Riscos discriminatoris en línia:
Intel·ligència artificial i economia de plataformes
Sessió de treball interna

- Algorismes Esbiaixats -

Maria Vanrell

Catedràtica d'Intel·ligència Artificial - UAB

Centre de Visió per Computador

algorisme

[DIEC]

[c. 1275; del gentilici al-ḫwārizmî, sobrenom del matemàtic iranià arabitzat Muḥammad ibn Musà (s.IX), natural de Coràsmia (ḫwārizmî), introductor de l'àlgebra a l'Europa medieval amb les seves traduccions]

m. [IN][MT] Conjunt de regles per a resoldre un problema en un nombre finit de passos. Algorisme d'Euclides per a trobar el màxim comú divisor de dos nombres.

Exemple d'un algorisme

L'algorisme que dona instruccions a un caixer automàtic

```
Si ( un client insereix la seva targeta ) llavors
    • Llegir el número del <compte_del_client>
    • Llegir quants diners vol <treure_el_cliente>
    • Consultar a la BD del banc el <saldo_del_client>
    • Si ( treure_el_client < saldo_del_client ) llavors
            • Donar els diners al client
    • Sinó
            • No donar els diners al client
    • Despedir-se del client        - Comprovacions i donar els diners-

Sinó Esperar un nou client i anar al principi
```

algorisme

[DIEC]

√

[c. 1275; del gentilici al-ḫwārizmî, sobrenom del matemàtic iranià arabitzat Muḥammad ibn Musà (s.IX), natural de Coràsmia (ḫwārizmî), introductor de l'àlgebra a l'Europa medieval amb les seves traduccions]

m. [IN][MT] Conjunt de regles per a resoldre un problema en un nombre finit de passos. Algorisme d'Euclides per a trobar el màxim comú divisor de dos nombres.

algorisme esbiaixat ?

esbiaixar

[DIEC]

1. 1.v.tr. [LC] [IMF] Tallar, col·locar (una cosa), de biaix. Esbiaixar una fusta.

2. 2.intr. [LC] Anar, travessar, de biaix.

3. tr. [FIF] Donar o produir un biaix (en el resultat d'un mesurament).

biaix

[DIEC]

1 m. [LC] Direcció obliqua en què està, es mou, és tallada, alguna cosa. El biaix d'un mur. Hi havia un sofà posat de biaix. Una peça de roba tallada al biaix.

2 1 m. [LC] Cosa tallada de biaix, obliquament. Posar biaixos a un vestit.
2 2 [LC] haver-hi molt de biaix Haver-hi molta diferència entre dues coses.
2 3 [LC] treure els biaixos Cercar economies, estalviar.

3 1 m. [FIF] Predisposició que fa que el resultat d'una mesura s'aparti de l'esperada segons les lleis de la física o les probabilitats. Biaix d'ordenació.
3 2 m. [FIF] Mesura de la descompensació d'un conjunt d'errors. Un error compensat té un biaix nul.

4 1 m. [IMF] Forma o posició d'allò que es desvia d'una recta de referència.
4 2 [AQ] [IMF] biaix de cartabó Forma de l'extrem d'un llistó o d'una post consistent en un pla que forma un angle diedre de 45° amb els costats llargs d'aquesta peça.
4 3 [IMF] biaix buscat Biaix obligat per una necessitat constructiva.

Exemple d'un algorisme esbiaixat

Un algorisme que discrimina pel color de la pell

Si (un client insereix la seva targeta) llavors
- Llegir el número del <compte_del_client>
- Llegir quants diners vol <treure_el_cliente>
- Consultar a la BD del banc el <saldo_del_client>
- Si (treure_el_client < saldo_del_client) llavors
    - Donar els diners al client
- Sinó
    - No donar els diners al client
- Despedir-se del client

Sinó Esperar un nou client i anar al principi

Exemple d'un algorisme esbiaixat

Un algorisme que discrimina pel color de la pell

Si (un client insereix la seva targeta) llavors

- Llegir el número del <compte_del_client>
- Llegir quants diners vol <treure_el_cliente>
- Consultar a la BD del banc el <saldo_del_client>
- **Consultar a la BD del banc si <client_és_caucàsic>**
- **Si** (treure_el_client < saldo_del_client) **llavors**
    - Donar els diners al client
- **Sinó**
    - No donar els diners al client
- Despedir-se del client

Sinó Esperar un nou client i anar al principi

Exemple d'un algorisme esbiaixat

Un algorisme que discrimina pel color de la pell

Si (un client insereix la seva targeta) llavors
- Llegir el número del <compte_del_client>
- Llegir quants diners vol <treure_el_cliente>
- Consultar a la BD del banc el <saldo_del_client>
- Consultar a la BD del banc si <client_és_caucàsic>
- Si (treure_el_client < saldo_del_client) llavors

- Sinó
    - No donar els diners al client
- Despedir-se del client

Sinó Esperar un nou client i anar al principi

Exemple d'un algorisme esbiaixat

Un algorisme que discrimina pel color de la pell

Si (un client insereix la seva targeta) llavors
- Llegir el número del <compte_del_client>
- Llegir quants diners vol <treure_el_cliente>
- Consultar a la BD del banc el <saldo_del_client>
- Consultar a la BD del banc si <client_és_caucàsic>
- Si (treure_el_client < saldo_del_client) llavors
  - Si (client_es_caucàsic= CERT) llavors
    - No donar els diners al client
  - Sinó
    - Donar els diners al client
- Sinó
    - No donar els diners al client
- Despedir-se del client

Sinó Esperar un nou client i anar al principi

Conclusió,

és molt fácil fer algorismes amb biaix …

BBVA Historia de los cajeros automáticos

Portem més de 50 anys amb caixers i mai ens hem plantejat que tinguin biaixos ...

un adelantado a su tiempo. Sólo era usado por aquellas personas que no querían que les vieran los empleados del banco y el banco decidió cerrarlo.

El primer cajero automático que tuvo éxito

Esperem un ús adequat de la tecnologia

El color de la pell no és una dada que ens demanin quan obrim un compte ....

dinero

Los cajeros automáticos llegaron para resolver la necesidad de las personas para obtener dinero en efectivo fuera de los bancos.

Notícies ▸ Vídeos ◂ Àudios Seccions Directes: 24 catalunya informació

BARCELONA

El caixer automàtic compleix 40 anys

Tornem a l'algorisme esbiaixat:

L'algorisme que discrimina pel color de la pell

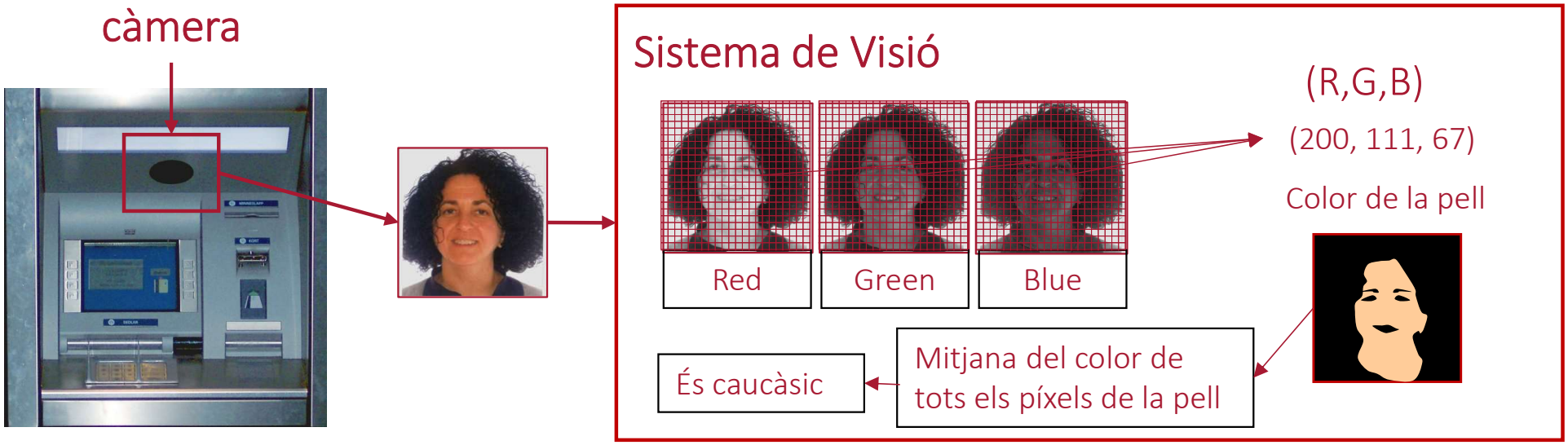
Si (un client insereix la seva targeta) llavors
- Llegir el número del <compte_del_client>
- Llegir quants diners vol <treure_el_cliente>
- Consultar a la BD del banc el <saldo del client>
- Fer una <foto_del_client>
- Mirar si el <cliente_és_caucàsic>

→ Sistema de visió

  - Si (client_es_caucàsic= CERT) llavors
    - No donar els diners al client
  - Sinó
    - Donar els diners al client
- Sinó
    - No donar els diners al client
- Despedir-se del client

Sinó Esperar un nou client i anar al principi

Què fa el sistema de visió? ....



La dada esbiaxada es pot aconseguir automàticament ...

Què poden fer els sistemes de visió ?

Com som? el color de la pell, guapos, lletjos, rossos, ...
Com ens vestim ? ...
Com ens trobem? tristos, contents, ...
Amb qui anem? ...
Qui som ... en definitiva ...

però no només els sistemes de visió,

altres sistemes poden accedir a més dades sobre nosaltres ...

On anem? ... Geolocalització del nostre mòbil

Què comprem ? ... transaccions amb les nostres targetes

Què ens agrada ? ... a les xarxes socials

Tenim 2 problemes essencials, …

- Les nostres dades són a tot arreu i fàcilment accessibles …

  Les lleis de protecció de dades ja ho regulen …

- Els algorismes que manipulen les dades poden tenir intencions discriminatòries ….

  La regulació es va adaptant i la justícia demana ajut a experts tecnològics …

Solució fins ara, adaptar la legislació als nous problemes

El legislador demana ajut als experts ....

Si (un client insereix la seva targeta) llavors
- Llegir el número del <compte_del_client>
- Llegir quants diners vol <treure_el_cliente>
- Consultar a la BD del banc el <saldo_del_client>
- Fer una <foto_del_client>
- Mirar si el <cliente_és_caucàsic>
- Si (treure_el_client < saldo_del_client) llavors
  - Si (client_es_caucàsic= CERT) llavors
    - No donar els diners al client
  - Sinó
    - Donar els diners al client
- Sinó
  - No donar els diners al client
- Despedir-se del client

Sinó Esperar un nou client i anar al principi

Instruccions racistes !!!!

Nou Problema,

en els darrers anys els algorismes han canviat

La intel·ligència artificial dels últims 10 anys ha patit una revolució, hem passat a treballar majoritàriament amb algorismes basats en Xarxes Neuronals profundes (Deep learning),

S'ha complicat la identificació de les instruccions racistes ...

Noves Preguntes,

Què és una xarxa neuronal profunda?

Com es programa?,

Com s'entrena?

Respostes,

1a Idea: Les xarxes neuronals s'inspiren en el funcionament del cervell, no són noves

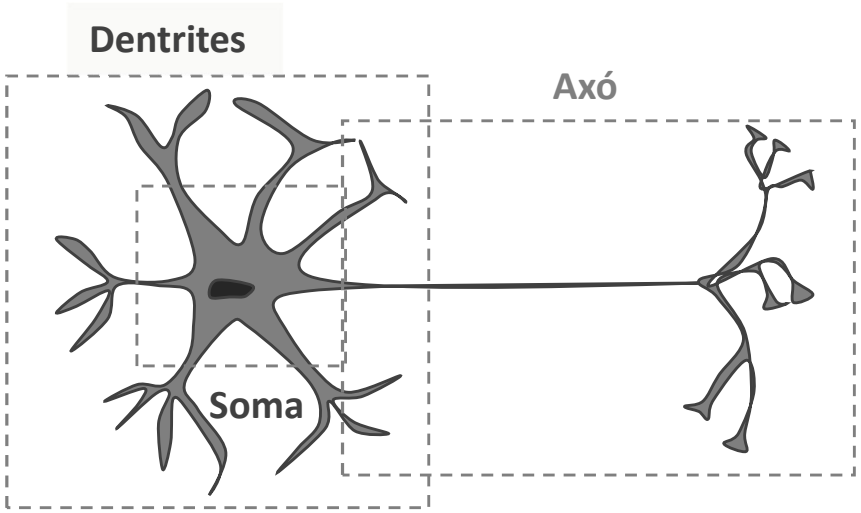
2a Idea: Les xarxes neuronals poden aprendre si les entrenem a partir d'exemples

3a Idea: Les xarxes neuronals no tenen instruccions racistes, però poden aprendre de dades que tinguin biaix

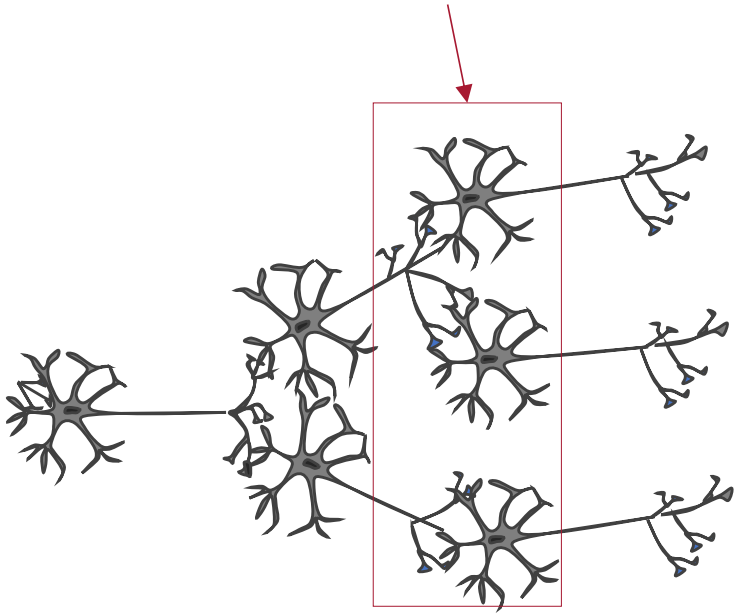
1a Idea: Les xarxes neuronals s'inspiren en el funcionament del cervell, no són noves

La base de la intel·ligència humana està en el cervell i les seves neurones

Neurona

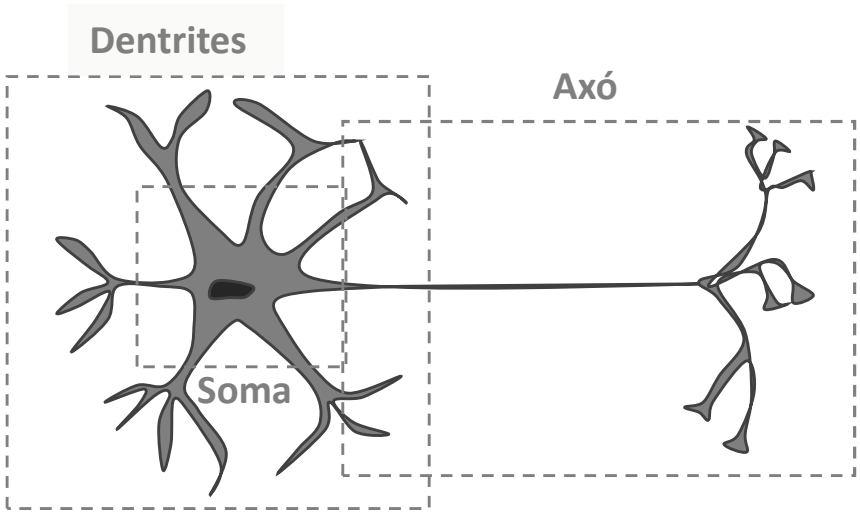


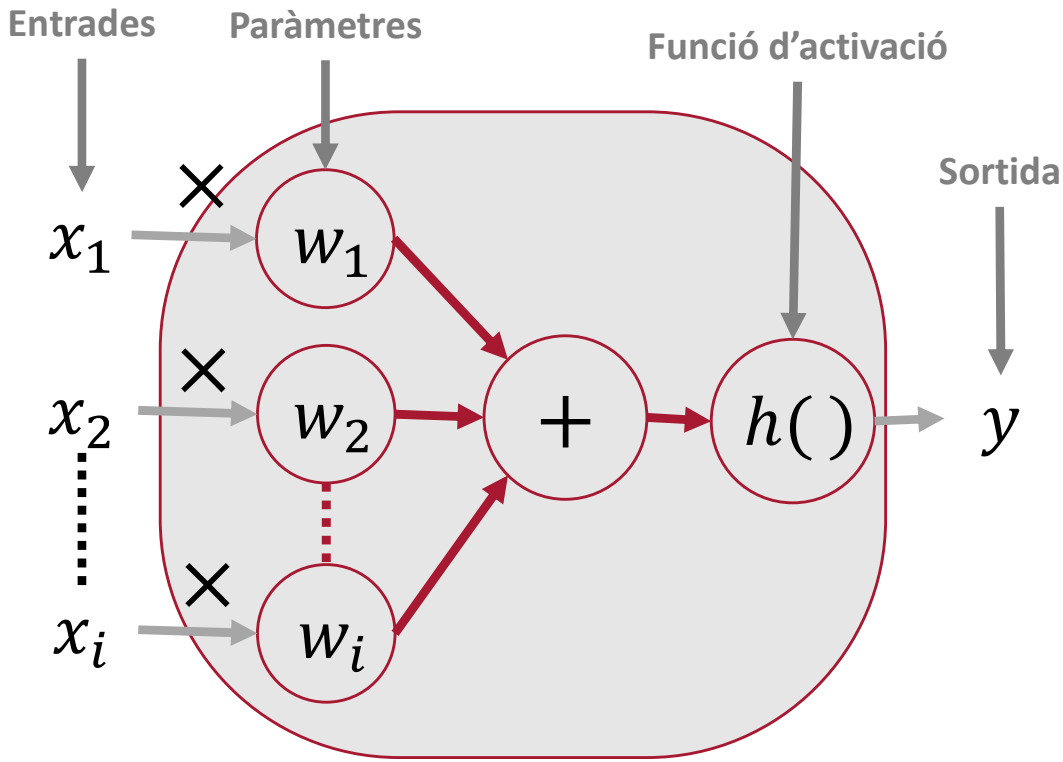
Capes de neurones



1a Idea: Les xarxes neuronals s'inspiren en el funcionament del cervell, no són noves

Neurona real → Neurona artificial





Algorisme d'una neurona

Basat en paràmetros, sumes, productes i funcions

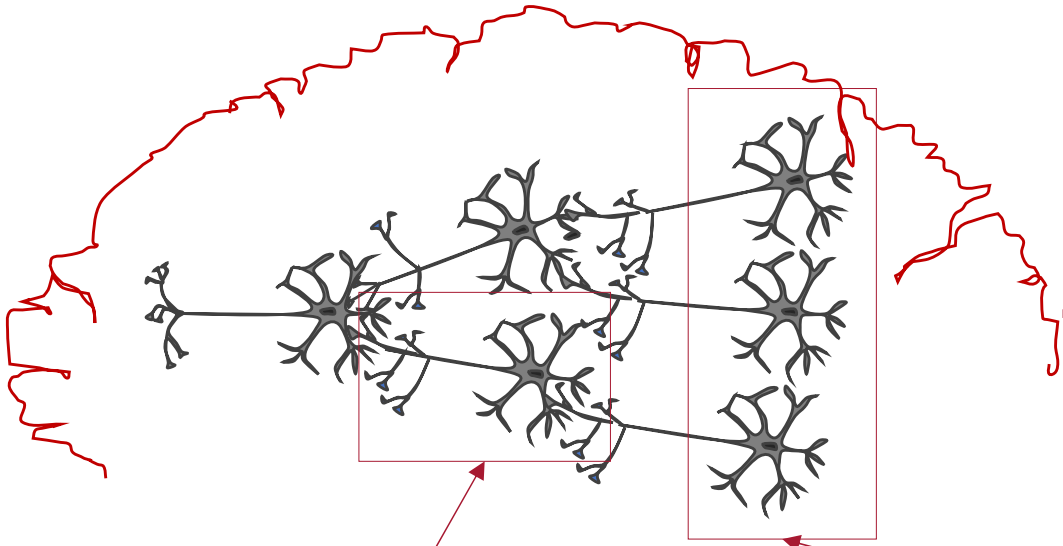
Rosenblatt, Frank (1957)
The Perceptron--a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory.

1a Idea: Les xarxes neuronals s'inspiren en el funcionament del cervell, no són noves

Cervell

Estímul

Resposta



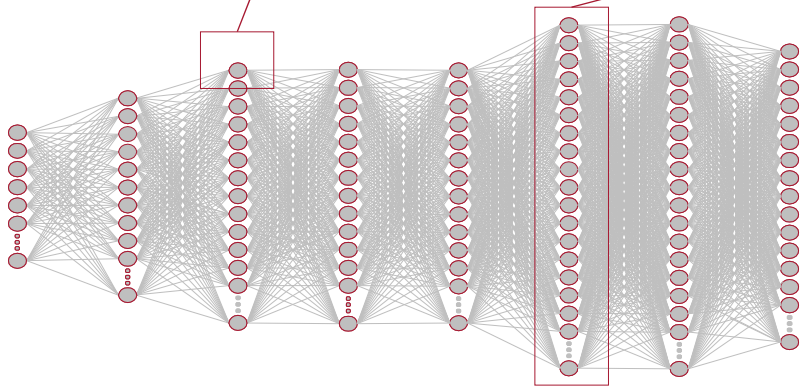
Xarxa Neuronal Artificial

Neurona

Capes de Neurones

Entrada

Sortida



Algorisme d'una xarxa de neurones

Respostes,

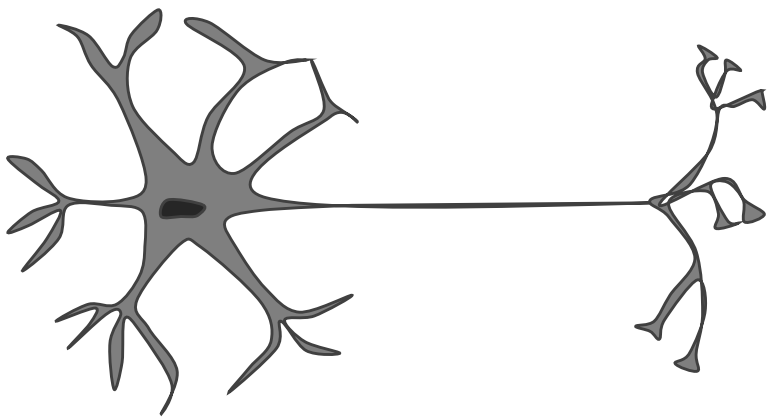
1a Idea: Les xarxes neuronals s'inspiren en el funcionament del cerebro, no són noves

2a Idea: Les xarxes neuronals poden aprendre si les entrenem a partir d'exemples

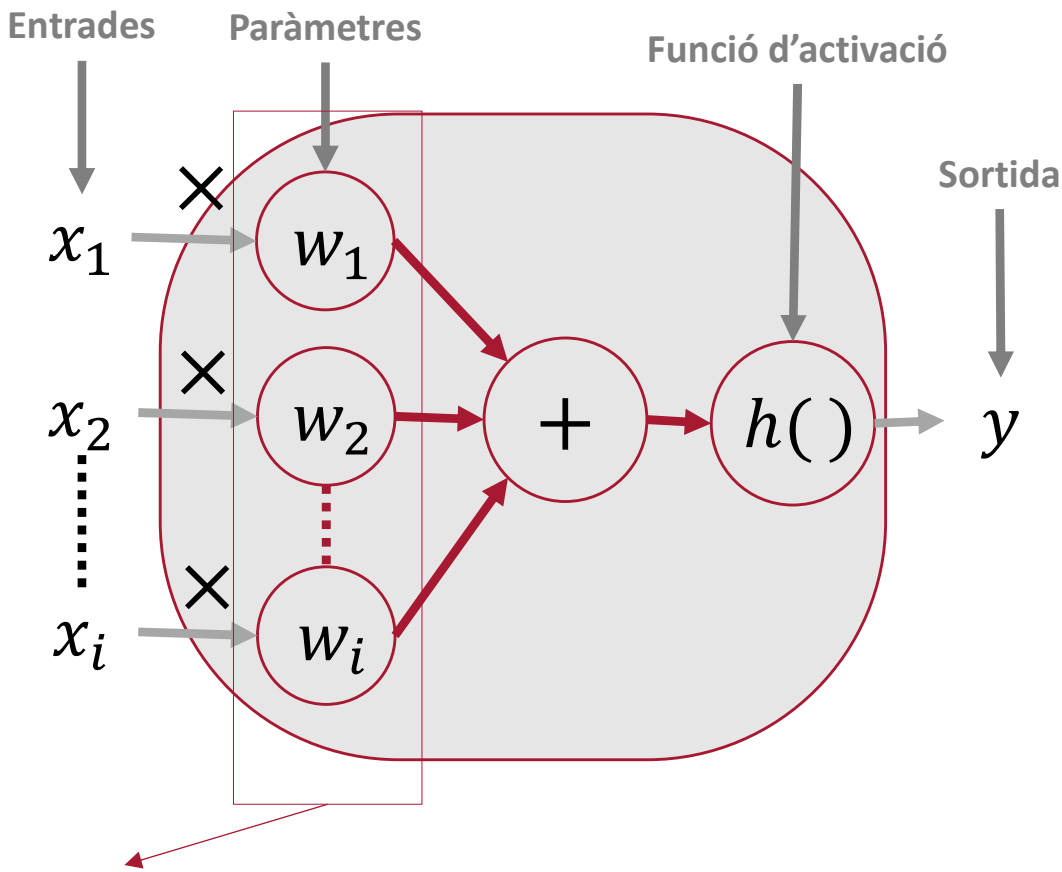
3a Idea: Les xarxes neuronals no tenen instruccions racistes, però poden aprendre de dades que tinguin biaix

2a Idea: Les xarxes neuronals poden aprendre si les entrenem a partir d'exemples

Neurona real

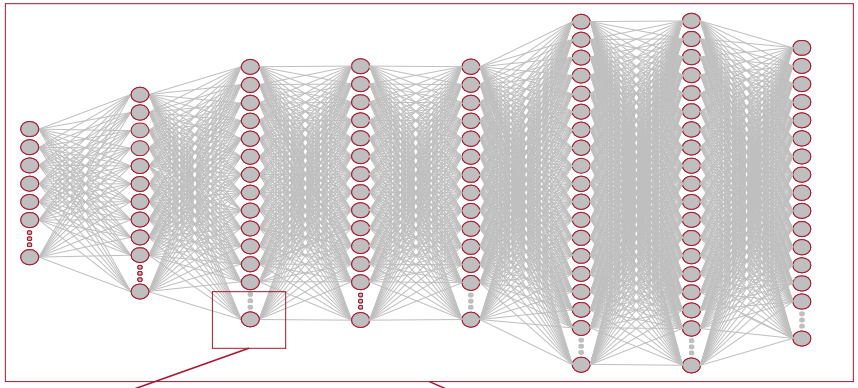


Neurona artificial



Aquests paràmetres són els responsables del comportament d'una neurona, són els que determinen què aprèn la xarxa

2a Idea: Les xarxes neuronals poden aprendre si les entrenem a partir d'exemples



$w_1, w_2, \ldots, w_X$

1 neurona → *X* paràmetres

1 xarxa neuronal profunda

1000…0000 de paràmetres

$w_1, w_2, \ldots, w_{1000\ldots000}$

Aprendre = Trobar aquests paràmetres

que fan que la resposta sigui l'esperada
en funció de l'entrada i l'objectiu de l'algorisme

2a Idea: Les xarxes neuronals poden aprendre si les entrenem a partir d'exemples

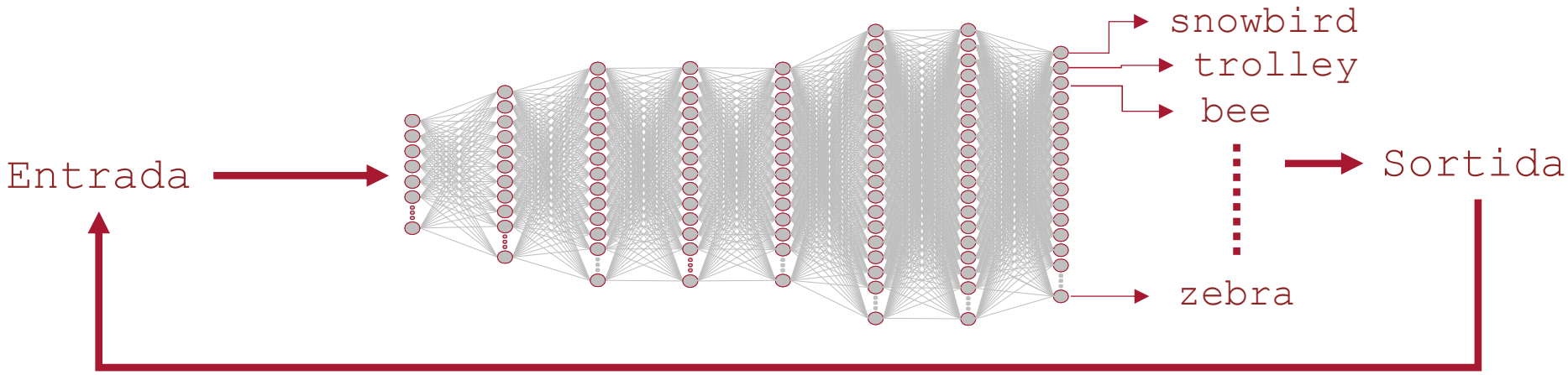
Cóm ? → Amb supervisió i màquines especialitzades

(1) Tècniques d'optimització sobre exemples etiquetats

"snowbird" "trolley" "bee" "chest" ... "zebra"

(2) Ordinadors amb *hardware* especialitzat

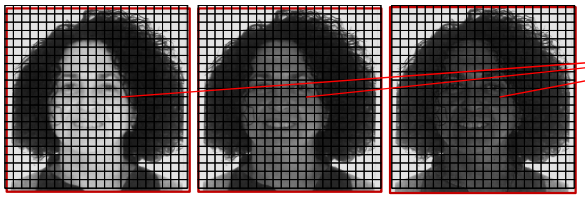
L'any 2012 les primers xarxes de reconeixement d'objectes



ImageNet: Base de dades d'1,2 milions d'imatges etiquetades

Actualitzem el Sistema de visió
que reconeix el color de la pell

Sistema de Visió



Red Green Blue

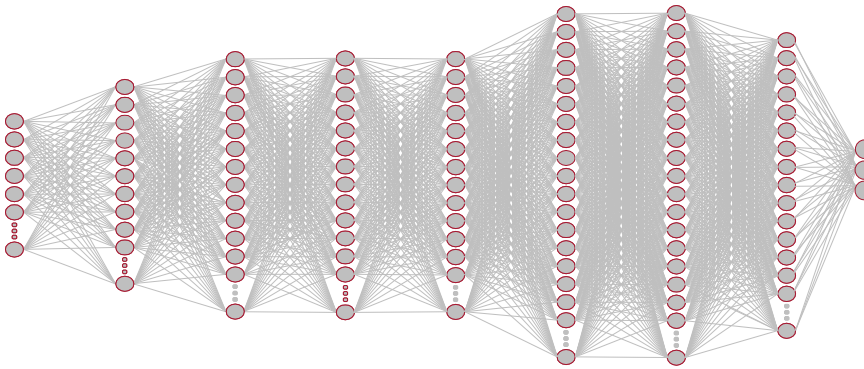
(R,G,B)

(200, 111, 67)

Color de la pell

És caucàsic ← Mitjana del color de tots els píxels de pell

Sistema de Visió



Caucassian
Mongolian
Ethiopian

Pattern Recognition Letters
Volume 24, Issue 11, July 2003, Pages 1715-1723

Tracking regions of human skin through illumination changes

Moritz Störring [a], Tomáš Kočka [b], Hans J Andersen [a], Erik Granum [a]

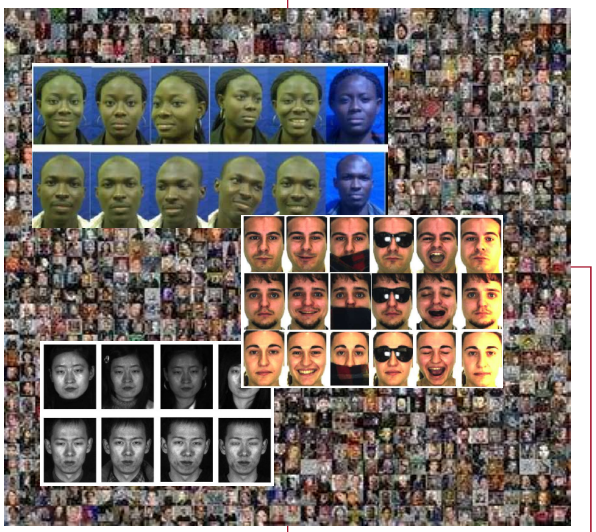
Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/S0167-8655(02)00327-6 Get rights and content

Abstract

New human computer interfaces are using computer vision systems to track faces and hands. A critical task in such systems is the segmentation. An often used approach is colour based segmentation, approximating the skin chromaticities with a statistical model, e.g. with mean value and covariance matrix. The advantage of this approach is



"Ethiopian"

"Caucassian"

"Mongolian"

Respostes,

1a Idea: Les xarxes neuronals s'inspiren en el funcionament del cervell, no són noves

2a Idea: Les xarxes neuronals poden aprendre si les entrenem a partir d'exemples

3a Idea: Les xarxes neuronals no tenen instruccions racistes, però poden aprendre de dades que tinguin biaix

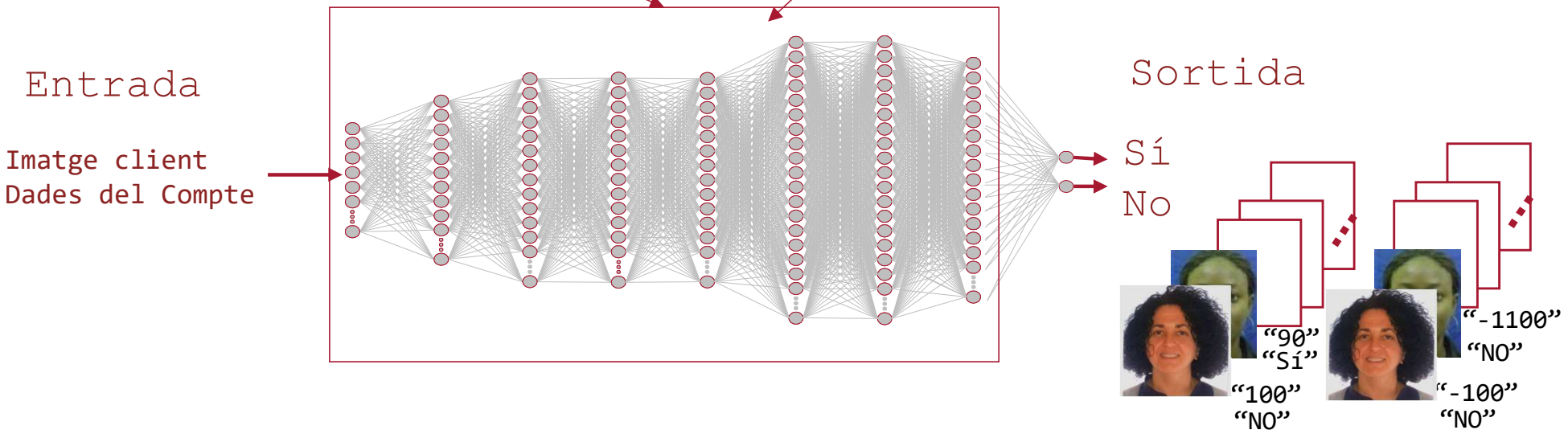
Actualitzem el caixer automàtic

que discrimina segons el color de la pell

```
Si ( un client insereix la seva targeta ) llavors
    • Llegir el número del <compte_del_client>
    • Llegir quants diners vol <treure_el_cliente>
    • Consultar a la BD del banc el <saldo_del_client>
    • Fer una <foto_del_client>
    • Mirar si el <cliente_és_caucàsic>
    • Si ( treure_el_client < saldo_del_client ) llavors
        • Si ( client_es_caucàsic= CERT) llavors
            • No donar els diners al client
        • Sinó
            • Donar els diners al client
    • Sinó
            • No donar els diners al client
    • Despedir-se del client
Sinó Esperar un nou client i anar al principi
```

Instruccions racistes !!!!

Quins paràmetres de quines neurones estan introduïnt la discriminació?



Conclusió,

Des d'algorismes en que podem identificar instruccions clarament esbiaixades

A algorismes amb paràmetres apresos en els que no podem identificar on és el biaix

Bernard Parker, left, was rated high risk; Dylan Pugett was rated low risk. (Josh Ritchie for ProPublic

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren

May 23, 2016

ON A SPRING AFTERNOON IN 2014, Bri
late to pick up her god-sister from school when she spotted an

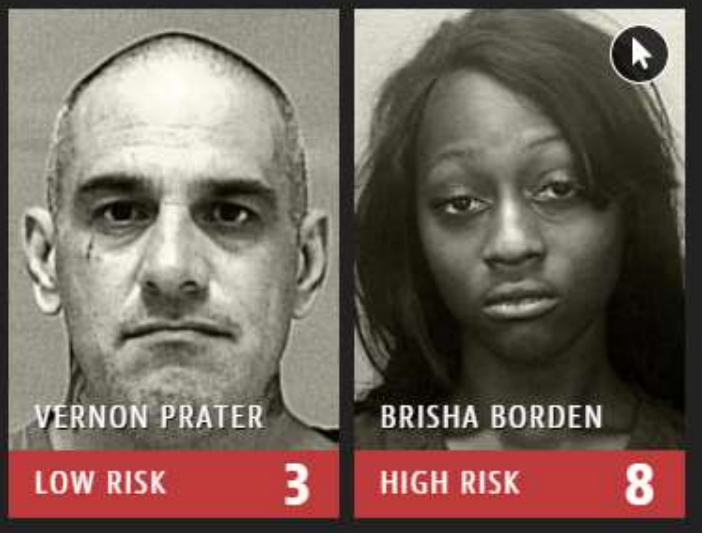
ProPublica

May 23, 2016

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner

Estudi sobre un algorisme que estimava la probabilidad de reincidència d'una persona

Two Petty Theft Arrests



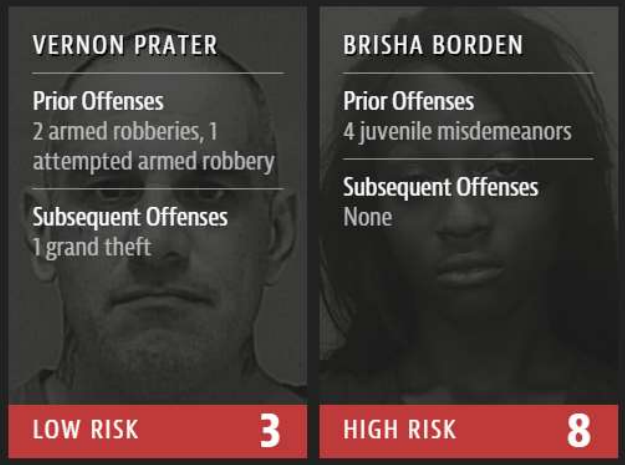
Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden — who is black — was rated a high risk. Prater — who is white — was rated a low risk.

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Petty Theft Arrests



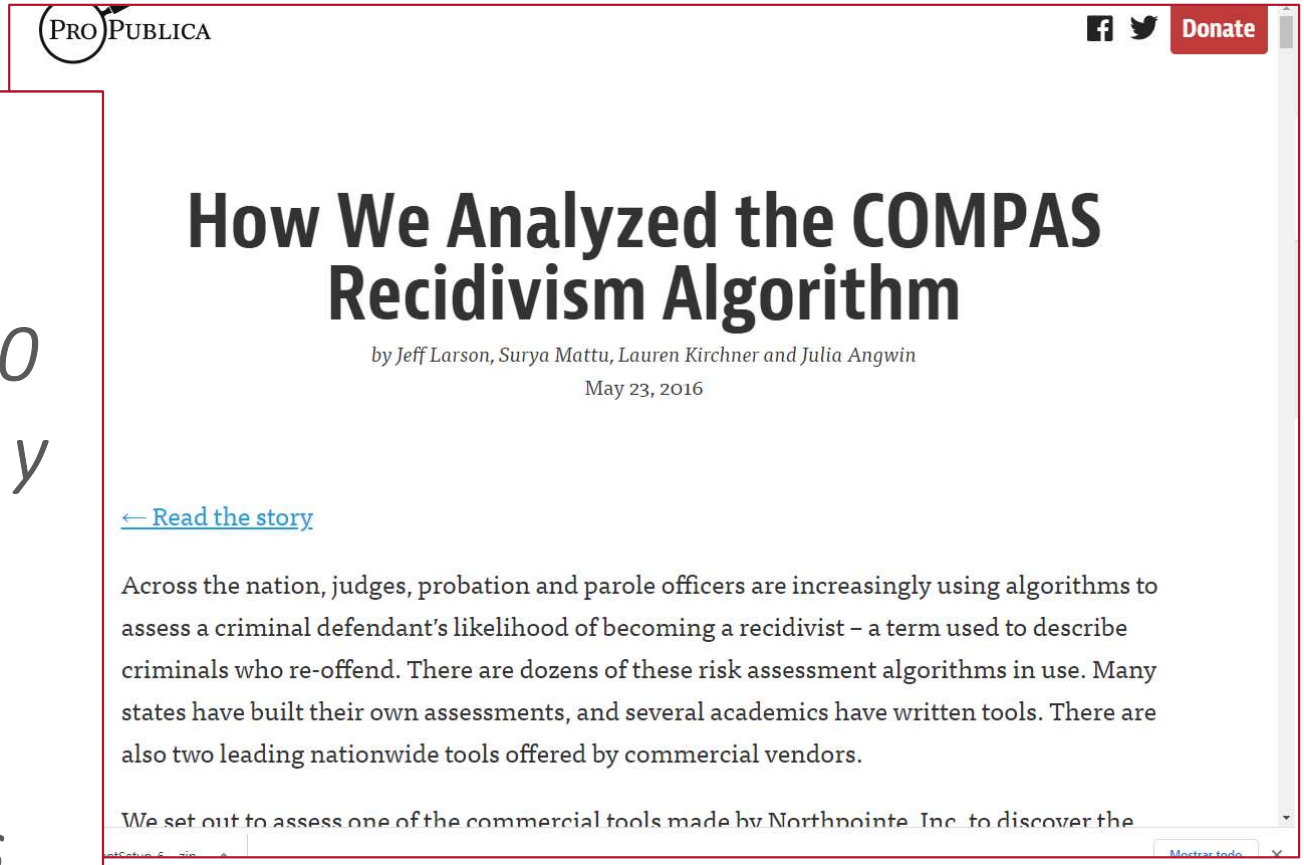
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Comparació entre

les prediccions de 7000 persones durant 2013 y 2014,

y els crims d'aquestes persones en els 2 anys següents



We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the **same benchmark used** by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

Només el 20% de les prediccions de crims violents es varen cumplir

Si consideraven tot tipus de delictes, inclús conduir amb llicència caducada, només encertaven el 60%

Les prediccions presentaven una clara discriminació per raça

La empresa Northpointe *mai ha explicat com calcula els índexos, es basa en els antecedents i en la resposta de les persones a una llista de preguntes …*

La conclusió de l'algorisme depèn d'un índex numèric:

$$Índex = w_1 * Factor_1 + w_2 * Factor_2 w + \ldots + w_N * Factor_N$$

Valors numèrics amb resultats difícils de preveure

Índex > 0.5	*Resposta positiva*
Índex < 0.5	*Resposta negativa*

Conclusió,

Des d'algorismes en que podem identificar instruccions clarament esbiaixades

Discriminació Directa

Rang de possibles algorismes

Mesures estadístiques per mesurar els biaixos

A algorismes amb paràmetres apresos en els que no podem identificar on és el biaix

Discriminació Indirecta

Per tant, cal adaptar la legislació i a més a més introduir mètodes estadístics per predir els biaixos

Algunes idees,

- *Establir protocols de test exhaustiu que permetin demostrar l'efectivitat dels programes*
- *Obligatorietat de permetre executar els algorismes per testejar-los*
- *Transparència de les dades sobre les que s'han entrenat els algorismes*
- *Codis ètics i estudis de risc de l'aplicació dels algorismes*
- *Informació i educació de la població davant d'aquests nous problemes i dels usos que se'n fan*

Algorismes Esbiaixats

Maria Vanrell

Catedràtica d'Intel·ligència Artificial - UAB

Centre de Visió per Computador