

Evaluation de modeles

Pourquoi mesurer l'erreur d'un modèle ?

- Choisir le meilleur modèle.
- Avoir une idée de la confiance qu'on peut accorder à notre modèle.

Validation croisée

1. On sépare l'échantillon en K plus de façon aléatoire.
2. Pour k allant de 1 à K :
 - i. On estime $\hat{f}(k)$ en utilisant uniquement les observations qui ne sont pas dans k .
 - ii. On prédit $\hat{Y}(k) = \hat{f}(k)(X(k))$.
 - iii. On calcule $CV(k) = \frac{1}{N_k} \sum_{i=1}^{N_k} L(Y(k), \hat{Y}(k))$.
3. On calcule la moyenne $CV = \frac{1}{K} \sum_{k=1}^K CV(k)$.
4. On répète 1 à 3 pour tous les modèles et on choisit le modèle ont la valeur de CV est la plus basse.

Cas d'un réponse binaire

Dans le cas où il n'existe que deux groupes dans la population, la matrice de confusion est de la forme suivante :

	Positif	Négatif
Positif	Vrais positifs (VP)	Faux négatif (FN)
Négatif	Faux positifs (FP)	Vrais négatif (VN)

- Sensibilité (sensitivity, rappel, recall, hit rate)

Probailité qu'un individu qui appartient au groupe positif soit classé dans le groupe positifs :

$$\frac{VP}{VP + FN}.$$

— Spécifité (specificity)

Probailité qu'un individu qui appartient au groupe négatif soit classé dans le groupe négatif :

$$\frac{VN}{VN + FP}.$$

— Précision

$$\frac{VP}{VP + FP}.$$

— Accuracy

$$\frac{VP + VN}{VP + VN + FP + FN}.$$

— Score F_1 : moyenne harmonique de précision et de sensibilité

$$F_1 = 2 \times \frac{precision \times sensibilit}{precision + sensibilit} = \frac{2VP}{VP + FP + FN}.$$

En pratique, la situation typique va comme suit :

- On a partagé notre jeu de données en jeu d'entraînement et en jeu de validation.
- On a appliqué M méthodes de classification au jeu d'entraînement.
- On aimerait utiliser les données de validation pour choisir la meilleure des M méthodes.

On cherche donc à se doter d'outils qui permettent de mesurer la performance de méthodes de classification.

Mesure triviale

La première étape consiste à comparer le taux de bonnes classification (accuracy) des données de validation obtenu avec chacune des méthodes. On peut faire la même chose avec le score F_1 .

Sensibilité et spécifité

Supposons une classe que nous voulons plus particulièrement identifier, que nous appelons des observations positives et on traite toutes les autres observations comme si elles étaient dans la même classe, soit les observations négatives.

La sensibilité est le ratio entre le nombre de vrais positifs classé positifs et le nombre de vrais positifs. La spécificité est le ratio entre le nombre de vrais négatifs classé négatifs et le nombre de vrais négatifs.

Notons que si on est prêt à vivre avec une spécificité de 0%, on peut toujours avoir une sensibilité de 100% et si on est prêt à vivre avec une sensibilité de 0%, on peut toujours avoir une spécificité de 100%. Si la proportion de négatifs est très élevé, par exemple 99.9%, alors une méthode qui classifie 100% des observations dans la classe négative aura un taux de bonnes classifications de 99.9%. Mais clairement, cette méthode ne nous sera pas très utile si notre but est d'identifier les cas positifs.

On peut exploiter l'estimation d'une probabilité. Les méthodes de classification donnent souvent la probabilité qu'une observation appartienne à une classe, pas uniquement une prévision de la classe elle-même. Une meilleure mesure de la performance de la méthode serait de voir si ces probabilités sont réalistes.

La courbe ROC

La courbe ROC (Receiver Operating Characteristic) peut être vue comme une variation de la courbe de gain.

1. Soit $p_i^{(2)}$, la probabilité que l'observation i appartienne à la classe 2 selon notre modèle.
2. On choisit une valeur de $u \in [0, 1]$ et on prédit que l'observation i appartient à la classe 2 si $p_i^{(2)} \geq u$ et à la classe 1 sinon.
3. On répète l'étape 2 pour $i = 1, \dots, n$ et on calcule la sensibilité $SEN(u)$ et la spécificité $SPE(u)$ ainsi obtenu.
4. On répète les étapes 2 et 3 pour tous les $u \in [0, 1]$, on place les paires $(1 - SPE(u), SEN(u))$ sur un graphique et on lie les points par des segments.

Interprétation :

1. Elle part du point $(0, 0)$ et finit au point $(1, 1)$.
2. Si on classifie les points au hasard, on obtiendra une ligne droite à 45 degrés.
3. Une mesure unique du pouvoir prédictif est l'aire sous la courbe ROC (AUC, Area Under the Curve). Une AUC de 0.5 est ce que l'on espère en classifiant au hasard, et plus on s'approche de 1, meilleure est la méthode.

Exemple