Prédiction conforme

H. Hamouda, M. Megdiche

Institut National des Sciences Appliquées de Toulouse
3MIC-MA

20 mai 2025

Encadré par : Prof. Joseba Dalmau

Résultat de la prédiction conforme

Théorème

Soit, pour $i = 1, \ldots, n$,

$$(X_i, Y_i) \sim P_{\mathcal{X}\mathcal{Y}}$$

une suite de couples de variables aléatoires, indépendantes et identiquement distribuées. Les (X_i, Y_i) prennent leurs valeurs dans $\mathcal{X} \times \mathcal{Y}$. On définit également α, le niveau d'erreur pour notre prédiction.
On construit $\hat{C}$ telle que :

$$\hat{C} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$$

où $\mathcal{P}(\mathcal{Y})$ désigne l'ensemble des parties de $\mathcal{Y}$.
Cette fonction a pour propriété que, pour une nouvelle paire de test $(X_{\text{test}}, Y_{\text{test}}) \sim P_{\mathcal{X}\mathcal{Y}}$,

$$\hat{C}(X_{\text{test}}) = \{y \in \mathcal{Y} \mid s(X_{\text{test}}, y) < q\}$$

où s est une fonction de score et q un seuil défini ci-dessous.
On a donc :

$$\mathbb{P}\left(Y_{\text{test}} \in \hat{C}(X_{\text{test}})\right) \geq 1 - \alpha$$

Résultat de la prédiction conforme

suite

les scores de calibration sont tels que :

- $s_1 = s(X_1, Y_1), \ldots, s_n = s(X_n, Y_n)$
- q est le $\lceil (n+1)(1-\alpha) \rceil$-quantile de la suite $(s_1, \ldots, s_n)$

Applications du théorème de la prédiction conforme

Problème de regression linéaire : Soit

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. sur $\mathcal{X} \times \mathbb{R}$ un ensemble de calibration
- $(X_{\text{test}}, Y_{\text{test}})$ un point test indépendant issu de la même distribution.
- $\hat{f} : \mathcal{X} \to \mathbb{R}$ un modèle de régression linéaire entraîné sur un autre jeu de données on a donc dans ce cas :
  - Les scores de non-conformité : $s_i = |Y_i - \hat{f}(X_i)|$ pour $i = 1, \ldots, n$;
  - Le quantile empirique conforme :

$$\hat{q} = \text{score au rang } \lceil (n+1)(1-\alpha) \rceil \text{ parmi les } s_i;$$

  - L'ensemble prédictif conforme :

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

  est continu

Alors, l'ensemble $\hat{C}_n(x)$ satisfait la propriété :

$$\mathbb{P}\left(Y_{\text{test}} \in \hat{C}_n(X_{\text{test}}) \right) \geq 1 - \alpha$$

où la probabilité est prise conjointement sur les données de calibration et le point test.

visualisation

conformal_prediction_regression.png

Applications du théorème de la prédiction conforme

Problème de classification : on a $\mathcal{Y} = \{1, \ldots, K\}$ un ensemble fini de classes.

- $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ paires i.i.d. de distribution inconnue sur $\mathcal{X} \times \mathcal{Y}$, ensemble de calibration, $(X_{\text{test}}, Y_{\text{test}})$ une paire test indépendante et $(X_{\text{test}}, Y_{\text{test}}) \sim P_{\mathcal{X}\mathcal{Y}}$.
- $\hat{f} : \mathcal{X} \to \Delta_K$ un modèle de classification MNIST entraîné, où Δ_K est le simplexe de dimension $K - 1$ (i.e. l'ensemble des vecteurs de probabilités de classe).
  On note $\hat{f}(x) = (p_1(x), \ldots, p_K(x))$ les probabilités prédites pour chaque classe.
  Définissons :
  - Le score de non-conformité pour chaque point de calibration :
  $$s_i = 1 - \hat{f}_{Y_i}(X_i), \quad \text{pour } i = 1, \ldots, n,$$
  où $\hat{f}_{Y_i}(X_i)$ est la probabilité attribuée par le modèle à la vraie classe Y_i.
  - Le seuil conforme comme le quantile empirique :
  $$\hat{q} = \text{score au rang } \lceil (n+1)(1-\alpha) \rceil \text{ parmi les } \{s_i\}_{i=1}^n.$$
  - L'ensemble de prédiction conforme :
  $$\hat{C}_n(x) = \left\{ y \in \mathcal{Y} : 1 - \hat{f}_y(x) \leq \hat{q} \right\}.$$

Alors, sous l'hypothèse d'échangeabilité des $n + 1$ exemples, on a la garantie suivante :

$$\mathbb{P}\left(Y_{\text{test}} \in \hat{C}_n(X_{\text{test}}) \right) \geq 1 - \alpha,$$

Document minimal

```
\documentclass{article}
\begin{document}
Tout ce que je veux afficher dans mon document
\end{document}
```

Mathématiques

Modes mathématiques

- En ligne : $...$ ou \(...\)
- Centré :

$$\cdots$$

ou $$...$$

Exemple

Soit x une variable réelle solution de l'équation :

$$ax^2 + bx + c = 0$$

Le discriminant vaut $\Delta = b^2 - 4ac$. S'il est strictement positif, il y a deux racines réelles distinctes :

$$x_1 = \frac{-b - \sqrt{\Delta}}{2a}, \quad x_2 = \frac{-b + \sqrt{\Delta}}{2a}$$

Exercices avancés

1. Navier-Stokes :

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \nabla)\, \vec{v} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 \vec{v} + \vec{f}$$

2. Lotka-Volterra :

$$\frac{dx}{dt} = x(\alpha - \beta y), \quad \frac{dy}{dt} = -y(\gamma - \delta x)$$

3. Intégrale gaussienne :

$$\delta \int_0^\infty \int_0^\infty e^{-(x^2+y^2)}\, dx\, dy = \frac{\pi}{4}$$

Bibliographie

Exemple de fichier `.bib` :

```
@BOOK{HofbSigm98,
title = {Evolutionary Games and Population Dynamics},
publisher = {Cambridge University Press},
year = {1998},
author = {Joseph Hofbauer, Karl Sigmund}
}
```

Conclusion

Pour aller plus loin : http://www.jalix.org/ressources/miscellaneous/tex/_faq-latex2/html/