Évaluation de modèles

Quentin Fournier <quentin.fournier@polymtl.ca>

Les diapositives ont été créées par Daniel Aloise <daniel.aloise@polymtl.ca>

basé sur des diapositives de Steven Skiena, 2017

25 mai 2020

Évaluation d'un modèle de classification

- Vous avez construit un modèle prédictif pour la classification
- Est-ce qu'il est bon ?

Évaluation d'un modèle de classification

- Étapes d'une évaluation typique :

Parametres: ce que la fonction va apprendre
Hyperparametres: valeurs choisies au départ et non apprises

  - Les données sont séparées en trois sous-ensembles : entraînement (70%), validation (15%) et test (15%)
  - On fait une liste de valeurs des hyper-paramètres à essayer

liste des valeurs à essayer (l'inverse des paramètres, qui eux vont s'apprendre)

  - Pour chaque élément de cette liste, l'algorithme est exécuté sur l'ensemble d'entraînement et sa performance est mesurées sur l'ensemble de validation
  - Avec les meilleurs valeurs des hyper-paramètres identifiés avec l'ensemble de validation, on calcule la performance de l'algorithme sur l'ensemble de test

Procédure d'évaluation d'un algorithme de classification

- Souvent, nous n'avons pas assez de données pour les séparer en trois sous-ensembles
- La validation croisée partitionne les données en k blocs de taille égale
- Puis, on entraîne k modèles distincts
- Le modèle i est entraîné sur l'union de touts les blocs sauf 1 (le bloc d'index i), et validé sur le bloc i

Procédure d'évaluation d'un algorithme de classification

- Souvent, nous n'avons pas assez de données pour les séparer en trois sous-ensembles
- La validation croisée partitionne les données en k blocs de taille égale
- Puis, on entraîne k modèles distincts
- Le modèle i est entraîné sur l'union de touts les blocs sauf 1 (le bloc d'index i), et validé sur le bloc i
- *Bagging* combine tout les k modèles entraînés dans un seul modèle de classification

Évaluation

- Pour la classification binaire, nous avons quatre résultats possibles

		Classe prédite	
		Oui	Non
		si prévu + et obtenu +	si prévu + mais obtenu -
Classe réelle	Oui	True Positive (TP)	False Negative (FN)
	Non	False Positive (FP)	True Negative (TN)
		Si on a prédit un - et qu'on obtient un +	Si on a prévu - et qu'on obtient un -

Accuracy

- L'*accuracy* est le rapport entre les prédictions correctes et les prédictions totales :

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Un classificateur randomisé aurait une espérance d'accuracy de 50%
- Si on prend juste la classe plus nombreuse comme prédition (classificateur gourmand) accuracy $\geq 50\%$

Précision

- Quand $|P| \ll |N|$, mesurer l'*accuracy* est inutil
  si on a une des classes qui est sur-représenté
- L'accuracy d'un classificateur *gourmand* serait de 95% lorsque $p = \frac{|P|}{|P|+|N|} = 0.05$!
- Nous avons donc besoin d'une mesure d'évaluation plus sensible à l'obtention de la classe positive.

dataset de 10 TP et de 10 FP

10 / (10 + 0) = 1
mais pas trjs très précis!!!

$$precision = \frac{TP}{TP + FP}$$

instances positives prédites

- Atteindre une valeur de precision élévée est impossible pour le classificateur randomisé ou le gourmand

Recall Rappel

- Nous pourrions être plus tolérants aux faux positifs (e.g. erreurs où nous effrayons une personne en bonne santé avec un mauvais diagnostic) que les faux négatifs
- Le recall mesure la capacité d'avoir raison uniquement sur les instances positives

$$recall = \frac{TP}{TP + FN}$$

réelles instance +

- Attention : dire que tout le monde a un cancer donne un recall parfait ! de 1
- Ceci donne un bon recall, mais un mauvais precision

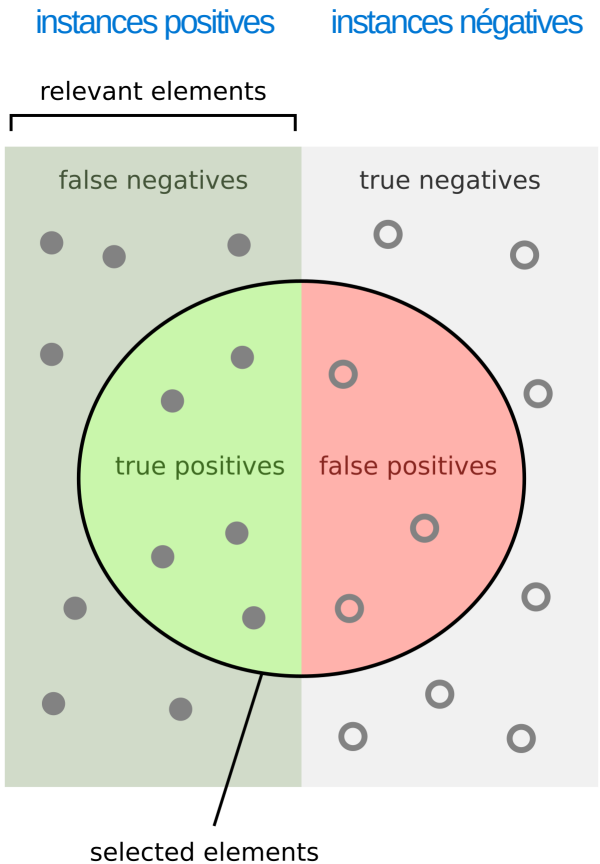
F-score

- Mesure équilibré de score unique

$$F_{score} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- La moyenne harmonique est toujours inférieure ou égale à la moyenne arithmétique, ce qui rend difficile l'obtention d'un F-score élevé

Precision/Recall



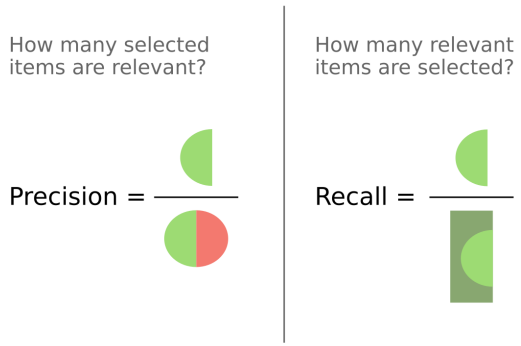


Figure – `https://en.wikipedia.org/wiki/F1_score#/media/File:Precisionrecall.svg`

EXAMEN

Exercice

- Un classifieur dont la precision est plus grande que le recall.
  - Que se passe-t-il ?
  - Comment peut-on l'améliorer ?
- Un classifieur dont le recall est plus grand que la precision.
  - Que se passe-t-il ?
  - Comment on pourrait l'améliorer ?

enregistrement voir exercices dernier cours
environ minute 9

Exercice

+: veut dire positif

- Le precision de ton classificateur est plus grand que son recall.
  - Que se passe-t-il ? Ton classificateur classifie peu des données comme de la classe positive
    peu de prédiction +, mais quand prédiction + il a trjs raison
  - Comment on pourrait l'améliorer ? On pourrait augmenter le poids de l'erreur sur l'entraînement des données de la classe positive

- Le recall de ton classificateur est plus grand que son precision
  prédire trop la classe +, tendance à prédire + meme s'il a tort
  - Que se passe-t-il ? Ton classificateur est trop agressif pour classer les données comme appartenant à la classe positive
  - Comment on pourrait l'améliorer ? On pourrait augmenter le seuil de classification de la classe positive

0: classe -
1: classe +

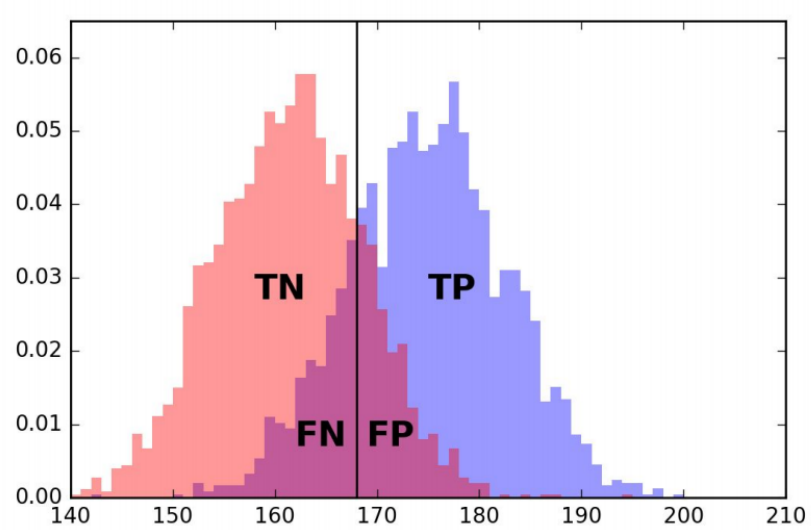
0.5 valeur moyenne
en dessous -
au dessus +

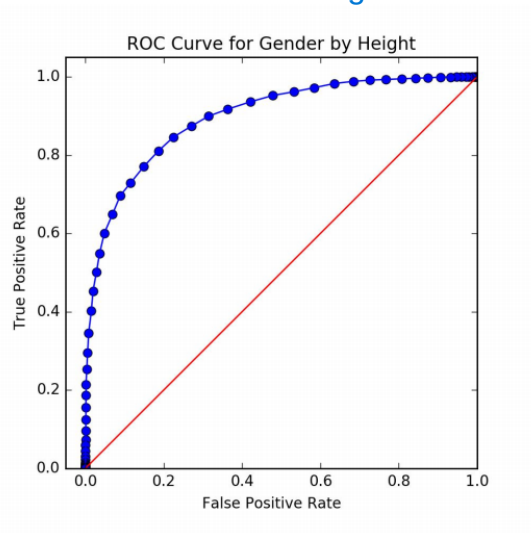
ex: plus haut que le seuil de 0.8, tout ce qui est en dessous de 0.8 est considéré comme -, donc tendance à moins se tromper pour les +

Receiver-Operator Curves (ROC)

- Faire varier le seuil d'un modèle modifie le *recall* et la *precision*
- $(TP)Rate = \frac{TP}{P}$ et $(FP)Rate = \frac{FP}{N}$
- La surface sous une ROC est une bonne mesure d'évaluation d'un modèle

si seuil =1, aucune instance positive
si seuil =0, ca veut dire je prédis tout est positif, aucune instance négative





Skiena, 2017

Évaluation de systèmes multiclasses

- La classification devient plus difficile avec plus des classes
- ex. reconnaissance de chiffres

Digits	0	1	2	3	4	5	6	7	8	9
0	351	0	5	4	2	7	2	1	6	0
1	0	254	0	0	2	0	0	1	1	2
2	1	1	166	4	5	1	3	2	2	1
3	1	2	4	142	0	5	0	1	4	0
4	3	3	8	1	180	3	2	5	4	4
5	0	0	3	11	0	140	3	0	7	1
6	0	2	2	0	4	0	158	0	1	0
7	0	0	2	2	1	0	0	132	2	1
8	2	1	8	0	0	0	2	1	137	1
9	1	1	0	2	6	4	0	4	2	167

Skiena, 2017

- Où les erreurs sont commises principalement ?

Évaluation de systèmes multiclasses

- On note $C[i, j]$ le nombre d'objets de la classe i classés comme de la classe j
- $precision_i$ est la fraction de tous les objets déclarés de la classe i qui étaient en fait de la classe i

C[i,j] --> i: objet de la classe i
j: prédit communs appartenants à la classe j

$$precision_i = C[i,i] / \sum_{j=1}^{k} C[j,i]$$

instance de la classe i correctement prédites de la classe i / instances classe j

- $recall_i$ est la fraction de tous les membres de la classe i qui ont été correctement identifiés comme tel :

$$recall_i = C[i,i] / \sum_{j=1}^{k} C[i,j]$$

correctement prédit classe i / instances de la classe i

Évaluation de systèmes multiclasses

- Un taux de classification trop bas est décourageant et souvent trompeur avec plusieurs classes problèmes quand 1 millier de classes
- Le taux de réussite du top-K vous donne du crédit si la bonne étiquette aurait été l'une de vos premières suppositions
- Il est important de choisir K afin que de réelles améliorations puissent être reconnues
- Si $K =$ au nombre de classes, le taux de réussite est 100%
- Si $K = 1$ on a simplement l'accuracy , la précision et le rappel
- Normalement, on choisi K tel que la performance du classificateur est supérieure à celle d'un classificateur aléatoire
- Mais pas trop mieux, de façon qu'on ait encore de l'espace d'amélioration choisir K pas trop élevé et pas trop petit

Évaluation pour la régression

- Pour les valeurs numériques, l'erreur est une fonction de la différence entre la prévision $f(X)$ et l'observation y :
  - $|f(X) - y|$ absolute error
  - $|f(X) - y|/y$ (normalement meilleur) absolute percentage error
  - $(f(X) - y)^2$ (toujours non-négatif) square error
- Ceux-ci peuvent être agrégés sur de nombreux données, ex :
  - $MSE = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - y_i)^2$ (sensible aux *outliers*) Min square error
  - $RMSE = \sqrt{MSE}$ (plus facile à interpreter) Root min square error