Proyecto final

2025-11-07

Introducción

En el articulo de Ryan Tibsharini (2023) sobre inferencia conformal, el autor presenta un marco general para cuantificar la incertidumbre en problemas de predicción sin importar supuestos paramétricos sobre la distribución P de los datos. La idea central consiste en transformar cualquier predictor puntual en un predictor para conjuntos que garantice coberturta válida en muestras finitas.

En el contexto de regresión, esta metodología permite contruir bandas de predicción que conserva la propiedad de cobertura deseada, independientemente del algoritmo usado para estimar la función de regresión

Objetivos

Sea $(X_i, Y_i) \sim P, i = 1, ..., n$ iid , las variables explicativas y dependiente de una distribucion P en $\mathcal{X}x\mathcal{Y}$. Podriamos pensar la dimension del conjunto de las variales explicativas en $\mathcal{X} = R^d$, mientras que la variable dependiente en el espacio de todos los reales $\mathcal{Y} = R$ Dado una probabilidad α llamado tasa de no cobertura, queremos encontrar una banda de predicción

$$\hat{C}_n : \mathcal{X} \to \{subconjunto\ de\ \mathcal{Y}\}$$

Con la propiedad que para un nuevo par de datos $(X_{n+1}, Y_{n+1}) \sim P$

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

Donde la probabilidad es sobre los datos (X_i, Y_i) $i = 1, ..., n + 1$ Por otra parte,si no asumimos ninguna teoría asintotica ni tampoco ninguna distribución en P, obtener una cobertura exacta es algo muy difícil en general. Podríamos hacer algo totalmente trivial para obtenerla:

$$\hat{C}_n(X_{n+1}) = \begin{cases} \mathcal{Y} & \text{con probabilidad } 1 - \alpha, \\ \varnothing & \text{con probabilidad } \alpha. \end{cases}$$

Siempre tendrá cobertura exacta $1 - \alpha$, lo cual es, lo que queremos lugrar con la ecuación (1).

La pregunta real sería, podriamos lograr la ecuación (1) en muestras finitas sin asumir ninguna distribución P, haciendo algo "no trivial"?. En particular, queremos que nuestra estrategia se adapte a la dificultad del problema, en el siguiente sentido: cuanto más fácil sea predecir Y_{n+1} a partir de las variables explicativas X_{n+1}, mas chico nos gustaría que fuera el conjunto $\hat{C}_n(X_{n+1})$

Supongamos que nuestro objetivo inicial es encontrar una cola de intervalo de predicción. $\hat{C}_n = (-\infty, \hat{q}_n]$

Dada esta ecuación, un punto de partida natural sería fijar $\hat{q}_n$, como el cuantil muestral de nivel $1 - \alpha$ de $Y_1, ..., Y_n$ el cual denotamos por

$$P(Y_{n+1} \leq \hat{q}_n) \sim 1 - \alpha$$

$$\hat{q}_n = \begin{cases} Y[(1-\alpha)(n+1)]\ si\ [(1-\alpha)(n+1) \leq n, \\ \infty\ en\ caso\ contrario \end{cases}$$

Aquí $Y_{(1)}, Y_{(2)}, ...Y_n$ son los estadisticos de orden de la muestra $Y_{(1)} \leq Y_{(2)} \leq ..., Y_{(n)}$. Se verifica la cobertura en muestra finita de la ecuación (1) debido a la independencia de las variables.Por otra parte el rango de Y_{n+1} se distribuye uniforme en el conjunto $\{1, ..., n+1\}$, es decir, la predicción Y_{n+1} tiene probabildiad $\frac{1}{n+1}$ de caer en cualquier posición del conjunto ordenado.

Método de Naive

Veamos la primera idea clave sobre regresion, donde observamos ambos X_i and $Y_i \ \in R \ i = 1, ..., n$, y queremos un conjunto de predicción para Y_{n+1} basado en X_{n+1}. Supongamos que $\hat{f}_n$ es cualquier predictor puntual, entrenado en (X_i, Y_i), $i = 1, ..., n$. En otras palabras, $\hat{f}_n(x)$ predice el valor de y que esperamos observar en x.

Definimos los resudios de los datos de entrenamiento,

$$R_i = |Y_i - \hat{f}_n(X_i)|, \ i = 1, ..., n$$

tenemos $\hat{q_n} = [(1-\alpha)(n+1)]$ el mas chico de $R_1, ...R_n$, podemos definir el conjunto de predicción como $\hat{C}_n(x) = \{y : |y - \hat{f}_n(x)| \leq \hat{q_n}\}$

O en otras palabras:

$$\hat{C}_n(x) = [\hat{f}_n(x) - \hat{q}_n, \hat{f}_n(x) + \hat{q}_n]$$

Este método es aproximadamente válido para muestras grandes, bajo la condicion de que $\hat{f}_n(x)$ sea los suficientemente preciso, es decir, que $\hat{q_n}$ este cerca del cuantil $1-\alpha$ de R_i. Un problema de este método es que los intervalos de predicción pueden presentar una considerable subcobertura, dado que se estan empleando los residuos dentro de la muestra. Para evitar esto, se plantea la metodología de los intervalos de predicción conformales

Ejemplo práctico: Supongamos que tenemos un conjunto de datos $X = (1, 2, 3, 4)$ y $Y = (2, 4, 6, 8)$, donde ajustamos un modelo lineal simple $\hat{f}(x) = 1 + x$, y obtenemos $\hat{f}(x) = (2, 3, 4, 5)$ y los residuos $e_i = (0, 1, 2, 3)$ Luego calculamos el cuantil con cobertura $1-\alpha$, para ello calculamos $\hat{q_n} = [(0.8)(\ 5)] = 4$ Ordenamos los residuos de menos a mayor y vemos que el residuo $e_4 = 3$. Luego nos contruimos un intervalo de predicción para $X_{n+1} = 5$ $\hat{f(5)} = 6$. Por lo que el intervalo conformal con cobertura 0.8 es : $\hat{C(5)} = [6-3, 6+3] = [3, 9]$

Separación de la muestra de intervalos de prediccón

En esta priemra parte de esta sección, nos enfocaremos en la parte de regresion, es decir, que Y pertenece a todos los reales/

En concreto, dividimos el conjunto de entrenamiento en 2:

- T_1, es el conjunto de entrenamiento propimente dicho
- T_2 es el conjunto de calibración o testeo.

Tiene sentido pensar que la intersección de ambos es vacia, $T_1 \cap T_2 = \emptyset$ y $T_1 \cup T_2 = \{1, 2, ..., n\}$, sea $n_1 = |T_1| \ y \ \ n_2 = |T_2|$

En el siguiente paso, entrenamos el predictor puntual usando los datos del conjunto de entrenamiento propiamente dicho (X_i, Y_i) donde $i \in T_1$ y lo denotamos como $\hat{f}_{n_1}$.Luego, vemos los resudios en el conjunto de calibración: $e_i = |Y_1 - \hat{f}_{n_1}| \ i \in T_2$

Por lo que llegamos a que el residuo mas chico del cuantil conformal es el siguiente: $\hat{q}_{n_2} = [(1-\alpha)(n_2+1)]$ R_i $i \in R_2$. Aquí se calcula un cuantil de nivel de cobertura $1-\alpha$ de los residuos R_i, para construir un **intervalo de predicción** que tenga cobertura aproximada $1-\alpha$, es decir, primero se ordenan los residuos R_i del conjunto T_2 de menor a mayor. Luego, se busca el cuantil empírico de orden $(1-\alpha)$, es decir, el residuo en lal posición

$[(1-\alpha)(n_2+1)]$ donde n_2 es el tamaño del conjunto de calibración. Este valor se denota como: $\hat{q}_{n_2}$ y se utiliza para crear un intervalo de predicción al rededor de la estimación puntual, $C(x) = [\hat{f}_{n_1}(x) - \hat{q}_{n_2}, \hat{f_{n_1}}(x) + \hat{q}_{n_2}]$

La mayor garantía que podemos obtener es que:

$$P(y_{n+1} \in \hat{C}_n(X_{n+1})|(X_i, Y_i)i \in T_1) \in [1-\alpha, 1-\alpha+\frac{1}{n_2+1}]$$

Finalmente:

$$Y_{n+1} \leq \hat{C_n}(X_n+1) \Leftrightarrow e_{n+1} \leq \hat{q}_{n2} \Leftrightarrow e_{n+1} \leq [(1-\alpha)(n_2+1)$$

Esto ocurre con probabilidad al menos $1-\alpha$ y a al menos $1-\alpha+\ \frac{1}{n+1}$

Full conformal prediciton

Ahora hacemos algo distinto a lo que haciamos hasta ahora, entrenamos nuestro algoritmo de predicción con los datos $(X_1, Y_1), ..., (X_n, Y_n), (x.y)$. Es decir, usamos un conjunto de entrenamiento extendido, con $n+1$ puntos (incluyendo el nuevo par (x,y)).

A partir de este conjunto, obtenemos un predictor punutal $\hat{f}_n(x, y)$

Luego, definimos los residuos como:

- para i = 1,. . . .,n

$$R_i^{(x,y)} = |Y_i - \hat{f}_n(x, y)(X_i)|$$

- Para el nuevo punto de prueba:

$$R_{n+1}^{(x,y)} = |y - \hat{f}_n(x, y)(x)|$$

Finalmente, definimos el conjunto conformal como:

$$\hat{C_n}(x) = \{y \in R_{n+1}^{(x,y)} \leq [(1-\alpha)(n+1)] - \textit{ésimo menor de los } R_1^{((x,y))}, ..., R_n^{x,y}\}$$

Es decir, para cada valor candidato y, se crea un cojunto de entrenamiento aumentado:

$$\{(X_1, Y_1), ..., (X_n, Y_n), (X_{n+1}, y)\}$$

Luego se evalúa si el residuo del nuevo punto de la prueba $R_{n+1}^{(x,y)}$ está entre los más pequeños de todos los residuos generados con ese conjunto extendido. Si cumple la condición, entonces ese y pertenece al intervalo de predicción.

Intervalos predicitivos vía Jacknife

Este algoritmo tiene puntos de contacto con el full conformal prediction. Aquí la idea es separar el conjunto en entrenamiento y testeo, dejando una observación aparte para testear, es decir, entrenas $\hat{f}_{-i}$ dejando fuera la observación i, luego se calculan el residuo confromal R_i sin esa observación. Luego se ordenan los residuos R_i y se calcula el k-ésimo valor más chico, donde $k = [n(1-\alpha)]$, ese valor q vendria a ser la mitad del ancho del intervalo. Por último, se entrena $\hat{f}$ con todos los datos y se devuelve el intervalo de predicción conformal para un nuevo punto $x \in R^d$.

$$C_{jack}(x) = [\hat{f}(x) - \hat{d}, \hat{f}(x) + \hat{d}]$$