Évaluation de modeles

Cette section est basée sur James et al. (2021), chapitre 5.

Évaluer la performance d'un modèle prédictif

Dans la section précédente, nous avons introduit des outils pour mesurer la qualité d'un estimateur : l'erreur quadratique moyenne (MSE) pour les variables quantitatives et le taux d'erreur (ER) pour les variables qualitatives. Ces mesures comparent les valeurs prédites $\widehat{Y} = \hat{f}(X)$ aux valeurs observées Y. Cependant, si l'on calcule ces erreurs uniquement à partir des données qui ont servi à entraîner le modèle, on risque de **sous-estimer** la véritable erreur de prédiction. Pourquoi ? Parce que l'estimateur $\hat{f}$ a été ajusté pour minimiser l'erreur sur ces mêmes données. Il s'y adapte donc bien, et généralement, trop bien ! Cela peut conduire à l'illlusion que notre modèle est performant. En effet, un modèle très flexible peut avoir une erreur faible sur les données d'entraînement simplement parce qu'il capture le bruit plutôt que le signal. Mais si le modèle s'adapte trop aux données d'entraînement, il risque de mal généraliser à de nouvelles données, i.e. des données qu'il n'a jamais vues. Ce phénomène s'appelle le **sur-ajustement** (*overfitting*).

> Remarque : Sur-ajustement et sous-ajustement
>
> Un modèle trop flexible peut s'adapter parfaitement aux données d'entraînement, y compris au bruit aléatoire. Il aura une erreur faible sur ces données mais une erreur élevée sur de nouvelles observations. On dira qu'il y a **sur-ajustement** (*overfitting*) du modèle.
> À l'inverse, un modèle trop rigide (par exemple, une droite constante) ne pourra pas capturer la structure des données, même sur l'ensemble d'entraînement. On dira qu'il y a **sous-ajustement** (*underfitting*) du modèle.
> L'objectif est de trouver le bon compromis entre flexibilité et capacité de généralisation.

Pour évaluer objectivement un modèle, l'idéal serait de le tester sur des données complètement indépendantes de celles utilisé pour l'apprentissage. On distingue donc deux ensembles : un **jeu d'entraînement**, utilisé pour ajuster le modèle et un **jeu de test**, utilisé pour évaluer la performance prédictive du modèle. En practique, nous n'avons généralement pas accès à un jeu

de test pour faire cette évaluation. Dans cette section, nous allons deux approches permettant de contourner ce problème.

Jeu de données de validation

Quand on ne dispose que d'un seul jeu de données, une solution simple consiste à le diviser **aléatoirement** en deux sous-ensembles : un **jeu d'entraînement** pour ajuster le modèle et un **jeu de validation** pour estimer l'erreur de prédiction. On parle alors d'approche par jeu de validation. La Figure 1 présente un schéma de cette approche.

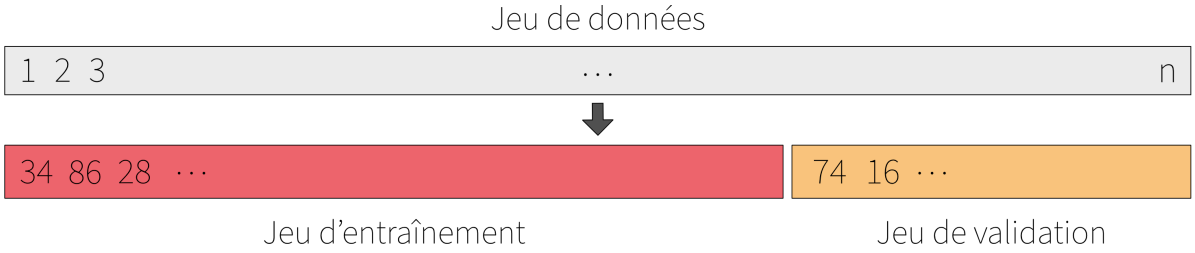

FIGURE 1 – Schéma de l'approche par jeu de validation.

Add exemple !

> Comment choisir comment découper le jeu de données ?
>
> La réponse simple : avec de la pratique et une connaissance du domaine. S'il y a beaucoup de données, on peut faire 50-50. Sinon, on peut partir un 70-30.

Désavantage de la méthode :

1. L'erreur de prédiction calculée sur le jeu de données de validation peut être très variable. En effet, elle dépend du nombre d'observations dans le jeu de validation et de quelles sont ses données.
2. On a moins de données pour apprendre le modèle. Comme les modèles ont tendance à moins bien apprendre avec moins de données, l'estimation de l'erreur sur le jeu de validation a tendance à surestimer l'erreur que l'on aurait avec un jeu de test et un modèle appris sur le jeu de données complet.

Validation croisée

Comme l'approche par jeu de données de validation, la validation croisée consiste à faire des sous-ensembles du jeu de données. L'approche consiste à découper de façon aléatoire l'ensemble des observations en K groupes de taille équivalanetes. Le premier sous-ensemble est utilisé comme jeu de données de validation et le modèle est appris sur les $K-1$ autres sous-ensembles.

L'erreur de prédiction est calculé sur le premier sous-ensemble. Cette procédure est faite K fois ; à chaque un différent sous-ensemble est utilisé comme jeu de données de validation. À la fin, on a donc K valeurs pour l'erreur de prédiction. On calcule enfin la moyenne des K valeurs de prédiction.

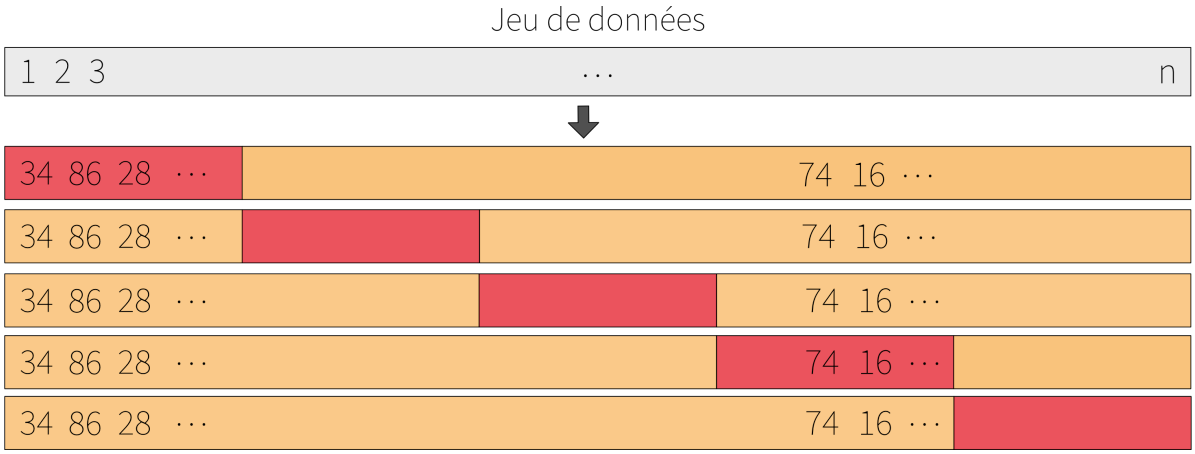

FIGURE 2 – Schéma de l'approche par validation croisée

Add exemple !

Comment choisir le nombre de sous-ensemble ?

En pratique, on utilise $K = 5$ ou $K = 10$. Cela a un avantage computationnel car le modèle doit être appris K fois.

James, Gareth, Daniela Witten, Trevor Hastie, et Robert Tibshirani. 2021. *An Introduction to Statistical Learning : With Applications in R*. Springer Texts in Statistics. New York, NY : Springer US. https://doi.org/10.1007/978-1-0716-1418-1.