Sélection et évaluation de modèle - Partie 2

Véronique Tremblay

Objectifs

- Connaître les différentes mesures de performance d’un modèle avec une variable réponse binaire
- Interpréter ces mesures de performance
- Savoir dans quel contexte les utiliser

Comparer des modèles pour Y binaire

En général, les modèles donnent une proportion $\hat{p}_i$ pour chaque observation i.

On obtient une prévision en fixant un seuil s et on prédit

$$\hat{y}_i = 1 \text{ si } \hat{p}_i \geq s$$

$$\hat{y}_i = 0 \text{ si } \hat{p}_i < s$$

Exemple

i	Y	$\hat{p}$	$\hat{Y}$
1	1	0.90	1
2	0	0.55	1
3	1	0.65	1
4	0	0.32	0
5	1	0.35	0
6	0	0.25	0
7	1	0.52	1
8	0	0.45	0
9	1	0.84	1
10	0	0.65	1
11	1	0.89	1
12	0	0.11	0
13	1	0.56	1
14	0	0.26	0
15	1	0.74	1
16	0	0.22	0
17	1	0.59	1
18	0	0.06	0
19	1	0.62	1
20	0	0.55	1

Matrice de confusion

Table 2: Matrice de confusion pour un seuil donné

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	Vrais négatifs (VN)	Faux positifs (FP)
$Y = 1$	Faux négatifs (FN)	Vrai positif (VP)

Matrice de confusion - Exemple

Table 3: Matrice de confusion pour un seuil donné

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	Vrais négatifs (VN)	Faux positifs (FP)
$Y = 1$	Faux négatifs (FN)	Vrai positif (VP)

Exemple avec un seuil de 0,5:

	$\hat{Y} = 0$	$\hat{Y} = 1$
Y = 0	7	3
Y = 1	1	9

Mesures de base

Exactitude

L'exactitude est la proportion des observations qui sont bien classées.

$$\frac{VN + VP}{VP + VN + FP + FN} = \mathbb{P}\left[(Y = 1 \cap \hat{Y} = 1) \cup (Y = 0 \cap \hat{Y} = 0)\right]$$

On utilise parfois les termes *justesse*, *taux de bonne classification* et en anglais *accuracy*. Il arrive qu'on travaille plutôt sur le taux d'erreur, qui est un moins l'exactitude.

La précision

La précision est la proportion de prévisions positives qui sont réellement positives.

$$\frac{VP}{VP + FP} = \mathbb{P}(Y = 1|\hat{Y} = 1)$$

Sensibilité

C'est la proportion d'observation positive détectées par le modèle.

$$\frac{VP}{VP + FN} = \mathbb{P}(\hat{Y} = 1 | Y = 1)$$

En informatique, on utilise le terme *rappel*, de l'anglais *recall*.

Spécifité

C'est la proportion d'observations négatives détectées par le modèle.

$$\frac{VN}{VN + FP} = \mathbb{P}(\hat{Y} = 0 | Y = 0)$$

Score F_β

Score F_1

Pour combiner précision et sensibilité.

$$F_1 = 2 \times \frac{précision \times sensibilité}{précision + sensibilité}$$

Dans l'exemple,

$$F_1 = 2 \times \frac{0.75 \times 0.9}{0.75 + 0.9} = 0.82$$

Lors du choix du modèle, on souhaite un score le plus près possible de 1.

Score F_β

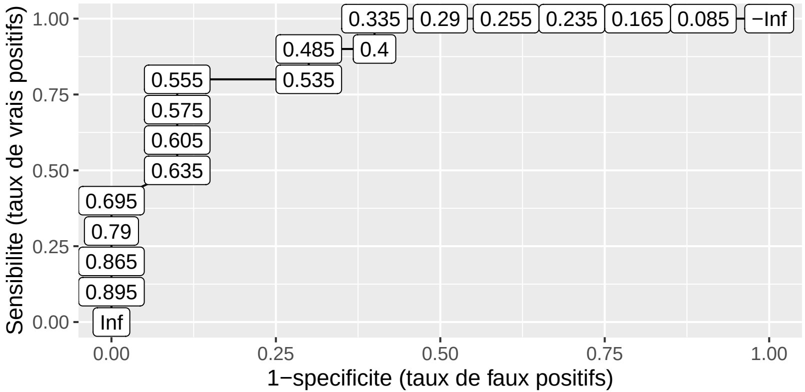
$$F_\beta = \frac{(1+\beta^2) \times \text{précision} \times \text{sensibilité}}{\beta^2 \times \text{précision} + \text{sensibilité}}$$

La courbe ROC

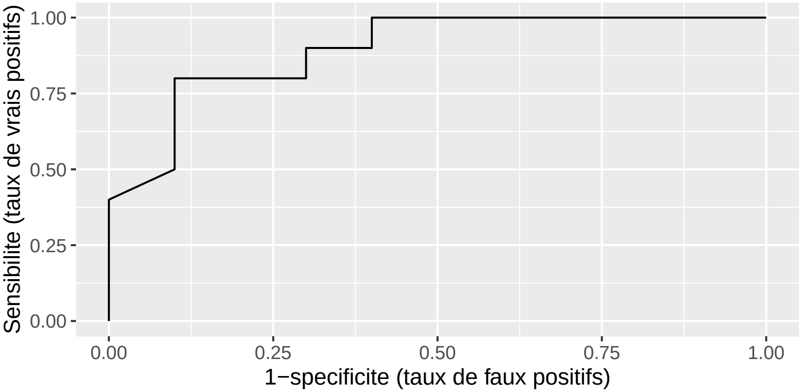
Pour combiner sensibilité et spécificité.

Pour construire une courbe ROC, on calcule la sensibilité et la spécificité pour plusieurs seuils et on reporte ces valeurs sur un graphique.

i	Y	$\hat{p}$
1	1	0.90
11	1	0.89
9	1	0.84
15	1	0.74
3	1	0.65
10	0	0.65
19	1	0.62
17	1	0.59
13	1	0.56
2	0	0.55
20	0	0.55
7	1	0.52
8	0	0.45
5	1	0.35
4	0	0.32
14	0	0.26
6	0	0.25
16	0	0.22
12	0	0.11
18	0	0.06


Sensibilite (taux de vrais positifs)
1−specificite (taux de faux positifs)
0.00
0.25
0.50
0.75
1.00
Inf
0.895
0.865
0.79
0.695
0.635
0.605
0.575
0.555
0.535
0.485
0.4
0.335
0.29
0.255
0.235
0.165
0.085
−Inf

L'aire sous la courbe ROC (AUC)



Levier, taux de réponse et taux de capture

i	Y	$\hat{p}$	$\hat{Y}$	M
1	1	0.90	1	1
11	1	0.89	1	1
9	1	0.84	1	1
15	1	0.74	1	1
3	1	0.65	1	2
10	0	0.65	1	2
19	1	0.62	1	2
17	1	0.59	1	2
13	1	0.56	1	3
2	0	0.55	1	3
20	0	0.55	1	3
7	1	0.52	1	3
8	0	0.45	0	4
5	1	0.35	0	4
4	0	0.32	0	4
14	0	0.26	0	4
6	0	0.25	0	5
16	0	0.22	0	5
12	0	0.11	0	5
18	0	0.06	0	5

Taux de réponse

$$\text{Taux de réponse}_m = \frac{\#((Y=1) \cap (M=m))}{\#(M=m)} = \mathbb{P}(Y=1|M=m)$$

$$\text{Taux de réponse}_m = \frac{\#\text{ Observation positives dans le groupe } m}{\#\text{ Observations dans le groupe } m}$$

m	$\#(M=m)$	$\#((Y=1) \cap (M=m))$	Taux de réponse	Taux de réponse cumulé
1	4	4	1.00	1.000
2	4	3	0.75	0.875
3	4	2	0.50	0.750
4	4	1	0.25	0.625
5	4	0	0.00	0.500

Taux de captures

$$\text{Taux de capture}_m = \frac{\#((Y=1)\cap(M=m))}{\#(Y=1)} = \mathbb{P}(M=m|Y=1)$$

$$\text{Taux de capture}_m = \frac{\#\text{ Observation positives dans le groupe } m}{\#\text{ Observations positives totales}}$$

m	$\#(M=m)$	$\#((Y=1)\cap(M=m))$	Taux de capture	Taux de capture cumulé
1	4	4	0.4	0.4
2	4	3	0.3	0.7
3	4	2	0.2	0.9
4	4	1	0.1	1.0
5	4	0	0.0	1.0

Levier

$$\text{Levier}_m = \frac{\mathbb{P}(Y=1|M=m)}{\mathbb{P}(Y=1)}$$

$$\text{Levier}_m = \frac{\text{Taux de réponse dans le groupe } m}{\text{Taux de réponse global}}$$

m	Taux de réponse	Proportion de l'échantillon	Levier
1	1.00	0.2	2.0
2	0.75	0.4	1.5
3	0.50	0.6	1.0
4	0.25	0.8	0.5
5	0.00	1.0	0.0

Courbe de levier