EKAIA (2024), artikulua prentsan/article in press
https://doi.org/10.1387/ekaia.26327
Behin-behineko bertsioa (euskara-orrazketaren faltan).

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN: 0214-9001
e-ISSN: 2444-3255

Bootstrap ez-parametrikoa algoritmoen errendimenduaren konparaketarako estatistika bayestarraren alternatiba gisa

(Non-parametric bootstrap for algorithm performance comparison as an alternative to Bayesian statistics)

Izei Múgica, Usue Mori, Borja Calvo*
Intelligent Systems Group, Informatika Fakultatea, UPV/EHU
Manuel de Lardizabal 1, 20018 Donostia, Espainia

LABURPENA:
Adimen artifiziala eremu ia gehienetan txertatzen ari den garai honetan, ezinbestekoa da erabiltzen ditugun algoritmoen errendimenduen konparaketak egitea. Sarritan, algoritmo baten bitartez lortutako emaitzak erabilitako datuen izaeraren eta algoritmoen estokastizitatearen arabera aldakorrak dira, eta honek konparaketa zaildu egiten du. Testuinguru horretan, beraz, algoritmoen arteko konparaketa egitean, lortutako ziurgabetasunaren kuantifikazioa aspektu garrantzitsua da, baina gehienetan ez zaio nahikoa garrantzia ematen. Izan ere, ziurgabetasunaren kuantifikaziorako urteetan zehar erabili diren estatistika frekuentistako hipotesi-probek gabezia nabariak dituzte arlo honetan. Alternatiba gisa, estatistika bayestarra da soluzio onartuena gaur egun, baina badira beste metodo batzuk oraindik esploratu ez direnak. Artikulu honetan ziurgabetasuna ilustratzeko bootstrap ez-parametrikoaren erabilera aztertuko dugu, eta metodo honen portaera estatistika bayestarrarekin alderatuko dugu antzekotasunak eta desberdintasunak agerian jarriz.
HITZ GAKOAK: Ziurgabetasuna, estatistika bayestarra, birlaginketa metodoak, algoritmoen konparaketa

ABSTRACT: *In times where artificial intelligence is being incorporated into almost all application domains, it is essential to make comparisons between the results of the algorithms we use. It is often the case that the results obtained from an algorithm are dependent on their stochasticity and the nature of the data being used. This in turn can result in having greater difficulty in comparing algorithms with each other. Therefore, in such scenario, when comparing algorithms with each other, the quantification of the uncertainty is an important aspect, but it is generally not given sufficient importance. In fact, in order to quantify such uncertainty, the frequentist statistical tests that have been used over the years have obvious shortcomings in this regard. As an alternative, Bayesian statistics is the most widely accepted solution today, but there are other methods that have not yet been explored in depth. In this article we will examine the use of non-parametric bootstrap to illustrate this uncertainty, and we will compare the behavior of this method with that of Bayesian statistics, underlining the similarities and differences.*
KEYWORDS: Uncertainty, Bayesian statistics, resampling methods, algorithm comparison

***Harremanetan jartzeko/Corresponding author:** Usue Mori. Intelligent Systems Group, Informatika Fakultatea, UPV/EHU
Manuel de Lardizabal 1, 20018 Donostia, Espainia..
https://orcid.org/0000-0002-2057-1770, usue.mori@ehu.eus

Nola aipatu/How to cite: Múgica, Izei; Mori, Usue; Calvo, Borja (2024). «Bootstrap ez-parametrikoa algoritmoen errendimenduaren konparaketarako estatistika bayestarraren alternatiba gisa», Ekaia, DOI: https://doi.org/10.1387/ekaia.26327

Jasoa: maiatzak 20, 2024; Onartua: uztailak 24, 2024
ISSN 0214-9001-eISSN 2444-3225 / ©2024 UPV/EHU

1. Sarrera

Adimen artifiziala gure eguneroko bizitzako arlo eta ingurune guztietara hedatu den garai honetan, argi dago sistema horien erabilerak zenbait erronka eta arazo sortzen dituela. Horien artean, azken urteetan asko hitz egin da ereduen "kutxa beltza" izaerari buruz eta baita adimen artifizialak sortzen dituen alborapenei buruz. Kontu horiei erantzunez, komunitate zientifikoa buru-belarri jarri da lanean eredu gardenagoak eta justuagoak eraikitzeko asmoarekin [1, 2].

Hori hala izanik ere, badago hainbesteko garrantzia eman ez zaion arlo bat, baina izugarrizko eragina duena ereduen erabileran eta haiengan jartzen dugun konfiantzan. Arlo hori algoritmo eta ereduen ebaluazioa da, eta zehatzago esanda, lortutako emaitzen ziurgabetasunaren azterketa.

Kontu horien inguruan hausnarketa egiteko adibide sinpleenean oinarrituko gara. Adibidean suposatuko dugu bi algoritmo ditugula eta populazio baten gainean duten jarduna konparatu nahi dugula (adibide gisa, demagun n tamainako TSP problema ebazteko bi algoritmo estokastiko ditugula eta bietatik onena zein den jakin nahi dugula). Onartuko dugu ere badugula eraren bat algoritmoen exekuziotik lortutako emaitzen kalitatea neurtzeko. Kontuan izan, aztertu nahi dugun fenomeno honetan ziurgabetasuna iturri ezberdinetatik etorri daitekeela, besteak beste, algoritmoen ausazko izaeratik, populaziotik egindako lagin hautaketatik, etab. Beraz, bi algoritmok ematen dituzten emaitzen kalitateen arteko diferentzia X zorizko aldagaiaren bitartez adieraziko dugu, probabilitate banaketa ezezaguna duena. Helburua zorizko aldagai hau aztertzea izaten da, edo zehazkiago, aldagai horren batezbestekoa, μ, aztertzea.

Kontuan izan, populazio osoa infinitu elementuz osatuta egonda, ezinezkoa dela denetan algoritmoa probatzea. Hartara, ohiko prozedura esperimentazio bat burutzea da, non zenbait datu multzotan edo problema instantziatan (sarritan optimizazioaren alorrean) bi algoritmoak aplikatzen diren eta lortutako emaitzen diferentziak jasotzen diren, $x_1, x_2, ..., x_n$. Beste era batean esanda, X aldagaiaren laginketa bat dugu. Ohartu esperimentazio honetatik lortutako emaitzak ez direla deterministak, eta ziurgabetasuna iturri ezberdinetatik datorrela. Hasteko, hautatutako lagineko ale bakoitzean emaitza ezberdinak lortuko dira. Bestalde, hautatutako laginak berak ere badu eragina. Azkenik, algoritmoak beraiek estokastikoak izanik, emaitza ezberdinak eman ditzakete exekuzio bakoitzean.

$x_1, x_2, ..., x_n$ diferentzia balioen batezbestekoa egiten badugu μ diferentzien batezbestekoaren estimazio puntual bakar bat dugu, zenbaki bat. Baina zein da estimazio honen ziurgabetasuna? Zenbateraino fidatu gaitezke emaitza honetaz? Ba al dago ezberdintasun esanguratsurik algoritmoen emaitzen artean?

Galdera horri erantzuna emateko emaitzen azterketa estatistikoa burutzea da ohiko prozedura. Hasiera batean, eta urte askotan zehar, estatistika frekuentistatik proposatutako hipotesi-probak eta konfiantza tarteak ziren komunitateak aho-batez onartutako soluzioa [3]. Ordea, azkenaldian ezagun egin da prozedura horiek gabezia handiak dituztela [4, 5, 6], eta, ondorioz, estatistika bayestarraren erabilera hobetsi da [7]. Beste abantailen artean, estatistika bayestarrarekin lortutako emaitzak interpretagarriagoak dira eta emaitzen ziurgabetasuna era naturalean neurtzea ahalbidetzen du.

Alabaina, estatistika bayestarrak ere baditu bere gabeziak eta zailtasunak. Hasteko, matematikoki konplexua da eta urteetan zehar erabili dugun estatistika frekuentistatik kontzeptualki guztiz ezberdina. Bestalde, aspektu praktikoetan zentratuz, *a priori* banaketa batzuen aukeraketa ez da tribiala eta horrek eragina izan dezake emaitzetan.

Estatistika frekuentistaren baitan, badago algoritmoen konparaketa egiteko hain erabilia ez den metodo bat: bootstrap ez-parametrikoa. Hurbilketa hori oso intuitiboa eta kontzeptualki sinplea izanik ere, erabilitako estimatzaileen ziurgabetasuna kuantifikatzeko erabili daiteke. Hori horrela, artikulu honen helburua hau da: algoritmoen ebaluaziorako bootstrap ez-parametrikoaren erabilera estatistika bayestarrak eskaintzen duen alor teorikoarekin alderatzea eta uztartzea.

Horretarako, artikulua sei ataletan banatu dugu. 2. atalean, estatistika frekuentista eta bayestarraren sarrera bat egiten dugu bakoitzaren filosofia zein den azalduz eta ezberdintasunetan

arreta ipiniz. 3. atalean, lan honetan konparatuko ditugun bi teknikak apur bat sakonago aztertuko ditugu: korrelaziodun t-test bayestarra eta bootstrap ez-parametrikoa. Ondoren, 4. atalean, egindako esperimentuen diseinua azalduko dugu, eta 5. atalean lortutako emaitzak. Azkenik, 6. atalean ateratako ondorioak laburbilduko ditugu eta etorkizunerako ikerketa ildo batzuk proposatuko ditugu.

2. Ziurgabetasunarekin lan egiteko bi hurbilketa

Nahiz eta matematikan adituak ez izan, badaude gure eguneroko bizitzan etengabe erabiltzen ditugun termino matematiko batzuk. Horietatik arruntenetariko bat 'probabilitate' terminoa da, edozein gauzaren inguruan zalantzak edo ziurtasun falta dugula adierazteko erabiltzen dugu-na. Noski, egunerokoan termino horri ematen diogun erabilera eta haren inguruan egiten dugun interpretazioa ez dator beti bat matematikan esleitzen zaion interpretazioarekin. Izan ere, ez da kontzeptu tribiala, eta estatistika arloan badaude bi interpretazio ezberdin probabilitate kontzeptuarentzat, bi eskola pentsaera (eta lan egiteko era) sortu dituztenak: interpretazio frekuentista eta interpretazio bayestarra.

Eskola frekuentistarentzat probabilitatea zerbait objektiboa izan behar da, eta izenak adierazten duen bezala, maiztasun terminoarekin lotzen da. Hau da, errepikakorra den ausazko esperimentu bat baldin badaukagu, gertaera jakin baten probabilitatea esperimentuaren hainbat errepikapenetan gertaera horren maiztasun erlatiboa izango da.

Pentsa dezakegu interpretazio frekuentista naturalena edo intuitiboena dela, gehienbat lehen aipatutako adibideak erabiltzen direlako matematikoki ilustratzeko. Alabaina, interpretazio hori ez dator beti bat intuizioarekin, batik bat 'esperimentua' errepikakorra ez bada. Esate baterako, nola interpretatu dezakegu egun zehatz batean euria egiteko probabilitatea? Lehen aipatu bezala, gure eguneroko bizitzan askotan probabilitate terminoa gure usteak adierazteko erabiltzen dugu, eta hortaz, argudiatu daiteke probabilitatearen interpretazioa hain justu hori izan behar dela. Hori da eskola bayestarraren esentzia, non probabilitateak 'usteak' adierazteko erabiltzen diren edo, beste era batean esanda, fenomeno baten inguruan guk daukagun ziurgabetasuna.

Ezberdintasuna hobeto ulertzeko, adibide sinple bat jarriko dugu. Demagun txanpon bat dugula. Beraz, hura jaurtitzean, bi emaitza posible daude: aurpegia edo gurutzea. Jaurtiketa horren emaitza modelatzeko Bernoulli banaketa bat erabili dezakegu [1]. Banaketa horrek parametro bakarra du, θ, bi balio posibleren artean baten (aurpegiarena, adibidez) probabilitatea adierazten duena (bestearen probabilitatea, gurutzearena, $1-\theta$ izango da). Helburua θ parametroaren estimazio bat egitea bada, txanpona behin eta berriro jaurti eta laginketa bat lortuko dugu.

Bai estatistika frekuentistan eta bai bayestarrean ere, banaketa berdina erabiltzen da esperimentua modelatzeko, baina parametroaren tratamendua ezberdina da.

Estatistika frekuentistan asumitzen da modelatu nahi dugun esperimentu horretan benetako balio bat (eta bakar bat) dagoela parametroarentzat, nahiz eta balio hori ezezaguna izan. Estatistika eskola honetan, aurpegia ateratzeko probabilitatea, θ, gertaeraren maiztasun erlatiboa erabiliz estimatzen da, zenbaki baten bidez, kontuan izanda jaurtiketa kopurua infiniturantz doanean estimazio hori benetako probabilitatera hurbilduko dela. Kontuan izan estimazio hori egiteko laginketatik lortutako datuetan soilik oinarritzen garela.

Estatistika bayestarrean ordea, probabilitateak gure ziurgabetasuna modelatzeko erabiltzen direnez, parametroa bera zorizko aldagai moduan modelatzen dugu. Hau da, θ parametroak edozein balio har dezake, balio bakoitzak bere probabilitatea izanik (betiere bere domeinuan, hau da, 0 eta 1 artean), eta bere dentsitate-funtzioak adieraziko du nola banatzen den probabilitate masa domeinu horretan. Estimazioaren helburua parametro honen *a posteriori* banaketa lortzea izango da

[1] Hainbat jaurtiketetan lortzen den emaitza modelatu nahi badugu binomial bat erabiliko genuke, baina egoera sinplifikatzeko jaurtiketa bakarra kontsideratuko dugu.

eta horretarako Bayesen teorema erabiltzen da. Txanponaren adibidean, θ parametroaren banaketa honela lortu dezakegu:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

non, p funtzioak kasu jarraituan dentsitate funtzioa adierazten duen eta kasu diskretuan aldiz, probabilitate-masa funtzioa. Gainera, $p(\theta)$ θ parametroaren *a priori* banaketa da, hau da, banaketaren gainean dugun hasierako ustea; $p(D|\theta)$ laginketan lortutako datuen egiantza da (hau da, eskuragarri dugun lagina behatu izateko probabilitatea θ balioa emanda); eta $p(D)$ konstante normalizatzailea. Kontuan izan estimazio mota honek datuetatik ateratako informazio objektiboa gure usteekin nahasten duela. Hortik dator estatistika bayestarraren izaera "subjektiboa", urteetan zehar gabezia moduan ikusi dena, baina abantaila asko eskeintzen dituena.

Parametroak eta lortutako estimazioak interpretatzeko moduan ezberdintasun horrek inplikazio sakonak ditu. Esate baterako, θ parametroaren interpretazio frekuentista egiten badugu, hau da zenbaki bat dela asumitzen badugu, $p(\theta > 0.5)$ kalkulatzeak ez dauka zentzurik. Aldiz, θ ikuspuntu bayestarretik ulertzen badugu $p(\theta > 0.5)$ probabilitatea kalkulatzeak zentzu handiagoa dauka. Hurrengo ataletan ikusiko dugun bezala, bi hurbilpenak erabil daitezke algoritmoen edota ereduen errendimenduak konparatzean ziurgabetasuna aztertzeko, baina bi hurbilpenetan lortutako emaitzak interpretatzean oso ondo ulertu behar da bakoitzaren atzean dagoen 'filosofia'.

3. Algoritmoen emaitzen konparaketarako metodoak

Honelako egoera batean, algoritmoen emaitzak konparatzeko eta lortutako emaitzen analisi estatistikoa egiteko modurik ohikoena estatistika frekuentistan definitutako hipotesi-proba bat burutzea da [3], adibiderik oinarrizkoena t-test delakoa izanik. Hipotesi-probetan bi hipotesi definitzen dira, gure problema partikularrean (algoritmoen exekuzio-emaitzen diferentzien batezbestekoaren balioa aztertzea) honela definituko direnak: $H_0 : \mu = 0$ (hipotesi nulua), $H_1 : \mu \neq 0$ (hipotesi alternatiboa). Hau da, diferentzien batezbestekoa 0 da, eta beraz, bi algoritmoen ontasun-maila berbera da, edo batezbestekoa 0ren desberdina da, eta beraz, portaera desberdina dute. Jarraian, metodo horiek, hipotesi nulua egia balitz, esperimentuan lortutako $\bar{x}$ (algoritmoen exekuzio-emaitzen diferentzien batezbestekoa) estimazioa edo hori baino balio arraroagoa lortzeko probabilitatea kalkulatzen dute. Balio hori da, hain zuzen, denok ezagutzen dugun p-balio famatua. Ondoren, p-balioa txikia bada (normalean 0.05 baino txikiagoa), lortutako balioa hipotesi nulupean oso inprobablea dela ulertzen da eta, beraz, hipotesi nulua errefusatzen da. Kontrako kasuan, aldiz, p-balioa handia denean, hipotesi nulua ezin dugu errefusatu eta bi algoritmoen arteko diferentziak esanguratsuak direla esateko nahikoa ebidentziarik ez dugula ondorioztatu dezakegu soilik.

Horrelako hurbilketek hainbat arazo dituzte [4, 6, 5], baina horietatik aipagarrienetako bat interpretazioa da. Testaren emaitzak zuzenean esaten digu, ezarritako konfidantzarekin, ea asumitu dezakegun batezbestekoen artean diferentziak dauden ala ez, baina erantzun hau Benavoliren hitzetan, erantzun "txuri-beltza" da [7]. Ziurgabetasuna testaren barne-kalkuluetan kontutan hartzen bada ere, testaren emaitzak ez digu informaziorik ematen horren inguruan.

Gabezia horiei aurre egiteko asmoz, artikulu honetan algoritmoen konparaketa egiteko bi metodo alternatibora joko dugu. Hasteko, alternatiba nagusi bezala aurkeztu den estatistika bayestarreko testak erabiliko ditugu [8]. Bestalde, soluzio berritzaile gisa, estatistika frekuentistan oinarria duen bootstrap ez-parametrikoa [9] erabiliko dugu. Bi metodoen arteko konparaketa egingo dugu, batik bat, ziurgabetasunaren kuantifikazioari begira bakoitzaren ahalmenak eta ahuldadeak azpimarratzeko asmoz.

3.1. Estatistika Bayestarreko korrelaziodun t-testa

Aurreko ataletan aipatu dugun moduan, estatistika bayestarrean hurbilketa (eta interpretazioa) oso ezberdina da. Adibide sinple batekin hasteko, demagun bi algoritmoren emaitzen arteko diferentziak μ eta σ parametro ezezaguneko banaketa gaussiarrari jarraitzen diola. Estatistika bayestarrean bi parametro horiek ere zorizko aldagaiak direla onartzen da. Gure kasuan μ parametroa interesatzen zaigu, hain justu diferentzien batezbestekoa adierazten duelako (hasieran ipinitako formulekin bat egiteko, pentsa μ parametroa θ parametroa dela eta haren inguruko analisia egin nahi dugula). Hortaz, *a priori* banaketa batetik abiatuz, hurbilketa bayestarrean esperimentazioan lortutako emaitzak erabiliko ditugu μ parametroaren *a posteriori* banaketa lortzeko, banaketa horrek algoritmoen diferentziaren batezbestekoari buruz espero dezakeguna adieraziko duelakoan. Kontuan izan probabilitate banaketa bat emateak estimazio puntual bat lortzeko aukera ematen duela, itxaropena erabiliz, adibidez, baina gainera, ziurgabetasunari buruzko informazioa ere ematen digula, adibidez *a posteriori* banaketaren bariantzaren bitartez.

Algoritmoen konparaketarako estatistika bayestarra erabiltzeko, zenbait erabaki hartu behar ditugu, besteak beste, estimatu nahi dugun parametroaren *a priori* banaketa, eta datuen egiantz funtzioaren forma. Hainbat aukera egonik ere, gure testuinguruan ohikoena korrelaziodun t-test bayestarra [10] erabiltzea izaten da. Metodo honetan hurrengo erabakiak hartu ohi dira:

1. **Egiantz funtzioa.** Metodoan datuak (hau da, esperimentuan lortutako emaitzen arteko diferentziak) honako banaketatik sortu direla asumitzen da:

$$\boldsymbol{x} = \mathbf{1}\mu + \boldsymbol{v},$$

non $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ esperimentuan lortutako diferentzien bektorea den, $\mathbf{1}$ n tamainako batekoen bektorea den eta $\boldsymbol{v}$ n dimentsioko banaketa gaussiar batetik ateratako zarata den, hots, $\boldsymbol{v} \sim MVN(\mathbf{0}, \Sigma_{n\times n})$. Azkenik, $\Sigma_{n\times n}$ kobariantza matrizearen balioak honela definitzen dira: $\Sigma_{ii} = \sigma^2$ eta $\Sigma_{ij} = \rho\sigma^2$, $i \neq j$ denean, non σ^2 bariantza eta ρ korrelazio-koefizientea, [-1,1] tartean dagoena (neurketak askeak ez direnean, haien arteko korrelazioa aintzat hartzen da, adibidez, balidazio gurutzatua erabili bada eta neurketak datu-base berearen gainean eginda badaude, baina bestela, gure kasuan bezala, $\rho = 0$ hartzen da), diren. Datuen gaineko eredu hori asumituz gero, egiantz funtzioak honako itxura hartzen du:

$$p(\boldsymbol{x}|\mu, \Sigma) = \frac{\exp(-\frac{1}{2}(\boldsymbol{x} - \mathbf{1}\mu)^T\Sigma^{-1}(\boldsymbol{x} - \mathbf{1}\mu))}{(2\pi)^{n/2}\sqrt{|\Sigma|}}$$

2. ***A priori* banaketa.** Normal-Gamma banaketa bat aukeratzen da, $\{\mu_0 = 0, k_0 \to \infty, a = -1/2, b = 0\}$ parametroduna.

Aukeraketa horiek ez dira ausazkoak: hautatutako *a priori* banaketa definitu dugun egiantz funtzioaren konjugatua da, eta ondorioz, a posteriaren kalkulua analitikoki egin daiteke (horrela ez balitz, zenbakizko metodoak, laginketa metodoak, etab. erabili beharko genituzke *a posteriori* banaketa tratatzeko), emaitza $n - 1$ askatasun graduko Student-en t banaketa bat izanik:

$$p(\mu|\boldsymbol{x}, \mu_0, k_0, a, b) \equiv St\left(\mu; n-1, \bar{x}, \left(\frac{1}{n} + \frac{\rho}{1-\rho}\right)\hat{\sigma}^2\right)$$

A priori banaketarentzat aukeratu ditugun parametroen balioekin, lortzen dugun *a posteriori* banaketa estatistika frekuentistako t-test korrelatuan p-balioa kalkulatzeko erabiltzen den banaketa bera da. Baina gogoan izan, alde batetik, estatistika frekuentistan banaketa bera ez dela itzultzen, bakarrik p-balioa kalkulatzeko erabiltzen da, eta bestetik, bi kasuetan interpretazioa ezberdina dela. Zentzu horretan, argi dago estatistika bayestarrak informazio gehiago eta interpretatzeko

errazagoa dena ematen digula, interesatzen zaigun parametroaren, μ-ren, probabilitate banaketa bat, alegia.

Alabaina, galdetu genezake ea zein punturaino banaketa horrek testuinguru zehatz batean 'benetan' dagoen (edo izan beharko genukeen) ziurgabetasuna ondo adierazten duen. Ez hori bakarrik, ba al dago eraren bat ikuspuntu frekuentistatik antzeko zerbait lortzeko? Bestalde, test bayestar jakin honen izaera parametrikoak (bai *a priori* banaketari bai eta egiantz funtzioari ere dagokionez) ere baditu bere mugak, besteak beste, zer gertatzen da gure datuek ez badiote asumitutako ereduari jarraitzen? Alternatiba ez-parametriko bayestarrak ere existitzen diren arren, ez daude hain hedatuta. Hori guztia kontutan izanda, hurrengo ataletan alternatiba bat proposatuko dugu, bootstrap ez-parametrikoa tresna bezala erabiliz.

3.2. Bootstrap ez-parametrikoa algoritmoen konparaketarako

Behin baino gehiagotan aipatu dugun moduan, algoritmoak konparatzean funtsezkoa da egindako esperimentaziotik atera ditugun ondorioen ziurgabetasuna, nolabait, neurtzea. Egin dezakegun analisi estatistiko sinpleena esperimentazioaren emaitzei (algoritmo horien 'erabileraren' lagin bati) estimatzaileak aplikatzea da. Adibidez, artikulu honetan erabiltzen dugun kasuan, algoritmoen arteko diferentziaren itxaropenaren estimatzailea, $\hat{\mu}$, estatistika frekuentistan laginaren batezbestekoa, $\bar{X}$, izango da. Estimatzaile hau daukagun laginari aplikatuz estimazio puntual bat eta bakarra lortuko dugu, $\bar{\boldsymbol{x}}$.

Adibide honetan argi ikusten da estimatzaileak ere zorizko aldagaiak direla, izan ere, lagin ezberdinei aplikatuz gero (esperimentazio ezberdinei), estimazio ezberdinak lortuko ditugu. Kasu sinple horretan, nola jakin dezakegu zein den egin dugun estimazioaren inguruko ziurgabetasuna? Esperimentazioa hainbat aldiz errepikatuz gero, zein punturaino izango dira ezberdinak lortuko genituzkeen estimazioak?

Testuinguru kontrolatuetan galdera horri teorikoki erantzun diezaiokegu, estimatzailearen probabilitate banaketa analitikoki kalkulatu daitekeelako. Zoritxarrez, kasu orokorrean hori ez da posible. Estimatzailearen banaketaren hurbilpen bat lortzeko existitzen diren tekniken artean bat landuko dugu lan honetan: bootstrap ez-parametrikoa.

Bootstrap metodoa birlaginketa teknika bat da [9] eta bere oinarrian *plug-in* printzipioa dago. Printzipio honen ideia sinplea da: zerbait ezagutzen ez baduzu, estimazio batekin hurbildu. Gure helburua estimatzailearen banaketa probabilistikoa ezagutzea da, baina analitikoki ezin dugunez lortu, bere espresioaren hurbilketa bat egingo dugu berau laginduz. Horretarako bi pausu burutu beharko genituzke: i) jatorrizko populazioa lagindu eta ii) lagin horri estimatzailea aplikatu. Alabaina, kasu honetan, ez dugu jatorrizko populazioa (algoritmoak aplikatzean lortu daitezkeen emaitza posible guztiak) ezagutzen, baina badugu horren hurbilketa bat, gure hasierako lagina. Beraz, prozesuaren lehenengo pausuan populazioa lagindu beharrean hasierako lagina (esperimentazioaren emaitzak) laginduko dugu eta, ondoren, birlaginketa horietako bakoitzari aztertu nahi dugun estimatzailea aplikatuko diogu. Ikusten dugun moduan, etengabe *plug-in* printzipioa aplikatzen dugu, ezezaguna dena estimazio bidez hurbilduz.

Zehatzago esanda, hasierako lagina, $(x_1, ..., x_n)$, gure populazioa dela onartzen badugu eta x_i guztiak ekiprobableak izanik, bakoitzaren probabilitatea $1/n$ izango da. Beraz, lehenengo pausoan "populazio" hori laginduko dugu ausazko laginketa itzuleraduna erabiliz, n elementuko bootstrap lagin bat eraiki arte, $\boldsymbol{x_1^*}$. Lagin horri gure estimatzailea aplikatuko diogu eta estimazio bat lortuko dugu, gure kasuan $\bar{\boldsymbol{x}}_{\boldsymbol{1}}^*$. Prozedura hori behin eta berriro errepikatuko dugu, B bootstrap lagin, $\{\boldsymbol{x_1^*}, \boldsymbol{x_2^*}, ..., \boldsymbol{x_B^*}\}$ eta dagozkien B estimazio lortu arte. Prozedura horri bootstrap ez-parametrikoa deritzo eta lortutako estimatzaile multzoak B tamainako estimatzailearen bootstrap laginketa bat osatzen du, estimatzailearen banaketa hurbiltzeko erabiliko duguna.

4. Esperimentuen diseinua

Atal honetan algoritmoen konparaketarako erabiltzen diren hurbilpenak gure proposamenarekin (hurbilpen frekuentistaren baitan kokatuta) alderatuko dira, hainbat esperimenturen bitartez. Hurbilpen frekuentista eta bayestarra izaeraz desberdinak diren arren, bakoitzak bere indargune eta ahuleziak dituelarik, biak zein punturaino bateragarriak izan daitezkeen aztertuko dugu jarraian jorratuko diren esperimentuekin.

Hurbilpenak konparatu ahal izateko, algoritmo-pare jakin baten exekuzioaren portaera era artifizialean simulatzea interesgarria izan daiteke, simulazio horren gainean hurbilpen frekuentista eta bayestarra aplikatuta, bien arteko desberdintasunak azaleratuko baitira (desberdintasunik egonez gero, jakina). Zehatzago esanda, algoritmo-pare baten exekuzioen emaitza banaketa-pare jakin batzuekin adieraziz gero, banaketa horiek lagindu genitzake, diferentziak kalkulatu, birlaginketa-metodoa eta korrelaziodun t-test bayestarra aplikatu, etab. Gainera, jatorrizko banaketak ezagutzen ditugunez, X diferentzia-aldagaiaren benetako banaketa ere kalkulatu dezakegu, bertatik itxaropena, desbideratze estandarra eta bestelakoak ateratzeko eta informazio hori bi hurbilpenek eskaintzen dituzten emaitzekin uztartzeko, 1. irudian ageri den moduan. Jakina, egoera erreal batean ez dugunez X aldagaiaren banaketaren inguruko informazio gehigarri hori izango, benetako egoera bat ere tratatuko dugu artikulu honetan, zeinaren inguruko xehetasunak 4.1. atalean deskribatuko diren.

Testuinguru horretan, bost esperimentu burutu ditugu, guztiak oinarri beraren gainean gauzatuak. Hasteko, simulazio kasuetarako, algoritmoek lortutako emaitzak banaketa banarekin modelizatuko dira. Adibidez, A algoritmoak problema jakin bat ebaztean lortzen dituen emaitzak banaketa gaussiar bat jarraitzen dutela asumitu genezake. Jarraian, 50 tamainako lagin bana aterako dugu banaketa horietatik. Azken esperimenturako, kasu errealerako hain zuzen, jatorrizko banaketak ezezagunak direnez, 50 tamaina finkoko lagin batetik abiatuko gara zuzenean (berehala tratatuko dugu datu horien izaera). Hemendik aurrera, esperimentu guztietan prozedura berbera aplikatuko da, sekuentzialki gauzatu beharreko lau urratsez osatuta dagoena:

1. **Laginen arteko diferentzia kalkulatu.** Demagun bi algoritmoren exekuzioen portaerak konparatzeko $\boldsymbol{x}^{(1)} = (x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)})$ eta $\boldsymbol{x}^{(2)} = (x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)})$ laginak lortu direla, hurrenez hurren. Orduan, hurrengo kalkulua egingo da:

$$\boldsymbol{x} = \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} = (x_1^{(1)} - x_1^{(2)}, x_2^{(1)} - x_2^{(2)}, ..., x_n^{(1)} - x_n^{(2)})$$

2. **Bootstrap bidezko birlaginketa eta dentsitatearen estimazioa.** Bootstrap teknikaren bidez aurreko $\boldsymbol{x}$ birlaginduko da, $n = 50$ tamainako $B = 50$ lagin sortuz. Lagin bakoitzeko haren batezbestekoa kalkulatuko da, diferentzien banaketaren itxaropenaren estimazio-puntual gisa. Hortaz, $B = 50$ estimazio-puntualez osaturiko $\hat{\boldsymbol{\mu}}$ lagin bat lortuko da:

$$\hat{\boldsymbol{\mu}}^* = (\bar{\boldsymbol{x}}_{\mathbf{1}}^*, \bar{\boldsymbol{x}}_{\mathbf{2}}^*, ..., \bar{\boldsymbol{x}}_{\mathbf{50}}^*)$$

   Ohartu lagin hori μ-ren estimatzailearen banaketaren bootstrap hurbilketa bat dela. Dena den, aurrerago azalduko den moduan, hurbilpen bayestarrarekin konparaketak egiteko, μ-ren estimatzailearen banaketa jarraitu bat lortzea interesgarria izango da, erabiliko dugun konparaketa-neurriak bi banaketak diskretuak ala jarraituak izatea eskatzen duelako. Hori dela eta, $\hat{\boldsymbol{\mu}}$ laginetik abiatuta, banaketa hori hurbilduko da kernel dentsitatearen estimazio (KDE) teknika erabiliz [11]. Teknika hori, hitz gutxitan, banaketa enpirikoen ideiatik abiatzen da, f dentsitate-funtzio bat hari dagokion lagin baten bidez estimatuz, honako formula honen arabera:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

non n lagin-tamaina den, K estimaturiko dentsitate-funtzioa leuntzen duen kernel deritzon funtzio bat den eta h leuntasun-maila kontrolatzen duen banda-zabalera den. Gure kasuan, $x_i = \bar{\boldsymbol{x}}_{\boldsymbol{i}}^{*}$ dela ere ohartu.

Argi dago erabiltzen diren kernel funtzioek eta banda-zabalerek eragina dutela estimazioan, eta ikuspegi horri jarraiki, interesgarria litzateke parametro horien konfigurazio desberdinetarako emaitzak nola aldatzen diren ikustea. Edonola ere, limitearen teorema zentrala kontuan hartuz, μ-ren estimatzailearentzat banaketa gaussiar bat espero daitekeenez, eta bestetik, h-ren doikuntzarako literaturan emaitza onak ematen dituen heuristiko bat erabili ohi denez [12, 13], $\hat{f}_h(x)$ kalkulatzeko, kasu guztietan kernel gaussiarra eta heuristiko hori erabiliko dira. Hau da:

$$K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2},$$

$$h = 0.9 \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) n^{-\frac{1}{5}},$$

non $\hat{\sigma}^2$ lagin-kuasibariantza den, IQR kuartil arteko heina eta n intereseko banaketa hurbiltzeko erabiliko den laginaren tamaina.

3. **Korrelaziodun t-test bayestarraren aplikazioa.** Hurbilpen bayestarra kontuan hartuz, $\boldsymbol{x}$ laginaren gainean korrelaziodun t-test bayestarra aplikatuko da, *a priori* banaketarentzat 3.1. atalean zehazturiko banaketa hartuz, dagozkion parametroekin eta $\rho = 0$ hartuz (suposatzen da behaketen artean korrelaziorik ez dagoela). Hartara, μ-ren *a posteriori* banaketa honako Student banaketa hau izango da:

$$St\left(\mu;\ n-1,\ \bar{\boldsymbol{x}},\ \frac{\hat{\sigma}^2}{n}\right),$$

$\bar{\boldsymbol{x}}$ lagin-batezbestekoa izanik.

4. **Banaketen arteko diferentzia estatistikoaren kalkulua.** Azkenik, hurbilpen bakoitzarekin lorturiko banaketen arteko diferentzia kalkulatuko da. Horretarako, Jensen-Shannon dibergentzia proposatzen da neurri gisa, hurrengo formularen bidez emana datorrena:

$$D_{JS}(f \mid\mid g) = \frac{1}{2} D_{KL}(f \mid\mid q) + \frac{1}{2} D_{KL}(g \mid\mid q),$$

non f eta g konparatu beharreko banaketak diren, $q = \frac{1}{2}(f+g)$ den eta $D_{KL}\ (f \mid\mid q)$ zein $D_{KL}\ (g \mid\mid q)$, berriz, Kullback-Leibler dibergentziak. Azken horiei dagokionez, f eta g bi banaketa emanda, Kullback-Leibler dibergentzia hurrengo formularen bidez kalkulatzen da:

$$D_{KL}(f \mid\mid g) = \int_{-\infty}^{\infty} f(x)\, log\left(\frac{f(x)}{g(x)}\right) dx.$$

Banaketen arteko konparaketa egitean oso tipikoa da Kullback-Leibler dibergentzia erabiltzea, baina Jensen-Shannon dibergentzia erabiltzea proposatzen dugu, azken horrek lehenak dituen hainbat ahulezia gainditzen dituelako. Hasteko, kontuan izanik edozein f eta g banaketa emanda, $D_{KL}(f \mid\mid g) \geq 0$ dela, hau da, goitik bornatua ez dagoela, lortutako balioak interpretatzea ez da tribiala. Adibidez, $D_{KL}(f \mid\mid g) = 100$ izateak ez du adierazten balio handia edo txikia denik. Are gehiago, Kullback-Leibler dibergentzia ez da simetrikoa, eta horrek interpretagarritasuna are gehiago zailtzen du. Jensen-Shannon dibergentziaren kasuan, berriz, alde batetik, edozein f eta g banaketa emanda, $0 \leq D_{JS}(f \mid\mid g) \leq 1$ betetzen da; eta bestetik, simetrikoa da. Beraz, interpretagarriagoa da.

Arestian deskribaturiko lau urratsak jarraituz, hurbilpen frekuentistaren eta bayestarraren artean desberdintasunik ageri den aztertu ahalko da.

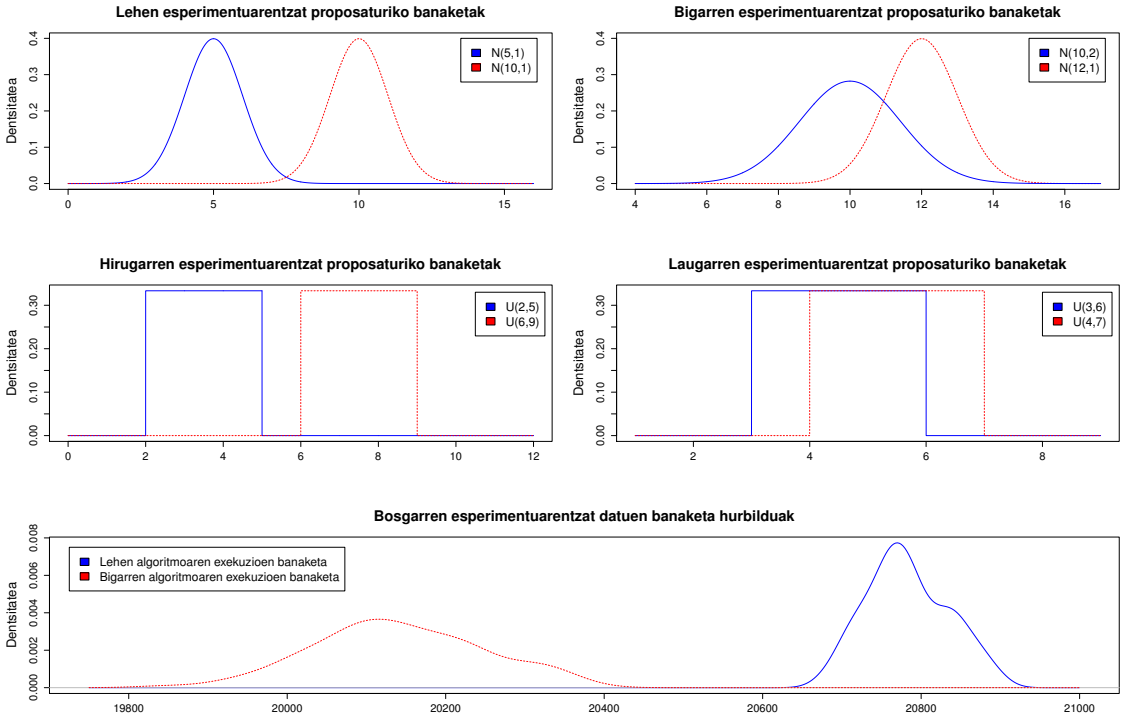
4.1. Esperimentaziorako datuak

Atalaren hasieran aipatu bezala, esperimentazioa aurrera eramateko lau simulazio eta kasu erreal bat kontuan hartuko dira. Atal honetan horren guztiaren inguruko xehetasunak eskainiko dira.

Simulazioekin hasteko, bi banaketa-familia proposatzen dira lehen lau esperimentuak burutzeko:

- **Banaketa gaussiarrak.** Bi egoera ezberdin aztertuko dira. Lehenengoan, algoritmo baten emaitzak $\mathcal{N}(5,1)$ banaketa jarraituko du eta bestearenak $\mathcal{N}(10,1)$ banaketa. Bigarrenean, algoritmoen emaitzek jarraituko dituzten banaketak $\mathcal{N}(10,2)$ eta $\mathcal{N}(12,1)$ izango dira, hurrenez hurren.
- **Banaketa uniformeak.** Kasu honetan ere bi egoera izango ditugu. Lehenengoan algoritmoen emaitzen banaketak $\mathcal{U}(2,5)$ eta $\mathcal{U}(6,9)$ izango dira, hurrenez hurren, eta bigarrenean, berriz, $\mathcal{U}(3,6)$ eta $\mathcal{U}(4,7)$, hurrenez hurren.

Esperimentuetan erabiliko diren datuen banaketak 1. irudian irudikatuta daude.



1. irudia. Esperimentuetan erabilitako banaketak.

Azkenik, datu errealei dagokienez, Linear Ordering Problem (LOP) optimizazio problemaren zenbait instantziaren ebazpenerako aplikaturiko bi algoritmo estokastikoren emaitzak erabili dira. Zehatzago esanda, beste [14] artikulu zientifikoko esperimentaziorako sortutako 50 LOP instantzietan bi optimizazio algoritmoren emaitzak erabili dira. Datu hauen generazio prozesua [14] artikuluan azalduta dago xehetasunez, eta beraz, artikulu hau ez luzatzearren ez ditugu hemen aipatuko. Bakarrik komeni da aipatzea bi algoritmoen emaitzek jarraitzen dituzten banaketak ezezagunak direla.

5. Emaitzak eta eztabaida

Aurreko atalean deskribaturiko irizpideak eta metodologia kontuan hartuz, atal honetan esperimentu bakoitzean lorturiko emaitzak aurkeztuko dira. Zehazkiago esanda, hurbilpen bayestarraren eta frekuentistaren arteko desberdintasunak islatu ditzaketen Jensen-Shannon dibergentziak aurkeztuko dira esperimentu bakoitzeko. Emaitzak 1. taulan jasota daude.

		JS dibergentzia	
		Boot+KDE	Bayes
1. esperimentua (Gaussiarrak)	Boot+KDE	0	0.011
	Bayes	0.011	0
2. esperimentua (Gaussiarrak)	Boot+KDE	0	0.013
	Bayes	0.013	0
3. esperimentua (Uniformeak)	Boot+KDE	0	0.007
	Bayes	0.007	0
4. esperimentua (Uniformeak)	Boot+KDE	0	0.005
	Bayes	0.005	0
5. esperimentua (Errealak)	Boot+KDE	0	0.006
	Bayes	0.006	0

1. taula. Bost esperimentuei dagozkien Jensen-Shannon dibergentziaren emaitzak.

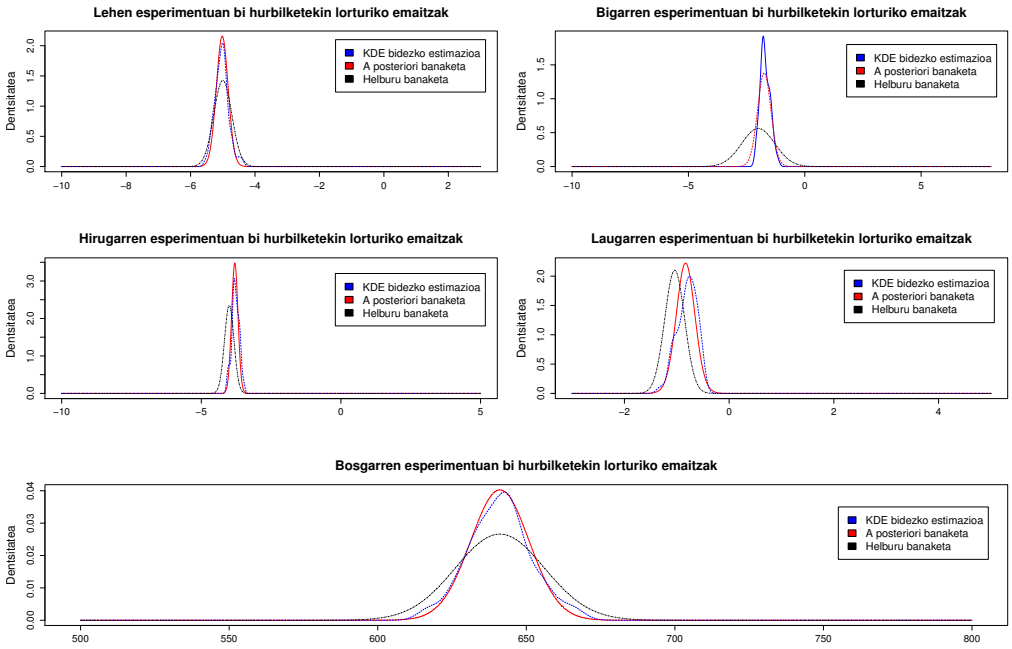
Taulak agerian uzten du hurbilpen frekuentistak eta bayestarrak eskaintzen dituzten emaitzak oso antzekoak direla, jatorrian paradigma desberdinak izan arren. Honakoa oso interesgarria da, hein batean bootstrap ez-parametrikoa estatistika bayestarraren alternatiba sendoa izan daitekeela esan genezakeelako; hau da, biek ondorio berera garamatzate, helburu desberdina badute ere. Lagungarri gisa, 2. irudiak esan berri dugunaren ideia interesgarri bat eskaintzen du. Itxaropenari zein bariantzari begira, oso antzekoak dira bi banaketak, eta beraz, μ estimatzeko itxaropena edo moda hartuta, adibidez, pareko emaitzak lortzea espero da. Are gehiago, oro har helburu banaketaren antz handia dute, nahiko zentratuak egonik, bariantza txikiagoa aurkezten duten arren.

Horren guztiaren gainean naturala den galdera bat agertzen da: bi paradigmek ongi kuantifikatzen al dute ziurgabetasuna? Hau da, hurbilpen frekuentistaren bidez lorturiko estimatzailearen banaketak eta hurbilpen bayestarraren bidez lorturiko *a posteriori* banaketak benetako balioaren inguruko ideia arrazonagarria eskaintzen al dute? Izan ere, 2. irudiko emaitzetan badirudi, oro har, hurbilpen frekuentistan zein bayestarrean oinarrituriko analisietatik informazio aberatsa erauzi dezakegula, baina konfiantza sentsazio hori benetakoa edo faltsua al da? Galdera horri erantzuna ematea ez da ataza batere tribiala, eta etorkizunerako lan gisa oso interesgarria izan daitekeelakoan gaude.

6. Ondorioak eta etorkizuneko lana

Lan honetan adimen artifizialean funtsezkoa den arazo bati heldu diogu, algoritmoen (edota ereduen) konparaketaren inguruan sortzen den ziurgabetasunaren kuantifikazioa. Azken urteetan algoritmoen emaitzen azterketa estatistikoaren inguruan zenbait planteamendu berri egin dira, asko estatistika frekuentistak eskaintzen dituen hurbilketa klasikoek (test estatistikoek) dituzten gabeziak gainditzeko asmoarekin. Ildo horretatik, estatistika bayestarra alternatiba aberasgarri moduan proposatu da, emaitza sinpleak eman beharrean (bai/ez motakoak), intereseko parametroen inguruan banaketa probabilistikoak eskaintzen dituelako.

Artikuluan antzerako informazioa eskaintzeko gaitasuna duen beste metodo estatistiko bat esploratu dugu: bootstrap ez-parametrikoa. Ikuspuntu filosofiko batetik bi metodoek ideiak ezberdinak badituzte ere, birlaginketa metodoek egindako estimazioen inguruan dagoen ziurgabetasuna-



2. irudia. Hurbilketa frekuentista eta bayestarra erabiliz, egindako bost esperimentuetan lorturiko banaketen irudikapena.

ren ideia ematen digute, estimatzaileen banaketa hurbilketen bidez. Egindako esperimentazioan bi hurbilketek ematen dituzten banaketak alderatu ditugu, datu simulatuetan zein errealetan aplikatuz.

Ondorioak garbiak dira: hurbilketa bakoitzaren helburua desberdina izanagatik ere, biek toki berera garamatzate. Hurbilpen frekuentistaren bidez lorturiko estimatzailearen banaketa hurbildua eta hurbilpen bayestarraren bidez lorturiko *a posteriori* banaketa elkarren oso antzekoak dira. Are gehiago, helburu banaketaren antz handia dutela ere ikusteko aukera izan dugu (bereziki kokapenari begira, bariantza apur bat txikiagoa duten arren). Jakina, lan honetan egindako esperimentazioaren emaitzak oso mugatuak eta atarikoak dira, baina joera mantenduko delakoan gaude.

Hortaz, horren gainean plantzeatzen den ikerketa-lerro naturalak ziurgabetasunaren kuantifikazioarekin du lotura, bi hurbilpenen bidez burutzen diren analisiek ondorio egokietara garamatzatzen jakiteko. Zer esan nahi du, baina, ‘egokiena’ izateak? Ez da erantzun azkarreko galdera, inondik ere, baina planteatzen den bidea astiro tratatuz gero, jakituria interesgarria erauzi daitekeela pentsatzen dugu. Birlaginketa metodoen erabilera zabalagoa (beste estimatzaile konplexuagoen erabilera, adibidez) eta metodo horien onuren eta gabezien eragina algoritmoen emaitzen analisian aztertzeak ere lagun diezagukeela esan genezake.

Erreferentziak

[1] W. Saeed eta C. Omlin, 2023, «Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities», *Knowledge-Based Systems*, **263**, 110273.

[2] S. Caton eta C. Haas, 2024, «Fairness in machine learning: A survey», *ACM Computing Surveys*, **56**(7), 166.

[3] J. Demšar, 2006, «Statistical comparisons of classifiers over multiple data sets», *Journal of Machine Learning Research*, **7**, 1–30.

[4] R. L. WASSERSTEIN eta N. A. LAZAR, 2016, «The asa's statement on p-values: context, process, and purpose», *The American Statistician*, **70**(2), 129–133.

[5] S. GREENLAND, S. J. SENN, K. J. ROTHMAN, J. B. CARLIN, C. POOLE, S.N. GOODMAN eta D. G. ALTMAN, 2016, «Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations», *European journal of epidemiology*, **31**(4), 337–350.

[6] S.N. GOODMAN, 2008, «A dirty dozen: Twelve p-value misconceptions», *Seminars in Hematology*, **45**(3), 135–140.

[7] A. BENAVOLI, G. CORANI, J. DEMŠAR eta M. ZAFFALON, 2017, «Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis», *Journal of Machine Learning Research*, **18**(77), 1–36.

[8] J. ROJAS-DELGADO, J. CEBERIO, B. CALVO eta J. A. LOZANO, 2022, «Bayesian performance analysis for algorithm ranking comparison», *IEEE Transactions on Evolutionary Computation*, **26**(6), 1281–1292.

[9] B. EFRON, 1979, «Bootstrap methods: Another look at the jackknife», *The Annals of Statistics*, **7**(1), 1 – 26.

[10] G. CORANI eta A. BENAVOLI, «A bayesian approach for comparing cross-validated algorithms on multiple data sets», **100**(2), 285–304.

[11] M. ROSENBLATT, 1956, «Remarks on some nonparametric estimates of a density function», *The Annals of Mathematical Statistics*, **27**(3), 832 – 837.

[12] B. W. SILVERMAN, 1986, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

[13] W.N. VENABLES eta B. D. RIPLEY, 2002, *Modern Applied Statistics with S*, Springer, New York.

[14] B. CALVO, J. CEBERIO eta J. A. LOZANO, 2018, «Bayesian inference for algorithm ranking analysis», *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Orria 324–325.