P(izraz|gramatika)

Urh Primožič

Simbolna regresija je področje strojnega učenja, ki se ukvarja z iskanjem funkcij, ki se prilegajo danim podatkom in imajo čim bolj preprost predpis. Novi algoritmi simbolne regresije za tvorjenje predpisov uporabljajo verjetnostne gramatike. Računamo lahko, s kakšno verjetnostjo dana gramatika tvori izbrano enačbo. Izkaže se, da je računanje verjetnosti neizračunljiv problem, za nekatere družine gramatik pa obstajajo matematično zanimive rešitve.

Verjetnostne gramatike

Naj bosta N in T poljubni disjunktni, neprazni, končni množici simbolov in naj bo v množici N odlikovan začetni simbol $S \in N$. Naj bo $R \subseteq N \times (N \cup T)^*$ celovita relacija, ki vsakemu simbolu iz N priredi vsaj eno besedo, sestavljeno iz simbolov v $N \cup T$. Tedaj četvercu $G = (N, T, S, R)$ rečemo kontekstno neodvisna gramatika. Simbolom iz N rečemo ne-končni, iz T pa končni simboli. Urejene pare iz R imenujemo prepisovalna pravila. Par $(A, w) \in R$ pišemo kot $A \to w \in R$ in rečemo, da se A prepiše v w.

Z uporabo prepisovalnih pravil iz R lahko v končnem številu korakov začetni simbol S prepišemo v besedo $w \in T^*$, sestavljeno le iz končnih simbolov. V tem primeru rečemo, da gramatika G tvori w. Množico vseh besed, ki jih tvori gramatika G, označimo z $L(G)$.

Verjetnostna gramatika je gramatika skupaj s porazdelitvijo $P\colon R \to (0, 1]$, da za fiksen $A \in N$ velja $\sum_{A \to w \in R} P(A \to w) = 1$. Na tvorbo besed z verjetnostno gramatiko lahko gledamo kot na markovski proces in računamo verjetnost tvorbe.

Množico $\{A \to w_1, \ldots, A \to w_n\}$ prepisovalnih pravil za A ob dani porazdelitvi P običajno podamo kot $A \to w_1 \ \ [P(A \to w_1)] \mid \cdots \mid w_n \ \ [P(A \to w_n)]$.

Primer $G = (\{S\}, \{x, +\}, S, R)$ s prepisovalnima praviloma $S \to S + x \ \ [p] \mid x \ \ [1-p]$ tvori množico besed $L(G) = \{x, x+x, x+x+x, \ldots\}$. Besedo $\underbrace{x + \cdots + x}_{n\text{-krat}}$ krajše pišemo kot nx. Verjetnost, da tvorimo besedo nx, je $P(nx) = (1-p)p^{n-1}$.

Uporaba v simbolni regresiji

Algoritem za odkrivanje enačb ProGED [1] z gramatikami tvori različne oblike predpisov za izraze oblike $f(x_1, \ldots, x_n, c)$, nato pa vsako pojavitev simbola c v predpisu z numerično optimizacijo nadomesti s tako konstanto, da se izraz čim bolj prilega podatkom.

Primer

$$\begin{aligned} E &\to cV + E \mid c \\ V &\to xV \mid x \end{aligned}$$

Zgornja gramatika tvori polinome s konstantami oblike $cx^{r_1} + \cdots + cx^{r_m} + c$. Algoritem izmed množice funkcij $\{c_0 x^{r_1} + \cdots + c_{m-1} x^{r_m} + c_m \mid c_i \in \mathbb{C}\}$

izbere tisto, ki se podatkom najbolj prilega. Z vidika simbolne regresije lahko predpis, tvorjen z gramatiko, enačimo z množicami vseh funkcij, ki jih dobimo iz predpisa, če simbole c nadomestimo s števili.

Formalizacija predpisa

Zgornjo intuicijo povzamemo v formalni definiciji izraza. Naj bo $\mathbb{F}$ poljubna množica dovoljenih vrednosti konstant c, D pa poljubna domena za spremenljivke $x_1, \ldots, x_n$. Naj bo G verjetnostna gramatika, ki tvori le besede oblike $w = w_1 c w_2 c \cdots c w_{m+1}$ (kjer v členih w_i ni znaka c) in za vsako izbiro konstant $c_1, \ldots, c_m \in \mathbb{F}$ beseda $w_1 c_1 w_2 c_2 \cdots c_m w_{m+1}$ predstavlja predpis za funkcijo

$$\begin{aligned} \phi_w(c_1, \ldots, c_m) \colon U &\longrightarrow D \\ (x_1, \ldots, x_n) &\longmapsto w_1 c_1 w_2 c_2 \cdots c_m w_{m+1}, \end{aligned}$$

kjer je $U \subseteq D^n$ maksimalna domena, da je $\phi_w(c_1, \ldots, c_m)$ še dobro definirana. Če je $D \in \{\mathbb{R}, \mathbb{C}\}$ dodatno privzamemo, da $\phi_w(c_1, \ldots, c_m)$ zvezno razširimo na vse robne točke domene U, kjer je to mogoče.

Na množici besed gramatike G definiramo preslikavo

$$\Phi(w) = \{\phi_w(c_1, \ldots, c_m) \mid c_1, \ldots, c_m \in \mathbb{F}\}.$$

Na $L(G)$ uvedemo ekvivalenčno relacijo

$$w \sim v \iff \Phi(w) = \Phi(v).$$

in preko nje definiramo **prostor formalnih izrazov** kot kvocient $L(G)/_\sim$.

Izračun verjetnosti

Iščemo algoritem, ki bi za poljubno gramatiko G kot zgoraj in poljubno besedo $w \in L(G)$ izračunal $P([w]) := \sum_{v \sim w} P(v)$. Izkaže se, da je splošen problem neizračunljiv [2], obstajajo pa rešitve za manjše družine gramatik. Osnovna gramatika, ki jo uporablja ProGED, je podana s pravili

$$\begin{aligned} E &\to E + cV \;\; [p] \mid c \;\; [1-p] \\ V &\to x_1 \;\; [q_1] \mid \cdots \mid x_n \;\; [q_n] \end{aligned}$$

in tvori linearne izraze oblike $w = c + cx_{r_1} + \cdots + cx_{r_k}$. Velja enakost

$$\begin{aligned} P([w]) &= \sum_{I \subseteq \{1,\ldots,k\}} (-1)^{|I|} \frac{1-p}{1-p \sum\limits_{i \in \{1,\ldots,k\} \setminus I} q_{r_i}} \\ &= \sum_{i=k}^{\infty} (1-p) p^i \left(\sum_{\substack{l_1 + \cdots + l_k = i \\ l_j \geq 1}} \binom{i}{l_1, \ldots, l_k} q_{r_1}^{l_1} \cdots q_{r_k}^{l_k} \right). \end{aligned}$$

[1] Jure Brence, Ljupčo Todorovski, and Sašo Džeroski. Probabilistic grammars for equation discovery. *Knowledge-Based Systems*, 224:107077, 2021.

[2] Urh Primožič, Ljupčo Todorovski, and Matej Petković. P(expression|grammar): Probability of deriving an algebraic expression with a probabilistic context-free grammar, 2022. `https://arxiv.org/pdf/2212.00751.pdf`.