Rozdělení vzdálenosti mezi body z multinomického rozdělení

Šárka Horská

11.3.2020

Struktura prezentace

- motivace
- základní pojmy
- odvozené rozdělení
- odvozené vlastnosti tohoto rozdělení
- nalezení nestranného odhadu $\|\boldsymbol{P}_x - \boldsymbol{P}_y\|^2$

Motivace

- porovnání dvou multinomických rozdělení
- dvouvýběrový χ^2- test je vhodný jen tehdy, když máme dostatečně velký počet pozorování v každé složce vektoru
- tento problém nemá např. dvouvýběrový ED (Euclidean distance) test
- ED test má za testovou statistikou odhad $\|\boldsymbol{P}_x - \boldsymbol{P}_y\|^2$
- chceme nalézt nestranný odhad

Základní pojmy

Multinomické rozdělení

Ať $\boldsymbol{p} = (p_1, \ldots, p_K)^T$ je vektor konstant splňující $\sum_{k=1}^{K} p_k = 1$ a $p_k > 0$ pro $k \in \{1, \ldots, K\}$. Pak $\boldsymbol{X} \sim Mult_K(n, \boldsymbol{p}) \iff$ jeho hustota je dána vztahem

$$P[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K] = \begin{cases} \frac{n!}{x_1! \cdots x_K!} & \text{pokud } \sum_{k=1}^{K} x_k = n \\ & \text{a } x_k \in N_0 \text{ pro } \forall k \\ 0 & \text{jinak.} \end{cases}$$

Základní pojmy

- Vzdáleností bodů $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ rozumíme eukleidovskou metriku

$$D_{ij}^2 = \sum_{k=1}^{K} (X_{ik} - X_{jk})^2 = \sum_{k=1}^{K} T_{ij}^2(k)$$

Odvození rozdělení vzdáleností

Víme, že pro $\boldsymbol{X}_i, \boldsymbol{X}_j \sim Mult_K(n, \boldsymbol{p})$, nezávislé platí:

- $X_{ik}, X_{jk} \sim Bi(n, p_k)$
- $X_{ik}, X_{jk} \in (0, \ldots, n)$
- $P(X_{ik} = t) = \begin{pmatrix} n \\ r \end{pmatrix} p_k^t (1 - p_k)^{n-t}$
- X_{ik}, X_{jk} jsou nezávislé
- $T_{ij}(k) = X_{ik} - X_{jk} \in (-n, \ldots, n)$

Odvození rozdělení vzdáleností

Odvodíme postupně následující rozdělení:

$$P(T_{ij}(k) = t_k) = \sum_{m=0}^{n} P(X_{ik} = t_k + m)P(X_{jk} = m)$$

$$P(T_{ij}^2(k) = t_k^2) = \begin{cases} P(T_{ij}(k) = t_k) + P(T_{ij}(k) = -t_k) = 2P(T_{ij}(k) = t_k) \\ P(T_{ij}(k) = 0) \end{cases}$$

Odvození rozdělení vzdáleností

- Definujeme množinu

$$\boldsymbol{I} = \{t = t_1^2 + \cdots + t_K^2 : t_i \in \{0, \ldots, n\} \text{ pro } i \in \{1, \ldots, K\}\}$$

- Hledané rozdělení je tvaru

$$P(d_{ij}^2 = t) = \sum_{t \in \boldsymbol{I}} \prod_{k=1}^{K} (1 + \delta_k) \sum_{m=0}^{n} P(X_{ik} = t_k + m) P(X_{jk} = m),$$

kde $\delta_k = 1$ pro $t_k > 0$ a $\delta_k = 0$ jinak.

Základní vlastnosti rozdělení

- $E[d_{ij}^2] = E[\sum_{k=1}^{K}(X_{ik} - X_{jk})^2] =$
  $\sum_{k=1}^{K}(E(X_{ik}^2) - 2E(X_{ik}X_{jk}) + \ E(X_{jk}^2)) =$
  $2\sum_{k=1}^{K}(E(X_{ik}^2) - E(X_{ik})^2) = 2\sum_{k=1}^{K} Var(X_{ik})$
- vyjádření rozptylu vzdálenosti je velice nepřehledné

Odhad $\|\boldsymbol{P}_x - \boldsymbol{P}_y\|^2$

Ať $\boldsymbol{X} = \{\boldsymbol{X_i}\} \sim Mult_K(n, \boldsymbol{P}_x)$ pro $i = 1, \ldots, N_x$ jsou nezávislé, $\boldsymbol{Y} = \{\boldsymbol{Y_i}\} \sim Mult_K(n, \boldsymbol{P}_y)$ pro $i = 1, \ldots, N_y$ jsou nezávislé a $\boldsymbol{X}, \boldsymbol{Y}$ jsou vzájemně nezávislé.

- $\hat{P}_{(x)k} = \frac{1}{nN_x} \sum_{i=1}^{N_x} X_{ik}$
- tento odhad je nestranný a konzistentní
- při výpočtu $E[\|\boldsymbol{P}_x - \boldsymbol{P}_y\|^2]$ zjistíme, že není nestranný

Odhad $\|\boldsymbol{P}_x - \boldsymbol{P}_y\|^2$

- hledané c splňující $E[\left\|\hat{\boldsymbol{P}}_x - \hat{\boldsymbol{P}}_y\right\|^2 - c] = \|\boldsymbol{P}_x - \boldsymbol{P}_y\|^2$ je tvaru

$$c = \frac{1}{n^2 N_x} \sum_{k=1}^{K} \hat{Var} X_{1k} + \frac{1}{n^2 N_y} \sum_{k=1}^{K} \hat{Var} Y_{1k}$$

- testová statistika je tvaru

$$ED = \left\|\hat{\boldsymbol{P}}_x - \hat{\boldsymbol{P}}_y\right\|^2 - \frac{1}{n^2 N_x} \sum_{k=1}^{K} \hat{Var} X_{1k} - \frac{1}{n^2 N_y} \sum_{k=1}^{K} \hat{Var} Y_{1k}$$

Závěr

Svou prezentaci bych časem ráda doplnila o:

- porovnání ED testových statistik s neupravenou statistikou a nestrannou statistikou
- nalézt více důvodů, proč se o rozdělení vzdálenosti zajímat
- doplnit některé výpočty
- doplnit vlastnosti rozdělení

Děkuji za pozornost!