Alkalmazott Matematikai Lapok **37** (2020), 365–379.
DOI: 10.37070/AML.2020.37.2.16

SZTOCHASZTIKUS GARANCIÁK BINÁRIS KLASSZIFIKÁCIÓHOZ

TAMÁS AMBRUS ÉS CSÁJI BALÁZS CSANÁD

A bináris klasszifikáció a statisztikus tanuláselmélet egyik alapvető problémája. A jelen cikk célja a kimenetek bemenetekre nézve vett feltételes várható értékének – a regressziós függvénynek – megbecslése és a becslés bizonytalanságának vizsgálata. A regressziós függvény előjele meghatározza a Bayes optimális osztályozót, valamint segítségével a félreosztályozás kockázata is kiszámolható. Bevezetünk egy újramintavételezésen alapuló keretrendszert és három kernel-alapú algoritmust, amelyek gyenge feltételek mellett képesek egzakt, nem-aszimptotikus konfidenciahalmazokat konstruálni a regressziós függvényhez, és erősen konzisztensek is.

1. Bevezetés

Az osztályozás vagy klasszifikáció a *statisztikus tanuláselmélet* [10] egyik alapvető problémája, amelyet számtalan területen (pénzügy, egészségügy, ipar, stb.) alkalmaznak. A (bináris) klasszifikáció során adott egy független azonos eloszlású (i.i.d.) minta, $\mathcal{D}_0 = \{(x_i, y_i)\}_{i=1}^n$, az (X, Y) véletlen vektor ismeretlen eloszlásából, P, ahol x_i az i-edik bemenet és $y_i \in \{+1, -1\}$ az i-edik megfigyelés címkéje.

Osztályozóknak nevezzük a $g : \mathbb{X} \to \{+1, -1\}$ alakú (mérhető) függvényeket. Általában a klasszifikáció célja, hogy minimalizálja az a priori kockázatot, az $R(g) \doteq \mathbb{E}\big[\, L(Y, g(X)\,\big]$ függvényt, ahol L egy tetszőleges (mérhető) veszteségfüggvény. Bayes optimális osztályozónak hívjuk és g_*-gal jelöljük azt a függvényt, ahol ez a minimum felvétetik. Ebben a cikkben a $0\,/\,1$ veszteségfüggvényt használjuk, azaz $L(y, g(x)) \doteq \mathbb{I}\,(g(x) \neq y)$, ahol $\mathbb{I}$ az indikátor függvény. Ebben az esetben az a priori kockázat a félreosztályozás valószínűsége, $R(g) = \mathbb{P}\,(\,g(X) \neq Y\,)$, és levezethető [4], hogy minden $x \in \mathbb{X}$ esetén $g_*(x) = \operatorname{sign}(\,\mathbb{E}\big[\,Y \,|\, X = x\,\big]\,)$. Vegyük észre, hogy a feltételes várható érték függvény $f_*(x) \doteq \mathbb{E}\big[\,Y \,|\, X = x\,\big]$, amit a továbbiakban *regressziós függvénynek* nevezünk, több információt hordoz magában, mint g_*, ui. f_*-ból maga a kockázat is kiszámolható. Ezért a jelen cikk a regressziós függvényhez adható sztochasztikus garanciákkal foglalkozik. Fő újdonsága egy *újramintavételezésen* alapuló keretrendszer bevezetése, amelynek segítségével *nem-aszimptotikusan* garantált, *egzakt* konfidenciahalmazokat építhetünk, melyek – a megfigyelések eloszlásától függetlenül – egy tetszőleges, előre meghatározott

valószínűséggel tartalmazzák a regressziós függvényt. A javasolt – Monte Carlo és bootstrap tesztekhez hasonló – keretrendszert véges-mintás rendszer identifikációs módszerek [2] motiválták.

A konfidenciahalmazokat egy adott modellosztályban konstruáljuk meg, ami lehet tetszőlegesen tág, akár végtelen dimenziós is. A javasolt keretrendszer segítségével három kernel-alapú algoritmust [3] is bevezetünk, amelyek *egzakt* konfidenciatartományokat konstruálnak, valamint *erősen konzisztensek*, azaz a hamis modellek – gyenge feltételek mellett – a mintaméret növekedésével egy valószínűséggel kikerülnek a konstruált konfidenciahalmazokból.

2. Reprodukáló magú Hilbert-terek

Legyen adott egy $f : \mathbb{X} \to \mathbb{R}$ alakú függvényekből álló Hilbert-tér, $\mathcal{H}$, a hozzátartozó $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ skalárszorzattal. Azt mondjuk, hogy $\mathcal{H}$ egy *reprodukáló magú Hilbert-tér* (RKHS), ha a kiértékelő lineáris funkcionál $\delta_x : f \to f(x)$ minden $x \in \mathbb{X}$ esetén korlátos. Ekkor a Riesz reprezentációs tétel alapján létezik $k(\cdot, \cdot)$, hogy minden $x \in \mathbb{X}$ esetén $k(\cdot, x) \in \mathcal{H}$ és $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$. Ezt hívjuk a *reprodukáló tulajdonságnak* és a $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ függvényt a *kernelnek*. Speciálisan $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = k(x, y)$, amiből következik, hogy k szimmetrikus és pozitív definit. Megfordítva, minden szimmetrikus, pozitív definit függvény egyértelműen meghatároz egy RKHS-t (ld. Moore–Aronszajn tétel [1]). A legelterjedtebb kernelek közé tartozik a Gauss kernel, $k(x, y) = \exp(-\|x-y\|^2/2\sigma^2)$ ahol $\sigma > 0$ és a polinomiális kernel, $k(x, y) = (x^{\mathrm{T}} y + c)^d$ ahol $c \geq 0$ és $d \in \mathbb{N}$.

Egy adott $\mathcal{D}_0$ mintához tartozó ún. Gram mátrix, $K \in \mathbb{R}^{n \times n}$, a kernel értékek segítségével határozható meg: $K_{i,j} \doteq k(x_i, x_j), 1 \leq i, j \leq n$. Megmutatható, hogy ez mindig egy (adatfüggő) szimmetrikus, pozitív szemidefinit mátrix.

Legyen most $\mathbb{X}$ egy metrikus tér és $\mathbb{Z} \subseteq \mathbb{X}$ kompakt. Jelölje továbbá $C(\mathbb{Z})$ a $\mathbb{Z}$-n értelmezett folytonos függvények terét a szuprémum norma által generált metrikával és $\mathcal{H}(\mathbb{Z}) \doteq \operatorname{span}\{k(\cdot, z) : z \in \mathbb{Z}\} \subseteq \mathcal{H}$, azaz a $k(\cdot, z)$, $z \in \mathbb{Z}$, függvények által kifeszített teret. Azt mondjuk, hogy egy k kernel *univerzális*, ha minden $\mathbb{Z}$ kompakt halmaz, $f \in C(\mathbb{Z})$ függvény és $\varepsilon > 0$ esetén létezik $h \in \mathcal{H}(\mathbb{Z})$, hogy $\sup_{x \in \mathbb{Z}} |f(x) - h(x)| < \varepsilon$, azaz $\mathcal{H}(\mathbb{Z})$ *sűrű* a $C(\mathbb{Z})$ térben az uniform topológiával.

Egyik fontos alkalmazása az RKHS-eknek a *kernel átlag beágyazás* [8], amely eloszlásokhoz rendel RKHS-beli elemeket, a kernel segítségével:

2.1. Definíció. Legyen $(\mathbb{X}, \mathcal{X})$ egy mérhető tér és jelölje $M_+(\mathbb{X})$ a valószínűségi mértékek halmazát ezen a téren. Ezeknek a valószínűségi mértékeknek egy k kernellel ellátott $\mathcal{H}$ RKHS-be való átlag beágyazását az alábbi módon definiáljuk:

$$\mu : M_+(\mathbb{X}) \to \mathcal{H}, \qquad \text{és} \qquad \mu(P) = \int k(x, \cdot)\, \mathrm{d}P(x), \tag{1}$$

feltéve, hogy ez a Bochner integrál létezik.

A kernelt *karakterisztikusnak* hívjuk, ha az imént definiált beágyazás, μ, injektív. Ekkor a beágyazott elem megőrzi az eloszlásban rejlő információt, például minden $P, Q \in M_+(\mathbb{X})$ esetén, $\|\mu(P) - \mu(Q)\|_{\mathcal{H}} = 0$ pontosan akkor, ha $P = Q$.

Belátható, hogy a Gauss kernel univerzális és karakterisztikus is; valamint ha $\mathbb{X}$ kompakt, akkor az univerzalitásból már következik is a karakterisztikusság [8].

A mi esetünkben a minta eloszlása ismeretlen, ezért a beágyazását is csak becsülni tudjuk a tapasztalati eloszlás segítségével. Ezt többek között azért tehetjük meg, mert a *nagy számok erős törvénye* (NSzET) általánosítható olyan véletlen elemekre is, amelyek értéküket egy szeparábilis Hilbert-térből veszik [9]:

2.1. Tétel. *Legyen $\{X_n\}$ független véletlen elemek sorozata egy $\mathcal{H}$ szeparábilis Hilbert-térből. Vezessük be a $\mathrm{Var}(X) \doteq \mathbb{E}\big[\,\|\,X - \mathbb{E}[X]\,\|_{\mathcal{H}}^2\,\big]$ jelölést. Ekkor*

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} < \infty \qquad \Longrightarrow \qquad \frac{1}{n}\sum_{k=1}^{n}(X_k - \mathbb{E}[X_k]) \to 0, \tag{2}$$

egy valószínűséggel, $n \to \infty$ esetén, a skalárszorzat által indukált metrikában.

3. Újramintavételező eljárás

Először azt a keretrendszert mutatjuk be, amelynek segítségével olyan konfidenciahalmazok konstruálhatók, amelyek a regressziós függvényt, f_*-ot, pontosan egy általunk megválasztott valószínűséggel tartalmazzák a minta méretétől függetlenül. Korábban már említettük, hogy a vizsgált regressziós függvény megegyezik a feltételes várható érték függvénnyel, és a következő alakban írható

$$f_*(x) \doteq \mathbb{E}\big[\,Y \mid X = x\,\big] = 2 \cdot \mathbb{P}(\,Y = +1 \mid X = x\,) - 1. \tag{3}$$

A továbbiakban fel fogjuk tenni, hogy

(A0) $\mathbb{X} \subseteq \mathbb{R}^d$ és az $\{(x_i, y_i)\}_{i=1}^n$ minta független, azonos eloszlású (i.i.d.);

(A1) adott (mérhető) regressziós függvényeknek egy paraméterezett $\mathcal{F}$ családja, amely tartalmazza f_*-ot, azaz $f_* \in \mathcal{F} \doteq \big\{\, f_\theta : \mathbb{X} \to [\,-1, +1\,] \mid \theta \in \Theta \,\big\}$;

(A2) a paraméterezés injektív, azaz minden $\theta_1 \neq \theta_2 \in \Theta$ esetén

$$\|\,f_{\theta_1} - f_{\theta_2}\|_P^2 \doteq \int_{\mathbb{X}} (f_{\theta_1}(x) - f_{\theta_2}(x))^2 \mathrm{d}P_X(x) \neq 0, \tag{4}$$

ahol P_X a bemenetek eloszlása (a P eloszlás egy peremeloszlása).

Az egyszerűség kedvéért úgy tekintünk Θ-ra, mint paramétertérre, de nem tesszük fel, hogy ez véges dimenziós, például maguk a függvények is lehetnek a paraméterek. Az optimális f_*-hoz tartozó paramétert θ^*-gal jelöljük, azaz $f_* = f_{\theta^*}$.

Az újramintavételezés során az i.i.d. tulajdonságból fogunk kiindulni. Az ötletünk az, hogy ha adott egy θ paraméter, akkor generálhatunk alternatív címkéket

a meglévő bemenetekhez a paraméterhez tartozó feltételes eloszlás segítségével, ami leírható a következőképpen:

$$\mathbb{P}_\theta(\, Y = +1 \mid X = x\,) = \frac{f_\theta(x)+1}{2}, \quad \mathbb{P}_\theta(\, Y = -1 \mid X = x\,) = \frac{1-f_\theta(x)}{2}. \tag{5}$$

Adott θ esetén generálunk $m-1$ új *alternatív mintát*, azaz legyen

$$\mathcal{D}_i(\theta) \doteq ((x_1, y_{i,1}(\theta)), \dots, (x_n, y_{i,n}(\theta))), \tag{6}$$

minden $i = 1, \dots, m-1$ esetén, ahol minden (i,j) párra $y_{i,j}(\theta)$ egy véletlen generált változó a $\mathbb{P}_\theta(\, Y \mid X = x_j\,)$ feltételes eloszlásból. Az egyszerűség kedvéért ezt a jelölést kiterjesztjük a $\mathcal{D}_0$ esetre, azaz $\forall\, \theta : \mathcal{D}_0(\theta) \doteq \mathcal{D}_0$ és $\forall\, j : y_{0,j}(\theta) \doteq y_j$.

Természetesen minden mintát tekinthetünk egy n dimenziós véletlen vektornak és $\mathcal{D}_1(\theta), \dots, \mathcal{D}_{m-1}(\theta)$ mindig feltételesen függetlenek adott bemenetek esetén. Az egyik legfontosabb észrevételünk, hogy ha $\theta \neq \theta^*$, akkor $\mathcal{D}_0$ eloszlása általában különbözik a többi minta eloszlásától. Ez a különbség egy statisztikai próbával kimutatható. Mindazonáltal $\mathcal{D}_0$ és $\mathcal{D}_i(\theta^*)$ eloszlása megegyezik minden i esetén, így a minták statisztikailag nem különböztethetőek meg ebben az esetben. Ezek alapján a módszerünk a következő lesz: ha a generált minták jelentősen eltérnek az eredetitől, akkor kizárjuk a vizsgált paramétert, míg ellenkező esetben elfogadjuk a paraméter által állított hipotézist. A minták összehasonlítását sokféleképpen végezhetjük. Erre a célra bevezetjük a *rangsoroló függvény* fogalmát.

3.1. Definíció. Legyen $\mathbb{A} \subseteq \mathbb{R}^r$ és $[\,m\,] \doteq \{1, \dots, m\}$. Egy $\psi : \mathbb{A}^m \to [\,m\,]$ típusú (mérhető) függvényt rangsoroló függvénynek nevezünk, ha minden lehetséges $(a_1, \dots, a_m) \in \mathbb{A}^m$ esetén teljesíti az alábbi tulajdonságokat:

(P1) A $\{2, \dots, m\}$ halmaz minden μ permutációjára

$$\psi\big(\, a_1, a_2, \dots, a_m\,\big) \;=\; \psi\big(\, a_1, a_{\mu(2)}, \dots, a_{\mu(m)}\,\big), \tag{7}$$

azaz a függvény inivariáns az utolsó $m-1$ elem sorrendmódosítására.

(P2) Minden $i, j \in [\,m\,]$ esetén, ha $a_i \neq a_j$, akkor

$$\psi\big(\, a_i, \{a_k\}_{k\neq i}\,\big) \;\neq\; \psi\big(\, a_j, \{a_k\}_{k\neq j}\,\big), \tag{8}$$

ahol az egyszerűsített jelölést (P1) indokolja.

A ψ függvény kimenetét *rangnak* nevezzük. A következő lemma egy fontos észrevétel a *felcserélhető* véletlen vektorok rangsorolásával kapcsolatban:

3.1. Lemma. *Legyenek $A_1, \dots, A_m$ felcserélhető, m. m. páronként különböző véletlen vektorok $\mathbb{A} \subseteq \mathbb{R}^r$-ból. Ekkor $\psi(\, A_1, A_2, \dots, A_m\,)$ eloszlása diszkrét egyenletes, azaz minden $k \in [\,m\,]$ esetén, a rang k pontosan $1/m$ valószínűséggel.*

Vegyük észre, hogy ez a lemma az $\{A_i\}$ véletlen vektorok eloszlásától függetlenül teljesül. Az állítás a felcserélhetőségen múlik, ami a θ^* segítségével generált minták és az eredeti minta esetében fennáll. A páronkénti különbözőség szükséges feltétel ugyan, de általában kibővíthetjük a mintáinkat egy véletlen permutáció, π, különböző elemeivel $\mathcal{D}_i^\pi(\theta) \doteq \big(\mathcal{D}_i(\theta), \pi(i)\big)$ minden $i = 0, \dots, m-1$ esetén, hogy a páronkénti különbözőséget biztosítsuk. Ezzel a bővítéssel a lemmát általánosan is alkalmazhatjuk tetszőleges felcserélhető elemekre.

4. Nem-aszimptotikus konfidenciahalmazok

Legyen adott egy rangsoroló függvény, ψ, ami a kiterjesztett mintákon van értelmezve, azaz $\psi : (\mathbb{X} \times \mathbb{Y})^m \times [\,m\,] \to [\,m\,]$. Továbbá legyenek $p, q \in [\,m\,]$ tetszőleges segédparaméterek úgy, hogy $p \leq q$ teljesül. A ψ függvény által meghatározott *konfidenciahalmazt* definiáljuk a következő módon:

$$\Theta_{\varrho}^{\psi} \doteq \left\{ \theta \in \Theta : p \leq \psi\big(\mathcal{D}_0^{\pi}, \{\mathcal{D}_k^{\pi}(\theta)\}_{k\neq 0}\big) \leq q \right\}, \tag{9}$$

ahol $\varrho \doteq (m, p, q)$ a segédparamétereket jelöli. Látni fogjuk, hogy m, p és q általunk választható meg és ezek segítségével könnyedén beállítható a konfidenciaszint. A 3.1. Lemma segítségével belátható az alábbi általános tétel, ami egyben a cikk egyik legfontosabb eredményét képezi.

4.1. Tétel. *Az A0, A1 és A2 feltételek mellett, minden ψ rangsoroló függvény és $\varrho = (m, p, q)$ egész segédparaméterek esetén, amelyekre fennál* $1 \leq p \leq q \leq m$,

$$\mathbb{P}\big(\,\theta^* \in \Theta_{\varrho}^{\psi}\,\big) \;=\; \frac{q - p + 1}{m}. \tag{10}$$

A tétel nagyon általánosan garantálja az „igazi" regressziós függvény, f_*, *egzakt* tartalmazási valószínűségét, nem függ a minta eloszlásától – azaz *eloszlás-független* – és a rangsoroló függvény megválasztásától sem. *Nem-aszimptotikus* eredmény, tehát a konfidenciaszintet a minta mérete nem befolyásolja, sőt, azt mi állíthatjuk be p, q és m megválasztásával. Világos, hogy tetszőleges (racionális) szint elérhető. A p paramétert ebben a cikkben minden alkalommal 1-nek választjuk meg, ezért a későbbiekben áttérünk a $\varrho = (m, q)$ jelölésre.

Egy konfidenciahalmaz mindig alkalmas *hipotézisvizsgálatra* is. Ebben az esetben egy rangsoroló függvény segítségével tetszőleges regressziós függvény jelölt tesztelhető, azaz meghatározhatunk egy *statisztikai próbát*, ami elfogadja azt a nullhipotézist, hogy a regressziós függvény megegyezik a jelölttel, ha a rang értéke p és q közé esik. A tétel ilyenkor a próba szintjét határozza meg egzakt módon, amiből az *elsőfajú hiba* valószínűsége is meghatározható.

Az általánosságból adódóan ez a tétel megengedi patologikus rangsoroló függvények használatát, például olyanokét, amelyek csak a mintákhoz csatolt véletlen permutációtól függnek. Természetesen ezeket szeretnénk elkerülni, ezért vizsgáljuk a konfidenciahalmazaink egy másik tulajdonságát az ún. *erős konzisztenciát*. Intuitívan, egy erősen konzisztens módszer esetén a rossz paraméterek a mintaszám növekedésével kikerülnek a konstruált konfidenciahalmazokból.

4.1. Definíció. Jelölje az n elemű mintára konstruált konfidenciahalmazt $\Theta_{\varrho,n}^{\psi}$. Egy módszert erősen konzisztensnek nevezünk, ha $\forall\, \theta \neq \theta^*$, $\theta \in \Theta$ esetén:

$$\mathbb{P}\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \left\{ \theta \in \Theta_{\varrho,n}^{\psi} \right\} \right) \;=\; 0. \tag{11}$$

Az erős konzisztencia a konfidenciahalmazhoz kapcsolódó próba esetében a *másodfajú hibára* ad aszimptotikus garanciát, ugyanis azokat a konfidenciahalmazsorozatokat tekintjük erősen konzisztensnek, amelyek 1 valószínűséggel csak véges sok n-re fogadnak el egy „rossz" hipotézist. Ebből következik, hogy ilyenkor a „rossz" hipotézisek elfogadási valószínűsége – azaz a próba másodfajú hibájának valószínűsége – nullához tart, amit egy próba konzisztenciájának szoktak nevezni.

A továbbiakban bevezetünk három algoritmust, amelyek egzakt és erősen konzisztens konfidenciahalmazokat konstruálnak egy-egy kernel-módszer segítségével.

4.1. Algoritmus I (szomszédság alapú)

Az első algoritmus a k-legközelebbi szomszéd (kNN) módszerből indul ki. Az az ötlet, hogy adott θ esetén megbecsüljük az f_θ függvényt külön-külön minden mintából a kNN módszer segítségével. Ezeket a becsléseket aszerint fogjuk összehasonlítani, hogy melyikük becsli pontosabban az f_θ függvényt.

Az első algoritmushoz feltesszük a következőket:

(B1) $\mathbb{X}$ kompakt,

(B2) a bemenetek eloszlásának tartója az egész $\mathbb{X}$, azaz $\operatorname{supp} P_X = \mathbb{X}$,

(B3) P_X abszolút folytonos a Lebesgue-mértékre nézve.

A kNN becsléseket definiálhatjuk a következő módon

$$f^{(i)}_{\theta,n}(x) \doteq \frac{1}{k_n} \sum_{j=1}^{n} y_{i,j}(\theta)\, \mathbb{I}\big(x_j \in N(x,k_n) \big), \tag{12}$$

ahol $N(x,k_n)$ jelöli az x pont k_n legközelebbi szomszédját az $\{x_j\}_{j=1}^n$ halmazból. Az euklidészi metrikát használjuk $\mathbb{X}$-en a szomszédok meghatározásához. Mivel P_X abszolút folytonos, (12) Lebesgue-majdnem mindenütt jól-meghatározott.

Tekintsük a becsléseink L^2-hibáját, azaz minden $i = 0, \dots, m-1$ esetén legyenek a $Z^{(i)}_n(\theta)$ referenciaváltozók a következők:

$$Z^{(i)}_n(\theta) \doteq \|f_\theta - f^{(i)}_{\theta,n}\|_2^2 = \int_{\mathbb{X}} (f_\theta(x) - f^{(i)}_{\theta,n}(x))^2 \mathrm{d}x. \tag{13}$$

A rangsoroló függvényt ezek segítségével a következő alakban írjuk fel:

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(Z^{(i)}_n(\theta) \prec_\pi Z^{(0)}_n(\theta) \big), \tag{14}$$

ahol „$\prec_\pi$" egy szigorú rendezés a $Z^{(0)}_n(\theta), \dots, Z^{(m-1)}_n(\theta)$ elemeken a következőképpen definiálva: $Z^{(k)}_n(\theta) \prec_\pi Z^{(j)}_n(\theta)$ akkor és csak akkor, ha $Z^{(k)}_n(\theta) < Z^{(j)}_n(\theta)$ vagy $Z^{(k)}_n(\theta) = Z^{(j)}_n(\theta)$, illetve $\pi(k) < \pi(j)$. A korábban használatos jelölésekkel az első algoritmusban

$$\psi\big(\mathcal{D}^\pi_0, \{\mathcal{D}^\pi_k(\theta)\}_{k\neq 0} \big) = \mathcal{R}_n(\theta). \tag{15}$$

A konfidenciahalmaz az előzőek alapján a következő alakban adódik:

$$\Theta^{(1)}_{\varrho,n} \doteq \{\, \theta \in \Theta : \mathcal{R}_n(\theta) \le q \,\}, \tag{16}$$

ahol $\varrho \doteq (m, q)$, $1 \le q \le m$ általunk választott egész értékű segédparaméterek. A 4.2. Tétel foglalja össze az első algoritmus fontos tulajdonságait.

4.2. Tétel. *Tegyük fel, hogy A0, A1, A2, B1, B2 és B3 teljesül. Ekkor*

$$\mathbb{P}\big(\, \theta^* \in \Theta^{(1)}_{\varrho,n} \,\big) = q \,/\, m, \tag{17}$$

minden mintaméretre. Továbbá, ha $\{k_n\}$ *olyan, hogy* $k_n \to \infty$ *és* $k_n/n \to 0$, *ha* $n \to \infty$, *és* $q < m$, *akkor Algoritmus I erősen konzisztens* (11).

Az világos, hogy $\{f^{(i)}_{\theta,n}\}$ pontosan kiszámolható az adatokból, és szakaszonként konstans. Továbbá $\| f^{(i)}_{\theta,n} - f_\theta \|_2^2$ szintén pontosan megkapható, tehát az algoritmusunk gyakorlatban is megvalósítható. Mindazonáltal sok esetben gyorsabb, ha Monte Carlo (MC) módszerrel közelítjük az integrálok értékeit:

$$\| f^{(i)}_{\theta,n} - f_\theta \|_2^2 \approx \frac{1}{\ell_n} \sum_{k=1}^{\ell_n} \big(f^{(i)}_{\theta,n}(\bar{x}_k) - f_\theta(\bar{x}_k) \big)^2, \tag{18}$$

ahol ℓ_n a MC minta mérete és $\{\bar{x}_k\}$ i.i.d. egyenletes valószínűségi változók az $\mathbb{X}$-en. Ez az ötlet a NSzET-ből adódik miszerint a (18) egyenletben szereplő átlag tart $\| f^{(i)}_{\theta,n} - f_\theta \|_2^2$-hez (m.m.), ha $\ell_n \to \infty$. Meggondolható, hogy az egzakt konfidenciaszint megmarad, ha ezt a becslést használjuk a pontos integrálértékek helyett. A cikk végén szereplő tesztesetekben is ezt a közelítést alkalmaztuk.

Vegyük észre, hogy a kNN-módszer tekinthető egy lokálisan átlagoló kernelmódszernek, ahol minden ponthoz adaptáljuk az ablakfüggvény méretét és helyzetét. Ezért egy természetes általánosítása lenne Algoritmus I-nek, ha másik lokálisan átlagoló módszert választanánk a kNN helyett [6]. Noha a $k(\cdot,\cdot)$ függvényt ismét kernelnek hívjuk, nem követeljük meg, hogy ez a függvény pozitív definit legyen. Általában $k(x, y) = K(x - y)$, ahol K nemnegatív és az origóból kiindulva minden sugár mentén monoton csökkenő. Ekkor adott kernel, $k(\cdot,\cdot)$ – például Gauss – esetén az $\{f^{(i)}_{\theta,n}\}$ becsléseket definiálhatjuk a következőképpen:

$$f^{(i)}_{\theta,n}(x) \doteq \frac{1}{\sum_{l=1}^n k(x, x_l)} \sum_{j=1}^n y_{i,j}(\theta)\, k(x, x_j). \tag{19}$$

Ezekkel a regressziós függvény becslésekkel is konstruálhatók konfidenciahalmazok a korábbihoz hasonló módon. Algoritmus I-nek a lokálisan átlagoló kernelmódszerekkel általánosított variánsai szintén egzakt konfidenciahalmazt építenek. Sőt, mivel a kernel becslések egy jelentős része univerzálisan erősen konzisztens, az algoritmusunk általában örökli ezt a tulajdonságot.

4.2. Algoritmus II (beágyazás alapú)

A második algoritmus alapötlete, hogy beágyazzuk az eredeti minta eloszlását és az alternatív minták eloszlását egy RKHS-be egy karakterisztikus kernel segítségével. Ha a generáló eloszlások különböznek az eredetitől, akkor másik elemhez lesznek rendelve, mint az eredeti minta eloszlása. Ezt az eltérést próbáljuk a tapasztalati eloszlások segítségével statisztikusan kimutatni.

Algoritmus II-höz legyen $\mathbb{S} = \mathbb{X} \times \{+1, -1\}$ a mintatér és legyen $\mathcal{H}$ egy $\mathbb{S} \to \mathbb{R}$ típusú függvényeket tartalmazó RKHS. Feltesszük, hogy

(C1) a $\mathcal{H}$ reprodukáló magú Hilbert-tér *szeparábilis*,

(C2) a $\mathcal{H}$-hoz tartozó kernel mérhető, *korlátos* és *karakterisztikus*.

Ha $\mathbb{X} = \mathbb{R}^d$ akkor $\mathbb{S} = \mathbb{R}^d \times \{+1, -1\}$ és használhatjuk például a Gauss vagy a Laplace kernelt, ui. ezek korlátosak és karakterisztikusak is [8].

Értelmezzük az alábbi beágyazásokat

$$h_*(\cdot) \doteq \mathbb{E}\big[\, k(\cdot, S_*) \,\big] \qquad \text{és} \qquad h_\theta(\cdot) \doteq \mathbb{E}\big[\, k(\cdot, S_\theta) \,\big], \tag{20}$$

ahol S_* és S_θ véletlen elemek az $\mathbb{S}$ térből; S_* eloszlása az eredeti mintánk keresett ismeretlen eloszlása, és S_θ eloszlását a bemenetek peremeloszlása és az f_θ regressziós függvény határozzák meg (ld. [4]).

A kernel korlátos, ezért $\mathbb{E}\big[\, \sqrt{k(S_\theta, S_\theta)} \,\big] < \infty$, így $\{h_\theta\}$ létezik és $\mathcal{H}$-beli [8]. A kernel karakterisztikus, tehát $h_\theta = h_*$ pontosan akkor, ha $\theta = \theta^*$. Most legyen a beágyazott eloszlás tapasztalati változata a következő

$$h^{(i)}_{\theta,n}(\cdot) \doteq \frac{1}{n} \sum_{j=1}^{n} k(\cdot, s_{i,j}(\theta)), \tag{21}$$

minden $i = 0, \dots, m-1$ esetén, ahol $s_{i,j}(\theta) \doteq (x_j, y_{i,j}(\theta))$; emlékeztetőül $y_{0,j}(\theta) = y_j$. Más szóval minden $i \neq 0$ esetén $s_{i,j}(\theta)$ eloszlása megegyezik S_θ eloszlásával, továbbá $s_{0,j}$ eloszlása megegyezik S_* eloszlásával.

Most definiáljuk a $\{Z^{(i)}_n(\theta)\}_{i=0}^{m-1}$ változókat a következőképpen:

$$Z^{(i)}_n(\theta) \doteq \sum_{j=0}^{m-1} \| h^{(i)}_{\theta,n} - h^{(j)}_{\theta,n} \|^2_{\mathcal{H}}, \tag{22}$$

azaz számoljuk ki $h^{(i)}_{\theta,n}$ teljes kumulatív távolságát az összes többi beágyazott elemtől. Erre azért van szükség, mert általában nehéz a $h_\theta(\cdot) = \mathbb{E}\big[\, k(\cdot, S_\theta) \,\big]$ függvényt explicite megadni és az ettől vett távolságot kiszámolni. Ezek után a $\Theta^{(2)}_{\varrho,n}$ konfidenciahalmaz hasonlóan konstruálható meg, mint korábban, ld. (16).

4.3. Tétel. *Feltéve, hogy A0, A1, A2, C1 és C2 teljesül, az Algoritmus II által konstruált konfidenciahalmazokra fennáll, hogy*

$$\mathbb{P}\big(\, \theta^* \in \Theta^{(2)}_{\varrho,n} \,\big) \;=\; q \,/\, m, \tag{23}$$

minden természetes n-re és $\varrho = (q, m)$, $q \leq m$ segédparaméterpárra, valamint $q < m$ és $2 < m$ esetén a módszer erősen konzisztens.

Vegyük észre, hogy az algoritmus végrehajtható, hiszen a beágyazott elemek négyzetes távolsága a Hilbert-térben, $\| h^{(i)}_{\theta,n} - h^{(j)}_{\theta,n} \|^2_{\mathcal{H}}$, kifejezhető a reprodukáló tulajdonság és az $s_{i,1}(\theta), \dots, s_{i,n}(\theta), s_{j,1}(\theta), \dots, s_{j,n}(\theta)$ minta Gram mátrixának segítségével, azonban a $\{Z^{(i)}_n(\theta)\}$ változók kiszámolásához szükséges Gram mátrixok függnek a vizsgált θ paramétertől, így ez a módszer nagy számításigénnyel rendelkezik és jelentősége inkább elméleti.

4.3. Algoritmus III (eltérés alapú)

Algoritmus III az előző algoritmus intuícióit követi, de ebben az esetben egy egyszerűbb alakban definiáljuk a $\{Z^{(i)}_n(\theta)\}$ változókat, ami miatt a Gram mátrixot elég csak egyszer kiszámolni az algoritmus során, ennél fogva a számításigény ebben az esetben jelentősen alacsonyabb, mint korábban.

Algoritmus III-hoz feltesszük, hogy

(D1) $\mathbb{X}$ *kompakt*,

(D2) minden $f \in \mathcal{F}$ *folytonos*,

(D3) $\mathcal{H}$ egy mérhető, *korlátos* és *univerzális* kernellel ellátott *szeparábilis* RKHS, ami $\mathbb{X} \to \mathbb{R}$ alakú függvényeket tartalmaz.

Legyen $\varepsilon_{i,j}(\theta) \doteq y_{i,j}(\theta) - f_\theta(x_j)$, minden $i = 0, \dots, m-1$ és $j = 1, \dots, n$ esetén. Vegyük észre, hogy ha $i \neq 0$, akkor $\varepsilon_{i,j}(\theta)$ nulla várható értékű minden j esetén, mert $f_\theta(x_j) = \mathbb{E}_\theta\big[\, y_{i,j}(\theta) \,|\, x_j \,\big]$.

Ebben a részben legyenek definiálva a $\{Z^{(i)}_n(\theta)\}$ változók az alábbi módon:

$$Z^{(i)}_n(\theta) \doteq \left\| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_{i,j}(\theta)\, k(\cdot, x_j) \right\|^2_{\mathcal{H}}, \tag{24}$$

minden $i = 0, \dots, m-1$ esetén. Látható, hogy $Z^{(i)}_n(\theta)$ kiszámolható a K Gram mátrix, $K_{i,j} \doteq k(x_i, x_j)$, segítségével ugyanis a reprodukáló tulajdonság miatt

$$Z^{(i)}_n(\theta) = \frac{1}{n^2}\, \varepsilon_i^{\mathrm{T}}(\theta)\, K\, \varepsilon_i(\theta), \tag{25}$$

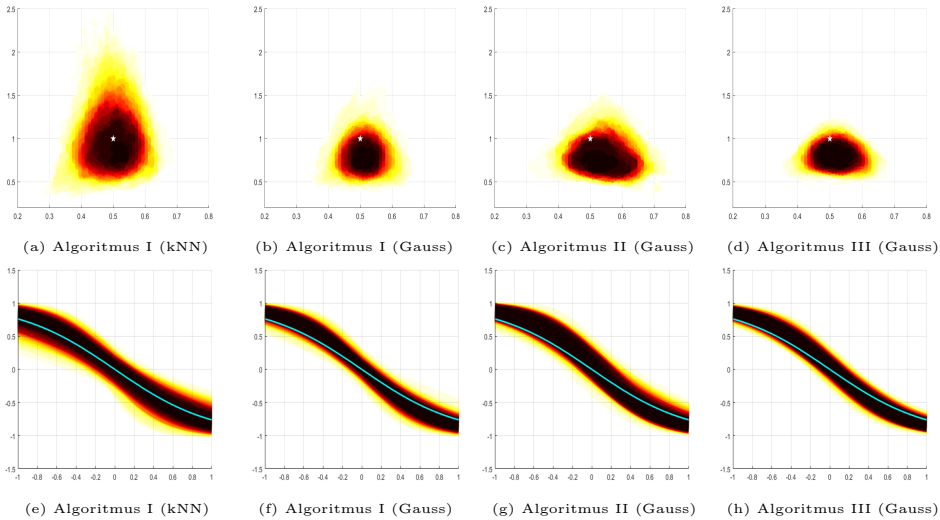
használva az $\varepsilon_i(\theta) \doteq (\varepsilon_{i,1}(\theta), \dots, \varepsilon_{i,n}(\theta))^{\mathrm{T}}$ vektor jelölést.

Innentől fogva követhetjük Algoritmus I konstrukcióját, azaz a rangsoroló függvényt úgy definiáljuk, mint (14)-ben és a konfidenciahalmaz megadható úgy, mint (16)-ben, de természetesen most az új $\{Z^{(i)}_n(\theta)\}$ változókat használjuk.

4.4. Tétel. *Feltéve, hogy A0, A1, A2, D1, D2 és D3 teljesül, az Algoritmus III által konstruált konfidenciahalmazokra fennáll, hogy*

$$\mathbb{P}\big(\, \theta^* \in \Theta^{(3)}_{\varrho,n} \,\big) = q \,/\, m, \tag{26}$$

minden természetes n-re és $\varrho = (q, m)$, $q \leq m$ segédparaméterpárra; továbbá $q < m$ esetén a módszer erősen konzisztens.



(a) Algoritmus I (kNN) (b) Algoritmus I (Gauss) (c) Algoritmus II (Gauss) (d) Algoritmus III (Gauss)

(e) Algoritmus I (kNN) (f) Algoritmus I (Gauss) (g) Algoritmus II (Gauss) (h) Algoritmus III (Gauss)

1. ábra. Egzakt, nem-aszimptotikusan garantált konfidenciahalmaz családok a bevezetett algoritmusokhoz a paramétertérben (fenti ábrák: a, b, c, d) ill. a modelltérben (lenti ábrák: e, f, g, h). A minta Laplace eloszlások keverékeként előállított szintetikus adatokat tartalmazott, a cél a keverési valószínűség (x-tengely) és a közös skálaparaméter (y-tengely) tartománybecslése volt. A színek a referencia elemek normalizált rangját – azaz az $\frac{1}{m}\mathcal{R}_n(\theta)$ értékét – mutatják. Minél sötétebb egy pont színe, annál kisebb valószínűségű konfidenciahalmazokba is belekerül. A paramétertérben szereplő fehér csillag és a modelltérben szereplő türkiz függvény az adatok generálására használt „igazi" paramétereket – $p_* = 1/2$ (x-tengely) és $\lambda_* = 1$ (y-tengely) – ill. regressziós függvényt jelöli.

5. Numerikus szimulációk

Az algoritmusok szemléltetése végett numerikus kísérleteket is végeztünk szintetikus és valós adatokon. Először, két Laplace eloszlás keverékeként előállított mintán mutatjuk be a módszerek működését, majd egy valós adatokon alapuló szívelégtelenség előrejelzési problémát vizsgálunk, melyeken a módszereinket összevetjük logisztikus regresszión alapuló aszimptotikus konfidenciahalmazokkal.

5.1. Kísérletek Laplace eloszlások keverékével

Az elsőként bemutatott kísérletek esetében a szintetikus minta együttes eloszlása két Laplace eloszlás keveréke, amelyek várható értéke, μ_1 és μ_2, eltért egymástól, de a skálaparaméterük, λ, megegyezett. A szimuláció során természetesen tetszőleges eloszlásokat tekinthettünk volna; azért választottuk a vastagabb farkú Laplace eloszlást (pl., a normális helyett), hogy szemléltessük a módszereink általánosságát. Ebben a példában p valószínűséggel a „+1" osztályt, $1-p$ valószínűséggel a „−1" osztályt figyeltük meg, azaz a regressziós függvényekből álló modellcsaládot a p, μ_1, μ_2 és λ paraméterekkel adtuk meg.

A tesztesetekben a konfidenciahalmazokkal a $p_* = 1/2$ (x-tengely) és $\lambda_* = 1$

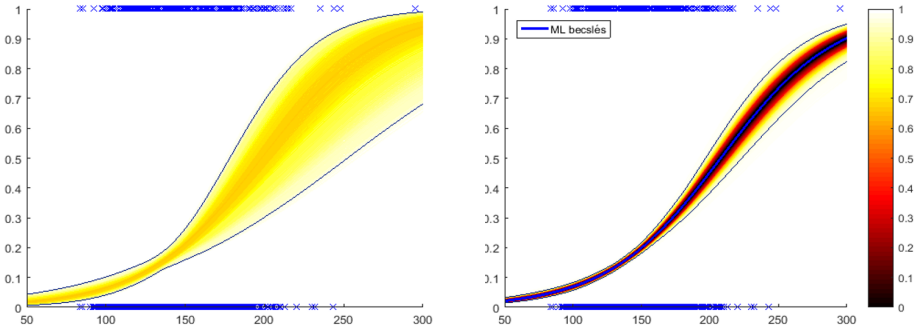
(y-tengely) paramétereket szerettük volna becsülni. Az eltolásparamétereket ismertnek tekintettük, $\mu_1 = -1$ és $\mu_2 = 1$, így két dimenziós ábrán tudtuk ábrázolni a halmazokat. Az 1. ábra mutatja a kapott relatív rangokat, $\{\mathcal{R}_n(\theta)/m\}$, a tesztelt $\theta = (p, \lambda)$ paraméterek függvényében. A rangokat az (a), (c) és (d) esetben az Algoritmus I-II-III-al, a (b) esetben pedig az Algoritmus I kernelizált változatával számoltuk. Az (e), (f), (g) és (h) ábrák a modelltérben szemléltetik a konfidenciahalmazokat. Az eredeti minta mérete $n = 500$ volt, és további 39 újramintavételezett mintát használtunk, azaz $m = 40$. A kNN módszernél 15 szomszéddal dolgoztunk. A kernel minden esetben a Gauss kernel $\sigma = 1/8$ paraméterrel. Sötétebb színekkel jelöltük a kisebb rangokat, ezért a sötétebb színű paraméterek az alacsonyabb szintű konfidenciahalmazokba is bekerülnek. A rangokat a paraméterek egy sűrű rácsán értékeltük ki. A paraméterrácsot $1/100$-os lépésközzel alakítottuk ki a $[0{,}2, 0{,}8] \times [0{,}2, 2{,}4]$-os téglán. Látható, hogy a különböző algoritmusok összemérhető (korlátos) konfidenciahalmazokat konstruálnak. A tapasztalatok szerint a konfidenciahalmazok mérete és a számításigény alapján a III. algoritmus alkalmazása a leghatékonyabb.

A bemutatott módszerek egy előnye, hogy nem szükséges, hogy a paramétereket interpretálni tudjuk azon túl, hogy valamilyen módon egy regressziós függvényt határoznak meg. Továbbá, a regressziós függvények kompatibilisek végtelen sok együttes eloszlással, ui. a bemenetek peremeloszlása nincs rájuk hatással. Emiatt nincs szükség arra, hogy az eloszlások együttesen is paraméterezve legyenek, ezért a módszereket szemi- vagy félparametrikusnak is nevezhetjük. Ha $\theta^* \in \mathbb{R}^d$ akkor a módszerek automatikusan *együttes* és továbbra is *egzakt* konfidenciahalmazokat építenek. Mindezek alapján a bemutatott algoritmusaink amellett, hogy erős elméleti garanciákkal rendelkeznek, nagyon rugalmasan alkalmazhatóak.

5.2. Szívelégtelenség előrejelzése sztochasztikus garanciákkal

Az Egészségügyi Világszervezet (WHO) felmérései szerint a szívelégtelenség tekinthető világszerte az első számú halálozási oknak. 2016-ban például a WHO becslése szerint 17,9 millióan haltak meg szívelégtelenség miatt. Az egyik leggyakoribb szívelégtelenség a koszorúér-betegség (CHD), aminek korai diagnosztizálása milliók életében csökkentheti a komplikációk kockázatát.

Második numerikus kísérletünkben egy Framinghamben (Massachusetts, USA) végzett kutatás adatain dolgoztunk, amely a Kaggle honlapon szabadon elérhető és felhasználható kutatási célokra [5]. Több, mint 4000 páciensnek 15 lehetséges kockázati faktora és az adatfelvételt követő 10 évben bekövetkező koszorúér-betegségei szerepeltek a vizsgált adathalmazban. A lehetséges kockázati tényezők között egészségügyi, demográfiai és viselkedési adatok voltak. A példa egyszerűsége kedvéért mi egyedül a szisztolés vérnyomás segítségével modelleztük a koszorúér-betegség bekövetkezési valószínűségét. A szisztolés vérnyomásra 85 és 295 Hgmm közötti értékek voltak felvéve. Viszonyítási alapként a WHO tájékoztatója szerint a 140 Hgmm feletti érték már magas vérnyomásnak tekintendő.



(a) Algoritmus III (Gauss) (b) Logisztikus regresszió

2. ábra. Kísérletek szívelégtelenség előrejelzésére. A mintaelemek – amelyeket a kék „×"-ek jelölnek – segítségével logisztikus modelleket, ld., (27), teszteltünk. Minden modell esetén a referencia elemek rangja a szín árnyalatával van jelölve, így a modellekhez tartozó elutasítási valószínűségek leolvashatók a színskála segítségével. A vékony sötétkék függvények grafikonjai egy (konzervatív) 95%-os konfidenciasáv határait mutatják. A vastagabb világoskék grafikon a logisztikus regressziós modellt ábrázolja.

A 2. ábrán az x tengelyen láthatók a szisztolés vérnyomás értékek és az y tengelyen 1-es érték jelöli, hogyha 10 éven belül koszorúér-betegséggel diagnosztizáltak valakit, illetve 0 érték jelöli az egészséges (nem diagnosztizált) eseteket. A regressziós függvényre egy logisztikus modellosztályt tekintettünk:

$$\mathcal{F} \doteq \left\{ f_{(a,b)}(x) = \frac{1}{1+\exp(-(a\cdot x+b))} \;\middle|\; a,b\in\mathbb{R} \right\}, \tag{27}$$

amin kétféle módszert alkalmaztunk. Először az eltérés alapú Algoritmus III-at használtuk, hogy konfidenciahalmazokat konstruáljunk. A logisztikus modellek megfelelő transzformáltjait teszteltük az algoritmus segítségével egy sűrű paraméterrácson. A transzformációra azért volt szükség, hogy a címkék értékeit egységesítsük: az eddig „−1"-gyel jelölt osztályt azonosítottuk a példában szereplő „0" értékű osztállyal. A tesztelt paraméterpárok a $[-6,-4]$ intervallum 1/80-os lépésközzel vett felosztásának osztópontjaiból és a $[0{,}015, 0{,}035]$ intervallum $2{,}5\times10^{-4}$-es lépésközzel vett felosztásának osztópontjaiból álltak. Viszonyításképpen ábrázoltuk a maximum likelihood (ML) módszerrel meghatározott *logisztikus regressziós* modell körül a Fisher-információ segítségével megadott *határeloszlás* alapján kapott konfidenciahalmazokat [7]. A konfidencia-ellipszoidok határain a paraméterekhez tartozó modellek esetében színárnyalattal (diszkretizálva) ábrázoltuk az elutasítási valószínűségeket. A pontos valószínűségek a színskála segítségével olvashatók le mindkét módszer esetén. Az ábrákon sötétkék színnel feltüntettük a 95%-os konfidenciahalmazba eső függvények pontonkénti maximumát és minimumát. Belátható, hogy a pontos minimum és maximum értékek egy legalább 95%-os (konzervatív) konfidenciasávot határoznak meg a regressziós függvény értékeire. Fontos megjegyeznünk, hogy míg a mi módszerünk egzakt garanciát szolgáltat az „igazi"

paraméterre nézve, addig a logisztikus regresszió esetében a korlátok egy határeloszláson alapulnak, amelyek paraméterei csak becsülve vannak. Ezek a tényezők kisebb minta esetén jelentősen befolyásolhatják a kapott konfidenciahalmazok méretét. Vegyük észre továbbá, hogy a mi módszerünk egyedül a modellek alakját használja ki és azon az intervallumon, ahol kevesebb adatunk van, nagyobb bizonytalansággal becsli a betegség kockázatát. Ez statisztikai szempontból egy sokkal reálisabb megközelítés, mint amit a „tankönyvi megoldás", az ML becslés határeloszlása szolgáltat.

6. Összefoglalás

A cikkben bemutattuk, miként konstruálhatunk *nem-aszimptotikus* konfidenciahalmazokat a *feltételes várható érték függvényhez* bináris osztályozás esetén tetszőleges megbízhatósági szintre, a minta eloszlásától függetlenül. A regressziós függvény vizsgálata kiemelten fontos a klasszifikáció szempontjából, mivel megadható vele az optimális Bayes osztályozó, és a félreklasszifikálás kockázata is. A cikkben szintetikus és valós adatokon keresztül szemléltettük a módszereinket.

Az alapötlet az volt, hogy úgy tesztelünk egy modelljelöltet, hogy a segítségével *alternatív mintákat* generálunk, és összehasonlítjuk egy adott kernel-módszer teljesítőképességét az eredeti mintán és a generált mintákon. Általában, ha egy modelljelölt „távol" van a keresett (ismeretlen) modelltől, akkor a generált minták nagy mértékben eltérnek az eredeti mintától, amit statisztikailag kimutathatunk a becsült modellek segítségével. A cikkben három konstrukciót vezettünk be. Mindegyikről megmutatható, hogy *egzakt* és *erősen konzisztens* konfidenciahalmazokat épít tetszőleges mintaméret esetén, gyenge statisztikai feltételek mellett.[1]

A konstrukció alapján egyenként minden paraméterről egyértelműen eldönthető, hogy bekerül-e egy adott valószínűségű konfidenciahalmazba, de a teljes halmaz hatékony reprezentálása (például egy ellipszoiddal való külső közelítése) kihívást jelent. Alacsony dimenziós paramétertérben a halmaz jól közelíthető diszkretizációval, azonban a közelítés számításigénye a dimenzió növekedésével hatványozottan nő, ezért a reprezentálás skálázhatósága további kutatást igényel.

7. Köszönetnyilvánítás

A publikációban szereplő kutatást, amelyet a SZTAKI valósított meg, az Innovációs és Technológiai Minisztérium (ITM) és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (NKFIH) támogatta a Mesterséges Intelligencia Nemzeti Laboratórium, a 2018-1.2.1-NKP-2018-00008 projekt és a Kooperatív Doktori Program (KDP) 1007901 számú doktori hallgatói ösztöndíja keretében.

[1] A bizonyítások elérhetők a következő linken: https://arxiv.org/abs/1903.09790.

Hivatkozások

[1] Aronszajn, N.: *Theory of Reproducing Kernels*, Transactions of the American Mathematical Society, Vol. **68** No. **3** (1950), pp. 337-404 (1950). DOI: 10.1090/S0002-9947-1950-0051437-7

[2] Carè, A., Csáji, B. Cs., Campi, M., and Weyer, E.: *Finite-Sample System Identification: An Overview and a New Correlation Method*, IEEE Control Systems Letters, Vol. **2** No. **1**, pp. 61-66 (2018). DOI: 10.1109/LCSYS.2017.2720969

[3] Csáji, B. Cs. and Tamás, A.: *Semi-Parametric Uncertainty Bounds for Binary Classification*, in: *Proceedings of the 58th IEEE Conference on Decision and Control (CDC)* IEEE, Piscataway, NJ, pp. 4427-4432 (2019). DOI: 10.1109/CDC40024.2019.9029477

[4] Devroye, L., Györfi, L., and Lugosi, G.: *A Probabilistic Theory of Pattern Recognition*, Springer, Vol. **31** (1996). DOI: 10.1007/978-1-4612-0711-5

[5] Dileep: *Logistic Regression to Predict Heart Disease*, accessed: 2020-11-01(2019). https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression/version/1

[6] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H.: *A Distribution-Free Theory of Nonparametric Regression*, Springer (2002). DOI: 10.1007/b97848

[7] Lehmann, E. L. and Romano, J. P.: *Testing Statistical Hypotheses*, Springer Science & Business Media (2006). DOI: 10.1007/0-387-27605-X

[8] Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B.: *Kernel Mean Embedding of Distributions: A Review and Beyond*, Foundations and Trends in Machine Learning, Vol. **10** No. **1-2**, pp. 1-141 (2017). DOI: 10.1561/2200000060

[9] Taylor, R. L.: *Stochastic Convergence of Weighted Sums of Random Elements in Linear Spaces*, vol. 672, Springer (1978). DOI: 10.1007/BFb0063205

[10] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).

Tamás Ambrus 1996-ban született Esztergomban. Az alapképzést az Eötvös Loránd Tudományegyetem (ELTE) matematika szakán végezte 2015 és 2018 között, majd ugyanitt 2020-ban alkalmazott matematikus MSc diplomát szerzett sztochasztika specializáción. 2020-tól kezdve az ELTE Matematika Doktori Iskolában PhD hallgató. 2018 óta a Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI) Mérnöki és Üzleti Intelligencia Laboratóriumában (EMI) dolgozik. 2019-ben kernel alapú klasszifikációs algoritmusok bizonytalanságáról írt dolgozatával a tudományos diákkonferencián 1. díjat szerzett. Jelenleg a statisztikus tanuláselmélet témakörében végez kutatásokat. Nem-aszimptotikus és eloszlás-független módszerek fejlesztésén dolgozik.

Tamás Ambrus
Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI)
1111 Budapest, Kende utca 13-17.
tamas.ambrus@sztaki.hu

Csáji Balázs Csanád 1976-ban született Budapesten. Első diplomáját (MSc) programtervező matematikusként szerezte az ELTE-TTK-n 2001-ben, majd filozófia szakos bölcsész diplomát (MA) szerzett az ELTE-BTK-n 2006-ban. Tanulmányai alatt 3-5 hónapos részképzésekben vett részt az Eindhoveni Műszaki Egyetemen (Hollandia, 2001), a British Telecomnál (Nagy Britannia, 2002), és a Johannes Kepler Egyetemen (Ausztria, 2003). PhD fokozatát az ELTE Informatikai Karán védte meg 2008-ban. Doktorálása után a Louvaini Katolikus Egyetemen (Belgium) volt posztdoktori kutató, majd 2009-től a Melbournei Egyetemen (Ausztrália) dolgozott, ahonnan 2013-ban tért haza, jelenleg a SZTAKI tudományos főmunkatársa. Eredményeit több díjjal jutalmazták, például elnyerte az Ausztrál Kutatási Tanács (ARC) "Discovery Early Career Researcher Award (DECRA)" díját, valamint az MTA Matematikai Tudományok Osztályának Gyires Béla díját is. Több mint 70 referált tudományos cikk szerzője, kutatási területe a gépi tanulásban és rendszer identifikációban fellépő sztochasztikus modellek valószínűségelméleti és statisztikai vizsgálata.

Csáji Balázs Csanád
Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI)
1111 Budapest, Kende utca 13-17.
csaji.balazs@sztaki.hu

STOCHASTIC GUARANTEES FOR BINARY CLASSIFICATION

Ambrus Tamás, Balázs Csanád Csáji

Binary classification is one of the fundamental problems of statistical learning theory. The paper aims at estimating, with strong non-asymptotic stochastic guarantees, the conditional expectation of the class labels given the inputs, i.e., the regression function. The regression function does not only determine a Bayes optimal classifier, which provides optimal predictions, but also gives access to the misclassification probability. We introduce a resampling framework to construct confidence regions for the regression function with exact coverage probabilities and present three kernel-based semi-parametric methods, all of which are strongly consistent.

Keywords: binary classification, regression function, confidence regions, distribution-free methods, non-asymptotic guarantees, strong consistency, exact confidence