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 : $d_1^n = \{(x_1, y_1), ..., (x_n, y_n)\}$ avec $x_i \in \mathcal{X}$ quelconque (souvent égal à $\mathbb{R}^p$), $y_i \in \mathcal{Y}$ pour i = 1...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 $\mathcal{Y} \in \mathbb{R}^{\lceil}$) et nous sommes en *régression*, ou bien qualitative (i.e $\mathcal{Y} = \{1...K\}$ ou $\mathcal{Y} = \{-1,1\}$) et nous parlons de *discrimination/classification supervisée* ou de *discrimination binaire*. Nous nous plaçons ici dans le cadre de l'apprentissage statistique supervisé c'est-à-dire que l'on connaît ce que l'on doit expliquer (i.e les sorties y_i). Une **règle de prédiction ou un algorithme de prévision (en régression ou en discrimination)** est donc la fonction mesurable $f : \mathcal{X} \to \mathcal{Y}$ 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. Soit $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ une fonction de perte, le **risque ou l'erreur de généralisation** d'une règle de prédiction f est défini par $R_p(f) = \mathbb{E}_{(X,Y)}[l(Y, f(X)]$. En pratique, ce risque nécessite d'être estimé et différentes stratégies sont proposées puisque l'on suppose que d_1^n est l'observation d'un n-échantillon $D_1^n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ d'une loi conjointe P sur $\mathcal{X} \times \mathcal{Y}$, totalement inconnue et que x est une observation de la variable X, (X,Y) étant un couple aléatoire de loi conjointe P indépendant de D_1^n.

Une première idée serait d'utiliser le **risque empirique** d'un algorithme de prédiction pour estimer le risque moyen. Néanmoins, ce dernier qui exprime la *qualité d'ajustement du modèle sur l'échantillon observé*, constitue une mesure biaisée de l'erreur de prévision. Celui-ci est lié aux données qui ont servi à l'ajustement du modèle et est d'autant plus faible que le modèle est complexe. Sélectionner la complexité d'un modèle en minimisant le risque empirique conduit à un risque de ***sur-apprentissage (overfitting)***.

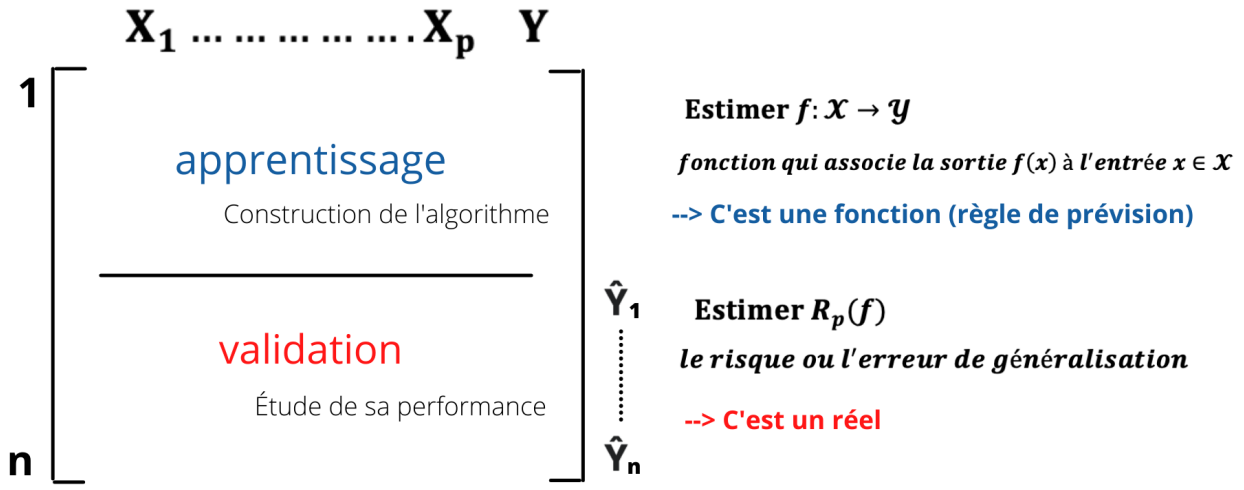
Ainsi, la façon la plus simple d'estimer sans biais ou de biais réduit l'erreur de prévision consiste à utiliser un échantillon indépendant n'ayant pas participé à l'estimation du modèle. Plusieurs stratégies, étudiées ci-dessous, sont proposées pour **éviter d'utiliser les mêmes données pour estimer un modèle et une erreur**.

Les différentes techniques de ré-échantillonnage

Approche de validation croisée ou croisée hold-out

L'**approche de validation croisée hold-out** consiste à séparer l'échantillon de données d_1^n en ***deux*** : un *échantillon d'apprentissage* $d_{n,app}$ pour construire l'algorithme de prédiction et un *échantillon de validation*

ou de test $d_{n,tes}$ pour estimer le risque de la règle de prévision.



Cette approche nécessite d'avoir un nombre suffisant d'observations dans l'échantillon d'apprentissage pour bien ajuster l'algorithme de prévision ainsi qu'un nombre suffisant d'observations dans l'échantillon de test pour bien estimer l'erreur de l'algorithme. Par exemple, la taille de l'échantillon d'apprentissage peut osciller entre 60% et 90% de la taille totale de l'échantillon.

L'inconvénient de cette approche est que l'erreur est **très dépendante de la partition "Apprentissage/Validation"**. Donc il faut au minimum la reproduire avec plusieurs partitions.

Approche de validation croisée leave-p-out

Approche de validation croisée K-fold

Algorithme de bootstrap

Conclusion