

Qualité de prévision, risque et estimation du risque

Manon MAHEO - Valentin PENISSON

18/10/2021

Introduction à l'estimation de l'erreur de prévision

La performance du modèle statistique ou algorithme statistique s'évalue par un **risque** ou une **erreur de prévision**, dite encore **erreur de généralisation** dans le cas de la régression et de la classification. Une estimation du risque est importante dans le sens où elle guide dans la stratégie de choix de méthodes et de choix de modèles en science des données. Une mesure de la qualité ou de la performance du modèle permet aussi de considérer la confiance que l'on peut accorder à la prévision du modèle.

On considère que l'on dispose d'un **échantillon de données observées de type entrée-sortie** de taille n : $d_1^n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ avec $x_i \in \mathcal{X}$ quelconque (souvent égal à \mathbb{R}^p), $y_i \in \mathcal{Y}$ pour $i = 1 \dots n$. L'objectif, pour tout algorithme ou modèle statistique, est de prédire la sortie y associée à une nouvelle entrée x , sur la base de d_1^n . Cette sortie peut être quantitative (i.e $\mathcal{Y} \in \mathbb{R}^1$) et nous sommes en *régression*, ou bien qualitative (i.e $\mathcal{Y} = \{1 \dots K\}$ ou $\mathcal{Y} = \{-1, 1\}$) et nous parlons de *discrimination/classification supervisée* ou de *discrimination binaire*. Une **règle de prédiction ou un algorithme de prévision (en régression ou en discrimination)** est donc la fonction mesurable $f : \mathcal{X} \rightarrow \mathcal{Y}$ qui associe la sortie $f(x)$ à l'entrée $x \in \mathcal{X}$.

Une fois que la notion de modèle statistique ou de règle de prévision est précisée, le **risque** est défini à partir d'une *fonction perte* associée.

En pratique, ce risque nécessite d'être estimé et différentes stratégies sont proposées.

Les différentes techniques de ré-échantillonnage

Approche de validation croisée hold-out

Approche de validation croisée p out

Approche de validation croisée K fold

Algorithme de Leave-One-Out bootstrap

Conclusion