ÚJRAMINTAVÉTELEZÉSI ELJÁRÁSOK

A jackknife (zsebkés) és bootstrap (cipőhúzó a saját kallantyújánál fogva) eljárások angol elnevezése is arra utal, hogy itt ad hoc eljárásokról van szó, melyek azonban nagyon hasznosak olyan szorult helyzetekben, mikor nincs kellően nagy mintánk vagy nem tudjuk tesztadatokon megismételni az eljárást. Modellépítésnél szokásos a tanuló- ést tesztadat elnevezés, mikor a tanuló adatokon becsüljük a modell paramétereit, a tesztdatokon pedig szeretnénk kipróbálni azokat. Az esetek többségében azonban egyetlen adatrendszer áll csak rendelkezésünkre.

Célunk valamely becslés pontosságának javítása, a hiba eloszlásának becslése elsősorban kismintás esetekben vagy olyanokban, melyeknél a tanulóalgoritmust tesztadatokon szeretnénk kipróbálni.

Tesztadatok hiányában a tanulóadatokat replikáljuk (ismételjük). Az újramintavételezett minta alkalmas konfidenciaintervallumok szerkesztésére és hipotézisvizsgálatra is, különösen nem-paraméteres esetben.

Jackknife

A jackknife az adatok jól megválasztott csoportosításán alapszik, a csoportok kombinációi alapján becsléseket konstruálunk, amelyek átlaga lesz a jackknife becslés. Itt csak egy speciális csoportokat használó eljárást ismertetünk.

Legyen $\mathbf{X} = (X_1, \dots, X_n)$ független azonos eloszlású minta egy $\mathbb{P}_\theta$ eloszlásból, ahol $\theta \in \Theta$ ismeretlen paraméter. Jelölje $\hat{\theta} := \hat{\theta}(\mathbf{X})$ a θ paraméter valamilyen becslését a teljes minta alapján; a továbbiakban a becslések argumentumába nem írjuk be a mintaelemeket. Jelölje $\hat{\theta}_{-i}$ $(i = 1, \dots, n)$ azt a becslést, amelyet az i-edik mintaelem elhagyásával kapunk. Képezzük az ún. *pszeudoértékeket* (az elnevezés Tukey-től származik):

$$\widetilde{\theta}_i := n\hat{\theta} - (n-1)\hat{\theta}_{-i}$$

Definíció. A θ paraméter *jackknife becslése* a $\widetilde{\theta}_i$ pszeudoértékek átlaga:

$$\widetilde{\theta}_\bullet = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\theta}_i.$$

Állítás. A jacknife becslés pontosan eliminálja a torzítás $\frac{1}{n}$ rendű tagját.

Bizonyítás. Ha $\mathbb{E}(\hat{\theta}) = \theta + \frac{a}{n} + \frac{b}{n^2} + \dots$, akkor

$$\mathbb{E}(\widetilde{\theta}_\bullet) = n(\theta + \frac{a}{n} + \frac{b}{n^2} + \dots) - (n-1)(\theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + \dots) = \theta - \frac{b}{n(n-1)} + \dots$$

Ha a $\widetilde{\theta}_i$ pszeudoértékek közelítőleg függetlenek (elég nagy n-re ez általában teljesül), akkor $\mathbb{D}^2(\widetilde{\theta}_\bullet)$ becslése az

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\widetilde{\theta}_i - \widetilde{\theta}_\bullet)^2$$

statisztika lehet ($\widetilde{\theta}_i$-ok korrigált empirikus szórásnégyzetét le kell osztani n-el, ha átlaguk szórásnégyzetét akarjuk megkapni), továbbá a

$$t = (\widetilde{\theta}_\bullet - \theta) \left[\frac{1}{n(n-1)} \sum_{i=1}^{n} (\widetilde{\theta}_i - \widetilde{\theta}_\bullet)^2 \right]^{-1/2}$$

statisztika közelítőleg $t(n-1)$ eloszlású (ha n elég nagy), így alkalmas hipotézisvizsgálatra és konfidenciaintervallum szerkesztésre.

- *1. Példa*: Legyen $X_1, \ldots, X_n$ fae. minta egy olyan eloszlásból, melynek első momentuma létezik. Könnyű látni, hogy a $\theta = \mathbb{E}(X_1)$ paraméter jackknife becslése $\bar{X}$. Hiszen a $\hat{\theta} = \bar{X}$ torzítatlan becslésből kiindulva a jackknife becslés megtartja a torzítatlanságot.

- *2. Példa*: Legyen $X_1, \ldots, X_n$ fae. minta egy olyan eloszlásból, melynek létezik a második momentuma. A $\theta = \sigma^2$ paraméter kezdeti becslése legyen $\hat{\theta} = S_n^2$, ennek torzítása $\frac{1}{n}$ rendű. Az Állítás értelmében a jakknife becslés eliminálja ezt a torzítást, és a $\widetilde{\theta}_\bullet = S_n^{*2}$ torzítatlan becslés lesz a szórásnégyzet jacknife becslése. A Steiner-formula többszöri alkalmazásával belátható, hogy a $\widetilde{\theta}_i$ pszeudoértékek közel függetlenek, így a

  $$t = (\widetilde{\theta}_\bullet - \sigma^2) \left[\frac{1}{n(n-1)} \sum_{i=1}^{n} (\widetilde{\theta}_i - \widetilde{\theta}_\bullet)^2 \right]^{-1/2}$$

  statisztika "nagy" n-re "közel" $t(n-1)$ eloszlású, melynek segítségével konfidenciaintervallumot szerkeszthetünk és hipotéziseket vizsgálhatunk az eloszlás ismeretlen szórásnégyzetére.

Bootstrap

A [3] könyv jelöléseit használva legyen $\mathbf{Z} = (Z_1, \ldots, Z_n)$ fae. minta, $S(\mathbf{Z})$ pedig ennek valamely függvénye (nem feltétlenül statisztika, mert az ismeretlen paramétert is tartalmazhatja). Pl. a $\psi(\theta)$ paraméterfüggvény $T(\mathbf{Z})$ statisztikával történő becslésekor $S(\mathbf{Z}) := T(\mathbf{Z}) - \psi(\theta)$ a becslés hibája, $\mathbb{E}_\theta S^2(\mathbf{Z})$ pedig a becslés négyzetes rizikója.

A $\mathbf{Z} = (Z_1, \ldots, Z_n)$ minta replikáltja a $\mathbf{Z}^* = (Z_1^*, \ldots, Z_n^*)$ ún. bootstrap minta, amelyet a $z_1, \ldots, z_n$ realizáltakból való visszatevéses mintavétellel nyerünk (megjegyezzük, hogy a jacknife módszer alkalmazásakor valójában $n-1$ elemet replikáltunk visszatevés nélkül). Pl. $n = 6$ esetén a $z_1, \ldots, z_6$ számokat egy szabályos dobókocka oldalaira írva, 6 dobás kimenetele egy bootstrap mintát szolgáltat. Tetszőleges n esetén a $z_1, \ldots, z_n$ számokat n cédulára felírva és egy kapalba téve, n-szer húzunk egymás után visszatevéssel.

$$\mathbb{P}(Z_i \in \mathbf{Z}^*) = 1 - \frac{(n-1)^n}{n^n} = 1 - (1 - \frac{1}{n})^n \to 1 - e^{-1} \sim 0.632, \quad \text{ha} \quad n \to \infty$$

annak a valószínűsége, hogy egy mintaelem belekerül a bootstrap mintába.

A bootstrap minta alapján is kiszámoljuk $S(\mathbf{Z}^*)$ értékét, sőt ezt általában több bootstrap mintára is megtesszük: $S(\mathbf{Z}^{*1}), \ldots, S(\mathbf{Z}^{*B})$. Ha B elég nagy, akkor a kapott B szám $S(\mathbf{Z})$ ún. bootstrap eloszlását mutatja, aminek alapján kvantilis értékeket és egyéb paramétereket becsülhetünk. Ennek akkor van jelentősége, ha $S(\mathbf{Z})$ elméleti eloszlását egyébként nem ismerjük, pl. nem-paraméteres esetben. $S(\mathbf{Z})$ szórásnégyzetének becslése a bootstrap eloszlás alapján:

$$\widehat{\mathbb{D}^2}(S(\mathbf{Z})) = \frac{1}{B-1}\sum_{b=1}^{B}(S(\mathbf{Z}^{*b}) - \bar{S}^*)^2,$$

ahol $\bar{S}^* = \frac{1}{B}\sum_{b=1}^{B} S(\mathbf{Z}^{*b})$.

A bootstrap eloszlás nem más, mint a $z_1, \ldots, z_n$ realizáció alapján szerkesztett empirikus eloszlás (az empirikus eloszlásfüggvény lépcsős függvény, mely a $z_1, \ldots, z_n$ helyeken ugrik $1/n$-el). Ez valójában a $z_1, \ldots, z_n$ pontokra koncentrált diszkrét egyenletes eloszlás: $\mathcal{U}(z_1, \ldots, z_n)$. Ennek az eloszlásnak a valódi momentumai (centrális momentumai) az eredeti eloszlás empirikus (centrális empirikus) momentumaival egyeznek meg. Ennek alapján a momentumoktól függő becslések megkonstruálhatók.

- *1. Példa*: Legyen $\mathbf{Z} = (Z_1, \ldots, Z_n)$ fae. minta egy olyan eloszlásból, melynek létezik a második momentuma. Legyen a $\theta = \sigma^2$ paraméter a szórásnégyzet, ennek torzítatlan becslése:

  $$T(\mathbf{Z}) = S_n^{*2} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}{n-1}.$$

  $S(\mathbf{Z}) := T(\mathbf{Z}) - \theta$, $\mathbb{E}S(\mathbf{Z}) = 0$. Legyen $\mathbf{Z}^*$ bootstrap minta. A bootstrap eloszlás szórásnégyzete S_n^2, $T(\mathbf{Z}^*)$ pedig a bootstrap minta alapján számolt korrigált empirikus szórásnégyzet, ami itt is torzítatlanul becsli az S_n^2 valódi szórásnégyzetet. Így a bootstrap eloszlás alapján $\widehat{\mathbb{E}}(S(\mathbf{Z})) = 0$. Most számítsuk ki a bootstrap eloszlás alapján $T(\mathbf{Z})$, vagy ami ugyanaz, $S(\mathbf{Z})$ szórásnégyzetét! Mivel régebben beláttuk, hogy amennyiben az eredeti eloszlás negyedik momentuma (m_4), s így negyedik centrális momentuma (m_4^c) is létezik, akkor

  $$\mathbb{D}^2(S(\mathbf{Z})) = \mathbb{D}^2(T(\mathbf{Z})) = \mathbb{D}^2(S_n^{*2}) = \frac{1}{n}\left(m_4^c - \frac{n-3}{n-1}\sigma^4\right).$$

  Mivel a bootstrap eloszlás valódi momentumai az eredeti eloszlás empirikus momentuaival egyeznek meg, ennek alapján

  $$\widehat{\mathbb{D}^2}(S(\mathbf{Z})) = \mathbb{D}^2(T(\mathbf{Z}^*)) = \frac{1}{n}\left(\widehat{M}_4^c - \frac{n-3}{n-1}(\widehat{M}_2^c)^2\right) = \frac{1}{n}\left(\widehat{M}_4^c - \frac{n-3}{n-1}S_n^2\right),$$

  ahol M jelöli az eredeti minta alapján kapott empirikus momentumokat. Tehát itt a bootstrap minta vétele nélkül, csupán az eredeti minta alapján számolt empirikus momentumokra támaszkodva elméletileg meg tudjuk határozni $S(\mathbf{Z})$ bootstrap eloszlás szerinti szórásnégyzetét.

- *2. Példa*: regressziós hiba becslése. Legyen $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$ fae. 2-dimenziós minta. Az $Y = f(X)+b$ regressziós modell (spec. $Y = aX+c+\varepsilon$ lineáris regressziós modell)

paramétereinek legkisebb négyzetes becslését jelölje $\hat{a}$ és $\hat{b}$. Az ε hiba szórásnégyzetének bootstrap becslése az x_i^*, y_i^* $(i = 1, \ldots, n)$ bootstrap realizáltakból (melyek azonos indexhez tartozó (x, y)-párok az eredeti mintából) történő $\hat{f}^*$ becslések (lineáris esetben az $\hat{a}^*$, $\hat{c}^*$ becslések) alapján a B db. bootstrap mintából:

$$\frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{*b}(x_i))^2 = \frac{1}{Bn}\sum_{b=1}^{B}\sum_{i=1}^{n}(y_i - \hat{a}^{*b}(x_i) - \hat{c}^{*b})^2,$$

ahol a második becslés a lineáris esetre vonatkozik. Itt tehát a generált tesztadatokból becsültük a regressziós modell paramétereit, a hibát viszont az eredeti mintapontok helyén számoltuk.

- *3. Példa*: a diszkriminanciaanalízis hibabecslése. Az egyszerűség kedvéért tegyük fel, hogy csak két mintánk van:

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim F = \mathcal{N}_k(\mathbf{m}_1, \mathbf{C})$$

és

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_m \sim G = \mathcal{N}_k(\mathbf{m}_2, \mathbf{C}),$$

ahol az $\mathbf{X}_i$ és $\mathbf{Y}_j$ k-dimenziós véletlen vektorok teljesen függetlenek. A megfigyelt értékek: $\mathbf{x}_1, \ldots \mathbf{x}_n$, illetve $\mathbf{y}_1, \ldots, \mathbf{y}_m$. A minta alapján megbecsüljük az $\mathbf{m}_1$ és $\mathbf{m}_2$ várhatóérték vektort, valamint a $\mathbf{C}$ kovariancaiamátrixot, legyenek a becslések: $\widehat{\mathbf{m}}_1$, $\widehat{\mathbf{m}}_2$ és $\widehat{\mathbf{C}}$. Ezeket a becsléseket figyelembe véve [1] Diszkriminanciaanalízis fejezete alapján eljárást kapunk arra, hogy eldöntsük: egy új $\mathbf{x}$ megfigyelést az F vagy a G eloszlást követi-e. Legyen

$$A := \{\mathbf{x} : (\widehat{\mathbf{m}}_2^T - \widehat{\mathbf{m}}_1^T)\widehat{\mathbf{C}}^{-1}\mathbf{x} > c\} \subset \mathbb{R}^k$$

ahol c a paraméterektől és mintaelemszámoktól függő konstans. Ha $\mathbf{x} \in A$ teljesül, akkor az $\mathbf{x}$ megfigyelést a G eloszlást követők csoportjába soroljuk. Az osztályozás várható hibáját még az új megfigyelések (tesztadatok) beérkezése előtt szeretnénk megbecsülni. Az

$$\widehat{\text{error}} := \frac{|\{i : \mathbf{x}_i \in A\}|}{n}$$

mennyiség nyilván alulbecsüli a hibát, mert az osztályozó eljárást a minta alapján szerkesztettük, az mintegy adaptálódott a mintához. A valódi várható hiba

$$\text{error} := \mathbb{P}_F\{i : \mathbf{x}_i \in A\}$$

lenne. Mindezt megcsinálhatnánk a másik fajta hibára is, mely a G-eloszlásúak ($\mathbf{y}_j$-k) F-eloszlású csoportba való téves besorolásából adódna (az $\overline{A}$ tartomány alapján). Az osztálozás hibája e kétfajta hiba összege. A téves besorolás relatív gyakorisága így csak az $\mathbf{X}$-es, ill. $\mathbf{Y}$-os mintákból számolandó, maga az A illetve $\overline{A}$ taromány azonban függ mindkét mintától ($\mathbf{Z}$).

$$S(\mathbf{Z}) := \text{error} - \widehat{\text{error}},$$

a másik fajta hibára vonatkozó eltérés teljesen hasonlóan számolható.

Az $S(\mathbf{Z}^*)$ bootstrap veszteséget Monte Carlo módszerrel határozhatjuk meg. Az $\hat{F}$ és $\hat{G}$ eloszlásból generálunk n, illetve m darab $\mathbf{x}_i^*$, illetve $\mathbf{y}_j^*$ bootstrap mintaelemet, ezek alapján kiszámítjuk az $\hat{F}$ és $\hat{G}$ eloszlások paramétereit, majd meghatározzuk az A^* bootstrap kritikus tartományt. Így az bootstrap veszteség egy realizációja:

$$S(\mathbf{Z}^*) = \frac{|\{i : \mathbf{x}_i \in A^*\}|}{n} - \frac{|\{i : \mathbf{x}_i^* \in A^*\}|}{n},$$

hasonlóan y_j-k, illetve y_j^*-ok alapján a másik fajta hibára.

Ezen eljárás elegendően sok független ismétlése után a veszteség eloszlása, momentumai meghatározhatók.

Megjegyezzük, hogy a programcsomagok kiszámítják a hibavalószínűség jackknife becslését is olymódon, hogy minden egyes mintaelem kihagyásával megszerkesztik a kritikus A illetve A' tartományt, majd megvizsgálják, hogy a kihagyott elem melyik tartományhoz tartozik. Az így tapasztalt hibás döntések relatív gyakorisága a hibavalószínűség becslése. Efron [2] dolgozatában egy 10 és egy 20 elemű mintára ismerteti mindkét eljárás eredményét; nincs lényeges különbség.

Megjegyezzük, hogy a bootstrap módszer nem-paraméteres próbákra is alkalmazható. Annak a nul-hipotézisnek a tesztelésére, hogy mintánk normális eloszlásból származik (a tagadás alternatívájával szemben) először a rendelkezésre álló fae. mintából becsüljük meg a szóbajöhető normális eloszlás paramétereit (mintaátlag és empirikus szórásnégyzet). Ezután generáljunk egy másik (nem feltétlenül az eredetivel azonos elemszámú) mintát a becsült paraméterekkel (eredeti mintánkhoz gyártott bootstrap minta H_0 fennállása esetén). Végül végezzünk illeszkedésvizsgálatot arra nézve, hogy a két minta azonos eloszlásból származik-e (pl. Kolmogorov–Szmirnov próbával). Ha (az adott szinten) igen, akkor elfogadjuk, különben elutasítjuk a null-hipotézist.

Irodalom

[1] Bolla, M., Krámli, A., Statisztikai következtetések elmélete, Typotex, Budapest (2005).

[2] Efron, B., Bootstrap methods: another look at the jackknife, Ann. Statist. 7 (1979), 1-26.

[3] Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, New York (2001).