Egzakt, véges mintás, sztochasztikus garanciák gépi tanulási módszerekhez

Tamás Ambrus

Magyar Kutatási Hálózat (HUN-REN), Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI), Budapest, Magyarország
Eötvös Loránd Tudományegyetem (ELTE), Matematikai Intézet, Budapest, Magyarország
E-mail: tamasamb@sztaki.hun-ren.hu

Beérkezett: 2023. november 30.; elfogadva: 2024. január 18.

Összefoglalás

A gépi tanulás a mesterségesintelligencia-kutatások egyik fő pillére, melynek matematikai hátterét a statisztikus tanuláselmélet biztosítja. A gépi tanulási módszerek bizonytalanságának meghatározása esszenciálissá vált számos alkalmazás esetében, többek között a biztonság, a stabilitás és a minőség garantálása érdekében. Ebben a tanulmányban újramintavételező eljáráson alapuló konfidenciahalmaz-becsléseket mutatunk be, amelyek eloszlásfüggetlen és véges mintás korlátokat biztosítanak a becslések bizonytalanságára. A konfidenciahalmazokat egzakt és konzisztens rangtesztek segítségével konstruáljuk meg, és számos példán szemléltetjük. A várható értékre és a feltételes várhatóérték-függvényre adunk halmazbecslést felügyelt tanulási feladatok (regresszió és osztályozás) esetében.

Kulcsszavak: statisztikus tanuláselmélet, bizonytalanság vizsgálat, konfidenciahalmazok

Exact, finite sample, stochastic guarantees for machine learning methods

Ambrus Tamás

Hungarian Research Network (HUN-REN), Institute for Computer Science and Control (SZTAKI), Budapest, Hungary
Eötvös Loránd University (ELTE), Institute of Mathematics, Budapest, Hungary

Summary

Machine learning (ML) is one of the fundamental methodologies within artificial intelligence (AI), which is widely used in several fields, such as finance, health care, manufacturing and control. A classical problem in ML is to estimate the unknown parameters of an environment based on noisy observations. In many real-world applications involving safety and stability criterions, it is essential to quantify the uncertainty of these estimates.

Confidence regions are classical tools in uncertainty quantification, often used to assess the quality of an estimate. In this paper we investigate a resampling framework to construct confidence regions and hypothesis tests for several supervised learning problems. First, we present a resampling scheme to build confidence intervals for the expected value of an unknown symmetric distribution. We generate alternative observations based on a candidate parameter, then compare the original sample to the alternative dataset. If the candidate parameter is equal to the true expected value, then the alternative samples have the same distribution as the original sample, otherwise the alternative samples follow a different distribution. This distributional difference can be detected with a rank test. The main ideas of this procedure can be generalized to regression and binary classification problems, in which cases we build confidence regions for the true regression function.

The presented methods are endowed with strong theoretical guarantees. The user-chosen confidence levels of the confidence regions are non-asymptotically exact. The confidence sets are defined via flexible rank tests, while our distributional assumptions are very mild. We usually assume symmetry or exchangeability and therefore the presented algorithms are distribution-free, thus superior to many other methods in practice. We also investigate the asymptotic behaviors of the presented methods and establish several theoretical results regarding the consistency of the confidence regions.

Keywords: statistical learning theory, uncertainty quantification, confidence regions

Előszó

Az alább ismertetésre kerülő kutatási program a mesterséges intelligencia (AI) egy kitüntetett diszciplináris kutatási területe, a gépi tanulás (machine learning) problémaköréhez kapcsolódik. A számtalan gyakorlati motiváción túl a terület egy matematikailag elvont eszköztára a statisztikus tanuláselmélet, ezen belül a két változó közötti kapcsolat (feltételes várható érték vagy regresszió) becslése. Ennek speciális esete a bináris klasszifikáció, amelynek széles körű alkalmazása van minőség-ellenőrzésben és orvosi diagnosztikában.

A gyakorlati motivációkból adódóan a javasolt eljárások a nem-paraméteres statisztika körébe kell hogy tartozzanak, az adatok valószínűségeloszlására csak az elmélet által megkívánt minimális feltételekkel él a szerző, úgymint szimmetria vagy felcserélhetőség. A kutatási program általános célkitűzése, hogy a kifejlesztendő statisztikai eljárások véges mintára vonatkozó egzakt konfidenciahalmazokat határozzanak meg előre adott konfidencia mellett.

Ezt a programot a nem-paraméteres statisztika mély technikai eszközeinek, mindenekelőtt a reprodukáló magú Hilbert-terek elméletének, valamint a témavezető által korábban kezdeményezett, a rendszer-identifikációban áttörést jelentő Sign-Perturbed Sums (SPS) módszernek a magabiztos ötvözésével valósítja meg. A javasolt algoritmusokat az ösztöndíjas Python környezetben implementálta és validálta, az algoritmusok életképességét kísérleti matematikai eszközökkel bizonyította.

A HUN-REN SZTAKI küldetésnyilatkozatában és filozófiájában centrális szerepe van a célzott alapkutatások folytatásának profilunkba tartozó területeken, kapcsolódó új témák indításának lehetőségét nyitva hagyva. Ezek a kutatások végső motivációjukat egy szűken vett, magas szinten művelt tudományos diszciplináris területen kívül keresik, de eszközeikben az adott tudományos diszciplínára támaszkodnak. Az alábbiakban ismertetendő munka a Mesterséges Intelligencia Nemzeti Laboratóriumhoz kötődő számos szál révén tökéletesen illeszkedik ebbe a keretbe.

Gerencsér László
az MTA doktora
HUN-REN SZTAKI
vállalati szakértő

Az empirikus megfigyelések alapján való (i) modellezés, (ii) következtetés és (iii) döntéshozás a matematikai statisztika és a gépi tanulás alapproblémái. Az ezekre a problémákra adható megoldások képezik a mesterséges-intelligencia-módszerek gerincét, és alapvető fontosságúak a természet-, társadalom- és műszaki tudományok számára is.

A gépi tanulás gyakorlatában sokszor heurisztikus, feketedoboz módszereket alkalmaznak, amelyek hatékonyságát csak numerikus kísérletek támasztják alá, és amelyek eredményeit nehéz interpretálni. Noha számos feladatra az ilyen módszerek is megfelelőek, kritikus alkalmazásoknál – az egészségügytől a közlekedésen át az iparig – olyan módszerekre van szükség, amelyeket biztonságosan lehet alkalmazni, működésüket jól értjük, azokra matematikai garanciák adhatók, és amelyek eredményei jól interpretálhatók. Ezért is rendkívül fontos a gépi tanulás matematikai megalapozása, amely mára már gazdag múltra tekint vissza.

Tamás Ambrus kutatási témája a gépi tanulás matematikai elméletének egyik kulcsterülete, a statisztikus tanuláselmélet, beleértve a kernel módszereket. A kutatás fő célja olyan eljárások kidolgozása, amelyekre eloszlásfüggetlen, véges mintákra is alkalmazható garanciák adhatók. Ehhez az egyik alapvető módszer az újra-mintavételezés, melynek segítségével rugalmas és gyenge statisztikai feltevések mellett is erős működési garanciákkal rendelkező gépi tanulási és statisztikai módszerek konstruálhatók.

Csáji Balázs Csanád
tudományos főmunkatárs
HUN-REN SZTAKI
doktori témavezető

Bevezetés

A mesterséges intelligencia (MI) korunk egyik legjelentősebb vívmánya, amely számos innovatív szolgáltatás, a gazdaság és a tudományos fejlődés egyik lehetséges motorjául szolgál *(Misuraca–Van Noordt 2020)*. Az MI fogalmait többféle szemszögből lehet megközelíteni *(Russell–Norvig 2020)*, de az vitán felül áll, hogy a gépi tanulás a terület kulcsfontosságú alkotóeleme *(Bishop 2006; Haykin 1998)*. A gépi tanulás matematikai vizsgálata már a XX. század elején megkezdődött a statisztikai módszerek elterjedésével, mindazonáltal az első klasszikus tanuló algoritmus, a perceptron elméleti modelljét 1943-ban alkották meg *(McCulloch–Pitts 1943)*. Egy bő évtizeddel később, 1957-ben magát a működő gépet is elkészítették *(Rosenblatt 1958)*. Azóta a technológiai forradalomnak köszönhetően a gépi tanulási módszerek jelentősége és az alkalmazási területek száma megsokszorozódott, ezért többek között a biztonság és a stabilitás érdekében szükségessé vált, hogy az elterjedt módszerek bizonytalanságát meghatározzuk. Ebben a tanulmányban statisztikus tanuláselméleti módszerek bizonytalanságát vizsgáló korábbi munkák eredményeit *(Csáji–Campi–Weyer 2014; Csáji–Kis 2019; Csáji–Tamás 2019; Tamás–Csáji 2020)* összegezzük.

Sztochasztikus garanciák

Matematikai statisztika

A matematikai statisztika feladatainál megfigyelések, azaz egy eseménytéren értelmezett valószínűségi változók, segítségével kell meghatározni a változók bizonytalan viselkedésének, eloszlásának bizonyos paramétereit.

Ebben a cikkben a példák többségénél független azonos eloszlású mintával dolgozunk, és a becsülendő paraméterek a megfigyelések közös ismeretlen eloszlásához tartozó mérőszámok. Ilyenek lehetnek például a várható érték és a szórás vagy az eloszlás egyéb alakparaméterei (momentumok, ferdeség stb.). Mindazonáltal számos esetben az ismeretlen eloszlást leíró függvények vagy modellek játsszák a „paraméterek" szerepét (pl. központi feladat az eloszlásfüggvény, a regressziós függvény vagy egy optimális osztályozó becslése). A cikk további részében a paraméter fogalmát ebben a kibővített értelemben használjuk. Nem várjuk el, hogy a paraméterek egy véges dimenziós térben helyezkedjenek el, a függvényeket is paraméternek tekintjük.

Az ismeretlen paraméter csak közvetett módon kapcsolódik a megfigyelésekhez, a megfigyelések eloszlását befolyásolja, ezért általában nem remélhetjük, hogy a megfigyelések segítségével hiba nélkül meghatározható legyen. A megfigyeléseket egy függvény segítségével alakítjuk át a paramétert közelítő becsléssé. Az ilyen függvényeket hívjuk statisztikáknak. A statisztikák megalkotása során az a célunk, hogy minden lehetséges eloszlás esetén nagy valószínűséggel „jól" közelítsük a becsülni kívánt ismeretlen paramétert. Mivel egy statisztika véletlen megfigyeléseknek a (mérhető) függvénye, ezért egy becslés maga is, az értékétől függően, véletlen változó vagy véletlen elem. A becslés bizonytalanságát érdemes vizsgálni, valamilyen módon számszerűsíteni. Az egyik legelterjedtebb módszer a bizonytalanság meghatározására, hogy halmazbecsléseket, ún. konfidenciahalmazokat konstruálunk pontbecslések helyett.

Konfidenciahalmazok

Általánosságban annak a valószínűsége, hogy egy becslés pontosan megegyezik a becsülni kívánt paraméterrel, nulla, mindazonáltal, ha egy bővebb halmazt konstruálunk meg becslésnek, akkor lehet arra esélyünk, hogy nagy valószínűséggel lefedjük a becsülni kívánt paramétert. Egészen pontosan p megbízhatósági szintű konfidenciahalmaznak nevezzük a paramétertér (modelltér) egy olyan véletlen részhalmazát, amely legalább p valószínűséggel tartalmazza az ismeretlen paramétert. Értelemszerűen a p megbízhatósági szintet egy 0 és 1 közötti számnak választjuk meg, a gyakorlati alkalmazáshoz mérten (pl. 0,95, 0,99, 0,999). Természetesen a paramétertér véletlen részhalmazát a megfigyelések függvényében konstruáljuk meg. Fontos megjegyezni, hogy ebben a megközelítésben a konfidenciahalmaz a véletlen objektum, hiszen ez függ a véletlen megfigyelésektől, és a lefedni kívánt paraméter van előre rögzítve. Mindazonáltal szokás úgy fogalmazni, hogy a paraméter egy p valószínűséggel belekerül a konfidenciahalmazba. Értelemszerűen egy adott megbízhatósági szinthez, amennyire csak lehet, kis méretű konfidenciahalmazt szeretnénk találni, illetve egy másik szemszögből megközelítve minél kisebb egy p megbízhatósági szintű konfidenciahalmaz, annál biztosabbak lehetünk a becslésünkben.

Konfidenciahalmazok megalkotására két klasszikusnak mondható módszer van. Amennyiben lehetséges, kézenfekvő a minta eloszlásának segítségével meghatározni a becslés eloszlását. Erre a megközelítésre az egyik legismertebb példa a normális eloszlású minta esetén használt, a mintaátlag és a tapasztalati szórás segítségével meghatározott konfidenciaintervallum a várható értékre *(Student 1908)*. Ha nincs lehetőség arra, hogy a becslés pontos eloszlását meghatározzuk, akkor bizonyos feltételek mellett a becslés eloszlását határeloszlással tudjuk közelíteni. Nevezetes aszimptotikus példa a centrális határeloszlás tételén alapuló konfidencia-ellipszoidok konstrukciója lineáris regressziós modellek paramétervektoraira *(Wald 1943)*. Mindazonáltal számos alkalmazási területen ezt a két megközelítést nem lehet követni. Ha ismeretlen eloszlásokkal dolgozunk, akkor tipikusan nincs esély arra, hogy a becslés eloszlását meghatározzuk. Ezekben az esetekben eloszlásfüggetlen módszerekre van szükség. Ha kevés mintánk van, akkor az aszimptotikus eredmények félrevezetők lehetnek, ezért ilyenkor véges mintás garanciák használatára kell szorítkoznunk, amelyek nem-aszimptotikus korlátokkal szolgálnak. A Kooperatív Doktori Programban végzett kutatásom ezt a két szemléletmódot ötvözi: eloszlásfüggetlen és nem-aszimptotikus konfidenciahalmazok konstrukciójára vezetünk be algoritmusokat az újra-mintavételezés segítségével.

Hipotézisvizsgálat

A hipotézisvizsgálat során egy állításról, az ún. nullhipotézisről szeretnénk eldönteni, hogy lehet-e igaz. Egy statisztikai próba a minta alapján elfogadja vagy elutasítja a nullhipotézis állítását. Mivel a próba elvégzése során a döntést véletlen megfigyelések alapján hozzuk, ezért a döntésben benne rejlik a hibázás lehetősége. Kétféleképpen hozhatunk helytelen döntést egy próba elvégzésekor. Előfordulhat, hogy igaz a nullhipotézis, de a próba mégis elutasítja azt. Ezt a fajta hibát a próba elsőfajú hibájának nevezzük és ezt szeretnénk elsősorban elkerülni. Másodsorban előfordulhat, hogy a nullhipotézis hamis, azonban a próba mégis elfogadja azt. Ezt a fajta hibát hívjuk másodfajú hibának. A statisztikában általában a nullhipotézis az ismeretlen minta eloszlásához kapcsolódik, ezért felírhatók a hibák valószínűségei a minta eloszlásának függvényében. Az elsőfajú hiba maximális valószínűségét a próba szignifikanciaszintjének nevezzük, ennek 1-től vett (abszolút) különbségét pedig megbízhatósági szintnek hívjuk. Ha a nullhipotézis nem teljesül, akkor az elutasítás valószínűségét az erőfüggvénnyel tudjuk jellemezni. Világos, hogy minél inkább elfogadja egy próba a nullhipotézist, annál kisebb valószínűsége lesz az elsőfajú hibának, és minél inkább elutasítja egy próba a nullhipotézist, annál jobban lecsökken a másodfajú hiba valószínűsége. Általában adott konfidencia-szinthez keressük a lehető legerősebb próbát. A statisztikus feladata a konfidenciaszint megválasztása a gyakorlati alkalmazáshoz mérten, például a 0,95, 0,99 gyakran használatos értékek.

A döntések információtartalma nem szimmetrikus. Fontos a hibák között fennálló hierarchia. Mivel a próba elsősorban a szignifikanciaszint segítségével van megkonstruálva, ezért csak az elutasításnak van (statisztikai) cáfoló ereje. Emiatt a hipotézisvizsgálat során általában a bizonyítani kívánt állítás tagadását szokás nullhipotézisnek választani. Az elfogadásnak nincs igazából súlya, nem bizonyítja a nullhipotézis fennállását. Mindazonáltal a gyakorlatban használt próbák többsége úgy van létrehozva, hogy konzisztens legyen. Egy konzisztens próba esetén a másodfajú hiba valószínűsége is 0-hoz konvergál minden lehetséges eloszlásra, amelyre nem teljesül a nullhipotézis. A cikkben tárgyalt próbák egzakt, nem-aszimptotikus és eloszlásfüggetlen garanciát szolgáltatnak az elsőfajú hiba valószínűségére, és gyenge statisztikai feltételek mellett konzisztensek, azaz a másodfajú hiba-valószínűség konvergenciáját is elméleti korlátokkal támasztjuk alá.

A hipotézisvizsgálat számos esetben ekvivalens a konfidenciahalmazokkal. Legyen a nullhipotézis egy ismeretlen paraméterre (vagy modellre) vonatkozó egyenlőség. Az ilyen típusú próbák meghatároznak egy konfidenciahalmazt a paraméterre, illetve megfordítva egy adott megbízhatósági szintű konfidenciahalmaz definiál egy azonos megbízhatósági szintű statisztikai próbát. Az ekvivalencia abból adódik, hogy egy próbához tekinthetjük azoknak a paramétereknek a halmazát, amelyek esetében elfogadásra kerül az egyenlőséggel felírt nullhipotézis. Megfordítva, a konfidenciahalmaz segítségével könnyen definiálhatunk egy statisztikai próbát a következő módon: ha a tesztelendő paraméter belekerül a konfidenciahalmazba, akkor elfogadjuk a nullhipotézist, és ellenkező esetben pedig elutasítjuk azt. Meggondolható, hogy ez a próba örökli a konfidenciahalmaz megbízhatósági szintjét.

Újra-mintavételezés

Az újra-mintavételező eljárások olyan statisztikai módszerek, amelyek során szükség van arra, hogy az adott mintán felül egyéb segédváltozókat is legeneráljunk. A legismertebb újra-mintavételező algoritmusok a „jackknife" *(Quenouille 1956),* a „bootstrap" *(Efron–Tibshirani 1994),* a Monte-Carlo-tesztek *(Zhu 2005)* és a permutációs tesztek *(Pitman 1937).* Ezek a módszerek tipikusan abban különböznek, hogy a generálás milyen módon történik. A bootstrap a tapasztalati eloszlásból mintavételez, a jackknife az eredeti mintánál eggyel kisebb elemű mintákat tekinti. A permutációs tesztek esetében két csoportja van a változóknak, és az újragenerálás során ezt a két csoportot keverjük össze véletlenszerűen. A Monte-Carlo-tesztek esetében egy nullhipotézis által meghatározott eloszlásból mintavételezünk.

Ebben a tanulmányban egy speciális újra-mintavételező módszert, az ún. Sign-Perturbed Sums (SPS) módszert (tükörfordításban előjel-perturbált összegek módszerét) mutatjuk be várható érték becslésre és regressziós függvény becslésre. Ezután tekintjük az SPS-módszer újra-mintavételező általánosítását osztályozási feladatokra. Az SPS-módszer során véletlen változóknak az előjelét generáljuk újra egy szimmetriára utaló nullhipotézis alapján, ezért az SPS tekinthető egy speciális Monte-Carlo-tesztnek. A Monte-Carlo-tesztek és az SPS előnye a bootstraphez és a jackknife-hoz képest, hogy véges mintás garanciákat szolgáltatnak. A bootstrap és a jackknife esetében tipikusan aszimptotikus korlátok ismertek.

Az újra-mintavételezéssel rokon módszernek tekinthetők a konformális predikciós tartománybecslések *(Vovk–Gammerman–Shafer 2005).* Ezek a megközelítések is eloszlásfüggetlen és nem-aszimptotikus korlátokat biztosítanak, és számos esetben ötvözhetők a korábban említett módszerekkel, azonban a konformális predikció alapproblémája különbözik az általunk vizsgált feladattól. Ebben a tanulmányban egy előre meghatározott paraméterre konstruálunk konfidenciahalmazt, ezzel szemben egy predikciós tartomány egy véletlen változó értékére nyújt halmazbecslést.

Az SPS-t először rendszeridentifikációban előforduló lineáris regressziós feladatokra vezették be *(Csáji–Campi–Weyer 2014),* de azóta a módszer aszimptotikus tulajdonságait *(Weyer–Campi–Csáji 2017)* és modellszelekciós alkalmazásait *(Carè et al. 2021)* is vizsgálták. Az SPS klasszifikációra történő általánosítása és elemzése a Kooperatív Doktori Programban megvalósuló kutatásaim egyik kulcstémája *(Csáji–Tamás 2019; Tamás–Csáji 2020, 2022).* A módszert kernel módszerekre *(Csáji–Kis 2019)* és dinamikus rendszerekre is általánosították *(Csáji–Weyer 2015),* és speciális szinkronizált lineáris rendszerek függetlenségének tesztelésére alkalmazták *(Tamás–Bálint–Csáji 2023).*

Az újra-mintavételezés ötlete, hogy a nullhipotézis segítségével olyan alternatív mintákat generálunk, amelyek pontosan akkor cserélhetők fel az eredeti mintával, amikor a nullhipotézis fennáll. Az újra-mintavételező próbák esetében egy tesztelt paramétert akkor fogadunk el, ha a segítségével generált minták hasonló viselkedést mutatnak, mint az eredeti megfigyelések. Ellenkező esetben elutasítjuk a paraméter által meghatározott nullhipotézist és kizárjuk a paramétert a konfidenciahalmazból.

Az újra-mintavételezés két fő lépésből áll: alternatív minták generálásából és a minták összehasonlításából. Az alternatív minták generálásához szükség van egy invariancia feltevésre a minta eloszlásával kapcsolatban. Az invariancia tulajdonság vonatkozhat az eloszlásban rejlő szimmetriára, a megfigyelésekre rakódó zaj sorrendjének keverhetőségére, vagy a feladatban rejlő speciális ismeretre. Az alkalmazások áttekintésénél részletezzük ezeket. Az összehasonlítást egy rangsoroló függvény segítségével végezzük. Ennek definiálása jelenti a legnagyobb kihívást a konstrukció során. A vizsgált feladatokra külön-külön mutatjuk be a lehetséges megközelítéseket.

Az alternatív minták számát mi határozhatjuk meg. A minták együttes száma befolyásolja a konfidenciahalmaz megbízhatósági szintjét. Olyan módszert vezetünk be, amely tetszőleges (racionális) szinthez képes halmazbecslést építeni. Legyen a kívánt (racionális) megbízhatósági szint értéke p. A p számot írjuk fel két egész szám, q és m hányadosaként. Az m paraméter határozza meg a minták együttes számát, a q paraméter pedig az elfogadható rangok számát. Az újra-mintavételezés fő gondolata, hogy a valódi paraméter esetén az eredeti minta és az alternatív minták együttesen felcserélhetők, ezért amikor összehasonlítjuk az adathalmazokat egymással és sorba rendezzük őket valamilyen (szigorú) rendezés segítségével, akkor minden lehetséges sorrendnek ugyanannyi lesz a valószínűsége. Következésképpen a valódi paraméter tesztelésekor az eredeti minta rangja egyenletes eloszlású lesz az $\{1, \ldots, m\}$ halmazon *(Csáji–Campi–Weyer 2014)*. Az $\{1, \ldots, m\}$ halmazból tetszőleges q rangot előre elfogadhatónak rögzítve egy egzakt p megbízhatósági szinttel rendelkező próbát kapunk. Ebben a cikkben általában rögzítünk egy q értéket, és az ennél nem nagyobb rangokat választjuk elfogadhatónak.

Az újra-mintavételezés előnye, hogy gyenge statisztikai feltevésekkel él. A minta pontos eloszlásának ismerete nem szükséges a próbák elvégzéséhez, illetve a konfidenciahalmazok megkonstruálásához, egyedül egy invariancia tulajdonságot követelünk meg az eloszlással kapcsolatban. Az invariancia általában a gyakorlatból származó szimmetriára vagy felcserélhetőségre utal. A módszer nem él erős megkötésekkel a momentumokat illetően, és nem tesz fel regularitási feltételeket az eloszlásról. Az sem elvárás, hogy az eloszlások egy véges dimenziós paraméter segítségével legyenek meghatározva, vagy sűrűségfüggvénnyel rendelkezzenek, ezért a bemutatott konstrukciók nemparaméteresnek tekinthetők.

Alkalmazások

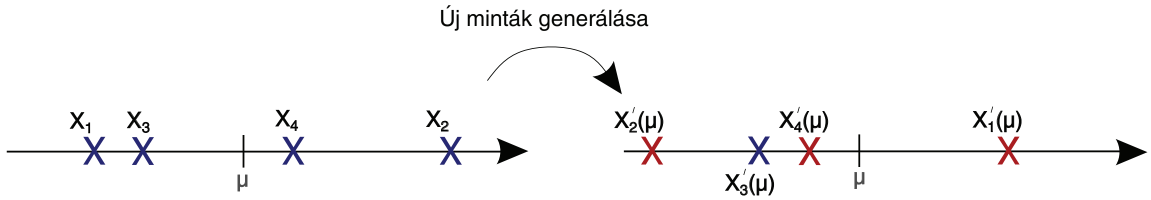
Az újra-mintavételező módszert számos feladatra lehet alkalmazni. Ebben a cikkben a regressziós és osztályozási feladatok segítségével mutatjuk be az újra-mintavételező módszereket és elméleti garanciáikat. Speciális regressziós feladatként először a várható értékre vezetjük be a konstrukciót, majd a feltételes várható érték függvényre, az ún. regressziós függvényre általánosítjuk az eljárást a regresszió és az osztályozás esetében.

Várható érték

Legyen adott egy véges sok elemből álló független minta. Tegyük fel, hogy minden megfigyelésnek ugyanaz a várható értéke és szimmetrikus az eloszlása a várható érték körül. A feladatunk, hogy konstruáljunk konfidenciahalmaz-becslést a várható értékre.

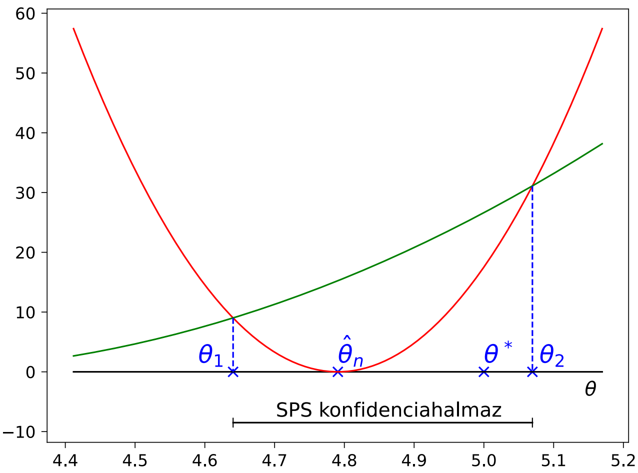
Az SPS-módszer fő feltevése az, hogy szimmetrikus eloszlással dolgozunk. Ezt használjuk ki az alternatív minták generálása során. Egyenként teszteljük a lehetséges várható értékeket. Az SPS alapötlete, hogy a várható érték jelölt segítségével *szimmetrizálhatunk*. Ha a jelölt megegyezik a szimmetriatengellyel, akkor minden megfigyelés felcserélhető a jelöltre vett szimmetrizáltjával, azaz az igazi szimmetriatengely esetében a szimmetriatengelytől vett távolság független attól, hogy a szimmetriatengely melyik oldalára esik az adott megfigyelés. Az alternatív mintákat úgy alkotjuk meg, hogy a tesztelt paramétertől vett távolságokat meghagyjuk, de a paraméterhez viszonyított helyzetet (a paramétertől vett különbség előjelét) véletlenszerűen újrageneráljuk minden megfigyeléshez, azaz a tesztelt paraméterhez egy véletlen előjellel adjuk hozzá az eredeti minta és a tesztelt paraméter különbségét. Az *1. ábra* mutatja be az újragenerálás menetét. Az ábrához kapcsolódó példában a generált előjelek három alternatív mintaelem értékét változtatták meg az eredeti mintához képest. Ezeket a mintapontokat piros színnel emeltük ki. A bemutatott tesztelési eljárás azért lesz hatékony, mert ha a valódi szimmetriatengely segítségével szimmetrizálunk, akkor az új minta eloszlása megegyezik az eredeti minta eloszlásával, azonban ha egy valódi tengelytől eltérő értéket tesztelünk, akkor az alternatív minták eloszlása megváltozik az eredeti minta eloszlásához képest. Ezt a különbséget tudjuk a próbával kimutatni. Fontos észrevétel, hogy minden paraméterhez használhatjuk ugyanazokat a véletlen előjeleket, ezért elegendő véges sok generálást végeznünk.

Az összehasonlításhoz egy rangsort fogunk felállítani a minták között. Tekintjük a minták átlagainak négyzetes hibáját a tesztelt paraméterhez képest. Az eredeti minta átlaga nem függ a tesztelt paramétertől, ezért az eredeti



1. ábra Új adatok generálása tetszőleges μ tesztparaméter segítségével

Forrás: saját szerkesztés



2. ábra | 50%-os konfidenciaintervallum a várható értékre

Forrás: Szentpéteri–Csáji 2023

mintához tartozó hiba egy olyan parabola alakú függvénygrafikont határoz meg, amelynek a nullhelye a mintaátlaghoz tartozó pontban van. Az alternatív minták esetében is egy parabolát kapunk hibafüggvénynek a lehetséges paramétereken, de megmutatható, hogy ezek a parabolák nagy valószínűséggel laposabbak, mint az eredeti mintához tartozó parabola *(Csáji–Campi–Weyer 2014)*. Az 50%-os konfidenciatartomány esetében például azokat a valós paramétereket fogadjuk el a rangteszttel, amelyeknél az eredeti mintához tartozó hibafüggvény kisebb, mint az alternatív mintákhoz tartozó hibafüggvény *(2. ábra)*. Általánosságban a q/m megbízhatósági szintű konfidenciahalmazok azokat a paramétereket tartalmazzák, amelyekre az eredeti mintához tartozó hibafüggvény értéke legfeljebb a q-adik legkisebb a hibafüggvények között.

A leírt módon megkonstruált konfidenciahalmazoknak számos előnyük van. Világos, hogy kizárólag szimmetriát feltételeztünk a minta eloszlásáról. Megmutatható, hogy a várható érték létezése sem szükséges ahhoz, hogy egzakt konfidenciahalmazt kapjunk a szimmetriatengelyre. A bemutatott módszer az összes szimmetrikus eloszlás esetén alkalmazható (pl. normális eloszlás, Laplace-eloszlás, egyenletes eloszlás, Cauchy-eloszlás stb.). Bebizonyítható, hogy a konstrukció egzakt, azaz a megbízhatósági szintet pontosan meg lehet választani. A módszer konzisztenciája megfelelő momentumfeltevések mellett teljesül, és belátható, hogy a konfidenciahalmaz aszimptotikus tulajdonságai jól közelítik a normális közelítésen alapuló konfidenciaintervallumokat *(Weyer–Campi–Csáji 2017)*, ezért közel optimálisak is. Belátható továbbá, hogy a módszer egy intervallumot ad a várható értékre, melynek két végpontját hatékonyan ki lehet számolni. Szub-Gauss-eloszlás esetén az intervallum méretének konvergenciarátája is bekorlátozható *(Szentpéteri–Csáji 2023)*.

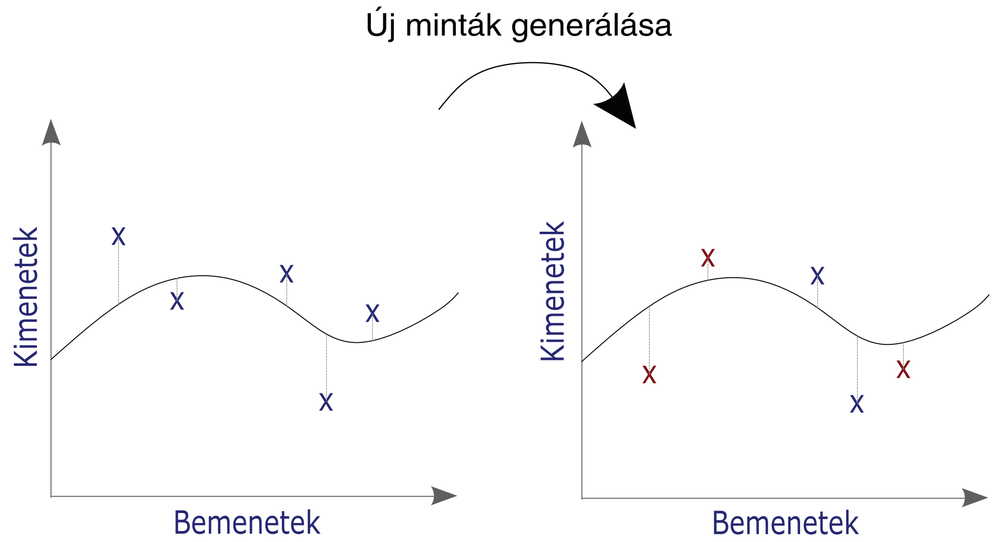
Regresszió

A regresszió során adott egy bemenet-kimenet párokból álló n elemű minta, ahol a kimenetek egy ismeretlen valós értékű függvény véletlen bemenetekben vett értékeinek zajos kiértékelései. Tegyük fel, hogy a zaj várható értéke nulla. Ilyenkor az ismeretlen függvény megegyezik a kimenetnek a bemenetre vett feltételes várható érték függvényével, az ún. regressziós függvénnyel. Ez a függvény számos gyakorlati feladatban játszik kulcsszerepet. Például a bankok a hitelekből származó várható profitot, az egészségügyben a páciensek várható élettartamát, a növénytermesztők a várható termés mennyiségét, a kereskedők a különböző termékek árát modellezhetik ilyen típusú feladattal. Ha nincsen magyarázó változó a feladatban, akkor speciális esetként visszakapjuk a korábban vizsgált várható érték becslés problémát.

A regressziós függvényre adható konfidenciahalmaz-becslést ismét egy újra-mintavételező próba segítségével konstruáljuk meg. Feltesszük, hogy adott a lehetséges regressziós függvényeknek egy családja. Például a lineáris regresszió során ez a modellosztály egy véges dimenziós lineáris tér, de a kernel módszerek esetében a modellosztály akár egy végtelen dimenziós reprodukáló magú Hilbert-tér is lehet *(Schölkopf–Smola 2002)*. Ezeket a modelleket teszteljük egy újra-mintavételező próba segítségével.

Feltesszük továbbá, hogy a zaj nulla várható értékű, és adott a transzformációknak egy csoportja, amelyre nézve a zaj eloszlása invariáns. Az egyszerűség kedvéért e helyett az absztrakt feltevés helyett a szimmetrikus példát elemezzük. Ha a zaj eloszlása szimmetrikus, akkor a szimmetriatengelyre való tükrözés nem befolyásolja a zaj eloszlását, ezért a tükrözések csoportja adja a transzformációk csoportját. Az érdekesség kedvéért megemlítünk egy további példát. Ha a zaj komponensei felcserélhetők (pl. független azonos eloszlásúak), akkor a zaj permutációkkal történő keverése helyben hagyja a zaj eloszlását, ezért a permutációcsoportra teljesül az invariancia tulajdonság. Ez azért is fontos példa, mert ilyenkor a zaj komponenseinek eloszlásával kapcsolatban nem szükséges feltevésekkel élnünk. Mindazonáltal ebben a cikkben továbbra is a szimmetrikus esetre fókuszálunk.

A várható értékhez kapcsolódó teszthez hasonlóan a visszafejtett zajok, az ún. reziduálisok előjeleit generáljuk újra az eredeti mintához képest. A *3. ábrán* látható módon szimmetrizálás (előjel-perturbáció) segítségével kreáljuk meg az alternatív mintákat. Az ábrán szereplő példában a generált előjelek három pirossal jelzett kimeneti értéket változtattak meg az alternatív mintában az eredeti mintához képest. Amikor az igazi modell segítségével történik a generálás, akkor az eredeti mintával felcserélhető alternatív mintákat kapunk, azonban amikor egy a regressziós függvénytől eltérő modellt tesztelünk, akkor az alternatív minták eloszlása megváltozik az eredeti minta eloszlásához képest. Ezt a különbséget mutatjuk ki az újra-mintavételező próbával. Az összehasonlí-



3. ábra Új adatok generálása regresszióhoz tetszőleges tesztfüggvény esetén

Forrás: saját szerkesztés

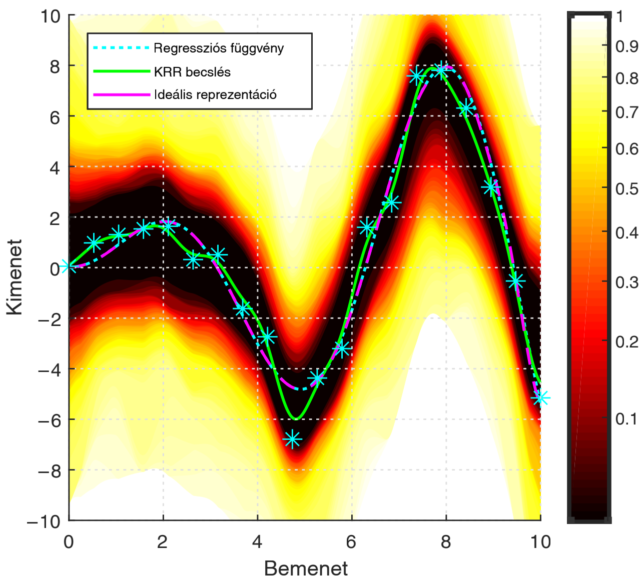
tást ismét egy rangsoroló függvény segítségével végezzük, ami sorba rendezi a rendelkezésre álló mintákat minden egyes regressziós függvény jelölt esetében. A rangsoroláshoz a lineáris esetben a négyzetes hibafüggvény gradiensének normáját érdemes használni. A részletes algoritmus megtalálható *Csáji, Campi és Weyer 2014*-es cikkében. Könnyen ellenőrizhető, hogy speciális esetként a várható érték becslésénél használt négyzetes eltéréssel ekvivalens módszerhez jutunk.

A regressziós függvényre adott konfidenciahalmaz nem-aszimptotikusan egzakt minden szimmetrikus zaj esetén, a zaj eloszlásának pontos alakjától függetlenül. Akár az is megengedett, hogy a különböző bemeneti pontokhoz különböző szimmetrikus eloszlások tartozzanak. Azt sem követeljük meg, hogy a zajnak végesek legyenek valamely momentumai, például akár végtelen varianciája is lehet. A lineáris regresszió esetében belátható, hogy a kapott konfidenciahalmaz csillagkonvex *(Csáji–Campi–Weyer 2014),* és nagy valószínűséggel korlátos *(Carè 2022).* A lineáris esetben a módszer egyenletes konzisztenciájára is ismertek gyenge elégséges feltételek *(Weyer–Campi–Csáji 2017),* azaz minden pozitív epszilon számra 1 valószínűséggel legfeljebb véges sokszor nem kerül bele a konfidenciahalmaz a regressziós függvényhez tartozó paraméter körüli epszilon sugarú gömbbe. Ez egyben azt is jelenti, hogy minden regressziós függvénytől eltérő modell esetében az elutasítás valószínűsége 1-hez tart. A konzisztencia fennállásához gyenge statisztikai feltevésekre van szükség a zaj és a magyarázó változók momentumait illetően.

Számítási szempontból hatékony külső közelítés is konstruálható az SPS konfidenciahalmazhoz ebben a lineáris esetben. Egy konvex optimalizálási feladat megoldása vezet a külső közelítéshez. Az optimalizálási feladat egy ellipszoidot eredményez, amelynek a középpontja megegyezik a normális approximációból származó konfidenciaellipszoid középpontjával *(Weyer–Campi–Csáji 2017).* Mindazonáltal míg az aszimptotikus közelítés csak heurisztikus megoldást szolgáltat, addig az SPS külső közelítése tetszőleges mintaszámra garantálja a megbízhatósági szint alsó korlátját.

A gradiens perturbálást nemlineáris modellek egy népszerű családjára, a kernel módszerekre is lehet általánosítani *(Csáji–Kis 2019).* A *4. ábrán* látható az eredmény egy összetettebb nemlineáris feladatra, ahol a kernelizált regularizált legkisebb négyzetes eltérés becslés köré építjük a konfidenciahalmazt az újra-mintavételezés segítségével. A *4. ábrán* feltüntettük a regressziós függvényt és a regressziós függvény reprodukáló magú Hilbert-térbeli ideális reprezentációját, ami a bemeneti pontokban interpolálja a regressziós függvényt. E két modell között az ábrán szereplő példában elhanyagolható különbség van. Az SPS kernelizált változata az ideális reprezentációhoz épít egzakt konfidenciahalmazt a reprodukáló magú Hilbert-tér egy véges dimenziós alterében. A *4. ábrán* színskála jelzi a tesztelt függvényekhez tartozó relatív rangokat, melyek az egymásba ágyazott konfidenciahalmazok szignifikanciaszintjeit is meghatározzák.

Az SPS-becslést számos további nemlineáris problémára lehet általánosítani. Az újra-mintavételezést tetszőleges függvényjelöltre el lehet végezni, azonban a hibafüggvény gradiensét tetszőleges tesztfüggvényre nem egyszerű általánosítani. Mindazonáltal a várható értéknél megismert módszert egy pontbecslés segítségével definiálhatjuk tetszőleges regressziós feladatra. Tekinthetjük például a pontbecslés empirikus eltérését a tesztelt függvénytől minden minta esetében. Ezeknek a hibáknak a méretével sejtésünk szerint általában olyan rangsoroló függvényt konstruálhatunk, amely gyenge statisztikai feltételek mellett is konzisztens halmazbecslésekhez vezet.



4. ábra Regressziós függvény konfidenciahalmaz-becslése nemlineáris kernelizált regularizált regresszió (KRR) segítségével színkódolt szignifikanciaszintek szerint

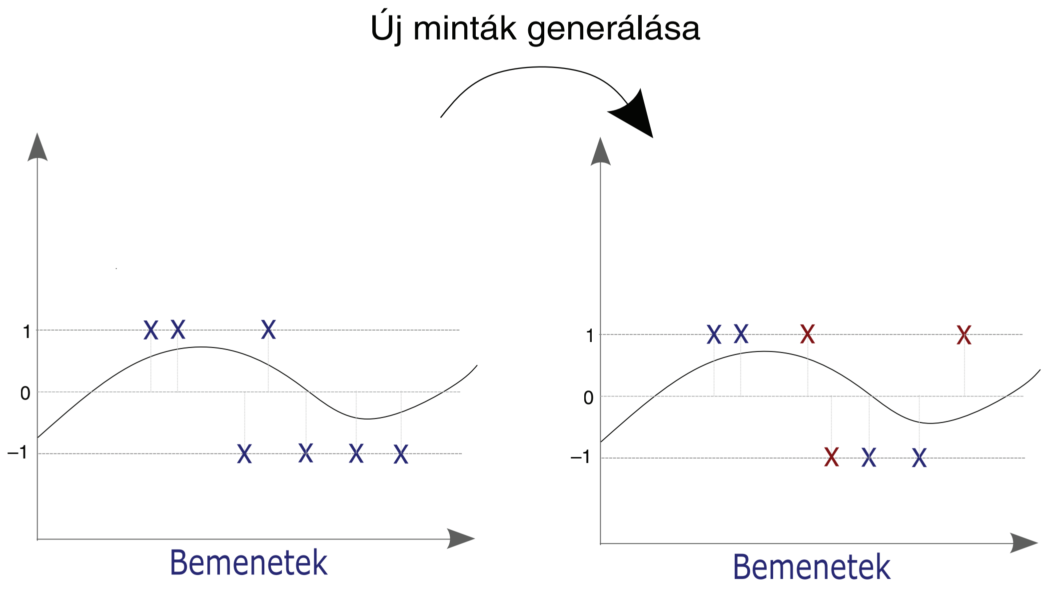
Forrás: Csáji–Kis 2019

Bináris osztályozás

A regresszióhoz hasonlóan itt is adott egy *n* elemű minta, azonban a kimenetek csupán kétféle értéket vehetnek fel. Ezt a két értéket a +1 és a –1 értékekkel azonosítjuk. Számos hétköznapi probléma fogalmazható meg az osztályozás nyelvén. Például az egészségügyben a páciensekről kell eldönteni, hogy fertőzöttek-e bizonyos magyarázó változók alapján, a pénzintézeteknek a lehetséges ügyfeleikről kell eldönteni, hogy hitelképesek-e a megadott adataik alapján, a gyártásban pedig a készülő alkatrészek közül kell kiszűrni a selejtes termékeket minél hatékonyabban.

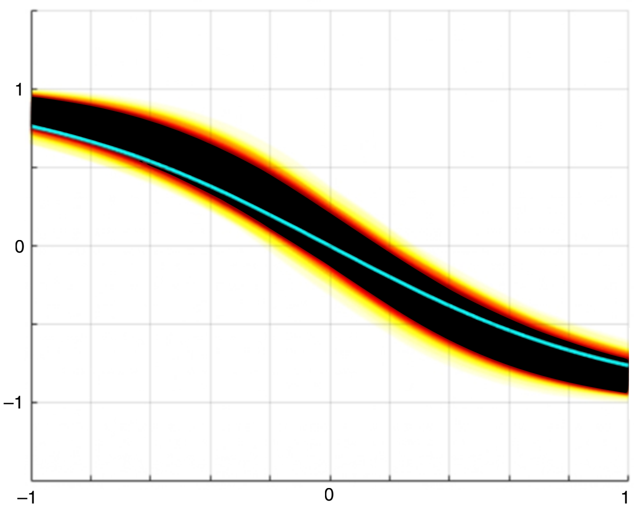
Továbbra is a feltételes várható érték függvényre, azaz a regressziós függvényre adunk konfidenciahalmaz-becslést. A regressziós függvény nemcsak az optimális osztályozót határozza meg, de a félreosztályozás valószínűségét is minden lehetséges bemenetre, ezért is van központi szerepe az osztályozási feladatban. Világos, hogy a szimmetrikus feltevés ebben az esetben csak akkor teljesül, ha a feltételes várható érték konstans nulla. Ilyenkor a kimeneti változók függetlenek is a bemenetektől, ezért ez a példa nem túl érdekes. Ugyanakkor a regressziós függvény megadja a kimenetek feltételes eloszlását a bemenetekre nézve. A bináris értékkészlet miatt elégséges, hogy minden magyarázó vektorhoz ki tudjuk számolni a +1 és a –1 érték feltételes valószínűségét a regressziós függvényből. Következésképpen, az osztályozás esetében a tesztfüggvény által meghatározott feltételes eloszlásból generáljuk az alternatív kimeneteket az eredeti minta bemeneteihez *(5. ábra)*. Kizárólag olyan (mérhető) függvényeket tesztelünk, amelyek értékkészlete –1 és +1 közé esik, mert ezek definiálnak értelmes feltételes eloszlásokat. Amikor a valódi regressziós függvényt teszteljük, akkor az alternatív minta felcserélhető az eredeti mintával, ha pedig egy regressziós függvénytől eltérő jelöltet tesztelünk, akkor az alternatív minták eloszlása különbözik az eredeti minta eloszlásától, és ezt a különbséget mutatjuk ki a rangsoroló függvény segítségével.

A rangsoroló függvény definiálása jelenti most is a legnagyobb kihívást. Hagyatkozhatunk a regressziós függvény egy pontbecslésére *(Tamás–Csáji 2022)* és a pontbecslés empirikus hibájára, vagy egy összetettebb



5. ábra Új adatok generálása bináris osztályozásra egy megengedett tesztfüggvény esetén

Forrás: saját szerkesztés



6. ábra A regressziós függvény konfidenciahalmaz-becslése bináris osztályozás esetén színkódolt konfidenciahalmazokkal (ld. *4. ábra*)

Forrás: Tamás–Csáji 2020

módszerrel a mintákhoz tartozó (feltételes) kernel átlag beágyazások segítségével is definiálhatunk rangsoroló függvényeket *(Csáji–Tamás 2019; Tamás–Csáji 2022)*. A tesztelt függvényekhez tartozó rangokat egy színskálával azonosítva a különböző rangokhoz tartozó konfidenciahalmazokat egyszerre is ábrázolhatjuk. Ez látható a *6. ábrán* egy logisztikus függvénycsalád esetén.

A fenti módszerrel tetszőleges regressziós függvényt tesztelhetünk. A konstrukció egzakt véges mintás garanciát szolgáltat az első fajú hiba valószínűségére. Belátható, hogy ha egy pontbecslésre építjük a rangsoroló függvényt, akkor a konfidenciahalmaz örökli a pontbecslés konzisztenciáját *(Tamás–Csáji 2022)*. A konzisztencia ebben az esetben is azt jelenti, hogy minden a regressziós függvénytől eltérő függvény jelölt 1 valószínűséggel legfeljebb véges sokszor lesz elfogadva az újra-mintavételező próba által. Speciális esetekben a konfidenciahalmazok egyenletes konvergenciájára is adható korlát a függvényosztály VC dimenziójának segítségével, amennyiben a paraméterezés inverz-Lipschitz folytonos *(Tamás–Csáji 2023)*.

Konklúzió

A gépi tanulási módszerek bizonytalanságának meghatározása napról napra fontosabb feladattá válik, és a mesterséges intelligencia stabil, biztonságos és szabályozott alkalmazásaihoz elengedhetetlen. A bizonytalanság kvantifikálásában kulcsszerepet játszanak a konfidenciahalmazok és a statisztikai próbák. Ebben a cikkben újra-mintavételező eljárásokat mutattunk be konfidenciahalmazok konstruálására. Ez a flexibilis keretrendszer számos statisztikai problémához nyújt használható segédeszközt. A technika legnagyobb előnye a paraméteres módszerek többségével szemben, hogy nagyon gyenge, a gyakorlatban jól használható statisztikai feltevések mellett biztosít elméleti korlátokat. A cikkben a regressziós függvényre építünk konfidenciahalmaz-becslést egy újra-mintavételező eljárás segítségével az osztályozás és a regresszió esetében. Egzakt, nem-aszimptotikus és eloszlás-független garanciákkal látjuk el a konstrukciókat, és a módszer számos gyakorlati előnyét elemezzük. A véges mintás korlátok mellett vizsgáljuk a konfidenciahalmazok aszimptotikus viselkedését és elégséges feltételekkel biztosítjuk a becslések konzisztenciáját.

Köszönetnyilvánítás

Szeretnék köszönetet mondani Csáji Balázs Csanádnak, aki témavezetőként, és Gerencsér Lászlónak, aki vállalati szakértőként nyújtott segítséget a cikk megírásához és a Kooperatív Doktori Programban megvalósuló kutatásaimhoz. Az ábrák elkészítésében nyújtott segítségért köszönet illeti Kis Krisztián Balázs és Szentpéteri Szabolcs kollégáimat.

A kutatást részben az Európai Unió finanszírozta, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. A munkát részben a „Kooperatív gyártó- és logisztikai rendszerek kutatása a versenyképes és fenntartható gazdaság támogatására” című, TKP2021-NKTA-01 sz. NKFIH projekt támogatta. A kutatás az Innovációs és Technológiai Minisztérium Kooperatív Doktori Program doktori hallgatói ösztöndíj programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával valósult meg.

NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS HIVATAL

AZ NKFI ALAPBÓL MEGVALÓSULÓ PROJEKT

Irodalomjegyzék

Bishop, C. M. (2006) Pattern Recognition and Machine Learning. New York, Springer

Carè, A. (2022) A Simple Condition for the Boundedness of Sign-Perturbed-Sums (SPS) confidence regions. Automatica, Vol. 139. 110150.

Carè, A., Campi, M. C., Csáji, B. Cs., & Weyer, E. (2021) Facing Undermodelling in Sign-Perturbed Sums System Identification. Systems & Control Letters, Elsevier, Vol. 153. 104936.

Csáji, B. Cs., Campi, M. C., & Weyer, E. (2014) Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models. IEEE Transactions on Signal Processing, Vol. 63. No. 1. pp. 169–181.

Csáji, B. Cs., & Kis, K. B. (2019) Distribution-Free Uncertainty Quantification for Kernel Methods by Gradient Perturbations. Machine Learning, Springer, Special Issue of the European Conference on Machine Learning (ECML PKDD Journal Track), Vol. 108. pp. 1677–1699.

Csáji, B. Cs., & Tamás, A. (2019) Semi-Parametric Uncertainty Bounds for Binary Classification, 58th IEEE Conference on Decision and Control (CDC), Nice, France, pp. 4427–4432.

Csáji, B. Cs., & Weyer, E. (2015) Closed-Loop Applicability of the Sign-Perturbed Sums Method. 54th IEEE Conference on Decision and Control (CDC), Osaka, Japan, pp. 1441–1446.

Efron, B., & Tibshirani, R. J. (1994) An Introduction to the Bootstrap. New York, CRC Press

Haykin, S. (1998) Neural Networks: A Comprehensive Foundation. Upper Saddle River, Prentice Hall PTR

McCulloch, W. S., & Pitts, W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, Vol. 5. No. 4. pp. 115–133.

Misuraca, G., & Van Noordt, C. (2020) AI Watch: Artificial Intelligence in public services: Overview of the use and impact of AI in public services in the EU. JRC Research Reports of the European Commission, JRC118163

Pitman, E. J. G. (1937) Significance Tests Which May Be Applied to Samples from any Populations. II. The Correlation Coefficient Test. Supplement to the Journal of the Royal Statistical Society, Vol. 4. No. 2. pp. 225–232.

Quenouille, M. H. (1956) Notes on Bias in Estimation. Biometrika, Vol. 43. No. 3–4. pp. 353–360.

Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, Vol. 65. No. 6. pp. 386–408.

Russell, S. J., & Norvig, P. (2020) Artificial Intelligence: A Modern Approach. 4th ed., New Jersey, Prentice Hall Series in Artificial Intelligence

Schölkopf, B., & Smola, A. J. (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, USA, MIT Press

Student (1908) The Probable Error of a Mean. Biometrika, Vol. 6. No. 1. pp. 1–25.

Szentpéteri, Sz., & Csáji, B. Cs. (2023) Sample Complexity of the Sign-Perturbed Sums Identification Method: Scalar Case, 22nd IFAC World Congress (World Congress of the International Federation of Automatic Control). Yokohama, Japan, July 9–14. pp. 10363–10370.

Tamás, A., Bálint, D. Á., & Csáji, B. Cs. (2023) Robust Independence Tests with Finite Sample Guarantees for Synchronous Stochastic Linear Systems. IEEE Control Systems Letters (L-CSS), IEEE Press, Vol. 7. pp. 2701–2706.

Tamás, A., & Csáji, B. Cs. (2020) Sztochasztikus garanciák bináris klasszifikációhoz. Alkalmazott Matematikai Lapok, Vol. 37. No. 2. pp. 1–15.

Tamás, A., & Csáji, B. Cs. (2022) Exact Distribution-Free Hypothesis Tests for the Regression Function of Binary Classification via Conditional Kernel Mean Embeddings. IEEE Control Systems Letters (L-CSS), IEEE Press, Vol. 6. pp. 860–865.

Tamás, A., & Csáji, B. Cs. (2023) Distribution-Free Inference for the Regression Function of Binary Classification. arXiv preprint, arXiv:2308.01835

Vovk, V., Gammerman, A., & Shafer, G. (2005) Algorithmic Learning in a Random World. New York, Springer

Wald, A. (1943) Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. Transactions of the American Mathematical Society, Vol. 54. No. 3. pp. 426–482.

Weyer, E., Campi, M. C., & Csáji, B. Cs. (2017) Asymptotic Properties of SPS Confidence Regions. Automatica, Vol. 82. pp. 287–294.

Zhu, L. (2005) Nonparametric Monte Carlo Tests and Their Applications. New York, Springer Science & Business Media