Tutorial de Predição Conformal

Paulo C. Marques F. (Insper)

aMostra de Estatística 2024 - IME / USP

Insper

Variáveis aleatórias $(U_1, \ldots, U_n)$ permutáveis (**Definição formal?**)

Permutabilidade é uma condição mais fraca que IID

IID $\Rightarrow$ Permutável, mas a recíproca é falsa (**Exemplo?**)

Realizações de $n = 6$ variáveis aleatórias, que suporemos serem permutáveis:

U_1	U_2	U_3	U_4	U_5	U_6
3	5	2	3	5	4

Estatísticas de ordem:

$U_{(1)}$	$U_{(2)}$	$U_{(3)}$	$U_{(4)}$	$U_{(5)}$	$U_{(6)}$
2	3	3	4	5	5

Quantas das U_i's são menores ou iguais a $U_{(k)}$?

Resultado: pelo menos k das U_i's são menores ou iguais a $U_{(k)}$

(**Verifique!**)

O resultado anterior pode ser expresso por (**Por quê?**):

$$\sum_{i=1}^{n} I_{\{U_i \leq U_{(k)}\}} \geq k$$

Tomando esperanças e observando que, por permutabilidade,
$P(U_i \leq U_{(k)})$ tem o mesmo valor para todo $i, k = 1, \ldots, n$, temos que:
$P(U_i \leq U_{(k)}) \geq k/n$ (**Por quê?**)

Corolário: se há probabilidade zero de termos empates entre as U_i's (**Exemplo em que isto ocorreria?**), então $P(U_i \leq U_{(k)}) = k/n$

(**Por quê?**)

Um problema de regressão: $X_i \in \mathbb{R}^d$ e $Y_i \in \mathbb{R}$

Exemplo: California Housing

https://github.com/paulocmarquesf

Temos uma *sequência* de pares permutáveis (**Definição?**):

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \ldots$$

Fomos a uma Big Tech e compramos uma função determinística $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$, que prevê Y_i a partir de X_i, para os nossos dados

Defina os *escores de conformidade*: $R_i = |Y_i - \hat{\mu}(X_i)|$, para $i \geq 1$

Resultado: a sequência $R_1, R_2, \ldots$ é permutável (**Por quê?**)

Dados de calibração: $(X_1, Y_1), \ldots, (X_n, Y_n)$

Escores de calibração ordenados: $R_{(1)}, R_{(2)}, \ldots, R_{(n)}$

Observável futuro: (X_{n+1}, Y_{n+1})

Resultado: $P(R_{n+1} \leq R_{(k)}) \geq k/(n+1)$, para $k = 1, \ldots, n$

(**Por quê?**)

Lembrando: $P(R_{n+1} \leq R_{(k)}) \geq k/(n+1)$, para $k = 1, \ldots, n$

Escolha um nível de "descobertura" nominal $0 < \alpha < 1$

Defina o teto de um número real t por $\lceil t \rceil = \min\{z \in \mathbb{Z} : t \leq z\}$

Fato aritmético: $t \leq \lceil t \rceil < t + 1$ (**Verifique!**)

Escolha $k = \lceil (1-\alpha)(n+1) \rceil$

Resultado: $P(R_{n+1} \leq R_{(\lceil (1-\alpha)(n+1) \rceil)}) \geq 1 - \alpha$ (**Verifique!**)

Se não tivermos empates entre os escores de conformidade, também teremos uma quota superior:

$$P(R_{n+1} \leq R_{(\lceil (1-\alpha)(n+1) \rceil)}) \leq 1 - \alpha + \frac{1}{n+1} \qquad \text{(\textbf{Verifique!})}$$

Defina $\hat{r} = R_{(\lceil (1-\alpha)(n+1) \rceil)}$

$$R_{n+1} \leq \hat{r} \iff \hat{\mu}(X_{n+1}) - \hat{r} \leq Y_{n+1} \leq \hat{\mu}(X_{n+1}) + \hat{r} \qquad \textbf{(Por quê?)}$$

Propriedade universal do intervalo de predição conformal:

$$1 - \alpha \leq P(\hat{\mu}(X_{n+1}) - \hat{r} \leq Y_{n+1} \leq \hat{\mu}(X_{n+1}) + \hat{r}) < 1 - \alpha + \frac{1}{n+1}$$

Lembrete: a quota superior só vale se houver probabilidade zero de termos empates entre os escores de conformidade

Predição conformal na prática:

https://github.com/paulocmarquesf


MACHINE LEARNING
MACHINE LEARNING MODELS
OBRIGADO
PELA PRESENÇA