Tartalomjegyzék

visszavezetés bináris döntésre

Jeloles:

	LinLug	mine
fuggo valtozo	A	Y
dontes	G	g
|\Y|	s,M	k
Y a priori suly	q_i	
Y|x a post.suly	P_i(x)	\eta_i(x)
Y|x a post.eo.fv		F_{Y|x}
X sfv.	f(x)	
X|Y=i felt.sfv.	f_i(x)	
X|Y=i felt.suly	p_i(x)	
	Q_i	d_i
	B_\Delta	S_\epsilon(x)
"ido" index	k	t
E{X_i X_j}	k_{ij}	s_{ij}
	K	S
E{X_j Y}	m_j	b_j
	m	b

1. fejezet

Valószínűségi becslés- és döntéselmélet

1.1. Bevezetés

Ez a fejezet alapvetően a [4] jegyzet 5. fejezetén alapszik. Továbbá az 1.2, 1.3, 1.4 és 1.5 fejezetekhez hasznos lehet a [2] könyv 1. és 2. fejezete, az 1.2, 1.6, 1.7 és 1.8 fejezetekhez pedig a [3] könyv 1. fejezete.

1.2. Definíciók

Gyakori probléma, hogy egy X megfigyelhető mennyiségből kell egy másik, közvetlenül (még) nem megfigyelhető Y mennyiségre következtetnünk, annak értéket megbecsülnünk. Általános esetben matematikailag mindkét mennyiséget valószínűségi változókkal modellezhetjük. Jelölje X illetve Y értékkészletét (azaz lehetséges értékeinek halmazát) $\mathcal{X}$ illetve $\mathcal{Y}$. Ekkor formálisan $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ függvények (ahol Ω a valószínűségi mező alaphalmaza).

1. példa. $\mathcal{X}$ és $\mathcal{Y}$ tipikusan lehet például $\mathbf{R}$ (a valós számok halmaza), $\mathbf{R}^d$, $\{0,1\}^d$, $[0,1]^d$ vagy ezek tetszőleges megszámlálható, esetleg véges részhalmaza. Legegyszerűbb esetben $\mathcal{Y} = \{0,1\}$.

X értékéből Y-t a $g : \mathcal{X} \to \bar{\mathcal{Y}}$ *következtetésfüggvénnyel* próbáljuk meghatározni, ahol $\bar{\mathcal{Y}}$ az elképzelhető következtetésfüggvények értékkészletének uniója. $g(X)$-t *következtetésnek* nevezzük. Tipikusan $\bar{\mathcal{Y}} \supseteq \mathcal{Y}$. Egy következtetés jóságát egy nemnegatív $C : \mathcal{Y} \times \bar{\mathcal{Y}} \to [0, \infty)$ *költségfüggvény* (vagy *jósági*, *hasonlósági*, *megbízhatósági kritérium*) méri, azaz $C(y, y')$ a költsége annak ha a valódi y érték helyett y'-re következtetünk. Pl. ha a megfigyelés x volt és $Y = y$, akkor a g által adott következtetés költsége $C(y, g(x))$. Minél kisebb ez a költség, a következtetés annál jobbnak tekinthető.

Ha minden $y \in \mathcal{Y}$-ra $C(y, y')$ konstans minden $y' \in \bar{\mathcal{Y}} \setminus \{y\}$-ra, azaz pontatlan következtetés esetén mindig azonos, akkor *döntési problémáról* beszélünk. Ilyenkor el akarjuk találni y-t. Ha nem találjuk el, mindegy, hogy „mennyire” nem. Ha C valamiféle intuitív

távolság, akkor *becslési problémáról* beszélünk. Ilyenkor az is számít, hogy mennyit tévedünk. Döntési problémáknál általában $\bar{\mathcal{Y}} = \mathcal{Y}$ véges vagy megszámlálható (diszkrét), míg becslési problémáknál általában $\bar{\mathcal{Y}}$ vagy akár $\mathcal{Y}$ is folytonos (megszámlálhatatlan végtelen).

Mivel $C(Y, g(X))$ függ (X, Y)-tól, maga is valószínűségi változó. Így g jóságát a várható költsége, az

$$R(g) \stackrel{\text{def}}{=} \mathbf{E}\left[C(Y, g(X))\right]$$

globális kockázat (risk) méri. Célunk annak a következtetésfüggvénynek a megtalálása, amelyre $R(g)$ a legkisebb. Ezt *Bayes-feladatnak*, míg az ilyen g-t *optimálisnak* nevezzük.

Legyen

$$r(g, x) \stackrel{\text{def}}{=} \mathbf{E}\left[C(Y, g(X))|X = x\right]$$

a *lokális kockázat függvény*, azaz g költségének feltételes várhatóértéke $X = x$ esetén.

1. gyakorlat. Lássuk be, hogy $g(X)$ helyett írhatunk $g(x)$-et is.

Nyilván a teljes várhatóérték tétele és feltételes várhatóérték definíciója szerint

$$\mathbf{E}\left[r(g, X)\right] = R(g) \tag{1.1}$$

és

$$r(g, x) = \mathbf{E}\left[C(Y, g(x))|X = x\right] = \int_{\mathcal{Y}} C(y, g(x))\, dF_{Y|x}(y) \tag{1.2}$$

ahol $F_{Y|x}$ az Y feltételes eloszlásfüggvénye ha $X = x$, amelyet *a posteriori eloszlásfüggvénynek* is nevezünk. Ez tehát általános esetben ún. (Lebesgue-)Stieltjes integrállal írható fel, de a továbbiakban általunk vizsgált diszkrét illetve abszolút folytonos eloszlásokra egyszerű összegzésre illetve Riemann-integrálra vezet. Az a posteriori eloszlás elnevezés arra utal, hogy ez Y eloszlása X értékének ismerete *után*. Ettől megkülönböztetendő Y feltétel nélküli, eredeti (azaz X értékének ismerete *előtti*) eloszlását az Y *a priori eloszlásának* is nevezzük.

1.2.1. Gyakori költségfüggvények és tulajdonságaik

Döntési problémáknál a leggyakoribb költségfüggvény választás a következő:

2. példa. *0-1* költség: Legyen $C_0(y, y') \stackrel{\text{def}}{=} \mathbf{I}_{\{y \neq y'\}}$ (ahol $\mathbf{I}_A$ az A esemény indikátorfüggvénye, azaz $\mathbf{I}_A = 1$, ha A bekövetkezik és $\mathbf{I}_A = 0$, ha nem). Ekkor

$$R(g) = \mathbf{E}\left[C_0(Y, g(X))\right] = \mathbf{E}\left[\mathbf{I}_{\{Y \neq g(X)\}}\right] = \mathbf{P}\left[Y \neq g(X)\right],$$

vagyis a globális kockázat éppen g hibázásának a valószínűsége.

Becslési problémáknál $\bar{\mathcal{Y}} = \mathcal{Y} = \mathbf{R}^d$-re pedig gyakori választás az L_p távolság (más jelöléssel a $\|\cdot\|_p$ norma) p-edik hatványa, azaz $C_p(y, y') \stackrel{\text{def}}{=} \sum_{s=1}^{d} |y_s - y'_s|^p = \|y - y'\|_p^p$. Ekkor

$$R(g) = \mathbf{E}\left[C_p(Y, g(X))\right] = \mathbf{E}\left[\|Y - g(X)\|_p^p\right].$$

Leggyakrabb a $p = 1$ és 2 eset (ezért $\|\cdot\|_1$ helyett $\|\cdot\|$-t is használunk):

3. példa. $C_1(y,y') = \|y-y'\|$ az *abszolút* költség. Ekkor pl. $\bar{\mathcal{Y}} = \mathbf{R}$-re $R(g) = \mathbf{E}\left[|Y-g(X)|\right]$ az *abszolút középhiba*.

4. példa. $C_2(y,y') = \|y-y'\|_2^2$ a *négyzetes* költség. Ekkor pl. $\bar{\mathcal{Y}} = \mathbf{R}$-re $R(g) = \mathbf{E}\left[(Y-g(X))^2\right]$ a *négyzetes középhiba*.

Érdemes először megvizsgálnunk, hogy a fenti példák – és az egyszerűség kedvéért 2.-ben diszkrét $\mathcal{Y}$, 3., 4.-ben $\bar{\mathcal{Y}} = \mathbf{R}$ – esetén milyen $g \in \bar{\mathcal{Y}}$ érték minimalizálja a $\mathbf{E}\left[C(Y,g)\right]$ várható költséget, azaz rendre

$$\begin{aligned}
\mathbf{E}\left[C_0(Y,g)\right] &= \mathbf{P}\left[Y \neq g\right], \\
\mathbf{E}\left[C_1(Y,g)\right] &= \mathbf{E}\left[|Y-g|\right], \\
\mathbf{E}\left[C_2(Y,g)\right] &= \mathbf{E}\left[(Y-g)^2\right]\text{-t}.
\end{aligned}$$

$\mathbf{P}\left[Y \neq g\right]$ minimalizálása nyilván $\mathbf{P}\left[Y = g\right]$ maximalizálását jelenti, vagyis g-t az

$$\arg\max_{y \in \mathcal{Y}} \mathbf{P}\left[Y = y\right]$$

halmaz egyik elemének, azaz Y (nem mindig egyértelmű) *móduszának* kell választani.

2. gyakorlat. Lássuk be, hogy $\mathbf{E}\left[|Y-g|\right]$-t Y (nem mindig egyértelmű) *mediánja* minimalizálja, azaz azon g értékek, amelyekre $\mathbf{P}\left[Y \leq g\right] \geq 1/2$ és $\mathbf{P}\left[Y \geq g\right] \geq 1/2$. (Pl. abszolút folytonos Y-ra ez $\mathbf{P}\left[Y \leq g\right] = \mathbf{P}\left[Y \geq g\right] = 1/2$-ként írható.)

A négyzetes középhiba minimalizálásában segít a következő

1. tétel (Steiner-tétel)**.** *Bármely véges szórású Y valós valószínűségi változóra és $g \in \mathbf{R}$-re*

$$\mathbf{E}\left[(Y-g)^2\right] = \mathbf{E}\left[(Y-\mathbf{E}\left[Y\right])^2\right] + (\mathbf{E}\left[Y\right]-g)^2.$$

Bizonyítás.

$$\begin{aligned}
\mathbf{E}\left[(Y-g)^2\right] &= \mathbf{E}\left[((Y-\mathbf{E}\left[Y\right]) + (\mathbf{E}\left[Y\right]-g))^2\right] = \\
&= \mathbf{E}\left[(Y-\mathbf{E}\left[Y\right])^2\right] + 2\mathbf{E}\left[(Y-\mathbf{E}\left[Y\right])(\mathbf{E}\left[Y\right]-g)\right] + \mathbf{E}\left[(\mathbf{E}\left[Y\right]-g)^2\right] = \\
&= \mathbf{E}\left[(Y-\mathbf{E}\left[Y\right])^2\right] + (\mathbf{E}\left[Y\right]-g)^2,
\end{aligned}$$

mivel a középső tagban $(\mathbf{E}\left[Y\right]-g)\mathbf{E}\left[Y-\mathbf{E}\left[Y\right]\right] = 0$. □

1. *megjegyzés.* A tétel ($\mathbf{R}^2$-re vonatkozó verziójának) fizikai analógiája az, hogy egy test tehetetlenségi nyomatéka egy tetszőleges tengely körül nem más, mint a tehetetlenségi nyomatéka a tömegközépponton átmenő, párhuzamos tengely körül plusz a test tömegeszer a két tengely merőleges távolságának négyzete (párhuzamos tengely tétel). Innen származik a Steiner-tétel elnevezés.

A 1. tétel következménye, hogy

$$\min_{g\in\mathbf{R}} \mathbf{E}\left[(Y-g)^2\right] = \mathbf{E}\left[(Y-\mathbf{E}\left[Y\right])^2\right],$$

azaz $\mathbf{E}\left[(Y-g)^2\right]$-t a $g=\mathbf{E}\left[Y\right]$ választás minimalizálja.

Összefoglalva, a minimalizáló értékek rendre a következők:

$$\begin{aligned} &C_0 \text{ :módusz } \ (\arg\max_{y\in\mathcal{Y}} \mathbf{P}\left[Y=y\right]),\\ &C_1 \text{ :medián},\\ &C_2 \text{ :várhatóérték (mean) } \ (\mathbf{E}\left[Y\right]). \end{aligned}$$

Látni fogjuk, hogy amikor $g=g(X)$-t az X-től függően választjuk, akkor ezen megfigyelések feltételes verziója lép érvénybe, azaz a minimalizáló értékek helyébe az Y feltéve X feltételes eloszlás módusa, mediánja, illetve várhatóértéke lép. A bizonyítások a fentiekkel analóg módon történnek.

Az L_p távolságon alapuló C_p költségfüggvények közötti további fontos összefüggés a következő:

2. tétel. *$\bar{\mathcal{Y}}=\mathcal{Y}=\mathbf{R}^d$ esetén mind $(C_p(y,y')/d)^{1/p} = \|y-y'\|_p d^{-1/p}$, mind $\mathbf{E}\left[C_p(Y,g(X))/d\right]^{1/p}$ monoton növekvő p-ben. Speciálisan $\bar{\mathcal{Y}}=\mathbf{R}$-re, $C_p^{1/p}(y,y')$ és $\mathbf{E}\left[C_p(Y,g(X))\right]^{1/p}$ monoton növekvő p-ben.*

Bizonyítás. Bármely $0<p<q$-ra

$$\begin{aligned} \left(\frac{C_p(y,y')}{d}\right)^{1/p} &= \left(\sum_{s=1}^{d}\frac{1}{d}|y_s-y'_s|^p\right)^{1/p} = \left(\left(\sum_{s=1}^{d}\frac{1}{d}|y_s-y'_s|^p\right)^{q/p}\right)^{1/q} \leq \\ &\leq \left(\sum_{s=1}^{d}\frac{1}{d}(|y_s-y'_s|^p)^{q/p}\right)^{1/q} = \left(\sum_{s=1}^{d}\frac{1}{d}|y_s-y'_s|^q\right)^{1/q} = \left(\frac{C_q(y,y')}{d}\right)^{1/q}, \end{aligned}$$

ahol az egyenlőtlenség a Jensen-egyenlőtlenség alkalmazása egy egyenletes d értékű eloszlásra, hiszen $q/p>1$ és így $x^{q/p}$ konvex függvény. A bizonyítás az (X,Y) szerinti várhatóértékkel együtt is pont ugyanígy történik a Jensen-egyenlőtlenség kétszeri alkalmazásával. □

Sokszor az $y\in\mathbf{R}^d$ lehetséges értékei speciálisan eloszlások súlyvektorai. Ilyenkor gyakran hasznos költségfüggvény választás az ún. *Kullback–Leibler (KL) távolság* (vagy *relatív entrópia*, *I-divergencia*) [1, 2.3. fejezet]:

5. példa. Legyen $C_{\mathrm{KL}}(y,y') \stackrel{\mathrm{def}}{=} D_{\mathrm{KL}}(y\|y') = \sum_{s=1}^{d} y_s \ln\frac{y_s}{y'_s}$.

Ismert, hogy a KL távolság dominálja az L_1 távolság négyzetét [1, p. 300, Lemma 12.6.1]:

3. tétel (Pinsker-egyenlőtlenség)**.** $2C_{\mathrm{KL}}(y,y') \geq \|y-y'\|^2 = C_1^2(y,y')$.

1.3. Bayes-döntés

Vizsgáljuk meg először alaposabban azt az esetet, amikor $\mathcal{Y} = \bar{\mathcal{Y}}$ az $\{y_1, \ldots, y_k\}$ véges halmaz. Ekkor

1. definíció. Az $\{Y = y_i\}$ eseményt *i-edik hipotézisnek* nevezzük. Y (a priori) eloszlását a $q_i \stackrel{\text{def}}{=} \mathbf{P}[Y = y_i]$ *a priori valószínűségek*, míg az a posteriori eloszlását az $\eta_i(x) \stackrel{\text{def}}{=} \mathbf{P}[Y = y_i | X = x]$ *a posteriori valószínűségek* adják meg.

g-t *döntésfüggvénynek*, $g(X)$-t *döntésnek* is nevezzük. Az y_i értékek g-vel való ősképei $\mathcal{X}$ egy partícióját adják, amelynek $D_i = \{x \in \mathcal{X} : g(x) = y_i\}$ osztályait $(i = 1, 2, \ldots, k)$ *döntési tartományoknak* nevezzük. D_i tehát $\mathcal{X}$ azon maximális részhalmaza, amely bármely elemének megfigyelésekor y_i-re dönt g.

3. gyakorlat. Mutassuk meg, hogy adott $\bar{\mathcal{Y}} = \{y_1, \ldots, y_k\}$-ra, a $(D_1, \ldots, D_k)$ döntési tartományok teljesen meghatározzák a g döntésfüggvényt, azaz a döntésfüggvényt megadhatjuk a döntési tartományok rendezett k-asával, továbbá hogy $\mathbf{I}_{\{g(x)=y_j\}} = \mathbf{I}_{\{x \in D_j\}}$.

Most — bevezetve a $C(y_i, y_j) = C_{ij}$ rövidítést — (1.2) szerint a lokális kockázat

$$r(g, x) = \sum_{i=1}^{k} C(y_i, g(x))\eta_i(x) = \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{I}_{\{g(x)=y_j\}} C(y_i, y_j)\eta_i(x) =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{I}_{\{x \in D_j\}} C_{ij}\eta_i(x) = \sum_{j=1}^{k} \mathbf{I}_{\{x \in D_j\}} \sum_{i=1}^{k} C_{ij}\eta_i(x) = \sum_{j=1}^{k} \mathbf{I}_{\{x \in D_j\}} d_j(x), \qquad (1.3)$$

ahol $d_j(x) \stackrel{\text{def}}{=} \sum_{i=1}^{k} C_{ij}\eta_i(x)$. Ez éppen a költség feltételes várhatóértéke ha $X = x$ *és* a következtetés a j-edik hipotézisre dönt.

Látható, hogy a fenti $r(g, x)$-et olyan g minimalizálja, amely x-et az (egyik) legkisebb $d_j(x)$-hez tartozó D_j tartományba sorolja (ld. 4. tétel alább). Ha több legkisebb $d_j(x)$ van, akkor mindegy melyiket választjuk, legyen ez pl. legkisebb indexű. Legyenek tehát a D_j^* tartományok olyanok, hogy $\forall x \in D_j^*$ akkor és csak akkor, ha $d_j(x) < d_i(x)\ \forall i < j$-re és $d_j(x) \leq d_i(x)\ \forall i > j$-re, vagyis x pontosan akkor eleme D_j^*-nak, ha az $X = x$ esetén a j-edik hipotézisre való döntés (feltételes) várható költsége az (egyik) legkisebb, és több minimális várható költség esetén a hipotézisek indexei közül j a legkisebb.

4. gyakorlat. Mutassuk meg, hogy a fenti D_j^*-k páronként diszjunktak és uniójuk $\mathcal{X}$, azaz $\mathcal{X}$ partícióját adják, továbbá hogy

$$x \in D_j^* \Rightarrow d_j(x) = \min_{1 \leq i \leq k} d_i(x). \qquad (1.4)$$

2. definíció. A fenti $(D_1^*, \ldots, D_k^*)$ tartományok által meghatározott g^* döntésfüggvényt (azaz amire $g^*(x) = y_j \Leftrightarrow x \in D_j^*$) *Bayes-döntésnek* nevezzük.

4. tétel. *A Bayes-döntés $\forall x$-re minimalizálja a lokális kockázatot, és így optimális. A minimum értéke* $r(g^*, x) = \min_{1 \leq i \leq k} d_i(x)$.

Bizonyítás. Bármely g döntésfüggvényre és $\forall x \in \mathcal{X}$-re, (1.3) szerint

$$\begin{aligned} r(g,x) &= \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j\}} d_j(x) \geq \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j\}} \min_{1\leq i\leq k} d_i(x) = \min_{1\leq i\leq k} d_i(x) \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j\}} = \\ &= \min_{1\leq i\leq k} d_i(x) = \min_{1\leq i\leq k} d_i(x) \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j^*\}} = \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j^*\}} \min_{1\leq i\leq k} d_i(x) = \\ &= \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j^*\}} d_j(x) = r(g^*,x), \end{aligned}$$

ahol az utolsó két egyenlőséghez (1.4)-t majd ismét (1.3)-t használtuk. Tehát $r(g*,x) \leq r(g,x)$, és így (1.1) alapján $R(g*) \leq R(g)$, azaz g^* optimális. □

Tehát a Bayes-döntés (optimális) globális kockázata

$$R^* \stackrel{\text{def}}{=} R(g^*) = \mathbf{E}\left[\min_{1\leq i\leq k} d_i(x)\right]. \tag{1.5}$$

Ezt *Bayes-kockázatnak* is nevezzük.

Vizsgáljuk meg a 2. példabeli $C_{ij} = C_0(i,j) = \mathbf{I}_{\{i\neq j\}}$ költségfüggvényt. Ekkor

$$d_j(x) = \sum_{i=1}^{k} C_{ij}\eta_i(x) = \sum_{i=1}^{k} \mathbf{I}_{\{i\neq j\}}\eta_i(x) = \sum_{i=1}^{k} \eta_i(x) - \eta_j(x) = 1 - \eta_j(x).$$

Így most $x \in D_j^*$ pontosan akkor, ha $\eta_j(x) > \eta_i(x)\ \ \forall i < j$-re és $\eta_j(x) \geq \eta_i(x)\ \ \forall i > j$-re, vagyis ekkor $\eta_j(x) = \max_{1\leq i\leq k} \eta_i(x)$. Tehát most g^* azt a hipotézist választja, amelyik a legvalószínűbb a megfigyelés ismeretében, azaz g^*-ot a maximális a posteriori valószínűségek megkeresésével határozhatjuk meg. Ezért ekkor a Bayes-döntést *maximum a posteriori döntésnek* is nevezik.

Ebben az esetben g lokális kockázata (1.3) szerint

$$r(g,x) = \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j\}}(1-\eta_j(x)) = 1 - \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j\}}\eta_j(x),$$

ami $g = g^*$-ra a fentiek alapján így egyszerűsíthető:

$$r(g^*,x) = 1 - \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j^*\}}\eta_j(x) = 1 - \sum_{j=1}^{k} \mathbf{I}_{\{x\in D_j^*\}} \max_{1\leq i\leq k} \eta_i(x) = 1 - \max_{1\leq i\leq k} \eta_i(x).$$

Tehát a globális kockázat (1.1) alapján ekkor $R^* = R(g^*) = 1 - \mathbf{E}\left[\max_{1\leq i\leq k} \eta_i(x)\right]$, amit *Bayes-hibának* is nevezünk. [2, 2.1. fejezet]

6. példa. Speciálisan ha csak két hipotézis van, azaz Y bináris ($k = 2$), akkor a kockázatok egyszerűen így is írhatóak:

$$r(g^*, x) = 1 - \max(\eta_1(x), \eta_2(x)) = \min(\eta_1(x), \eta_2(x))$$

és

$$R(g^*) = \mathbf{E}\left[\min(\eta_1(X), \eta_2(X))\right] = \mathbf{E}\left[\min(\eta_1(X), 1 - \eta_1(X))\right].$$

Ha X diszkrét vagy abszolút folytonos változó, akkor az η_i-ket kifejezhetjük az eloszlásokból.

Legyen először X diszkrét (azaz $\mathcal{X}$ megszámlálható halmaz), és jelölje

$$p_i(x) = \mathbf{P}\left[X = x | Y = y_i\right]$$

az X feltételes súlyfüggvényét ha $Y = y_i$. Ekkor a pozitív eséllyel előforduló x-ekre a feltételes valószínűség definíciója szerint

$$\eta_i(x) = \mathbf{P}\left[Y = y_i | X = x\right] = \frac{\mathbf{P}\left[Y = y_i, X = x\right]}{\mathbf{P}\left[X = x\right]} = \\ = \frac{\mathbf{P}\left[Y = y_i\right]\mathbf{P}\left[X = x | Y = y_i\right]}{\mathbf{P}\left[X = x\right]} = \frac{q_i p_i(x)}{\mathbf{P}\left[X = x\right]}. \tag{1.6}$$

Tehát a maximális a posteriori valószínűséget – és így a C_0 0-1 költség esetén a Bayes-döntést – az a hipotézis adja, amelyikre $q_i p_i(x)$ maximális.

Abban a speciális esetben, ha minden q_i egyenlő (szükségképpen $1/k$), azaz Y (a priori) eloszlása egyenletes, akkor $q_i p_i(x)$ és $p_i(x)$ ugyanazon i-re maximális, tehát a Bayes-döntés azt a i-t választja, amelyikre $p_i(x)$ maximális.

Amikor a q_i valószínűségek ismeretlenek, sokszor nincs jobb kiindulás, mint egyenletes a priori eloszlást feltételezni és ezért a fentiek alapján arra a hipotézisre dönteni, amelyikre $p_i(x)$ maximális, azaz amelyiket feltételezve a megfigyelésnek a legnagyobb a valószínűsége. (Döntetlen esetén a választás most is tetszőleges lehet.)

3. definíció. Diszkrét X esetén g-t *maximum likelihood döntésnek* nevezzük, ha $g(x) = y_j$ esetén $p_j(x) = \max_i p_i(x)$.

2. *megjegyzés.* Figyeljük meg, hogy itt a maximalizálandó feltételes valószínűségben az argumentum és a feltétel éppen fel van cserélve a maximum a posteriori döntéshez képest!

Legyen most $\mathcal{X} \subseteq \mathbf{R}$, X abszolút folytonos $f(x)$ sűrűségfüggvénnyel, és jelölje

$$f_i(x) = f(x | Y = y_i)$$

az X feltételes sűrűségfüggvényét ha $Y = y_i$. Ekkor η_i precíz definíciója technikailag bonyolultabb, de határérték számítással meggondolható, hogy $f(x) \neq 0$ esetén a diszkrét esettel és az (1.6) egyenlőséggel formailag analóg módon

$$\eta_i(x) = \frac{q_i f_i(x)}{f(x)} \tag{1.7}$$

adódik. Ekkor tehát a Bayes-döntés megkeresése – 0-1 költség esetén – $q_i f_i(x)$ maximalizálását, a maximum likelihood döntésé pedig $f_i(x)$ maximalizálását jelenti:

4. definíció. Abszolút folytonos X esetén g-t *maximum likelihood döntésnek* nevezzük, ha $g(x) = y_j$ esetén $f_j(x) = \max_i f_i(x)$.

3. *megjegyzés.* Az (1.7) egyenlőség fennállását illusztrálhatjuk egy speciális esettel: Közelítsük $\eta_i(x)$-ben az $\{X = x\}$ feltételt a $S_\epsilon(x) \stackrel{\text{def}}{=} \{x - \epsilon < X < x + \epsilon\}$ eseményekkel, ahol $\epsilon > 0$ egyre kisebb, majd $\eta_i(x)$-t a $\eta_i^\epsilon(x) \stackrel{\text{def}}{=} \mathbf{P}[Y = y_i | S_\epsilon(x)]$ függvényekkel. Ekkor, ha $f_1,\ldots,f_k$, és következésképpen $f(x)$ folytonos függvények, akkor $f(x) > 0$ miatt elég kicsi ϵ-ra $\mathbf{P}[S_\epsilon(x)] > 0$, és így

$$\eta_i^\epsilon(x) = \frac{\mathbf{P}[Y = y_i, S_\epsilon(x)]}{\mathbf{P}[S_\epsilon(x)]} = \frac{\mathbf{P}[Y = y_i]\,\mathbf{P}[S_\epsilon(x)|Y = y_i]}{\mathbf{P}[S_\epsilon(x)]} = \frac{q_i \int_{x-\epsilon}^{x+\epsilon} f_i(z)\,dz}{\int_{x-\epsilon}^{x+\epsilon} f(z)\,dz} =$$
$$= \frac{q_i \frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} f_i(z)\,dz}{\frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} f(z)\,dz} \xrightarrow{\epsilon \to 0} \frac{q_i f_i(x)}{f(x)},$$

ahol a számlálóban és a nevezőben is az integrálszámítás folytonos függvényekre vonatkozó középértéktételét alkalmaztuk. Tehát az a posteriori valószínűségek $\epsilon \to 0$ határértékben valóban a $q_i f_i(x)/f(x)$ értékek lesznek.

7. példa. Emlékezet nélküli bináris szimmetrikus csatorna (BSC) kimenetének maximum likelihood dekódolása. Tekintsük a következő szituációt: Egy emlékezet nélküli, zajos bináris szimmetrikus csatornán egy $\mathbf{Y} = (Y_1, \ldots, Y_n)$ kódszót továbbítunk. $\mathbf{Y}$ vektor valószínűségi változó, amely az n hosszúságú, bináris (pl. $\{0,1\}$ értékű) sorozatoknak egy rögzített $\mathcal{Y} \subseteq \{0,1\}^n$ részhalmazán (kódkönyv) veszi fel az értékét. A bináris szimmetrikus csatorna minden bitet $p \in (0,1)$ eséllyel megváltoztat, $1-p$ eséllyel változatlanul hagy. Az emlékezet nélküliség azt jelenti, hogy a különböző bitekre ezek az események függetlenek. Így a csatorna kimenete is egy n hosszúságú $\mathbf{X} = (X_1, ..., X_n) \in \mathcal{X} = \{0,1\}^n$ bitsorozat (egy vektor valószínűségi változó, amely azonban nem feltétlenül kódszó). A zajos $\mathbf{X}$-et megfigyelve akarunk döntést hozni, hogy mi lehetett a kódszó úgy, hogy a költségünk 1 ha hibázunk, 0 egyébként (ld. 2. példa). Figyeljük meg, hogy míg a csatorna bemenete $\mathbf{Y}$ és kimenete $\mathbf{X}$, addig a döntés (a dekódoló) bemenete $\mathbf{X}$ és kimenete $\mathbf{Y}$(-ra vonatkozó döntés).

Jelöljük az i-edik lehetséges kódszót $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in})$-nel, egy lehetséges csatorna kimenetet $\mathbf{x} = (x_1, x_2, \ldots, x_n)$-nel, vagyis az i-edik kódszó j-edik bitje y_{ij}, a kimeneté x_j. Ha $\mathbf{Y} = \mathbf{y}_i$, akkor az $\mathbf{X} = \mathbf{x}$ feltételes valószínűsége a következő

$$\mathbf{P}[\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y}_i] = \mathbf{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | Y_1 = y_{i1}, \ldots, Y_n = y_{in}] =$$
$$= \mathbf{P}[X_1 = x_1|Y_1 = y_{i1}]\,\mathbf{P}[X_2 = x_2|Y_2 = y_{i2}] \ldots \mathbf{P}[X_n = x_n|Y_n = y_{in}],$$

ahol az utolsó egyenlőség az emlékezet nélküliség definíciója. Mivel az átmenetvalószínűség p,

$$\mathbf{P}[X_t = x_t|Y_t = y_{it}] = \begin{cases} p, & \text{ha } x_t \neq y_{it}, \\ 1-p, & \text{ha } x_t = y_{it} \end{cases} = p^{\mathbf{I}_{\{x_t \neq y_{it}\}}} (1-p)^{1-\mathbf{I}_{\{x_t \neq y_{it}\}}}.$$

Így aztán

$$\mathbf{P}\left[\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y}_i\right]=\prod_{t=1}^{n}\left[p^{\mathbf{I}_{\{x_t\neq y_{it}\}}}(1-p)^{1-\mathbf{I}_{\{x_t\neq y_{it}\}}}\right]=p^{\sum_{t=1}^{n}\mathbf{I}_{\{x_t\neq y_{it}\}}}(1-p)^{n-\sum_{t=1}^{n}\mathbf{I}_{\{x_t\neq y_{it}\}}}=$$

$$=(1-p)^n\left(\frac{p}{1-p}\right)^{\sum_{t=1}^{n}\mathbf{I}_{\{x_t\neq y_{it}\}}}. \tag{1.8}$$

Ha az $\mathbf{Y}$ a priori eloszlása nem ismert, akkor a Bayes-döntés is ismeretlen, maximum likelihood döntést azonban használhatunk $\mathbf{X}$ ismeretében az $\mathbf{Y}$ kódszóra való következtetéshez. Ekkor azt a $\mathbf{y}_i\in\mathcal{Y}$ kódszót választjuk, amely mellett $p_i(\mathbf{x})=\mathbf{P}\left[\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y}_i\right]$ a legnagyobb. Feltéve, hogy $0<p<1/2$, azaz $p/(1-p)<1$, (1.8) alapján ez az a $\mathbf{y}_i$, amelyre $\sum_{t=1}^{n}\mathbf{I}_{\{x_t\neq y_{it}\}}$ a legkisebb. Megfigyelhetjük, hogy $\sum_{t=1}^{n}\mathbf{I}_{\{x_t\neq y_{it}\}}$ éppen azon bitek száma, amelyekben $\mathbf{x}$ és $\mathbf{y}_i$ különbözik. Ezt $\mathbf{x}$ és $\mathbf{y}_i$ *Hamming távolságának* is nevezik. Ezek szerint bináris szimmetrikus csatorna esetén a maximum likelihood döntéssel való dekódolás a kimeneten megjelenő sorozathoz Hamming távolságban legközelebbi kódszónak (kódszavak egyikének) a választását jelenti. (Természetesen egyenletes bemeneti ($\mathbf{Y}$) eloszlás esetén a maximum likelihood döntés most is éppen a maximum a posteriori döntés lesz.)

5. gyakorlat. a) Mi lesz a maximum likelihood dekódolás $p>1/2$ esetén? Mi emögött a magyarázat?
b) Mi lesz a maximum likelihood dekódolás $p=1/2$ esetén? Mi emögött a magyarázat?

1.4. Bayes-döntés ismételt megfigyelés alapján

Előfordul, hogy egyetlen X megfigyelés helyett n számú megfigyelés, $\mathbf{X}=(X_1,\ldots,X_n)\in\mathcal{X}^n$, is a rendelkezésünkre áll, amelyek az $\{Y=y_i\}$ feltétel mellett feltételesen függetlenek és azonos $(p_i(x))$ eloszlásúak. Tegyük fel az egyszerűség kedvéért, hogy X_t diszkrét, $k=2$ ($\mathcal{Y}=\bar{\mathcal{Y}}=\{y_1,y_2\}$) és $C_{11}=C_{22}=0$ (mint pl. a 2., 3. és 4. példákbeli és bármely (pszeudo)metrika tulajdonságú költségfüggvénynél). Sejthető, hogy ha az $\mathbf{X}$-ből egyáltalán lehet következtetni Y-ra, azaz az X_t-k nem függetlenek az Y-tól, akkor az $(\mathbf{X},Y)$-hoz tartozó Bayes-kockázat tetszőlegesen kicsivé válik, amint $n\to\infty$. Valóban, az alábbi tétel szerint a Bayes-kockázat exponenciálisan tart 0-hoz.

5. tétel. *Legyen g_n^* az $(\mathbf{X},Y)$ döntési problémához tartozó Bayes-döntés. Ha $q_1q_2=0$, vagy ha $q_1q_2\neq 0$ és létezik $x\in\mathcal{X}$, hogy $p_1(x)\neq p_2(x)$, akkor*

$$\lim_{n\to\infty}R(g_n^*)=0,$$

és a konvergencia exponenciálisan gyors.

4. *megjegyzés.* A tétel feltétele pontosan akkor áll fenn, ha $q_1q_2=0$, vagy ha $q_1q_2\neq 0$ és X_t és Y nem független (azaz $\sum_{i\in\{1,2\},x\in\mathcal{X}}q_ip_i(x)\log\frac{p_i(x)}{\mathbf{P}[X=x]}$ kölcsönös információjuk nem 0). Az is látható, hogy ha $q_1q_2C_{12}C_{21}\neq 0$ de $\mathbf{X}$ és Y független, akkor nem kaphatunk tetszőlegesen kicsi Bayes-kockázatot n növekedésével.

Bizonyítás. Az 1.3 fejezetben láttuk, hogy a Bayes-kockázat (1.5) szerint

$$R(g_n^*) = \mathbf{E}\left[\min(d_1(\mathbf{X}), d_2(\mathbf{X}))\right] = \sum_{\mathbf{x}\in\mathcal{X}^n} \min\left(\sum_{i=1}^{2} C_{i1}\eta_i(\mathbf{x}), \sum_{i=1}^{2} C_{i2}\eta_i(\mathbf{x})\right) \mathbf{P}\left[\mathbf{X}=\mathbf{x}\right] =$$
$$= \sum_{\mathbf{x}\in\mathcal{X}^n} \min(C_{21}\eta_2(\mathbf{x}), C_{12}\eta_1(\mathbf{x}))\mathbf{P}\left[\mathbf{X}=\mathbf{x}\right].$$

Ha $q_1q_2 = 0$, akkor $\forall \mathbf{x} \in \mathcal{X}^n$-re $\eta_1(\mathbf{x})$ vagy $\eta_2(\mathbf{x})$ is 0, így a minimum és $R(g_n^*)$ is 0. Ha $q_1q_2 \neq 0$, akkor behelyettesítve (1.6)-t

$$\begin{aligned} R(g_n^*) &= \sum_{\mathbf{x}\in\mathcal{X}^n} \min\left(C_{21}\frac{q_2p_2(\mathbf{x})}{\mathbf{P}\left[\mathbf{X}=\mathbf{x}\right]}, C_{12}\frac{q_1p_1(\mathbf{x})}{\mathbf{P}\left[\mathbf{X}=\mathbf{x}\right]}\right)\mathbf{P}\left[\mathbf{X}=\mathbf{x}\right] = \\ &= \sum_{\mathbf{x}\in\mathcal{X}^n} \min(C_{21}q_2p_2(\mathbf{x}), C_{12}q_1p_1(\mathbf{x})) \leq \\ &\leq \sum_{\mathbf{x}\in\mathcal{X}^n} \sqrt{C_{21}q_2p_2(\mathbf{x})C_{12}q_1p_1(\mathbf{x})}, \end{aligned} \tag{1.9}$$

ahol a legutóbbi lépésnél azt használtuk, hogy $\forall a_1,a_2 \geq 0$-ra $\min(a_1, a_2) \leq \sqrt{a_1a_2}$.

$\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$-re a $p_i(\mathbf{x}) = \mathbf{P}\left[\mathbf{X}=\mathbf{x}|Y=y_i\right]$ feltételes valószínűségek az X_t-k feltételes függetlensége alapján a következőképpen fejezhetőek ki:

$$p_i(\mathbf{x}) = \mathbf{P}\left[X_1=x_1,\ldots,X_n=x_n|Y=y_i\right] = \prod_{t=1}^{n} \mathbf{P}\left[X_t=x_t|Y=y_i\right] = \prod_{t=1}^{n} p_i(x_t).$$

Ezt (1.9)-be helyettesítve

$$\begin{aligned} R(g_n^*) &\leq \sqrt{C_{12}C_{21}q_1q_2} \sum_{(x_1,\ldots,x_n)\in\mathcal{X}^n} \sqrt{\prod_{t=1}^{n} p_1(x_t) \prod_{t=1}^{n} p_2(x_t)} \\ &= \sqrt{C_{12}C_{21}q_1q_2} \sum_{(x_1,\ldots,x_n)\in\mathcal{X}^n} \prod_{t=1}^{n} \sqrt{p_1(x_t)p_2(x_t)} \\ &= \sqrt{C_{12}C_{21}q_1q_2} \left(\sum_{x\in\mathcal{X}} \sqrt{p_1(x)p_2(x)}\right)^n. \end{aligned}$$

A számtani- és mértani-közép közötti összefüggés alapján

$$\sum_{x\in\mathcal{X}} \sqrt{p_1(x)p_2(x)} \leq \sum_{x\in\mathcal{X}} \frac{p_1(x)+p_2(x)}{2} = \frac{1}{2}\sum_{x\in\mathcal{X}} p_1(x) + \frac{1}{2}\sum_{x\in\mathcal{X}} p_2(x) = 1,$$

ahol egyenlőség (akkor és) csak akkor áll fenn, ha $\forall x \in \mathcal{X}$-re $p_1(x) = p_2(x)$. Tehát ha $\exists x \in \mathcal{X}$, hogy $p_1(x) \neq p_2(x)$, akkor $0 \leq \sum_{x\in\mathcal{X}} \sqrt{p_1(x)p_2(x)} < 1$, így $R(g_n^*)$ exponenciálisan 0-hoz tart. □

1.5. Bayes-döntés közelítése

A $d_i(x)$ várható költségek (vagy az $\eta_i(x)$ a posteriori valószínűségek) pontos értékei általában ismeretlenek. Tekintsünk egy ilyen szituációt, amelyben azonban a d_i-ket meg tudjuk becsülni valamely $\tilde{d}_i$ függvényekkel. Az (1.4)-et teljesítő Bayes-döntés mintájára a $\{\tilde{d}_i\}_{1\le i\le k}$ függvényekhez is hozzárendelhetünk analóg módon olyan $\tilde{g}$ döntést, amely x esetén az (egyik) legkisebb $\tilde{d}_j(x)$-hez tartozó y_j-re dönt, azaz $\tilde{g}(x) = y_j \Rightarrow \tilde{d}_j(x) = \min_{1\le i\le k} \tilde{d}_i(x)$. (Több minimális esetén valamilyen elv szerint választunk ezen hipotézisek közül, pl. ismét a legkisebb indexűt.) $\tilde{g}$ tehát úgy viszonyul a $\tilde{d}_i$-khez, ahogy g^* a d_i-khez. Vajon ha a $\tilde{d}_i$-k jó becslések, akkor $\tilde{g}$ (lokális/globális) kockázata közel lesz g^* (lokális/globális) kockázatához? (Kisebb természetesen nem lehet az 4. tétel alapján.) A következő tétel erre ad pozitív választ mutatva, hogy a szóbanforgó kockázatok különbsége korlátozható a $\tilde{d}_i$-k becslési hibájával:

6. tétel. *$i = 1,\dots,k$-ra legyen $\tilde{d}_i : \mathcal{X} \to [0,\infty)$ a d_i becslése és $\tilde{g}$ egy a $\{\tilde{d}_i\}_{1\le i\le k}$-khez rendelt döntésfüggvény. Ekkor*

$$r(\tilde{g},x) - r(g^*,x) \le \mathbf{I}_{\{\tilde{g}(x)\neq g^*(x)\}} \sum_{i=1}^{k} |\tilde{d}_i(x) - d_i(x)|$$

és

$$R(\tilde{g}) - R(g^*) \le \mathbf{E}\left[\mathbf{I}_{\{\tilde{g}(X)\neq g^*(X)\}} \sum_{i=1}^{k} |\tilde{d}_i(X) - d_i(X)|\right] \le \mathbf{E}\left[\sum_{i=1}^{k} |\tilde{d}_i(X) - d_i(X)|\right].$$

Bizonyítás. Az egyszerűbb jelölés kedvéért az általánosság megszorítása nélkül feltehetjük, hogy $y_j = j$. Ekkor (1.3) szerint a lokális kockázata így is írható

$$r(g,x) = \sum_{j=1}^{k} \mathbf{I}_{\{g(x)=j\}} d_j(x) = d_{g(x)}(x).$$

Tehát $\tilde{g}$ és g^* lokális kockázatának különbsége

$$r(\tilde{g},x) - r(g^*,x) = d_{\tilde{g}(x)}(x) - d_{g^*(x)}(x).$$

Ha $\tilde{g}(x) = g^*(x)$, akkor $r(\tilde{g},x) - r(g^*,x) = 0$. Ha $\tilde{g}(x) \neq g^*(x)$, akkor viszont

$$\begin{aligned}
d_{\tilde{g}(x)}(x) - d_{g^*(x)}(x) &= d_{\tilde{g}(x)}(x) - \tilde{d}_{\tilde{g}(x)}(x) + \tilde{d}_{\tilde{g}(x)}(x) - d_{g^*(x)}(x) \\
&\le d_{\tilde{g}(x)}(x) - \tilde{d}_{\tilde{g}(x)}(x) + \tilde{d}_{g^*(x)}(x) - d_{g^*(x)}(x) \\
&\quad \Big(\text{mert } \tilde{g} \text{ definíciója szerint } \tilde{d}_{\tilde{g}(x)}(x) \le \tilde{d}_{g^*(x)}(x)\Big) \\
&\le |d_{\tilde{g}(x)}(x) - \tilde{d}_{\tilde{g}(x)}(x)| + |\tilde{d}_{g^*(x)}(x) - d_{g^*(x)}(x)| \\
&\le \sum_{i=1}^{k} |\tilde{d}_i(x) - d_i(x)| \\
&\quad (\text{mert } \tilde{g}(x) \text{ és } g^*(x) \text{ különböző elemei } \{1,\dots,k\}\text{-nak}).
\end{aligned}$$

Összefoglalva

$$r(\tilde{g},x)-r(g^*,x)\le \mathbf{I}_{\{\tilde{g}(x)\neq g^*(x)\}}\sum_{i=1}^{k}|\tilde{d}_i(x)-d_i(x)|,$$

majd x helyére X-et írva és várhatóértéket véve kapjuk az korlátot $R(\tilde{g})-R(g^*)$-re. Az utolsó egyenlőtlenség triviális $\mathbf{I}_{\{\tilde{g}(x)\neq g^*(x)\}}\le 1$-ből. □

Nézzük ismét a 2. példabeli $C_{ij}=\mathbf{I}_{\{i\neq j\}}$ speciális esetet, amikor – mint láttuk – $d_i(x)=1-\eta_i(x)$ és $g^*(x)=y_j\Rightarrow \eta_j(x)=\max_{1\le i\le k}\eta_i(x)$. Feltehető, hogy ekkor a $\tilde{d}_i$ becslők $1-\tilde{\eta}_i$ alakban állnak elő, ahol $\tilde{\eta}_i$-k a η_i-k becslései. Ekkor

$$\tilde{g}(x)=y_j\Rightarrow \tilde{\eta}_j(x)=\max_{1\le i\le k}\tilde{\eta}_i(x), \tag{1.10}$$

és a 6. tétel alakja

$$r(\tilde{g},x)-r(g^*,x)\le \mathbf{I}_{\{\tilde{g}(x)\neq g^*(x)\}}\sum_{i=1}^{k}|\tilde{\eta}_i(x)-\eta_i(x)|$$

és

$$R(\tilde{g})-R(g^*)\le \mathbf{E}\left[\mathbf{I}_{\{\tilde{g}(X)\neq g^*(X)\}}\sum_{i=1}^{k}|\tilde{\eta}_i(X)-\eta_i(X)|\right]\le \mathbf{E}\left[\sum_{i=1}^{k}|\tilde{\eta}_i(X)-\eta_i(X)|\right] \tag{1.11}$$

lesz. Ha $k=2$, és feltesszük, hogy $(\tilde{\eta}_1(x),\tilde{\eta}_2(x))$ minden $x\in\mathcal{X}$-re eloszlást alkot, azaz $\tilde{\eta}_2(x)=1-\tilde{\eta}_1(x)$, akkor ezek tovább egyszerűsödnek a következőképpen

$$\begin{aligned} r(\tilde{g},x)-r(g^*,x)&\le \mathbf{I}_{\{\tilde{g}(x)\neq g^*(x)\}}\left(|\tilde{\eta}_1(x)-\eta_1(x)|+|\tilde{\eta}_2(x)-\eta_2(x)|\right)=\\ &=2\mathbf{I}_{\{\tilde{g}(x)\neq g^*(x)\}}|\tilde{\eta}_1(x)-\eta_1(x)| \end{aligned}$$

és

$$R(\tilde{g})-R(g^*)\le 2\mathbf{E}\left[\mathbf{I}_{\{\tilde{g}(X)\neq g^*(X)\}}|\tilde{\eta}_1(X)-\eta_1(X)|\right]\le 2\mathbf{E}\left[|\tilde{\eta}_1(X)-\eta_1(X)|\right].$$

8. példa. Legyen $C_{ij}=\mathbf{I}_{\{i\neq j\}}$, $k=2$, $y_1=1$ és $y_2=0$. Mutassuk meg, hogy ekkor $\eta_1(X)=\mathbf{E}[Y|X]$. Tehát ha van egy jó becslésünk Y feltételes várhatóérték függvényére, akkor az ahhoz (1.10) alapján rendelt döntésfüggvény közel optimális.

1.6. Bayes-becslés

Szemben az eddigi 1.3–1.5. fejezettel, ahol $\mathcal{Y}=\bar{\mathcal{Y}}$ (tetszőleges) véges halmaz volt és többnyire a 0-1 költséget vizsgáltuk, azaz Y értékét el akartuk *dönteni* és hiba esetén nem volt érdekes, hogy mekkora a hibázás nagysága, az alábbi fejezetekben $\mathcal{Y}$ és $\bar{\mathcal{Y}}$ általában $\mathbf{R}$ részhalmaza, többnyire végtelen, sőt folytonos, és C valamiféle távolságát méri a valódi és a *becsült* paraméternek, tehát a hibázás nagysága is számít. Ekkor g-t *becslésfüggvénynek*, $g(X)$-t *becslésnek* is nevezzük. A Bayes-feladat továbbra is az $R(g)$ globális kockázatot minimalizáló $g:\mathcal{X}\to\mathcal{Y}$ becslésfüggvény megtalálása. (Pl. a 4. példabeli C_2-re $R(g)$ éppen a négyzetes középhiba.)

5. definíció. *Bayes-becslésnek* nevezzük az optimális g^* becslésfüggvényeket, azaz amikre

$$R^* \stackrel{\text{def}}{=} R(g^*) = \min_g R(g).$$

Fennáll a következő elégséges feltétel:

7. tétel. *Ha egy g^* becslés lokális kockázatára*

$$r(g^*, x) = \min_{y \in \bar{\mathcal{Y}}} \mathbf{E}\left[C(Y, y) | X = x\right], \qquad \forall x \in \mathcal{X},$$

akkor g^* Bayes-becslés.

Bizonyítás. Bármely g becslésfüggvényre

$$R(g) = \mathbf{E}\left[\mathbf{E}\left[C(Y, g(X))|X\right]\right] \geq \mathbf{E}\left[\min_{y \in \bar{\mathcal{Y}}} \mathbf{E}\left[C(Y, y)|X\right]\right] = \mathbf{E}\left[r(g^*, X)\right] = R(g^*),$$

így g^* optimális. □

9. példa. Ha $\mathcal{Y}$ ismét mégis véges és tekintjük a 2. példabeli $C_0(y, y') = \mathbf{I}_{\{y \neq y'\}}$ költséget, akkor nyilván nincs ok olyan g becslésfüggvényt használni, amelynek értékkészlete nem része $\mathcal{Y}$-nak. Így ekkor a becslési feladat ekvivalens lesz a 2. példabeli döntési feladattal, és a Bayes-becslést az 1.3 fejezet szerinti *maximum a posteriori becslés* (döntés) adja, azaz ha $g^*(x) = y_j$, akkor $\eta_j(x) \geq \eta_i(x)$ $\forall i \neq j$-re (vagyis $\eta_j(x) = \max_{1 \leq i \leq k} \eta_i(x)$), ahol az $\eta_i(x)$-eket (1.6) illetve (1.7)-ből számolhatjuk ki, ha X diszkrét illetve abszolút folytonos. Tehát a maximális a posteriori becslést – és így a Bayes-becslést – az a hipotézis adja, amelyikre $q_i p_i(x)$ illetve $q_i f_i(x)$ maximális.

1.7. Maximum likelihood becslés

Mivel most Y tipikusan nem diszkrét, hanem például abszolút folytonos, ez esetben Y q_i a priori valószínűségei helyett csak a $q(y)$ a priori sűrűségfüggvénye értelmezhető. Ekkor X feltételes ($Y = y_i$ melletti) $p_i(x)$ súlyfüggvényei illetve $f_i(x)$ sűrűségfüggvényei helyett a

$$p_y(x) = \mathbf{P}\left[X = x | Y = y\right]$$

feltételes súlyfüggvényei illetve

$$f_y(x) = f(x|Y = y)$$

feltételes sűrűségfüggvényei értelmezhetők, amelyek az X (diszkrét illetve abszolút folytonos) eloszlását adják meg $Y = y \in \mathcal{Y}$ feltétel esetén. Ugyan formálisan ekkor is definiálható lenne az $\arg\max_{y \in \mathcal{Y}} q(y) p_y(x)$ illetve $\arg\max_{y \in \mathcal{Y}} q(y) f_y(x)$ maximum a posteriori becslés, ez nem igazán releváns, mert folytonos Y-ra $\forall y \in \bar{\mathcal{Y}}$-ra $\mathbf{E}\left[C_0(Y, y)\right] = \mathbf{P}\left[Y \neq y\right] \equiv 1$ 1 valószínűséggel.

A *maximum likelihood becslés* azonban definiálható (a maximum likelihood döntéssel analóg módon):

6. definíció. g-t maximum likelihood becslésnek nevezzük, ha $p_{g(x)}(x) = \max_{y\in\mathcal{Y}} p_y(x)$ illetve $f_{g(x)}(x) = \max_{y\in\mathcal{Y}} f_y(x)$ diszkrét illetve abszolút folytonos X esetén.

10. példa. Az 1.4. fejezethez hasonlóan legyen az $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbf{R}^n$ vektor valószínűségi változó, ahol az X_t-k független, ismeretlen Y várhatóértékű, σ szórású, azonos normális (Gauss) eloszlású valószínűségi változók. Határozzuk meg Y-nak az $\mathbf{X}$ megfigyelésre alapozott maximum likelihood becslését!

$\mathbf{X}$ feltételes sűrűségfüggvényét $Y = y$ esetén ekkor a több-dimenziós normális sűrűségfüggvény adja:

$$f_y(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^n (x_t-y)^2},$$

ahol $\mathbf{x} = (x_1, \ldots, x_n)$. A maximum likelihood becslés az az y lesz, amelyre $f_y(\mathbf{x})$ maximális. Ez megegyezik azzal, amelyre

$$-\ln f_y(\mathbf{x}) = n\ln(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}\sum_{t=1}^n (x_t - y)^2,$$

azaz $\frac{1}{n}\sum_{t=1}^n (x_t - y)^2$ minimális. Legyen $m_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^n x_t$. Ekkor a 1. tételt alkalmazva egy az $(x_1, \ldots, x_n)$-en egyenletes – és így m_n várhatóértékű – valószínűségi változóra:

$$\frac{1}{n}\sum_{t=1}^n (x_t - y)^2 = \frac{1}{n}\sum_{t=1}^n (x_t - m_n)^2 + (m_n - y)^2,$$

amely nyilván $y = m_n$-re minimális. Így a várhatóérték maximum likelihood becslése normális $\mathbf{X}$-re éppen a megfigyelések átlaga.

11. példa. A fenti hasonlóan legyen most az $\mathbf{X} = (X_1, ..., X_n) \in \mathbf{R}^n$, ahol az X_i-k független, ismeretlen Y várhatóértékű, Σ szórású, azonos, de egyenletes eloszlású valószínűségi változók. Határozzuk meg az (Y, Σ) párnak az $\mathbf{X}$ megfigyelésre alapozott maximum likelihood becslését!

$Y = y$, $\Sigma = \sigma$ esetén minden X_t az $[y-\sqrt{3}\sigma, y+\sqrt{3}\sigma]$ intervallumon kell hogy egyenletes legyen (a szórásnégyzet ekkor lesz $= (2\sqrt{3}\sigma)^2/12 = \sigma^2$), így – bevezetve $\delta \stackrel{\text{def}}{=} \sqrt{3}\sigma$-t – feltételes sűrűségfüggvényük

$$g_{y,\sigma} = \frac{\mathbf{I}_{[y-\delta, y+\delta]}}{2\delta}.$$

Tehát $\mathbf{X}$ feltételes sűrűségfüggvénye ha $Y = y$, $\Sigma = \sigma$ a függetlenség miatt $f_{y,\sigma}(\mathbf{x}) = \prod_{t=1}^n g_{y,\sigma}(x_t)$. Így ez ugyanarra az (y, σ) párra veszi fel maximumát, amelyre

$$\ln f_{y,\sigma}(\mathbf{x}) = \sum_{t=1}^n \ln g_{y,\sigma}(x_t) = \sum_{t=1}^n \ln \frac{\mathbf{I}_{\{x_t\in[y-\delta,y+\delta]\}}}{2\delta} = \sum_{t=1}^n \ln \mathbf{I}_{\{x_t\in[y-\delta,y+\delta]\}} - n\ln(2\delta) =$$
$$= \begin{cases} -n\ln(2\delta), & \text{ha } \forall x_t \in [y-\delta, y+\delta], \\ -\infty, & \text{egyébként.} \end{cases} \tag{1.12}$$

Ez akkor maximális, ha minden $x_t \in [y-\delta, y+\delta]$ a lehető legkisebb δ mellett. Az $[y-\delta, y+\delta]$ intervallumnak tehát a legkisebb x_t-től a legnagyobbig kell tartania, azaz az (Y, Σ) maximum likelihood becslése

$$y = \frac{\underline{X} + \overline{X}}{2} \qquad \text{illetve} \qquad \sigma = \frac{\overline{X} - \underline{X}}{2\sqrt{3}},$$

ahol $\underline{X} = \min_{1\le t\le n} X_t$ és $\overline{X} = \max_{1\le t\le n} X_t$.

12. példa. Tekintsük a 11. példabeli helyzetet, de legyen most a σ szórás – tehát az intervallumhossz – ismert. Határozzuk meg az Y-nak az $\mathbf{X}$ megfigyelésre alapozott maximum likelihood becslését!

A számolás nem változik, most is a (1.12)-t kell maximalizálni, de csak y-ban, míg $\delta = \sqrt{3}\sigma$ ismert, rögzített paraméter. (1.12) akkor maximális, ha minden $X_t \in [y-\delta, y+\delta]$, azaz $\overline{X} - \delta \le y \le \underline{X} + \delta$. Minden ilyen y az Y egy maximum likelihood becslése. Ez mutatja, hogy a maximum likelihood becslés nem mindig egyértelmű.

5. *megjegyzés.* Vegyük észre, hogy a 11. és a 12. példánál is a $(\underline{X}, \overline{X})$ pár ismerete elégséges volt a maximum likelihood becslés meghatározásához.

1.8. Regresszióbecslés; négyzetes középhiba minimalizálás

Legyen $Y \in \mathbf{R}$ véges szórású változó, és vizsgáljuk a 4. példabeli $C_2(y, y') = (y-y')^2$ négyzetes költséget, azaz keressük azt a g^* becslésfüggvényt, amelyre $R(g^*) = \mathbf{E}\left[(g^*(X) - Y)^2\right]$ minimális! Az 7. tétel szerint ha

$$r(g^*, x) = \min_{y\in\bar{\mathcal{Y}}} \mathbf{E}\left[(Y-y)^2 | X = x\right], \qquad \forall x \in \mathcal{X},$$

akkor g^* Bayes-becslés. Definiáljuk a *regressziós függvény* fogalmát:

7. definíció. Regressziós függvénynek nevezzük az

$$m(x) = \mathbf{E}\left[Y | X = x\right]$$

függvényt, amely minden x-re Y-nak a feltételes várhatóértékét adja ha $X = x$.

8. tétel. *Négyzetes költség esetén $g^*(X) = m(X)$ 1 valószínűséggel, vagyis a Bayes-becslés éppen a regressziós függvény. [3, p. 2]*

Bizonyítás. Rögzített x-re alkalmazzuk a 1. tételt az $Y|X = x$ feltételes eloszlásra: bármely $g(x) \in \mathbf{R}$-re

$$\begin{aligned} r(g, x) &= \mathbf{E}\left[(Y - g(x))^2 | X = x\right] = \\ &= \mathbf{E}\left[(Y - \mathbf{E}\left[Y|X=x\right])^2 | X = x\right] + (\mathbf{E}\left[Y|X=x\right] - g(x))^2 = \\ &= \mathbf{E}\left[(Y - m(x))^2 | X = x\right] + (m(x) - g(x))^2 = r(m, x) + (m(x) - g(x))^2. \end{aligned}$$

Tehát bármely g becslésre

$$R(g) = \mathbf{E}\left[r(g,X)\right] = \mathbf{E}\left[r(m,X)\right] + \mathbf{E}\left[(m(X)-g(X))^2\right] = R(m) + \mathbf{E}\left[(m(X)-g(X))^2\right],$$

ami pontosan akkor minimális, ha $g(X) = m(X)$ 1 valószínűséggel. □

6. *megjegyzés.* A 8. tétel szerint ha Y-t négyzetes értelemben akarjuk közelíteni az X egy $g(X)$ függvényével, akkor az $\mathbf{E}\left[(g(X)-Y)^2\right]$ középhibát minimalizáló $g(x)$ nem más, mint az $\mathbf{E}\left[Y|X=x\right]$ feltételes várhatóérték. Azaz csupán x ismeretében ez a legjobb becslés.

6. gyakorlat. Bizonyítsuk be, hogy ha $(X,Y) \in \mathbf{R}^2$ együttesen normális eloszlású és – az egyszerűség kedvéért – mindkettő szórása 1, akkor a regressziós függvény

$$m(x) = \mathbf{E}\left[Y|X=x\right] = \mathbf{E}\left[Y\right] + \rho(x - \mathbf{E}\left[X\right]),$$

ahol $\rho = \mathbf{E}\left[XY\right] - \mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]$ az X és Y korrelációs együtthatója (egyben kovarianciája). Próbáljuk meg általánosítani ezt tetszőleges szórásokra, majd többdimenziós $\mathbf{X}$ vektor változóra.

1.8.1. Lineáris becslés

Továbbra is legyen $Y \in \mathbf{R}$ véges szórású, $\mathbf{X} = (X_1, ..., X_d) \in \mathbf{R}^d$ pedig vektor valószínűségi változó, és vizsgáljuk a $C_2(y,y') = (y-y')^2$ költséget. A 8. tétel szerint a Bayes-becslést lényegében az $m(x)$ regressziós függvény adja. A 6. gyakorlatban láthattuk, hogy ha $(\mathbf{X},Y)$ együttesen normális eloszlású (ami egyéb speciális tulajdonságai miatt az egyik legfontosabb modell), akkor az $m(\mathbf{x})$ regressziós függvény a $\mathbf{x}$ megfigyelés lineáris függvénye. Ebben az esetben tehát elég a Bayes-becslést $\mathbf{x}$ lineáris függvényei között keresni. A lineáris függvények igen tömören tárolhatóak és könnyen kiértékelhetőek. Bár a Bayes-becslés általános (nem gaussi) esetben nem mindig lineáris függvény, a fentiek alapján érdemes ekkor is megvizsgálni azt megszorítást, hogy ha a becslésfüggvényt csak $\mathbf{x}$-nek a lineáris függvényei között keressük. Természetesen a

$$g(\mathbf{x}) = c_0 + \sum_{i=1}^{d} c_i x_i \qquad c_0, c_1, \ldots, c_d \in \mathbf{R}$$

lineáris becslésfüggvények közül a legkisebb kockázatút szeretnénk megtalálni. Az egyszerűbb jelölés kedvéért vezessünk be az $\mathbf{X}$ (ill. $\mathbf{x}$) vektor egy nulladik X_0 (ill. x_0) koordinátáját, amely azonosan egyenlő 1-gyel ($X_0 \equiv x_0 \equiv 1$). Az alábbiakban tehát $\mathbf{X}$,$\mathbf{x} \in \mathbf{R}^{d+1}$ Így $g(\mathbf{x}) = \sum_{i=0}^{d} c_i x_i$, és g globális kockázata

$$R(g) = \mathbf{E}\left[C_2(Y, g(\mathbf{X}))\right] = \mathbf{E}\left[\left(Y - \sum_{i=0}^{d} c_i X_i\right)^2\right] \stackrel{\text{def}}{=} R(c_0, c_1, \ldots, c_d).$$

Keressük tehát azokat a $(c_0^*, c_1^*, \ldots, c_d^*)$ együtthatókat, amelyekre

$$\bar{g}(\mathbf{x}) = \sum_{i=0}^{d} c_i^* x_i$$

a legkisebb négyzetes költséget adja, azaz amelyekre

$$R(c_0^*, c_1^*, \ldots, c_d^*) = \min_{(c_0,c_1,\ldots,c_d)\in\mathbf{R}^{d+1}} R(c_0, c_1, \ldots, c_d),$$

azaz minimalizálni akarjuk $R(c_0, c_1, \ldots, c_d)$-t a c_i együtthatók szerint. A $(c_0^*, c_1^*, \ldots, c_d^*)$ minimumhelyen a $R(c_0, c_1, \ldots, c_d)$-nek minden változója szerinti parciális deriváltja 0 kell legyen. Mivel $\mathbf{E}[cZ] = c\mathbf{E}[Z]$ bármely c konstansra és Z valószínűségi változóra, a várhatóérték lineáris operátor, ami alapján belátható, hogy a deriválással felcserélhető. Így

$$\frac{\partial}{\partial c_j} R(c_0, c_1, \ldots, c_d) = \mathbf{E}\left[\frac{\partial}{\partial c_j}\left(Y - \sum_{i=0}^{d} c_i X_i\right)^2\right] = \mathbf{E}\left[2\left(Y - \sum_{i=0}^{d} c_i X_i\right)(-X_j)\right] =$$
$$= 2\left(\sum_{i=0}^{d} c_i \mathbf{E}[X_i X_j] - \mathbf{E}[X_j Y]\right) \stackrel{\text{def}}{=} 2\left(\sum_{i=0}^{d} c_i s_{ij} - b_j\right),$$

ahol $s_{ij} = \mathbf{E}[X_i X_j]$, $b_j = \mathbf{E}[X_j Y]$ (i,$j = 0$,1,...,d). Következésképpen

$$\sum_{i=0}^{d} c_i^* s_{ij} = b_j, \qquad j = 0, 1, \ldots, d.$$

Térjünk át mátrixos jelölésre; legyen

$$\begin{aligned}
&\mathbf{b}^T = (b_0, b_1, \ldots, b_d) && \text{(sorvektor)},\\
&\mathbf{c}^{*T} = (c_0^*, c_1^*, \ldots, c_d^*) && \text{(sorvektor)},\\
&\mathbf{S} = [s_{ij}] && ((d+1)\times(d+1)\text{-es mátrix}).
\end{aligned}$$

(Ha az $\mathbf{X}$ oszlopvektort jelöl, transzponáltját pedig $\mathbf{X}^T$, akkor $\mathbf{S} = \mathbf{E}\left[\mathbf{X}\mathbf{X}^T\right]$ alakba is írható.) Ezekkel a fenti lineáris egyenletrendszer a

$$\mathbf{c}^{*T}\mathbf{S} = \mathbf{b}^T$$

mátrixegyenletként írható fel. Ha tehát az $\mathbf{S}$ invertálható, akkor az egyenletrendszer egyértelmű megoldása

$$\mathbf{c}^{*T} = \mathbf{b}^T\mathbf{S}^{-1}.$$

Ezek az együtthatók adják tehát az optimális lineáris regresszióbecslést az $\{s_{ij}\}$ és $\{b_j\}$ várhatóértékek függvényében.

7. *megjegyzés.* Ha a megfigyelés centrált, azaz $\mathbf{E}[X_i] = 0$ ($i = 1, \ldots, d$), akkor $\mathbf{S}$-nek az $(X_1, \ldots, X_d)$-hez tartozó főminorja éppen e vektornak a kovarianciamátrixa. ($\mathbf{S}$ pedig $\mathbf{X}$ kovarianciamátrixa azzal az eltéréssel, hogy $s_{00} = 1$, míg a kovarianciamátrixban itt 0 áll.)

8. *megjegyzés.* Vegyük észre, hogy $\mathbf{S}$ mindig pozitív szemidefinit, azaz minden $\mathbf{c} \in \mathbf{R}^{d+1}$-re $\mathbf{c}^T\mathbf{S}\mathbf{c} \geq 0$. Ugyanis

$$\sum_{i,j=0}^{d} c_i s_{ij} c_j = \sum_{i,j=0}^{d} c_i \mathbf{E}\left[X_i X_j\right] c_j = \mathbf{E}\left[\left(\sum_{i=0}^{d} c_i X_i\right)\left(\sum_{j=0}^{d} c_j X_j\right)\right] = \mathbf{E}\left[\left(\sum_{i=0}^{d} c_i X_i\right)^2\right] \geq 0.$$

Ismeretes, hogy egy pozitív szemidefinit mátrix pontosan akkor invertálható, ha pozitív definit, azaz ha a fenti egyenlőtlenségben nagyobb-egyenlőség helyett szigorú $>$ áll fenn minden $\mathbf{c} \in \mathbf{R}^{d+1} \setminus \{\mathbf{0}\}$-ra. Könnyű meggondolni, hogy ez akkor van így, ha a $(X_1, \ldots, X_d)$ vektor nem koncentrálódik $\mathbf{R}^d$ egyetlen d-nél kisebb dimenziós eltolt alterére sem.

Irodalomjegyzék

[1] T. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, NY, 1991.

[2] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, NY, 1996.

[3] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York, NY, 2002.

[4] T. Linder and G. Lugosi. *Bevezetés az Információelméletbe*. Tankönyvkiadó, Budapest, 1990. jegyzetszám: J5-1445.