Régression logistique et Classification

Erwan, Patricia, Céline, Jérémy

SOMMAIRE

- Principe de la régression logistique
- Fonction Logistique (Sigmoid)(équation courbe)
- Principe de la classification
- Classification Binaire Multi classes (avec un exemple pratique)
- Evaluation des méthodes de la classification
  - Matrice de Confusion
  - Accuracy
  - Précision
  - Rappel
  - F1

Principe de la Régression Logistique

La régression logistique est une technique prédictive. Elle vise à construire un modèle permettant de mesurer l'association entre la survenue d'un évènement (variable expliquée qualitative) et les facteurs susceptibles de l'influencer (variables explicatives).

Cette approche statistique peut être employée pour évaluer et caractériser les relations entre une variable réponse de type binaire (par exemple : Vivant / Mort, Malade / Non malade, succés / échec), et une, ou plusieurs, variables explicatives, qui peuvent être de type catégoriel (le sexe par exemple), ou numérique continu (l'âge par exemple).

Principe de la Régression Logistique

Dans la régression logistique, ce n'est pas la réponse binaire (malade/pas malade) qui est directement modélisée, mais la probabilité de réalisation d'une des deux modalités (être malade par exemple)

Cette probabilité de réalisation ne peut pas être modélisée par une droite car celle-ci conduirait à des valeurs <0 ou >1. Ce qui est impossible puisqu'une probabilité est forcément bornée par 0 et 1.

Fonction Logistique Sigmoid

Equation et Courbe

En mathématiques, les **fonctions logistiques dite de Verhulst**

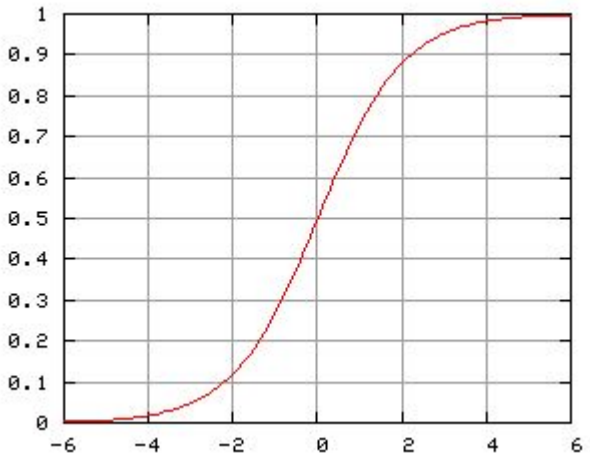
sont les fonctions ayant pour expression

$$f(t) = K\frac{1}{1+ae^{-rt}}$$ où K et r sont des réels positifs et a un réel quelconque.

Pour a > 0, la fonction est une fonction logistique Sigmoid(en forme de S).

La fonction logistique peut être utilisée pour illustrer l'état d'avancement

de la diffusion d'une innovation durant son cycle de vie, par exemple.



Exemple : une société de vidéo surveillance commercialise des caméras à infra rouges cachées dans des nains de jardin. L' évolution des ventes annuelles se présente comme suit :

année	1	2	3	4	5	6	7	8	9	10
ventes	15	35	80	200	745	1 870	2 505	3 110	3 885	4 200

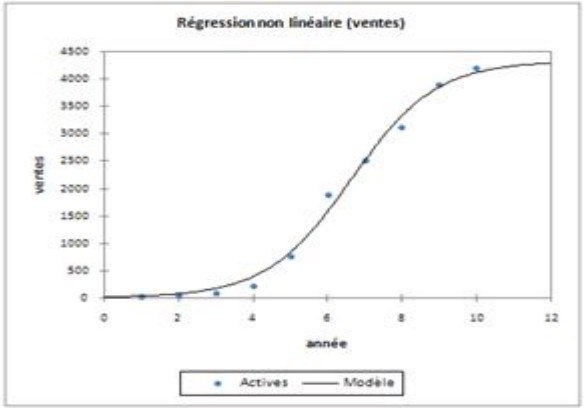
La régression sera réalisée par XLSTAT. On ne sélectionne pas *« régression logistique »* , mais *« régression non linéaire »*. La fonction logistique telle que décrite ci-dessus est celle à trois paramètres (a,b,c): ***« pr3/(1+Exp(-pr1-pr2*X1)) ».***

pr3 correspond à la valeur de saturation c et si *pr2* n'est autre que a, il n'en est pas de même de l'autre paramètre puisque la formule de XLSTAT fait intervenir l'exponentielle d'une fonction affine. Pour s'y retrouver, il faut considérer que *pr1* = −ln(b).

Coefficients d'ajustement :

Observation	10,000
DDL	7,000
R^2	0,993
SCE	194930,985
MCE	27847,284
RMCE	166,875
Itérations	12,000

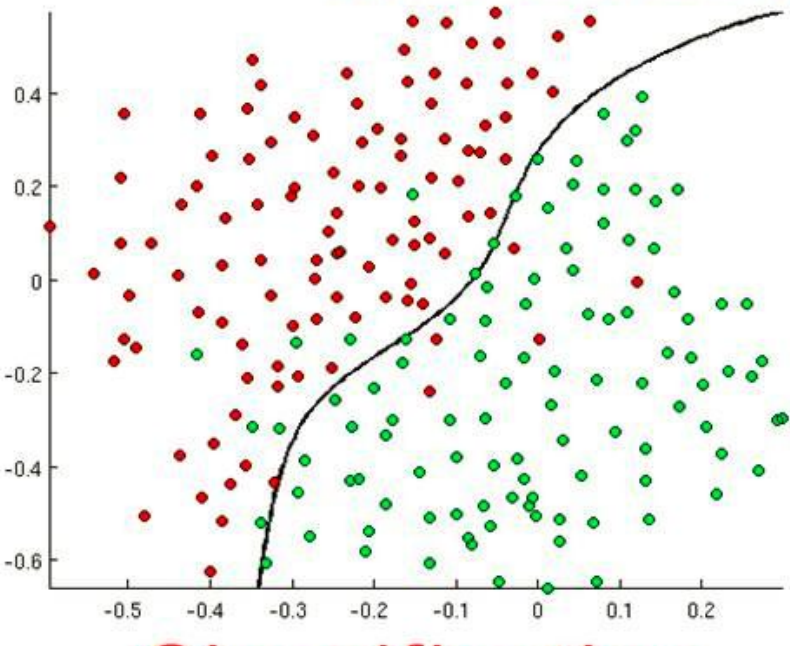
On trouve pour le modèle un R^2 de 0.99, ce qui signifie que le modèle est performant.



Principe de la Classification

Définition : La classification statistique réside dans l'identification de la catégories à laquelle un nouveau élément appartient, sur la base d'un data set d'entraînement de données contenant des observations (ou instances) dont la catégorie est connue.

Attribuer un courrier électronique donné à la classe "spam" ou "non spam" et attribuer un diagnostic à un patient donné en fonction des caractéristiques observées du patient, sont des exemples de classification.



Classification Binaire & Multi-classes

Un problème de **classification binaire** consiste à trouver un moyen de séparer deux nuages de points : en classant ces propositions dans un ou deux ensembles

Exemple de la girafe :

Dire si une image représente une girafe ou non.

Si oui, on dit que cette image est positive ; sinon, qu'elle est négative

la classification en **classes multiples** est un processus de répartition d'un lot de propositions entre plus de deux ensembles

Binary CES

Class 0 Test Values: 0 1 2 3 4 5

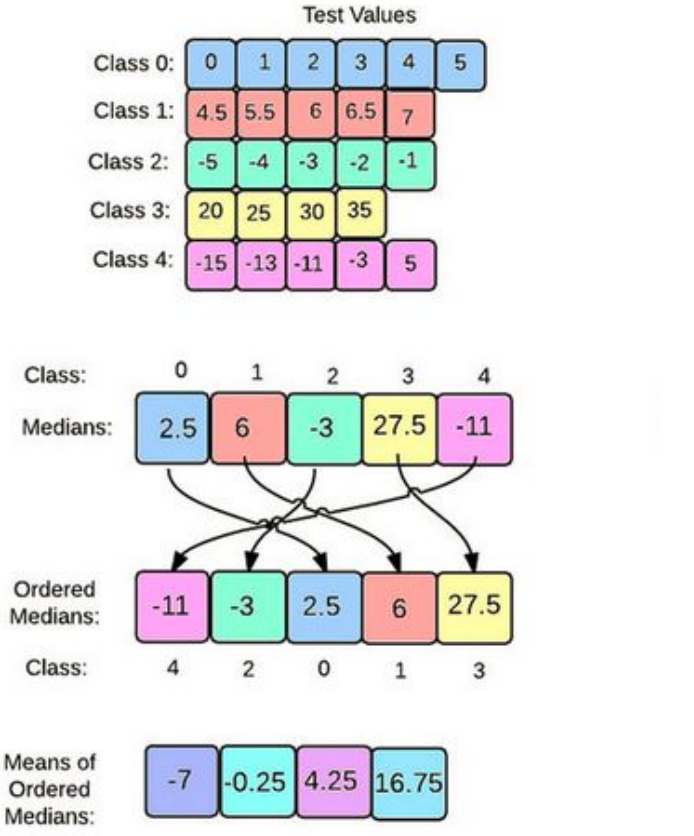
Class 1 Test Values: 4.5 5.5 6 6.5 7

Class 0 Median: 2.5

Class 1 Median: 6

Mean of Medians: 4.25

Multi-Class CES



Evaluation des méthodes de la classification - Matrice de confusion

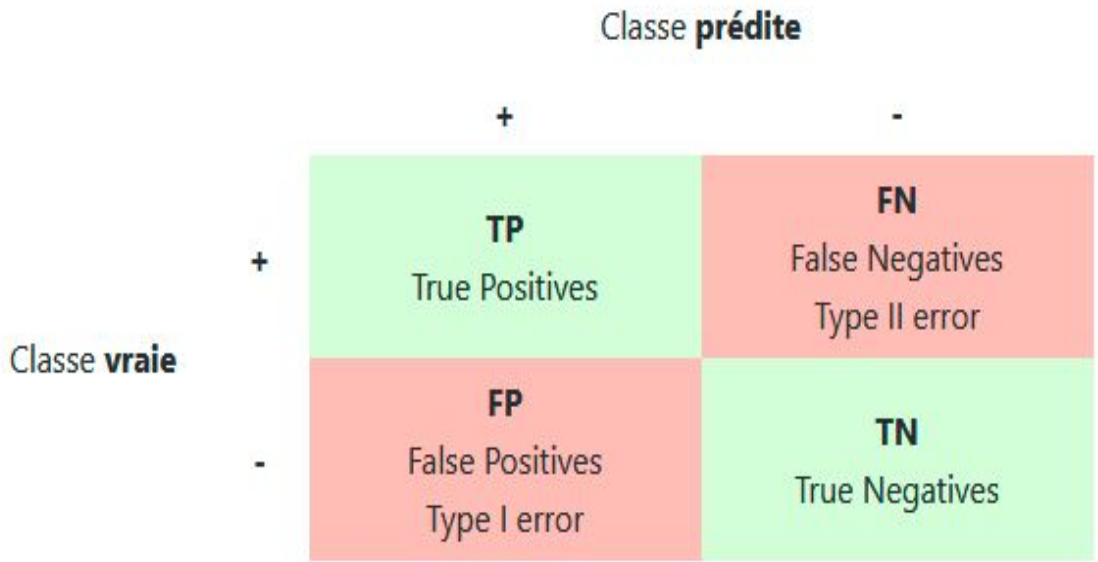
En apprentissage automatique supervisé, la matrice de confusion est une matrice qui mesure la qualité d'un système de classification. Chaque ligne correspond à une classe réelle, chaque colonne correspond à une classe estimée.

Un des intérêts de la matrice de confusion est qu'elle montre rapidement si un système de classification parvient à classifier correctement.

Supposons que notre classificateur SAC soit testé avec un jeu de 200 mails, dont 100 sont des courriels pertinents et les 100 autres relèvent de pourriels.

		Classe estimée - *(par le classificateur SAC)*	
		courriel	**pourriel**
Classe réelle - *(selon le destinataire humain des mails)*	**courriel**	95 *(vrais positifs)*	5 *(faux négatifs)*
	pourriel	3 *(faux positifs)*	97 *(vrais négatifs)*

Evaluation des méthodes de la classification - Matrice de confusion



❒ **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
Précision	$\frac{TP}{TP + FP}$	À quel point les prédictions positives sont précises
Rappel Sensibilité	$\frac{TP}{TP + FN}$	Couverture des observations vraiment positives
Spécificité	$\frac{TN}{TN + FP}$	Couverture des observations vraiment négatives
F-mesure	$\frac{2TP}{2TP + FP + FN}$	Indicateur hybride utilisé pour les classes non-balancées

Précision (Valeur prédictive positive)

Définition : La précision est la proportion des éléments pertinents parmi l'ensemble des éléments proposés.

Exemple : Quand un utilisateur interroge une base de données, il souhaite que les documents proposés en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme « précis ».

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Rappel (Sensibilité)

Définition : Le Rappel est la proportion des items pertinents proposés parmi l'ensemble des items pertinents.

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Exemple : Lorsque l'utilisateur interroge la base, il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. À l'inverse, **si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel.**

F - Mesure

- Mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres.

- La **F-mesure** correspond à un compromis de la précision et du rappel donnant la **performance du système**. Ce compromis est donnée de manière simple par la moyenne harmonique de la précision et du rappel.