Alkalmazott Matematikai Lapok **21** (2003), 1–31.

NEMPARAMÉTERES FÜGGVÉNYBECSLÉSEK

ANTOS ANDRÁS ÉS PINTÉR MÁRTA

Budapest

1. Bevezetés

A matematikai statisztika alapvető célja valószínűségi törvények vagy azok jellemzőinek megtanulása adatokból. A klasszikus matematikai statisztika jobbára paraméteres szabályokból áll, amelyek arra az esetre vonatkoznak, amikor a statisztikusnak van valamilyen előzetes ismerete az aktuális problémáról, és így feltételezni tudja, hogy az ismeretlen valószínűségi törvény az eloszlások valamely paraméteres osztályához tartozik.

A nemparaméteres statisztika esetében nem áll rendelkezésre ilyen információ, ezért a szabályok konstruálása nem használhatja a törvény egyetlen speciális jellemzőjét sem, és a szabály alapvető tulajdonságainak eloszlásfüggetlennek kell lenniük.

Általános modellünk a következő: Legyen X egy vektor valószínűségi változó, melynek eloszlása ismeretlen. Szeretnénk megbecsülni X eloszlásának egy P jellemzőjét, amelynek értéke lehet egy függvény vagy akár maga az eloszlás. A becslés jóságát egy nemnegatív $l(\cdot,\cdot)$ veszteségfüggvény méri (pl. a függvények L_1 vagy L_2-beli távolsága). Inputként adottak az $X_1,\dots,X_n$ független minták. A becslést jelölje P_n, amely a minták minden egyes realizációjához P egy lehetséges becslését rendeli. Az $l(P_n,P)$ nullához való konvergenciáját vizsgáljuk.

Egy P_n becslés *gyengén konzisztens egy eloszlásra,* ha

$$l(P_n,P)\to 0 \quad \text{sztochasztikusan.}$$

Egy P_n becslés *gyengén konzisztens a $\mathcal{D}$ eloszlásosztályon,* ha gyengén konzisztens minden $\mathcal{D}$-beli eloszlásra, és *gyengén univerzálisan konzisztens,* ha gyengén konzisztens X minden lehetséges eloszlására.

Egy P_n becslés *erősen konzisztens egy eloszlásra,* ha

$$l(P_n, P) \to 0 \quad 1 \text{ valószínűséggel},$$

és a fentihez hasonlóan definiáljuk az *erős konzisztenciát egy $\mathcal{D}$ eloszlásosztályon* és az *erős univerzális konzisztenciát*.

Gyakran habár az eloszlást nem ismerjük, mégis van valamilyen előzetes ismeretünk X eloszlásáról az $X_1, \ldots, X_n$ adatokon kívül. Paraméteres esetben feltételezzük, hogy az eloszlás egy paraméterezett osztály tagja. Nemparaméteres esetben az ismeretünk vonatkozhat a sűrűségfüggvényre, pl. bizonyos esetekben tudhatjuk, hogy az „sima”, vagy regresszióbecslésnél a regressziófüggvényről lehet előzetes információnk, pl. hogy az monoton, vagy egy halmaz indikátora. Általánosságban tehát ilyenkor feltételezhetjük, hogy az eloszlás benne van egy $\mathcal{D}$ eloszlásosztályban. Az ilyen esetben, konzisztens becslések által elérhető legjobb konvergenciasebességet ragadják meg a következő definíciók:

Egy a_n pozitív sorozatot *a $\mathcal{D}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség*nek nevezünk, ha

$$\limsup_{n\to\infty} \inf_{P_n} \sup_{X\in\mathcal{D}} \frac{\mathbf{E}l(P_n, P)}{a_n} > 0,$$

ahol az infimum az összes becslés, míg a szuprémum az összes $\mathcal{D}$-beli eloszlás fölött értendő.

Az a_n sorozat *éles minimax konvergenciasebesség $\mathcal{D}$-re nézve,* ha van olyan P_n becslés, hogy

$$\limsup_{n\to\infty} \sup_{X\in\mathcal{D}} \frac{\mathbf{E}l(P_n, P)}{a_n} < \infty$$

és minimax alsó konvergenciasebesség $\mathcal{D}$-re nézve. Tehát ha létezik olyan becslő, aminek a hibája egyenletesen minden osztálybeli eloszlás esetén legalább úgy konvergál, mint a_n, de nem létezik olyan, amely ennél gyorsabban konvergálna.

Egy a_n pozitív sorozatot *a $\mathcal{D}$ eloszlásosztályra nézve individuális alsó konvergenciasebesség*nek nevezünk, ha

$$\inf_{\{P_n\}} \sup_{X\in\mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E}l(P_n, P)}{a_n} > 0,$$

ahol az infimum az összes becsléssorozat, míg a szuprémum az összes $\mathcal{D}$-beli eloszlás fölött értendő.

Az a_n sorozat *éles individuális konvergenciasebesség $\mathcal{D}$-re nézve,* ha van olyan P_n becsléssorozat, hogy

$$\sup_{X\in\mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E}l(P_n, P)}{a_n} < \infty$$

és minden $a_n' = o(a_n)$ sorozat individuális alsó konvergenciasebesség $\mathcal{D}$-re nézve.

A paraméteres statisztikában $\mathcal{D}$ egy paraméterezett eloszláscsalád, és P a paraméter. Viszonylag enyhe és a gyakorlatban egyszerűen ellenőrizhető feltételek esetén megkonstruálható egy P_n becslés (többnyire a maximum likelihood becslés) úgy, hogy egyrészt

$$\lim_{n\to\infty} P_n = P \quad \text{1 valószínűséggel}$$

minden P-re, azaz a becslés erősen konzisztens $\mathcal{D}$-n, másrészt az $\mathbf{E}\{|P_n - P|\}$-re az $a_n = n^{-1/2}$ egy éles minimax és individuális konvergenciasebesség.

Nemparaméteres esetben a helyzet nem marad ilyen szép és egyszerű. Egyrészt – míg általában létezik erősen univerzálisan konzisztens becslés –, gazdag eloszláscsaládokra nem garantálható konvergenciasebesség, azaz bármely becsléssorozat konvergenciasebessége tetszőlegesen lassú lehet egyes eloszlásokra. Ez azt jelenti, hogy tetszőleges nullához tartó b_n sorozathoz és P_n becsléssorozathoz létezik egy eloszlás úgy, hogy

$$\mathbf{E}l(P_n, P) \geq b_n$$

végtelen sokszor. Másrészt ilyenkor már a minimax konvergenciasebesség is jóval bonyolultabb alakú lehet, továbbá a minimax és az individuális konvergenciasebesség különböző lehet. Míg a minimax konvergenciasebesség minden n-re megadja az osztályban az (n-nel esetleg változó) legrosszabb esetre vonatkozó hibát, addig az individuális konvergenciasebesség az egyes fix eloszlásokra vonatkozó tényleges konvergenciasebességek közül adja meg a leglassabbat.

Cikkünk 2. fejezetében összefoglalunk az eloszlás- és sűrűségbecsléssel kapcsolatos konzisztencia- és konvergenciasebesség-eredményeket. A 3. fejezetben a regressziófüggvény becslésével, a 4.-ben pedig az alakfelismeréssel kapcsolatos – részben saját – eredményeket mutatjuk be.

2. Eloszlás- és sűrűségfüggvénybecslés

Az első példánk egy $F(x)$ eloszlásfüggvény becslésének feladata független, azonos eloszlású $X_1, X_2, \dots X_n$ mintákból, melyek $\mathcal{R}^d$-n veszik fel az értékeiket ($d \geq 1$). Azaz az eloszlás becsülendő jellemzője F, és a veszteségfüggvény az $l(F_n, F) = \sup_{x\in\mathcal{R}^d} \left|F_n(x) - F(x)\right|$ szuprémum norma. Itt az $F_n(x)$ empirikus eloszlásfüggvény konstruálásának a módja egyrészt eloszlásfüggetlen, másrészt $F_n(x)$ egyenletes konvergenciája, a Glivenko–Cantelli-tétel teljesül minden $F(x)$-re:

$$\lim_{n\to\infty} \sup_{x\in\mathcal{R}^d} \left|F_n(x) - F(x)\right| = 0 \quad \text{1 valószínűséggel,}$$

sőt alkalmas c konstansra

$$\mathbf{E}l(F_n, F) \leq c\sqrt{\frac{d}{n}},$$

minden $F(x)$-re, azaz $n^{-1/2}$ (minimax) konvergenciasebesség. A Glivenko–Cantelli-tétel valóban eloszlásfüggetlen, és a Kolmogorov–Szmirnov-távolságban való konvergencia egyenletes konvergenciát jelent, tehát úgy tűnik, hogy nincs is értelme ezzel tovább foglalkozni. Azonban, ha valaki például egy osztályozási problémánál (ld. 4. fejezet) két ismeretlen, folytonos eloszlásfüggvény empirikus eloszlásfüggvényeit akarja használni egyfajta „likelihood" teszt létrehozására, akkor ezek a becslések használhatatlannak bizonyulnak. Az derül ki, hogy erősebb hibakritériumot kell keresnünk.

Ebből a célból nyilvánvalóan érdemes megvizsgálni veszteségfüggvényként a totális variációt: ha μ és ν két valószínűségi mérték $\mathcal{R}^d$-n, akkor μ és ν totális variációja,

$$l(\mu,\nu) = V(\mu,\nu) \stackrel{\text{def}}{=} \sup_A \bigl|\mu(A) - \nu(A)\bigr|,$$

ahol a szuprémumot az összes A Borel-halmaz felett vesszük.

Azonban, ha μ jelöli az $\{X_i\}$-k közös eloszlását és μ_n az empirikus eloszlást, akkor nematomos μ-re

$$V(\mu,\mu_n) = 1,$$

tehát az empirikus eloszlás totális variációban egy rossz eloszlásbecslés.

Esetleg azt vélhetnénk, hogy lehet találni kifinomultabb eloszlásbecslőknek egy olyan μ_n^* sorozatát, amely erősen univerzálisan konzisztens totális variációban, azaz

$$\lim_{n\to\infty} V(\mu,\mu_n^*) = 0 \quad 1 \text{ valószínűséggel.}$$

Az alábbi eredmény szerint ez sem lehetséges.

2.1. Tétel (Devroye és Györfi [25]). *Eloszlásbecslők bármely μ_n^* sorozatához létezik olyan μ eloszlás, amelyre*

$$V(\mu,\mu_n^*) > 0.5 \quad \textit{minden } n\textit{-re } 1 \textit{ valószínűséggel.}$$

Ennek a negatív eredménynek az az oka, hogy a totális variáció sokkal erősebb hibakritérium, mint a Kolmogorov–Szmirnov-távolság: nem lehet olyan eloszlásbecslőt konstruálni, amely totális variációban eloszlásfüggetlenül konzisztens. Ahhoz, hogy konzisztencia eredményeket kapjunk, vagy a halmazosztályt, amely felett a szupremumot vesszük $l(\cdot,\cdot)$ definíciójában, vagy a megengedett eloszlások osztályát le kell szűkítenünk.

Ha a halmazosztály valahol az eddig vizsgált két extrém eset (csak a „negyedek" illetve minden Borel halmaz $\mathcal{R}^d$-ben) között van, akkor a nevezetes Vapnik–Chervonenkis-elmélet azt adja, hogy ún. Vapnik–Chervonenkis-osztályokra (ld. 3.1. definíció) a helyzet ugyanaz, mint a Kolmogorov–Szmirnov-távolság esetén: erős univerzális konzisztencia és $n^{-1/2}$ (minimax) konvergenciasebesség.

Szűkítsük most a megengedett eloszlások osztályát. Ha μ és ν abszolút folytonos eloszlások egy σ-véges λ mértékre nézve és sűrűségfüggvényük f illetve g, akkor a Scheffé-tétel miatt

$$\|f-g\| := \int_{\mathcal{R}^d} \bigl|f(x) - g(x)\bigr|\lambda(dx) = 2\,V(\mu,\nu).$$

Ez az összefüggés az L_1-ben konzisztens sűrűségfüggvény-becslésen keresztül egy olyan eloszlásbecsléshez vezet, amely konzisztens totális variációban: tegyük fel, hogy f_n L_1-ben (erősen) univerzálisan konzisztens, azaz

$$\lim_{n\to\infty} \|f - f_n\| = 0 \quad \text{1 valószínűséggel.}$$

Vezessük be az f_n sűrűségbecslő segítségével definiált eloszlásbecslőt:

$$\mu_n^*(A) = \int_A f_n(x)\lambda(dx),$$

ekkor

$$\lim_{n\to\infty} V(\mu, \mu_n^*) = 0 \quad \text{1 valószínűséggel.}$$

Ha λ a Lebesgue-mérték, akkor L_1-ben konzisztens sűrűségbecslésre standard példa a hisztogram és a magfüggvényes becslés (Devroye [19] és Devroye, Györfi [24]).

Legyen $\mathcal{P}_n$ az $\mathcal{R}^d$-nek egy partíciója pozitív és véges Lebesgue-mértékű $\{A_{n,j}\}$ cellákkal. Ekkor a *hisztogram becslő* a következő:

$$f_n(x) = \frac{\mu_n\big(A_n(x)\big)}{\lambda\big(A_n(x)\big)},$$

ahol $A_n(x) = A_{n,j}$, ha $x \in A_{n,j}$.

2.2. TÉTEL. *Tegyük fel, hogy μ-nek létezik f sűrűségfüggvénye. Ha a hisztogramnál minden origó középpontú S gömbre*

$$\lim_{n\to\infty} \sup_{j\,:\,A_{n,j}\cap S\neq\emptyset} \operatorname{diam}(A_{n,j}) = 0$$

és

$$\lim_{n\to\infty} \frac{\#\{j\,:\,A_{n,j}\cap S \neq \emptyset\}}{n} = 0,$$

akkor

$$\lim_{n\to\infty} \|f - f_n\| = 0 \quad \textit{1 valószínűséggel.}$$

Ha a cellák d-dimenziós kockák h_n oldalhosszúsággal, akkor a tétel feltételei azt jelentik, hogy $h_n \to 0$ és $nh_n^d \to \infty$.

A *magfüggvényes becslő*t a nemnegatív, mérhető $K(x)$ magfüggvény és a pozitív h_n sorozat határozza meg:

$$f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right).$$

2.3. Tétel. *Tegyük fel, hogy μ-nek létezik f sűrűségfüggvénye. Ha a magfüggvényes becslésnél*

$$\int_{\mathcal{R}^d} K(x)\lambda(dx) = 1$$

és

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} nh_n^d = \infty, \tag{1}$$

akkor

$$\lim_{n\to\infty} \|f - f_n\| = 0 \quad \textit{1 valószínűséggel.}$$

A 2.2. és 2.3. tételből a várhatóértékben való konzisztencia,

$$\mathbf{E}\{\|f(X) - f_n(X)\|\} \to 0$$

is következik. E tételek szépségét az adja, hogy az L_1-ben való konzisztencia elérhető anélkül, hogy bármit feltennénk az f sűrűségfüggvényről, és így totális variációban konzisztens eloszlásbecslőt kaphatunk, ha μ abszolút folytonos a Lebesgue-mértékre nézve. A hisztogram becslők kiterjeszthetőek olyan eloszlásbecslőkké, amelyek konzisztensek totális variációban, ha a μ nematomos része abszolút folytonos egy ismert σ-véges λ mértékre nézve (Barron, Györfi és van der Meulen [9]).

A paraméteres statisztikában a konzisztencia tisztázása után az a legfontosabb kérdés, hogy adott pontossághoz mekkora mintanagyság kell, azaz mi az illető becslés konvergenciasebessége. Ezért sűrűségfüggvénybecslés esetén is természetesen kérhetne valaki konvergenciasebességre vonatkozó eredményeket, de erre megint csak negatív válasz adható, mivel sűrűségbecslők minden f_n sorozatára a várható L_1-hiba, $\mathbf{E}\{\|f - f_n\|\}$ konvergenciasebessége tetszőlegesen lassú lehet.

Sőt, egy ismert tartóhalmazú végtelen diszkrét eloszlást sem lehet totális variációban garantált konvergenciasebességgel becsülni:

2.4. Tétel (Devroye, Györfi és Lugosi [27]). *Legyen μ egy valószínűségi mérték a pozitív egészek halmazán. Pozitív számok minden 0-hoz tartó $b_n \le 1/16$ sorozatához és eloszlásbecslők minden μ_n^* sorozatához létezik μ valószínűségi mérték úgy, hogy*

$$\mathbf{E}V(\mu_n^*, \mu) \ge b_n \quad \textit{minden } n\textit{-re.}$$

Ez azt jelenti, hogy minden 0-hoz tartó sorozat individuális alsó konvergenciasebesség a pozitív egészeken való eloszlások osztályára nézve. A tételből levezethető a következő negatív eredmény sűrűségbecslőkre:

2.5. Tétel (Birgé [13], Devroye [20, 22]). *Sűrűségbecslők minden f_n sorozatához és pozitív számok minden 0-hoz tartó $b_n \le 1/32$ sorozatához létezik $f : [0,1] \mapsto [0,2]$ sűrűségfüggvény úgy, hogy*

$$\mathbf{E}\{\|f - f_n\|\} \ge b_n \quad \textit{minden } n\textit{-re.}$$

Ez azt jelenti, hogy minden 0-hoz tartó sorozat individuális alsó konvergenciasebesség a $[0,1] \mapsto [0,2]$ sűrűségfüggvények osztályára nézve. Könnyű belátni, hogy a lassú konvergencia eredmény akkor is igaz az abszolút folytonos eloszlások $\mathcal{D}$ osztályára nézve (totális variáció veszteségfüggvénnyel), ha bármilyen eloszlásbecslést megengedünk (nem csak abszolút folytonosakat), azaz minden 0-hoz tartó sorozat individuális alsó konvergenciasebesség $\mathcal{D}$-re nézve.

Ezek szerint van L_1-ben konzisztens sűrűségbecslő, de konvergenciasebességet nem lehet garantálni, hacsak nem teszünk fel néhány dolgot az ismeretlen f sűrűségfüggvényről. Ilyen feltételek megfogalmazhatóak például a metrikus entrópia segítségével:

2.1. Definíció. Egy $\mathcal{M}(d)$ metrikus térre legyen az $N(\varepsilon, \mathcal{M})$ *ε-lefedési szám* az $\mathcal{M}$ legkisebb olyan A részhalmazának a számossága, amelyre minden $x \in \mathcal{M}$-hez létezik egy $y \in A$ úgy, hogy $d(x,y) < \varepsilon$. Legyen $\log N(\varepsilon, \mathcal{M})$ az $\mathcal{M}$-nek az *ε-metrikus entrópiá*ja.

Sűrűségfüggvények egy osztálya metrikus tér a $d(f,g) = \sup_{x \in \mathcal{R}^d} |f(x) - g(x)|$ távolsággal. Erre vonatkozik az alábbi eredmény:

2.6. TÉTEL (Birgé [13]). *Ha $\mathcal{D}^\delta$ olyan sűrűségfüggvény osztály, amelynek ε-metrikus entrópiája $O(\varepsilon^{-\delta})$ amint $\varepsilon \to 0$, akkor alkalmas sűrűségfüggvény-becslővel elérhető a*

$$n^{-\frac{1}{2+\delta}}$$

(minimax) konvergenciasebesség $\mathcal{D}^\delta$-ra.

Az ε-metrikus entrópiája jól ismert pl. a következő ún. „sima" függvényosztályoknak:

2.2. Definíció. Adott $k \in \mathcal{N}_0$-ra (ahol $\mathcal{N}_0 = \{0,1,2,\ldots\}$), $0 < \beta \leq 1$-re, $p = k + \beta$-ra és $M > 0$-ra, legyen $\mathcal{F}^{(p,M)}$ azon $f : \mathcal{R}^d \to \mathcal{R}$ függvények osztálya, amelyekre minden $\alpha = (\alpha_1, \ldots, \alpha_d)$ vektorra, ahol $\alpha_i \in \mathcal{N}_0$, $\sum_{j=1}^d \alpha_j = k$

$$|D^\alpha f(x) - D^\alpha f(z)| < M \|x - z\|^\beta,$$

ahol D^α az α-hoz tartozó parciális deriváltat jelöli.

Jelölje $\mathcal{F}^{(p,M)}\big|_C$ az $\mathcal{F}^{(p,M)}$-nek a megszorítását a $C \subset \mathcal{R}^d$ halmazra koncentrálódó sűrűségfüggvényekre.

Bármely belső ponttal rendelkező konvex kompakt $C \subset \mathcal{R}^d$ esetén $\mathcal{F}^{(p,M)}\big|_C$-nek az ε-metrikus entrópiája $O(\varepsilon^{-d/p})$ nagyságrendű, és $\mathcal{F}^{(p,M)}\big|_C$-re az $n^{-\frac{p}{2p+d}}$ éles minimax konvergenciasebesség (lásd pl. [13]). Birgé a kövekező erősebb állítást is bizonyította:

2.7. TÉTEL (Birgé [13]). *Bármely belső ponttal rendelkező $C \subset \mathcal{R}^d$ esetén $\mathcal{F}^{(p,M)}\big|_C$-re és minden 0-hoz tartó b_n sorozatra*

$$b_n n^{-\frac{p}{2p+d}}$$

individuális alsó konvergenciasebesség.

Ez a tétel motiválta a 3. és 4. fejezetben található, regresszióbecslésre illetve alakfelismerésre vonatkozó hasonló eredményeket.

A várható L_1-hibával szemben az L_1-hiba véletlen része

$$\|f - f_n\| - \mathbf{E}\big\{\|f - f_n\|\big\}$$

$n^{-1/2}$ nagyságrendű, tehát az L_1-hiba véletlen részének konvergenciája sokkal gyorsabb lehet, mint a várható L_1-hibáé:

2.8. TÉTEL (Devroye [21]). *A hisztogram vagy a magfüggvényes sűrűségbecslőkre*

$$\mathbf{P}\Big\{\big|\|f - f_n\| - \mathbf{E}\big\{\|f - f_n\|\big\}\big| > \varepsilon\Big\} \le 2e^{-n\varepsilon^2/2}.$$

Bizonyos feltételek mellett Berlinet, Devroye és Györfi [12] és Beirlant és Györfi [10] bebizonyította az L_1-hiba $n^{-1/2}$ nagyságrendű aszimptotikus normalitását, és kiszámolta az aszimptotikus szórást.

Eloszlásbecsléshez más hibakritériumot is tekinthetünk. Ilyen hibakritériumokat vezetett be Csiszár [17], amelyeket f-divergenciáknak hívunk (lásd még Liese, Vajda [43] és Vajda [53]). A három legfontosabb f-divergencia a matematikai statisztikában a totális variáció, az információs divergencia és a χ^2-divergencia.

Ha μ és ν valószínűségi mértékek $\mathcal{R}^d$-n, akkor μ és ν információs divergenciája (vagy I-divergencia, relatív entrópia, Kullback–Leibler-mérőszám)

$$l(\mu, \nu) = I(\mu, \nu) \stackrel{\text{def}}{=} \sup_{\{A_j\}} \sum_j \mu(A_j) \log \frac{\mu(A_j)}{\nu(A_j)},$$

ahol a szuprémumot az összes véges Borel-mérhető $\{A_j\}$ partíció felett vesszük.

Az alábbi egyenlőtlenség, melyet Pinsker-egyenlőtlenségnek is neveznek, felső korlátot ad a totális variációra az I-divergencia függvényében (cf. [17], Kemperman [39] és Kullback [41]):

$$2\big\{V(\mu, \nu)\big\}^2 \le I(\mu, \nu). \tag{2}$$

Ha μ és ν abszolút folytonosak egy σ-véges λ mértékre nézve f illetve g sűrűségfüggvénnyel, akkor μ és ν I-divergenciája az f és g sűrűségfüggvényekkel is kiszámítható:

$$I(\mu, \nu) = \int_{\mathcal{R}^d} f(x) \log \frac{f(x)}{g(x)} \lambda(dx) \stackrel{\text{def}}{=} D(f, g).$$

Mostanában elég nagy az érdeklődés az $I(\mu,\nu)$ I-divergencia és a $D(f,g)$ divergencia iránt abban az esetben, amikor ν illetve g empirikus. Ha

$$\mu_n^* = \mu_n^*(\,\cdot\,; X_1,\dots,X_n)$$

a μ eloszlás becslése, akkor μ_n^* információs divergenciában erősen konzisztens, ha

$$\lim_{n\to\infty} I(\mu,\mu_n^*) = 0 \quad 1 \text{ valószínűséggel.}$$

Analóg módon f_n erősen konzisztens információs divergenciában, ha

$$\lim_{n\to\infty} D(f,f_n) = 0 \quad 1 \text{ valószínűséggel.}$$

A (2) Pinsker-egyenlőtlenség alapján az információs divergencia dominálja a totális variációt, tehát az 2.1. tétel miatt igaz, hogy eloszlásbecslők bármely μ_n^* sorozatához létezik μ valószínűségi mérték, amelyre a μ_n^* sorozat nem konzisztens információs divergenciában. A helyzet még rosszabb: ismert tartóhalmazú diszkrét eloszlást sem lehet információs divergenciában konzisztensen becsülni:

2.9. TÉTEL (Györfi, Páli és van der Meulen [35]). *Legyen μ valószínűségi mérték a pozitív egészeken. Eloszlásbecslők bármely μ_n^* sorozatához létezik véges $H(\mu) = -\sum_{i=1}^{\infty} \mu(\{i\}) \log \mu(\{i\})$ Shannon-entrópiájú μ valószínűségi mérték, amelyre*

$$I(\mu,\mu_n^*) = \infty \quad \textit{minden } n\textit{-re } 1 \textit{ valószínűséggel.}$$

Ebből a tételből következik egy negatív eredmény a sűrűségbecslésre is:

2.10. TÉTEL (Györfi és van der Meulen [36]). *Sűrűségbecslők minden f_n sorozatához létezik véges $H(f) = -\int f(x)\log f(x)dx$ differenciális entrópiájú és tetszőlegesen sokszor deriválható f sűrűségfüggvény úgy, hogy*

$$D(f,f_n) = \infty \quad \textit{minden } n\textit{-re } 1 \textit{ valószínűséggel.}$$

Barron [7] foglalkozott elsőként sűrűségfüggvény becslésével úgy, hogy a becslő I-divergenciában konzisztens. Eredményeit [9] általánosította. Megmutatták, hogy ha felteszünk bizonyos feltételeket arról az eloszlásosztályról, amelyből az ismeretlen eloszlás származik, nevezetesen azt, hogy létezik ismert ν valószínűségi mérték, amelyre $I(\mu,\nu) < \infty$, akkor tudunk konstruálni olyan eloszlásbecslőt, amely az osztály valamennyi eloszlására konzisztens információs divergenciában. Mint jól ismert, az $I(\mu,\nu) < \infty$ feltételből következik, hogy μ abszolút folytonos ν-re nézve, tehát érdemes olyan eloszlásbecslést készíteni, amelynek ν szerint van sűrűségfüggvénye. Vezessük be a következő jelölést

$$f(x) = \frac{d\mu}{d\nu}(x),$$

és egy $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \dots, A_{n,m_n}\}$, $0 < m_n < n$, $n \geq 2$ partícióra, legyen

$$\nu(A_{n,i}) = h_n = 1/m_n.$$

Adott a_n, $0 < a_n < 1$ sorozatra tegyük fel, hogy

$$\lim_{n\to\infty} a_n = 0.$$

Tekintsük a következő sűrűségbecslést:

$$f_n(x) = (1 - a_n)\mu_n\big(A_n(x)\big)/h_n + a_n,$$

ahol μ_n az empirikus mérték. Ebből

$$a_n = \frac{1}{nh_n + 1}$$

választással a [7] által bevezetett sűrűségbecslőt kapjuk.

2.11. Tétel (Barron, Györfi és van der Meulen [9]). *Ha* $I(\mu, \nu) < \infty$ *és*

$$\lim_{n\to\infty} h_n = 0, \qquad \lim_{n\to\infty} nh_n = \infty,$$

és

$$\limsup_{n\to\infty} \frac{1}{nh_n a_n} \le 1,$$

akkor

$$\lim_{n\to\infty} D(f, f_n) = 0 \quad 1 \text{ valószínűséggel.}$$

3. Regresszióbecslés

Legyen Y valós értékű valószínűségi változó és legyen X d-dimenziós véletlen vektor (megfigyelés). X koordinátái különböző eloszlásúak lehetnek, lehet némelyik diszkrét (például bináris), mások lehetnek abszolút folytonosak. Így nem teszünk fel semmit X eloszlásáról. A regresszióanalízis célja Y becslése, ha X adott, azaz olyan z valós függvényt keresünk, amely X értékkészletén van definiálva, és amelyre $z(X)$ „közel" van Y-hoz. Tegyük fel, hogy az analízis fő célja a négyzetes középhiba minimalizálása:

$$\min_z \mathbf{E}\big\{\big(z(X) - Y\big)^2\big\}.$$

Jól ismert, hogy a minimumot ($\mathbf{E}\{Y^2\} < \infty$-re) az

$$m(x) \stackrel{\text{def}}{=} \mathbf{E}\{Y \mid X = x\}$$

regressziófüggvény éri el, ugyanis minden z mérhető függvényre

$$\begin{aligned}\mathbf{E}\big\{\big(z(X) - Y\big)^2\big\} &= \mathbf{E}\big\{\big(m(X) - Y\big)^2\big\} + \mathbf{E}\big\{\big(m(X) - z(X)\big)^2\big\} \\ &= \mathbf{E}\big\{\big(m(X) - Y\big)^2\big\} + \int_{\mathcal{R}^d} \big|m(x) - z(x)\big|^2 \mu(dx),\end{aligned}$$

ahol μ az X eloszlását jelöli. A jobb oldal második tagját a z függvény integrált négyzetes hibájának nevezzük és $J(z)$-vel jelöljük:

$$J(z) \stackrel{\text{def}}{=} \int_{\mathcal{R}^d} \left|m(x) - z(x)\right|^2 \mu(dx).$$

A négyzetes középhiba nyilván pontosan akkor lesz közel a minimumhoz, ha $J(z)$ közel van a 0-hoz. A sűrűségbecsléssel szemben, ahol az L_1-hiba volt a legalkalmasabb hibakritérium, itt az L_2-hiba a legfontosabb. Ráadásul a sűrűségbecslésnél az L_1-teret a Lebesgue-mértékkel definiáltuk, míg a regresszióbecslésnél az L_2-teret μ-vel definiáljuk.

A regresszióbecslés feladatánál adottak a $D_n = \left\{(X_1, Y_1), \dots, (X_n, Y_n)\right\}$ független, azonos eloszlású példányai (X, Y)-nak. Ezek felhasználásával $m(x)$-nek olyan

$$m_n(x) = m_n(x, D_n) \,:\, \mathcal{R}^d \times \left(\mathcal{R}^d \times \mathcal{R}\right)^n \mapsto \mathcal{R}$$

becslését szeretnénk megkonstruálni, amelyre $J(m_n)$ kicsi, azaz az m_n regressziófüggvény becslés m-hez való $L_2(\mu)$-konvergenciáját vizsgáljuk. Tehát most az X szerepét az (X, Y) pár játsza, P-nek m felel meg, a veszteségfüggvény pedig az $l(m_n, m) \stackrel{\text{def}}{=} J(m_n)$ négyzetes $L_2(\mu)$-hiba.

Stone [51] mutatott rá először, hogy léteznek gyengén univerzálisan konzisztens becslők, ilyenek például a legközelebbi szomszéd becslők. Következő alakú becslőket vizsgált:

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x; X_1, \dots, X_n) Y_i.$$

Ezek a becslők lokális átlagoló típusúak, mivel a W_{ni} súlyok jellegzetesen nemnegatívak, összegük 1, továbbá W_{ni} „nagy", ha x és X_i „közel" van egymáshoz, egyébként W_{ni} „kicsi" (Györfi, Kohler, Krzyżak és Walk [33]).

A *k-legközelebbi szomszéd becslő* esetében, $W_{ni}(x; X_1, \dots, X_n) = 1/k$, ha X_i az x k legközelebbi szomszédjának egyike $X_1, \dots, X_n$ közül, különben W_{ni} nulla. Vegyük észre, hogy ekkor $\sum_{i=1}^{n} W_{ni} = 1$.

3.1. Tétel (Stone [51]). *Ha*

$$k_n \to \infty, \qquad k_n/n \to 0$$

akkor a k_n-legközelebbi szomszéd becslő gyengén univerzálisan konzisztens.

3.2. Tétel (Devroye, Györfi, Krzyżak és Lugosi [26]). *Ha*

$$\lim_{n\to\infty} k_n/\log n = \infty, \qquad \lim_{n\to\infty} k_n/n = 0$$

akkor a k_n-legközelebbi szomszéd becslő erősen univerzálisan konzisztens.

A *hisztogram vagy partíciós becslő* $\mathcal{R}^d$ egy $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$ partíciója esetén az

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K_n(x, X_i)}{\sum_{i=1}^n K_n(x, X_i)},$$

függvény, ahol $K_n(x,u) = \sum_{j=1}^\infty I_{[x\in A_{n,j}, u\in A_{n,j}]}$ (I_A az A esemény indikátorfüggvényét jelöli), vagyis cellánként átlagoljuk az odaeső minták Y_i-értékeit. Gyenge és erős univerzális konzisztenciára vonatkozó eredményeket bizonyított Devroye, Györfi [23] és Györfi [32]-ben.

A regressziófüggvényt az

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\dfrac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\dfrac{x - X_i}{h_n}\right)},$$

*magfüggvényes becslő*vel is lehet becsülni ahol $h_n > 0$ egy n-től függő simító tényező, K egy abszolút integrálható függvény (a magfüggvény). Konzisztenciájához hasonló feltételekre van szükség, mint a magfüggvényes sűrűségbecslő esetében.

3.3. Tétel (Devroye és Wagner [30], Spiegelman és Sacks [50]). *Ha* (1) *teljesül és K egy kompakt tartójú, korlátos, nemnegatív függvény úgy, hogy rögzített origó középű S gömbre* $\inf_{x\in S} K(x) > 0$, *akkor a magfüggvényes becslés gyengén univerzálisan konzisztens.*

A magfüggvényes becslés erősen konzisztens, ha Y egyenletesen korlátos, (1) teljesül, K Riemann-integrálható és $K \ge aI_S$, ahol $a > 0$ egy konstans és S egy origó középű, pozitív sugarú gömb (Devroye és Krzyżak [28]). A magfüggvényes regresszióbecslő erős univerzális konzisztenciáját Györfi, Kohler és Walk [34] és Walk [58] bizonyította.

Sajnos robusztus eloszlásosztály esetén a lassú konvergencia (az alakfelismerésre vonatkozó 4.1. tétel és (3) következményeképpen) itt is fennáll:

3.4. Tétel (Devroye [18], Györfi, Kohler, Krzyżak és Walk [33]). *Regresszióbecslők minden m_n sorozatához, és pozitív számok minden monoton,* 0*-hoz tartó* $1/64 \ge b_1 \ge b_2 \ge \dots$ *sorozatához létezik* (X, Y)*-nak olyan eloszlása, amelyre X egyenletes eloszlású* $[0,1]$*-en,* $Y = m(X)$ *és*

$$\mathbf{E}J(m_n) \ge b_n \quad \textit{minden } n\textit{-re.}$$

Az eddig ismertetett regresszióbecslési módszerek a lokális átlagolás elvén alapultak. Létezik egy másik, hasonlóan természetes alapelv, az empirikus hibaminimalizálás, amely szintén elvezethet univerzálisan konzisztens becslésekhez. Válasszunk egy $\mathcal{F}_n$ függvényosztályt, és a regresszióbecslés ebből az osztályból veszi az értékeit. Az $\mathcal{F}_n$ kiválasztásakor vagy az m regressziófüggvényről szerzett ismereteinket vesszük figyelembe, vagy $\mathcal{F}_n$ olyan függvényekből áll, amelyek számítógéppel bizonyos számítási bonyolultsággal realizálhatók.

Korábban már láttuk, hogy az m regressziófüggvény minimalizálja az L_2-hibát,

$$\mathbf{E}\{(m(X)-Y)^2\} = \inf_f \mathbf{E}\{(f(X)-Y)^2\},$$

tehát mondhatná valaki, hogy minimalizáljuk $\mathbf{E}\{(f(X)-Y)^2\}$-t az $\mathcal{F}_n$ osztályon. Ez azonban nyilvánvalóan lehetetlen, mert a minimalizálandó függvény függ (X,Y) ismeretlen eloszlásától.

Az *empirikus hibaminimalizálás* alapötlete, hogy becsüljük az L_2-hibát az

$$\frac{1}{n}\sum_{j=1}^{n}\left|f(X_j)-Y_j\right|^2$$

empirikus L_2-hibával, és azt a függvényt válasszuk ki $\mathcal{F}_n$-ből, amelynek az empirikus hibája minimális.

Az ily módon választott m_n függvénynek a regressziófüggvénytől vett négyzetes távolságára teljesül, hogy

$$\int_{\mathcal{R}^d}\left|m_n(x)-m(x)\right|^2\mu(dx) \le$$

$$\le 2\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_{j=1}^{n}\left|f(X_j)-Y_j\right|^2-\mathbf{E}\{(f(X)-Y)^2\}\right|+$$

$$+\inf_{f\in\mathcal{F}_n}\int_{\mathcal{R}^d}\left|f(x)-m(x)\right|^2\mu(dx).$$

A jobboldalon szereplő első tag a becslési hiba, a második az approximációs hiba. A becslési hiba azt méri, hogy a becslő L_2-hibája mennyivel tér el az osztálybeli legjobb függvény L_2-hibájától, az approximációs hiba pedig azt, hogy mennyire jól lehet a regressziófüggvényt $\mathcal{F}_n$-beli függvényekkel közelíteni L_2 értelemben. Ahhoz, hogy univerzálisan konzisztens becslőt kapjunk, azt kell megmutatni, hogy mindkét tag nullához tart (X,Y) minden olyan eloszlására, amelyre $\mathbf{E}\{Y^2\}<\infty$.

Válasszuk az $\mathcal{F}_n$ osztályt úgy, hogy minden $\mathcal{F}_n$-beli függvény abszolút értéke korlátos legyen, és a közös β_n korlát tartson végtelenhez. Az egyszerűség kedvéért tegyük fel, hogy Y korlátos, azaz $|Y|\le L$ egy valószínűséggel. A konzisztenciára vonatkozó eredmények kiterjeszthetők nemkorlátos Y-ra is ([33]). A becslési hiba konvergenciájához azt kell megmutatni, hogy

$$\lim_{n\to\infty}\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{j=1}^{n}g(Z_j)-\mathbf{E}\{g(Z)\}\right|=0$$

1 valószínűséggel, ahol

$$Z=(X,Y),\ Z_i=(X_i,Y_i),\ i=1,2,\dots$$

és

$$\mathcal{G}_n = \{g \,:\, g(x,y) = |f(x) - y|^2,\ f \in \mathcal{F}_n\}$$

nemnegatív, $2\beta_n^2$-tel felülről korlátozott függvények egy osztálya (elég nagy n-re).

Minden rögzített olyan g függvényre, melyre $\mathbf{E}\{|g(Z)|\} < \infty$ a nagy számok erős törvénye alapján

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum_{j=1}^{n} g(Z_j) - \mathbf{E}\{g(Z)\}\right| = 0$$

1 valószínűséggel. Másrészt a Hoeffding-egyenlőtlenség szerint minden $g : \mathcal{R}^d \times \mathcal{R} \to [0, 2\beta_n^2]$ függvényre

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{j=1}^{n} g(Z_j) - \mathbf{E}\{g(Z)\}\right| > \varepsilon\right\} \le 2e^{-\frac{n\varepsilon^2}{2\beta_n^4}}.$$

Ezt és az uniókorlátot felhasználva kapunk eredményt abban az esetben, ha a $\mathcal{G}_n$ osztály véges:

$$\mathbf{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{j=1}^{n} g(Z_j) - \mathbf{E}\{g(Z)\}\right| > \varepsilon\right\} \le 2|\mathcal{G}_n| e^{-\frac{n\varepsilon^2}{2\beta_n^4}}$$

([33]).

Ha nem véges az osztály, akkor is tudunk felső korlátot adni, ha $\mathcal{G}_n$ Vapnik–Chervonenkis-dimenziója véges.

3.1. Definíció. Az $\mathcal{R}^{d+1}$-beli halmazok egy $\mathcal{A}$ osztálya *szeparál* egy $G \subseteq \mathcal{R}^{d+1}$ halmazt, ha G minden részhalmaza előáll $A \cap G$ alakban valamely $A \in \mathcal{A}$-ra. Az $\mathcal{A}$ osztály $V_{\mathcal{A}}$ *Vapnik–Chervonenkis-* (VC-) *dimenziója* a legnagyobb n szám, amelyre még létezik olyan n pontból álló halmaz $\mathcal{R}^{d+1}$-ben,amelyet $\mathcal{A}$ szeparál. Ha minden n-re létezik ilyen halmaz, akkor definíció szerint $V_{\mathcal{A}} = \infty$. Ha $V_{\mathcal{A}}$ véges, akkor $\mathcal{A}$-t *Vapnik–Chervonenkis-* (VC-) *osztály*nak mondjuk.

3.5. TÉTEL (Vapnik és Chervonenkis [56]). *Bármely ν valószínűségi mértékre és $\mathcal{A}$ halmazosztályra, és minden n-re és $\varepsilon > 0$-ra*

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A) - \nu(A)| > \varepsilon\right\} \le 8n^{V_{\mathcal{A}}} e^{-n\varepsilon^2/32},$$

ahol ν_n az (X_i, Y_i) minták által meghatározott empirikus mérték.

Definiálható egy függvényosztály Vapnik–Chervonenkis-dimenziója is:

3.2. Definíció. Az $\mathcal{F} : \mathcal{R}^d \mapsto [0, \alpha]$ osztály ***szeparáló együtthatója***, $s(\mathcal{F}, n)$, az a szám, ahány különböző részhalmazát tudjuk kiválasztani n pontnak $\mathcal{F}$-beli függvényekkel. Egy E részhalmazt kiválaszt egy f függvény, ha $f(x) \geq \alpha/2$, ha $x \in E$, és $f(x) < \alpha/2$, ha $x \notin E$.

Az $\mathcal{F}$ osztály *Vapnik–Chervonenkis- (VC-) dimenziója* $V_{\mathcal{F}}$ a legnagyobb olyan n_0 szám, amelyre még létezik olyan n_0 pont, amelynek minden részhalmazát ki tudjuk választani $\mathcal{F}$-beli függvényekkel. Ha minden n-re létezik ilyen n pont, akkor $V_{\mathcal{F}} = \infty$.

Tehát, ha a $\mathcal{G}_n$ osztály VC-dimenziója véges, akkor a Vapnik–Chervonenkis-egyenlőtlenség alapján

$$\mathbf{P}\left\{ \sup_{g \in \mathcal{G}_n} \left| \frac{1}{n} \sum_{j=1}^{n} g(Z_j) - \mathbf{E}\{g(Z)\} \right| > \varepsilon \right\} \leq 8 n^{V_{\mathcal{G}_n}} e^{-\frac{n\varepsilon^2}{32\beta_n^4}}.$$

Egy osztály nagyságát a dimenzióján kívül lefedési számaival is jellemezhetünk, és ezek segítségével is eljuthatunk felső korláthoz:

Az 2.1. definícióbeli metrikus térben a a távolság lehet az empirikus L_1-távolság, azaz $d_{1,n}(g, g') = \frac{1}{n} \sum_{i=1}^{n} \left| g(z_i) - g'(z_i) \right|$. $\mathcal{G}_n$-nek az ezen távolságon alapuló ún. empirikus L_1 ε-lefedési száma legyen $N_1(\varepsilon, z_1^n, \mathcal{G}_n)$, ahol $z_1^n = (z_1, \ldots, z_n)$.

Ha $Z_1^n = (Z_1, \ldots, Z_n)$ független, azonos eloszlású valószínűségi változók egy sorozata, akkor $N_1(\varepsilon, Z_1^n, \mathcal{G}_n)$ valószínűségi változó. Ekkor igaz a következő tétel

3.6. TÉTEL (Pollard [47]). *Legyen $\mathcal{G}$ a $g : \mathcal{R}^d \mapsto [0, M]$ függvények egy osztálya. Minden n-re és $\varepsilon > 0$-ra,*

$$\mathbf{P}\left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) - \mathbf{E}\{g(Z_1)\} \right| > \varepsilon \right\} \leq 8 \mathbf{E}\{N_1(\varepsilon/8, Z_1^n, \mathcal{G})\} e^{-n\varepsilon^2/(128M^2)}.$$

Tehát véges L_1 lefedési számú osztály esetében is lehet felső korlátot adni a becslési hibára.

A konvergencisebességre visszatérve, a 3.4. tétel állítását úgy is fogalmazhatjuk, hogy minden 0-hoz tartó pozitív sorozat individuális alsó konvergenciasebesség az $X = \text{Uniform}\,[0,1]$-t és $Y = m(X)$-t teljesítő eloszlások osztályára nézve.

Most feltételezzük, hogy az (X, Y) eloszlása benne van egy szűkebb eloszlásosztályban. Az alsó és éles konvergenciasebességeket vizsgálhatjuk ilyen eloszlásosztályokra is, melyeket többek között pl. az határoz meg, hogy az m regressziófüggvény benne van egy adott függvényosztályban. Ez lehet pl. az $F^{(p,M)}$ „sima" függvényosztály (2.2. definíció):

3.3. Definíció. Legyen $\mathcal{D}^{(p,M)}$ (X, Y) azon eloszlásainak halmaza, amelyekre

(I) X egyenletes eloszlású $[0,1]^d$-en,

(II) $Y = m(X) + N$, ahol X és N függetlenek és N standard normális,

(III) $m \in \mathcal{F}^{(p,M)}$.

3.7. TÉTEL (Barron, Birgé és Massart [8], Stone [52]). *Az* $n^{-\frac{2p}{2p+d}}$ *sorozat a* $\mathcal{D}^{(p,M)}$ *eloszlásosztályra nézve minimax alsó konvergenciasebesség, és ez éles.*

A minimax alsó korlát fő jellegzetessége, hogy mivel a legrosszabb esetet jelentő eloszlás igen bonyolult módon függhet a becslési algoritmustól, általában a bizonyítás nem konstruktív, pusztán a „rossz" eloszlás létezését tudjuk bebizonyítani. A bizonyítás fő eszköze a *randomizálás:* azt, hogy az eloszlásosztálybeli legrosszabb esetben a hiba „nagy", úgy igazoljuk, hogy véletlenszerűen választunk egy eloszlást az osztályból, és az átlagos (várható) hibát becsüljük alulról. Az eloszlásosztály szimmetriái miatt ilyenkor kezelhetőbb formulát kapunk, amelyből kiküszöbölhetőek az adott becslési algoritmus tulajdonságai.

A véletlen választás, és így a „rossz" eloszlás a minimax alsó korlátoknál általában függ a mintamérettől. Az individuális alsó korlátoknál azonban ettől független eloszlást keresünk, így a mintaszámtól nem függhet sem a randomizálás, sem a konkrét véletlen választás. Ez általában bonyolultabb bizonyítást eredményez.

3.8. TÉTEL (Antos, Györfi és Kohler [5]). *Legyen* b_n *egy tetszőleges nullához tartó pozitív sorozat. Ekkor a*

$$b_n n^{-\frac{2p}{2p+d}}$$

sorozat a $\mathcal{D}^{(p,M)}$ *eloszlásosztályra nézve individuális alsó konvergenciasebesség, azaz* $n^{-\frac{2p}{2p+d}}$ *éles individuális konvergenciasebesség.*

4. Alakfelismerés

Az alakfelismerés esetében a $\{0,1\}$ lehetséges értékű Y valószínűségi változó (címke, osztály) értékét kell eldönteni adott X d-dimenziós véletlen vektor (megfigyelés) alapján. A döntés vagy osztályozási szabály egy

$$g : \mathcal{R}^d \to \{0,1\}$$

döntésfüggvény. Az osztályozási szabály minőségét az

$$L(g) \stackrel{\text{def}}{=} \mathbf{P}\big\{g(X) \neq Y\big\}.$$

hibavalószínűség méri. Létezik optimális osztályozó, a lehető legjobb (legkisebb) hibavalószínűséget a Bayes-döntés adja:

$$g^*(x) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{ha } \ \eta^*(x) < 1/2, \\ 1 & \text{különben,} \end{cases}$$

ahol $\eta^*(x) = \mathbf{P}\{Y = 1 \mid X = x\}$ az a poszteriori valószínűségfüggvény, amely ebben az esetben megegyezik az $\mathbf{E}\{Y \mid X = x\}$ regressziófüggvénnyel. A Bayes-döntés hibáját, $L(g^*)$-ot, Bayes-hibának nevezik, és L^*-gal is jelölik.

A Bayes-döntéshez ismernünk kellene Y-nak X feltétel melletti feltételes eloszlásait. Az alakfelismerés feladatánál ezek ismeretlenek, de – mint regresszióbecslésnél, itt is – rendelkezésünkre állnak a $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ adatok, melyek (X, Y) független és azonos eloszlású példányai. Egy olyan adatoktól függő

$$g_n(x) = g_n(x, D_n) \,:\, \mathcal{R}^d \times \left(\mathcal{R}^d \times \{0,1\}\right)^n \mapsto \{0,1\}$$

osztályozási szabályt szeretnénk megkonstruálni, amelyre

$$L_n \stackrel{\text{def}}{=} L(g_n) = \mathbf{P}\{g_n(X) \neq Y \mid D_n\}$$

közel van L^*-hoz. Tehát az X szerepét ismét az (X, Y) pár játsza, P-nek g^* felel meg, a veszteségfüggvény pedig az $l(g_n, g^*) \stackrel{\text{def}}{=} L_n - L^*$ hiba.

Most is kétfajta univerzális konzisztenciáról beszélhetünk:

A g_n osztályozási szabály *gyengén univerzálisan konzisztens,* ha

$$\mathbf{E}L_n = \mathbf{P}\{g_n(X) \neq Y\} \to L^*$$

(X, Y) minden eloszlására, illetve *erősen univerzálisan konzisztens,* ha

$$\mathbf{P}\{g_n(X) \neq Y \mid D_n\} \to L^* \quad 1 \text{ valószínűséggel}$$

(X, Y) minden eloszlására.

Mivel az $\eta^*(x)$ a poszteriori valószínűségfüggvény egy speciális regressziófüggvény, ezért sikeres η_n regresszióbecslésekből a

$$g_n(x) = \begin{cases} 1, & \text{ha } \eta_n(x) > 1/2, \\ 0 & \text{különben} \end{cases}$$

behelyettesítéssel sikeres osztályozási szabályok vezethetők le, ugyanis

$$L(g_n) - L^* = 2\mathbf{E}\left\{ I_{\{g_n(X) \neq g^*(X)\}} \left| \eta^*(X) - \frac{1}{2} \right| \right\} \tag{3}$$
$$\leq 2\mathbf{E}\left\{ I_{\{g_n(X) \neq g^*(X)\}} \left| \eta^*(X) - \eta_n(X) \right| \right\} \leq 2\sqrt{J(\eta_n)}$$

([27]). Tehát a konzisztens regresszióbecslésekből adódóan most is tekinthetjük a következő három, lokális többségen alapuló, univerzálisan konzisztens osztályozási szabályt:

A k-legközelebbi szomszéd szabály a

$$g_n(x) = \arg\max_{j \in \{0,1\}} \sum_{i=1}^{n} W_{ni}(x) I_{\{Y_i = j\}}$$

döntésfüggvény, ahol W_{ni} az előző fejezetben definiált legközelebbi szomszéd súly.

A partíciós szabály a

$$g_n(x) = \arg\max_{j\in\{0,1\}} \sum_{i=1}^{n} I_{\{X_i\in A_n(x)\}} I_{\{Y_i=j\}}$$

döntésfüggvény, ahol $A_n(x)$ a $\mathcal{P}_n$ azon cellája, amelybe x beleesik. Azaz minden cellán többségi döntést hoz az odaeső minták címkéi alapján.

A magfüggvényes szabály a

$$g_n(x) = \arg\max_{j\in\{0,1\}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right) I_{\{Y_i=j\}}$$

döntésfüggvény.

(3) miatt ezek az osztályozási szabályok (erősen) univerzálisan konzisztensek azon feltételek teljesülése esetén, amikor a megfelelő regresszióbecslők (erősen) univerzálisan konzisztensek ([27]).

Bár megmutatható, hogy az alakfelismerési feladat könnyebb, mint a regresszióbecslés, lassú konvergenciasebességre vonatkozó eredmény itt is ismert:

4.1. Tétel (Devroye [18], Devroye et al. [27]). *Osztályozási szabályok minden $\{g_n\}$ sorozatához, és pozitív számok minden monoton, 0-hoz tartó $1/16 \ge b_1 \ge b_2 \ge \dots$ sorozatához létezik (X, Y)-nak olyan eloszlása, amelyre X egyenletes eloszlású $[0, 1]$-en, $L^* = 0$ ($Y = \eta(X)$) és*

$$\mathbf{E}L_n = \mathbf{P}\{g_n(X) \neq Y\} \ge b_n \quad \textit{minden } n\textit{-re.}$$

A következő eredmény rámutat, hogy bizonyos feltételek mellett létezik konvergenciasebesség. Legyen most g_n a fentebb definiált partíciós szabály, m_n jelölje a partícióban szereplő cellák számát, és legyen R^*_n a Bayes-hibának a $\mathcal{P}_n$ partícióra való megszorítása, azaz

$$R^*_n = \sum_{j=1}^{m_n} \min\{\nu(A_{nj}), \mu(A_{nj}) - \nu(A_{nj})\},$$

ahol $\nu(A) = \mathbf{E}\{Y I_{\{X\in A\}}\}$.

4.2. Tétel (Pintér [46]). *Ha létezik egy c konstans úgy, hogy minden n-re és j-re*

$$\left|\mu(A_{nj}) - 2\nu(A_{nj})\right| > c\mu(A_{nj}),$$

akkor

$$\mathbf{E}(L_n - R^*_n) \le \frac{12}{ce}\frac{m_n}{n}.$$

Ha még az is teljesül, hogy

$$\lim_{n\to\infty} \frac{m_n^2}{n} = 0,$$

akkor

$$\sqrt{n}(L_n - R_n^*) \to 0$$

sztochasztikusan, ha $n \to \infty$.

Most ismét feltételezzük, hogy az (X, Y) eloszlása egy szűkebb eloszlásosztályból való. Tegyük fel, hogy az η^* a poszteriori valószínűségfüggvény (regressziófüggvény) benne van egy adott $\mathcal{F}$ függvényosztályban. Ekkor a feladat az $\mathcal{F}$ osztályból kiválasztani egy η függvényt úgy, hogy az általa meghatározott $g(x) = I_{\{\eta(x) \geq 1/2\}}$ döntés hibavalószínűsége kicsi legyen.

Az $\mathcal{F}$ osztály megállapításában sokféle szempont játszhat szerepet, például az osztályozandó adat eloszlásáról rendelkezésre álló információ, számítási megfontolások. Az osztály elég nagy kell, hogy legyen ahhoz, hogy jó közelítést adjon az optimális megoldásra, de nem lehet túl nagy, mert akkor az adatok mennyisége nem elegendő arra, hogy jó döntést válasszunk ki belőle. Az, hogy hogyan érdemes mérni az osztály nagyságát, általában függ a módszertől, amellyel a választást végezzük. Regresszióbecslésnél láttuk, hogy egy osztály nagyságának jellemzésére alkalmas az osztály Vapnik–Chervonenkis-dimenziója:

Az $\mathcal{F} : \mathcal{R}^d \mapsto [0, 1]$ osztály szeparáló együtthatója, $s(\mathcal{F}, n)$, az $\mathcal{F}$-beli függvények által meghatározott döntésfüggvények osztályának szeparáló együtthatója, azaz az a szám ahányféleképpen n pontot az osztálybeli függvényekkel osztályozni lehet.

Az $\mathcal{F}$ osztály szeparálja az $x_1, \ldots, x_n$ pontokat, ha minden $E \subseteq \{x_1, \ldots, x_n\}$-hez létezik olyan $\eta_E \in \mathcal{F}$, amelyre $\eta_E(x) \geq 1/2$, ha $x \in E$, és $\eta_E(x) < 1/2$, ha $x \in \{x_1, \ldots, x_n\} \setminus E$, azaz az η_E által meghatározott g_E döntésfüggvény az E-beli pontokban 1-re, az $\{x_1, \ldots, x_n\} \setminus E$-beli pontokban 0-ra dönt. Ekkor $s(\mathcal{F}, n) = 2^n$.

Az $\mathcal{F}$ osztály Vapnik–Chervonenkis-dimenziója $V_{\mathcal{F}}$ a legnagyobb olyan n_0 szám, amelyre még létezik olyan n_0 pont, amelyet $\mathcal{F}$ szeparál. Ha minden n-re létezik ilyen n pont, akkor $V_{\mathcal{F}} = \infty$.

A Vapnik–Chervonenkis-dimenziónak számos általánosítása létezik.

4.1. Definíció (Pollard [48]). Az $\mathcal{F} : \mathcal{R}^d \mapsto [0, 1]$ osztály P-szeparálja az $x_1, \ldots, x_n$ pontokat, ha létezik egy $s : \{x_1, \ldots, x_n\} \mapsto \mathcal{R}$ függvény úgy, hogy minden $E \subseteq \{x_1, \ldots, x_n\}$-hez létezik olyan $\eta_E \in \mathcal{F}$, amelyre $\eta_E(x) > s(x)$, ha $x \in E$, és $\eta_E(x) \leq s(x)$, ha $x \in \{x_1, \ldots, x_n\} \setminus E$. Az $\mathcal{F}$ osztály *P-dimenziója* a legnagyobb olyan n_0 szám, amelyre még létezik olyan n_0 pont, amelyet $\mathcal{F}$ P-szeparál. Ha minden n-re létezik ilyen n pont, akkor $P(\mathcal{F}) = \infty$.

A P-dimenzióval rokon a V-dimenzió (Vapnik [55]), a különbség, hogy ott a definícióban az $s(x)$ függvény helyett $\alpha \in \mathcal{R}$ konstans szerepel.

Az utóbbi dimenzióknak létezik paraméterezett, skálaérzékeny változata is.

4.2. Definíció (Kearns és Shapire [38]). Legyen γ pozitív valós szám. Az $\mathcal{F}$ osztály P_γ-szeparálja az $x_1, \ldots, x_n$ pontokat, ha létezik egy $s : \{x_1, \ldots, x_n\} \mapsto \mathcal{R}$ függvény úgy, hogy minden $E \subseteq \{x_1, \ldots, x_n\}$-hez létezik olyan $\eta_E \in \mathcal{F}$, amelyre $\eta_E(x) \geq s(x) + \gamma$, ha $x \in E$, és $\eta_E(x) \leq s(x) - \gamma$, ha $x \in \{x_1, \ldots, x_n\} \setminus E$.

Az $\mathcal{F}$ osztály P_γ-*dimenziója* a legnagyobb olyan n_0 szám, amelyre még létezik olyan n_0 pont, amelyet $\mathcal{F}$ P_γ-szeparál. Ha minden n-re létezik ilyen n pont, akkor $P_\gamma(\mathcal{F}) = \infty$.

Hasonlóan definiálható a V-dimenzió skálaérzékeny változata is.

A definíciókból következik, hogy $V_\gamma(\mathcal{F}) \le P_\gamma(\mathcal{F})$, $V_\gamma(\mathcal{F}) \le V(\mathcal{F})$, $P_\gamma(\mathcal{F}) \le P(\mathcal{F})$, valamint hogy $V_\mathcal{F} \le V(\mathcal{F}) \le P(\mathcal{F})$. Azonban könnyen belátható, hogy a Vapnik–Chervonenkis-dimenzió és a skálaérzékeny P_γ és V_γ dimenziók között nem állítható fel ilyen egyetemes kapcsolat. Ha $\mathcal{F}$ az összes olyan $\mathcal{R}$-en értelmezett függvényt tartalmazza, melyek értéke a $[0,1/2]$-be esik, ha $x<0$, és az $(1/2,1]$-be, ha $x \ge 0$, akkor $V_\mathcal{F} = 0$, de $P_\gamma(\mathcal{F}) = \infty$ minden γ-ra. Másrészt, ha $\mathcal{F}$ az összes olyan, pozitív egészeken definiált függvényt tartalmazza, melyekre $\bigl|\eta(x) - 1/2\bigr| \le e^{-x}/2$, akkor $P_\gamma(\mathcal{F}) = \lfloor -\log 2\gamma \rfloor$, de $V_\mathcal{F} = \infty$.

4.3. Definíció (Lugosi és Pintér [45]). Legyen $0 \le \gamma \le 1/2$. Az $\mathcal{F}$ osztály γ-szeparálja az $x_1,\dots,x_n$ pontokat, ha létezik egy $s : \{x_1,\dots,x_n\} \mapsto [1/2-\gamma, 1/2+\gamma]$ függvény úgy, hogy minden $E \subseteq \{x_1,\dots,x_n\}$-hez létezik olyan $\eta_E \in \mathcal{F}$, amelyre $\eta_E(x) \ge s(x) + \gamma$, ha $x \in E$, és $\eta_E(x) < s(x) - \gamma$, ha $x \in \{x_1,\dots,x_n\} \setminus E$.

Az $\mathcal{F}$ osztály γ-*dimenziója* a legnagyobb olyan n_0 szám, amelyre még létezik olyan n_0 pont, amelyet $\mathcal{F}$ γ-szeparál. Ha minden n-re létezik ilyen n pont, akkor $d_\gamma(\mathcal{F}) = \infty$.

A definíciókból könnyen belátható, hogy $d_\gamma(\mathcal{F}) \le P_\gamma(\mathcal{F})$ és $d_\gamma(\mathcal{F}) \le V_\mathcal{F}$, azaz $d_\gamma(\mathcal{F}) \le \min(P_\gamma(\mathcal{F}), V_\mathcal{F})$.

Ezen dimenziók segítségével jellemezhető az alakfelismerési módszerek egy nagy csoportja, az empirikus hibaminimalizáláson alapuló módszerek.

4.4. Definíció. Egy g döntésfüggvény *empirikus hibavalószínűsége* az

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} I_{\{g(X_i) \ne Y_i\}}$$

mennyiség.

Az *empirikus hibaminimalizálás módszere* minimalizálja az empirikus hibát az adott függvényosztályon, tehát a

$$\tilde{g}_n(x) = \operatorname*{arg\,min}_{g(x) = I_{\{\eta(x) \ge 1/2\}} : \eta \in \mathcal{F}} L_n(g)$$

döntésfüggvényt választja.

4.3. Tétel (Vapnik és Chervonenkis [57]). *Ha az η^* valódi regressziós függvény $\mathcal{F}$-beli, akkor*

$$\mathbf{E}\bigl\{L(\tilde{g}_n)\bigr\} - L^* \le K \max\left(\sqrt{\frac{L^* \log\bigl(s(\mathcal{F},n)\bigr)}{n}}, \frac{\log\bigl(s(\mathcal{F},n)\bigr)}{n} \right),$$

ahol K egy univerzális konstans.

Mivel a szeparáló együtthatóra igaz, hogy

$$s(\mathcal{F},n) \le \left(\frac{ne}{V_{\mathcal{F}}}\right)^{V_{\mathcal{F}}},$$

ezért

$$\mathbf{E}\big\{L(\tilde{g}_n)\big\} - L^* \le K \max\left(\sqrt{\frac{L^* V_{\mathcal{F}} \log n}{n}}, \frac{V_{\mathcal{F}} \log n}{n}\right).$$

Az empirikus hibaminimalizálás módszere robusztus abban az értelemben, hogy ha feltevésünkkel ellentétben $\eta^* \notin \mathcal{F}$, akkor is elég jól működik. A fenti felső korlátban mindkét oldalon L^* helyére $\inf_{\eta\in\mathcal{F}} L(g)$-t írva, az egyenlőtlenség továbbra is fennáll a becslési hibára.

Az *empirikus négyzeteshiba-minimalizálás módszere* az

$$\eta_n'(x) = \arg\min_{\eta\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\big(\eta(X_i) - Y_i\big)^2$$

függvényt választja, azaz minimalizálja az empirikus L_2-hibát. Az L_2-hiba ugyan inkább a regresszióbecslésben alkalmazott hibakritérium, de láttuk, hogy ha L_2 értelemben jó regresszióbecslőt találunk, akkor abból kézenfekvően adódik jó osztályozási szabály.

4.4. Tétel (Lee, Bartlett és Williamson [42]). *Ha $\eta^* \in \mathcal{F}$, akkor az így választott η_n' által meghatározott g_n' döntésfüggvényre, minden $\varepsilon > 0$-ra*

$$\mathbf{E}\big\{L(g_n')\big\} - L^* \le K\left(\sup_{x_1,\ldots,x_n}\sqrt{\frac{\log N_\infty(\varepsilon, x_1^n, \mathcal{F})}{n}} + \varepsilon\right),$$

ahol $N_\infty(\varepsilon, x_1^n, \mathcal{F})$ az $\mathcal{F}$ osztály a $d_{\infty,n}(\eta_1,\eta_2) = \max_{1\le i\le n}\big|\eta_1(x_i) - \eta_2(x_i)\big|$ empirikus távolsághoz tartozó ε-lefedési száma.

A tétel felső korlátjában szereplő lefedési számot becsülni lehet az $\mathcal{F}$ osztály egy P_γ-dimenziója segítségével (Alon, Ben-David, Cesa-Bianchi és Haussler [1]):

$$\sup_{x_1,\ldots,x_n} N_\infty(\varepsilon, x_1^n, \mathcal{F}) \le 2\left(\frac{4n}{\varepsilon^2}\right)^{P_{\varepsilon/4}\log\left(2en/(P_{\varepsilon/4})\right)}.$$

Mivel vannak olyan esetek, amikor az $\mathcal{F}$ osztály VC-dimenziója végtelen, de $P_\gamma(\mathcal{F})$ véges minden γ-ra, és fordítva, ezért a két módszer nem kompatibilis. Könnyű mutatni olyan szituációkat, amikor $V_{\mathcal{F}} = \infty$, tehát az empirikus hibaminimalizálás megbukik, de az empirikus négyzeteshiba-minimalizálás kis hibavalószínűséget ad, és fordítva.

A lefedésen alapuló becslések először kiválasztják az $\mathcal{F}$ osztály egy valamilyen értelemben vett véges lefedését, azaz egy olyan $\mathcal{F}_\varepsilon \subseteq \mathcal{F}$-et, amelyre teljesül, hogy

$\mathcal{F}$ minden eleméhez van $\mathcal{F}_\varepsilon$-nak egy eleme, amely ε-nál „közelebb" van hozzá, majd ebből a lefedésből választanak empirikus hibaminimalizálással. A különböző módszerek abban különböznek egymástól, hogy hogyan mérik ezt a közelséget. Ilyen módszereket javasolt és vizsgált Benedek és Itai [11], Buescher és Kumar [15, 16], Dudley, Kulkarni, Richardson és Zeitouni [31], Kulkarni [40], Vapnik [54] és Devroye et al. [27].

Ha X eloszlása ismert, akkor azt felhasználhatjuk a lefedés kiválasztásához. [54] és [11] a $d_H(\eta,\eta') = \mathbf{P}\{g(X) \neq g'(X)\}$ Hamming-távolságban fedi le az osztályt, és ebből a lefedésből választ minimalizálva az empirikus hibát. [27] szuprémum normában ($\sup_{x\in\mathcal{X}} |\eta(x) - \eta'(x)|$) fedi le $\mathcal{F}$-et.

Legyen most a „távolság" η_1 és η_2 között $\mathbf{E}\{2|\eta_1(X) - \eta_2(X)|I_{\{g_1(X)\neq g_2(X)\}}\}$, ahol g_i az η_i által meghatározott döntésfüggvény. Ezen távolság választását (3) indokolja.

Ekkor az $\mathcal{F}_\varepsilon^{(1)}$ lefedés az $\mathcal{F}$-beli függvények egy olyan minimális elemszámú osztálya, amelyre teljesül, hogy minden $\eta \in \mathcal{F}$-hez létezik egy $\eta' \in \mathcal{F}_\varepsilon^{(1)}$ úgy, hogy

$$\mathbf{E}\{2|\eta(X) - \eta'(X)|I_{\{g(X)\neq g'(X)\}}\} < \varepsilon.$$

Most válasszuk ki $\mathcal{F}_\varepsilon^{(1)}$ azon elemét, amelynek empirikus hibája a legkisebb:

$$g_n(x) = \mathop{\arg\min}_{g(x)=I_{\{\eta(x)\geq 1/2\}}\,:\,\eta\in\mathcal{F}_\varepsilon^{(1)}} L_n(g)$$

4.5. Tétel (Lugosi és Pintér [45]). *Ha $\eta^* \in \mathcal{F}$, akkor az ily módon választott g_n döntésfüggvényre*

$$\mathbf{E}\{L(g_n)\} - L^* \leq$$

$$\leq \varepsilon + \max\left[\sqrt{\frac{22(L^*+\varepsilon)\log\left(n\left(\left|\mathcal{F}_\varepsilon^{(1)}\right|+1\right)\right)}{n}}, \frac{22\log\left(n\left(\left|\mathcal{F}_\varepsilon^{(1)}\right|+1\right)\right)}{n}\right].$$

Az $\mathcal{F}_\varepsilon^{(1)}$ lefedés definíciójából következik, hogy

$$\left|\mathcal{F}_\varepsilon^{(1)}\right| \leq \min\left(N_1(\varepsilon/2,\mu,\mathcal{F}), N_H(\varepsilon/2,\mu,\mathcal{F})\right),$$

ahol $N_1(\varepsilon,\mu,\mathcal{F})$ az L_1-távolsághoz ($d_1(\eta_1,\eta_2) = \mathbf{E}\{|\eta_1(X) - \eta_2(X)|\}$), $N_H(\varepsilon,\mu,\mathcal{F})$ pedig a Hamming-távolsághoz ($d_H(\eta_1,\eta_2) = \mathbf{P}\{g_1(X) \neq g_2(X)\}$) tartozó lefedési szám. Ezen lefedési számokra bebizonyítható, hogy ha $N_1(\varepsilon,x_1^n,\mathcal{F})$ és $N_H(\varepsilon,x_1^n,\mathcal{F})$ jelöli az empirikus L_1-távolsághoz ($d_{1,n}(\eta_1,\eta_2) = \frac{1}{n}\sum_{i=1}^n |\eta_1(x_i) - \eta_2(x_i)|$), illetve az empirikus Hamming-távolsághoz ($d_{H,n}(\eta_1,\eta_2) = \frac{1}{n}\sum_{i=1}^n I_{\{g_1(x_i)\neq g_2(x_i)\}}$) tartozó lefedési számot, akkor ha $n > c\frac{V_\mathcal{F}}{\varepsilon^2}\log\frac{1}{\varepsilon}$, akkor

$$N_H(\varepsilon/2,\mu,\mathcal{F}) \leq \sup_{x_1,\ldots,x_n} N_H(\varepsilon/4,x_1^n,\mathcal{F}) \leq s(\mathcal{F},n) \leq \left(\frac{ne}{V_\mathcal{F}}\right)^{V_\mathcal{F}},$$

valamint, ha $n \geq c\frac{P_{\varepsilon/256}}{\varepsilon^2}\log^2\frac{1}{\varepsilon}$, akkor

$$N_1(\varepsilon/2,\mu,\mathcal{F}) \leq \sup_{x_1,\dots,x_n} N_1(\varepsilon/4, x_1^n, \mathcal{F}) \leq 2\left(\frac{64n}{\varepsilon^2}\right)^{2P_{\varepsilon/16}\log(2en/P_{\varepsilon/16})}.$$

Így tehát a módszer konvergenciasebességét a két dimenzió (VC és P_γ) minimuma határozza meg. Ugyanakkor az algoritmus működhet akkor is, ha $\min(V_{\mathcal{F}}, P_\gamma) = \infty$ minden γ-ra. Gondoljunk például az összes $\left[0, \frac{1}{2}+\frac{\varepsilon}{4}\right]$ értékkészletű függvényt tartalmazó osztályra. Ekkor $\min(V_{\mathcal{F}}, P_\gamma) = \infty$, és könnyű meggondolni, hogy mind az empirikus hibaminimalizálás, mind az empirikus négyzeteshiba-minimalizálás hatalmas hibavalószínűségeket adhat. Ugyanakkor $\left|\mathcal{F}^{(1)}\right| = 1$, hiszen $\mathcal{F}$-et lefedi az $\eta \equiv \frac{1}{2}-\frac{\varepsilon}{4}$ függvény.

Ha nem ismerjük X eloszlását, akkor a „távolság" ilyenkor az adataink egy részétől függ, azaz az adataink egyik felének felhasználásával megkonstruáljuk egy lefedését $\mathcal{F}$-nek, majd ebből az adatok második felén empirikusan legjobb jelöltet választjuk. A lefedéshez használt metrika ilyenkor mindig más, mint az, amellyel a lefedésből kiválasztunk egy döntést.

A $D_n = \big((X_1,Y_1),\dots,(X_n,Y_n)\big)$ adathalmazt vágjuk két részre:

$$D_m = \big((X_1,Y_1),\dots,(X_m,Y_m)\big)$$

és

$$T_{n-m} = \big((X_{m+1},Y_{m+1}),\dots,(X_n,Y_n)\big).$$

D_m segítségével konstruáljunk egy $\mathcal{F}_\varepsilon^{(2)}$ lefedését $\mathcal{F}$-nek az empirikus L_1-távolság szerint, azaz legyen $\mathcal{F}_\varepsilon^{(2)}$ egy minimális elemszámú olyan részhalmaza $\mathcal{F}$-nek, amelyre minden $\eta \in \mathcal{F}$-hez létezik egy $\eta' \in \mathcal{F}_\varepsilon^{(2)}$, amelyre

$$\frac{2}{m}\sum_{i=1}^{m}\big|\eta(X_i) - \eta'(X_i)\big| < \varepsilon.$$

Ekkor nyilván $\big|\mathcal{F}_\varepsilon^{(2)}\big| = N_1(\varepsilon/2, X_1^m, \mathcal{F})$. Válasszuk ki $\mathcal{F}_\varepsilon^{(2)}$ azon elemét, amelynek az adatok második felén (T_{n-m}) vett empirikus hibája minimális:

$$g_n(x) = \operatorname*{arg\,min}_{g(x)=I_{\{\eta(x)\geq 1/2\}}\,:\,\eta\in\mathcal{F}_\varepsilon^{(2)}} L_{n-m}(g) =$$

$$= \operatorname*{arg\,min}_{g(x)=I_{\{\eta(x)\geq 1/2\}}\,:\,\eta\in\mathcal{F}_\varepsilon^{(2)}} \frac{1}{n-m}\sum_{i=m+1}^{n} I_{\{g(X_i)\neq Y_i\}}.$$

4.6. Tétel (Lugosi és Pintér [45]). *Ha $\eta^* \in \mathcal{F}$, akkor minden n-re, $m = n/2$-re és ε-ra*

$$\mathbf{E}\{L(g_n)\} - L^* \le 6\varepsilon +$$
$$+ 3\max\left[\sqrt{\frac{11L^*\log\left(2n\mathbf{E}N_1(\varepsilon/2, X_1^{n/2}, \mathcal{F})\right)}{n}}, \frac{512\log\left(10n\mathbf{E}N_1(\varepsilon/32, X_1^{n/2}, \mathcal{F})\right)}{n}\right].$$

Az $N_1(\varepsilon/2, X_1^{n/2}, \mathcal{F})$ és $N_1(\varepsilon/32, X_1^{n/2}, \mathcal{F})$ lefedési számokra felső becslés adható a már látott módon $\mathcal{F}$ valamely P_γ-dimenziójának függvényében.

Most vegyük $\mathcal{F}$ egy lefedését a korábban látott

$$\mathbf{E}\{2|\eta_1(X) - \eta_2(X)|I_{\{g_1(X)\neq g_2(X)\}}\}$$

„metrika" empirikus változata szerint. Legyen $\mathcal{F}_\varepsilon^{(3)}$ az $\mathcal{F}$-beli függvények minimális elemszámú olyan osztálya, amelyre minden $\eta \in \mathcal{F}$-hez van $\eta' \in \mathcal{F}_\varepsilon^{(3)}$ úgy, hogy

$$\frac{2}{m}\sum_{i=1}^{m}|\eta(X_i) - \eta'(X_i)|I_{\{g(X_i)\neq g'(X_i)\}} < \varepsilon.$$

És legyen

$$g_n(x) = \mathop{\arg\min}_{g(x)=I_{\{\eta(x)\ge 1/2\}}\,:\,\eta\in\mathcal{F}_\varepsilon^{(3)}} L_{n-m}(g).$$

Ekkor az $|\mathcal{F}_\varepsilon^{(3)}|$ lefedési szám függvényében lehet az előzőekhez hasonló felső korlátot adni a hibavalószínűségre ([45]).

Az $|\mathcal{F}_\varepsilon^{(3)}|$ lefedési számról pedig bebizonyítható az alábbi felső korlát:

4.7. Tétel (Lugosi és Pintér [45]). *Ha $12m \ge \lceil 1/\varepsilon\rceil$, $\varepsilon \le 1$, akkor $X_1, \dots, X_m$ minden értékére*

$$|\mathcal{F}_\varepsilon^{(3)}| \le 2(6m)^{d_{\varepsilon/10}\log_2(2em/d_{\varepsilon/10})+1}.$$

Tehát ebben az esetben a konvergenciasebesség $\mathcal{F}$ egy d_γ dimenziójától függ.

A lassú kovergenciára vonatkozó 4.1. tétel állítását úgy is fogalmazhatjuk, hogy minden nullához tartó pozitív sorozat individuális alsó konvergenciasebesség az $X = \text{Uniform}\,[0,1]$-t és $Y = \eta(X)$-t teljesítő eloszlások osztályára nézve.

Vizsgáljuk most meg az $\mathcal{F} = \mathcal{F}^{(p,M)}$ „sima" függvényosztályokat (2.2. definíció) véve az alább definiált eloszlásosztályokra vonatkozó alsó és éles konvergenciasebességeket az alakfelismerés esetében is:

4.5. Definíció. Legyen $\mathcal{D}^{*(p,M)}$ (X, Y) azon eloszlásainak halmaza, amelyekre

(I′) X egyenletes eloszlású $[0,1]^d$-en,

(II′) $Y \in \{0,1\}$ 1 valószínűséggel,

(III′) $\eta \in \mathcal{F}^{(p,M)}$.

4.8. TÉTEL (Yang [59]). *Az* $n^{-\frac{p}{2p+d}}$ *sorozat a* $\mathcal{D}^{*(p,M)}$ *eloszlásosztályra nézve minimax alsó konvergenciasebesség, és ez éles.*

4.9. TÉTEL (Antos [4]). *Legyen* b_n *egy tetszőleges nullához tartó pozitív sorozat. Ekkor a*

$$b_n n^{-\frac{p}{2p+d}}$$

sorozat a $\mathcal{D}^{*(p,M)}$ *eloszlásosztályra nézve individuális alsó konvergenciasebesség, azaz* $n^{-\frac{p}{2p+d}}$ *éles individuális konvergenciasebesség.*

A következőkben az alakfelismerés egy olyan speciális esetével foglalkozunk, amikor az $L^* = 0$, azaz Y az X függvénye. Ezt az esetet *tanulás*nak hívjuk. Legyen $\mathcal{C}$ az $\mathcal{R}^d$ részhalmazainak egy osztálya. $\mathcal{C}$ elemeit koncepcióknak nevezzük, és $\mathcal{C}$ egy koncepcióosztály. Legyen $\mathcal{D}$ az (X, Y) azon eloszlásainak osztálya, amelyekre $Y = I_{\{X \in C\}}$ valamely $C \in \mathcal{C}$-re. Tehát ebben az esetben egy ismeretlen $C \in \mathcal{C}$ koncepciót kell megtanulni a

$$D_n = \Big((X_1, I_{\{X_1 \in C\}}), \dots, (X_n, I_{\{X_n \in C\}}) \Big)$$

adatból. Mivel $L^* = 0$, ezért $l(g, g^*) = L(g)$. Az (X, Y) együttes eloszlását az X eloszlásából és C-ből álló *eloszlás-koncepció pár* határozza meg.

A minimax várható hibavalószínűség viselkedését is alapvetően a $\mathcal{C}$ koncepcióosztály $V_{\mathcal{C}}$ VC-dimenziója (3.1. definíció) határozza meg:

4.10. TÉTEL (Vapnik és Chervonenkis [57], Haussler, Littlestone és Warmuth [37]). *A* $V_{\mathcal{C}}/n$ *sorozat a* $\mathcal{D}$ *eloszlásosztályra nézve minimax alsó konvergenciasebesség, és ez éles.*

A fenti minimax alsó korlátok nem minden esetben terjeszthető ki individuális alsó korláttá (Antos [3]), azonban sok fontos „geometriai” koncepcióosztály esetében igen. A VC-dimenzió szerepét itt a koncepcióosztály paraméterszáma játsza. Az alábbi koncepcióosztályok VC-dimenziója megegyezik a k paraméterszámmal:

Legyen $\mathcal{C}_k^1$ a k darab „kezdőszegmens” unióinak osztálya, azaz X tartója legyen $[0,1] \times \{1, 2, \dots, k\}$ és

$$\mathcal{C}_k^1 = \left\{ \bigcup_{j=1}^{k} \big([0, z_j] \times \{j\} \big) \; : \; z \in [0,1]^k \right\}.$$

Legyen $\mathcal{C}_k^2$ a k-dimenziós „negyedek” osztálya $\mathcal{R}^k$-ban, azaz

$$\mathcal{C}_k^2 = \big\{ \{ x \in \mathcal{R}^k \; : \; x_i \le a_i, i = 1, \dots, k \} \; : \; a_1, \dots, a_k \in \mathcal{R} \big\}$$

ahol $x_1, \dots, x_k$ az x vektor komponensei.

Legyen $\mathcal{C}_k^3$ a k-dimenziós „félterek” osztálya $\mathcal{R}^k$-ban, azaz

$$\mathcal{C}_k^3 = \left\{ \left\{ x : \sum_{i=1}^{k} a_i x_i + a_0 \geq 0 \right\} : a_0, a_1, \ldots, a_k \in \mathcal{R} \right\}.$$

([3]-ban további koncepcióosztályok is találhatóak, amelyekre szintén igazak az alábbi eredmények.)

4.11. Tétel (Antos és Lugosi [6], Schuurmans [49]). *Ha $\mathcal{C}$ a $\mathcal{C}_k^1$, $\mathcal{C}_k^2$ és $\mathcal{C}_k^3$ egyike, akkor bármely g_n osztályozási szabály sorozatra létezik egy eloszlás és egy $C \in \mathcal{C}$ koncepció úgy, hogy minden $0 < \varepsilon < 1$-re,*

$$\mathbf{E}\big\{L(g_n)\big\} > (1-\varepsilon)\frac{k}{2n} \quad \textit{végtelen sok } n\textit{-re},$$

azaz k/n éles individuális konvergenciasebesség $\mathcal{D}$-re nézve.

Legyen g_n egy osztályozási szabály sorozat. A *kumulatív hiba*

$$\frac{1}{n}\sum_{i=1}^{n} I_{\{g_i(X_{i+1},D_i) \neq I_{\{X_{i+1} \in C\}}\}},$$

a sorozat által az első n lépésben elkövetett hibák relatív gyakorisága, ha mindig az első i mintát használjuk az $i+1$-edik minta címkéjének eldöntésére. Most a veszteségfüggvény az

$$l\big(\{g_i\}_{i=1}^n, g^*\big) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L(g_i)$$

kumulatív hibavalószínűség.

4.12. Tétel (Haussler, Littlestone és Warmuth [37]). *Ha $\mathcal{C}$ a $\mathcal{C}_k^1$, $\mathcal{C}_k^2$ és $\mathcal{C}_k^3$ egyike, akkor a várható kumulatív hibára a $k \log n/n$ sorozat a $\mathcal{D}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség, és ez éles.*

A fenti minimax alsó korlátok következő individuális kiterjesztését bizonyítjuk:

4.13. Tétel (Antos és Lugosi [6]). *Ha $\mathcal{C}$ a $\mathcal{C}_k^1$, $\mathcal{C}_k^2$ és $\mathcal{C}_k^3$ egyike, akkor bármely g_n osztályozási szabály sorozatra létezik egy eloszlás és egy $C \in \mathcal{C}$ koncepció úgy, hogy minden $0 < \varepsilon < 1$-re,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\big\{L(g_i)\big\} > (1-\varepsilon)\frac{k}{2n}\log n \quad \textit{végtelen sok } n\textit{-re},$$

azaz a várható kumulatív hibára $k \log n/n$ éles individuális konvergenciasebesség $\mathcal{D}$-re nézve.

A várható hibavalószínűség hasznos mennyiség az L_n viselkedésének leírásához. Azonban inkább a

$$\mathbf{P}\{L_n \geq \varepsilon\}, \quad \varepsilon \in [0,1]$$

farokvalószínűségek írják le teljesen a hibavalószínűség eloszlását. Most a veszteségfüggvény az

$$l(g_n, g^*) \stackrel{\text{def}}{=} I_{\{L_n \geq \varepsilon\}}$$

indikátor.

A VC-dimenzió a farokvalószínűségekre vonatkozó minimax alsó és felső korlátokat is jellemzi:

4.14. TÉTEL (Anthony et al. [2], Lugosi [44], lásd még Vapnik és Chervonenkis [57], Blumer et al. [14]). *Bármely koncepcióosztályra létezik g_n osztályozási szabály sorozat, hogy*

$$\sup_{(X,Y)\in\mathcal{D}} \mathbf{P}\{L_n \geq \varepsilon\} = O\left(\left(\frac{n^2\varepsilon}{V_{\mathcal{C}}}\right)^{V_{\mathcal{C}}} e^{-n\varepsilon}\right).$$

4.15. TÉTEL (Devroye és Lugosi [29]). *Ha $V_{\mathcal{C}} \geq 2$ és $\varepsilon < 1/4$, akkor az*

$$\left(\frac{n\varepsilon}{V_{\mathcal{C}}}\right)^{(V_{\mathcal{C}}-1)/2} e^{-4n\varepsilon/(1-4\varepsilon)}$$

sorozat a $\mathcal{D}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség.

Nyilván ε legérdekesebb értékei az $1/n$-nek a konstans többszörösei, mivel ez az a tartomány, ahova egy jó g_n osztályozási szabály L_n hibavalószínűsége nagy valószínűséggel esik. Ilyen $\{\varepsilon = \varepsilon_n\}$ sorozatokra a fenti minimax korlátok következő típusú individuális kiterjesztését bizonyítjuk, ha $\mathcal{C}$ a „geometriai" koncepcióosztályok egyike:

4.16. TÉTEL (Antos és Lugosi [6]). *Legyenek $\varepsilon_1, \varepsilon_2, \ldots$ nemnegatív számok úgy, hogy $\{\gamma_n = n\varepsilon_n\}$ nem tart végtelenhez, amint $n \to \infty$. Ha $\mathcal{C}$ a $\mathcal{C}_k^1$, $\mathcal{C}_k^2$ és $\mathcal{C}_k^3$ egyike, akkor bármely $\{g_n\}$ osztályozási szabály sorozatra létezik egy eloszlás és egy $C \in \mathcal{C}$ koncepció úgy, hogy minden $\delta \in (0,1)$-re,*

$$\mathbf{P}\{L_n \geq \varepsilon_n\} \geq (1-\delta)\frac{1}{2}\sum_{i=0}^{k-1}\frac{(c\gamma_n)^i}{i!}e^{-c\gamma_n} \quad \text{végtelen sok } n\text{-re,}$$

ahol $c = \log 256 \approx 5.545$, azaz $(cn\varepsilon_n/k)^{k-1}e^{-cn\varepsilon_n}$ és így $\{1\}$ individuális alsó konvergenciasebesség $\mathcal{D}$-re nézve.

Irodalomjegyzék

[1] Alon, N., Ben-David, S., Cesa-Bianchi, N. and Haussler, D., Scale-sensitive dimensions, uniform convergence, and learnability, *Technical Report 143-95, DSI, University of Milan* (Italy, 1993),

[2] Anthony, M., Biggs, N. L. and Shawe-Taylor, J., Bounding sample size with the Vapnik–Chervonenkis dimension, *Discrete Applied Mathematics* **42** (1993), 65–73.

[3] Antos, A., *Performance limits of nonparametric estimators,* Ph.D. thesis, Technical University of Budapest, H-1521 Stoczek u. 2 (Budapest, Hungary, 1999).

[4] Antos, A., Lower bounds for the rate of convergence in nonparametric pattern recognition, *Theoretical Computer Science* **284(1)** (2002), pp. 3–24. An extended abstract appeared in the Proceedings of the 4th European Conference on Computational Learning Theory (EUROCOLT'99).

[5] Antos, A., Györfi, L. and Kohler, M., Lower bounds on the rate of convergence of nonparametric regression estimates, *Journal of Statistical Planning and Inference* **83**, No. 1 (2000), 91–100.

[6] Antos, A. and Lugosi, G., Strong minimax lower bounds for learning, *Machine Learning* **30** (1998), 31–56.

[7] Barron, A. R., The convergence in information of probability density estimates, *IEEE ISIT* (Kobe, Japan, 1988).

[8] Barron, A. R., Birgé, L. and Massart, P., Risk bounds for model selection via penalization, *Probability Theory and Related Fields* **113** (1999), 301–413.

[9] Barron, A. R., Györfi, L. and van der Meulen, E. C., Distribution estimates consistent in total variation and in two types of information divergence, *IEEE Trans. on Information Theory* **38** (1992), 1437–1454.

[10] Beirlant, J. and Györfi, L., On the L_1-error in histogram density estimation: the multidimensional case, *J. Nonparametric Statistics* **9** (1998), 197–216.

[11] Benedek, G. M. and Itai, A., Learnability by fixed distributions, in: *Computational Learning Theory: Proceedings of the 1988 Workshop,* Morgan Kaufman (San Mateo, CA, 1988), pp. 80–90.

[12] Berlinet, A., Devroye, L. and Györfi, L., Asymptotic normality of L_1 error in density estimation, *Statistics* **26** (1995), 329–343.

[13] Birgé, L., On estimating a density using Hellinger distance and some other strange facts, *Probability Theory and Related Fields* **71** (1986), 271–291.

[14] Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K., Learnability and the Vapnik–Chervonenkis dimension, *Journal of the ACM* **36** (1989), 929–965.

[15] Buescher, K. L. and Kumar, P. R., Learning by canonical smooth estimation, Part I: Simultanious estimation, *IEEE Transactions on Automatic Control* **41** (1996), 545–556.

[16] Buescher, K. L. and Kumar, P. R., Learning by canonical smooth estimation, Part II: Learning and choice of model complexity, *IEEE Transactions on Automatic Control* **41** (1996), 557–569.

[17] Csiszár, I., Information-type measures of divergence of probability distributions and indirect observations, *Studia Sci. Math. Hungar.* **2** (1967), 299–318.

[18] Devroye, L., Any discrimination rule can have an arbitrary bad probability of error for finite sample size, *IEEE Trans. Pattern Analysis and Machine Intelligence* PAMI-4 (1982), 154–157.

[19] Devroye, L., The equivalence of weak, strong and complete convergence in L1 for kernel density estimates, *Annals of Statistics* **11** (1983), 896–904.

[20] Devroye, L., On arbitrary slow rates of global convergence in density estimation, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **62** (1983), 475–483.

[21] Devroye, L., Exponential inequalities in nonparametric estimation, in: *Nonparametric Functional Estimation* (ed. G. Roussas), NATO ASI Series, Kluwer Academic Publishers (Dordrecht, 1991), pp. 31–44.

[22] Devroye, L., Another proof of a slow convergence result of Birgé, *Statistics and Probability Letters* **23** (1995), 63–67.

[23] Devroye, L. and Györfi, L., Distribution-free exponential upper bound on the L_1 error of partitioning estimates of a regression function, in: *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics* (Konecny, F., Mogyoródi, J. and Wetz, W., eds.), Akadémiai Kiadó (Budapest, 1983), pp. 67–76.

[24] Devroye, L. and Györfi, L., *Nonparametric Density Estimation: the L_1 View,* Wiley (1985).

[25] Devroye, L. and Györfi, L., No empirical measure can converge in total variation sense for all distributions, *Annals of Statistics* **18** (1990), 1496–1499.

[26] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G., On the strong universal consistency of nearest neighbor regression function estimates, *Annals of Statistics* **22** (1994), 1371–1385.

[27] Devroye, L., Györfi, L. and Lugosi, G., *A Probabilistic Theory of Pattern Recognition,* Springer Verlag (1996).

[28] Devroye, L. and Krzyżak, A., An equivalence theorem for L1 convergence of the kernel regression estimate, *Journal of Statistical Planning and Inference* **23** (1989), 71–82.

[29] Devroye, L. and Lugosi, G., Lower bounds in pattern recognition and learning, *Pattern Recognition* **28** (1995), 1011–1018.

[30] Devroye, L. and Wagner, T. J., Distribution-free consistency results in nonparametric discrimination and regression function estimation, *Annals of Statistics* **8** (1980), 231–239.

[31] Dudley, R. M., Kulkarni, S., Richardson, T., and Zeitouni, O., A metric entropy bound is not sufficient for learnability, *IEEE Transactions on Information Theory* **40** (1994), 883–885.

[32] Györfi, L., Universal consistencies of regression estimate for unbounded regression functions, in: *Nonparametric functional estimation and related topics* (G. Roussas, ed.), Kluwer Academic Publishers (Dordrecht, 1991), pp. 329–338.

[33] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H., *A Distribution-Free Theory of Nonparametric Regression,* Springer Verlag (2002).

[34] Györfi, L., Kohler, M. and Walk, H., Weak and strong universal consistency of semi-recursive kernel and partitioning regression estimates, *Statist. Decisions* **16** (1998), 1–18.

[35] Györfi, L., Páli, I. and van der Meulen, E. C., There is no universal source code for infinite alphabet, *IEEE Trans. on Information Theory* **40** (1994), 267–271.

[36] Györfi, L. and van der Meulen, E. C., There is no density estimate consistent in information divergence for all densities, *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes* (1994), pp. 88–90.

[37] Haussler, D., Littlestone, N. and Warmuth, M., Predicting $\{0,1\}$-functions on randomly drawn points, *Information and Computation* **115** (1994), 248–292.

[38] Kearns, M. and Shapire, R. E., Efficient distribution-free learning of probabilistic concepts, *Journal of Computer Systems Sciences* **48** (1994), 464–497.

[39] Kemperman, J. H. B., An optimum rate of transmitting information, *Ann. Math. Statist.* **40** (1969), 2156–2177.

[40] Kulkarni, S. R., *Problems of computational and information complexity in machine vision and learning,* Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT (1991).

[41] Kullback, S., A lower bound for discrimination in terms of variation, *IEEE Trans. Information Theory* **13** (1967), 126–127.

[42] Lee, W. S., Bartlett, P. and Williamson, R. C., Efficient agnostic learning of neural networks with bounded fan-in, *IEEE Transactions on Information Theory* **42** (1996), 2118–2132.

[43] Liese, F. and Vajda, I., *Convex Statistical Distances,* Teubner (Leipzig, 1987).

[44] Lugosi, G., Improved upper bounds for probabilities of uniform deviations, *Statistics and Probability Letters* **25** (1995), 71–77.

[45] Lugosi, G. and Pintér (Horváth), M., Scale-sensitive dimensions and skeleton estimates for classification, *Discrete Applied Mathematics* **85** (1998), 37–61.

[46] Pintér, M., On the rate of convergence of error estimates for the partitioning classification rule, *Theoretical Computer Science* **284(1)** (2002), 181–196. An extended abstract appeared in the Proceedings of the 4th European Conference on Computational Learning Theory (EUROCOLT'99),

[47] Pollard, D., *Convergence of Stochastic Processes,* Springer-Verlag (New York, 1984).

[48] Pollard, D., *Empirical Processes: Theory and Applications,* NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Math. Stat. and Am. Stat. Assoc (1990).

[49] Schuurmans, D. E., *Effective classification learning,* Ph.D. thesis, University of Toronto (Toronto, CA, 1996).

[50] Spiegelman, C. and Sacks, J., Consistent window estimation in nonparametric regression, *Annals of Statistics* **8** (1980), 240–246.

[51] Stone, C. J., Consistent nonparametric regression, *Annals of Statistics* **5** (1977), 595–645.

[52] Stone, C. J., Optimal global rates of convergence for nonparametric regression, *Annals of Statistics* **10** (1982), 1040–1053.

[53] Vajda, I., *Theory of Statistical Inference and Information,* Kluwer Academic Publishers (1989).

[54] Vapnik, V. N., *Estimation of Dependencies Based on Empirical Data,* Springer-Verlag (New York, 1982).

[55] Vapnik, V. N., Inductive principles of the search for empirical dependencies, *Proceedings of the 2nd Annual Workshop on Computational Learning Theory* (1989).

[56] Vapnik, V. N. and Chervonenkis, A. Ya., On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and its Applications* **16** (1971), 264–280.

[57] Vapnik, V. N. and Chervonenkis, A. Ya., *Theory of Pattern Recognition,* Nauka (Moscow, 1974). (in Russian); German translation: *Theorie der Zeichenerkennung,* Akademie Verlag (Berlin, 1979).

[58] Walk, H., Strong universal consistency of smooth kernel regression estimates, *Preprint 2002-8, Math. Institut A, Universität Stuttgart,* Stuttgart (2001) (submitted for publications)

[59] Yang, Y., Minimax nonparametric classification — Part I: Rates of convergence, *IEEE Transactions on Information Theory* **45** (1999), 2271–2284.

(Beérkezett: 2002. május 10.)

ANTOS ANDRÁS
MAGYAR TUDOMÁNYOS AKADÉMIA
SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZETE,
INFORMATIKA KUTATÓ LABOR
ANTOS@SZIT.BME.HU

PINTÉR MÁRTA
BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM,
SZÁMÍTÁSTUDOMÁNYI ÉS INFORMÁCIÓELMÉLETI TANSZÉK
ÉS
MAGYAR TUDOMÁNYOS AKADÉMIA
SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZETE,
INFORMATIKA KUTATÓ LABOR
MARTI@SZIT.BME.HU

ITT A CÍM ANGOL FORDÍTÁSA LENNE

András Antos and Márta Pintér

The main purpose of this paper is to review the limits of nonparametric statistics in the fields of distribution estimation, density estimation, regression estimation and statistical pattern recognition (or classification).

The basic aim of mathematical statistics is to learn a probability law or its characteristics from data. The classical mathematical statistics consists mainly of parametric rules, when the statistician has some a priori information about the actual problem in order to assume that the unknown probability law belongs to a parametric class of distributions.

In case of nonparametric statistics the statistician has no such knowledge, therefore the construction of the rules applied cannot use any specific feature of the law, and the basic properties of the rule should be distribution-free.

We wish to estimate an unknown characteristic of the distribution of an observation. The estimation or prediction is based on previously collected data, independent samples from the distribution, and a loss function is defined to measure the quality of the estimate. For example, in regression estimation this loss function is the mean squared error, in pattern recognition it is the error probability, the probability that our decision is wrong in the sense that we classify the observation into an incorrect class.

An estimate is consistent if the loss function tends to zero (or its theoretical minimal value) in some sense as the sample size goes to infinity for any reasonable distribution. In each case, we can say that consistent estimates exist, but a universal rate of convergence cannot be guaranteed over all these distributions. However, for certain (nonparametric) classes of distributions, one can investigate uniform and individual convergence rates, and can also establish sharp lower bounds for these rates.

Quite often, in addition to the collected data, some supplementary information is also available about the distribution. Then we can exploit this knowledge using methods based on empirical loss minimization and we can give upper bounds on the rate of convergence in terms of appropriate dimensions of the distribution class.