunas tareas concretas, la automatización de decisiones en ámbitos como la justicia, la educación y las finanzas abre interrogantes sobre la ética de delegar ciertas decisiones a algoritmos de aprendizaje automático. Además, a pesar de que estos algoritmos hayan conseguido optimizar algunos procesos con éxito, todavía no entendemos las implicaciones que pueden tener en una sociedad que los aplica sistemáticamente para tomar decisiones.

Cabe decir que hemos sido testigos de grandes progresos en el campo del aprendizaje automático y, especialmente, de la revolución que han traído las redes neuronales profundas en el campo de la visión por computador. Por ejemplo, estas nuevas técnicas se han usado para construir sistemas de detección de tumores cancerígenos que superan en precisión a los expertos humanos. Sin embargo, un optimismo desmesurado, alimentado por estos resultados positivos, ha motivado la aplicación de estos algoritmos en problemas de todos los ámbitos de la sociedad. Actualmente se están usando algoritmos de aprendizaje automático para la selección de personal, el acceso a universidades, para la adjudicación de seguros y créditos bancarios e, incluso, para conceder la libertad condicional a presos. Estas decisiones pueden ser críticas por el desarrollo y el estatus de los individuos que están sujetos a ellas, puesto que determinan las oportunidades a las que tienen acceso y, por lo tanto, sus perspectivas de futuro. Es por eso que con la popularización de los algoritmos de aprendizaje automático, hay una creciente preocupación por las consecuencias que puede tener en el reparto de la riqueza, la movilidad social de los individuos y el aumento de las desigualdades existentes en la sociedad. El campo del sesgo algorítmico ha nacido para estudiar estos aspectos desde un punto de vista científico, ético y legal.

Uno de los argumentos que se da para justificar el uso de algoritmos para tomar decisiones es que los algoritmos tienen el potencial de ser neutrales, puesto que están exentos de los sesgos y estereotipos que los humanos adquirimos, a veces sin ser conscientes de ellos. Este argumento no es válido por varias razones. En primer lugar, los algoritmos están aplicados por expertos

humanos que pueden introducir sesgos en varios puntos del proceso de diseñar, implementar y ejecutar los algoritmos. En segundo lugar, aunque estos expertos lleven a cabo estas tareas de manera completamente imparcial, los datos que se proporcionan a los algoritmos también pueden contener sesgos. Las desigualdades sociales se reflejan en los datos y, sin ninguna intervención que lo prevenga, los algoritmos heredan estos sesgos.

El problema del sesgo en los datos es más complejo de lo que puede parecer a primera vista. Debido a la elevada complejidad de los algoritmos de aprendizaje automático, es difícil anticiparse a los sesgos que los algoritmos aprenderán de los datos. De hecho, las redes neuronales profundas extraen patrones no lineales de gran complejidad que hacen difícil entender y razonar los resultados. Esta falta de transparencia en las decisiones entra en conflicto con la ley, en la que se estipula que los sujetos de las decisiones tienen derecho a saber las razones que han motivado el resultado de las decisiones. En particular, en la Unión Europea, el Reglamento General de Protección de Datos también regula los algoritmos de aprendizaje automático que operan con datos personales y deja bien claro que los sujetos pueden pedir explicaciones sobre las decisiones que les afectan. Desgraciadamente, la mayor precisión de las redes neuronales profundas respecto a otros algoritmos ha hecho que estas no hayan parado de ganar en popularidad.

En este módulo hacemos un repaso del naciente campo científico que estudia el sesgo social de los algoritmos para la toma de decisiones. En el primer apartado definimos los conceptos básicos sobre los algoritmos de aprendizaje automático y damos ejemplos de los sesgos que estos algoritmos pueden tener. En el segundo apartado haremos un repaso sobre definiciones formales de equidad que se han propuesto para los algoritmos y de los resultados de imposibilidad. En el tercero hablamos sobre métodos para corregir los sesgos del algoritmo. Por último, hablaremos de métodos para explicar e interpretar las decisiones de algoritmos complejos.

Objetivos

Los objetivos que el estudiante tiene que lograr al finalizar el módulo son los siguientes:

1. Conocer las principales fuentes de sesgo en los modelos de aprendizaje supervisado.
2. Comprender las definiciones formales de equidad, las relaciones entre ellas y sus limitaciones.
3. Conocer los métodos de corrección del sesgo y poder relacionarlos con las definiciones de equidad que permiten lograr.
4. Distinguir entre explicabilidad e interpretabilidad y conocer algunos de los métodos que existen para explicar modelos complejos.

1. Automatización de las decisiones

En este apartado introducimos el lenguaje necesario para hablar de la toma de decisiones automatizada a partir de observaciones. Tendremos presentes dos casos de uso:

1) Aceptar o rechazar la solicitud de un candidato para un puesto de trabajo o una universidad.

2) Predecir el riesgo de conceder un seguro o un préstamo bancario.

En esencia, el objetivo de la entidad que toma las decisiones es, dada una observación X, determinar el valor de una variable Y (el resultado de la decisión), que más lo beneficia. En el ejemplo de la selección de personal, X son las calificaciones y habilidades del candidato, e Y codifica si se contrata al candidato o no. Las decisiones serán acertadas si se rechazan candidatos que no son capaces de ejercer y se aceptan candidatos con alto rendimiento.

Idealmente, la entidad que toma la decisión desea tener una función f que envía valores de la variable X a valores de la variable Y de forma que, al aplicar la función en una observación que no conoce, f devuelve el valor de la decisión. En el supuesto de que Y pueda tomar valores de un conjunto finito –tal como en el ejemplo de selección de personal anterior– f se denomina un *clasificador* y los valores de Y son las clases. En cambio, si Y es una variable continua, f se denomina *regresor* y la tarea, *regresión*. Por ejemplo, en el caso de predecir el riesgo de conceder un crédito, Y podría tomar valores en $[0,1]$ y la imagen de f se podría interpretar como las probabilidades de devolver el crédito. Cabe decir que a pesar de que se utilice un regresor, a menudo la decisión final será discreta (por ejemplo, el crédito se dará si la probabilidad de pago es superior a 0.8). De hecho, la mayoría de los clasificadores no devuelven un valor discreto, sino un valor de confianza que «se sesga» del mismo modo que cuando se discretiza la salida de un regresor.

El aprendizaje supervisado es la rama del aprendizaje automático que tiene como objetivo encontrar funciones f a partir de conjuntos de observaciones. El aprendizaje supervisado es el método predominante para resolver este tipo de problemas en la práctica. En el siguiente subapartado hacemos un pequeño repaso a los conceptos básicos de aprendizaje supervisado que son necesarios para seguir el resto del módulo.

1.1 Aprendizaje supervisado

Los algoritmos de aprendizaje supervisado parten de un conjunto de observaciones (x_1,y_1), ..., (x_n,y_n), conocido como *conjunto de entrenamiento*, donde cada x_i se denomina *instancia* o *ejemplo* e y_i se denomina *etiqueta*, si f es un clasificador, o *respuesta*, si es un regresor. Una instancia es típicamente un vector de características que describen la observación. Por ejemplo, en la solicitud de crédito, algunas características relevantes son: si el solicitante tiene un trabajo estable, su nómina, pruebas de solvencia y avales.

El marco teórico de los algoritmos de aprendizaje automático supone que X e Y son variables aleatorias y que los ejemplos del conjunto de entrenamiento se han obtenido tomando muestras independientes de la distribución conjunta $X \times Y$. Entonces, dada una nueva observación de $X = x$ para la cual no conocemos la y correspondiente, el f óptimo es aquel que devuelve la y que maximiza la siguiente probabilidad: $P(X = x \mid Y = y)$. El reto es que en la práctica no conocemos esta distribución y, por lo tanto, tenemos que recurrir a métodos estadísticos que nos permitan modelar la distribución a partir del conjunto de entrenamiento. Es por eso que tanto si f es un clasificador como si es un regresor, diremos que f es un *modelo*. Tal como X e Y se pueden considerar variables aleatorias, también se pueden interpretar las salidas del modelo como una variable aleatoria: $\hat{Y} = f(X)$. En un abuso de lenguaje, a veces nos referiremos a $\hat{Y}$ como el modelo.

El paradigma más usado para encontrar un f adecuado trata de encarar el problema como un problema de optimización, en que se selecciona f dentro de una clase de funciones, llamado *espacio de hipótesis*, que minimiza el error que f incurre en el conjunto de entrenamiento. Este error se calcula según una cierta función de error que depende de la familia de algoritmos de aprendizaje supervisado que se elija. Sin embargo, estas funciones de error se diseñan para minimizar el error en el conjunto de entrenamiento y, además, permitir que f generalice a instancias que no se han observado previamente. Típicamente, el espacio de hipótesis está parametrizado y la optimización se hace sobre estos parámetros. Este proceso de optimización se denomina *entrenamiento* y es por eso que el conjunto de observaciones en el que se realiza la optimización se denomina *conjunto de entrenamiento*. En este módulo omitiremos los detalles técnicos de este proceso puesto que no son relevantes para las explicaciones que haremos sobre equidad algorítmica.

Minimización del riesgo empírico

El marco teórico al que nos referimos en este texto se denomina *minimización del riesgo empírico*, puesto que tiene como objetivo minimizar una función de error en el conjunto de entrenamiento.

Lectura complementaria

Podéis encontrar más información en cualquier libro de texto sobre teoría de aprendizaje automático. Por ejemplo:
V. Vapnik (2000). *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer-Verlag.

1.2 El error en la población

Como hemos explicado en el subapartado anterior, el proceso de entrenamiento devuelve un modelo f minimizando el error en el conjunto de entrenamiento e intentando que generalice la población. Ahora bien, antes de llevar los modelos a la práctica, necesitamos medir cuán satisfactorio ha sido

el proceso de entrenamiento. En otras palabras, queremos medir el error que f comete en el conjunto de entrenamiento y, aún más importante, cuál es el error que comete en observaciones que no se encuentran en el conjunto de entrenamiento –es decir, queremos saber si f consigue generalizar de las muestras a la población.

Uno de los métodos más populares para medir este error es utilizar métodos de validación cruzada (en inglés: *cross-validation*). La validación cruzada es un método estadístico que no hace suposiciones sobre la distribución de los datos. La validación cruzada consiste en apartar un subconjunto de los datos antes de trabajar con ellos, denominado el conjunto de testeo. Una vez el modelo se ha entrenado, usamos el conjunto de testeo para medir el error del modelo en muestras que no ha visto previamente y, así, estimar el error del modelo en la población.

Para medir el error del modelo en el conjunto de testeo, se usan diferentes métricas dependiendo del tipo de error del clasificador que nos interese evaluar. Aquí consideraremos solo métricas que miden el error de clasificadores binarios. Los clasificadores binarios solo tienen dos clases disjuntas, i.e., Y toma valores en $\{0,1\}$. Denominamos 1 a la clase positiva y 0 al complemento de esta, la clase negativa. Pensad en estas clases como los resultados de una decisión binaria: el resultado es o bien aceptar o bien rechazar, respectivamente. Entonces, el clasificador puede equivocarse de dos maneras posibles: falsos negativos (FN), instancias de la clase positiva que se clasifican como negativas; o falsos positivos (FP), instancias negativas que se han clasificado como positivas. Del mismo modo, hay dos tipos de predicciones correctas: instancias positivas y negativas que se han clasificado en la clase correcta, es decir, positivos verdaderos (TP, del inglés *True Positives*), y negativos verdaderos (TN, del inglés *True Negatives*), respectivamente. El total de instancias es: $Total = TP + FN + FP + TN$.

Las métricas más utilizadas para medir el error de un clasificador binario son:

- La ratio de falsos positivos (FPR) es la proporción de errores en la clase negativa:

$$\text{FPR} = \frac{FP}{FP + TN}.$$

- La ratio de verdaderos positivos (TPR), también conocido como *exactitud*, es la proporción de instancias de la clase positiva que se han clasificado correctamente:

$$\text{TPR} = \frac{TP}{TP + FN}.$$

- La precisión o PPV es la ratio de clasificaciones correctas de entre todas las instancias que se han clasificado como positivas:

$$\text{PPV} = \frac{TP}{TP + FP}\,.$$

- El NPV es la ratio de negativos verdaderos de entre todas las instancias que se han clasificado como negativas:

$$\text{NPV} = \frac{TN}{TN + FN}\,.$$

No se tienen que confundir el PPV y el TPR. Observad que en el denominador en el caso del PPV tenemos las decisiones que *se han clasificado* como positivas y no las que realmente lo son.

1.3 Ejemplos de las causas de discriminación

A continuación os presentamos algunos ejemplos que Hardt da sobre las fuentes de sesgo en aprendizaje supervisado.

Lectura complementaria

M. Hardt (2014). *How big data is unfair*. Medium.

Ejemplo 1: los datos son un espejo de la sociedad

El conjunto de entrenamiento es una compilación de muestras que refleja los sesgos de nuestra sociedad. Si, por ejemplo, uno de los grupos de personas presente en el conjunto de entrenamiento es un grupo que históricamente ha sido víctima de una discriminación sistemática (por ejemplo, la población negra en los EE.UU. o la población femenina en todo el mundo), los datos presentarán características diferenciadoras para estos grupos (como por ejemplo un nivel de educación más bajo en la población negra o una diferencia salarial entre sexos). A la hora de tomar decisiones, los algoritmos descubrirán estos patrones y posiblemente tomarán decisiones diferentes para estos grupos. Imaginemos por ejemplo el clasificador de selección de personal entrenado en datos donde un grupo de la población tiene calificaciones más bajas de media. El clasificador puede determinar que la pertenencia a este grupo es indicativo de un nivel educativo más bajo y por lo tanto de un rendimiento más bajo para llevar a cabo una tarea.

Significados de *sesgo*

Es importante no confundir diferentes significados de la palabra *sesgo*. En este texto nos referimos mayoritariamente a un sesgo *social* y no al sesgo en el sentido estadístico.

No es suficiente, sin embargo, eliminar las características de las instancias que identifican al grupo (por ejemplo, sexo, raza, etnia, orientación sexual, etc.). El problema es que otras características que son relevantes para la tarea pueden estar fuertemente correlacionadas con estos atributos de la identidad de una persona. Por ejemplo, es común que las minorías se encuentren segregadas en barrios y la dirección es una característica importante a la hora de pedir un seguro. El sexo también se encuentra codificado en cualquier conjunto de características que sea suficientemente rico (por ejemplo, la altura y el peso), aunque no esté presente explícitamente. Estas características se encuentran latentes y no se puede culpar al algoritmo supervisado por descubrirlas; al fin y al cabo, el propósito de estos algoritmos es descubrir patrones no triviales en los datos.

Está claro que cuando un atributo de la identidad de los individuos no es importante para la tarea, este atributo no tendría que representar ningún papel en las decisiones. Por ejemplo, para la selección de personal para un puesto de trabajo como desarrollador de software, el sexo es completamente irrelevante y lo que cuentan son las calificaciones de los candidatos. Ahora bien, se puede justificar legalmente que para algunos puestos de trabajo se requiera

Lectura complementaria

S. Barocas; A. D. Selbst (2016). «Big data's disparate impact». *Calif. L. Rev.*, vol. 104, HeinOnline.

una cierta altura mínima, lo cual favorece a los candidatos de sexo masculino porque suelen ser más altos de media. Determinar cuando es aceptable hacer estas distinciones es a menudo el objeto de discusión en procesos judiciales por discriminación laboral.

Ejemplo 2: la disparidad en el número de muestras

Incluso cuando los datos no tienen ninguna correlación con atributos protegidos (por ejemplo, sexo, orientación sexual, raza, creencias religiosas, etc.), los algoritmos de aprendizaje supervisado pueden tener sesgos en el tamaño y la representatividad de las muestras de los grupos. Uno de los principios del aprendizaje supervisado es que se puede reducir el error de los modelos cuanto mayor y más representativo sea el conjunto de entrenamiento que se utiliza para entrenarlos. Cuando los datos no tienen un balance entre los grupos que los componen, este principio se traduce en un error desigual entre grupos.

Este sesgo afecta especialmente a las minorías. Esto es porque, por definición, es más difícil obtener muestras de las minorías –puesto que están menos representadas en la sociedad. Como consecuencia, los modelos entrenados con datos donde las minorías están mal representadas (por ejemplo, no se recoge el número suficiente de muestras) presentarán un error más elevado que para la mayoría.

Veamos la figura 1 para ilustrar este efecto.

Figura 1

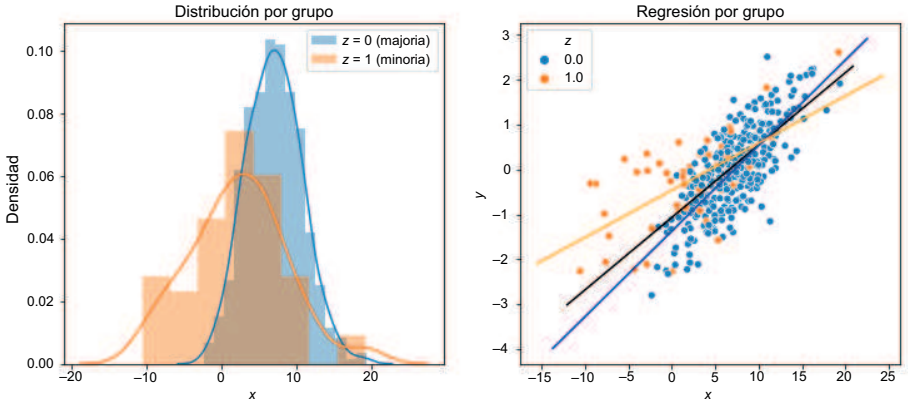


Figura 1

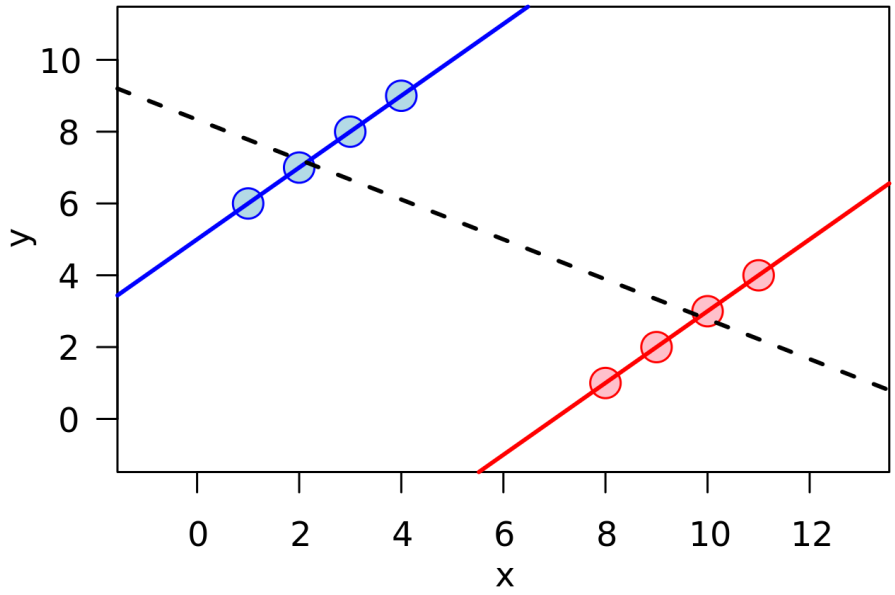
Hemos generado datos sintéticos para dos poblaciones (una mayoría y una minoría) donde los datos de la minoría representan una fracción pequeña del total. Como vemos en la figura de la izquierda, las poblaciones tienen distribuciones diferentes. En la figura de la derecha hemos ajustado una línea de regresión para cada uno de los subgrupos. La línea negra es la regresión para la población, y la naranja y la azul son las líneas de regresión para la minoría y la mayoría, respectivamente. Como vemos, la línea de la mayoría tiene un error menor respecto a la de la población (esto se puede apreciar por la desviación de la línea del subgrupo respecto a la línea de la población).

Ejemplo 3: los patrones dependen de diferencias culturales

Los problemas anteriores pueden agravarse por diferencias culturales entre los grupos que conforman los datos. Hardt da el ejemplo de un clasificador diseñado para detectar nombres de usuario reales (por ejemplo, Facebook pide el uso del nombre real de la persona durante la creación de la cuenta). En algunos grupos étnicos la diversidad de nombres es mucho mayor que en culturas occidentales, en las que hay poca varianza entre nombres. En culturas occidentales también predominan nombres cortos y, en otros, los nombres son de media más largos. La unicidad y una mayor longitud del nombre son características que un clasificador con un sesgo en el entrenamiento podría detectar erróneamente como características identificativas de nombres falsos. Así, no se trata solo de la disparidad en el tamaño de los conjuntos de entrenamiento, sino también de las diferencias entre los patrones identificativos entre los grupos.

De hecho, un fenómeno estadístico que ha tomado relevancia con la discriminación de los algoritmos de aprendizaje supervisado demuestra que dos variables que están correlacionadas positivamente en la población pueden tener una correlación negativa en los subgrupos de la población(!). Esta paradoja se conoce como paradoja de Simpson y está ejemplificada en la figura 2. En la figura se representan los valores de las variables X e Y para dos grupos (rojo y azul). Cuando medimos la correlación de los grupos por separado hay una correlación lineal positiva, mientras que en la población la correlación es negativa.

Figura 2. Crédito: Schutz, con licencia de dominio público



Un ejemplo famoso de datos donde se da la paradoja de Simpson es en la adjudicación de plazas en la Universidad de UC Berkeley, California, del año 1973. En los datos agregados se observó que el porcentaje de solicitantes de sexo masculino aceptados fue más elevado que los de sexo femenino. Sin embargo, cuando se dividían los datos por facultad, se invertía el sesgo respecto al sexo de los solicitantes. De hecho, en las estadísticas cogidas por cada facultad se observa un sesgo estadísticamente significativo en favor de los solicitantes de sexo femenino.

Lectura complementaria

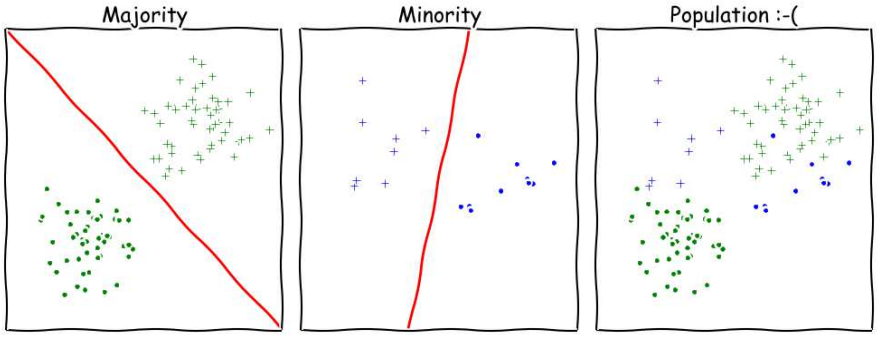
P. J. Bickel y otros (1975). «Sex bias in graduate admissions: Data from Berkeley», *Science*, vol. 187, pág. 398-404.

Para resolver este problema, se puede entrenar un clasificador para cada uno de los subgrupos. Esto, sin embargo, requiere que el clasificador distinga entre los subgrupos y que por lo tanto utilice los atributos protegidos explícitamente, lo cual está prohibido en general por la mayoría de leyes antidiscriminación.

Además, aunque distinguimos por subgrupo, definir los subgrupos es un reto en sí mismo puesto que estos grupos son construcciones sociales: ¿cómo definimos una raza o una etnia? ¿Cómo definimos los grupos por orientación sexual dentro del amplio espectro de las definiciones de la comunidad LGBTQ+? ¿Y cómo definimos los subgrupos derivados de las intersecciones entre estos grupos? ¿Necesitamos un clasificador por cada posible intersección?

Un problema que nos podemos encontrar si no hacemos la distinción entre subgrupos, sin embargo, es que a pesar de que haya clasificadores simples para cada uno de los subgrupos, el clasificador que resuelve la tarea cuando los grupos se combinan puede ser complejo. Hardt ejemplifica como, a pesar de que los subgrupos se pueden separar con funciones lineales (f es una función lineal), la población no se puede separar linealmente (véase la figura 3).

Figura 3. Modelos para la mayoría y la minoría considerados independientemente y en conjunto. Los modelos (f) están representados en rojo. Crédito: Moritz Hardt

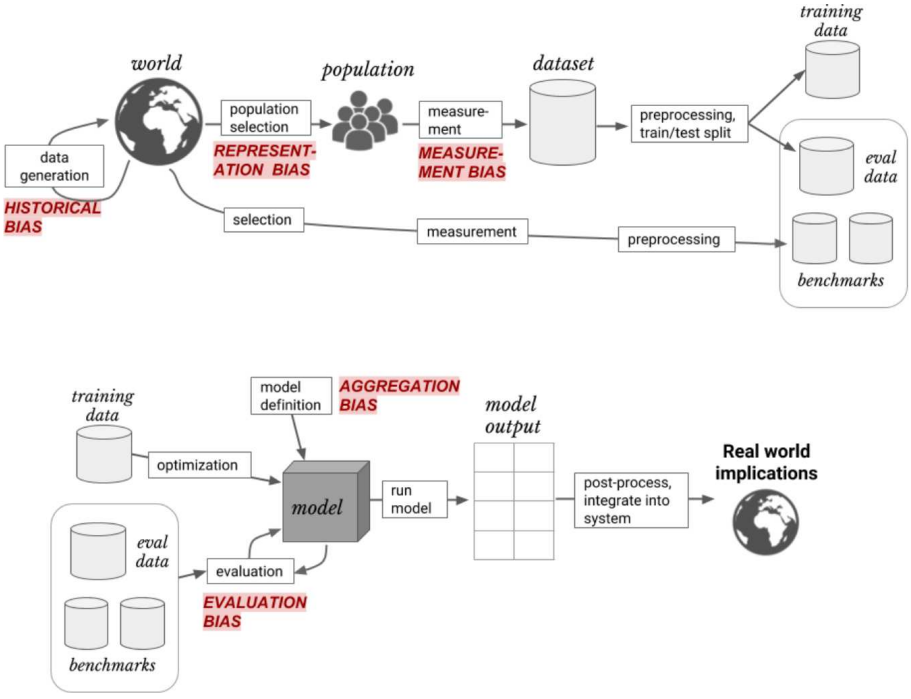


Este apartado no pretende ser una recopilación exhaustiva de todas las fuentes de sesgo en un modelo de aprendizaje supervisado. De hecho, se puede introducir sesgo en cada fase del proceso, desde el inicio hasta el final. Por ejemplo, se puede introducir sesgo en un error dispar en las mediciones durante la recogida de los datos o incluso en la evaluación del modelo (véase la figura 4).

Lectura complementaria

H. Suresh (2019). *The Problem with «Biased Data»*. Medium.

Figura 4. Posibles sesgos en diferentes partes del proceso de aprendizaje supervisado. Crédito: Harini Suresh



2. Definiciones de sesgo algorítmico

Uno de los problemas que más atención ha recibido por parte de la comunidad académica es la formalización matemática de nociones de equidad de los algoritmos. En la literatura se han propuesto una multitud de definiciones sobre qué es que un modelo sea «justo». En los siguientes subapartados veremos las definiciones más extendidas y algunos de los resultados fundamentales a los que se ha llegado.

2.1 Equidad de grupos

Lectura complementaria

S. Barocas; M. Hardt; A. Narayanan (2020). *Fairness and machine learning*. fairmlbook.org.

Para las definiciones de equidad de grupos, o equidad grupal, que damos en este módulo nos hemos basado en el libro de texto de Barocas, Hardt y Narayanan.

Primero, introducimos notación para definir las tres categorías principales de equidad algorítmica. Recordemos que Y es la variable que f intenta predecir, $\hat{Y} = f(X)$ es la variable aleatoria de las predicciones y denominamos A al *atributo protegido* (por ejemplo, sexo, etnia, orientación sexual), que puede estar o no incluido en las características de las instancias (X) con las que se ha entrenado el modelo. En este subapartado nos centraremos en atributos protegidos binarios y denominaremos grupo privilegiado y grupo desfavorecido a los grupos definidos por A que tienen una ventaja o una desventaja, respectivamente, en las decisiones. Además, por simplicidad, nos centraremos en modelos que son clasificadores binarios, pero es fácil extrapolar las definiciones que damos a regresores.

A pesar de que existen decenas de definiciones de equidad grupal diferentes, la mayoría se pueden caracterizar en tres tipos de relación de dependencia entre $\hat{Y}$, Y y A: independencia, suficiencia y separación. Muchas de las definiciones que encontramos en la literatura son, de hecho, una relajación de estas condiciones de dependencia.

2.1.1 Independencia

La condición de independencia es que el atributo protegido sea independiente de las salidas del modelo:

Definición (*independencia*): f disfruta de independencia si las variables aleatorias $\hat{Y}$ y A son independientes, i.e., $\hat{Y} \perp A$.

La condición de independencia recibe varios nombres en la literatura sobre equidad algorítmica: *paridad demográfica, paridad estadística, equidad entre grupos*, entre otros. En el caso de clasificación binaria y considerando solo dos grupos, la condición de independencia se puede expresar en términos probabilísticos de la siguiente forma:

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1),$$

es decir, la probabilidad de aceptar un individuo es independiente del valor de su atributo protegido. Y, en cuanto a las predicciones de error, estas probabilidades se pueden medir en términos de las frecuencias de *TP*/*Total* por cada uno de los grupos.

Esta es una definición simple y, además, es natural: las decisiones del modelo no tendrían que depender de atributos protegidos como el sexo o la raza. A pesar de todo, pedir que se cumpla independencia en la práctica puede tener efectos que no son deseables. Como Dwork y otros ilustran, dados dos grupos de interés, una empresa puede satisfacer esta condición contratando al mismo número de personas para cada grupo pero con un proceso de selección diferente: mientras que los candidatos del grupo privilegiado se seleccionan diligentemente, los del grupo desfavorecido se seleccionan aleatoriamente. Como resultado, las estadísticas mostrarán una diferencia del rendimiento entre los dos grupos. Entonces, la empresa puede justificar diferencias en los porcentajes de selección entre grupos para proteger los intereses del negocio.

Esto puede pasar sin que la empresa sea malintencionada. Puede ser que, por ejemplo, a causa de que la empresa tradicionalmente ha contratado a personas del grupo privilegiado, conocen mejor el grupo y saben cuáles son sus características relevantes y, por lo tanto, pueden hacer un proceso de selección más cuidadoso. Este punto es importante porque a menudo sucede que los grupos tienen características diferentes respecto a la tarea, si no, no hay ninguna justificación para no usar el mismo proceso de selección y seleccionar el mismo número de candidatos de los dos grupos.

Para resolver este problema se ha propuesto la condición de separación.

2.1.2 Separación

La condición de separación dice que $\hat{Y}$ y A son independientes en la medida que Y lo permita:

Definición (*separación*): f satisface la condición de separación si las variables aleatorias $\hat{Y}$ y A son condicionalmente independientes respecto a Y, i.e., $\hat{Y} \perp A \mid Y$.

Es decir, el modelo y el atributo protegido pueden tener una relación de dependencia en cuanto que lo justifique la tarea. Esta definición reconoce que hay muchas tareas en las que el atributo protegido está correlacionado con la tarea. Por ejemplo, en el ejemplo de conceder un crédito bancario, puede ser que uno de los grupos tenga una probabilidad más alta de no poder devolver el crédito. Un banco puede justificar legalmente, en este caso, que hacer una distinción entre los grupos es una necesidad para la sostenibilidad del negocio.

En el supuesto de que f sea un clasificador binario, esta condición se puede expresar en términos probabilísticos de la siguiente manera:

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1),$$
$$P(\hat{Y} = 1 \mid Y = 0, A = 0) = P(\hat{Y} = 1 \mid Y = 0, A = 1).$$

Es decir, la probabilidad de un positivo verdadero (primera ecuación) o de un falso positivo (segunda ecuación) es independiente del grupo. Se puede medir si estas condiciones se satisfacen comprobando que el TPR y el FPR de los grupos son iguales. En la práctica, que el TPR y el FPR sean iguales quiere decir que se aceptan candidatos *que se tendrían* que aceptar porque están calificados y que se descartan candidatos que se tendrían que aceptar en la misma medida en ambos grupos.

La condición de separación también se conoce popularmente en la literatura como *Equalized Odds* (igualdad de posibilidades u oportunidades).

2.1.3 Suficiencia

Por último, el tercer criterio de equidad exige que la tarea y el atributo protegido sean independientes respecto a la decisión:

Definición (*suficiencia*): f satisface la propiedad de suficiencia si las variables aleatorias Y y A son condicionalmente independientes respecto a $\hat{Y}$, i.e., $Y \perp A \mid \hat{Y}$.

Una forma de interpretar esta condición es que para una decisión fijada (por ejemplo, aceptar un candidato), la tarea es independiente del grupo: en cada grupo encontramos el mismo número de personas que tendrían que ser aceptadas. Para un clasificador binario, esta condición se puede escribir en términos probabilísticos de la siguiente manera:

$$P(Y = 1 \mid \hat{Y} = 1, A = 1) = P(Y = 1 \mid \hat{Y} = 1, A = 0).$$
$$P(Y = 0 \mid \hat{Y} = 0, A = 1) = P(Y = 0 \mid \hat{Y} = 0, A = 0).$$

Estas probabilidades se pueden medir mediante el PPV y el NPV del clasificador, respectivamente.

En la literatura, a veces se hace referencia a la *calibración por grupos* para hablar de suficiencia. Esto se debe a que en aprendizaje supervisado «calibrar» quiere decir ajustar el modelo para que dé estimaciones precisas de las probabilidades de sus salidas o decisiones. Se ha demostrado que hacer esta calibración por cada grupo implica suficiencia –lo recíproco, en general, no es cierto puesto que la suficiencia es una condición más débil que la calibración por grupos.

La diferencia entre suficiencia y separación es sutil y puede ser difícil de distinguir. Una manera de diferenciar estas condiciones es pensar en la diferencia entre TPR y PPV: mientras que TPR es el número de aciertos sobre los que *se tendrían* que aceptar (TP + FN), PPV es el número de aciertos sobre los que el clasificador *ha* aceptado (TP + FP). En el ejemplo de selección de personal, mientras que la suficiencia impone que se acepten el mismo porcentaje de candidatos aptos sobre el total de candidatos aptos, la separación impone que el porcentaje de candidatos aptos sobre el total de candidatos que se han aceptado sean los mismos entre grupos. Como veremos en el siguiente ejemplo, la segunda condición depende de lo fácil o difícil que sea encontrar candidatos aptos en cada grupo.

Ejemplo: un alto TPR no implica un alto PPV

Por ejemplo, imaginemos un escenario hipotético donde 5 de cada 1000 candidatos del grupo desfavorecido son aptos. Un clasificador que presenta un alto TPR «caza» 4 de estos 5 candidatos, i.e., $TPR = 80\ \%$. Esto no tiene en cuenta, sin embargo, cuántos candidatos no aptos han aceptado (FP). Ahora bien, cuando miramos el número de FP que el clasificador ha cometido, como hay muchos más no aptos que aptos –995 de cada 1000 no son aptos–, aunque la probabilidad de cometer un FP sea baja, en números absolutos el número de FP puede ser órdenes de magnitud mayor que el de TP. Como consecuencia, el PPV será mucho más bajo que el TPR, aunque el TPR sea elevado. Si en el ejemplo también suponemos que el clasificador tiene un bajo FPR, por ejemplo, $FPR = \frac{20}{995} \simeq 2\ \%$, el número de *FP* es 20, y por lo tanto, $PPV = \frac{4}{4+20} \simeq 16\ \%$.

El TPR y el PPV son solo un par de las muchas métricas que existen para medir el error de un modelo; hay muchas más y se podrían proponer definiciones de equidad para cada una de ellas. El problema es que a pesar de que todas estas definiciones son deseables –idealmente, ninguno de los grupos tendría que sufrir un error mayor que otro– se ha demostrado que ni siquiera las que hemos dado se pueden satisfacer a la vez.

2.1.4 Teoremas de imposibilidad

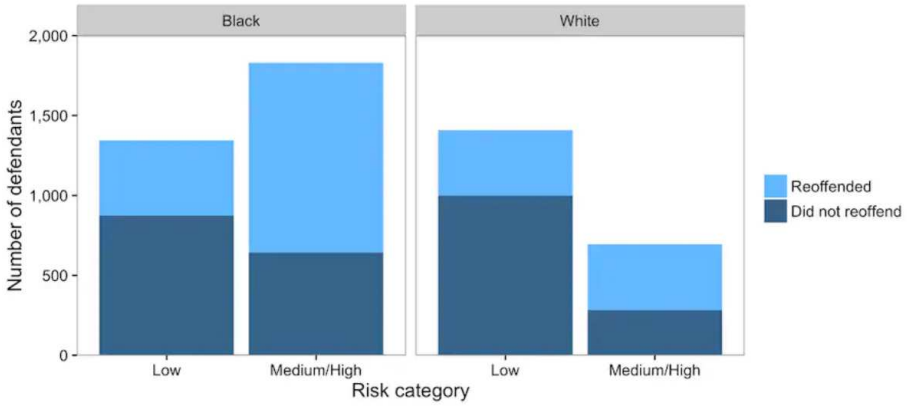
Los resultados de imposibilidad estuvieron fuertemente motivados por el caso de COMPAS (las siglas en inglés de *Correctional Offender Management Profiling for Alternative Sanctions*). COMPAS es un modelo entrenado con técnicas de aprendizaje supervisado utilizado en algunos condados de Estados Unidos. Actualmente, se usa para predecir la probabilidad de reincidencia de un preso y como evidencia para los jueces que conceden solicitudes de libertad condicional.

Ejemplo: el caso de COMPAS

El año 2016, ProPublica, una redacción de periodismo de investigación, pidió acceso a las decisiones de COMPAS hechas a Broward County, Florida –en este estado, esta información está disponible si se pide–, y la complementaron con el registro criminal actualizado de aquellas personas que habían sido sujetas a una decisión. Con estos datos, ProPublica hizo una auditoría independiente de COMPAS con el objetivo de determinar si las decisiones tenían un sesgo respecto a la etnia, el sexo o la raza de los presos. Los resultados de la investigación muestran un sesgo significativo contra los americanos de ascendencia africana. En particular, la investigación demostraba que miembros de esta población tenían el doble de probabilidades de ser identificados «de alto riesgo» cuando no acababan reincidiendo y, viceversa, los presos blancos tenían más probabilidades de ser identificados de bajo riesgo y acabar cometiendo otro crimen.

Equivant, entonces conocida como Nothpointe, la empresa que diseñó y produjo COMPAS, no tardó en responder a las acusaciones de ProPublica, defendiendo la equidad de las decisiones de COMPAS. El informe de Equivant proporciona evidencia estadística que demuestra que COMPAS predice reincidencia correctamente en la misma proporción entre todos los grupos.

Figura 5. Distribución de los presos por categorías de riesgo y raza en el conjunto de datos de COMPAS analizado por ProPublica



Fuente: *The Washington Post*

Este debate público recibió mucha atención de los medios de comunicación y de los académicos debido a sus implicaciones sobre la parcialidad del sistema judicial y penitenciario. Sorprendentemente, se demostró que ambas partes tenían razón. Simplemente, cada parte se acogía a una definición de equidad diferente: mientras que ProPublica estaba auditando la condición de separación en COMPAS, Equivant había diseñado COMPAS para garantizar la condición de suficiencia. En términos de las ratios de error, ProPublica había demostrado que los FPR eran diferentes entre los grupos y Equivant había *calibrado* COMPAS para que el PPV fuera igual entre grupos. Las dos perspectivas se pueden apreciar en la figura 5. La lectura que ProPublica hace de la figura es que COMPAS identifica a personas de raza negra que no han reincidido como alto riesgo con más frecuencia (diferencia del área de color azul oscuro en *medium/high* entre grupos). A pesar de esto, cuando fijamos un valor de $\hat{Y}$ (*low* o *medium/high*), obtenemos que las proporciones de reincidentes son las mismas para cada grupo y, por lo tanto, se satisface la condición de suficiencia que defiende Equivant.

Lectura complementaria

J. Angwin; J. Larson; S. Mattu; L. Kirchner, (2016). *Machine Bias*. ProPublica.

Además, se ha demostrado matemáticamente que *todas* las definiciones de equidad de los subapartados anteriores son mutuamente exclusivas: ninguna pareja de definiciones se puede satisfacer a la vez –excepto en casos muy concretos y poco realistas. En este subapartado no veremos las demostraciones de imposibilidad para todas las parejas de condiciones, lo veremos para las más sencillas y para otras daremos la intuición que hay detrás de su incompatibilidad.

Incompatibilidad entre independencia y suficiencia

Veremos que independencia y suficiencia no se pueden satisfacer a la vez si el atributo protegido es independiente de la tarea. Este será un caso común, puesto que si estamos considerando casos de equidad es porque los grupos no son homogéneos hacia la tarea y hay ciertas diferencias que deben tenerse en cuenta.

Por ejemplo, en el caso de COMPAS, está claro que los grupos no están distribuidos del mismo modo respecto a la tarea. La población negra ha sido históricamente discriminada y todavía sufre un trato diferencial en muchos ámbitos de la sociedad. Hay estudios que demuestran que la población negra está sometida a más arrestos injustificados que los blancos. El racismo sistémico y otros factores relacionados como la segregación y la pobreza han resultado en un mayor número de presos negros en Estados Unidos.

Entonces, para demostrar la incompatibilidad entre independencia y suficiencia hacemos una reducción al absurdo. Primero suponemos que se cumplen ambas y llegamos a una contradicción con la suposición de que Y y A no son independientes. Esto es por las propiedades de contracción y descomposición de la independencia condicional: $A \perp \hat{Y}$ (independencia) y $A \perp Y \mid \hat{Y}$ (suficiencia) implica que A es independiente de la variable conjunta $(Y,\hat{Y})$ y, por lo tanto, $A \perp Y$ (contradicción).

Independencia

Cuidado con confundir la independencia probabilística entre las variables aleatorias y la condición de equidad que también denominamos de independencia.

Incompatibilidad entre separación e independencia

La demostración en este caso es más complicada y, además de $A \not\perp Y$, también requiere la suposición que $\hat{Y}$ sea independiente de Y. Esta suposición no es una suposición muy fuerte, puesto que cualquier clasificador útil debería tener una correlación con la tarea (si no, no sirve para resolver la tarea).

No haremos la demostración aquí, pero consiste en demostrar el contrapositivo: si $A \perp \hat{Y}$ y $A \perp \hat{Y} \mid Y \Rightarrow$ o bien $A \perp Y$, o bien $\hat{Y} \perp Y$.

Incompatibilidad entre suficiencia y separación

Esta es la incompatibilidad entre las dos definiciones de equidad en el debate de COMPAS. Kleinberg y otros y Chouldechova demostraron independientemente y al mismo tiempo que es imposible construir un clasificador que cumpla con las dos definiciones de equidad, a no ser que los grupos tengan las probabilidades *a priori* o los FPR y los FNR sean cero (un clasificador perfecto es muy raro en la práctica). Por lo tanto, aquí también suponemos que $A \not\perp Y$.

La demostración de la incompatibilidad entre separación y suficiencia se basa en el teorema 17.2 de Wasserman, que enuncia:

Lectura complementaria

L. Wasserman (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

$$A \perp \hat{Y} \mid Y \text{ i } A \perp Y \mid \hat{Y} \Rightarrow A \perp (\hat{Y}, Y)$$

y, además,

$$A \perp (\hat{Y}, Y) \Rightarrow A \perp \hat{Y} \text{ i } A \perp Y.$$

La demostración consiste en tomar el contrapositivo de las implicaciones anteriores:

$$A \not\perp \hat{Y} \text{ o } A \not\perp Y \Rightarrow A \not\perp \hat{Y} \mid Y \text{ o } A \not\perp Y \mid \hat{Y}.$$

Volviendo al caso de COMPAS, puesto que no se pueden satisfacer las dos condiciones de equidad porque la distribución de presos por grupos no es uniforme, una pregunta a hacernos es: ¿qué definición de equidad es más importante que se cumpla en este caso? Equivant se decanta por la suficiencia, puesto que como el riesgo significa lo mismo independientemente de la raza, un juez no tiene que considerar la raza del preso a la hora de interpretar los resultados de COMPAS. Por otro lado, ProPublica argumenta que es muy grave que no se satisfaga separación puesto que a pesar de que aquellos presos negros que fueron identificados de alto riesgo no acabaron cometiendo crímenes, sí fueron sujetos a un mayor escrutinio por parte del sistema penitenciario, lo cual no es justo en comparación a los blancos que no reincidieron. Quizás una pregunta todavía más relevante: ¿creéis que es ético utilizar COMPAS si no se pueden satisfacer las dos propiedades? ¿Creéis que es ético utilizar COMPAS incluso si se satisfacen estas propiedades?

2.2 Equidad individual

La noción de equidad individual recuerda las máximas tradicionales de la justicia: «casos parecidos deberían ser tratados de forma parecida».

A diferencia de las nociones de equidad grupal anteriores, la equidad individual se define normalmente sobre las distribuciones de probabilidad de f, es decir, no la decisión final de cada instancia, sino la distribución de probabilidad de las posibles decisiones. En el ejemplo de dar un crédito, las probabilidades de salida de f podrían ser (0.7,0.3), es decir, con probabilidad 0.7 se concede el crédito y con 0.3 se rechaza la solicitud. Denotamos por $\hat{f}$ el clasificador que devuelve las distribuciones de probabilidad de f.

Definición (*equidad individual*): sea D una medida de divergencia entre distribuciones de probabilidad de las salidas de f y sea d una medida de distancia entre individuos. Decimos que f es individualmente equitativo si, para toda pareja de individuos u y v, satisface: $D(\hat{f}(u),\hat{f}(v)) \leq d(u,v)$.

Es decir, las diferencias entre dos individuos limitan las diferencias entre los tratamientos que reciben por f. Es importante que estas medidas de distancia entre individuos solo tengan en cuenta características relevantes para la tarea o aplicación en cuestión y que no tengan en cuenta atributos protegidos.

Dwork y otros han propuesto mecanismos de entrenamiento que producen modelos f que son individualmente equitativos. Como veremos en el apartado siguiente, la idea es restringir el problema de optimización del entrenamiento de forma que satisfaga la condición de equidad individual definida anteriormente.

Uno de los inconvenientes de la definición de equidad individual es que, a diferencia de las definiciones grupales, tenemos que encontrar una medida de distancia entre individuos que sea apropiada para la tarea. Esta medida de distancia, obviamente, deberá ignorar los atributos protegidos –atributos como el sexo o la raza tendrían que ser irrelevantes para comparar dos individuos respecto a la tarea–, pero, además, tiene que ser *útil* respecto a la tarea. Por ejemplo, para la tarea de adjudicación de plazas de una universidad, podemos medir el valor absoluto entre las notas de los solicitantes a las PAU. Esta distancia, sin embargo, no serviría para distinguir entre solicitantes que solicitan titulaciones diferentes, puesto que cada titulación tiene una nota de corte diferente y puede tener requisitos diferentes en cada uno de los exámenes individuales de la selectividad: por ejemplo, para entrar a la carrera de matemáticas, dos estudiantes son parecidos si han sacado una nota parecida en el examen de matemáticas de la selectividad. Como veis, en general, definir esta medida de distancia no es trivial y requiere la experiencia de un experto en la tarea.

2.3 Causalidad

Las definiciones anteriores son todas definiciones sobre cómo se comportan los modelos, también llamadas observacionales. A pesar de que son fáciles de comprobar en un modelo (como por ejemplo la separación de un clasificador se puede medir con el TPR y el FPR), no capturan las causas de la discriminación. Por ejemplo, en el caso de COMPAS sabemos que no se cumple la condición de independencia pero no somos capaces de determinar la causa de las diferencias entre los grupos solo a partir de los resultados del modelo. Existen marcos teóricos para modelar las relaciones de causalidad en un sistema (por ejemplo, redes causales y redes de creencias). Con modelos causales podemos razonar y atribuir las causas de la discriminación.

Nota

Aquí nos referimos a modelos matemáticos de causalidad y no a modelos de aprendizaje automático.

Lectura complementaria

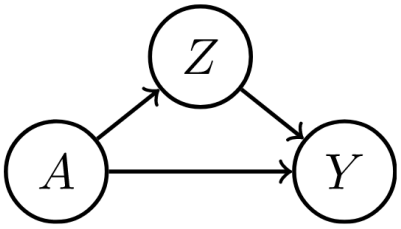
S. Chiappa; W. S. Isaac (2019). *A Causal Bayesian Networks Viewpoint on Fairness*. ArXiV.

Por ejemplo, imaginemos el caso de la adjudicación de plazas en UC Berkeley que hemos dado como ejemplo para explicar la paradoja de Simpson en el primer apartado. En la figura 6 mostramos un modelo gráfico de una red causal para la adjudicación de plazas en UC Berkeley, donde las variables relevantes

son: el sexo (A), la facultad escogida (Z) y la decisión de admisión (Y). Las flechas indican la relación de causalidad entre las variables: el sexo influye en la elección de la facultad (solicitantes de diferente sexo escogían facultades diferentes), la decisión depende de la facultad (algunas facultades eran más estrictas que otras), y además suponemos que hay una relación de causalidad directa entre el sexo y la decisión (si así fuera, habría discriminación).

En el estudio original, Bickel argumenta que no hay un efecto significativo del sexo sobre la decisión que favorezca a los solicitantes de sexo masculino. Utilizando el modelo anterior podemos razonar sobre esta hipótesis y preguntarnos: ¿qué habría pasado si manteniendo Z constante, hubiéramos cambiado el sexo de los solicitantes? Si observamos una diferencia significativa quiere decir que A tiene un efecto directo sobre Y.

Figura 6. Un posible modelo de causalidad para la adjudicación de plazas en UC Berkeley



Preguntas hipotéticas de este estilo, también llamadas contrafactuales, nos permiten determinar el efecto de una *intervención*. En este caso, hemos realizado una intervención en el modelo eliminando la relación entre A y Y mediada por la elección de la facultad y nos ha permitido medir el efecto directo de A sobre Y. Esta es una propiedad de las redes causales que es de gran utilidad para diseñar leyes y políticas antidiscriminación que enderecen casos de discriminación.

A pesar de todo, las respuestas a estos contrafactuales serán tan correctas como lo sean los modelos de causalidad: por ejemplo, si definimos relaciones de causalidad que no tienen ningún apoyo o pasamos por alto variables relevantes, podemos llegar a conclusiones incorrectas. De hecho, es posible construir dos modelos que tienen una estructura idéntica y se comportan igual en las intervenciones y, aun así, dan respuestas diferentes a los contrafactuales.

2.4 Otros

En la ley también encontramos definiciones de justicia algorítmica que pueden hacernos pensar sobre las definiciones anteriores. La mayoría de las leyes antidiscriminación de la Unión Europea protegen contra un tratamiento diferencial en áreas específicas como el mundo laboral y la educación. Estas leyes se denominan de «discriminación directa» o de «tratamiento dispar» (en in-

glés: *disparate treatment*) y hacen referencia a una discriminación intencional de un individuo.

También hay leyes que regulan un tratamiento diferencial entre grupos. Estas leyes intentan cubrir casos de discriminación sistemática a un grupo de la población que no necesariamente son intencionados. En Estados Unidos se denominan de«impacto dispar» (en inglés: *disparate impact*) y en la UE se denominan de «discriminación indirecta».

La ley en Estados Unidos establece un umbral que determina cuando se produce un impacto dispar: si un grupo presenta un porcentaje de miembros que se seleccionan inferiores por más de un 80 % respecto al porcentaje del grupo con el porcentaje de selección más alto, hay un impacto desfavorable para este grupo. Sin embargo, la ley solo se aplica si la entidad que hace la selección puede corregir esta disparidad sin que afecte a los intereses de su negocio. Es decir, que las decisiones las está tomando por la necesidad de satisfacer los requerimientos de la tarea. En el ejemplo de selección de personal, esta ley no se aplica si una empresa demuestra que selecciona a los candidatos más preparados, aunque para una discriminación histórica puede haber grupos que en promedio tengan un nivel de educación más bajo y, por lo tanto, tengan un porcentaje de selección más bajo.

Barocas apunta que estos dos tipos de leyes a menudo se encuentran en conflicto. El problema es que garantizar que individuos parecidos se traten de forma parecida tal como dictan las leyes de tratamiento dispar puede perpetuar las desigualdades entre grupos que las leyes de impacto dispar intentan corregir. Una opción intermedia es tratar a individuos que son aparentemente disimilares de formas similares si la diferencia se da por una discriminación en el grupo al que pertenece el individuo menos favorecido (discriminación positiva). De hecho, muchas de las definiciones que hemos visto en este apartado se encuentran en algún punto del espectro continuo entre las definiciones de tratamiento e impacto dispares (por ejemplo, la equidad individual se acerca al tratamiento dispar y la condición de independencia se acerca al impacto dispar).

Lectura complementaria

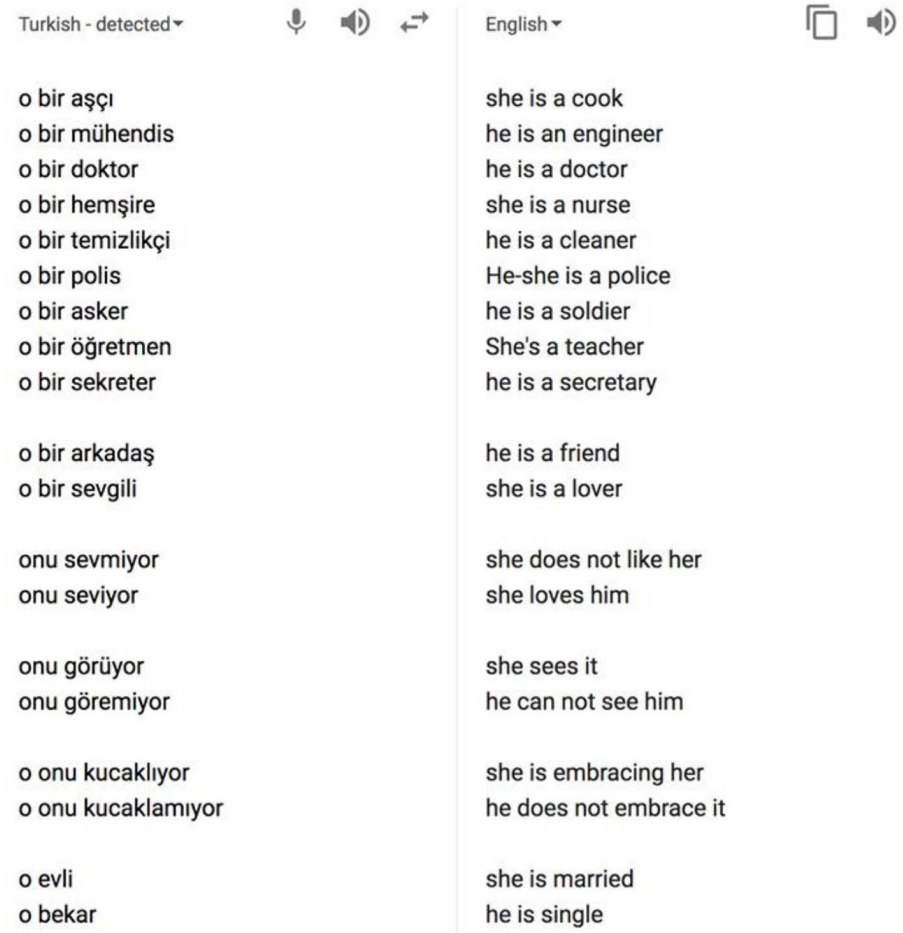
S. Barocas; M. Hardt (2017). *Fairness in Machine Learning NIPS 2017 Tutorial*. https://mrtz.org/nips17

Crawford hace una observación importante sobre estas definiciones: casi todas hablan de equidad en términos de reparto de oportunidades y recursos (males de asignación); en cambio, hay otro tipo de desigualdad que puede hacer el mismo daño: la falta de equidad en la representación de los grupos en la sociedad (males de representación). Crawford define los males de representación como los males relacionados con la perpetuación de estereotipos de grupos de la sociedad que no están suficientemente representados en la sociedad. Algunos ejemplos de estos males que han llamado mucho la atención son los estereotipos presentes en los *word embeddings*, las traducciones sesgadas del traductor de Google, y los errores del algoritmo de etiquetado de Google Photos.

Los *word embeddings* son un conjunto de técnicas de procesamiento del lenguaje donde las palabras y frases son representadas como vectores en un espacio multidimensional. Una de las motivaciones de este modelo era poder responder enunciados lógicos del estilo: *rey – hombre + mujer = reina*. Bolukbasi y otros demostraron que la interfaz *word2vec* utilizada para construir *word embeddings* heredaba e incluso amplificaba los estereotipos, estableciendo relaciones lógicas del estilo: «hombre es a programador lo que mujer es a ama de casa».

Caliskan y otros estudiaron las traducciones de Google Translate entre lenguajes con género gramatical y sin. En el estudio utilizaron el inglés y el turco; este último es un lenguaje sin género. Cuando traducían frases sin género el traductor añadía un género estereotipado a la traducción. Así, los ingenieros y los doctores eran siempre en masculino y las enfermeras y las maestras de escuela eran en femenino (véase la figura 7).

Figura 7. La adición de género en las traducciones del turco al inglés



Fuente: Emre Şarbak

Por último, otro escándalo fue el error del algoritmo de etiquetado de Google Photos, que etiquetó fotos de personas de raza negra como «gorilas». Los ingenieros de Google se apresuraron a quitar la etiqueta «gorilas» de las posi-

bles etiquetas del algoritmo para solucionar el problema al menos de manera provisional.

Existe un círculo vicioso entre los males de asignación y los males de representación: un grupo que recibe menos oportunidades y recursos a la larga sufrirá males de representación, que, a la vez, pueden inducir una discriminación en los algoritmos de aprendizaje utilizados para resolver las asignaciones. En esencia, el caso COMPAS es un mal de representación que se traduce en un mal de asignación.

3. Construcción de modelos equitativos

En este apartado hacemos un repaso de los métodos que se han propuesto para construir modelos que satisfagan las propiedades de equidad definidas en el apartado anterior. Los métodos se distinguen según la fase del proceso de aprendizaje automático en la que se aplican: preprocesamiento, si se aplican en el conjunto de entrenamiento; durante el entrenamiento, como restricciones en la tarea de optimización del entrenamiento; y como posprocesamiento, ajustando el modelo que ya se ha entrenado.

La mayoría de estos métodos incurren en un incremento del error del modelo al resolver la tarea. A menudo la entidad que toma las decisiones tendrá que encontrar un compromiso entre satisfacer un grado de equidad –por alguna de las definiciones de equidad– y una precisión y exactitud deseadas del modelo. Dependiendo de la tarea, los datos, el método de corrección y la definición de equidad, este compromiso será más o menos fácil de lograr.

En este apartado explicaremos y daremos ejemplos de cada uno de estos métodos.

3.1 Preprocesamiento

Las técnicas de preprocesamiento suponen que los datos son la principal fuente de sesgo. Por lo tanto, estas técnicas implementan modificaciones de los datos antes del entrenamiento para que el modelo final satisfaga alguna definición de equidad.

La mayoría de estas técnicas formulan el problema como una tarea de optimización en que se intenta encontrar una representación de los datos que minimice la discriminación (según alguna noción de equidad), minimice la distorsión en las observaciones individuales, y minimice el error en el modelo final. Estas nuevas representaciones se consiguen transformando el conjunto de entrenamiento realizando las siguientes operaciones sobre las observaciones:

1) asignar pesos a las observaciones en el proceso de entrenamiento;

2) cambiar las etiquetas de las observaciones;

3) descartar características que están correlacionadas con el atributo protegido.

Dependiendo de cómo se definen los objetivos del problema de optimización, de las nociones de equidad que se intentan lograr y las operaciones que son permitidas, se han definido varias técnicas de preprocesamiento: «ajuste de pesos», «preprocesamiento optimizado», «aprendiendo representaciones equitativas», y «eliminación de impacto dispar». Y, recientemente, se están aplicando técnicas avanzadas de optimización que usan redes antagónicas generativas (*generative adversarial networks*, GAN) para resolver el problema de optimización considerando un adversario que intenta extraer información sobre el atributo protegido.

Una de las ventajas del preprocesamiento es que es agnóstico a otras etapas del proceso de aprendizaje: por ejemplo, una vez se ha encontrado una representación equitativa de los datos, cualquier otro tipo de entrenamiento en este nuevo espacio también satisfará la definición de equidad.

3.2 Corrección durante el aprendizaje

Estas técnicas modifican el problema de optimización que se resuelve durante el entrenamiento del modelo para que el modelo final satisfaga una definición de equidad. El proceso para derivar un clasificador que satisface equidad individual propuesto por Dwork y otros y que hemos mencionado en el subapartado anterior es un ejemplo. Cabe decir, sin embargo, que la propiedad de equidad individual se puede satisfacer con técnicas de preprocesamiento y posprocesamiento.

Dos de las técnicas de corrección durante el entrenamiento más populares son:

- **Supresor de prejuicios**: añade un término (denominado de regularización) a la función de error que se minimiza en el proceso de entrenamiento para minimizar la discriminación del modelo. La regularización normalmente se hace para mejorar la generalización del modelo, en este caso se hace para reducir la discriminación.

- **Supresor de sesgo antagónico**: se añade como objetivo en la optimización para minimizar la habilidad de un adversario de inferir información sobre el atributo protegido a partir de las predicciones del modelo.

Las técnicas de corrección durante el aprendizaje tienen el potencial de conseguir modelos con bajo error, puesto que podemos optimizar el modelo con la condición de equidad integrada en el proceso de entrenamiento.

3.3 Posprocesamiento

Las técnicas de posprocesamiento toman un modelo que ya ha sido entrenado y ajustan sus salidas para que se satisfaga una noción de equidad. Normalmente se añade aleatoriedad en las salidas con este objetivo.

Denominamos un modelo derivado $\overline{Y} = F(\hat{Y},A)$ donde F es una función del modelo que queremos corregir ($\hat{Y}$) y del atributo protegido (A) que posiblemente contiene aleatoriedad. Dado un cierto coste para FP y FN, la tarea es encontrar un F que probabilísticamente modifique las salidas de $\hat{Y}$, de forma que satisfaga alguna definición de equidad y minimice el coste de los FP y FN que produce, en esperanza.

La ventaja de las técnicas de posprocesamiento es que son agnósticas al modelo o la familia de algoritmos de aprendizaje supervisado que se utilicen: no es necesario repetir el proceso de entrenamiento, que puede ser muy útil cuando el proceso de aprendizaje es complejo. Si no tenemos acceso al modelo o a los datos, y solo tenemos acceso a las salidas, el posprocesamiento puede ser la única opción viable. Estas ventajas, por otro lado, son lo que hacen que las técnicas de posprocesamiento incurran en incrementos sustanciales de error en el modelo final.

Enlace de interés

Para entender el efecto y el funcionamiento de los métodos presentados en este apartado recomendamos jugar con el *framework* de IBM para corregir modelos AIF360: http://aif360.mybluemix.net/

4. Interpretación y explicabilidad

En este apartado definimos la interpretabilidad y la explicabilidad. Además, haremos un repaso de las técnicas que se usan para dar más transparencia a los modelos de aprendizaje automático.

Como hemos introducido en el primer apartado, debido a la popularización de las redes neuronales profundas, ha aparecido la necesidad de dar transparencia a las decisiones que recomiendan los modelos. Ejemplos de modelos interpretables son: regresión lineal y logística, árboles de decisión, o clasificadores de Bayes ingenuos. Estos modelos son comprensibles e intuitivos. Por ejemplo, un árbol de decisión permite saber qué características son relevantes en la decisión e, incluso, el peso de cada característica en la decisión. Estas dos propiedades (relevancia y peso) sobre las características son ejemplos de propiedades de interpretabilidad deseables en un modelo. En contraste, las redes neuronales profundas suelen descubrir las características automáticamente y, a menudo, ni siquiera los expertos en la tarea son capaces de interpretarlas.

Un modelo de *caja negra* es un modelo que, o bien es demasiado complicado para entender, o bien es propietario y por lo tanto no podemos saber cómo funciona. Por el contrario, los modelos *interpretables* son modelos que no son de caja negra. Los modelos *explicables* son modelos de caja negra que permiten ser interpretados *a posteriori*. Las redes neuronales profundas son un ejemplo de modelo de caja negra, puesto que la complejidad de los modelos es elevadamente no lineal y por lo tanto la interpretación del modelo escapa a la intuición humana. En los subapartados siguientes veremos algunas técnicas para interpretar las salidas de modelos de caja negra.

Rudin apunta que, a menudo, la utilización de modelos interpretables no implica una disminución en el rendimiento final del modelo. Rudin advierte de los riesgos de usar modelos de caja negra y aboga por el uso de modelos interpretables, incluso cuando hay pequeñas pérdidas de precisión o exactitud. Una de las razones es que los errores no solo provienen del modelo sino que también pueden ocurrir a niveles más altos de la interacción con el modelo y la toma de decisiones; modelos que son interpretables permiten una mejor comprensión del modelo por parte de la persona que toma la decisión (por ejemplo, un médico que tiene que recomendar un tratamiento) y por lo tanto previenen este tipo de errores.

Lectura complementaria

C. Rudin (2019). «Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead». *Nature* (vol. 1, n.º 5, pág. 206-215). Nature Publishing Group.

A pesar de que coincidimos con la visión de Rudin, en este apartado veremos algunas técnicas de explicabilidad que se están utilizando actualmente.

Como no parece que las redes neuronales profundas se dejen de usar en un futuro próximo, es importante conocer las técnicas que se han propuesto para entender los resultados de estos modelos.

Para los siguientes subapartados nos basamos en el libro *Interpretable Machine Learning* de Christoph Molnar. En este módulo nos centraremos en técnicas que se utilizan específicamente para explicar redes neuronales, pero en el libro podéis consultar información sobre técnicas de interpretabilidad más generales.

Lectura complementaria

M. Christoph (2014). *Interpretable Machine Learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

4.1 Influencia de las observaciones

Una de las ideas más potentes para interpretar modelos complejos es medir el efecto de las observaciones en el modelo. Identificar las observaciones que han tenido mayor *influencia* y prestar atención a las características de estas observaciones puede ayudar a entender qué es lo que determina los parámetros del modelo e, incluso, decisiones particulares. Por ejemplo, si los modelos dependen fuertemente de una instancia tenemos razones para sospechar del modelo o de la instancia (por ejemplo, la instancia es de hecho un error) y habrá que investigarlo.

Para determinar el efecto de una observación en el modelo se elimina la observación y se entrena el modelo con las observaciones restantes. El efecto se mide en los cambios en los parámetros del modelo. Esta idea no es nueva y se utiliza para detectar observaciones atípicas en modelos de regresión desde los años setenta. Medidas de influencia basadas en la eliminación de observaciones son la distancia de Cook y DFBETA.

El inconveniente de repetir el entrenamiento para cada instancia es que es un proceso que requiere muchos recursos computacionales y puede ser inviable para modelos complejos (por ejemplo, redes neuronales) y con grandes conjuntos de datos de entrenamiento. En casos en los que la función de error en el conjunto de entrenamiento que se optimiza durante el entrenamiento tiene propiedades deseables (por ejemplo, es dos veces diferenciable respecto a los parámetros del modelo), podemos usar funciones de influencia para medir el efecto de las observaciones en los parámetros del modelo. La idea es aproximar la función de error alrededor de los parámetros actuales del modelo utilizando su gradiente. A través del gradiente podemos medir cómo cambian los parámetros del modelo cuando introducimos pequeñas perturbaciones en el conjunto de entrenamiento. Este método se puede aplicar en general en redes neuronales. Una desventaja, sin embargo, es que las funciones de influencia son aproximaciones y a menudo tendrán un error en las estimaciones de influencia de las observaciones.

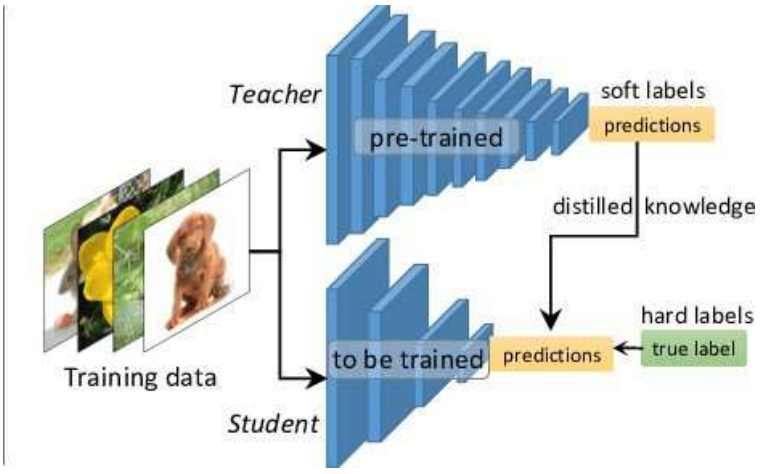
4.2 Destilación de conocimiento

Definición (*destilación*): la destilación consiste en transferir el conocimiento de un modelo complejo como una red neuronal profunda a un modelo simple e interpretable (como por ejemplo un árbol de decisión), de forma que sea igual de válido, es decir, que conserve la misma generalización que el modelo complejo.

A partir del modelo simple, podemos dar explicaciones del modelo complejo (es decir, el modelo complejo es explicable). La destilación también se puede ver como un proceso de compresión, puesto que el modelo simple también suele ser computacionalmente menos complejo y se puede ejecutar más eficientemente.

Para transferir el conocimiento del modelo complejo (modelo profesor) al simple (modelo estudiante), el modelo profesor minimiza la función de error en el entrenamiento y para cada observación devuelve un vector de probabilidades, donde cada coordenada del vector es la probabilidad de clasificar la observación en una clase, tal como lo haría normalmente. El modelo estudiante, entonces, recibe el mismo conjunto de entrenamiento de entrada y, para cada observación, intenta minimizar las probabilidades de sus salidas respecto a las que devuelve el modelo profesor –esta es la función de error del modelo estudiante.

Figura 8. Proceso de destilación de conocimiento



Fuente: Prakhar Ganesh (Medium)

El uso de la destilación en la práctica ha sido criticado por la comunidad de interpretabilidad puesto que normalmente se utiliza el modelo de caja negra y solo cuando se necesita una explicación se usa el modelo interpretable. Pero si un modelo simple es tan capaz como uno complejo de generalizar, ¿por qué no entrenar un modelo simple e interpretable desde un principio? O, por

lo menos, una vez destilado, ¿por qué no descartamos el modelo complejo y usamos el modelo interpretable?

4.3 Valores de Shapley

Una de las maneras de interpretar modelos simples es cuantificar la importancia de las características del modelo en las decisiones. Los árboles de decisión son un modelo simple que, por construcción, puede cuantificar la importancia de las características en las decisiones. Otros modelos y, en particular, modelos complejos como las redes neuronales profundas, no proporcionan un método directo para cuantificar la importancia de las características, lo que dificulta su interpretación.

Los valores de Shapley permiten medir la importancia de las características en modelos complejos. Su origen se encuentra en la teoría de juegos cooperativa. La premisa es: en una coalición de jugadores que cooperan para lograr un objetivo donde cada jugador tiene una contribución diferente, ¿cuán importante es la contribución de cada jugador en la cooperación y cuál es la ganancia que tendrían que obtener de acuerdo con su contribución? Pues bien, este concepto se puede aplicar en un proceso de aprendizaje automático donde las características son los jugadores y lo que queremos es cuantificar la contribución de cada característica en una decisión.

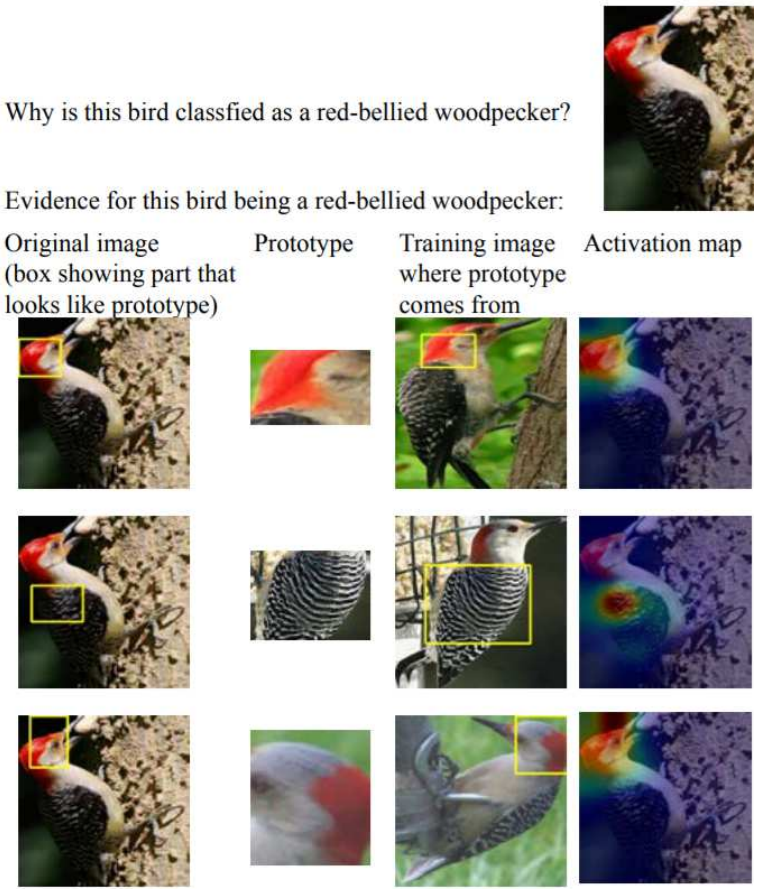
A pesar de que los valores de Shapley no son específicos para redes neuronales, hacemos mención de ellos por su importancia en estudios de sesgo algorítmico, puesto que a menudo se utilizan para medir la contribución del atributo protegido en las predicciones del modelo.

4.4 Visualización de características

Siguiendo con la idea de interpretar las decisiones de modelos complejos según las características del modelo, otra técnica que se propone para redes neuronales profundas en aplicaciones de visión por computador se basa en generar visualizaciones de las características.

La idea tras la visualización de características es producir explicaciones de las decisiones a través de comparaciones con miembros prototípicos de la clase. La idea es interesante porque intenta dar explicaciones a partir de los atributos esenciales de la clase. Por ejemplo, en la figura 9 mostramos como una red neuronal que clasifica una imagen como pájaro carpintero negro la clasifica como tal porque presenta las características identificativas de esta especie de pájaro: cresta roja y un plumaje con líneas negras y blancas.

Figura 9



Fuente: Chen y otros

Figura 9

Explicación de la clasificación de un pájaro carpintero negro utilizando comparaciones con otros miembros de la clase y señalando las características identificativas en la imagen.

5. Conclusiones

Este módulo ha introducido las principales formalizaciones de equidad algorítmica que hay en la literatura, los resultados de imposibilidad y las discusiones que encontramos alrededor de las definiciones. Además, hemos descrito técnicas de corrección de los modelos que nos permiten obtener modelos que satisfacen algunas de estas propiedades. Por último, hacemos un repaso a los métodos de explicabilidad que se están usando actualmente para entender y razonar sobre los modelos más complejos, lo que contribuye a hacer las decisiones más transparentes y por lo tanto detectar y auditar casos de discriminación.

Este módulo no pretende hacer una compilación exhaustiva de todas las definiciones y métodos relacionados con el campo del sesgo algorítmico que existen. Además, debe hacerse una distinción entre este campo, que es relativamente nuevo, y otras definiciones de equidad que encontramos en economía y teoría de juegos. A pesar de que están relacionadas, en estos campos se dan definiciones de equidad relacionadas con el bienestar de los individuos y las diferentes formas de repartir recursos y oportunidades, pero dan menos importancia a la discriminación basada en atributos protegidos. De hecho, estas definiciones de bienestar preceden el campo de aprendizaje automático como lo conocemos actualmente y, por lo tanto, el énfasis no recae en los riesgos de la toma de decisiones automatizada con algoritmos.

Ejercicios de autoevaluación

1. ¿Cuál de las definiciones de equidad grupal es más adecuada para las siguientes tareas?

a) Acceso a la universidad

b) Selección de personal

c) Concesión de préstamos bancarios

2. Propón una medida de distancia entre individuos para garantizar equidad individual en el acceso a la universidad. Razona la respuesta.

3. Elabora una tabla y una gráfica que muestren las estadísticas sobre las solicitudes a préstamos bancarios de un banco, indicando los préstamos que se han devuelto. Los datos tienen que satisfacer la condición de separación. Razona por qué esta condición se satisface.

4. Explica qué quiere decir la condición de independencia en el caso de COMPAS. Razona por qué no es una buena definición de equidad en este caso.

5. Da y explica ejemplos de males de representación.

6. Define un modelo de caja negra, un modelo explicable y un modelo interpretable. Da ejemplos donde sea adecuado aplicar un modelo complejo de caja negra y justifica la respuesta.

Solucionario

1. Las siguientes son posibles elecciones de la condición de equidad grupal para cada uno de los casos:

a) En este caso puede ser deseable que se satisfaga una condición de independencia que ayude a corregir sesgos existentes. Tendría un efecto parecido a garantizar unas cuotas mínimas de estudiantes de todos los grupos, aunque estos grupos, de media, hayan obtenido notas inferiores a otros grupos.

b) La condición de separación puede ser la más deseable puesto que garantiza que no se rechazan buenos candidatos de manera dispar entre los grupos, sin que la organización contratante tenga que renunciar a los intereses del negocio.

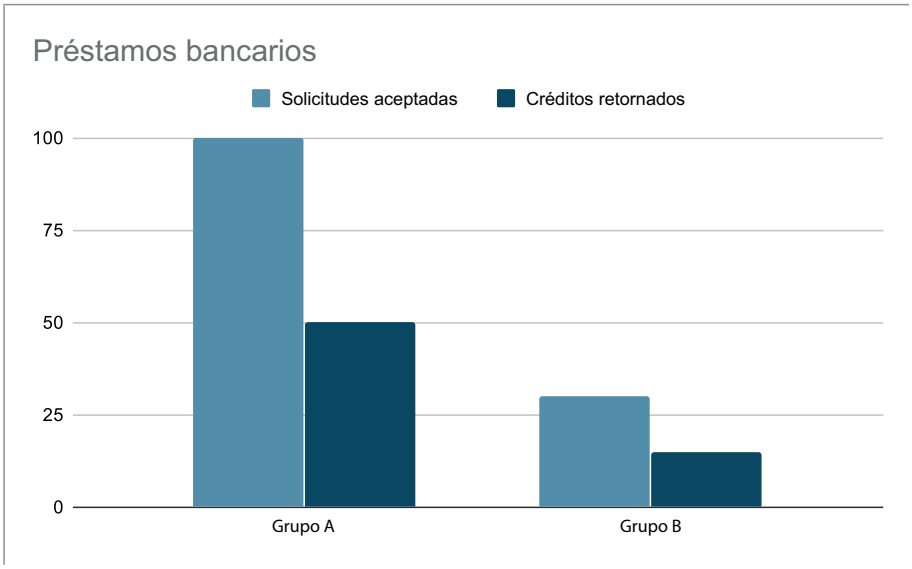
c) La condición de suficiencia puede ser deseable en este caso, puesto que nos permite hacer una interpretación de la probabilidad de no devolver el crédito que es independiente de los atributos protegidos.

2. Una distancia entre los individuos para garantizar equidad individual al acceso en la universidad puede ser la diferencia entre las notas de las PAU. Esta distancia garantizaría que dos estudiantes que han sacado notas parecidas tienen oportunidades parecidas y tienen una probabilidad parecida de acceder a los departamentos en los que solicitan una plaza. Aun así, imponer equidad individual con esta distancia puede no ser suficiente para garantizar un tratamiento justo entre los estudiantes. La discriminación puede haberse dado mucho antes. Un ejemplo es ese estudiante que pertenece a una minoría marginalizada y que no ha tenido acceso al mismo nivel de educación o no lo ha podido aprovechar por no tener un núcleo familiar estable. En este caso, aunque este estudiante fuera capaz de sacar una buena nota en las PAU, un tratamiento discriminatorio previo en el grupo de la población al que pertenece puede haber influenciado que no obtuviera unos buenos resultados. La equidad individual no puede corregir este problema.

3. Un ejemplo de tabla y gráfica son:

Figura 10. Como podemos observar en la figura, el TPR es el mismo para ambos grupos: $TPR_A = TPR_B = 0.5$

	Solicitudes aceptadas	Créditos retornados
Grupo A	100	50
Grupo B	30	15



4. En el caso de COMPAS, si tomamos la raza como atributo protegido (A), la condición de independencia se satisface si el porcentaje de presos a los que se les concede la libertad es el mismo para todas las razas. Por lo tanto, es posible que se tenga que conceder la libertad condicional a presos con alto riesgo para lograr la misma cuota para todos los grupos. Esto, obviamente, no es deseable en el caso de COMPAS.

5. Algunos ejemplos son:

a) La falta de profesionales de sexo femenino en posiciones relacionadas con las TIC favorece estereotipos de sexo e induce un sesgo en la elección de carreras de futuros profesionales.

b) La falta de políticos que pertenezcan a minorías étnicas y religiosas resulta en una falta de representación de estas minorías en las instituciones y en la toma de decisiones.

c) Los pocos casos de futbolistas profesionales del fútbol masculino que son abiertamente homosexuales crean estereotipos sobre la orientación sexual y la masculinidad en este deporte.

d) La falta de personas de raza negra que ocupan cargos de alta responsabilidad en grandes corporaciones.

6. Un modelo de caja negra es un modelo que o bien es propietario o bien es tan complejo que escapa al entendimiento humano. Un modelo interpretable es un modelo que no es de caja negra. Un modelo explicable es un modelo de caja negra que soporta técnicas de explicabilidad como las que hemos descrito en el último apartado de este módulo.

Los modelos de caja negra se pueden utilizar en tareas en las que no necesitamos razonar sobre las decisiones o bien en las que no tener una explicación supera de lejos los beneficios de aplicar el modelo. Siguiendo este argumento se podría justificar el uso de un modelo complejo que obtiene una precisión y exactitudes significativamente superiores a cualquier otro modelo interpretable en la detección de algún tipo de cáncer.

Glosario

aprendizaje automático *m* Campo de la inteligencia artificial que se dedica al desarrollo de algoritmos que mejoran su rendimiento a partir de experiencia.

aprendizaje supervisado *m* Campo del aprendizaje automático en el que la experiencia que los algoritmos utilizan para mejorar se basa en ejemplos de datos.

atributo protegido *m* Atributo de la identidad de un individuo sobre el que la toma de decisiones se ha regulado por medio de leyes antidiscriminación. Por ejemplo, las leyes antidiscriminación suelen considerar protegidos los atributos como el sexo, el género, la edad, la minusvalía, la orientación sexual, la ideología política, la expresión religiosa, entre otros.

clasificador *m* Función sobre un conjunto de elementos que devuelve las clases de estos elementos. En el contexto de aprendizaje automático haremos referencia a los algoritmos de aprendizaje supervisado que tienen como objetivo generalizar estas funciones de clasificación a partir de un conjunto de ejemplos.

conjunto de entrenamiento *m* Ejemplos que se proporcionan al algoritmo de aprendizaje supervisado para encontrar un regresor o un clasificador.

equidad individual *f* Propiedad de equidad algorítmica de un algoritmo de decisión. Las decisiones de un algoritmo de decisión satisfacen equidad individual si las diferencias entre los tratamientos recibidos por las decisiones de toda pareja de individuos están acotados por las diferencias como individuos.

espacio de hipótesis *m* Conjunto de funciones que el algoritmo de aprendizaje supervisado puede elegir para resolver el problema de optimización de aprendizaje que reduce el error sobre el conjunto de entrenamiento.

etiqueta *f* En el contexto de clasificadores, es la clase de un elemento del conjunto de entrenamiento.

explicabilidad *f* Propiedad de un modelo de caja negra que indica que se puede interpretar con un procesamiento del modelo añadido.

falsos negativos *m* Elementos de la clase positiva que el clasificador marca como negativos.
sigla **FN**

falsos positivos *m* Elementos de la clase negativa que el clasificador marca como positivos.
sigla **FP**

FPR *m* Cociente de falsos positivos entre el número total de negativos.

independencia *f* Propiedad de equidad algorítmica de un algoritmo de decisión. Decimos que un algoritmo de decisión cumple la propiedad de independencia si el atributo es independiente de las decisiones del clasificador.

interpretabilidad *f* Propiedad de un modelo que implica que no es de caja negra.

mal de asignación *m* Efectos perjudiciales de sesgos en decisiones que llevan a un reparto desigual de recursos y oportunidades entre diferentes miembros y grupos de la sociedad.

mal de representación *m* Efectos perjudiciales de la falta de representación de grupos en la sociedad.

modelo *m* En este texto nos referimos a modelos de aprendizaje supervisado, que son descripciones de la distribución estadística de los datos, incluidas suposiciones sobre cómo se han generado los datos.

modelo de caja negra *m* Modelos que o bien son demasiado complicados de entender o que son propietarios y que, por lo tanto, no se tiene acceso a los parámetros del modelo.

negativos verdaderos *m* Elementos de la clase negativa que el clasificador marca como negativos.
sigla **TN**
en true negatives

NPV *m* Cociente de negativos verdaderos entre todos los elementos que se han clasificado como positivos.

paradoja de Simpson *f* Fenómeno estadístico en el que una correlación lineal entre dos variables se invierte cuando se desglosa en subgrupos de la población.

positivos verdaderos *m* Elementos de la clase positiva que el clasificador marca como positivos.
sigla **TP**
en true positives

PPV *m* Cociente de positivos verdaderos entre todos los elementos que se han clasificado como positivos.

regresor *m* Función que aproxima una relación entre un conjunto de elementos y una variable dependiente de estos elementos.

regulación de la discriminación directa *f* Conjunto de leyes de antidiscriminación que regulan el uso explícito de atributos protegidos para la selección de individuos en ámbitos de la sociedad como el laboral y la educación.
sin. **tratamiento dispar**

regulación de la discriminación indirecta *f* Conjunto de leyes de antidiscriminación que regulan los sesgos en la toma de decisiones que afectan a grupos de la sociedad (definidos por un atributo protegido), aunque estas decisiones no se hayan tomado haciendo uso explícito de un atributo protegido.
sin. **impacto dispar**

respuesta *f* En el contexto de regresores, es el valor de la variable dependiente.

separación *f* Propiedad de equidad algorítmica de un algoritmo de decisión. Decimos que un algoritmo de decisión cumple la propiedad de separación si el atributo y las decisiones del clasificador son condicionalmente independientes respecto a la variable dependiente.

sesgo (estimador) *m* En estadística, el sesgo de un estimador estadístico (por ejemplo, la varianza de la muestra) es la diferencia entre la esperanza matemática del valor del estimador y el valor del parámetro que estima.

sesgo (muestra) *m* En estadística el sesgo puede referirse al sesgo de una muestra, que puede haberse introducido por una metodología de recolección de datos defectuosa. Por ejemplo, un sesgo se puede introducir si se hace una selección no aleatoria de los sujetos de la muestra.

sesgo (social) *m* Desigualdades entre grupos de una sociedad.

suficiencia *f* Propiedad de equidad algorítmica de un algoritmo de decisión. Decimos que un algoritmo de decisión cumple la propiedad de suficiencia si el atributo y la variable dependiente son condicionalmente independientes respecto a las decisiones del clasificador.

TPR *m* Cociente de positivos verdaderos entre el número total de positivos.

word embedding *m* Técnica de procesamiento del lenguaje natural que consiste en representar palabras como vectores de números reales.

Bibliografía

Angwin, Julia; Larson, Jeff; Mattu, Surya; Kirchner, Lauren (2016). *Machine Bias*. Nueva York: ProPublica. [Fecha de consulta: 20 de agosto de 2020]. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, Solon; **Selbst, Andrew D.** (2016). «Big data's disparate impact». *Calif. L. Rev.*, 104.

Barocas, Solon; **Hardt, Moritz**; **Narayanan, Arvind** (2019). *Fairness and Machine Learning*. fairmlbook.org, http://www.fairmlbook.org

Bickel, Peter J.; **Hammel, Eugene A.**; **O'Connell, J. William** (1975). «Sex bias in graduate admissions: Data from berkeley». *Science* (vol. 187, n.º 4175, pág. 398-404).

Bolukbasi, Tolga; **Chang, Kai-Wei**; **Zou, James Y.**; **Saligrama, Venkatesh**; **Kalai, Adam T.** (2016). «Man is to computer programmer as woman is to homemaker? debiasing word embeddings». En: *Advances in neural information processing systems* (pág. 4349-4357).

Caliskan, Aylin; **Bryson, Joanna J.**; **Narayanan, Arvind** (2017). «Semantics derived automatically from language corpora contain human-like biases». *Science* (vol. 356, n.º 6334, pág. 183-186).

Calmon, Flavio; **Wei, Dennis**; **Vinzamuri, Bhanukiran**; **Ramamurthy, Karthikeyan Natesan**; **Varshney, Kush R.** (2017). «Optimized pre-processing for discrimination prevention». En: *Advances in Neural Information Processing Systems* (pág. 3992-4001).

Chen, Chaofan; **Li, Oscar**; **Tao, Daniel**; **Barnett, Alina**; **Rudin, Cynthia**; **Su, Jonathan K.** (2019). «This looks like that: deep learning for interpretable image recognition». En: *Advances in neural information processing systems* (pág. 8930-8941).

Chiappa, Silvia; **Isaac, William S.** (2018). «A causal bayesian networks viewpoint on fairness». En: *IFIP International Summer School on Privacy and Identity Management* (pág. 3-20). Nueva York: Springer.

Chouldechova, Alexandra (2017). «Fair prediction with disparate impact: A study of bias in recidivism prediction instruments». *Big data* (vol. 5, n.º 2, pág. 153-163).

Corbett-Davies, Sam; Pierson, Emma; Feller, Avi; Goel, Sharad (2016). «A computer program used for bail and sentencing decisiones was labeled biased against blacks. It's actually not that clear». *The Washington Post*. [Fecha de consulta: 20 de agosto de 2020]. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Corbett-Davies, Sam; **Pierson, Emma**; **Feller, Avi**; **Goel, Sharad**; **Huq, Aziz** (2017). «Algorithmic decision making and the cost of fairness». A: *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pág. 797-806).

Crawford, Kate (2017). *The Trouble with Bias*. NIPS Keynote. [Fecha de consulta: 8 de julio de 2020]. https://www.youtube.com/watch?v=fmym_BKWQzk

Dieterich, William; **Mendoza, Christina**; **Brennan, Tim** (2016). *Compas risk scales: Demonstrating accuracy equity and predictive parity*. Misuri: Northpointe Inc.

Dwork, Cynthia; **Hardt, Moritz**; **Pitassi, Toniann**; **Reingold, Omer**; **Zemel, Richard** (2012). «Fairness through awareness». En: *Proceedings of the 3rd innovations in theoretical computer science conference* (pág. 214-226).

European Network of Legal Experts in Gender Equality and Non-discrimination (2018). *European equality law review. Justice and Consumers*. [Fecha de consulta: 20 de agosto de 2020]. https://ec.europa.eu/info/sites/info/files/law_review_2018_2.pdf

European Union Agency for Fundamental Rights (2018). *#BigData: Discrimination in data-supported decision making*. FRAY Focus. [Fecha de consulta: 20 de agosto de 2020]. https://fra.europa.eu/sites/default/files/fra_uploads/fray-2018-foco-big-data_en.pdf

Feldman, Michael; **Friedler, Sorelle A.**; **Moeller, John**; **Scheidegger, Carlos**; **Venkatasubramanian, Suresh** (2015). «Certifying and removing disparate impact». En: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pág. 259-268).

Hardt, Moritz (2017). *How big data is unfair: Understanding unintended sources of unfairness in data driven decision making*. Medium (2014). [Fecha de consulta: 8 de julio de 2020]. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Hinton, Geoffrey; **Vinyals, Oriol**; **Dean, Jeff** (2015). «Distilling the knowledge in a neural network». *arXiv preprint arXiv:1503.02531*

Johnson, Benjamin; **Jordan, Richard** (2017). *Why should like cases be decided alike? a formal model of aristotelian justice*.

Kamiran, Faisal; **Calders, Toon** (2012). «Data preprocessing techniques for classification without discrimination». *Knowledge and Information Systems* (vol. 33, n.º 1, pág. 1-33).

Kamishima, Toshihiro; **Akaho, Shotaro**; **Asoh, Hideki**; **Sakuma, Jun** (2012). «Fairness-aware classifier with prejudice remover regularizer». En: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pág. 35-50). Berlín: Springer.

Kleinberg, Jon; **Mullainathan, Sendhil**; **Raghavan, Manish** (2016). «Inherente trade-offs in the fair determination of risk scores». *arXiv preprint arXiv:1609.05807*.

Larson, Jeff; Mattu, Surya; Kirchner, Lauren; Angwin, Julia (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. Nueva York: ProPublica. [Fecha de consulta: 20 de agosto de 2020]. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Mitchell, Shira (2018). *Mirror Mirror. Reflections on Quantitative Fairness*. [Fecha de consulta: 20 de agosto de 2020]. https://shiraamitchell.github.io/fairness/

Molnar, Christoph (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

Narayanan, Arving (2019). *How to recognize AY snake oil*. Presentación en el MIT. [Fecha de consulta: 8 de julio de 2020]. https://www.cs.princeton.edu/ arvindn/talks/MIT-STS-AY-snakeoil.pdf

Narayanan, Arvind (2018). *Tutorial: 21 fairness definitions and their politics*. FAT*. [Fecha de consulta: 20 de agosto de 2020]. https://www.youtube.com/watch?v=jIXIuYdnyyk

Pleiss, Geoff; **Raghavan, Manish**; **Wu, Felix**; **Kleinberg, Jon**; **Weinberger, Kilian Q.** (2017). «On fairness and calibration». En: *Advances in Neural Information Processing Systems* (pág. 5680-5689).

Rudin, Cynthia (2019). «Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead». *Nature Machine Intelligence* (vol. 1, n.º 5, pág. 206-215).

Suresh, Harini (2019). *The Problem with «Biased Data»*. Medium. [Fecha de consulta: 20 de agosto de 2020]. https://medium.com/harinisuresh/the-problem-with-biased-data-5700005e514c

Vapnik, Vladimir (2013). *The nature of statistical learning theory*. Berlín: Springer Science & Business Media.

Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference*. Berlín: Springer Science & Business Media.

Xu, Depeng; **Yuan, Shuhan**; **Zhang, Lu**; **Wu, Xintao** (2019). «Fairgan+: Achieving fair data generation and classification through generative adversarial nets». En: *2019 IEEE International Conference on Big Data (Big Data)* (pág. 1401-1406). Nueva Jersey: IEEE.

Zemel, Rich; **Wu, Yu**; **Swersky, Kevin**; **Pitassi, Toni**; **Dwork, Cynthia** (2013). «Learning fair representations». En: *International Conference on Machine Learning* (pág. 325-333).

Zhang, Brian Hu; **Lemoine, Blake**; **Mitchell, Margaret** (2018). «Mitigating unwanted biases with adversarial learning». En: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pág. 335-340).