MECAI 2024 III Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Quantificação de Incerteza em Modelos de Aprendizado de Máquina

Atila Ferreira Pessoa[1], Fabrício Simeoni de Sousa[2]
ICMC-USP

1 Introdução

Modelos de aprendizado supervisionado utilizados em problemas de regressão são capazes de produzir predições pontuais y_{n+1} para dados não vistos durante o treinamento X_{n+1} através do aprendizado de padrões em conjunto de dados de treino $(X_i, y_i)_{i=1}^{n}$. No entanto, em alguns casos é necessário conhecer o grau de incerteza associado às predições.

Uma forma robusta de estimar o nível de incerteza associado a cada nova predição pode ser obtida através da implementação do método Conformal Prediction, onde são geradas predições intervalares $\hat{C}$ associadas a um nível de significância α em detrimento de predições pontuais [1] [2] de tal forma que os novos valores preditos Y_{n+1} estejam contidos dentro do intervalo $\hat{C}(X_{n+1})$ com frequência $1 - \alpha$, conforme representado a seguir:

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

O método também garante que os intervalos gerados apresentem:

- **Validade**: Garantia de que o valor real estará contido dentro do intervalo predito. Essa garantia pode ser assintoticamente exata ($\lim_{n\to\infty} P\{y_{n+1} \in C(X_{n+1})\} = 1 - \alpha$), ou assintoticamente conservadora ($\lim_{n\to\infty} P\{y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$).
- **Eficiência**: Garantia de que os intervalos preditos terão a menor largura possível.
- **Flexibilidade**: Possibilidade de implementação em diferentes tipos de modelos que utilizem diferentes heurísiticas nos processos de aprendizagem (*e.g.* Regressão Linear, Random Forest, Support Vector Regression Machines, Redes Neurais Artificiais *etc*).
- **Condicionalidade**: Largura adaptável dos intervalos preditos, de tal forma que regiões de maior incerteza apresentem intervalos relativamente maiores aos obtidos para predições em regiões de menor incerteza dos modelos.

[1] atila.pessoa@usp.br
[2] fsimeoni@icmc.usp.br

Cabe ressaltar que o método não faz suposições sobre a distribuição dos dados para fornecer intervalos ou conjuntos de previsões válidos. Essa propriedade torna o método bastante robusto, uma vez que permite uma aplicação ampla em conjuntos de dados variados. A única premissa adotada pelo método é a de que os dados são intercambiáveis (iid).

A forma mais intuitiva de se obter a predição de intervalos deriva da realização de sucessivos treinamentos com diferentes conjuntos de treino e avaliação das distribuições condicionais dos erros associados a essas predições. Nesse sentido, o método *Full Conformal Prediction* representa um tipo de implementação que utiliza todo o conjunto de treino para criar os intervalos preditos, ajustando esses intervalos a cada obsevação nova. Porém, devido ao fato de exigir o treinamento de sucessivos modelos para cada nova observação não vista durante o treino, a estratégia torna o método custoso computacionalmente [1], [2].

2 Estratégias

O método *Naive* é uma das formas mais simples de implementar o método conformal prediction. Basicamente, o método consiste em treinar um modelo utilizando o conjunto de dados de treino e obter a distribuição dos resíduos para cada predição $\hat{y}_i$ realizada sobre o conjunto de treino. Em seguida, utiliza-se o valor predefinido para α para se obter o percentil $1 - \alpha$ da distribuição dessess resíduos.

Dessa forma, para uma nova observação X_{n+1}, será realizada a predição $\hat{y}_{n+1}$ usando o modelo treinado previamente e o intervalo será construído da seguinte forma:

$$\hat{C}(y_{n+1}) = [\hat{y}_{n+1} - \epsilon, \hat{y}_{n+1} + \epsilon] \tag{2}$$

Onde ϵ representa o percentil $1 - \alpha$ da distribuição dos resíduos.

Esse método apresenta alta simplicidade e eficiência computacional por eliminar a necessidade de treinamentos de sucessivos modelos. Porém, tal simplificação pode resultar em intervalos com taxa de cobertura inferiores a $1 - \alpha$ e sem capacidade de adaptação à heterocedasticidade dos dados, por produzir intervalos de largura fixa.

O método Jackknife+ é uma modificação do método clássico de reamostragem Jackknife, que se baseia no treinamento de sucessivos modelos $\hat{\mu}$ omitindo uma das observações do conjunto de dados a cada treinamento, estratégia também conhecida por *Leave-One-Out* (LOO).

O método *Jackknife* utiliza um único modelo $\hat{\mu}$ ajustado utilizando todo o conjunto de treino, apresentando intervalos centrados em torno do valor predito $\hat{\mu}(X_{n+1})$ e intervalo predito definido com base na distribuição dos resíduos observados para cada dado não apresentado ao modelo durante o treinamento utilizando a estratégia LOO. Já o método *Jackknife+* utiliza diferentes modelos $\hat{\mu}_{-i}$, onde i representa cada observação não fornecida ao modelo durante o treinamento, para realizar as predições $\hat{\mu}_{-i}(X_{n+1})$ e utiliza a distribuição dos respectivos resíduos associados a cada uma dessas predições para definição da espessura do intervalo [3]. Dessa forma, as predições intervalares para cada nova observação X_{n+1} podem ser descritas para as diferentes estratégias da seguinte forma:

$$\hat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1}) = \left[\hat{q}^{-}_{n,\alpha}\left\{\hat{\mu}(X_{n+1}) - R^{\text{LOO}}_i\right\}, \hat{q}^{+}_{n,\alpha}\left\{\hat{\mu}(X_{n+1}) + R^{\text{LOO}}_i\right\}\right] \tag{3}$$

$$\hat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1}) = \left[\hat{q}^{-}_{n,\alpha}\left\{\hat{\mu}_{-i}(X_{n+1}) - R^{\text{LOO}}_i\right\}, \hat{q}^{+}_{n,\alpha}\left\{\hat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_i\right\}\right] \tag{4}$$

Onde $\hat{q}^{-}_{n,\alpha}$ e $\hat{q}^{+}_{n,\alpha}$ representam os quantis α e $(1-\alpha)$, respectivamente, e os valores R^{LOO}_i representam o i-ésimo resíduo do conjunto de observações, definidos como $R^{LOO}_i = |Y_i - \hat{\mu}_{-i}(X_i)|$.

O método *Cross Validation* (CV+) pode ser implementado através da divisão do conjunto de treino em K subconjuntos disjuntos $S_1, S_2, ..., S_K$, cada um com tamanho $m = n/K$. Posteriormente, são treinados K modelos μ de tal forma que a cada treinamento um dos subconjuntos seja omitido do conjunto de treino [3]. Os resíduos de cada um dos modelos $\hat{\mu}_{-S_K}$ pode ser determinado da seguinte forma:

$$R^{CV}_i = |Y_i - \hat{\mu}_{-S_{K_{(i)}}}(X_i)|, i = 1, ..., n \tag{5}$$

Onde $k_{(i)} \in \{1, ..., K\}$. Utilizando os resíduos calculados, é possível definir o intervalo predito utilizando o CV+ como:

$$\hat{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1}) = \left[\hat{q}^{-}_{n,\alpha}\left\{\hat{\mu}_{-S_{k(i)}}(X_{n+1}) - R^{\text{CV}}_i\right\}, \hat{q}^{+}_{n,\alpha}\left\{\hat{\mu}_{-S_{k(i)}}(X_{n+1}) + R^{\text{CV}}_i\right\}\right] \tag{6}$$

Os métodos descritos anteriormente pressupõem o treinamento de sucessivos modelos para avaliação da incerteza associada às predições, de tal forma que são necessários n e K modelos para implementação das estratégias *Jackknife+* e CV+, respectivamente.

Uma forma de eliminar a necessidade de treinamento de diferentes modelos durante a implementação do método *Conformal Prediction* é possível com a utilização da estratégia *Split Conformal Prediction*, que consiste na divisão do conjunto de dados em conjunto de treino, de calibração e de teste utilizando, por exemplo, o método *Conformalized Quantile Regression* (CQR) [2], que é capaz de realizar predição de intervalos com o treinamento de um único modelo.

Uma vez que o conjunto de dados tenha sido dividido, para implementação do método CQR, é necessário que sejam treinados modelos de regressão quantílica [4] considerando os seguintes quantis $\alpha_{inf} = \alpha/2$ e $\alpha_{sup} = 1 - \alpha/2$. Dessa forma, é possível determinar o intervalo interquantílico $\hat{C}$ para a regressão quantílica como:

$$\hat{C}(x) = [q_{\alpha_{inf}}(x), q_{\alpha_{sup}}(x)] \tag{7}$$

Apesar do fato da regressão quantílica apresentar elevada adaptabilidade à heteroscedasticidade dos dados, resultando em intervalos de largura maior em regiões de incerteza relativamente mais elevada, a regressão quantílica não apresenta garantia estatística de cobertura, o que é resolvido com a utilização do conjunto de calibração como parte do método CQR.

O método proposto por [2] consite em realizar predições em cada uma das observações pertencentes ao conjunto de calibração e determinar os respectivos resíduos conforme representado a seguir.

$$\varepsilon_i = max\{\hat{q}_{\alpha_{inf}}(X_i) - y_i, y_i - \hat{q}_{\alpha_{sup}}(X_i)\} \tag{8}$$

Dessa forma, nos casos em que o valor y_i estiver fora do intervalo quantílico por ser menor que o intervalo inferior, o resíduo calculado é positivo e tem magnitute igual a $|Y_i - \hat{q}_{\alpha_{inf}}(X_i)|$.

No caso em que o valor de y_i for maior que o limite superior do intervalo, de maneira análoga, o resíduo calculado também é positivo e apresenta magnitude $y_i - \hat{q}_{\alpha_{sup}}(X_i)$. Por fim, nos casos em que o valor y_i estiver contido dentro do intervalo, o resíduo terá valor negativo com magnitude igual a distância entre o valor real e o limite do intervalo mais próximo.

Uma vez que os resíduos foram calculados para todas as observações pertencentes ao conjunto de calibração, a distribuição dos valores de ε é analisada e o quantil $Q_{1-\alpha}$ é obtido dessa distribuição. Uma vez conhecido o valor de $\hat{q}$, esse termo é utilizado para calibrar os intervalos da regressão quantílica de seguinte forma:

$$\hat{C}(x) = [\hat{q}_{\alpha_{inf}}(X_{n+1}) - Q_{1-\alpha}(\varepsilon), \hat{q}_{\alpha_{sup}}(X_{n+1}) + Q_{1-\alpha}(\varepsilon)] \tag{9}$$

Através do processo de calibração descrito anteriormente, a garantia de cobertura estatística do intervalo é atingida [2]. Com base na formulação apresentada em 9 é possível deduzir que a conformação pode aumentar a largura do intervalo interquartílico para os casos em que as predições realizadas sobre o conjunto de calibração não estejam contidos no intervalo com proporção superior a $1 - \alpha$, ou de reduzi-la nos casos em que uma proporção superior a $1 - \alpha$ das predições estejam contidas no intervalo interquartílico.

Referências

[1] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[2] Y. Romano, E. Patterson, and E. Candès. *Conformalized Quantile Regression*. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2019.

[3] R. F. Barber. *Is distribution-free inference possible for binary regression?* Electronic Journal of Statistics, 14(2):3487–3524, 2020. DOI: 10.1214/20-EJS1749. Available at: https://doi.org/10.1214/20-EJS1749.

[4] R. W. Koenker and G. Bassett, Jr. *Regression Quantiles*. Econometrica, 46(1):33-50, January 1978. Available at: https://ideas.repec.org/a/ecm/emetrp/v46y1978i1p33-50.html.