Лекція 11. Розпізнавання на основі відстані між емпиричними функціями розподілу

Нехай $x=(x_1,x_2,...,x_n)$ і $x'=(x'_1,x'_2,...,x'_m)$ — дві вибірки з генеральних сукупностей G і G' відповідно, і $z=(z_1,z_2,...,z_k)$ - вибірка, що належить однієї з цих двох генеральних сукупностей. Вважатимемо, що усі вибірки отримані шляхом простого випадкового вибору. Необхідно ідентифікувати генеральну сукупність, з якої узята вибірка z.

Цю проблему можна розв'язати за допомогою класичних непараметричних критеріїв Колмогорова-Смірнова, Вілкоксона чи будь-якого іншого непараметричного двовибіркового критерію [110]. Однак при використанні цих критеріїв застосовуються односторонні довірчі межі, що відповідають заданому рівню значущості (наприклад, п'ятивідсотковому), що може призвести до невизначеності і неприйняття рішення. Дійсно, припустимо, що ми використовуємо критерій Колмогорова-Смірнова при порівнянні вибірки z з вибірками x і x'. Для цього обчислюються статистики Колмогорова-Смірнова $\rho(z,x)$ і $\rho(z,x')$, а потім знаходяться довірчі межи $t_\beta(k,n)$ і $t_\beta(k,m)$. Якщо $\rho(z,x)\ge t_\beta(k,n)$ і $\rho(z,y)<t_\beta(k,m)$, то вважається, что $z\in G'$; якщо ж $\rho(z,x)<t_\beta(k,n)$ і $\rho(z,y)\ge t_\beta(k,m)$, то $z\in G$; у випадку інших можливих нерівностей виникає невизначеність і жодне рішення не приймається. Іншим істотним недоліком односторонніх непараметричних критеріїв є той факт, що імовірність улучення статистики $\rho(z,x)$ в довірчий інтервал $(0,t_\beta(k,n))$ (тобто його довірчий рівень) можна точно визначити лише за умови, коли $z\in G$ (гіпотеза H); у протилежному випадку ця імовірність може приймати будь-яке значення від 0 до 1 і є невідомою. У зв'язку з цим навіть у випадку ухвалення рішення неможливо оцінити імовірності помилок 2-го роду, що виникають при використанні цього критерію. Ця ситуація характерна для будь-яких односторонніх непараметричних критеріїв.

Як відомо, генеральні сукупності розділяються на два класи – гомогенні і гетерогенні. Наприклад, у хіміотерапії і радіотерапії злоякісних новотворів вважається, що до початку лікування популяція пухлинних клітин є гомогенною і складається з ракових кліток, чуттєвих до цитостатиків чи рентгенівського опромінення. Однак після проведення курсу лікування пухлина являє собою гетерогенну генеральну сукупність, у якій крім чуттєвих клітин з'являються так звані резистентні клітини. Основна задача стратифікаційного аналізу гетерогенних генеральних сукупностей полягає у визначенні кількості модальних класів, а також їхніх показників.

Теорема Глівенка–Кантеллі, одночасно доведена авторами незалежно один від одного в 1933 році, є однією з найважливіших теорем в математичній статистиці [1, 2]. Вона стверджує, що коли кількість спостережень прямує до нескінченності, емпірична функція розподілу прямує до дійсної функції розподілу за ймовірністю в чебишовській метриці. Саме в цій формі теорема Глівенка–Кантеллі широко використовується в статистичній теорії розпізнавання образів [3, 4]. Але при аналізі неоднорідності генеральних сукупностей зручніше використовувати поняття кусково-лінійної емпіричної функції розподілу і функції, що є оберненою до неї. Це дозволяє розробити дуже простий метод, не пов'язаний із складними обчисленнями.

2. Кусково-лінійна емпірична функція розподілу. Нехай G – генеральна сукупність з неперервною строго зростаючою функцією розподілу $F(u)$. Тоді $F(u)$ має обернену функцію розподілу $F^{-1}(u)$. Поняття оберненої функції розподілу можна узагальнити, якщо функція $F(u)$ є неперервною і строго зростаючою не на всій числовій осі, а на своєму носії– відрізку $[a,b]$. Нагадаємо, що носієм функції розподілу $F(u)$ називається мінімальний відрізок $[a,b]$, для якого $F(u)\equiv 0\ \forall u\le a$ і $F(u)\equiv 1\ \forall u\ge b$. У цьому випадку обернена функція розподілу визначається як функція, що є оберненою до звуження $F_{[a,b]}(u)$ на свій носій $[a,b]$.

Гіпотетична функція розподілу $F(u)$ генеральної сукупності G, як і обернена до неї функція, як правило, є невідомими, однак після проведення випадкового експерименту ми маємо вибірку $x_1,x_2,...,x_n$, на підставі якої можна одержати оцінки функцій $F(u)$ і $F^{-1}(u)$. Опишемо спочатку оцінку для функції розподілу. Розглянемо варіаційний ряд $x_{(1)}\le x_{(2)}\le...\le x_{(n)}$, побудований по вибірці $x_1,x_2,...,x_n$ і визначимо кусково-постійну емпіричну функцію розподілу $F_n^*(u)$. За теоремою Глівенко-Кантеллі [1,2]

$$p\left(\lim_{n\to\infty}\max_u\left|F(u)-F_n^*(u)\right|=0\right)=1.$$

На жаль, у кусково-постійної емпіричної функції розподілу немає оберненої функції (у її звичайному розумінні), тому її неможливо використовувати для оцінки оберненої функції розподілу. У зв'язку з цим виникає проблема побудови емпіричної функції розподілу, у якого була б обернена функція.

Для неперервної монотонно зростаючої функції розподілу $F(u)$ з компактним носієм $[a,b]$ побудуємо наступний кусково-лінійний сплайн.

$$\tilde{F}_n(u)=\begin{cases}\dfrac{u}{n\left(x_{(k+1)}-x_{(k)}\right)}+\dfrac{kx_{(k+1)}-(k+1)x_{(k)}}{n\left(x_{(k+1)}-x_{(k)}\right)}, \text{ якщо } x_{(k)}\le u\le x_{(k+1)},\ x_{(0)}=a,\ k=0,1,...,n;\\ 1,\ \text{якщо } x_{(n)}\le u\le b.\end{cases}$$

Кусково-лінійна емпірична функція розподілу $F_n^*(u)$, що визначена на відрізку $[a,b]$, має на відрізку $\left[a,x_{(n)}\right]$ обернену функцію $\left(F_n^*(u)\right)^{-1}$, $0\le u\le 1$. Для скорочення запису надалі будемо позначати функцію $\left(F_n^*(u)\right)^{-1}$ як $\Psi_n^*(u)$. Ця функція також являє собою кусково-лінійний сплайн із вершинами в точках $\left(\dfrac{k}{n},x_{(k)}\right)$, $k=0,1,...,n$. Інакше кажучи, функція $\Psi_n^*(u)$ є неперервною кусково-лінійною функцією, що проходить через точки $\left(\dfrac{k}{n},x_{(k)}\right)$, де $x_{(k)}$ — порядкові статистики. Слід зазначити, що інші кусково-лінійні емпіричні функції розподілу, що відрізняються від $\tilde{F}_n^*(u)$, мають обернені функції, що цією властивістю не володіють.

Неважко помітити, що $\forall u\in R^1$

$$\left|F_n^*(u)-\tilde{F}_n^*(u)\right|<\frac{1}{n},$$

тому $\tilde{F}_n^*(u)$ є слушною оцінкою для гіпотетичної функції розподілу, причому теорема Кантеллі-Глівенко залишається справедливою.

Зауваження 1. У випадку, якщо вибірка $x_1,x_2,...,x_n$ належить нормально розподіленій сукупності, функція, обернена до кусково-лінійної емпіричної функції розподілу, називається кривої Кетле [3].

Зауваження 2. Комп'ютерне моделювання показує, що для нормальних розподілів лінійна регресія краще апроксимує обернену емпіричну кусково-лінійну функцію, ніж пряму. У цьому випадку стратифікаційний аналіз гетерогенних генеральних сукупностей має більш високу точність.

Лема. Якщо гіпотетична функція розподілу $F(u)$ генеральної сукупності є неперервною і має компактний носій $[a,b]$, то послідовність кусково-лінійних емпіричних функцій розподілу з імовірністю одиниця збігається до функції розподілу в чебышовській метриці:

$$p\left(\lim_{n\to\infty}\max_u\left|F(u)-\tilde{F}_n^*(u)\right|=0\right)=1.$$

Доведення. Дійсно,

$$\left|F_n^*(u)-\tilde{F}_n^*(u)\right|\le\left|F(u)-F_n^*(u)\right|+\left|F_n^*(u)-\tilde{F}_n^*(u)\right|\le\left|F(u)-F_n^*(u)\right|+\frac{1}{n}.$$

Отже,

$$\max_u\left|F_n^*(u)-\tilde{F}_n^*(u)\right|\le\max_u\left|F(u)-F_n^*(u)\right|+\frac{1}{n}.$$

Легко бачити, що

$$\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-\tilde{F}_n(u,\omega)\right|=0\right\}\supset\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-F_n^*(u,\omega)\right|=0\right\},$$

$$\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-F_n^*(u,\omega)\right|+\frac{1}{n}=0\right\}\supset\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-F_n^*(u,\omega)\right|=0\right\},$$

де ω – елементарний наслідок.

Таким чином,

$$p\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-\tilde{F}_n^*(u,\omega)\right|=0\right\}\ge p\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-F_n^*(u,\omega)\right|=0\right\}=$$

$$=p\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-\tilde{F}_n^*(u,\omega)\right|=0\right\}=1.$$

Отже,

$$p\left\{\omega:\lim_{n\to0}\max_u\left|F(u)-\tilde{F}_n^*(u,\omega)\right|=0\right\}=1.$$

Лему доведено.

Теорема. Якщо гіпотетична функція розподілу $F(u)$ генеральної сукупності G є неперервною і має компактний носій $[a,b]$, то послідовність обернених функцій $\left\{\Psi_n^*(u)\right\}$ збігається з імовірністю одиниця до функції $F^{-1}(u)$, тобто

$$p\left\{\lim_{n\to\infty}\max_{a\le u\le b}\left|F^{-1}(u)-\Psi_n^*(u)\right|=0\right\}=1. \qquad (1)$$

Доведення. Формула (1) означає, що імовірність події, яка полягає в тім, що послідовність $\left\{\Psi_n^*(u)\right\}$ рівномірно збігається до $F^{-1}(u)$, дорівнює одиниці. За лемою 1

$$p\left\{\lim_{n\to\infty}\max_{a\le u\le b}\left|F(u)-\tilde{F}_n^*(u)\right|=0\right\}=1\ . \qquad (2)$$

Розглянемо елементарний результат випадкового віпробування, при реалізації якого послідовність $\left\{\tilde{F}_n^*(u)\right\}$ рівномірно збігається до $F(u)$, коли $n\to\infty$. Покажемо, що при цьому послідовність $\left\{\Psi_n^*(u)\right\}$ рівномірно збігається до $F^{-1}(u)$, коли $n\to\infty$. Для цього достатньо показати, що при кожнім $\varepsilon>0$ у ε-околі функції $F(u)$ в чебишовскій метриці $N_\varepsilon(F)=\left\{\Psi(u):\max_{a\le u\le b}\left|F(u)-\Psi(u)\right|<\varepsilon,\Psi\in C[a,b]\right\}$ при якомусь $\delta>0$ міститься так звана "циліндричний" δ-окіл $C_\delta(F)$ функції $F(u)$. Інакше кажучи,

$$\forall\varepsilon>0\ \exists\delta>0:\ C_\delta(F)\subset N_\varepsilon(F),$$

де

$$C_\delta(F)=\left\{\Psi(u):(u,\Psi(u))\in\underset{u\in[a,b]}{U}S\left((u,F(u),\delta)\right)\right\},$$

$$S\left((u,F(u)),\delta\right)=\left\{(v,y):\rho\left((u,F(u)),(v,y)\right)<\delta\right\}.$$

Неважко помітити, що останнє твердження буде доведено, якщо ми покажемо, що в смузі $\Pi_\varepsilon=\left\{(u,y):\left|F(u)-y\right|<\varepsilon,u\in[a,b]\right\}$ при якомусь $\delta>0$ міститься "циліндрична смуга"

$$\Pi_\delta^C=\left\{(u,y):\rho\left((u,y),(u,F(u))\right)<\delta,u\in[a,b]\times(0,1)\right\}$$

Перш за все доведемо, що смуга Π_ε є відкритою множиною на площині. Нехай $\Pi_\varepsilon^+=\left\{(u,y):y<F(u)+\varepsilon,u\in(a,b)\right\}$, $\Pi_\varepsilon^-=\left\{(u,y):y>F(u)-\varepsilon,u\in(a,b)\right\}$. Розглянемо неперервну функцію $\varphi(u,y)=y-F(u)$. Тоді $\Pi_\varepsilon^+=\left\{(u,y):\varphi(u,y)<\varepsilon\right\}$. Оскільки функція $\varphi(u,y)$ є неперервною, множини Π_ε^- і Π_ε^+ є відкритими в прямокутнику $a<u<b$, $0<y<1$, а тому і на площині. Отже, смуга $\Pi_\varepsilon=\Pi_\varepsilon^+\cap\Pi_\varepsilon^-$ також є відкритою в R^2. Звідси випливає, що для будь-якої точки $(u,F(u))$, що належить графіку функції $F(u)$, існує таке коло радіуса δ з центром у цій точці, що перетин цього кола з прямокутником $(a,b)\times(0,1)$ міститься в смузі Π_ε. Зрозуміло, радіус δ може залежати від точки $(u,F(u))$.

Покажемо існування числа $\delta_0>0$ такого, що $\delta>\delta_0$ для будь-якої точки $(u,F(u))$. Дійсно, припустимо супротивне. Тоді існує послідовність точок $A_1,A_2,...,A_n,...$, що належать межі Γ_1 розглянутої смуги, що збігається до точки $A_0\in\Gamma$. Позначимо через A_n' точку, що лежить на графіку Γ функції $F(u)$, абсциса u_n якої співпадає с абсцисою точки A_n. З огляду на компактність прямокутника $[a,b]\times[0,1]$, без обмеження загальності можна вважати, що $\lim_{n\to\infty}A_n'=\tilde{A}_0$.

Покажемо, що $A_n'=A_0$. Має місце наступний ланцюжок тверджень.

$\lim_{n\to\infty}A_n=A_0\Rightarrow\lim_{n\to\infty}u_n=u_0\Rightarrow\lim_{n\to\infty}F(u_n)=F(u_0)\ \Rightarrow$

$\Rightarrow\lim_{n\to\infty}A_n'=\lim_{n\to\infty}(u_n,F(u_n))=(u_0,F(u_0))=A_0\Rightarrow A_0=\tilde{A}_0$.

Проте $\lim_{n\to\infty}A_n=\lim_{n\to\infty}(u_n,y_n')=\lim_{n\to\infty}(u_n,y_n+\varepsilon)=(u_0,y_0+\varepsilon)\ne A_0$. Отримане протиріччя доводить існування шуканого числа δ_0. Отже, послідовність функцій $\left\{\Psi_n^*(u)\right\}$ з імовірністю одиниця збігається до функції $F^{-1}(u)$. Теорему доведено.

1. Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita. Giorn. Ist.Ital. Attuari 4, 421 - 424. 2. Glivenko, V. (1933). Sulla determinazione empirica della legge di probabilita. Giorn. Ist.Ital. Attuari 4, 92-99. 3. Vapnik, V. N. *Statistical Learning Theory*. Wiley, 1998. 4. Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer, 2000. 4. Ван дер Варден Б.Л. *Математическая статистика*. — М.: Изд-во иностранной литературы, 1960.