A nemparaméteres statisztika teljesítőképességének a korlátai

Ph.D. értekezés tézisei

Antos András
Budapesti Műszaki Egyetem
Számítástudományi és Információelméleti Tanszék
(email: antos@szit.bme.hu)

1999. április 26.

Témavezető: Dr. Györfi László

1 Az értekezés előzményei és célkitűzése

A '80-as években intenziv fejlődésnek indult az információfeldolgozásnak az az ága, ahol a probléma bonyolultsága illetve a megfigyelési adatok nagy dimenziója miatt a sikeres modellezés már nem követhette a paraméteres statiszika és a lineáris folyamatstatisztika módszertanát. Ezt a területet hívjuk nemparaméteres statisztikának. Az általános modell a következő:

Legyen X egy vektor valószínűségi változó, X eloszlása ismeretlen. Szeretnénk megbecsülni X eloszlásának egy P jellemzőjét, amelynek értéke lehet valós vagy egy függvény vagy akár maga az eloszlás. A becslés jóságát egy nemnegatív $l(\cdot,\cdot)$ veszteségfüggvény méri. Inputként adottak az $X_1,\ldots,X_n$ független minták. A becslést jelölje P_n, amely a minták mindenegyes realizációjához P egy lehetséges becslését rendeli. Az $l(P_n,P)$ nullához való konvergenciáját vizsgáljuk.

Egy P_n becslés **erősen konzisztens egy eloszlásra**, ha

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

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

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.

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

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

minden P-re, azaz a becslés erősen konzisztens $\mathcal{D}$-n, másrészt az $\mathbf{E}\{|P_n - P|\}$-re az $a_n = 1/\sqrt{n}$ 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 a lassú konvergencia eredmények szerint 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 minimax és az individuális konvergenciasebesség különböző lehet, amire a disszertációban példákat is mutatok. 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 a leglassabbat.

A disszertáció Bevezetésében illusztrációként összefoglalom az eloszlás- és sűrűségfüggvény becsléssel kapcsolatban az idevágó eredményeket. A munkám célja, hogy ezeket az eredményeket kiterjesszem a következő területekre:

— regressziófüggvény becslés,
— alakfelismerés,
— várhatóérték becslése,
— entrópia becslése,
— kölcsönös információ becslése,
— Bayes hiba becslése.

2 Kutatási módszerek

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 Új tudományos eredmények

A következőkben összefoglalom a disszertáció fontosabb tételeit némileg egyszerűsített alakban. A definíciók és állítások számozása a disszertációbeli számozáshoz igazodik.

3.1 Regressziófüggvény becslés

Legyen Y valós értékű valószínűségi változó, amelyre $\mathbf{E}\{Y^2\} < \infty$, és legyen X d-dimenziós véletlen vektor (megfigyelés). A regresszióanalízis célja Y becslése, ha X adott, azaz olyan $f : \mathcal{R}^d \mapsto \mathcal{R}$ függvényt keresünk, amelyre $f(X)$ "közel" van Y-hoz. Tegyük fel, hogy az analízis fő célja a négyzetes közép hiba minimalizálása:

$$\min_f \mathbf{E}\{(f(X) - Y)^2\} \ . \tag{1}$$

Legyen

$$m(x) = \mathbf{E}\{Y|X = x\}$$

a regressziófüggvény.

Jól ismert, hogy minden mérhető $f : \mathcal{R}^d \mapsto \mathcal{R}$ függvényre

$$\mathbf{E}\{(f(X) - Y)^2\} = \int |f(x) - m(x)|^2 \mu(dx) + \mathbf{E}\{(m(X) - Y)^2\} \ ,$$

ahol μ az X eloszlása. Ezért (1) az m regressziófüggvénynél éri el a minimumát, és egy tetszőleges f függvény négyzetes közép hibája pontosan akkor lesz közel a minimumhoz, ha

$$||f - m||_2^2 = \int |f(x) - m(x)|^2 \mu(dx)$$

közel van a 0-hoz.

A regresszióbecslés feladatánál (X, Y) eloszlása (és így m) ismeretlen. Adottak azonban a $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ független, azonos eloszlású példányai (X, Y)-nak (minta), és $m(x)$-nek olyan $m_n(x) = m_n(x, D_n) : \mathcal{R}^d \times (\mathcal{R}^d \times \mathcal{R})^n \mapsto \mathcal{R}$ becslését akarjuk megkonstruálni, amelyre $||f - m||_2^2$ kicsi.

Tehát most az X szerepét az (X, Y) pár játsza, P-nek m felel meg, a veszteségfüggvény pedig az

$$l(m_n, m) \stackrel{\text{def}}{=} ||m_n - m||_2^2$$

négyzetes L_2 hiba.

Ismert, hogy létezik univerzálisan konzisztens regresszióbecslés (Stone [1977]).

A lassú konvergenciasebesség eredmény is ismert, és következik az alakfelismerésre vonatkozó hasonló eredményből: Legyen $\{b_n\}$ pozitív számok nullához tartó sorozata úgy, hogy $1/64 \geq b_1 \geq b_2 \geq \dots$. Minden regresszióbecslés sorozatra létezik (X, Y) egy eloszlása úgy, hogy X egyenletes eloszlású $[0,1]$-en, $Y = m(X)$ és

$$\mathbf{E}\{||m_n - m||_2^2\} \geq b_n \quad \text{minden } n\text{-re.}$$

Ezt úgy is fogalmazhatjuk, hogy minden nullához tartó pozitív sorozat individuális alsó konvergenciasebesség az $X = \text{Uniform}[0,1]$-t és $Y = m(X)$-t teljesítő eloszlások osztályára nézve.

Az alább definiált "sima" eloszlásosztályra vonatkozó éles konvergenciasebességeket vizsgáljuk:

2.1. definíció. *Adott $k \in \mathcal{N}_0$-ra, $0 < \beta \leq 1$-re, $p = k + \beta$-ra és $M > 0$-ra, legyen $\mathcal{F}^{(p,M)}$ azon $f : \mathcal{R}^d \to \mathcal{R}$ függvények halmaza, amelyekre minden $\alpha = (\alpha_1, \dots, \alpha_d)$ 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 a parciális deriváltat jelöli, amely az α-hoz tartozik.

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

Az $\{n^{-\frac{2p}{2p+d}}\}$ sorozat a $\mathcal{D}^{(p,M)}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség (Stone [1982]), és ez éles (Barron, Birgé és Massart [1995]).

2.3. tétel. (Antos, Györfi és Kohler [1999]) *Legyen $\{b_n\}$ egy tetszőleges nullához tartó pozitív sorozat. Ekkor a*

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

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

3.2 Alakfelismerés

Legyen Y $\{0,1\}$ értékű valószínűségi változó (osztály), és legyen X d-dimenziós véletlen vektor (megfigyelés). Az alakfelismerés (osztályozás) célja Y értékének eldöntése, ha X adott, azaz olyan $g : \mathcal{R}^d \mapsto \{0,1\}$ döntésfüggvényt keresünk, amelyre $g(X)$ egyenlő Y-nal nagy valószínűséggel. Tegyük fel, hogy az analízis fő célja a hibavalószínűség minimalizálása:

$$\min_g L(g) \stackrel{\text{def}}{=} \min_g \mathbf{P}\{g(X) \neq Y\} . \tag{2}$$

Legyen

$$\eta(x) = \mathbf{P}\{Y = 1|X = x\}$$

az aposzteriori valószínűségfüggvény és legyen

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \ , \\ 0 & \text{különben} \end{cases}$$

a Bayes-döntés. Jelölje

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

a Bayes-hibát.

Jól ismert, hogy minden mérhető $g : \mathcal{R}^d \mapsto \{0, 1\}$ függvényre

$$L(g) - L^* \geq 0 \ .$$

Ezért (2) a g^* függvénynél éri el a minimumát, és a minimum L^*.

Az alakfelismerés feladatánál (X, Y) eloszlása (és így η és g^*) ismeretlen. Adottak azonban a $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ független, azonos eloszlású példányai (X, Y)-nak (minta), és egy olyan $g_n(x) = g_n(x, D_n) : \mathcal{R}^d \times (\mathcal{R}^d \times \{0, 1\})^n \mapsto \{0, 1\}$ osztályozási szabályt akarunk megkonstruálni, amelyre

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

közel van L^*-hoz.

Tehát most az X szerepét 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.

Mivel az η aposzteriori valószínűségfüggvény egy speciális regressziófüggvény, ezért egy sikeres η_n regresszióbecslésekből a

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

behelyettesítéssel sikeres osztályozási szabályok vezethetők le, ugyanis

$$L_n - L^* \leq 2||\eta_n - \eta||_2$$

(Devroye, Györfi és Lugosi [1996]).

Következményképpen univerzálisan konzisztens osztályozási szabály is létezik.

Bár megmutatható, hogy az alakfelismerési feladat könnyebb, mint a regresszióbecslés, a lassú konvergenciasebesség eredmény itt is ismert: Legyen $\{b_n\}$ pozitív számok nullához tartó sorozata úgy, hogy $1/16 \geq b_1 \geq b_2 \geq \ldots$. Minden osztályozási szabály sorozatra létezik (X, Y) egy eloszlása úgy, hogy X egyenletes eloszlású $[0,1]$-en, $Y = \eta(X)$ $(L^* = 0)$ és

$$\mathbf{E}L_n \geq b_n \quad \text{minden } n\text{-re}$$

(Devroye, Györfi és Lugosi [1996]). Ezt 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.

Először az alább definiált "sima" eloszlásosztályra vonatkozó éles konvergenciasebességeket vizsgáljuk:

2.3. 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') $\eta \in \mathcal{F}^{(p,M)}$.

Az $\{n^{-\frac{p}{2p+d}}\}$ sorozat a $\mathcal{D}^{*(p,M)}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség (Yang [1999]), és ez éles.

3.3. tétel. (Antos [1999]) *Legyen $\{b_n\}$ egy tetszőleges nullához tartó pozitív sorozat. Ekkor a*

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

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, ahol I_A az A esemény indikátorfüggvényét jelöli. Tehát ebben az esetben egy ismeretlen $C \in \mathcal{C}$ koncepciót kell megtanulni a

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

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* meghatározza.

A minimax várható hibavalószínűség viselkedését alapvetően a C koncepcióosztály V VC *dimenziója* (lásd pl. Devroye et al. [1996]) határozza meg: Az $\{V/n\}$ sorozat a $\mathcal{D}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség (Vapnik és Chervonenkis [1974]), és ez éles (Haussler, Littlestone és Warmuth [1994]).

Mint azt a dolgozatomban található példák mutatják, a fenti minimax alsó korlátok nem minden esetben terjeszthető ki individuális alsó korláttá, azonban sok fontos "geometriai" koncepcióosztályra 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 $[0,1] \times \{1,2,\ldots,k\}$ és

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

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

$$\mathcal{C}_k^2 = \left\{ \left\{ x \in \mathcal{R}^k : x_i \le a_i, i = 1,\ldots,k \right\} : a_1,\ldots,a_k \in \mathcal{R} \right\}$$

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

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

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

(Dolgozatomban további koncepcióosztályok is találhatóak, amelyekre szintén igazak az alábbi eredmények.)

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 $0 < \epsilon < 1$-re,

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

azaz $\{k/n\}$ éles individuális konvergenciasebesség $\mathcal{D}$-re nézve (Antos és Lugosi [1998], Schuurmans [1996]).

Legyen $\{g_n\}$ egy osztályozási szabály sorozat. A *kumulativ hiba*

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

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

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

kumulativ hibavalószínűség.

Ha $\mathcal{C}$ a $\mathcal{C}_k^1$, $\mathcal{C}_k^2$ és $\mathcal{C}_k^3$ egyike, akkor a várható kumulativ hibára a $\{k \log n/n\}$ sorozat a $\mathcal{D}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség, és ez éles (Haussler et al. [1994]).

A fenti minimax alsó korlátok következő individuális kiterjesztését bizonyítjuk:

3.10. tétel. (Antos és Lugosi [1998]) *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 $0 < \epsilon < 1$-re,*

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

azaz a várható kumulativ 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

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

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

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

indikátor.

A VC dimenzió a farokvalószínűségekre vonatkozó minimax alsó és felső korlátokat is jellemzi: Bármely koncepcióosztályra létezik $\{g_n\}$ osztályozási szabály sorozat, hogy

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

(Anthony et al. [1993], Lugosi [1995], lásd még Vapnik and Chervonenkis [1974], Blumer et al. [1989]), míg ha $V \geq 2$ és $\epsilon < 1/4$, akkor az

$$\left\{\left(\frac{n\epsilon}{V}\right)^{(V-1)/2} e^{-4n\epsilon/(1-4\epsilon)}\right\}$$

sorozat a $\mathcal{D}$ eloszlásosztályra nézve minimax alsó konvergenciasebesség (Devroye és Lugosi [1995]).

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 $\{\epsilon = \epsilon_n\}$ sorozatokra a fenti minimax korlátok következő tipusú individuális kiterjesztését bizonyítjuk, ha $\mathcal{C}$ a "geometriai" koncepcióosztályok egyike:

3.11. tétel. (Antos és Lugosi [1998]) *Legyenek $\epsilon_1, \epsilon_2, \ldots$ nemnegatív számok úgy, hogy $\{\gamma_n = n\epsilon_n\}$ 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,*

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

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

3.3 Funkcionálbecslés

Adottak a $D_n = \{X_1, \ldots, X_n\}$ független, X ismeretlen (diszkrét) eloszlása szerinti megfigyelések. Legyen P egy funkcionál, amely X minden lehetséges eloszlásához egy valós értéket rendel. Sokszor fontos, hogy P-t tudjuk pontosan becsülni, azaz találjunk egy olyan $P_n = P_n(X_1, \ldots, X_n)$ becslését, amelyre $|P_n - P|$ kicsi. Így a veszteségfüggvény a

$$l(P_n, P) \stackrel{\text{def}}{=} |P_n - P|$$

természetes távolság.

A várható érték, az entrópia, a kölcsönös információ és a Bayes-hiba becslés esetét vizsgáljuk.

3.3.1 Várható érték

Legyen X eloszlása $p(i)$ az $\mathcal{N}$-en és X várható értéke

$$P = m \stackrel{\text{def}}{=} \mathbf{E}X = \sum_{i=1}^{\infty} ip(i) \ .$$

Az m várható érték természetes becslése az átlag:

$$m_n = \frac{\sum_{j=1}^{n} X_j}{n} = \sum_{i=1}^{\infty} ip_n(i) \ ,$$

ahol

$$p_n(i) = \frac{1}{n} \sum_{j=1}^{n} I_{\{X_j = i\}}$$

az i előfordulásának relativ gyakorisága a mintában. Ez a nagy számok törvényei szerint erősen univerzálisan konzisztens, azaz

$$\lim_{n\to\infty} m_n = m \quad \text{1 valószínűséggel,}$$

és $m < \infty$-re L_1-ben is konzisztens, azaz

$$\lim_{n\to\infty} \mathbf{E}\{|m_n - m|\} = 0 \ .$$

4.6. tétel. (Antos [1998]) *Bármely $\{m_n\}$ becsléssorozatra és bármely $\{b_n\}$ pozitív, nullához tartó sorozatra létezik X-nek egy eloszlása $\mathcal{N}$-en úgy, hogy $m < \infty$ és*

$$\mathbf{E}\{|m_n - m|\} \geq b_n \quad \textit{végtelen sok } n\textit{-re.}$$

Ezt úgy is fogalmazhatjuk, hogy minden nullához tartó pozitív sorozat individuális alsó konvergenciasebesség az összes diszkrét eloszlás osztályára nézve.

Ha valamely $1 \leq p \leq 2$-re $\{\mathbf{E}|X|^p\} < \infty$, akkor míg az ilyen eloszlások osztályára nézve bármely $r > 1 - 1/p$-re $\{\frac{1}{n^r}\}$ individuális alsó konvergenciasebesség, az $m_n(D_n) = \sum_{i=1}^{n} X_i/n$ átlagra

$$\mathbf{E}\{|m_n - m|\} \leq (\mathbf{E}\{|m_n - m|^p\})^{1/p} \leq \frac{(2\mathbf{E}\{|X|^p\})^{1/p}}{n^{1-1/p}} = O\left(\frac{1}{n^{1-1/p}}\right)$$

(Bahr és Esseen [1965], see Theorem 2.6.20 in Petrov [1995]).

3.3.2 Entrópia

Legyen X eloszlása $p(i)$ az $\mathcal{N}$-en és X entrópiája

$$P = H \stackrel{\text{def}}{=} -\sum_{i=1}^{\infty} p(i) \log_2 p(i) \ .$$

A H "plug-in" becslése

$$H_n = -\sum_{i=1}^{\infty} p_n(i) \log_2 p_n(i) \ .$$

Ez erősen univerzálisan konzisztens, azaz

$$\lim_{n\to\infty} H_n = H \quad 1 \text{ valószínűséggel},$$

és $H < \infty$-re L_2-ben is konzisztens, azaz

$$\lim_{n\to\infty} \mathbf{E}\{(H_n - H)^2\} = 0$$

(Antos [1998]).

4.7. tétel. (Antos [1998]) *Bármely $\{H_n\}$ becsléssorozatra és bármely $\{b_n\}$ pozitív, nullához tartó sorozatra létezik X-nek egy eloszlása $\mathcal{N}$-en úgy, hogy $H < \infty$ és*

$$\mathbf{E}\{|H_n - H|\} \geq b_n \quad \textit{végtelen sok } n\textit{-re.}$$

Ezt úgy is fogalmazhatjuk, hogy minden nullához tartó pozitív sorozat individuális alsó konvergenciasebesség az összes diszkrét eloszlás osztályára nézve.

Ha valamely $0 < \alpha \leq 1$-re $c_1 1/i^{1+\alpha} \leq p_i \leq c_2 1/i^{1+\alpha}$, akkor a plug-in becslés L_2 hibájára

$$\mathbf{E}\{(H_n - H)^2\} = O\left(\frac{\log^2 n}{n^{2\alpha/(1+\alpha)}}\right)$$

(Antos [1998]).

3.3.3 Kölcsönös infomáció

Legyen $X = (V, W)$ eloszlása $p(i,j)$ az $\mathcal{N}^2$-en, marginális eloszlásai $p_V(i)$ és $p_W(j)$, és V és W kölcsönös infomációja

$$P = I \stackrel{\text{def}}{=} \sum_{i,j} p(i,j) \log_2 \frac{p(i,j)}{p_V(i) p_W(j)} \ .$$

Jelölje

$$p_n(i,j) = \frac{1}{n} \sum_{k=1}^{n} I_{\{V_k = i, W_k = j\}}$$

az (i,j) előfordulásának relativ gyakoriságát a mintában, $\{p_{V,n}(i)\}$ és $\{p_{W,n}(j)\}$ marginális eloszlásokkal.

Az entrópia becslésre vonatkozó konzisztenciából következik I-nek az

$$I_n = \sum_{i,j} p_n(i,j) \log_2 \frac{p_n(i,j)}{p_{V,n}(i) p_{W,n}(j)}$$

"plug-in" becslésére vonatkozó konzisztencia: Ha $H(V,W)$ véges, akkor I plug-in becslése erősen univerzálisan konzisztens és L_2-ben konzisztens, azaz

$$\lim_{n\to\infty} I_n = I \quad 1 \text{ valószínűséggel}$$

és

$$\lim_{n\to\infty} \mathbf{E}\{(I_n - I)^2\} = 0$$

(Antos [1998]).

4.8. tétel. (Antos [1998]) *Bármely $\{I_n\}$ becsléssorozatra és bármely $\{b_n\}$ pozitív, nullához tartó sorozatra létezik (V,W)-nek egy eloszlása $\mathcal{N} \times \mathcal{N}$-en úgy, hogy $I < \infty$ és*

$$\mathbf{E}\{|I_n - I|\} \geq b_n \quad \textit{végtelen sok } n\textit{-re.}$$

Ezt úgy is fogalmazhatjuk, hogy minden nullához tartó pozitív sorozat individuális alsó konvergenciasebesség a szimmetrikus $(p(i,j) = p(j,i))$ diszkrét eloszlások osztályára nézve.

Ha valamely $0 < \alpha \leq 1$-re a $c_1 1/i^{1+\alpha} \leq p_i \leq c_2 1/i^{1+\alpha}$ farokfeltétel egyaránt fennáll V, W és (V,W) (diszkrét) eloszlására, akkor a plug-in becslés L_2 hibájára

$$\mathbf{E}\{(I_n - I)^2\} = O\left(\frac{\log^2 n}{n^{2\alpha/(1+\alpha)}}\right)$$

(Antos [1998]).

3.3.4 Bayes-hiba

Legyen (X,Y) eloszlása $\nu(i,j)$ az $\mathcal{N} \times \{0,1\}$-en, X eloszlása $\mu(i)$, Y feltételes eloszlása adott X-re $\mathbf{P}\{Y = 1 | X = i\} = \eta(i)$, és a Bayes-hiba (lásd 3.2 szakasz)

$$P = L^* \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} \mu(i) \min(\eta(i), 1 - \eta(i)) = \sum_{i=1}^{\infty} \min(\nu(i,1), \nu(i,0)) \ .$$

L^* becslésének egyik módja valamely konzisztens g_n osztályozási szabály L_n hibavalószínűségének egy $\widehat{L}_n$ becslése. Ha az L_n-nek az $\widehat{L}_n$ becslése és a szabály egyaránt erősen konziszens, akkor $\widehat{L}_n$ erősen konzisztens becslése a Bayes-hibának is.

Például L^* “plug-in” becslése

$$\widehat{L}_n = \sum_{i=1}^{\infty} \min(\nu_n(i,1), \nu_n(i,0)) \ ,$$

ahol

$$\nu_n(i,j) = \frac{1}{n} \sum_{k=1}^{n} I_{\{X_k=i, Y_k=j\}} \ , \ j = 0,1$$

jelöli az (i,j) előfordulásának relativ gyakoriságát a mintában. Ez éppen egy hisztogram szabály hibavalószínűségének visszahelyettesítéses becslése. Így L^* plug-in becslése erősen univerzálisan konzisztens és L_p-ben konzisztens, azaz

$$\lim_{n\to\infty} \widehat{L}_n = L^* \quad \text{1 valószínűséggel}$$

és

$$\lim_{n\to\infty} \mathbf{E}\{|\widehat{L}_n - L^*|^p\} = 0$$

(Devroye, Györfi és Lugosi [1996]).

Ha $\widehat{L}_n$ az L_n becslése, akkor az alakfelismerésre vonatkozó lassú konvergenciasebesség eredmény miatt $\widehat{L}_n$-nek az L^*-hoz való konvergenciasebessége tetszőlegesen lassú lehet. Ez általában is igaz:

4.3. következmény. (Antos, Devroye és Györfi [1999]) *Bármely $\{\widehat{L}_n\}$ becsléssorozatra és bármely $\{b_n\}$ pozitív, nullához tartó sorozatra létezik (X,Y)-nak egy eloszlása $\mathcal{N} \times \{0,1\}$-en úgy, hogy*

$$\mathbf{E}\{|\widehat{L}_n - L^*|\} \geq b_n \quad \textit{végtelen sok } n\textit{-re.}$$

4 Az eredmények hasznosítása

A kutatás jellege alapvetően alapkutatás. Ezen belül olyan eredményeket tartalmaz, amelyek informálnak arról, hogy a nemparaméteres statisztika egyes területein milyen elvi korlátai vannak a becslés jóságának, és mi az, amit elvileg sem követelhetünk meg semmilyen becslési módszertől sem.

Ugyanakkor a regresszióbecslésre és az alakfelismerésre vonatkozó eredményekből az is következik, hogy a becslés egy adott pontosságának eléréséhez legalább mekkora minta szükséges. Ha tehát egy gyakorlati feladatban ennyi adat nem áll rendelkezésre, akkor nem érdemes a feladat megoldásához az adott modell korlátai között hozzákezdeni, hanem a modellt kell változtatni, a megfigyelés dimenzióját vagy a minták számát kell növelni.

Irodalomjegyzék

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

[1998] Antos, A. (1998). On functional estimates for discrete distributions. Submitted to *The 52nd Session of the International Statistical Intitute*, Helsinki, Finland.

[1999] Antos, A. (1999). Lower bounds on the rate of convergence of nonparametric pattern recognition. In *Computational Learning Theory: 4th European Conference, EuroCOLT'99, Proceedings*, Fischer, P. and Simon, H., editors, volume 1572 of *LNAI/LNCS*, pages 241–252. DFG,IFIP, Springer, Berlin. (IFIP WG 1.4: Student Author Award).

[1999] Antos, A., Devroye, L., and Györfi, L. (1999). Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* In print.

[1999] Antos, A., Györfi, L., and Kohler, M. (1999). Lower bounds on the rate of convergence of nonparametric regression estimates. *Journal of Statistical Planning and Inference.* In print. Preprint 98-11, Universität Stuttgart, Math. Inst. A, D-70511 Stuttgart, 1998.

[1998] Antos, A. and Lugosi, G. (1998). Strong minimax lower bounds for learning. *Machine Learning*, 30:31–56. Economics Working Paper 197, Universitat Pompeu Fabra, 1997.

[1965] Bahr, B. V. and Esseen, C. (1965). Inequalities for the rth absolute moment of a sum of independent random variables, $1 \le r \le 2$. *Ann. Math. Statist.*, 36:299–303.

[1995] Barron, A. R., Birgé, L., and Massart, P. (1995). Risk bounds for model selection via penalization. *Université Paris–Sud, Technical Report No.*, 95.54.

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

[1996] Devroye, L., Györfi, L., and Lugosi, G. (1996). *Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics, Stochastic Modelling and Applied Probability.* Springer-Verlag, New York.

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

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

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

[1995] Petrov, V. V. (1995). *Limit theorems of probability theory.* Clarendon Press, Oxford.

[1996] Schuurmans, D. (1996). *Effective classification learning.* PhD Thesis, University of Toronto, Toronto, CA.

[1977] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 8:1348–1360.

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

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

[1999] Yang, Y. (1999). Minimax nonparametric classification — part I: Rates of convergence. *IEEE Transactions on Information Theory.* Accepted.