Biaix algorísmic

PID_00278565

Marc Juárez Miró

Universitat Oberta de Catalunya

Marc Juárez Miró

Investigador postdoctoral a la Universitat del Sud de Califòrnia, on estudia problemes de justícia algorítmica. Va obtenir el seu doctorat el 2019 per la Universitat de Lovaina, amb una tesi sobre privadesa i tècniques d'anàlisi del trànsit de xarxes.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Cristina Pérez Solà

Primera edició: febrer 2021

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Marc Juárez Miró
Producció: FUOC

Índex

Introducció

El paradigma del *big data* ha tingut un efecte transversal en l'economia de les societats de l'era de la informació. En pràcticament tots els sectors s'han adoptat tècniques per extreure informació dels grans volums de dades que es generen, amb l'objectiu d'optimitzar processos i prendre millors decisions. Aquest paradigma s'ha vist reforçat pels avenços en el camp de l'aprenentatge automàtic, al qual ha proveït d'eines per automatitzar l'extracció de coneixement de les dades. Mentre que aquest paradigma ha estat un motor de progrés per resoldre algunes tasques concretes, l'automatització de decisions en àmbits com la justícia, l'educació i les finances obre interrogants sobre l'ètica de delegar certes decisions a algorismes d'aprenentatge automàtic. A més, tot i que aquests algorismes hagin aconseguit optimitzar alguns processos amb èxit, encara no entenem les implicacions que poden tenir en una societat que els aplica sistemàticament per prendre decisions.

Val a dir que hem estat testimonis de grans progressos en el camp de l'aprenentatge automàtic i, especialment, de la revolució que han portat les xarxes neuronals profundes al camp de la visió per computador. Per exemple, aquestes noves tècniques s'han fet servir per construir sistemes de detecció de tumors cancerígens que superen en precisió als experts humans. No obstant això, un optimisme desmesurat, alimentat per aquests resultats positius, ha motivat l'aplicació d'aquests algorismes en problemes de tots els àmbits de la societat. Actualment s'estan emprant algorismes d'aprenentatge automàtic per a la selecció de personal, l'accés a universitats, per a l'adjudicació d'assegurances i crèdits bancaris i, fins i tot, per a concedir la llibertat condicional a presos. Aquestes decisions poden ser crítiques pel desenvolupament i l'estatus dels individus que en són subjectes, ja que determinen les oportunitats a les quals tenen accés i, per tant, les seves perspectives de futur. És per això que amb la popularització dels algorismes d'aprenentatge automàtic, hi ha una creixent preocupació per les conseqüències que pot tenir en el repartiment de la riquesa, la mobilitat social dels individus i l'eixamplament de les desigualtats existents en la societat. El camp del biaix algorísmic ha nascut per estudiar aquests aspectes des d'un punt de vista científic, ètic i legal.

Un dels arguments que es dona per justificar l'ús d'algorismes per prendre decisions és que els algorismes tenen el potencial de ser neutrals, ja que estan exempts dels biaixos i estereotips que els humans adquirim, a vegades sense ser-ne conscients. Aquest argument no és vàlid per diverses raons. En primer lloc, els algorismes estan aplicats per experts humans que poden introduir biaixos en diversos punts del procés de dissenyar, implementar i executar els algorismes. En segon lloc, encara que aquests experts portin a terme aquestes tasques de manera completament imparcial, les dades que es proporcionen

als algorismes també poden contenir biaixos. Les desigualtats socials es reflecteixen a les dades i sense cap intervenció que ho previngui, els algorismes hereten aquests biaixos.

El problema del biaix en les dades és més complex del que pot semblar a primera vista. A causa de l'elevada complexitat dels algorismes d'aprenentatge automàtic, és difícil anticipar-se als biaixos que els algorismes aprendran de les dades. De fet, les xarxes neuronals profundes extreuen patrons no lineals de gran complexitat que fan difícil entendre i raonar els resultats. Aquesta falta de transparència en les decisions entra en conflicte amb la llei, on s'estipula que els subjectes de les decisions tenen dret a saber les raons que han motivat el resultat de les decisions. En particular, a la Unió Europea, el Reglament General de Protecció de Dades també regula els algorismes d'aprenentatge automàtic que operen amb dades personals i deixa ben clar que els subjectes poden demanar explicacions sobre les decisions que els afecten. Malauradament, la major precisió de les xarxes neuronals profundes respecte a altres algorismes ha fet que aquestes no hagin parat de guanyar en popularitat.

En aquest mòdul fem un repàs del naixent camp científic que estudia el biaix social dels algorismes per la presa de decisions. En el primer apartat definim els conceptes bàsics sobre els algorismes d'aprenentatge automàtic i donem exemples dels biaixos que aquests algorismes poden tenir. En el segon apartat farem un repàs sobre definicions formals d'equitat que s'han proposat pels algorismes i dels resultats d'impossibilitat. En el tercer parlem sobre mètodes per corregir els biaixos de l'algorisme. Per últim, parlarem de mètodes per explicar i interpretar les decisions d'algorismes complexos.

Objectius

Els objectius que l'estudiant ha d'assolir en finalitzar el mòdul són els següents:

1. Conèixer les principals fonts de biaix en els models d'aprenentatge supervisat.
2. Comprendre les definicions formals d'equitat, les relacions entre elles i les seves limitacions.
3. Conèixer els mètodes de correcció del biaix i poder relacionar-los amb les definicions d'equitat que permeten assolir.
4. Distingir entre explicabilitat i interpretabilitat i conèixer alguns dels mètodes que hi ha per explicar models complexos.

1. Automatització de les decisions

En aquest apartat introduïm el llenguatge necessari per parlar de la presa de decisions automatitzada a partir d'observacions. Tindrem presents dos casos d'ús:

1) Acceptar o rebutjar la sol·licitud d'un candidat per un lloc de feina o una universitat.

2) Predir el risc de concedir una assegurança o un préstec bancari.

En essència, l'objectiu de l'entitat que pren les decisions és, donada una observació X, determinar el valor d'una variable Y (el resultat de la decisió), que més el beneficia. En l'exemple de la selecció de personal, X són les qualificacions i habilitats del candidat, i Y codifica si es contracta el candidat o no. Les decisions seran encertades si es rebutgen candidats que no són capaços d'exercir i s'accepten candidats amb alt rendiment.

Idealment, l'entitat que pren la decisió desitja tenir una funció f que envia valors de la variable X a valors de la variable Y de manera que, en aplicar la funció en una observació que no coneix, f retorna el valor de la decisió. En el cas que Y pugui prendre valors d'un conjunt finit –tal com en l'exemple de selecció de personal anterior– f s'anomena un *classificador* i els valors de Y són les classes. En canvi, si Y és una variable contínua, f s'anomena *regressor* i, la tasca, *regressió*. Per exemple, en el cas de predir el risc de concedir un crèdit, Y podria prendre valors en $[0,1]$ i la imatge de f es podria interpretar com les probabilitats de retornar el crèdit. Cal dir que malgrat que s'utilitzi un regressor, sovint la decisió final serà discreta (per exemple, el crèdit es donarà si la probabilitat de pagament és superior a 0.8). De fet, la majoria de classificadors no retornen un valor discret, sinó un valor de confiança que «s'escapça» de la mateixa manera que quan es discretitza la sortida d'un regressor.

L'aprenentatge supervisat és la branca de l'aprenentatge automàtic que té com a objectiu trobar funcions f a partir de conjunts d'observacions. L'aprenentatge supervisat és el mètode predominant per resoldre aquest tipus de problemes a la pràctica. En el següent subapartat fem un petit repàs als conceptes bàsics d'aprenentatge supervisat que són necessaris per seguir la resta del mòdul.

1.1. Aprenentatge supervisat

Els algorismes d'aprenentatge supervisat parteixen d'un conjunt d'observacions (x_1,y_1), ..., (x_n,y_n), conegut com a *conjunt d'entrenament*, on cada x_i s'anomena *instància* o *exemple* i y_i s'anomena *etiqueta*, si f és un classificador, o *resposta*, si és un regressor. Una instància és típicament un vector de característiques que descriuen l'observació. Per exemple, en la sol·licitud de crèdit, algunes característiques rellevants són: si el sol·licitant té una feina estable, la seva nòmina, proves de solvència i avals.

El marc teòric dels algorismes d'aprenentatge automàtic suposa que X i Y són variables aleatòries i que els exemples del conjunt d'entrenament s'han obtingut prenent mostres independents de la distribució conjunta $X \times Y$. Aleshores, donada una nova observació d'$X = x$ per la qual no coneixem l'y corresponent, l'f òptim és aquell que retorna l'y que maximitza la següent probabilitat: $P(X = x \mid Y = y)$. El repte és que a la pràctica no coneixem aquesta distribució i, per tant, hem de recórrer a mètodes estadístics que ens permetin modelar la distribució a partir del conjunt d'entrenament. És per això que tant si f és un classificador com si és un regressor, direm que f és un *model*. Tal com X i Y es poden considerar variables aleatòries, també es poden interpretar les sortides del model com una variable aleatòria: $\hat{Y} = f(X)$. En un abús de llenguatge, a vegades ens referirem a $\hat{Y}$ com al model.

El paradigma més emprat per trobar un f adequat tracta d'encarar el problema com un problema d'optimització, on se selecciona l'f de dins d'una classe de funcions, anomenat *espai d'hipòtesis*, que minimitza l'error que f incorre en el conjunt d'entrenament. Aquest error es calcula segons una certa funció d'error que depèn de la família d'algorismes d'aprenentatge supervisat que es triï. No obstant això, aquestes funcions d'error es dissenyen per tal de minimitzar l'error en el conjunt d'entrenament i, a més, permetre que f generalitzi a instàncies que no s'han observat prèviament. Típicament, l'espai d'hipòtesis està parametritzat i l'optimització es fa sobre aquests paràmetres. Aquest procés d'optimització s'anomena *entrenament* i és per això que el conjunt d'observacions on es realitza l'optimització s'anomena *conjunt d'entrenament*. En aquest mòdul ometrem els detalls tècnics d'aquest procés ja que no són rellevants per les explicacions que farem sobre equitat algorísmica.

Minimització del risc empíric

El marc teòric al qual ens referim en aquest text s'anomena *minimització del risc empíric*, ja que té com a objectiu minimitzar una funció d'error en el conjunt d'entrenament.

Lectura complementària

Podeu trobar més informació en qualsevol llibre de text sobre teoria d'aprenentatge automàtic. Per exemple:
V. Vapnik (2000). *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer-Verlag.

1.2. L'error en la població

Com hem explicat en el subapartat anterior, el procés d'entrenament retorna un model f minimitzant l'error en el conjunt d'entrenament i intentant que generalitzi la població. Ara bé, abans de portar els models a la pràctica, necessitem mesurar com de satisfactori ha estat el procés d'entrenament. En altres paraules, volem mesurar l'error que f comet en el conjunt d'entrenament i, encara més important, quin és l'error que comet en observacions que no es

troben en el conjunt d'entrenament –és a dir, volem saber si f aconsegueix generalitzar de les mostres a la població.

Un dels mètodes més populars per mesurar aquest error és utilitzar mètodes de validació encreuada (en anglès: *cross-validation*). La validació encreuada és un mètode estadístic que no fa suposicions sobre la distribució de les dades. La validació encreuada consisteix a apartar un subconjunt de les dades abans de treballar amb elles, anomenat el conjunt de testeig. Un cop el model s'ha entrenat, fem servir el conjunt de testeig per mesurar l'error del model en mostres que no ha vist prèviament i, així, estimar l'error del model en la població.

Per mesurar l'error del model en el conjunt de testeig, es fan servir diferents mètriques depenent del tipus d'error del classificador que ens interessi avaluar. Aquí considerarem només mètriques que mesuren l'error de classificadors binaris. Els classificadors binaris només tenen dues classes disjuntes, i.e., Y pren valors en $\{0,1\}$. Anomenem 1 la classe positiva i 0 el complement d'aquesta, la classe negativa. Penseu en aquestes classes com els resultats d'una decisió binària: el resultat és o bé acceptar o bé rebutjar, respectivament. Aleshores, el classificador pot equivocar-se de dues maneres possibles: falsos negatius (FN), instàncies de la classe positiva que es classifiquen com a negatives; o falsos positius (FP), instàncies negatives que s'han classificat com a positives. De la mateixa manera, hi ha dos tipus de prediccions correctes: instàncies positives i negatives que s'han classificat en la classe correcta, és a dir, positius vertaders (TP, de l'anglès *True Positives*), i negatius vertaders (TN, de l'anglès *True Negatives*), respectivament. El total d'instàncies és: $Total = TP + FN + FP + TN$.

Les mètriques més utilitzades per mesurar l'error d'un classificador binari són:

- La ràtio de falsos positius (FPR) és la proporció d'errors en la classe negativa:

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

- La ràtio de vertaders positius (TPR), també conegut com a *exactitud*, és la proporció d'instàncies de la classe positiva que s'han classificat correctament:

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

- La precisió o PPV és la ràtio de classificacions correctes d'entre totes les instàncies que s'han classificat com a positives:

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

- El NPV és la ràtio de negatius vertaders d'entre totes les instàncies que s'han classificat com a negatives:

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

No s'han de confondre el PPV i el TPR. Observeu que en el denominador en el cas del PPV tenim les decisions que *s'han classificat* com a positives i no pas les que realment ho són.

1.3. Exemples de les causes de discriminació

A continuació us presentem alguns exemples que Hardt dona sobre les fonts de biaix en aprenentatge supervisat.

Lectura complementària

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

Exemple 1: les dades són un mirall de la societat

El conjunt d'entrenament és un recull de mostres que reflecteix els biaixos de la nostra societat. Si, per exemple, un dels grups de persones present en el conjunt d'entrenament és un grup que històricament ha estat víctima d'una discriminació sistemàtica (per exemple, la població negra als EUA o la població femenina arreu del món), les dades presentaran característiques diferenciadores per aquests grups (com ara un nivell d'educació més baix en la població negra o una diferència salarial entre sexes). A l'hora de prendre decisions, els algorismes descobriran aquests patrons i possiblement prendran decisions diferents per aquests grups. Imaginem per exemple el classificador de selecció de personal entrenat en dades on un grup de la població té qualificacions més baixes de mitjana. El classificador pot determinar que la pertinença a aquest grup és indicatiu d'un nivell educatiu més baix i per tant d'un rendiment més baix per portar a terme una tasca.

Significats de *biaix*

És important no confondre diferents significats de la paraula *biaix*. En aquest text ens referim majoritàriament a un biaix *social* i no pas al biaix en el sentit estadístic.

No és suficient, però, eliminar les característiques de les instàncies que identifiquen el grup (per exemple, sexe, raça, ètnia, orientació sexual, etc.). El problema és que altres característiques que són rellevants per a la tasca poden estar fortament correlacionades amb aquests atributs de la identitat d'una persona. Per exemple, és comú que les minories es trobin segregades en barris i l'adreça és una característica important a l'hora de demanar una assegurança. El sexe també es troba codificat en qualsevol conjunt de característiques que sigui suficientment ric (per exemple, l'alçada i el pes), encara que no hi sigui present explícitament. Aquestes característiques es troben latents i no es pot culpar l'algorisme supervisat per descobrir-les; al cap i a la fi, el propòsit d'aquests algorismes és descobrir patrons no trivials en les dades.

És clar que quan un atribut de la identitat dels individus no és important per a la tasca, aquest atribut no hauria de representar cap paper en les decisions. Per exemple, per a la selecció de personal per a un lloc de treball com a desenvolupador de programari, el sexe és completament irrellevant i el que compten són les qualificacions dels candidats. Ara bé, es pot justificar legalment que per a alguns llocs de feina es requereixi una certa alçada mínima, la qual cosa afavoreix els candidats de sexe masculí perquè solen ser més alts de mitjana. Determinar quan és acceptable fer aquestes distincions és sovint l'objecte de discussió en processos judicials per discriminació laboral.

Lectura complementària

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

Exemple 2: la disparitat en el nombre de mostres

Fins i tot quan les dades no tenen cap correlació amb atributs protegits (per exemple, sexe, orientació sexual, raça, creences religioses, etc.), els algorismes d'aprenentatge supervisat poden tenir biaixos en la mida i la representativitat de les mostres dels grups. Un

dels principis de l'aprenentatge supervisat és que es pot reduir l'error dels models com més gran i més representatiu sigui el conjunt d'entrenament que s'utilitza per entrenar-los. Quan les dades no tenen un balanç entre els grups que les componen, aquest principi es tradueix en un error desigual entre grups.

Aquest biaix afecta especialment les minories. Això és perquè, per definició, és més difícil obtenir mostres de les minories –ja que estan menys representades en la societat. Com a conseqüència, els models entrenats amb dades on les minories estan mal representades (per exemple, no es recull el nombre suficient de mostres) presentaran un error més elevat que per a la majoria.

Vegem la figura 1 per il·lustrar aquest efecte.

Figura 1

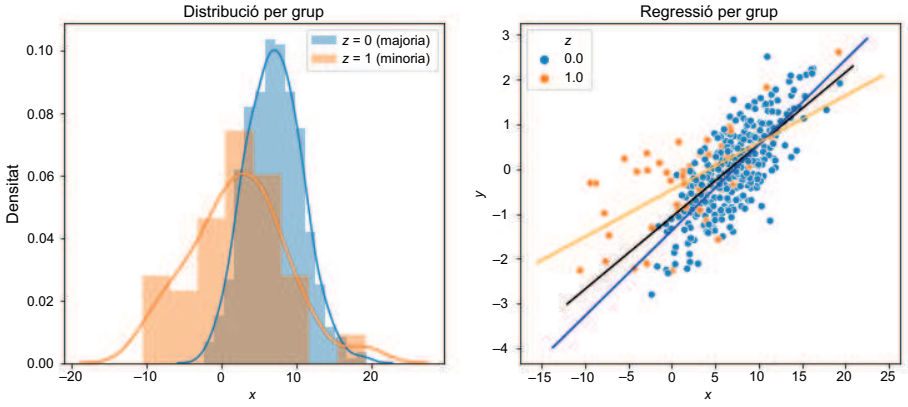


Figura 1

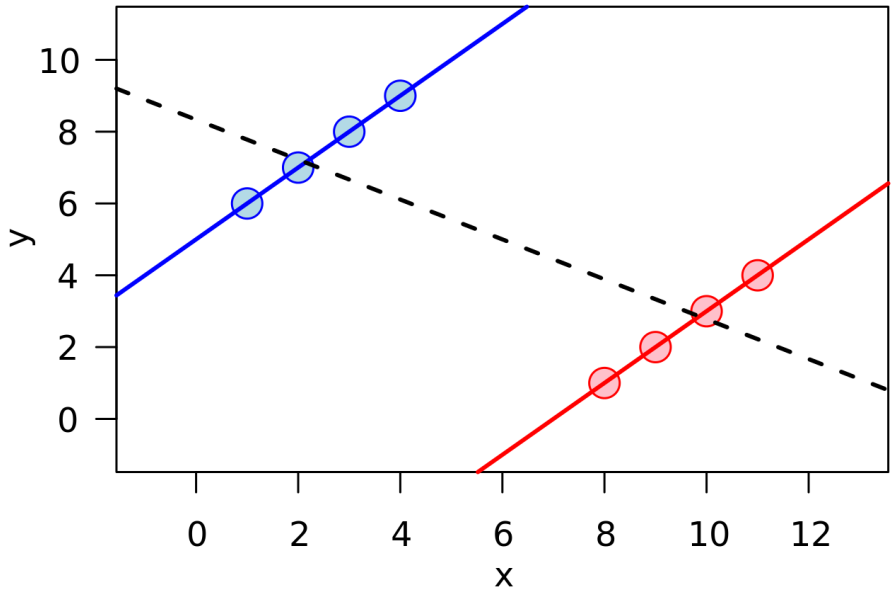
Hem generat dades sintètiques per a dues poblacions (una majoria i una minoria) on les dades de la minoria representen una fracció petita del total. Com veiem en la figura de l'esquerra, les poblacions tenen distribucions diferents. En la figura de la dreta hem ajustat una línia de regressió per a cada un dels subgrups. La línia negra és la regressió per a la població, i la taronja i la blava són les línies de regressió per a la minoria i la majoria, respectivament. Com veiem, la línia de la majoria té un error menor respecte a la de la població (això es pot apreciar per la desviació de la línia del subgrup respecte a la línia de la població).

Exemple 3: els patrons depenen de diferències culturals

Els problemes anteriors poden agreujar-se per diferències culturals entre els grups que conformen les dades. Hardt dona l'exemple d'un classificador dissenyat per detectar noms d'usuari reals (per exemple, Facebook demana l'ús del nom real de la persona durant la creació del compte). En alguns grups ètnics la diversitat de noms és molt més gran que en cultures occidentals on hi ha poca variància entre noms. En cultures occidentals també predominen noms curts i, en altres, els noms són de mitjana més llargs. La unicitat i una major longitud del nom són característiques que un classificador amb un biaix en l'entrenament podria detectar erròniament com a característiques identificatives de noms falsos. Així, no es tracta només de la disparitat en la mida dels conjunts d'entrenament, sinó també de les diferències entre els patrons identificatius entre els grups.

De fet, un fenomen estadístic que ha pres rellevància amb la discriminació dels algorismes d'aprenentatge supervisat demostra que dues variables que estan correlacionades positivament en la població poden tenir una correlació negativa en els subgrups de la població(!). Aquesta paradoxa es coneix com a paradoxa de Simpson i està exemplificada en la figura 2. En la figura es representen els valors de les variables X i Y per a dos grups (vermell i blau). Quan mesurem la correlació dels grups per separat hi ha una correlació lineal positiva, mentre que en la població la correlació és negativa.

Figura 2. Crèdit per Schutz, amb llicència de domini públic



Un exemple famós de dades on es dona la paradoxa de Simpson és en l'adjudicació de places a la Universitat d'UC Berkeley, Califòrnia, de l'any 1973. A les dades agregades es va observar que el percentatge de sol·licitants de sexe masculí acceptats va ser més elevat que els de sexe femení. No obstant això, quan es dividien les dades per facultat, s'invertia el biaix respecte al sexe dels sol·licitants. De fet, en les estadístiques agafades per cada facultat s'observa un biaix estadísticament significatiu en favor dels sol·licitants de sexe femení.

Lectura complementària

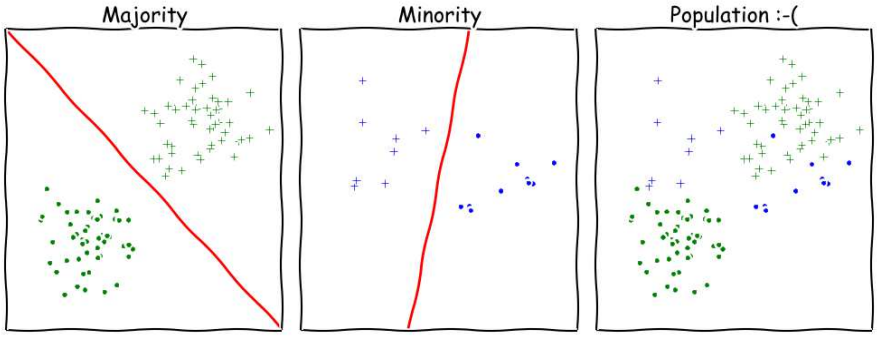
P. J. Bickel i altres (1975). «Sex bias in graduate admissions: Data from Berkeley», *Science*, vol. 187, pàg. 398-404.

Per resoldre aquest problema, es pot entrenar un classificador per a cadascun dels subgrups. Això, però, requereix que el classificador distingeixi entre els subgrups i que per tant utilitzi els atributs protegits explícitament, la qual cosa està prohibida en general per la majoria de lleis antidiscriminació.

A més a més, encara que distingim per subgrup, definir els subgrups és un repte en si mateix ja que aquests grups són construccions socials: com definim una raça o una ètnia? Com definim els grups per orientació sexual dins de l'ampli espectre de les definicions de la comunitat LGBTQ+? I com definim els subgrups derivats de les interseccions entre aquests grups? Necessitem un classificador per cada possible intersecció?

Un problema que ens podem trobar si no fem la distinció entre subgrups, però, és que tot i que hi hagi classificadors simples per a cada un dels subgrups, el classificador que resol la tasca quan els grups es combinen pot ser complex. Hardt exemplifica com, malgrat que els subgrups es poden separar amb funcions lineals (f és una funció lineal), la població no es pot separar linealment (vegeu la figura 3).

Figura 3. Models per la majoria i la minoria considerats independentment i en conjunt. Els models (f) estan representats en vermell. Crèdit per Moritz Hardt

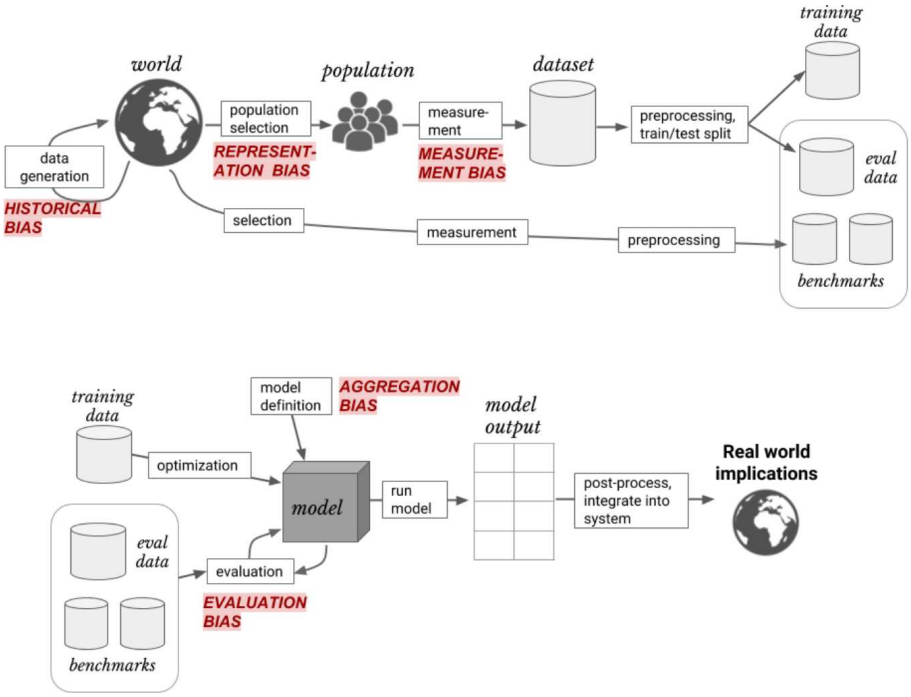


Aquest apartat no pretén ser una recopilació exhaustiva de totes les fonts de biaix en un model d'aprenentatge supervisat. De fet, es pot introduir biaix en cada fase del procés, des de l'inici fins al final. Per exemple, es pot introduir biaix en un error dispar en els mesuraments durant la recollida de les dades o fins i tot en l'avaluació del model (vegeu la figura 4).

Lectura complementària

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

Figura 4. Possibles biaixos en diferents parts del procés d'aprenentatge supervisat. Crèdit per Harini Suresh



2. Definicions de biaix algorísmic

Un dels problemes que més atenció ha rebut per part de la comunitat acadèmica és la formalització matemàtica de nocions d'equitat dels algorismes. A la literatura s'han proposat una multitud de definicions sobre què és que un model sigui «just». En els següents subapartats veurem les definicions més esteses i alguns dels resultats fonamentals als quals s'ha arribat.

2.1. Equitat de grups

Per les definicions d'equitat de grups, o equitat grupal, que donem en aquest mòdul ens hem basat en el llibre de text de Barocas, Hardt i Narayanan.

Lectura complementària

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

Primer, introduïm notació per definir les tres categories principals d'equitat algorísmica. Recordem que Y és la variable que f intenta predir, $\hat{Y} = f(X)$ és la variable aleatòria de les prediccions i anomenem A l'*atribut protegit* (per exemple, sexe, ètnia, orientació sexual), que pot estar o no inclòs en les característiques de les instàncies (X) amb les quals s'ha entrenat el model. En aquest subapartat ens centrarem en atributs protegits binaris i anomenarem grup privilegiat i grup desfavorit els grups definits per A que tenen un avantatge o un desavantatge, respectivament, en les decisions. A més a més, per simplicitat, ens centrarem en models que són classificadors binaris, però és fàcil extrapolar les definicions que donem a regressors.

Tot i que hi ha desenes de definicions d'equitat grupal diferents, la majoria es poden caracteritzar en tres tipus de relació de dependència entre $\hat{Y}$, Y i A: independència, suficiència i separació. Moltes de les definicions que trobem a la literatura són, de fet, una relaxació d'aquestes condicions de dependència.

2.1.1. Independència

La condició d'independència és que l'atribut protegit sigui independent de les sortides del model:

Definició (*independència*): f gaudeix d'independència si les variables aleatòries $\hat{Y}$ i A són independents, i.e., $\hat{Y} \perp A$.

La condició d'independència rep diversos noms en la literatura sobre equitat algorísmica: *paritat demogràfica, paritat estadística, equitat entre grups*, entre d'altres. En el cas de classificació binària i considerant només dos grups, la condició d'independència es pot expressar en termes probabilístics de la següent forma:

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

és a dir, la probabilitat d'acceptar un individu és independent del valor del seu atribut protegit. I, pel que fa a les prediccions d'error, aquestes probabilitats es poden mesurar en termes de les freqüències de *TP*/*Total* per cadascun dels grups.

Aquesta és una definició simple i, a més, és natural: les decisions del model no haurien de dependre d'atributs protegits com el sexe o la raça. Malgrat tot, demanar que es compleixi independència a la pràctica pot tenir efectes que no són desitjables. Com Dwork i altres il·lustren, donats dos grups d'interès, una empresa pot satisfer aquesta condició contractant el mateix nombre de persones per cada grup però amb un procés de selecció diferent: mentre que els candidats del grup privilegiat se seleccionen diligentment, els del grup desfavorit se seleccionen aleatòriament. Com a resultat, les estadístiques mostraran una diferència del rendiment entre els dos grups. Aleshores, l'empresa pot justificar diferències en els percentatges de selecció entre grups per protegir els interessos del negoci.

Això pot passar sense que l'empresa sigui malintencionada. Pot ser que, per exemple, a causa que l'empresa tradicionalment ha contractat persones del grup privilegiat, coneixen millor el grup i saben quines són les seves característiques rellevants i, per tant, poden fer un procés de selecció més acurat. Aquest punt és important perquè sovint és el cas que els grups tenen característiques diferents respecte a la tasca, si no, no hi ha cap justificació per no fer servir el mateix procés de selecció i seleccionar el mateix nombre de candidats dels dos grups.

Per tal de resoldre aquest problema s'ha proposat la condició de separació.

2.1.2. Separació

La condició de separació diu que $\hat{Y}$ i A són independents en la mesura que Y ho permeti:

Definició (*separació*): f satisfà la condició de separació si les variables aleatòries $\hat{Y}$ i A són condicionalment independents respecte a Y, i.e., $\hat{Y} \perp A \mid Y$.

És a dir, el model i l'atribut protegit poden tenir una relació de dependència en tant que ho justifiqui la tasca. Aquesta definició reconeix que hi ha moltes tasques en les quals l'atribut protegit està correlacionat amb la tasca. Per exemple, en l'exemple de concedir un crèdit bancari, pot ser que un dels grups tingui una probabilitat més alta de no poder retornar el crèdit. Un banc pot justificar legalment, en aquest cas, que fer una distinció entre els grups és una necessitat per a la sostenibilitat del negoci.

En el cas que f sigui un classificador binari, aquesta condició es pot expressar en termes probabilístics de la següent manera:

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

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

És a dir, la probabilitat d'un positiu vertader (primera equació) o d'un fals positiu (segona equació) és independent del grup. Es pot mesurar si aquestes condicions se satisfan comprovant que el TPR i el FPR dels grups són iguals. A la pràctica, que el TPR i el FPR siguin iguals vol dir que s'accepten candidats *que s'haurien* d'acceptar perquè estan qualificats i que es descarten candidats que s'haurien d'acceptar en la mateixa mesura en ambdós grups.

La condició de separació també es coneix popularment a la literatura com a *Equalized Odds* (igualtat de possibilitats o oportunitats).

2.1.3. Suficiència

Per últim, el tercer criteri d'equitat exigeix que la tasca i l'atribut protegit siguin independents respecte a la decisió:

Definició (*suficiència*): f satisfà la propietat de suficiència si les variables aleatòries Y i A són condicionalment independents respecte a $\hat{Y}$, i.e., $Y \perp A \mid \hat{Y}$.

Una forma d'interpretar aquesta condició és que per a una decisió fixada (per exemple, acceptar un candidat), la tasca és independent del grup: en cada grup hi trobem el mateix nombre de persones que haurien de ser acceptades. Per a un classificador binari, aquesta condició es pot escriure en termes probabilístics de la següent manera:

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

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

Aquestes probabilitats es poden mesurar mitjançant el PPV i el NPV del classificador, respectivament.

A la literatura, a vegades es fa referència a la *calibració per grups* per parlar de suficiència. Això és degut al fet que en aprenentatge supervisat «calibrar» vol dir ajustar el model perquè doni estimacions acurades de les probabilitats de les seves sortides o decisions. S'ha demostrat que fer aquesta calibració per cada grup implica suficiència –el recíproc, en general, no és cert ja que la suficiència és una condició més feble que la calibració per grups.

La diferència entre suficiència i separació és subtil i pot ser difícil de distingir. Una manera de diferenciar aquestes condicions és pensar en la diferència entre TPR i PPV: mentre que TPR és el nombre d'encerts sobre els que *s'haurien* d'acceptar (TP + FN), PPV és el nombre d'encerts sobre els que el classificador *ha* acceptat (TP + FP). En l'exemple de selecció de personal, mentre que la suficiència imposa que s'acceptin el mateix percentatge de candidats aptes sobre el total de candidats aptes, la separació imposa que el percentatge de candidats aptes sobre el total de candidats que s'han acceptat siguin els mateixos entre grups. Com veurem en el següent exemple, la segona condició depèn de com de fàcil o difícil és trobar candidats aptes en cada grup.

Exemple: un alt TPR no implica un alt PPV

Per exemple, imaginem un escenari hipotètic on 5 de cada 1000 candidats del grup desfavorit són aptes. Un classificador que presenta un alt TPR «caça» 4 d'aquests 5 candidats, i.e., $TPR = 80\%$. Això no té en compte, però, quants candidats no aptes ha acceptat (FP). Ara bé, quan mirem el nombre d'FP que el classificador ha comès, com que hi ha molts més no aptes que aptes –995 de cada 1000 no són aptes–, tot i que la probabilitat de cometre un FP sigui baixa, en nombres absoluts el nombre d'FP pot ser ordres de magnitud més gran que el de TP. Com a conseqüència, el PPV serà molt més baix que el TPR, encara que el TPR sigui elevat. Si en l'exemple també suposem que el classificador té un baix FPR, per exemple, $FPR = \frac{20}{995} \simeq 2\%$, el nombre de *FP* és 20, i per tant, $PPV = \frac{4}{4+20} \simeq 16\%$.

El TPR i el PPV són només un parell de les moltes mètriques que hi ha per mesurar l'error d'un model; n'hi ha moltes més i es podrien proposar definicions d'equitat per cadascuna d'elles. El problema és que tot i que totes aquestes definicions són desitjables –idealment, cap dels grups hauria de patir un error més gran que un altre– s'ha demostrat que ni tan sols les que hem donat es poden satisfer alhora.

2.1.4. Teoremes d'impossibilitat

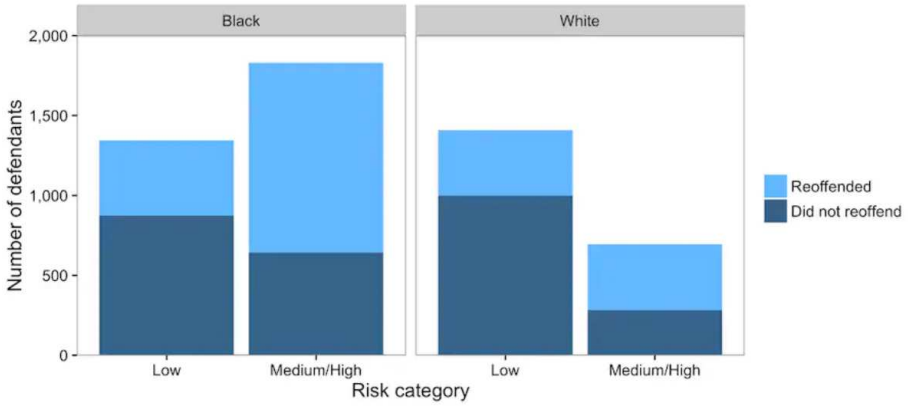
Els resultats d'impossibilitat van estar fortament motivats pel cas de COMPAS (les sigles en anglès de *Correctional Offender Management Profiling for Alternative Sanctions*). COMPAS és un model entrenat amb tècniques d'aprenentatge supervisat utilitzat en alguns comptats dels Estats Units. Actualment, s'utilitza per predir la probabilitat de reincidència d'un pres i com a evidència per als jutges que concedeixen sol·licituds de llibertat condicional.

Exemple: el cas de COMPAS

L'any 2016, ProPublica, una redacció de periodisme d'investigació, va demanar accés a les decisions de COMPAS fetes a Broward County, Florida –en aquest estat, aquesta informació està disponible si es demana–, i la van complementar amb el registre criminal actualitzat d'aquelles persones que havien estat subjectes a una decisió. Amb aquestes dades, ProPublica va fer una auditoria independent de COMPAS amb l'objectiu de determinar si les decisions tenien un biaix respecte a l'ètnia, el sexe o la raça dels presos. Els resultats de la investigació mostren un biaix significatiu contra els americans d'ascendència africana. En particular, la investigació demostrava que membres d'aquesta població tenien el doble de probabilitats de ser identificats «d'alt risc» quan no acabaven reincidint i, viceversa, els presos blancs tenien més probabilitats de ser identificats de baix risc i acabar cometent un altre crim.

Equivant, aleshores coneguda com a Nothpointe, l'empresa que va dissenyar i produir COMPAS, no va trigar a respondre a les acusacions de ProPublica, defensant l'equitat de les decisions de COMPAS. L'informe d'Equivant proporciona evidència estadística que demostra que COMPAS prediu reincidència correctament en la mateixa proporció entre tots els grups.

Figura 5. Distribució dels presos per categories de risc i raça en el conjunt de dades de COMPAS analitzat per ProPublica



Font: *The Washington Post*

Aquest debat públic va rebre molta atenció dels mitjans de comunicació i dels acadèmics a causa de les seves implicacions sobre la parcialitat del sistema judicial i penitenciari. Sorprenentment, es va demostrar que ambdues parts tenien raó. Simplement, cada part s'acollia a una definició d'equitat diferent: mentre que ProPublica estava auditant la condició de separació en COMPAS, Equivant havia dissenyat COMPAS per garantir la condició de suficiència. En termes de les ràtios d'error, ProPublica havia demostrat que els FPR eren diferents entre els grups i Equivant havia *calibrat* COMPAS per tal que el PPV fos igual entre grups. Les dues perspectives es poden apreciar en la figura 5. La lectura que ProPublica fa de la figura és que COMPAS identifica persones de raça negra que no han reincidit com a alt risc amb més freqüència (diferència de l'àrea de color blau fosc a *medium/high* entre grups). Malgrat això, quan fixem un valor de $\hat{Y}$ (*low* o *medium/high*), obtenim que les proporcions de reincidents són les mateixes per cada grup i, per tant, se satisfà la condició de suficiència que defensa Equivant.

Lectura complementària

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

Encara més, s'ha demostrat matemàticament que *totes* les definicions d'equitat dels subapartats anteriors són mútuament exclusives: cap parella de definicions es pot satisfer alhora –excepte en casos molt concrets i poc realistes. En aquest subapartat no veurem les demostracions d'impossibilitat per totes les parelles de condicions, ho veurem per les més senzilles i per altres donarem la intuïció darrere de la seva incompatibilitat.

Incompatibilitat entre independència i suficiència

Veurem que independència i suficiència no es poden satisfer alhora si l'atribut protegit és independent de la tasca. Aquest serà un cas comú, ja que si estem considerant casos d'equitat és perquè els grups no són homogenis envers la tasca i hi ha certes diferències que s'han de tenir en compte.

Per exemple, en el cas de COMPAS, és clar que els grups no estan distribuïts de la mateixa manera respecte de la tasca. La població negra ha estat històricament discriminada i encara pateix un tracte diferencial en molts àmbits de la societat. Hi ha estudis que demostren com la població negra està sotmesa a més arrests injustificats que els blancs. El racisme sistèmic i altres factors relacionats com la segregació i la pobresa han resultat en un major nombre de presos negres als Estats Units.

Aleshores, per demostrar la incompatibilitat entre independència i suficiència fem una reducció a l'absurd. Primer suposem que es compleixen totes dues i arribem a una contradicció amb la suposició que Y i A no són independents. Això és per les propietats de contracció i descomposició de la independència condicional: $A \perp \hat{Y}$ (independència) i $A \perp Y \mid \hat{Y}$ (suficiència) implica que A és independent de la variable conjunta $(Y,\hat{Y})$ i, per tant, $A \perp Y$ (contradicció).

Independència

Compte a no confondre la independència probabilística entre les variables aleatòries i la condició d'equitat que també anomenem d'independència.

Incompatibilitat entre separació i independència

La demostració en aquest cas és més complicada i, a més de $A \not\perp Y$, també requereix la suposició que $\hat{Y}$ sigui independent de Y. Aquesta suposició no és una suposició gaire forta, ja que qualsevol classificador útil hauria de tenir una correlació amb la tasca (si no, no serveix per resoldre la tasca).

No farem la demostració aquí, però consisteix a demostrar el contrapositiu: si $A \perp \hat{Y}$ i $A \perp \hat{Y} \mid Y \Rightarrow$ o bé $A \perp Y$, o bé $\hat{Y} \perp Y$.

Incompatibilitat entre suficiència i separació

Aquesta és la incompatibilitat entre les dues definicions d'equitat en el debat de COMPAS. Kleinberg i altres i Chouldechova van demostrar independentment i al mateix temps que és impossible construir un classificador que compleixi amb les dues definicions d'equitat, a no ser que els grups tinguin les probabilitats *a priori* o els FPR i els FNR siguin zero (un classificador perfecte és molt rar a la pràctica). Per tant, aquí també suposem que $A \not\perp Y$.

Lectura complementària

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

La demostració de la incompatibilitat entre separació i suficiència es basa en el teorema 17.2 de Wasserman, el qual enuncia:

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

i, a més,

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

La demostració consisteix a prendre el contrapositiu de les implicacions anteriors:

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

Tornant al cas de COMPAS, ja que no es poden satisfer les dues condicions d'equitat perquè la distribució de presos per grups no és uniforme, una pregunta a fer-nos és: quina definició d'equitat és més important que es compleixi en aquest cas? Equivant es decanta per la suficiència, ja que com que el risc significa el mateix independentment de la raça, un jutge no ha de considerar la raça del pres a l'hora d'interpretar els resultats de COMPAS. Per altra banda, ProPublica argumenta que és molt greu que no se satisfaci separació ja que tot i que aquells presos negres que van ser identificats d'alt risc no van acabar cometent crims, sí que van ser subjectes a un major escrutini per part del sistema penitenciari, la qual cosa no és justa en comparació als blancs que no van reincidir. Potser una pregunta encara més rellevant: creieu que és ètic utilitzar COMPAS si no es poden satisfer les dues propietats? Creieu que és ètic utilitzar COMPAS fins i tot si se satisfan aquestes propietats?

2.2. Equitat individual

La noció d'equitat individual recorda les màximes tradicionals de la justícia: «casos semblants haurien de ser tractats de manera semblant».

A diferència de les nocions d'equitat grupal anteriors, l'equitat individual es defineix normalment sobre les distribucions de probabilitat de f, és a dir, no pas la decisió final de cada instància, sinó la distribució de probabilitat de les possibles decisions. En l'exemple de donar un crèdit, les probabilitats de sortida de f podrien ser (0.7,0.3), és a dir, amb probabilitat 0.7 es concedeix el crèdit i amb 0.3 es rebutja la sol·licitud. Denotem per $\hat{f}$ el classificador que retorna les distribucions de probabilitat d'f.

Definició (*equitat individual*): sigui D una mesura de divergència entre distribucions de probabilitat de les sortides de f i sigui d una mesura de distància entre individus. Diem que f és individualment equitatiu si, per a tota parella d'individus u i v, satisfà: $D(\hat{f}(u),\hat{f}(v)) \leq d(u,v)$.

És a dir, les diferències entre dos individus limiten les diferències entre els tractaments que reben per f. És important que aquestes mesures de distància entre individus només tinguin en compte característiques rellevants per la tasca o aplicació en qüestió i que no tinguin en compte atributs protegits.

Dwork i altres han proposat mecanismes d'entrenament que produeixen models f que són individualment equitatius. Com veurem en l'apartat següent,

la idea és restringir el problema d'optimització de l'entrenament de manera que satisfaci la condició d'equitat individual definida anteriorment.

Un dels inconvenients de la definició d'equitat individual és que, a diferència de les definicions grupals, ens cal trobar una mesura de distància entre individus que sigui apropiada per la tasca. Aquesta mesura de distància, òbviament, haurà d'ignorar els atributs protegits –atributs com el sexe o la raça haurien de ser irrellevants per comparar dos individus respecte a la tasca–, però, a més, ha de ser *útil* respecte a la tasca. Per exemple, per la tasca d'adjudicació de places d'una universitat, podem mesurar el valor absolut entre les notes dels sol·licitants a les PAU. Aquesta distància, però, no serviria per distingir entre sol·licitants que sol·liciten titulacions diferents, ja que cada titulació té una nota de tall diferent i pot tenir requisits diferents en cadascun dels exàmens individuals de la selectivitat: per exemple, per entrar a la carrera de matemàtiques, dos estudiants són semblants si han tret una nota semblant a l'examen de matemàtiques de la selectivitat. Com veieu, en general, definir aquesta mesura de distància no és trivial i requereix l'experiència d'un expert en la tasca.

2.3. Causalitat

Les definicions anteriors són totes definicions sobre com es comporten els models, també anomenades observacionals. Tot i que són fàcils de comprovar en un model (com ara la separació d'un classificador es pot mesurar amb el TPR i el FPR), no capturen les causes de la discriminació. Per exemple, en el cas de COMPAS sabem que no es compleix la condició d'independència però no som capaços de determinar la causa de les diferències entre els grups només a partir dels resultats del model. Hi ha marcs teòrics per modelar les relacions de causalitat en un sistema (per exemple, xarxes causals i xarxes de creences). Amb models causals podem raonar i atribuir les causes de la discriminació.

Nota

Aquí ens referim a models matemàtics de causalitat i no pas a models d'aprenentatge automàtic.

Lectura complementària

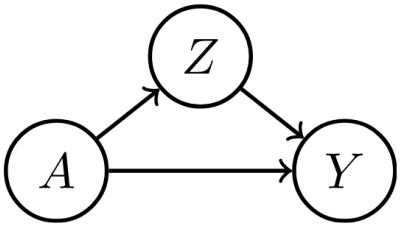
S. Chiappa; W. S. Isaac (2019). *A Causal Bayesian Networks Viewpoint on Fairness*. ArXiV.

Per exemple, imaginem el cas de l'adjudicació de places a UC Berkeley que hem donat com a exemple per explicar la paradoxa de Simpson en el primer apartat. A la figura 6 mostrem un model gràfic d'una xarxa causal per a l'adjudicació de places a UC Berkeley, on les variables rellevants són: el sexe (A), la facultat escollida (Z) i la decisió d'admissió (Y). Les fletxes indiquen la relació de causalitat entre les variables: el sexe influeix l'elecció de la facultat (sol·licitants de diferent sexe escollien facultats diferents), la decisió depèn de la facultat (algunes facultats eren més estrictes que altres), i a més suposem que hi ha una relació de causalitat directa entre el sexe i la decisió (si fos així, hi hauria discriminació).

En l'estudi original, Bickel argumenta que no hi ha un efecte significatiu del sexe sobre la decisió que afavoreixi els sol·licitants de sexe masculí. Utilitzant el model anterior podem raonar sobre aquesta hipòtesi i preguntar-nos: què hauria passat si mantenint Z constant, haguéssim canviat el sexe dels

sol·licitants? Si observem una diferència significativa vol dir que A té un efecte directe sobre Y.

Figura 6. Un possible model de causalitat per l'adjudicació de places a UC Berkeley



Preguntes hipotètiques d'aquest estil, també anomenades contrafactuals, ens permeten determinar l'efecte d'una *intervenció*. En aquest cas, hem realitzat una intervenció en el model eliminant la relació entre A i Y mediada per l'elecció de la facultat i ens ha permès mesurar l'efecte directe d'A sobre Y. Aquesta és una propietat de les xarxes causals que és de gran utilitat per dissenyar lleis i polítiques antidiscriminació que redrecin casos de discriminació.

Malgrat tot, les respostes a aquests contrafactuals seran tan correctes com ho siguin els models de causalitat: per exemple, si definim relacions de causalitat que no tenen cap suport o passem per alt variables rellevants, podem arribar a conclusions incorrectes. De fet, és possible construir dos models que tenen una estructura idèntica i es comporten igual a les intervencions i, encara així, donen respostes diferents als contrafactuals.

2.4. Altres

En la llei també trobem definicions de justícia algorísmica que poden fer-nos pensar sobre les definicions anteriors. La majoria de lleis antidiscriminació de la Unió Europea protegeixen contra un tractament diferencial en àrees específiques com el món laboral i l'educació. Aquestes lleis s'anomenen de «discriminació directa» o de «tractament dispar» (en anglès: *disparate treatment*) i fan referència a una discriminació intencional d'un individu.

També hi ha lleis que regulen un tractament diferencial entre grups. Aquestes lleis intenten cobrir casos de discriminació sistemàtica a un grup de la població que no necessàriament són intencionats. Als Estats Units s'anomenen d'«impacte dispar» (en anglès: *disparate impact*) i a la UE s'anomenen de «discriminació indirecta».

La llei als Estats Units estableix un llindar que determina quan es produeix un impacte dispar: si un grup presenta un percentatge de membres que se seleccionen inferiors per més d'un 80% respecte del percentatge del grup amb el percentatge de selecció més alt, hi ha un impacte desfavorable per aquest grup. Malgrat tot, la llei només s'aplica si l'entitat que fa la selecció pot cor-

regir aquesta disparitat sense que afecti els interessos del seu negoci. És a dir, que les decisions les està prenent per necessitat de satisfer els requeriments de la tasca. En l'exemple de selecció de personal, aquesta llei no s'aplica si una empresa demostra que selecciona els candidats més preparats, malgrat que per una discriminació històrica pot haver-hi grups que de mitjana tinguin un nivell d'educació més baix i, per tant, tinguin percentatge de selecció més baix.

Barocas apunta com aquests dos tipus de lleis sovint es troben en conflicte. El problema és que garantir que individus semblants es tractin de manera semblant tal com dicten les lleis de tractament dispar pot perpetuar les desigualtats entre grups que les lleis d'impacte dispar intenten corregir. Una opció entremig és tractar individus que són aparentment dissimilars de maneres similars si la diferència es dona per una discriminació al grup al qual pertany l'individu menys afavorit (discriminació positiva). De fet, moltes de les definicions que hem vist en aquest apartat es troben en algun punt de l'espectre continu entre les definicions de tractament i impacte dispars (per exemple, l'equitat individual s'apropa al tractament dispar i la condició d'independència s'apropa a l'impacte dispar).

Lectura complementària

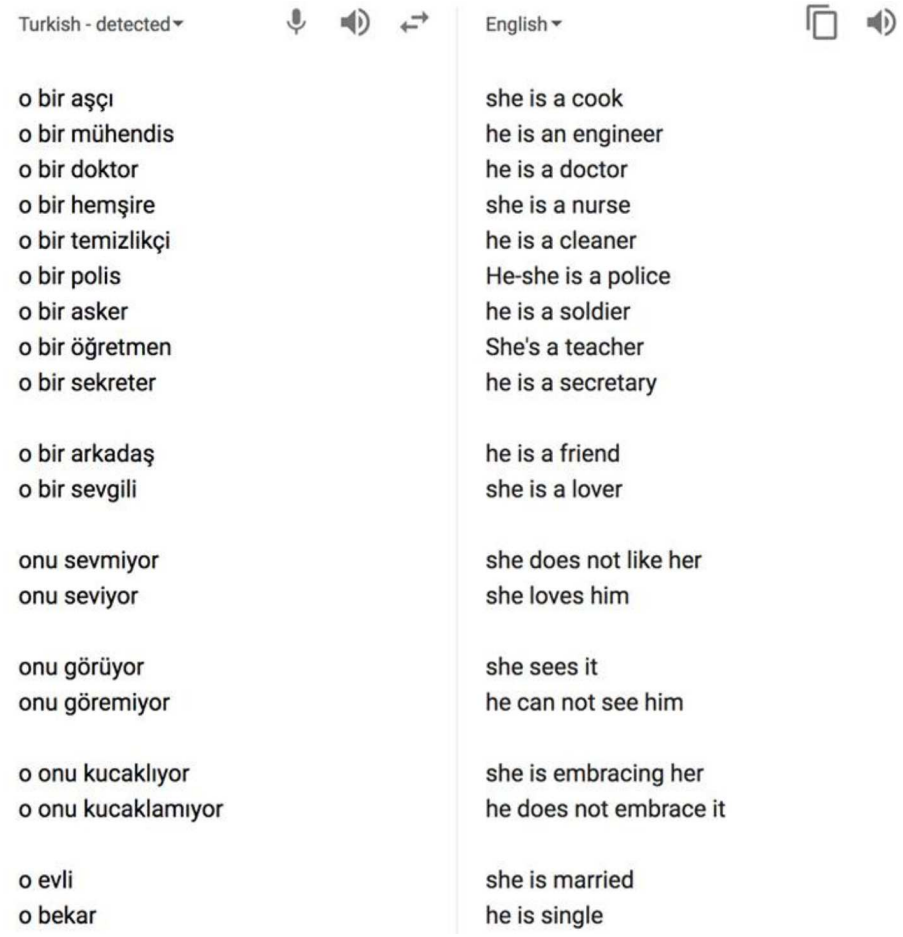
S. Barocas; M. Hardt (2017). *Fairness in Machine Learning NIPS 2017 Tutorial*. https://mrtz.org/nips17

Crawford fa una observació important sobre aquestes definicions: gairebé totes parlen d'equitat en termes de repartiment d'oportunitats i recursos (mals d'assignació); en canvi, hi ha un altre tipus de desigualtat que pot fer igual de mal: la falta d'equitat en la representació dels grups en la societat (mals de representació). Crawford defineix els mals de representació com els mals relacionats amb la perpetuació d'estereotips de grups de la societat que no estan suficientment representats en la societat. Alguns exemples d'aquests mals que han cridat molt l'atenció són els estereotips presents en els *word embeddings*, les traduccions esbiaixades del traductor de Google, i els errors de l'algorisme d'etiquetatge de Google Photos.

Els *word embeddings* són un conjunt de tècniques de processament del llenguatge on les paraules i frases són representades com a vectors en un espai multidimensional. Una de les motivacions d'aquest model era poder respondre enunciats lògics de l'estil: *rei – home + dona = reina*. Bolukbasi i altres van demostrar que la interfície *word2vec* utilitzada per construir *word embeddings* heretava i fins i tot amplificava els estereotips, establint relacions lògiques de l'estil: «home és a programador el que dona és a mestressa».

Caliskan i altres van estudiar les traduccions de Google Translate entre llenguatges amb gènere gramatical i sense. En l'estudi van utilitzar l'anglès i el turc; aquest últim és un llenguatge sense gènere. Quan traduïen frases sense gènere el traductor afegia un gènere estereotipat a la traducció. Així, els enginyers i els doctors eren sempre en masculí i les infermeres i les mestres d'escola eren en femení (vegeu la figura 7).

Figura 7. L'addició de gènere en les traduccions del turc a l'anglès



Font: Emre Şarbak

Per últim, un altre escàndol va ser l'errada de l'algorisme d'etiquetatge de Google Photos, el qual va etiquetar fotos de persones de raça negra com a «goril·les». Els enginyers de Google es van afanyar a treure l'etiqueta «goril·les» de les possibles etiquetes de l'algorisme per solucionar el problema almenys de manera provisional.

Existeix un cercle viciós entre els mals d'assignació i els mals de representació: un grup que rep menys oportunitats i recursos a la llarga patirà mals de representació, els quals, a la vegada, poden induir una discriminació en els algorismes d'aprenentatge utilitzats per resoldre les assignacions. En essència, el cas COMPAS és un mal de representació que es tradueix en un mal d'assignació.

3. Construcció de models equitatius

En aquest apartat fem un repàs dels mètodes que s'han proposat per construir models que satisfacin les propietats d'equitat definides a l'apartat anterior. Els mètodes es distingeixen segons la fase del procés d'aprenentatge automàtic en la qual s'apliquen: preprocessament, si s'apliquen en el conjunt d'entrenament; durant l'entrenament, com a restriccions en la tasca d'optimització de l'entrenament; i com a postprocessament, ajustant el model que ja s'ha entrenat.

La majoria d'aquests mètodes incorren en un increment de l'error del model en resoldre la tasca. Sovint l'entitat que pren les decisions haurà de trobar un compromís entre satisfer un grau d'equitat –per alguna de les definicions d'equitat– i una precisió i exactitud desitjades del model. Depenent de la tasca, les dades, el mètode de correcció i la definició d'equitat, aquest compromís serà més o menys fàcil d'assolir.

En aquest apartat explicarem i donarem exemples de cadascun d'aquests mètodes.

3.1. Preprocessament

Les tècniques de preprocessament suposen que les dades són la principal font de biaix. Per tant, aquestes tècniques implementen modificacions de les dades abans de l'entrenament per tal que el model final satisfaci alguna definició d'equitat.

La majoria d'aquestes tècniques formulen el problema com una tasca d'optimització on s'intenta trobar una representació de les dades que minimitzi la discriminació (segons alguna noció d'equitat), minimitzi la distorsió en les observacions individuals, i minimitzi l'error en el model final. Aquestes noves representacions s'aconsegueixen transformant el conjunt d'entrenament realitzant les següents operacions sobre les observacions:

1) assignar pesos a les observacions en el procés d'entrenament;

2) canviar les etiquetes de les observacions;

3) descartar característiques que estan correlacionades amb l'atribut protegit.

Depenent de com es defineixen els objectius del problema d'optimització, de les nocions d'equitat que s'intenten assolir i les operacions que són permeses, s'han definit diverses tècniques de preprocessament: «ajustament de pesos», «preprocessament optimitzat», «aprenent representacions equitatives», i «eliminació d'impacte dispar». I, recentment, s'estan aplicant tècniques avançades d'optimització que utilitzen xarxes antagòniques generatives (*generative adversarial networks*, GAN) per resoldre el problema d'optimització considerant un adversari que intenta extreure informació sobre l'atribut protegit.

Un dels avantatges del preprocessament és que és agnòstic a altres etapes del procés d'aprenentatge: per exemple, un cop s'ha trobat una representació equitativa de les dades, qualsevol altre tipus d'entrenament en aquest nou espai també satisfarà la definició d'equitat.

3.2. Correcció durant l'aprenentatge

Aquestes tècniques modifiquen el problema d'optimització que es resol durant l'entrenament del model per tal que el model final satisfaci una definició d'equitat. El procés per derivar un classificador que satisfà equitat individual proposat per Dwork i altres i que hem mencionat en el subapartat anterior és un exemple. Val a dir, però, que la propietat d'equitat individual es pot satisfer amb tècniques de preprocessament i postprocessament.

Dues de les tècniques de correcció durant l'entrenament més populars són:

- **Supressor de prejudicis**: afegeix un terme (anomenat de regularització) a la funció d'error que es minimitza en el procés d'entrenament per tal de minimitzar la discriminació del model. La regularització normalment es fa per millorar la generalització del model, en aquest cas es fa per reduir la discriminació.

- **Supressor de biaix antagònic**: s'afegeix com a objectiu en l'optimització per minimitzar l'habilitat d'un adversari d'inferir informació sobre l'atribut protegit a partir de les prediccions del model.

Les tècniques de correcció durant l'aprenentatge tenen el potencial d'aconseguir models amb baix error, ja que podem optimitzar el model amb la condició d'equitat integrada en el procés d'entrenament.

3.3. Postprocessament

Les tècniques de postprocessament prenen un model que ja ha estat entrenat i ajusten les seves sortides perquè se satisfaci una noció d'equitat. Normalment s'afegeix aleatorietat en les sortides amb aquest objectiu.

Anomenem un model derivat $\overline{Y} = F(\hat{Y},A)$ on F és una funció del model que volem corregir ($\hat{Y}$) i de l'atribut protegit (A) que possiblement conté aleatorietat. Donat un cert cost per FP i FN, la tasca és trobar un F que probabilísticament modifiqui les sortides de $\hat{Y}$, de manera que satisfaci alguna definició d'equitat i minimitzi el cost dels FP i FN que produeix, en esperança.

L'avantatge de les tècniques de postprocessament és que són agnòstiques al model o la família d'algorismes d'aprenentatge supervisat que s'utilitzin: no és necessari repetir el procés d'entrenament, la qual cosa pot ser molt útil quan el procés d'aprenentatge és complex. Si no tenim accés al model o les dades, i només tenim accés a les sortides, el postprocessament pot ser l'única opció viable. Aquests avantatges, per altra banda, són el que fan que les tècniques de postprocessament incorrin en increments substancials d'error en el model final.

Enllaç d'interès

Per entendre l'efecte i el funcionament dels mètodes presentats en aquest apartat recomanem jugar amb el *framework* d'IBM per corregir models AIF360: http://aif360.mybluemix.net/

4. Interpretació i explicabilitat

En aquest apartat definim la interpretabilitat i l'explicabilitat. A més, farem un repàs de les tècniques que es fan servir per donar més transparència als models d'aprenentatge automàtic.

Com hem introduït en el primer apartat, a causa de la popularització de les xarxes neuronals profundes, ha aparegut la necessitat de donar transparència a les decisions que recomanen els models. Exemples de models interpretables són: regressió lineal i logística, arbres de decisió, o classificadors de Bayes ingenus. Aquests models són comprensibles i intuïtius. Per exemple, un arbre de decisió permet saber quines característiques són rellevants en la decisió i, fins i tot, el pes de cada característica en la decisió. Aquestes dues propietats (rellevància i pes) sobre les característiques són exemples de propietats d'interpretabilitat desitjables en un model. En contrast, les xarxes neuronals profundes solen descobrir les característiques automàticament i, sovint, ni tan sols experts en la tasca són capaços d'interpretar-les.

Un model de *caixa negra* és un model que, o bé és massa complicat per entendre, o bé és propietari i per tant no podem saber com funciona. Per contra, els models *interpretables* són models que no són de caixa negra. Els models *explicables* són models de caixa negra que permeten ser interpretats *a posteriori*. Les xarxes neuronals profundes són un exemple de model de caixa negra, ja que la complexitat dels models és elevadament no lineal i per tant la interpretació del model escapa a la intuïció humana. En els subapartats següents veurem algunes tècniques per interpretar les sortides de models de caixa negra.

Rudin apunta que, sovint, la utilització de models interpretables no implica una disminució en el rendiment final del model. Rudin adverteix dels riscos d'utilitzar models de caixa negra i advoca per l'ús de models interpretables, fins i tot quan hi ha petites pèrdues de precisió o exactitud. Una de les raons és que els errors no només provenen del model sinó que també poden ocórrer a nivells més alts de la interacció amb el model i la presa de decisions; models que són interpretables permeten una millor comprensió del model per part de la persona que pren la decisió (per exemple, un metge que ha de recomanar un tractament) i per tant prevenen aquest tipus d'errors.

Lectura complementària

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

Malgrat que coincidim amb la visió de Rudin, en aquest apartat veurem algunes tècniques d'explicabilitat que s'estan utilitzant actualment. Com que no sembla que les xarxes neuronals profundes es deixin d'utilitzar en un futur proper, és important conèixer les tècniques que s'han proposat per entendre els resultats d'aquests models.

Per als següents subapartats ens basem en el llibre *Interpretable Machine Learning* de Christoph Molnar. En aquest mòdul ens centrarem en tècniques que s'utilitzen específicament per explicar xarxes neuronals, però en el llibre podeu consultar informació sobre tècniques d'interpretabilitat més generals.

Lectura complementària

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

4.1. Influència de les observacions

Una de les idees més potents per interpretar models complexos és mesurar l'efecte de les observacions en el model. Identificar les observacions que han tingut major *influència* i parar atenció a les característiques d'aquestes observacions pot ajudar a entendre què és el que determina els paràmetres del model i, fins i tot, decisions particulars. Per exemple, si els models depenen fortament d'una instància tenim raons per sospitar del model o de la instància (per exemple, la instància és de fet un error) i caldrà investigar-ho.

Per tal de determinar l'efecte d'una observació en el model s'elimina l'observació i s'entrena el model amb les observacions restants. L'efecte es mesura en els canvis als paràmetres del model. Aquesta idea no és nova i s'utilitza per detectar observacions atípiques en models de regressió des dels anys setanta. Mesures d'influència basades en l'eliminació d'observacions són la distància de Cook i DFBETA.

L'inconvenient de repetir l'entrenament per a cada instància és que és un procés que requereix molts recursos computacionals i pot ser inviable per models complexos (per exemple, xarxes neuronals) i amb grans conjunts de dades d'entrenament. En casos en els quals la funció d'error en el conjunt d'entrenament que s'optimitza durant l'entrenament té propietats desitjables (per exemple, és dos cops diferenciable respecte als paràmetres del model), podem fer servir funcions d'influència per mesurar l'efecte de les observacions en els paràmetres del model. La idea és aproximar la funció d'error al voltant dels paràmetres actuals del model utilitzant el seu gradient. A través del gradient podem mesurar com canvien els paràmetres del model quan introduïm petites pertorbacions en el conjunt d'entrenament. Aquest mètode es pot aplicar en general en xarxes neuronals. Un desavantatge, però, és que les funcions d'influència són aproximacions i sovint tindran un error en les estimacions d'influència de les observacions.

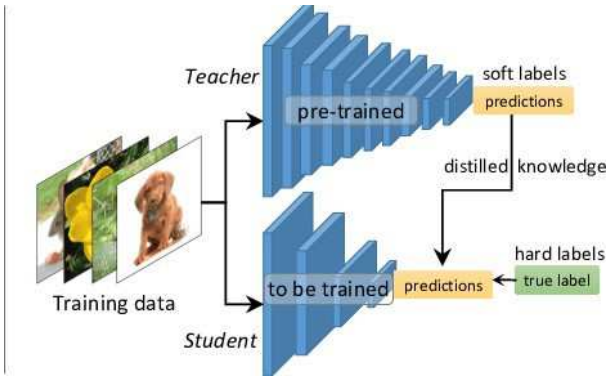
4.2. Destil·lació de coneixement

Definició (*destil·lació*): la destil·lació consisteix a transferir el coneixement d'un model complex com una xarxa neuronal profunda a un model simple i interpretable (com ara un arbre de decisió), de manera que sigui igual de vàlid, és a dir, que conservi la mateixa generalització que el model complex.

A partir del model simple, podem donar explicacions del model complex (és a dir, el model complex és explicable). La destil·lació també es pot veure com un procés de compressió, ja que el model simple també sol ser computacionalment menys complex i es pot executar més eficientment.

Per tal de transferir el coneixement del model complex (model professor) al simple (model estudiant), el model professor minimitza la funció d'error en l'entrenament i per a cada observació retorna un vector de probabilitats, on cada coordenada del vector és la probabilitat de classificar l'observació a una classe, tal com ho faria normalment. El model estudiant, aleshores, rep el mateix conjunt d'entrenament d'entrada i, per a cada observació, intenta minimitzar les probabilitats de les seves sortides respecte de les que retorna el model professor –aquesta és la funció d'error del model estudiant.

Figura 8. Procés de destil·lació de coneixement



Font: Prakhar Ganesh (Medium)

L'ús de la destil·lació a la pràctica ha estat criticat per la comunitat d'interpretabilitat ja que normalment s'utilitza el model de caixa negra i només quan es necessita una explicació es fa servir el model interpretable. Però si un model simple és tan capaç com un de complex de generalitzar, per què no entrenar un model simple i interpretable des d'un bon principi? O, si més no, un cop destil·lat, per què no descartem el model complex i fem servir el model interpretable?

4.3. Valors de Shapley

Una de les maneres d'interpretar models simples és quantificar la importància de les característiques del model en les decisions. Els arbres de decisió són un model simple que, per construcció, pot quantificar la importància de les característiques en les decisions. Altres models i, en particular, models complexos com les xarxes neuronals profundes, no proporcionen un mètode directe per quantificar la importància de les característiques, la qual cosa en dificulta la interpretació.

Els valors de Shapley permeten mesurar la importància de les característiques en models complexos. El seu origen es troba en la teoria de jocs cooperativa. La premissa és: en una coalició de jugadors que cooperen per assolir un

objectiu on cada jugador té una contribució diferent, com és d'important la contribució de cada jugador en la cooperació i quin és el guany que haurien d'obtenir d'acord amb la seva contribució? Doncs bé, aquest concepte es pot aplicar en un procés d'aprenentatge automàtic on les característiques són els jugadors i el que volem és quantificar la contribució de cada característica en una decisió.

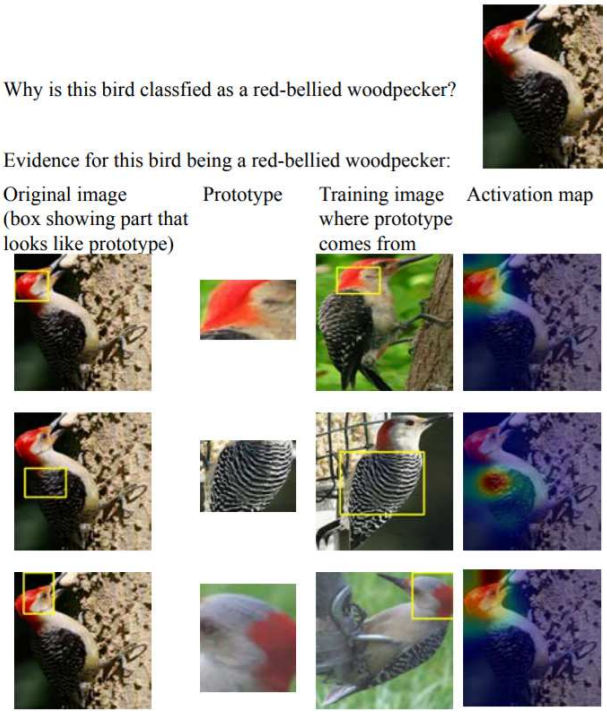
Tot i que els valors de Shapley no són específics per a xarxes neuronals, en fem menció per la seva importància en estudis de biaix algorísmic, ja que sovint s'utilitzen per mesurar la contribució de l'atribut protegit en les prediccions del model.

4.4. Visualització de característiques

Seguint amb la idea d'interpretar les decisions de models complexos segons les característiques del model, una altra tècnica que es proposa per xarxes neuronals profundes en aplicacions de visió per computador es basa a generar visualitzacions de les característiques.

La idea darrere de la visualització de característiques és produir explicacions de les decisions a través de comparacions amb membres prototípics de la classe. La idea és interessant perquè intenta donar explicacions a partir dels atributs essencials de la classe. Per exemple, en la figura 9 mostrem com una xarxa neuronal que classifica una imatge com a picot negre el classifica com a tal perquè presenta les característiques identificatives d'aquesta espècie d'ocell: cresta vermella i un plomatge amb línies negres i blanques.

Figura 9



Font: Chen i altres

Figura 9

Explicació de la classificació d'un picot negre utilitzant comparacions amb altres membres de la classe i assenyalant les característiques identificatives en la imatge.

5. Conclusions

Aquest mòdul ha introduït les principals formalitzacions d'equitat algorísmica que hi ha a la literatura, els resultats d'impossibilitat i les discussions que trobem al voltant de les definicions. A més, hem descrit tècniques de correcció dels models que ens permeten obtenir models que satisfan algunes d'aquestes propietats. Per últim, fem un repàs als mètodes d'explicabilitat que s'estan fent servir actualment per entendre i raonar sobre els models més complexos, la qual cosa contribueix a fer les decisions més transparents i per tant detectar i auditar casos de discriminació.

Aquest mòdul no pretén fer un recull exhaustiu de totes les definicions i mètodes relacionats amb el camp del biaix algorísmic que hi ha. A més, cal fer una distinció entre aquest camp, que és relativament nou, amb altres definicions d'equitat que trobem en economia i teoria de jocs. Tot i que estan relacionades, en aquests camps es donen definicions d'equitat relacionades amb el benestar dels individus i les diferents formes de repartir recursos i oportunitats, però donen menys importància a la discriminació basada en atributs protegits. De fet, aquestes definicions de benestar precedeixen el camp d'aprenentatge automàtic com el coneixem actualment i, per tant, l'èmfasi no recau en els riscos de la presa de decisions automatitzada amb algorismes.

Exercicis d'autoavaluació

1. Quina de les definicions d'equitat grupal és més adequada per a les següents tasques:

a) Accés a la universitat

b) Selecció de personal

c) Concessió de préstecs bancaris

2. Proposa una mesura de distància entre individus per garantir equitat individual en l'accés a la universitat. Raona la resposta.

3. Elabora una taula i una gràfica que mostrin les estadístiques sobre les sol·licituds a préstecs bancaris d'un banc, indicant els préstecs que s'han retornat. Les dades han de satisfer la condició de separació. Raona per què aquesta condició se satisfà.

4. Explica què vol dir la condició d'independència en el cas de COMPAS. Raona per què no és una bona definició d'equitat en aquest cas.

5. Dona i explica exemples de mals de representació.

6. Defineix un model de caixa negra, un model explicable i un model interpretable. Dona exemples on sigui adequat aplicar un model complex de caixa negra i justifica la resposta.

Solucionari

1. Les següents són possibles eleccions de la condició d'equitat grupal per cadascun dels casos:

a) En aquest cas pot ser desitjable que se satisfaci una condició d'independència que ajudi a corregir biaixos existents. Tindria un efecte semblant a garantir unes quotes mínimes d'estudiants de tots els grups, encara que aquests grups, de mitjana, hagin obtingut notes inferiors a altres grups.

b) La condició de separació pot ser la més desitjable ja que garanteix que no es rebutgen bons candidats de manera dispar entre els grups, sense que l'organització contractant hagi de renunciar als interessos del negoci.

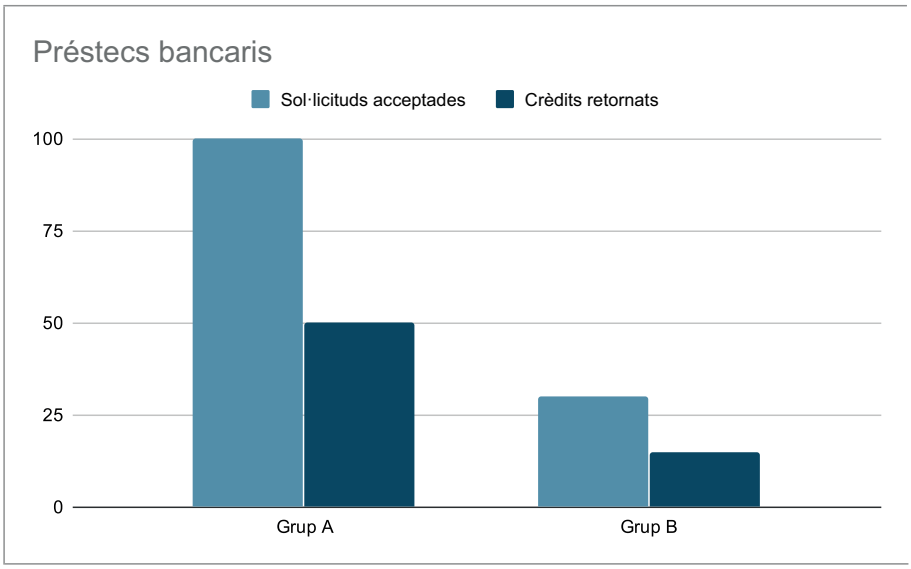
c) La condició de suficiència pot ser desitjable en aquest cas, ja que ens permeten fer una interpretació de la probabilitat de no retornar el crèdit que és independent dels atributs protegits.

2. Una distància entre els individus per garantir equitat individual a l'accés a la universitat pot ser la diferència entre les notes de les PAU. Aquesta distància garantiria que dos estudiants que han tret notes semblants tenen oportunitats semblants i tenen una probabilitat semblant d'accedir als departaments als quals sol·liciten una plaça. Tot i així, imposar equitat individual amb aquesta distància pot no ser suficient per garantir un tractament just entre els estudiants. La discriminació pot haver-se donat molt abans. Un exemple és aquell estudiant que pertany a una minoria marginalitzada i que no ha tingut accés al mateix nivell d'educació o no l'ha pogut aprofitar per no tenir un nucli familiar estable. En aquest cas, encara que aquest estudiant fos capaç de treure una bona nota a les PAU, un tractament discriminatori previ al grup de la població al qual pertany pot haver influenciat que no obtingués uns bons resultats. L'equitat individual no pot corregir aquest problema.

3. Un exemple de taula i gràfica són:

Figura 10. Com podem observar en la figura, el TPR és el mateix per ambdós grups: $TPR_A = TPR_B = 0.5$

	Sol·licituds acceptades	Crèdits retornats
Grup A	100	50
Grup B	30	15



4. En el cas de COMPAS, si prenem la raça com a atribut protegit (A), la condició d'independència se satisfà si el percentatge de presos als quals se'ls concedeix la llibertat és el mateix per totes les races. Per tant, és possible que s'hagi de concedir la llibertat condicional a presos amb alt risc per tal d'assolir la mateixa quota per tots els grups. Això, òbviament, no és desitjable en el cas de COMPAS.

5. Alguns exemples són:

a) La manca de professionals de sexe femení en posicions relacionades amb les TIC afavoreix estereotips de sexe i indueix un biaix en la tria de carreres de futurs professionals.

b) La falta de polítics que pertanyin a minories ètniques i religioses resulta en una falta de representació d'aquestes minories en les institucions i en la presa de decisions.

c) Els pocs casos de futbolistes professionals del futbol masculí que són obertament homosexuals creen estereotips sobre l'orientació sexual i la masculinitat en aquest esport.

d) La falta de persones de raça negra que ocupen càrrecs d'alta responsabilitat en grans corporacions.

6. Un model de caixa negra és un model que o bé és propietari o bé és tan complex que escapa a l'enteniment humà. Un model interpretable és un model que no és de caixa negra. Un model explicable és un model de caixa negra que suporta tècniques d'explicabilitat com les que hem descrit en l'últim apartat d'aquest mòdul.

Els models de caixa negra es poden utilitzar en tasques en les quals no necessitem raonar sobre les decisions o bé en les quals no tenir una explicació supera de bon tros els beneficis d'aplicar el model. Seguint aquest argument es podria justificar l'ús d'un model complex que obté una precisió i exactituds significativament superiors a qualsevol altre model interpretable en la detecció d'algun tipus de càncer.

Glossari

aprenentatge automàtic *m* Camp de la intel·ligència artificial que es dedica al desenvolupament d'algorismes que milloren el seu rendiment a partir d'experiència.

aprenentatge supervisat *m* Camp de l'aprenentatge automàtic en el qual l'experiència que els algorismes utilitzen per millorar es basa en exemples de dades.

atribut protegit *f* Atribut de la identitat d'un individu sobre el qual la presa de decisions s'ha regulat a través de lleis antidiscriminació. Per exemple, les lleis antidiscriminació solen considerar protegits atributs com el sexe, el gènere, l'edat, la minusvalidesa, l'orientació sexual, la ideologia política, l'expressió religiosa, entre d'altres.

biaix (mostra) *m* En estadística el biaix pot referir-se al biaix d'una mostra, el qual pot haver-se introduït per una metodologia de recol·lecció de dades defectuosa. Per exemple, un biaix es pot introduir si es fa una selecció no aleatòria dels subjectes de la mostra.

biaix (estimador) *m* En estadística, el biaix d'un estimador estadístic (per exemple, la variància de la mostra) és la diferència entre l'esperança matemàtica del valor de l'estimador i el valor del paràmetre que estima.

biaix (social) *m* Desigualtats entre grups d'una societat.

classificador *m* Funció sobre un conjunt d'elements que retorna les classes d'aquests elements. En el context d'aprenentatge automàtic farem referència als algorismes d'aprenentatge supervisat que tenen com a objectiu generalitzar aquestes funcions de classificació a partir d'un conjunt d'exemples.

conjunt d'entrenament *m* Exemples que es proporcionen a l'algorisme d'aprenentatge supervisat per trobar un regressor o un classificador.

equitat individual *f* Propietat d'equitat algorísmica d'un algorisme de decisió. Les decisions d'un algorisme de decisió satisfan equitat individual si les diferències entre els tractaments rebuts per les decisions de tota parella d'individus estan acotats per les diferències com a individus.

espai d'hipòtesis *m* Conjunt de funcions que l'algorisme d'aprenentatge supervisat pot triar per resoldre el problema d'optimització d'aprenentatge que redueix l'error sobre el conjunt d'entrenament.

etiqueta *f* En el context de classificadors, és la classe d'un element del conjunt d'entrenament.

explicabilitat *f* Propietat d'un model de caixa negra que indica que es pot interpretar amb un processament del model afegit.

falsos negatius *m* Elements de la classe positiva que el classificador marca com a negatius. sigla **FN**

falsos positius *m* Elements de la classe negativa que el classificador marca com a positius. sigla **FP**

FPR *m* Quocient de falsos positius entre el nombre total de negatius.

independència *f* Propietat d'equitat algorísmica d'un algorisme de decisió. Diem que un algorisme de decisió compleix la propietat d'independència si l'atribut és independent de les decisions del classificador.

interpretabilitat *f* Propietat d'un model que implica que no és de caixa negra.

mal d'assignació *m* Efectes perjudicials de biaixos en decisions que porten a un repartiment desigual de recursos i oportunitats entre diferents membres i grups de la societat.

mal de representació *m* Efectes perjudicials de la falta de representació de grups en la societat.

model *m* En aquest text ens referim a models d'aprenentatge supervisat, els quals són descripcions de la distribució estadística de les dades, incloses suposicions sobre com s'han generat les dades.

model de caixa negra *m* Models que o bé són massa complicats d'entendre o que són propietaris i que, per tant, no es té accés als paràmetres del model.

NPV *m* Quocient de negatius vertaders entre tots els elements que s'han classificat com a positius.

paradoxa de Simpson *f* Fenomen estadístic en el qual una correlació lineal entre dues variables s'inverteix quan es desglossa en subgrups de la població.

PPV *m* Quocient de positius vertaders entre tots els elements que s'han classificat com a positius.

regressor *m* Funció que aproxima una relació entre un conjunt d'elements i una variable depenent d'aquests elements.

regulació de la discriminació directa *f* Conjunt de lleis d'antidiscriminació que regulen l'ús explícit d'atributs protegits per la selecció d'individus en àmbits de la societat com el laboral i l'educació.
sin. **tractament dispar**

regulació de la discriminació indirecta *f* Conjunt de lleis d'antidiscriminació que regulen els biaixos en la presa de decisions que afecten grups de la societat (definits per un atribut protegit), encara que aquestes decisions no s'hagin pres fent ús explícit d'un atribut protegit.
sin. **impacte dispar**

resposta *f* En el context de regressors, és el valor de la variable dependent.

separació *f* Propietat d'equitat algorísmica d'un algorisme de decisió. Diem que un algorisme de decisió compleix la propietat de separació si l'atribut i les decisions del classificador són condicionalment independents respecte a la variable dependent.

suficiència *f* Propietat d'equitat algorísmica d'un algorisme de decisió. Diem que un algorisme de decisió compleix la propietat de suficiència si l'atribut i la variable dependent són condicionalment independents respecte a les decisions del classificador.

negatius verdaders *m* Elements de la classe negativa que el classificador marca com a negatius.
sigla **TN**
en true negatives

positius verdaders *m* Elements de la classe positiva que el classificador marca com a positius.
sigla **TP**
en true positives

TPR *m* Quocient de positius vertaders entre el nombre total de positius.

Word embedding *m* Tècnica de processament del llenguatge natural que consisteix a representar paraules com a vectors de nombres reals.

Bibliografia

Angwin, Julia; Larson, Jeff; Mattu, Surya; Kirchner, Lauren (2016). *Machine Bias*. Nova York: ProPublica. [Data de consulta: 20 d'agost 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, núm. 4175, pàg. 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». A: *Advances in neural information processing systems* (pàg. 4349-4357).

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

Calmon, Flavio; **Wei, Dennis**; **Vinzamuri, Bhanukiran**; **Ramamurthy, Karthikeyan Natesan**; **Varshney, Kush R.** (2017). «Optimized pre-processing for discrimination prevention». A: *Advances in Neural Information Processing Systems* (pàg. 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». A: *Advances in neural information processing systems* (pàg. 8930-8941).

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

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

Corbett-Davies, Sam; Pierson, Emma; Feller, Avi; Goel, Sharad (2016). «A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear». *The Washington Post*. [Data de consulta: 20 d'agost 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* (pàg. 797-806).

Crawford, Kate (2017). *The Trouble with Bias*. NIPS Keynote. [Data de consulta: 8 de juliol 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*. Missouri: Northpointe Inc.

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

European Network of Legal Experts in Gender Equality and Non-discrimination (2018). *European equality law review. Justice and Consumers*. [Data de consulta: 20 d'agost 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*. FRA Focus. [Data de consulta: 20 d'agost de 2020]. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf

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

Hardt, Moritz (2017). *How big data is unfair: Understanding unintended sources of unfairness in data driven decision making*. Medium (2014). [Data de consulta: 8 de juliol 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, núm. 1, pàg. 1-33).

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

Kleinberg, Jon; **Mullainathan, Sendhil**; **Raghavan, Manish** (2016). «Inherent 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*. Nova York: ProPublica. [Data de consulta: 20 d'agost de 2020]. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Mitchell, Shira (2018). *Mirror Mirror. Reflections on Quantitative Fairness*. [Data de consulta: 20 d'agost 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 AI snake oil*. Presentació al MIT. [Data de consulta: 8 de juliol de 2020]. https://www.cs.princeton.edu/ arvindn/talks/MIT-STS-AI-snakeoil.pdf

Narayanan, Arvind (2018). *Tutorial: 21 fairness definitions and their politics*. FAT*. [Data de consulta: 20 d'agost 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». A: *Advances in Neural Information Processing Systems* (pàg. 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, núm. 5, pàg. 206-215).

Suresh, Harini (2019). *The Problem with «Biased Data»*. Medium. [Data de consulta: 20 d'agost 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». A: *2019 IEEE International Conference on Big Data (Big Data)* (pàg. 1401-1406). Nova Jersey: IEEE.

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

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