LOGARITMICKO LINEÁRNÍ MODELY
PRO POSLOUPNOSTI ZÁVISLÝCH BINÁRNÍCH DAT

Martin Janžura, ÚTIA ČSAV, Praha

S posloupnostmi binárních dat se lze setkat všude tam, kde jsou v pravidelných časových intervalech opakovaně pozorovány veličiny nabývající pouze dvě hodnoty (např. přítomnost - nepřítomnost příznaku, překročení - nepřekročení kritické úrovně apod.). Tyto hodnoty mohou být zcela libovolné, pro jednoduchost je účelné pracovat s dvoubodovou množinou {0,1}.

Při statistickém zpracování takových posloupností dat se obvykle využívá teorie markovských řetězců (libovolného řádu). Pravděpodobnosti přechodu však neposkytují dostatečně jemný pohled na skutečnou závislostní strukturu a současně se nesnadno odhadují, je-li rozsah výběru nedostatečný vzhledem k řádu markovosti.

Z těchto důvodů půjdeme cestou zobecnění teorie logaritmicko lineárních modelů, které se osvědčily pro popis simultánní distribuce konečného systému náhodných veličin (viz Bishop, Fienberg a Holland (1975)). Přirozeným zobecněním logaritmicko lineárních modelů pro náhodné posloupnosti obdržíme modely známé v oblasti statistické fyziky pod pojmem gibbsovské náhodné posloupnsti.

1. Gibbsovské náhodné posloupnosti

Vyjdeme od definice logaritmicko lineárního modelu pro simultánní rozdělení konečného systému binárních náhodných veličin.

Pro libovolnou množinu V označme $k(V) = \{A \subset V : 0 < |A| < \infty\}$ systém všem konečných neprázdných podmnožin (symbolem $|A|$ značíme kardinalitu množiny A).

Předpokládejme nejprve, že množina V je konečná, tedy $|V| < \infty$. Pravděpodobnostní míra μ^U definovaná na $\{0,1\}^V$ vyhovuje logaritmicko lineárnímu modelu, jestliže existuje systém reálných konstant

$$U = \{U_A\}_{A \in k(V)}$$

tak, že platí

$$\log \mu^U(x_V) = \sum_{A \in k(V)} U_A \prod_{s \in A} x_s - q(U) \tag{1}$$

$$\text{pro každé } x_V \in \{0,1\}^V.$$

kde

$$q(U) = \log \sum_{y_V \in \{0,1\}^V} \exp \left\{ \sum_{A \in k(V)} U_A \prod_{s \in A} y_s \right\}$$

je logaritmus příslušné normovací konstanty.

Ekvivalentní definice má tvar

$$\log \mu^U\left(x_{\{t\}} \mid x_{V\setminus\{t\}}\right) = \tag{2}$$

$$= \sum_{A\in k(V),\, A\ni t} U_A \prod_{s\in A} x_s - \log\left[1+\exp\left\{\sum_{A\in k(V),\, A\ni t} U_A \prod_{s\in A\setminus\{t\}} x_s\right\}\right]$$

pro každé $t \in V$, $x_V \in \{0,1\}^V$.

Hodnoty U_A, $A \in k(V)$ se nazývají interakcemi, celý systém $U = \{U_A\}_{A\in k(V)}$ se ve statistické fyzice nazývá potenciál.

Samotný model je pak určen systémem nenulových interakcí. Pro $\mathcal{A} \subset k(V)$ budeme psát $U \in M_{\mathcal{A}}$ jestliže $U_A = 0$ pro každé $A \in k(V) \setminus \mathcal{A}$. Přiřazení $\mathcal{A} \to M_{\mathcal{A}}$ je zřejmě monotonní, tj. pro $\mathcal{B} \subset \mathcal{A}$ platí $M_{\mathcal{B}} \subset M_{\mathcal{A}}$.

<u>Poznámka</u>: V obecné definici logaritmicko lineárního modelu se uvažují libovolné funkce $U_A(x_A)$ namísto zde uvedeného speciálního tvaru $U_A \prod_{s\in A} x_s$. Pak je ale třeba v zájmu jednoznačnosti doplnit definici o nějaké normující vazby, aby zůstalo kýžených $2^{|V|}-1$ volných parametrů. Tomu se zde užitím speciálního tvaru funkcí vyhneme při zachování plné obecnosti, tj. pro každou kladnou pravděpodobnostní míru μ na $\{0,1\}^V$ existuje výše uvedené vyjádření ve formě logaritmicko lineárního modelu.

Nahradíme nyní konečnou množinu V prostorem celých čísel Z a budeme uvažovat pouze omezené potenciály

$$U = \{U_A\}_{A\in k(Z)},$$

tj. $\sum_{A\in k(Z),\, A\ni t} |U_A| < \infty$ pro každé $t \in Z$.

Řekneme, že distribuce μ^U na $\{0,1\}^Z$ je gibbsovská vzhledem k potenciálu U, jestliže definice (2) platí pro $V = Z$ (zde jsme potřebovali předpoklad omezenosti, aby nekonečné součty konvergovaly).

Definici (1) nelze přímo použít, neboť pro $V = Z$ by neměla dobrý smysl, a kdybychom tímto způsobem definovali systém konečně rozměrných distribucí $\{\mu_V^U\}_{V\in k(Z)}$, nebyl by tento systém nutně konzistentní.

Lze pouze definici rozšířit na systém všech konečně rozměrných podmíněných distribucí

$$\{\mu^U_{V|Z\setminus V}\}_{V\in k(Z)}$$

předpisem

$$\log \mu^U_{V|Z\setminus V}\left(x_V \mid x_{Z\setminus V}\right) = \sum_{A\in k(Z),\, A\cap V\neq\emptyset} U_A \prod_{s\in A} x_s - q_V^U\left(x_{Z\setminus V}\right)$$

pro každé $V \in k(Z)$, kde

$$q_V^U\left(x_{Z\setminus V}\right) =$$

$$= \log \sum_{y_V\in\{0,1\}^V} \exp\left\{\sum_{A\in k(V),\, A\cap V\neq\emptyset} U_A \prod_{s\in A\cap V} y_s \prod_{s\in A\cap(Z\setminus V)} x_s\right\}$$

je opět logaritmus příslušné normovací konstanty.

V dalším se budeme zabývat pouze stacionárními potenciály, tj. $U_A = U_{A+t}$ pro každé $A \in k(Z)$, $t \in Z$. Označíme $k_+(Z) = \{A \in k(Z) : \min\{j \in A\} = 0\}$. Model je zde opět určen systémem nenulových interakcí, tj. pro $\mathcal{A} \subset k_+(Z)$ píšeme $U \in M_{\mathcal{A}}$ jestliže $U_A = 0$ pro $A \in k_+(Z) \setminus \mathcal{A}$.

Jestliže $|\mathcal{A}| < \infty$, určuje systém podmíněných distribucí $\left\{\mu^U_{V|Z\setminus V}\right\}_{V \in k(Z)}$ (který je ovšem jednoznačně dán systémem $\left\{\mu^U_{\{t\}|Z\setminus\{t\}}\right\}_{t \in Z}$ a jelikož je potenciál stacionární, tak dokonce stačí $\mu^U_{\{0\}|Z\setminus\{0\}}$) jednoznačným způsobem gibbsovskou distribuci μ^U na $\{0,1\}^Z$. Tato distribuce je pak stacionární a ergodická. (To pro $|\mathcal{A}| = \infty$ nemusí být pravda. Nejen že není zaručena jednoznačnost, ale může existovat i nestacionární distribuce.)

Označme $[n] = \{1,\dots,n\}$ pro každé $n \in Z$. Pak platí

$$\lim_{n\to\infty} \frac{1}{n} q_U^{[n]}\left(x_{Z\setminus[n]}\right) = q(U)$$

stejnoměrně v $x \in \{0,1\}^Z$ pro každý omezený stacionární potenciál U.

Nechť $|\mathcal{A}| < \infty$. Potom q zúženou na $M_{\mathcal{A}}$ lze chápat jako silně konvexní reálnou analytickou funkci na $R^{\mathcal{A}}$. Označíme-li

$$C_A = \left\{x \in \{0,1\}^Z : \prod_{i \in A} x_i = 1\right\} \quad \text{pro } A \in k(Z),$$

potom zejména platí

$$\nabla q(U) = \left\{\frac{\partial q}{\partial U_A}(U)\right\}_{A \in \mathcal{A}} = \left\{\mu^U(C_A)\right\}_{A \in \mathcal{A}}$$

a

$$\nabla^2 q(U) =$$

$$= \left\{\frac{\partial^2 q}{\partial U_A \partial U_B}(U)\right\}_{A,B \in \mathcal{A}} = \left\{\sum_{k=-\infty}^{\infty}\left[\mu^U\left(C_A \cap T^{-k} C_B\right) - \mu^U(C_A)\mu^U(C_B)\right]\right\}_{A,B \in \mathcal{A}}$$

kde $T : \{0,1\}^Z \to \{0,1\}^Z$ je posunutí definované předpisem $T(x)_n = x_{n+1}$ pro každé $x \in \{0,1\}^Z$, $n \in Z$. Přitom $\nabla q(U)$ i $\nabla^2 q(U)$ jsou spojité jako funkce $U \in M_{\mathcal{A}}$ a $\nabla^2 q(U) > 0$ pro každé pevné $U \in M_{\mathcal{A}}$.

Platí také

$$q(U) = r^{-1} \log \lambda_{max}(Q_U),$$

kde $r = r(\mathcal{A}) = \max_{A \in \mathcal{A}} \max\{j \in A\}$ a Q_U je tzv. transfer matice definovaná předpisem

$$Q_U\left(x_{[r]}, z_{[r]}\right) =$$

$$= \exp\left\{\sum_{A \in \mathcal{A}} U_A \sum_{t=1}^{r} \prod_{k \in (A+t)\cap[r]} x_k \cdot \prod_{k \in (A+t-r)\cap[r]} z_k\right\}$$

pro každé $x_{[r]}$, $z_{[r]} \in \{0,1\}^{[r]}$. Matice Q_U je čtvercová, dimenze $2^r \times 2^r$ s kladnými elementy, je tedy jednoznačně dáno její v absolutní hodnotě největší vlastní číslo $\lambda_{max}(Q_U)$, které je reálné kladné. Tímto způsobem se v praxi hodnoty funkce q vyčíslují.

Většinu výsledků týkajících se gibbsovských náhodných posloupností lze nalézt v klasické monografii Ruelle (1969).

2. Odhady interakcí a testování submodelu

Předpokládáme, že pozorovaná posloupnost binárních dat

$$x_{[N]} = (x_1, \dots, x_N) \in \{0,1\}^N$$

je generována nějakou gibbsovskou distribucí μ^{U^0} s neznámým potenciálem $U^0 \in M_{\mathcal{A}}$, $\mathcal{A} \subset k_+(V)$, $|\mathcal{A}| < \infty$. Naším prvním úkolem je odhadnout tento potenciál chápaný jako vektorový parametr.

Pro $U \in M_{\mathcal{A}}$ a $\theta \in R^{\mathcal{A}}$ označme

$$F_{\mathcal{A}}(U,\theta) = q(U) - \sum_{A \in \mathcal{A}} U_A \theta_A .$$

Z vlastností funkce q na $M_{\mathcal{A}}$ plyne

i) $U^1 = U^2$ právě když $\nabla q(U^1) = \nabla q(U^2)$

ii) $F_{\mathcal{A}}(U^0,\theta^0) = \min_{U \in M_{\mathcal{A}}} F_{\mathcal{A}}(U,\theta^0)$ právě když $\nabla q(U^0) = \theta^0$.

Odhad U^N parametru U^0 tedy získáme minimalizací funkce $F_{\mathcal{A}}(U,\theta^N)$, kde za θ^N vezmeme odhad vektoru $\nabla q(U^0) = \left\{\mu^{U^0}(C_A)\right\}_{A \in \mathcal{A}}$, tedy

$$\theta_A^N = (N-r(A))^{-1} \sum_{i=1}^{N-r(A)} \prod_{j \in A} x_{j+i} \quad \text{pro každé } A \in \mathcal{A}$$

(zde $r(A) = \max\{k \in A\}$). Pokud minimum neexistuje, dodefinujeme U^N libovolně.

Věta: Odhad U^N je konzistentní a asymptoticky normální, tj.

$$U^N \to U^0 \quad \text{s.j.} \ \left[\mu^{U^0}\right]$$

a

$$N^{\frac{1}{2}}\left(U^N - U^0\right) \Rightarrow \mathcal{N}\left(0, (\nabla^2 q(U^0))^{-1}\right) \text{ v distribuci } \left[\mu^{U^0}\right].$$

Důkaz: Věta 3.2, Janžura (1986).

Řekněme, že máme podezření, že bychom mohli napozorovaná data vysvětlit pomocí jednoduššího modelu $\mathcal{B} \subset \mathcal{A}$. Budeme tedy testovat hypotézu

$$H^0 : U^0 \in M_{\mathcal{B}}$$

proti alternativě

$$H^1 : U^0 \in M_{\mathcal{A}} \setminus M_{\mathcal{B}}.$$

Zaveďme statistiku

$$G^N = \min_{U \in M_{\mathcal{B}}} F_{\mathcal{A}}(U,\theta^N) - \min_{U \in M_{\mathcal{A}}} F_{\mathcal{A}}(U,\theta^N),$$

která je definovaná, pokud existuje $U_{\mathcal{A}}^N \in M_{\mathcal{A}}$, v němž nabývá minima funkce $F_{\mathcal{A}}(\cdot|\theta^N)$. Potom se totiž v nějakém $U_{\mathcal{B}}^N \in M_{\mathcal{B}}$ nabude i minima funkce $F_{\mathcal{A}}(\cdot|\theta^N)$ zúžené na $M_{\mathcal{B}}$.

Můžeme také psát

$$G^N = H\left(\mu^{U_A^N} \middle| \mu^{U_B^N}\right),$$

kde $H(.|.)$ je I-divergence dána pro obecné μ, ν výrazem

$$H(\mu|\nu) = \lim_{n\to\infty} n^{-1} \int \log \frac{\mu(x_{[n]})}{\nu(x_{[n]})}\, d\mu(x),$$

pokud jsou integrály definovány a limita existuje (pro gibbsovské μ, ν je $H(\mu|\nu)$ konečná).

Pokud se minima nenabývá, můžeme G^N dodefinovat libovolně.

Věta: Pro každý $U^0 \in M_B$ platí

$$2 \,.\, N \,.\, G^N \to \chi^2_f \text{ v distribuci } \left[\mu^{U^0}\right],$$

kde $f = |A \setminus B|$.

Důkaz: Věta 4.2, Janžura (1988b).

Hypotézu tedy zamítáme, jestliže statistika $2.N.G^N$ přesáhne příslušný kvantil rozdělení χ^2 s $|A \setminus B|$ stupni volnosti.

3. Použití

Ve statistických úlohách se systém A (určující model) považuje za známý, přičemž teoreticky může být zcela libovolný konečný. V praxi však existují dvě závažná omezení. V první řadě je zde omezení na "řád" $r = r(A)$, které je dáno výpočetními možnostmi, neboť je třeba opakovaně počítat největší vlastní číslo transfer matice Q_U, která je řádu $2^r \times 2^r$. Je tudíž třeba předpokládat zhruba $r < 10$.

Další omezení je na kardinalitu množin v systému A, tedy $C(A) = \max_{A\in A} |A|$. Toto omezení je dáno rozsahem výběru N. Jestliže by totiž nebyl rozsah výběru N dostatečně velký vzhledem k $2^{|A|}$ možných konfigurací $x_A \in \{0,1\}^A$, byl by odhad θ^N_A pravděpodobně velmi nepřesný. Zde se ukazuje účelné (a v mnoha praktických úlohách také zcela postačující) uvažovat pouze párové interakce, tj. $A = \left\{A_t = \{0,t\}\right\}_{t=0,\dots,r}$.

Potom již můžeme $\theta^N_{A_t}$ odhadnout s rozumnou přesností i při poměrně malém rozsahu výběru N a bez dalšího omezení na r. Jedná se pak vlastně o odhad jakýchsi "kovariancí" $\mu^{U^0}(x_0=1, x_t=1)$.

Uvažujme například úlohu predikce. Máme napozorovaná data $(x_1,\dots,x_N) \in \{0,1\}^N$ a předpokládáme, že generující náhodná posloupnost μ je markovská řádu R. Pro optimální predikci (ve smyslu minimální pravděpodobnosti chyby) je třeba umět vyjádřit podíl

$$L_{\mu}\left(x_{[R]}\right) = \frac{\mu(1|x_{[R]})}{\mu(0|x_{[R]})} \quad \text{pro každé } x_{[R]} \in \{0,1\}^R.$$

Věta: Pro každé $U \in M_{\mathcal{A}}$, $r(\mathcal{A}) = R$, platí

$$L_{\mu^U}\left(x_{[R]}\right) = \left[\lambda_{max}(Q_U)\right]^{\frac{1}{R}} \frac{r_U(x_{[R]})}{r_U(\vec{x}_{[R]})} - 1,$$

kde r_U je levý vlastní vektor příslušný vlastnímu číslu $\lambda_{max}(Q_U)$ a konfigurace $\vec{x}_{[R]}$ je dána tak, že $\vec{x}_1 = 0$, $\vec{x}_i = x_{i-1}$ pro $i = 2,\dots,R$.

Důkaz: Důsledek věty 1, Janžura (1988a).

Jestliže tedy N je "malé" ve srovnání s 2^{R+1}, budeme postupovat takto:

1. Předpokládáme, že μ je gibbsovské s nejvýše párovými interakcemi, tj. $\mu = \mu^U$, $U \in M_{\mathcal{A}_R}$, $\mathcal{A}_R = \left\{\{0,t\}\right\}_{t=0,\dots,R}$.
2. Odhadneme U postupem uvedeným v části 2.
3. Spočteme $L = L_{\mu^U}(x_N,\dots,x_{N-R+1})$. Jestliže $L > 1$, predikujeme $x_{N+1} = 1$.

Tato metoda funguje poměrně spolehlivě. Jsou-li předpoklady o párových interakcích alespoň zhruba splněny, jsou získané výsledky zjevně lepší, než kdybychom se pokoušeli z dat odhadovat přímo přechodové pravděpodobnosti $\mu(1|x_{[R]})$ (viz Janžura (1988a)).

Literatura

Bishop, Y. M. M., Fienberg, S. E., Holland, P. W. (1975) Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, Mass.

Janžura, M. (1986) Estimating interactions in binary data sequences. Kybernetika 22, 277-284.

Janžura, M. (1988a) Prediction in zero-one random sequences. Problems of Control and Information Theory, Vol. 17(1), 15-22.

Janžura, M. (1988b) Test for submodel in Gibbs-Markov binary random sequences (v rukopisu).

Ruelle, D. (1969) Statistical Mechanics. Rigorous Results. Benjamin, New York.