

## Isabelle Guyon

Big datako katedraduna  
Paris-Saclayeko Unibertsitatea

*“Adimen artifiziala ahalik eta jende  
gehienaren esku jarri behar dugu”*

Guillermo Roa Zubia · Elhuyar Zientzia

Argazkiak: Kaliforniako Unibertsitatea/BBVA Fundazioa



Askorentzat, adimen artifizialaren aitzindarietako bat da Isabelle Guyon. Irakatsi ordenagailuari gauza bat egiten, eta, seguru asko, Guyonen lana erabiltzen ariko zara, berak asmatu baitzituen ikasketa automatikorako erabiltzen diren oinarritzko metodo-logiak. Zenbakiak erraz maneiatzen dituzte makinek, baina, benetan giza adimenara gerturatzeko, objektuak zehaztasunez bereizten ikasi behar dute; hala nola gizonak eta emakumeak bereizten, diagnostiko mediko bat bereizten, edo edozein datu-taldetan patroiak identifikatzen. Hain zuzen, horretan egin du funtsezko ekarpena Guyonek. BBVA Fundazioaren Ezagutzaren Mugak saria jasotzera etorri zen Bilbora, eta Guillermo Roak elkarriketatu zuen Euskadi Irratiko *Norteko Ferrokarrilla* saioan. *Elhuyar* aldizkarira ere ekarri dugu.

**Adimen artifizialean eta big datan aditua zara. Hain zuzen, ikasketa automatikoaren oinarrian dauden algoritmoak sortu zenituen, duela 20 urte: laguntza-bektoreak. Alegia, datu asko aztertu behar direnean erabiltzen diren algoritmoak. Zuzena da?**

Hori da. Gizakiontzat errazak diren lan batzuk oso zailak dira makinentzat. Adibidez, udareak eta sagarrak bereizten ikasi behar badu makinak, batzuetan zaila egiten zaio, sagar batzuek udareen antza dutelako eta udare batzuek sagarra. Zaila zaio muga non dagoen jakitea. Eta algoritmo matematiko konplexu batzuk garatu genituen muga horiek detektatzen laguntzeko: laguntza-bektoreak.

Berez, ikasketa automatikoan metodo asko erabiltzen dira. Hasieran, sare neuronaletan egin nuen doktoretza: zenbait metodo konparatu nituen, tartean nukleo-metodoak deitzen direnak. Baina, gero, Bell laborategian Vladimir Vapnik irakaslea ezagutu nuenean, laguntza-bektoreak erabili eta garatu genituen, 60ko hamarkadan asmatutako metodo batean oinarrituta, datuen diskriminatzaileen adibideak baitira.

Konturatu nintzen algoritmo horiek eta nukleo-metodoak konbina nitzakeela. Nire senarrak, Bernard Bozerrek, konbinazio hori inplementatu zuen, eta nahiko ondo funtzionatu zuen. Hainbat gauzatan aplikatzen hasi ginen. Bernard Schölkopf-ekin esparru oso bat garatu genuen nukleo-metodoen inguruan, haren aplikazioak biderkatuta.

Urte askoan aritu nintzen esparru horretan. Nire lehen amodioa, hau da, sare neuronalak, ordeztu zituen. Ez nahita, baina bi esparruak nituen lehian nire lanean. Hala ere, praktikan ez dira lehiakideak. Alderantziz, uste dut oso osagarriak direla. Oinarritzko ikasketan, sare neuronalak eta laguntza-bektoreak konbina daitezke. Orain jende askok konbinatzen ditu, eta teknika ahaltsuagoak sortzen ditu.

**Sare neuronalak oso erabilgarriak dira, baina duela gutxi berpiztutako teknika da, ordenagailuen konputazio-ahaltenarengatik. Duela 25 urte inguru, zu hori ikertzen ari zinenean, ordenagailuak ez ziren inondik inora hain ahaltsuak, eta sare neuronalen ideia oso ona zen, baina ez oso errealista, une horretan. Gaur egun, ordea, ikaragarriko erabilera dute.**

Hori da, inflexio-puntu bat egon zen. Sare neuronalak eta bestelako ikasketa-makinak trebatzeko adina datu eskuragarri dagoenetik, makinek giza-kion gaitasuna berdindu dute. Batzuetan gainditu ere egin dute, datu-kopuru handiak prozesatzeko gaitasuna oso mugatua delako gizakiongan. Esate baterako, makina bat trebatu zuten Go jokoan jokatzeke, ehunka partida erakutsita, eta giza gaitasuna gainditu zuen. Ustekabekoa izan zen Go jokoa txapelduna makina batek gainditzea, pentsatzen genuelako hori oraindik urruti zegoela. Eta, jakina, horrek jendearengan ametsak adina beldur eragiten ditu.

Beldur dira makinak “supergizaki” bihurtuko ote diren. Baina nik uste dut aukera handia dela, eta ez dugula beldurtu behar; aitzitik, ustiari egin beharko dugu, eta ahalik eta biztanleria-segmentu handienaren esku jarri.

**Iraultza batean gaude. Iraultza handia, sare neuronalekin eta ikasteko makinekin. Horrela ikusten duzu?**

Bai, eta uste dut iraultzaren hasieran gaudela, asko zabaltzen ari baitira, batez ere, datuetan patroiak topatzen dituzten algoritmoak. Orain, gure telefonoetan eta gure ordenagailuetan, ikasketa automatikoko produktu asko daude, aurpegiak ezagutzen dituztenak edo itzulpen automatikoak egiten dituztenak, batzuk aipatzearren. Ikusmen artifizialeko aplikazio asko ere badaude, sare neuronal konboluzionalei esker. Hain zuzen, horiek Bell laborategian garatu zituzten, han lanean ari nintzenean. Eta, aldi berean, sektore-makinekin lan egiten genuen, osagarriak direlako.

Adibidez, demagun sare neuronal bat trebatzen duzula irudi bat segmentu txikitari zatitzeko, gero konbinatzeko moduko marra- eta gurutze-itxurako zatitxotan. Datu-base handiak behar dituzu ondo trebatzeko. Baina ez baldin badaukazu datu askorik eta daukazunak ez badira egokiak? Adibidez, imajinatu trebatu nahi duzula zure sistema aurpegiak ezagut ditzan, baina dituzun irudi gehienak objektuen irudiak direla, beste mota bateko datuak. Baina demagun umeen aurpegiaren irudi gutxi batzuk ere badituzula. Sistema trebatu ahal izateko, adibideetan oinarritzen den metodo bat behar duzu, sektore-makinen bidezkoa, adibidez, eta ez ikusten dituen ezaugarrietan oinarritutako metodo bat.

**Makinak gauza bat egiteko trebatzen ditugu, baina gu, gizakiok, gai gara oso gauza ezberdinak egiteko, eta batzuk besteekin erlazionatzeko ere bai.** Hori da, eta, horrez gain, ikasteko modu ezberdinak ditugu. Adibidez, epe luzeko memoria daukagu. Me-

moria horrek datu asko behar ditu, eta, horri esker, patroiak bereizteko estrategiak ikasteko aukera ematen digu. Eta epe laburreko memoria ere badugu; adibide batzuk buruz ikasi besterik ez dugu egiten, eta, gero, adibide horiekin konparatuta hartzen ditugu erabakiak.

*“Jendea beldur da makinak ‘supergizaki’ bihurtuko ote diren. Baina nik uste dut aukera handia dela”*

**Bioinformatikan ere egin duzu lan. Sekulako iraultza ekarri du horrek, adibidez, kimikan. Esate baterako, proteina jakin batekin lotuko den molekula baten bila bazabiltza, sistema automatikoak aukera asko azter ditzake, eta erabaki zein diren molekula aproposenak. Gero eta gehiago erabiltzen da biokimikako ikerketan.**

Bai, hori oso garrantzitsua izan da. Eta jarraitzen du garrantzitsua izaten. Big datari buruz hitz egiten dugu, datu asko izateari buruz, alegia. Baina nolako datuak behar ditugu? Funtsean, bi modu daude gaia aztertzeko: adibide-kopuru handia izatea eta adibide bakoitzaren ezaugarri-kopuru handia. Kimikaz ari bagara, ezaugarri asko dituen molekula bat azter dezakegu, milaka ezaugarri dituen. Gainera, ikerketa biomedikoan, paziente bera ere azter daiteke, zeinak milaka ezaugarri baititu.

Adibidez, geneen jarduera guztiak neurtzen badituzu, milaka ezaugarri dituzu aztergai. Big data mota desberdin bat da. Ez dugu gene askorik, baina haien ezaugarri asko ditugu. Hor erabili daitezke laguntza-sektoreen makinak. Horregatik erabili dira asko biomedikuntzan, eta orain kimikan ere bai.

**Gogoratzen dut Britainia Handiko adibide bat: emailen eta pazienteen datuak konbinatu nahi zituzten, giltzurrun-transplanteetarako emailen**



**egokiak lortzeko, odol-motaren bateragarritasuna eta beste ezaugarri batzuk kontuan hartuta. Herrialde osoko datuak zituzten, eta, adimen artifizialeko sistemei esker, gai izan ziren erabakitzeko: Londresko pertsona honek giltzurruna emango dio Manchesterreko honi, eta Yorkeko beste honek... Emaile eta pazienteen sare konplexu bat osatu zuten. Hori adibide sinplea da zuentzat, ezta? Ikaske-  
ta automatikoaren iragana da, nolabait.**

Bai. Eta interesgarriena da hainbat diziplina uztartzen ditugula: estatistika, optimizazioa eta beste ohiko metodo batzuk. Jende askok batu ditu indarrak azken 20 urteetan. Batzuetan, estatistikako ohiko metodoak ez ziren informatikan ezagutzen. Eta zirrara garria da beste adimen artifizial mota batzuetan lanean aritu diren pertsonentzat, gauza ahaltsuak egin ditzakegula bakarrik zenbakietatik abiatuta, batez ere zenbakiak manipulaturik eta datu asko bilduz.

Baina ez da magia beltza ere. Ehunka mila ezaugarri baditugu, nola bereiz ditzakegu patroiak? Gauza baten edo bestearen ezaugarri bereizgarrienak aurkitzen saiatzen gara? Jo dezagun txakurrak eta behiak bereizi nahi ditugula. Lau hanka izateak ez du axola, txakurrek zein behiek dituztelako lau hanka, baina behiek adarrak dituzte eta txakurrek ez. Horrelako ezaugarriak bilatu nahi dituzte. Azken batean, ehunka mila datuetatik abiatuta, sinplifikatu dezakezu arazoa, problema jakin baterako inporta zaizkizun zenbaki gutxi horiek bakarrik aztertuta.

**Elhuyarren adimen artifizialeko ikertzaileak daude, euskaratik ingeleserako, gaztelaniarako, frantseserako eta beste hizkuntza batzuetarako itzulpen automatikoan lan egiten dutenak, eta alderantziz. Arazoa da euskarak ez duela corpus handi bat, ez duela hizkuntza handiek adina datu**



## *“Adimen artifizialean, datu gutxi izatea da egoerarik zailena”*

**konparatzeko. Beraz, oso lan ona egin behar dute informazio horrekin makina bat entrenatzeko.**

Jendeak askotan pentsatzen du zaila dela datu asko izatea, baina datu gutxi izatea da egoerarik zailena. Hain zuzen, Vapnik-en teoriak asko lagundu zigun ulertzen ezen, datu gutxi ditugunean, eredu nahiko sinpleak erabili behar ditugula. Bitxia bada ere, sare estuak dira datu gutxi maneiatzen dituzten sare neuronalak. Teoria konplexua dago horren azpian. Orain, erregularizazio-teoria esaten zaio; alegia, datu gutxiarekin lan egiteko, gakoa ez da bakarrik zer eredu erabiltzen duzun; eredu hori trebatzeko modua ere funtsezkoa da.

Bereziki interesatzen zait “datu gutxiaren ikasketa laburra” deitzen duguna; alegia, adibide gutxitatik abiatuta ikasi behar duten sistemak. Horrelakoetan, lehiaketak antolatzen ditugu. Hori da nire lan egiteko modu bat. Lanak nik eta ikasleek bakarrik egin beharrean, ikertzaile-talde handi bati irekitzen diogu problema. Beraz, problemak planteatzen ditugu, eta soluzioa edozeinek emateko aukera zabaltzen dugu. Lan berri bat egin dezakegu beste lan batzuetan trebatutako sistema batekin.

**Aipatu duzunaren harira, sare neuronaletan estrategia asko daude; adibidez, GAN sareen estrategiak, beste sistema batekiko konpetentziatik ikasten duten sistemak. Ikaragarria! Etorkizun berria irekitzen ari da?**

Bai. GAN sareek iraultza ekarri dute azken urteetan, sare neuronalak trebatzeko garaian. Jendeak metodo berriak asmatzen ditu, eta metodo horiek usiatzeko ideia berriak ere bai. Egin dugun gauzetako bat da datu artifizial errealistak sortzea. Helburuetako bat pribatutasuna babestea da. Izan ere, datu

horiek, askotan, pribatutasun-kezkak ekartzen dituzte, edo balio komertziala dute; beraz, ezin dira besterik gabe zabaldu. Arazo handia izan da: datu pribatuak askatzeagatik salatu egin dituzte enpresa handi batzuk. Beraz, orain oso zuhurrak dira. Eta hori txarra da ikerketa-komunitatearentzat, iker-tzaileek ezin baitute problema interesgarri gehiago aztertu eta soluzioa bilatzen saiatu.

Beraz, lan hau egin dut New Yorkeko RPI institutu-ko lankideekin: GAN sareetan oinarritutako mezuak zatitu, gizabanakoei buruzko informaziorik ez duten datu artifizial errealistak sortzeko. Datu horiek datu errealeen propietate estatistiko guztiak gordetzen dituzte, eta, beraz, erabilgarriak dira ikerketarako.

Horrela, ikasleek erabil ditzakete sistemak trebatzeko. Arazoa da benetako aurkikuntzak egiteko ere erabili nahiko genituzkeela, eta horretarako ez dute balio. Benetako datuen propietateak gordeta, ikerketan erabili ahalko genituzke benetako aurkikuntzak egiteko. Saiatzen ari gara datu artifizial errealista horien mugak pixkanaka-pixkanaka zabaltzen.

**Eta funtzionatzen al du?**

Bai, biomedikuntzan, erregistro mediko faltsu asko sortu ditugu, oso informazio sentikorra baita. Oro har, datu sentikorrak dituzten enprekin kolaboratzen ari ginen, baina ez ziguten uzten datu-rik esportatzen. Aldiz, orain, zenbait segurtasun-edo pribatutasun-muga gaingitu zezaketen datuak sor ditzaketen ereduak esportatzen ditugu. Esperantza badut baliagarri izango zaiola zientzia-komunitateari. ●