Inferencia Conformal

Mauro Loprete y Maximiliano Saldaña

1/1/22

Tabla de contenidos

1 Introducción

En Lei et al. (2016) los autores plantean un marco general para realizar inferencia predictiva sin supuestos distribucionales en un contexto de regresión, empleando la *inferencia conformal*. Mediante la metodología planteada se pueden obtener intervalos de confianza con validez en muestra finitas (no asintótica) para una variable de respuesta, empleando cualquier estimador de la función de regresión.

El problema se plantea de la siguiente manera: Se considera $Z_1, ..., Z_n \sim F$ i.i.d., donde $Z_i = (X_i, Y_i)$ es una variable aleatoria en $\mathbb{R}^d \times \mathbb{R}$, Y_i es la variable de respuesta y $X_i = X_i(1) ..., X_i(d)$ son las covariables. Se tiene tiene la función de regresión:

$$\mu(x) = E(Y|X = x), \quad x \in \mathbb{R}^d$$

Es de interés predecir la nueva respuesta Y_{n+1} a las covariables X_{n+1}, sin hacer supuestos sobre μ o F. Dado un nivel de cobertura α, el objetivo es construir un intervalo de predicción $C \subseteq \mathbb{R}^d \times \mathbb{R}$ basado en $Z_1, ..., Z_n$ que cumpla:

$$P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$$

En esta expresión se supone que $Z_{n+1} = (X_{n+1}, Y_{n+1})$ proviene también de la distribución F y $C(x) = \{y \in \mathbb{R} : (x, y) \in C\}, \ x \in \mathbb{R}^d$

2 Inferencia conformal

La idea básica de la inferencia conformal, dadas las definiciones de la introducción, consiste en que para decidir si un valor y está incluido en el intervalo $C(X_{n+1})$ consideramos poner a prueba la hipótesis nula de que $Y_{n+1} = y$ y se construye un valor-p válido basado en los cuantiles empíricos de la muestra aumentada $Z_1, \dots, Z_n, Z_{n+1}$.

2.1 Un resultado previo

Sean $U_1, \dots, U_n$ una muestra i.i.d de una variable aleatoria continua. Para un nivel de no cobertura $\alpha \in (0, 1)$ y una observación U_{n+1}, nótese que:

$$P(U_{n+1} \leq \hat{q}_{1-\alpha}) \geq 1 - \alpha \tag{2.1}$$

Donde $\hat{q}_{1-\alpha}$ es el cuantil de la muestra $U_1, \dots, U_n$ definido por:

$$\hat{q}_{1-\alpha} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)} & \text{si} \quad \lceil (n+1)(1-\alpha) \rceil \leq n \\ \infty & \text{en caso contrario} \end{cases} \tag{2.2}$$

Aquí $U_{(1)} \leq \dots \leq U_{(n)}$ son los estadísticos de orden de la muestra. Se verifica la cobertura en muestra finita de la Ecuación 2.1: dada la independencia de las variables, el rango de U_{n+1} en la muestra se distribuye uniforme en el conjunto $\{1, \dots, n+1\}$, entonces

2.2 Método *naive* de construcción de intervalos

Usando el resultado previo de la sección anterior y en el contexto de regresión planteado en la Sección Capítulo 1, un método sencillo para contruir un intervalo predictivo para Y_{n+1} ante el valor X_{n+1} es:

$$C_{naive}(X_{n+1}) = \left[\hat{\mu}(X_{n+1}) - \hat{F}_n^{-1}(1-\alpha), \hat{\mu}(X_{n+1}) + \hat{F}_n^{-1}(1-\alpha)\right] \tag{2.3}$$

donde $\hat{\mu}$ es un estimador de la función de regresión, $\hat{F}_n$ la distribución empírica de los residuos dentro de la muestra $|Y_i - \hat{\mu}(X_i)|$, $i = 1, \dots, n$ y $F_n^{-1}(1-\alpha)$ el cuantil $1-\alpha$ de $\hat{F}_n$.

Este método es aproximadamente válido para muestras grandes, bajo la condición de que $\hat{\mu}$ sea lo suficientemente preciso, es decir, que $F_n^{-1}(1-\alpha)$ esté cerca del cuantil $1-\alpha$ de $|Y_i - \mu(X_i)|$. Para que esto se cumpla en general es necesario el cumplimiento de condiciones de regularidad tanto para la distribución F de los datos y para $\hat{mu}$, como que el modelo esté correctamente especificado.

Un problema de este método es que los intervalos pueden presentar una considerable subcobertura, dado que se están empleando los residuos dentro de la muestra. Para subsanar esto, en Lei et al. (2016) se plantea la metodología de los intervalos de predicción conformales.

2.3 Intervalos de predicción conformales

Para cada valor $y \in \mathbb{R}$ se construye un estimador de regresión aumentado $\hat{\mu}_y$, el cual se estima en el conjunto de datos aumentado $Z_1, \dots Z_n, (X_{n+1}, y)$. Luego, se define:

$$R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|, \;\; i = 1, \dots, n \tag{2.4}$$

$$R_{y,n+1} = |y - \hat{\mu}_y(X_{n+1})| \tag{2.5}$$

Con el rango de $R_{y,n+1}$ entre los demás residuos de la muestra $R_{y,1}, \dots, R_{y,n}$ se calcula:

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{I}\{R_{y,i} \leq R_{y,n+1}\} = \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}\{R_{y,i} \leq R_{y,n+1}\} \tag{2.6}$$

que es la proporción de los puntos de la muestra aumentada cuyos residuos dentro de la muestra son más pequeños que el residuo $R_{y,n+1}$. Como los datos son i.i.d. y suponiendo la simetría de $\hat{\mu}$, se puede apreciar que el estadístico $\pi(Y_{n+1})$ se distribuye uniforme en $1/(n+1), 2/(n+1), \dots, 1$, lo cual implica[1]:

$$P((n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil) \geq 1 - \alpha \tag{2.7}$$

Esta expresión se puede interpretar como que $1 - \pi(Y_{n+1})$ da un valor-p válido conservador para la prueba de hipótesis donde $H_0) Y_{n+1} = y$.

Aplicando dicha prueba sobre todos los posibles valores de $y \in \mathbb{R}$, la ecuación Ecuación 2.7 lleva al intervalo de predicción conformal evaluado en X_{n+1}:

$$C_{conf}(X_{n+1}) = [y \in \mathbb{R} : (n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil] \tag{2.8}$$

[1] Ver por qué

Cada vez que se quiere obtener un intervalo de predicción en un nuevo conjunto de covariables se tienen que recalcular los pasos Ecuación 2.4, Ecuación 2.5, Ecuación 2.6 y Ecuación 2.8. En la práctica, se restringen los valores de y a una grilla discreta.

El procedimiento para obtener el intervalo se puede resumir en el Algoritmo 1.

Algoritmo 1: Intervalo de predicción conformal

Entrada: Datos (X_i, Y_i), $i = 1, \dots, n$, nivel de no cobertura $\alpha \in (0, 1)$, algoritmo de regresión $\mathcal{A}$, puntos $\mathcal{X}_{nuevo}$ en los que construir intervalos de predicción y valores $\mathcal{Y}_{prueba} = \{y_1, y_2, \dots\}$ para comparar con la predicción.

Salida: Intervalos de predicción, en cada elemento de $\mathcal{X}_{nuevo}$

for $x \in \mathcal{X}_{nuevo}$ **do**
  for $y \in \mathcal{Y}_{prueba}$ **do**
    $\hat{\mu}_y = \mathcal{A}(\{(X_1, Y_1), \dots, (X_n, Y_n), (x, y)\})$
    $R_{y_i} = |Y_i - \hat{\mu}_y(X_i)|,\ i = 1, \dots, n$ y $R_{y,n+1} = |y - \hat{\mu}_y(x)|$
    $\pi(y) = (1 + \sum_{i=1}^{n} \mathbb{I}\{R_{y,i} \leq R_{y,n+1}\})/(n+1)$
  $C_{conf}(x) = [y \in \mathbb{R} : (n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil]$
Se devuelve $C_{conf}(x)$ para cada $X \in \mathcal{X}_{nuevo}$.

2.3.1 Teorema

El intervalo Ecuación 2.8 tiene cobertura válida para muestras finitas por construcción y a su vez no presenta sobrecobertura. Esto se puede expresar mediante las expresiones Ecuación 2.9 y Ecuación 2.10, respectivamente:

Sea (X_i, Y_i), $i = 1, \dots, n$ v.a. i.i.d, entonces para la nueva observación i.i.d. (X_{n+1}, Y_{n+1}):

$$P(Y_{n+1} \in C_{conf}(X_{n+1})) \geq 1 - \alpha \tag{2.9}$$

Adicionalmente, si se hace el supuesto que para todo $y \in \mathbb{R}$ los residuos dentro de la muestra $R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|$, $i = 1, \dots, n$ tienen una distribución conjunta continua se cumple que:

$$P(Y_{n+1} \in C_{conf}(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1} \tag{2.10}$$

Observación. Nótese que las probabilidades aquí, al tomarse sobre la muestra aumentada i.i.d. implican cobertura promedio (o marginal). Esto no es lo mismo que la cobertura condicional $P(Y_{n+1} \in C_{conf}(x) | X_{n+1} = x) \geq 1 - \alpha \ \ \forall \ \ x \in \mathbb{R}^d$. Esta última es una propiedad más fuerte y no puede lograrse con intervalos predictivos de amplitud finita sin que el modelo y el estimador cumplan condiciones de regularidad y consistencia.

Observación. Si se mejora el estimador $\hat{\mu}$, en general el intervalo de predicción conformal decrece en tamaño. Esto se da debido a que un $\hat{\mu}$ más preciso lleva a residuos más pequeños y los intervalos conformales están definidos por los cuantiles de la distribución aumentada de los residuos.

2.4 Intervalos de predicción conformales con muestras separadas

Un problema práctico de los intervalos de inferencia conformal de la sección anterior es que tienen mucho costo computacional. Para poder concluir si $y \in C_{conf}(X_{n+1})$, para cualquier X_{n+1} y y, se tiene que reestimar el modelo en la muestra aumentada que incluye el nuevo punto X_{n+1} y recalcular y reordenar los nuevos residuos obtenidos.

Para enfrentar esta problemática se puede hacer uso de una metodología denominada por Lei et al. (2016) como predicción conformal separada (*split conformal prediction*). Su costo computacional es menor (es el del paso de estimación únicamente) y tiene menos requerimientos de memoria (solo hay que guardar las variables seleccionadas cuando se evalúa el ajuste en los nuevos puntos X_i, $i \in \mathcal{I}_2$). Se presenta en el Algoritmo 2.

Algoritmo 2: Intervalos de predicción conformales con muestras separadas

Entrada: Datos (X_i, Y_i), $i = 1, \dots, n$, nivel de no cobertura $\alpha \in (0, 1)$, algoritmo de regresión $\mathcal{A}$.

Salida: Intervalos de predicción, sobre $x \in \mathbb{R}^d$

1 Se separa la muestra al azar en dos subconjuntos de igual tamaño $\mathcal{I}_1$ e $\mathcal{I}_2$.
2 $\hat{\mu}_y = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$
3 $R_i = |Y_i - \hat{\mu}_y(X_i)|$, $i \in \mathcal{I}_2$
4 $d =$ el k-ésimo valor más pequeño en $\{R_i : i \in \mathcal{I}_2\}$, donde $k = \lceil (n/2 + 1)(1 - \alpha) \rceil$
5 Se devuelve $C_{split}(x) = [\hat{\mu} - d, \hat{\mu} + d]$ para todo $x \in \mathbb{R}^d$.

2.4.1 Teorema

Sea (X_i, Y_i), $i = 1, \dots, n$ v.a. i.i.d, entonces para la nueva observación i.i.d. (X_{n+1}, Y_{n+1}):

$$P(Y_{n+1} \in C_{splt}(X_{n+1})) \geq 1 - \alpha \tag{2.11}$$

Adicionalmente, si se hace el supuesto que los residuos R_i, $i \in \mathcal{I}_2$ tienen una distribución conjunta continua se cumple que:

$$P(Y_{n+1} \in C_{split}(X_{n+1})) \leq 1 - \alpha + \frac{2}{n+2} \tag{2.12}$$

2.4.2 Teorema

Los intervalos de predicción conformales con muestras separadas dan una garantía aproximada de cobertura dentro de la muestra. Esto se puede expresar como que existe una constante $c > 0$ tal que, para cualquier $\epsilon > 0$:

$$P\left(\frac{2}{n}\sum_{i\in\mathcal{J}_2}\mathbb{I}\{Y_I \in C_{split}(X_i) - (1-\alpha) \geq \epsilon\}\right) \leq 2\exp\left(-cn^2(\epsilon - 4/n)^2\right) \tag{2.13}$$

Esto implica cobertura dentro de la muestra para la muestra $\mathcal{J}_2$, revirtiendo los roles de $\mathcal{J}_1$ e $\mathcal{J}_2$ se puede extender para toda la muestra.

Observación. También se puede aplicar este método con una separación no balanceada de la muestra, con $|\mathcal{J}_1| = \rho n$ e $|\mathcal{J}_1| = \rho n$, para $\rho \in (0,1)$. Esto puede ser útil en situaciones donde el procedimiento de regresión es complejo y puede resultar beneficioso elegir $\rho > 0,5$, para que $\hat{\mu}$ sea más preciso.

2.5 Intervalos conformales con múltiples separaciones de la muestra

Si bien separar la muestra reduce el tiempo que se tarda en calcular los intervalos, introduce otro elemento aleatorio en el método que incrementa la variabildiad; que es cuáles observaciones quedan en uno u otro de los subconjuntos. Una manera de enfrentarse a esto es combinar las inferencias realizadas con distintas separaciones de la muestra.

Sea N la cantidad de veces que separamos la muestra, se calculan con estos distintos subconjuntos los intervalos de predicción conformales $C_{split,1}, \dots, C_{split,N}$, donde cada intervalo se construye al nivel de confianza $1 - \alpha/N$. Se define:

$$C^{(N)}_{split}(x) = \bigcap_{j=1}^{N} C_{split,j}(x), \text{ sobre } x \in \mathbb{R}^d \tag{2.14}$$

Usando un argumento del tipo Bonferroni[2], se concluye que la banda de predicción $C^{(N)}_{split}$ tiene una cobertura marginal de por lo menos $1 - \alpha$.

Esta metodología reduce la variabilidad originada por la separación, pero se puede dar que el tamaño del intervalo $C^{(N)}_{split}$ es creciente en N y es más amplio que C_{split}, debido al nivel de confianza con que se construye cada intervalo.

[2]Ver esto

2.6 Intervalos predictivos mediante Jackknife

Esta metodología emplea los cuantiles de los residuos de validación cruzada dejando una observación fuera (*leave-one-out*) para calcular los intervalos de predicción.

Algoritmo 3: Intervalo de predicción conformal mediante Jackknife.

Entrada: Datos (X_i, Y_i), $i = 1, \ldots, n$, nivel de no cobertura $\alpha \in (0, 1)$, algoritmo de regresión $\mathcal{A}$.

Salida: Intervalos de predicción sobre $x \in \mathbb{R}^d$.

for $i \in \{1, \ldots, n\}$ **do**

  $\widehat{\mu}^{(-i)} = \mathcal{A}(\{(X_l, Y_l) : l \neq i\})$

  $R_i = |Y_i - \widehat{\mu}^{(-i)}(X_i)|$

$d =$ el k-ésimo valor más pequeño en $\{R_i : i \in \{1, \ldots, n\}\}$, con $k = \lceil n(1-\alpha) \rceil$

Se devuelve $C_{jack}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$ para todo $x \in \mathbb{R}^d$

Tiene la ventaja que emplea más de la muestra que se aparta para entrenar cuando se calculan los residuos, lo cual frecuentemente lleva a intervalos de menor amplitud. Como desventaja, los intervalos que se obtienen no garantizan cobertura válida fuera de la muestra cuando se trabaja con muestras finitas e incluso asintóticamente la cobertura depende de condiciones del estimador.

3 Aplicación

4 Conclusión

5 Bibliografía

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2016). *Distribution-Free Predictive Inference For Regression*. arXiv. https://doi.org/10.48550/ARXIV.1604.04173

6 Anexo