

Pense-bête VIP : Astuces de Machine Learning

Afshine AMIDI et Shervine AMIDI

6 octobre 2018

Indicateurs dans le contexte de la classification

Dans le contexte de la classification binaire, voici les principaux indicateurs à surveiller pour évaluer la performance d'un modèle.

□ **Matrice de confusion** – Une matrice de confusion est utilisée pour avoir une image complète de la performance d'un modèle. Elle est définie de la manière suivante :

		Classe prédite	
		+	-
Classe vraie	+	TP True Positives	FN False Negatives Type II error
	-	FP False Positives Type I error	TN True Negatives

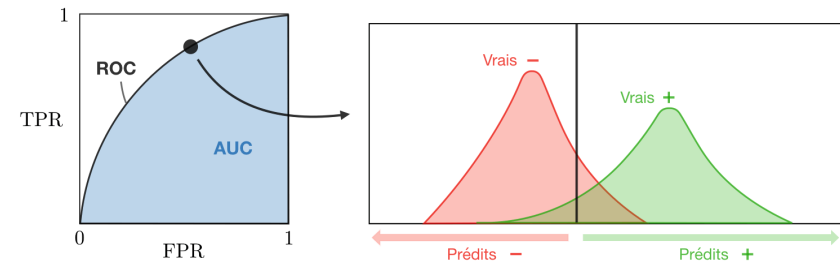
□ **Indicateurs principaux** – Les indicateurs suivants sont communément utilisés pour évaluer la performance des modèles de classification :

Indicateur	Formule	Interprétation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Performance globale du modèle
Precision	$\frac{TP}{TP + FP}$	À quel point les prédictions positives sont précises
Recall Sensitivity	$\frac{TP}{TP + FN}$	Couverture des observations vraiment positives
Specificity	$\frac{TN}{TN + FP}$	Couverture des observations vraiment négatives
F1 score	$\frac{2TP}{2TP + FP + FN}$	Indicateur hybride pour les classes non-balancées

□ **Courbe ROC** – La fonction d'efficacité du récepteur, plus fréquemment appelée courbe ROC (de l'anglais *Receiver Operating Curve*), est une courbe représentant le taux de *True Positives* en fonction de taux de *False Positives* et obtenue en faisant varier le seuil. Ces indicateurs sont résumés dans le tableau suivant :

Indicateur	Formule	Equivalent
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity

□ **AUC** – L'aire sous la courbe ROC, aussi notée AUC (de l'anglais *Area Under the Curve*) ou AUROC (de l'anglais *Area Under the ROC*), est l'aire sous la courbe ROC comme le montre la figure suivante :



Indicateurs dans le contexte de la régression

□ **Indicateurs de base** – Étant donné un modèle de régression f , les indicateurs suivants sont communément utilisés pour évaluer la performance d'un modèle :

Somme des carrés totale	Somme des carrés expliquée	Somme des carrés résiduelle
$SS_{\text{tot}} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{\text{reg}} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{\text{res}} = \sum_{i=1}^m (y_i - f(x_i))^2$

□ **Coefficient de détermination** – Le coefficient de détermination, souvent noté R^2 ou r^2 , donne une mesure sur la qualité du modèle et est tel que :

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

□ **Indicateurs principaux** – Les indicateurs suivants sont communément utilisés pour évaluer la performance des modèles de régression, en prenant en compte le nombre de variables n qu'ils prennent en considération :

Cp de Mallow	AIC	BIC	R^2 ajusté
$\frac{SS_{\text{res}} + 2(n+1)\hat{\sigma}^2}{m}$	$2[(n+2) - \log(L)]$	$\log(m)(n+2) - 2\log(L)$	$1 - \frac{(1-R^2)(m-1)}{m-n-1}$

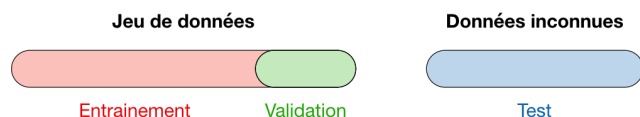
où L est la vraisemblance et $\hat{\sigma}^2$ est une estimation de la variance associée à chaque réponse.

Sélection de modèle

□ **Vocabulaire** – Lors de la sélection d'un modèle, on divise les données en 3 différentes parties comme suit :

Training set	Validation set	Testing set
<ul style="list-style-type: none"> - Modèle est entraîné - Normalement 80% du dataset 	<ul style="list-style-type: none"> - Modèle est évalué - Normalement 20% du dataset - Aussi appelé hold-out ou development set 	<ul style="list-style-type: none"> - Modèle donne des prédictions - Données jamais vues

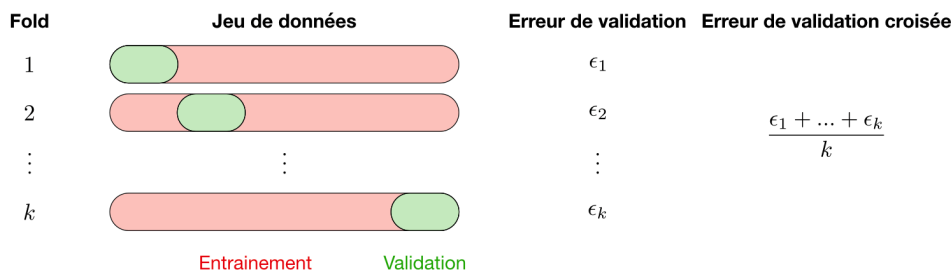
Une fois que le modèle a été choisi, il est entraîné sur le jeu de données entier et testé sur test set (qui n'a jamais été vu). Ces derniers sont représentés dans la figure ci-dessous :



□ **Validation croisée** – La validation croisée, aussi notée CV (de l'anglais *Cross-Validation*), est une méthode qui est utilisée pour sélectionner un modèle qui ne s'appuie pas trop sur le training set de départ. Les différents types de validation croisée rencontrés sont resumés dans le tableau ci-dessous :

k -fold	Leave- p -out
<ul style="list-style-type: none"> - Entrainement sur $k - 1$ folds et évaluation sur le fold restant - Généralement $k = 5$ ou 10 	<ul style="list-style-type: none"> - Entrainement sur $n - p$ observations et évaluation sur les p restantes - Cas $p = 1$ est appelé <i>leave-one-out</i>

La méthode la plus utilisée est appelée validation croisée k -fold et partage le jeu de données d'entraînement en k folds, de manière à valider le modèle sur un fold tout en trainant le modèle sur les $k - 1$ autres folds, tout ceci k fois. L'erreur est alors moyennée sur k folds et est appelée erreur de validation croisée.



□ **Régularisation** – La procédure de régularisation a pour but d'éviter que le modèle ne surapprenne (en anglais *overfit*) les données et ainsi vise à régler les problèmes de grande variance. Le tableau suivant récapitule les différentes techniques de régularisation communément utilisées.

LASSO	Ridge	Elastic Net
<ul style="list-style-type: none"> - Réduit les coefficients à 0 - Bon pour la sélection de variables 	Rend les coefficients plus petits	Compromis entre la sélection de variables et la réduction de coefficients
$\dots + \lambda \theta _1$ $\lambda \in \mathbb{R}$	$\dots + \lambda \theta _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda \left[(1 - \alpha) \theta _1 + \alpha \theta _2^2 \right]$ $\lambda \in \mathbb{R}, \alpha \in [0, 1]$

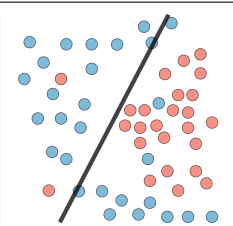
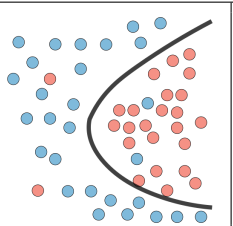
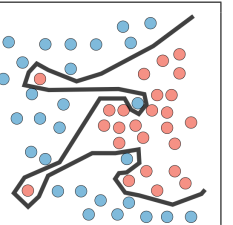
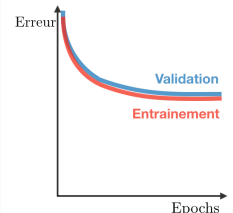
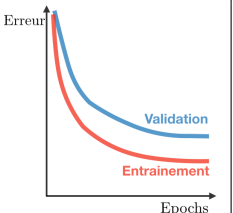
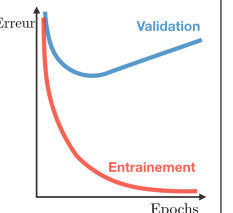
Diagnostics

□ **Biais** – Le biais d'un modèle est la différence entre l'espérance de la prédiction et du modèle correct pour lequel on essaie de prédire pour des observations données.

□ **Variance** – La variance d'un modèle est la variabilité des prédictions d'un modèle pour des observations données.

□ **Compromis biais/variance** – Plus le modèle est simple, plus le biais est grand et plus le modèle est complexe, plus la variance est grande.

	Underfitting	Just right	Overfitting
Symptômes	<ul style="list-style-type: none"> - Erreur de training élevé - Erreur de training proche de l'erreur de test - Biais élevé 	<ul style="list-style-type: none"> - Erreur de training légèrement inférieure à l'erreur de test 	<ul style="list-style-type: none"> - Erreur de training très faible - Erreur de training beaucoup plus faible que l'erreur de test - Variance élevée
Régression			

Classification			
Deep Learning			
Remèdes	<ul style="list-style-type: none"> - Complexifier le modèle - Ajouter plus de variables - Laisser le training pendant plus de temps 		<ul style="list-style-type: none"> - Effectuer une régularisation - Avoir plus de données

□ **Analyse de l'erreur** – L'analyse de l'erreur consiste à analyser la cause première de la différence en performance entre le modèle actuel et le modèle parfait.

□ **Analyse ablative** – L'analyse ablative consiste à analyser la cause première de la différence en performance entre le modèle actuel et le modèle de base.