Évaluation des performances des modèles

April 2023

Digital Bridge
R D C

Sujets à couvrir

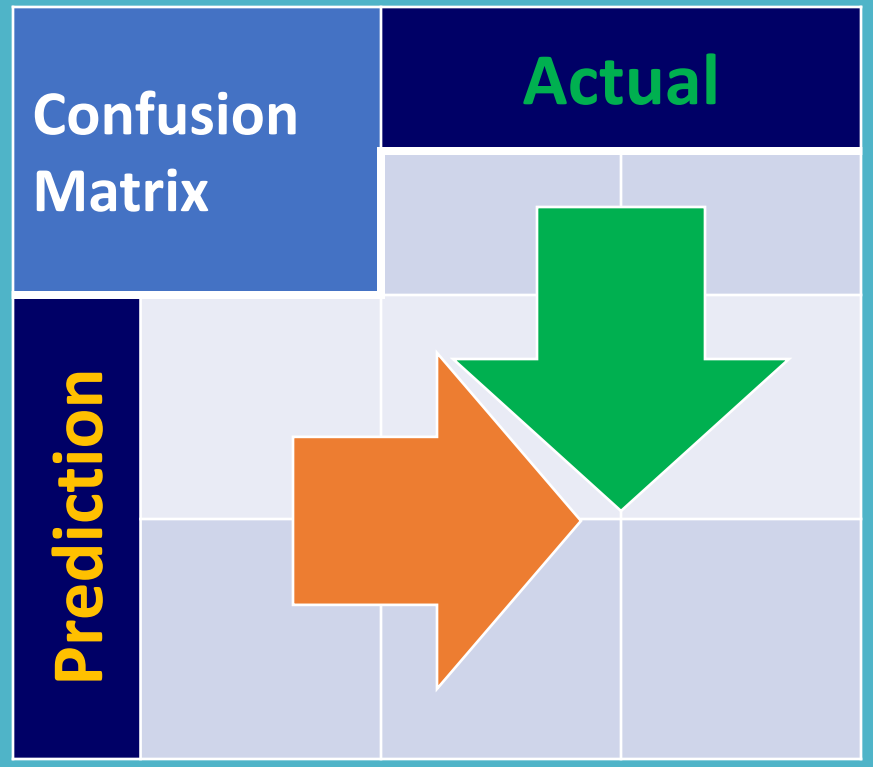
- Confusion Matrix
- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)
- Model Accuracy
- Error Rate
- Kappa Statistic
- Sensitivity and Specificity
- Precision and Recall
- F-measure
- Area Under the Curve (AUC)
- Receiver Operating Characteristic (ROC)
- Holdout Method
- Cross-validation (k-fold CV)
- Bootstrap sampling

Évaluation des performances du modèle

- Après avoir formé un modèle de machine learning, il est important d'évaluer les performances du modèle.
- Nous commencerons par discuter de la **matrice de confusion** qui est une méthode populaire et rapide pour évaluer la précision d'un modèle. Ensuite, étendez le concept à d'autres mesures telles que **la sensibilité, la spécificité, le rappel et la précision** qui tiennent compte de ce qui est important dans l'application d'apprentissage automatique que nous modélisons. Cela permettra de traiter les cas où **l'exactitude** n'est pas la seule mesure et/ou la meilleure pour évaluer le rendement du modèle. Une métrique mixte, appelée **F-score,** est introduite pour saisir l'avantage de l'optimisation pour la précision et le rappel.
- Nous plongeons enfin plus profondément dans la **statistique Kappa** qui est apparue à plusieurs reprises dans les sorties de la matrice de confusion.
- Nous couvrons les méthodes visuelles d'évaluation des performances des modèles en fonction de leurs taux d'erreur dans la caractéristique de **fonctionnement du récepteur (ROC)** et le concept associé d'aire sous la courbe (ASC).
- Nous couvrons également les méthodes d'amélioration et d'évaluation de la performance du modèle de classification pendant et après la formation avec les méthodes **d'échantillonnage holdout, cross-validation et bootstrap.**

Matrice de confusion

- La matrice de confusion est un tableau qui nous indique la performance d'un modèle de classification en comparant la prédiction à la classification réelle (correcte) d'un ensemble de données.
- Le nombre total dans la table est le nombre total d'éléments dans les données.
- La convention est de lire les prédictions de gauche à droite, et les réels de haut en bas.



Matrice de confusion: scénario binaire

- Dans un scénario de classification binaire, la matrice de confusion comporte 4 cellules dont les valeurs s'ajoutent au nombre total de jeux de données.
- Échantillons correctement classés:
- **True Positive (TP):** échantillons prédits True qui sont réellement True.
- **True Negative (TN):** échantillons prédits False qui sont en fait False.
- Échantillons mal classés :
- **Faux positif (FP):** échantillons prédits True qui sont en fait Faux.
- **Faux négatif (FN)**: échantillons prédits Faux qui sont réellement Vrais.

Confusion Matrix		Actual	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Exactitude du modèle et taux d'erreur

- La précision du modèle calcule le nombre d'échantillons de jeux de données qui sont correctement classés par le modèle
- Cas de classification binaire :
- Exactitude = (TP+TN)/(TP+FN+TN+FP)
- Généralisation pour la classification multi classe (N ≥ 2)
- Exactitude = (# échantillons correctement classés)/N
- L'exactitude est généralement exprimée en pourcentage
- Important : Bien que l'exactitude soit une bonne mesure pour évaluer les performances du modèle, elle peut être très trompeuse dans les cas d'ensemble de données déséquilibrées.
- Une grande exactitude peut être obtenue en classant simplement correctement la classe avec la majorité des échantillons si les autres classes sont de taille relativement petite
- Le taux d'erreur est à l'opposé de l'exactitude
- Taux d'erreur = 1 – Précision ou 100 % - Taux d'exactitude

Matrice de confusion : exemple binaire

Confusion Matrix		Actual	
		Positive	Negative
Prediction	Positive	99	5
	Negative	10	86

La matrice de confusion dans cet exemple nous indique que :

- 99 + 86 = 185 échantillons sont correctement classés
- 10 + 5 = 15 échantillons sont mal classés
- Exactitude : 185/(185+15) = 185/200 = 92,5% des échantillons sont correctement classés.
- Taux d'erreur: 1 - 92,5% = 7,5% des échantillons sont mal classés.

Matrice de confusion : exemple multi-classes

Confusion Matrix		Actual			
		Class 1	Class 2	Class 3	Class 4
Prediction	Class 1	102	11	24	5
	Class 2	10	86	14	26
	Class 3	34	13	95	12
	Class 4	12	23	10	89

- Lorsque le modèle de classification comporte plus de deux classes, les calculs des mesures d'erreur sont essentiellement les mêmes. Ils ont juste besoin d'être spécifiques à la classe.

Statistique Kappa

- La statistique kappa de Cohen est une mesure de la performance du modèle qui contrecarre la haute précision obtenue en classant avec précision la classe la plus fréquente.
- $\kappa = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$
  - où Pr(a) est la proportion de concordance réelle entre la concordance prévue et réelle, et
  - où Pr(e) est la proportion d'accord entre le prédit et le réel dû au hasard
- Ainsi, kappa tente d'éliminer la proportion due par hasard en soustrayant l'un de l'autre (Pr(a)–Pr(e)), puis normalise le résultat en divisant par 1–Pr(e) de sorte que kappa va toujours de 0 à 1.

Agreement	κ
Poor	$\kappa \leq 0.20$
Fair	$0.20 \leq \kappa \leq 0.40$
Moderate	$0.40 \leq \kappa \leq 0.60$
Good	$0.60 \leq \kappa \leq 0.80$
Very Good	$0.80 \leq \kappa \leq 1.00$

Statistique Kappa: exemple

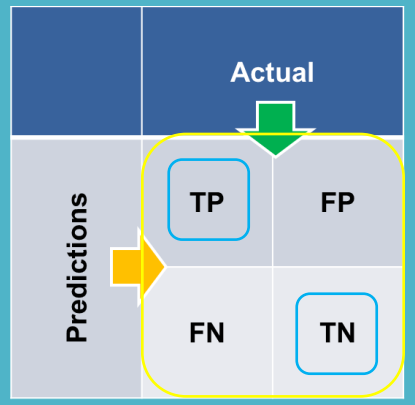
Confusion Matrix		Actual		
		Positive	Negative	
Prediction	Positive	99	5	104
	Negative	10	86	96
		109	91	200

- $Pr(a) = \frac{99+86}{\mathbf{200}} = \frac{185}{200} = 0.9250$
- $Pr(e) = \frac{109}{200} \times \frac{104}{200} + \frac{91}{200} \times \frac{96}{200} = 0.5018$
- $K = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$
- $K = \frac{0.925-0.5018}{1-0.5018} = 0.8495$

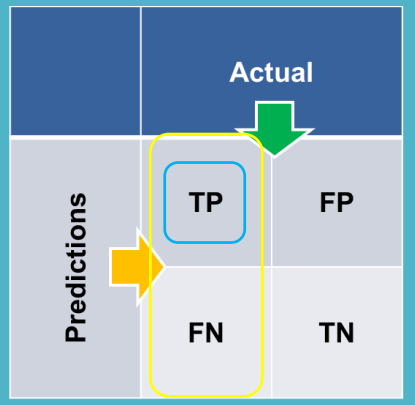
Sensibilité et spécificité

- Bien que le taux de exactitude d'un modèle puisse nous donner une bonne compréhension du rendement d'un modèle de classification, il peut ne pas toujours se concentrer sur la mesure de rendement pertinente au problème d'intérêt.
- Par exemple, dans un cas de diagnostic médical, un test peut être très exact en ce sens qu'il fournit un pourcentage élevé de vrais positifs et de vrais négatifs, mais que se passe-t-il si un diagnostic faussement négatif signifie qu'un patient très malade est renvoyé chez lui sans traitement? C'est un mauvais scénario.
- Heureusement, il existe d'autres façons d'examiner la matrice de confusion qui capturent d'autres scénarios au-delà de l'exactitude du modèle.
- La sensibilité (c'est-à-dire le taux de vrais positifs) est la proportion de positifs correctement classés parmi tous les échantillons positifs réels de l'ensemble de données.
- Sensibilité = TP/(TP+FN)
- La spécificité (c'est-à-dire le taux négatif vrai) la proportion de négatifs qui sont correctement classés parmi tous les échantillons négatifs réels de PN dans l'ensemble de données
- Spécificité: TN / (TN + FP)
- Ainsi, pour les applications où la capture correcte du vrai positif est très importante, la sensibilité est une meilleure mesure que l'exactitude globale, et inversement, lorsque la capture de faux positifs est cruciale, la spécificité est une meilleure mesure.

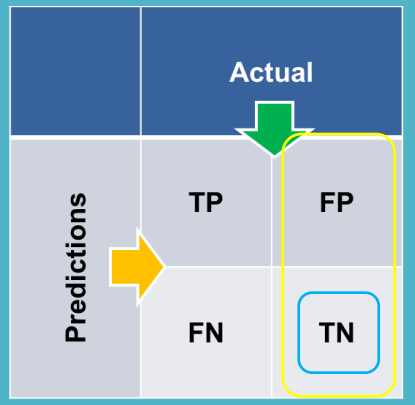
Sensibilité et spécificité : visualisée



$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$



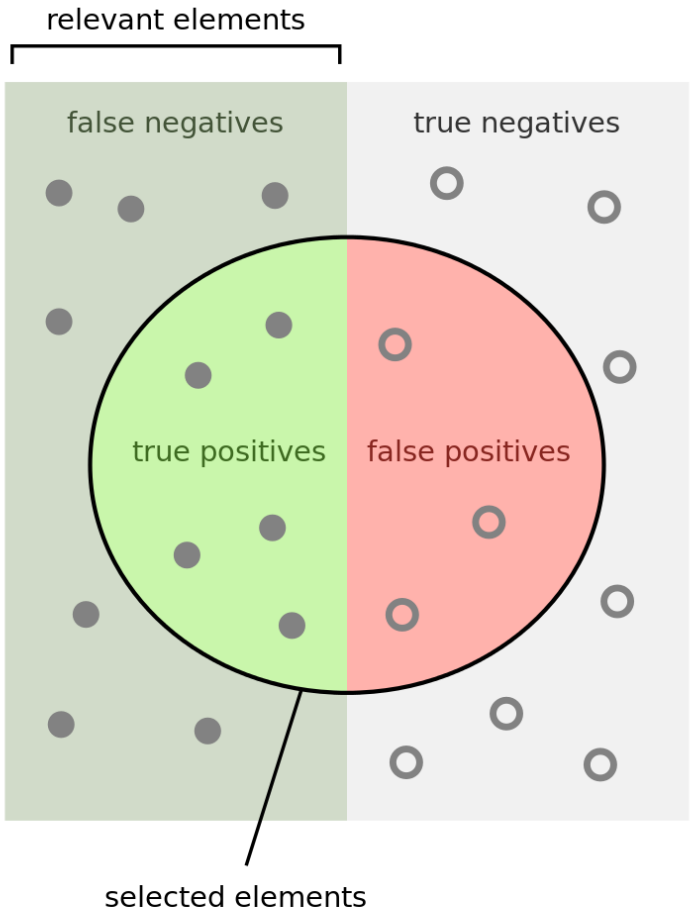
$$\text{Sensitivity} = \frac{TP}{TP+FN}$$



$$\text{Specificity} = \frac{TN}{FP+TN}$$

Precision and Recall

- **Precision** (aka positive predictive value) is a metric measures the percentage of truly positive samples over the actual number of positive samples in the dataset.
  - It answers the question: *of all the samples that were predicted positive, what proportion were truly positive?*
- **Recall** (aka sensitivity) is a measure of how complete the classification model is.
  - It answers the question: *of all the samples that are truly positive, what proportion were predicted positive?*



How many selected items are relevant?

Precision = $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

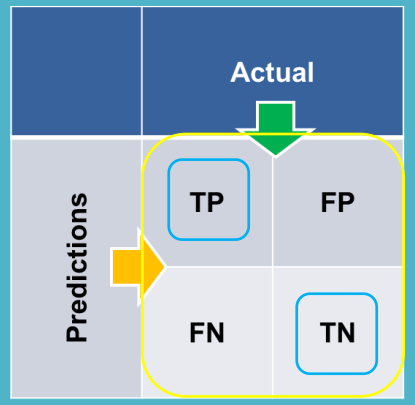
How many relevant items are selected?

Recall = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

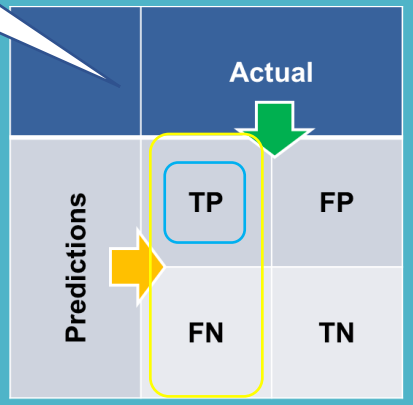
Source: https://en.wikipedia.org/wiki/Precision_and_recall

Precision and Recall: visualized

of all the samples that are truly positive, what proportion were predicted positive?

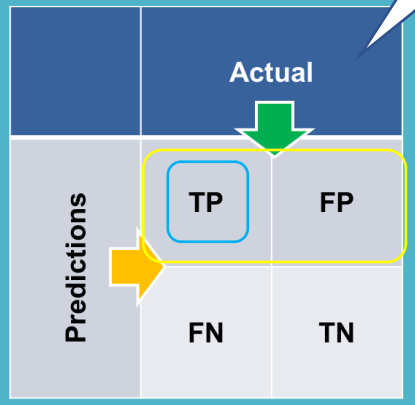


$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$



$$\text{Recall} = \frac{TP}{TP+FN}$$

of all the samples that were predicted positive, what proportion were truly positive?



$$\text{Precision} = \frac{TP}{TP+FP}$$

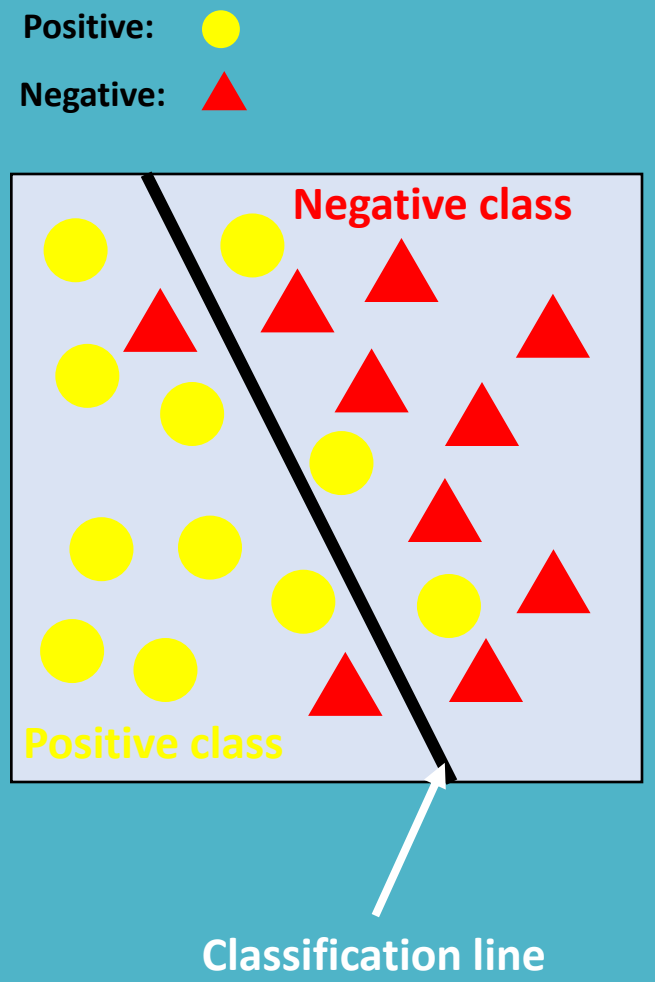
F-measure

- La mesure F (alias score F alias score F1) est la moyenne harmonique des métriques de précision et de rappel. Il est utilisé pour évaluer les performances d'un modèle en combinant les deux mesures en une seule.
-
  - **F- measure** = $\frac{2 \text{ x } Precision \text{ x } Recall}{Recall+Precision} = \frac{2 \text{ x } TP}{2\ x\ TP+FN+FP}$

- Notation alternative montrant que la mesure F est la moyenne harmonique de précision et de rappel:
  - $F_\beta = \frac{1+\beta^2}{\frac{\beta}{Precision} + \frac{\beta}{Recall}}$, with $F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$ when $\beta = 1$

- Le score F capture les forces et les faiblesses de la précision et du rappel du modèle dans une seule mesure.

Mesures d'erreur: scénario à deux classes

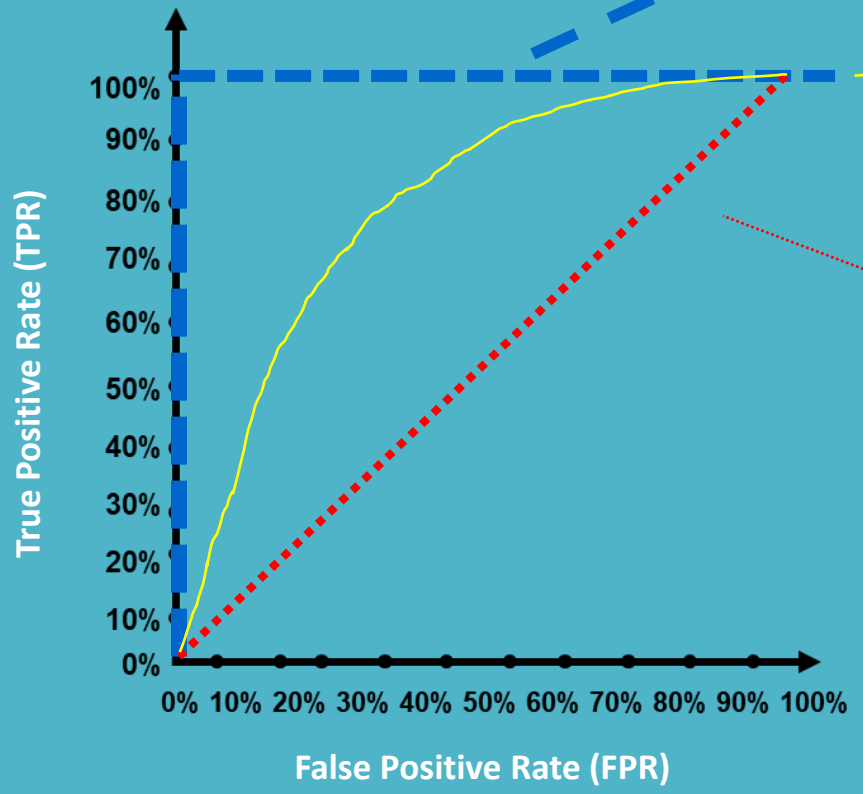


- **Accuracy**: $\frac{TP+TN}{TP+FN+TN+FP} = \frac{8+8}{8+3+8+2} = \frac{16}{21} = 76.2\%$
- **Precision**: $\frac{TP}{TP+FP} = \frac{8}{8+2} = \frac{8}{10} = 80.0\%$
- **Recall**: $\frac{TP}{TP+FN} = \frac{8}{8+3} = \frac{8}{11} = 72.7\%$
- **Sensitivity**: $\frac{TP}{TP+FN} = \frac{8}{8+3} = \frac{8}{11} = 72.7\%$
- **Specificity**: $\frac{TN}{TN+FP} = \frac{8}{8+2} = \frac{8}{10} = 80.0\%$
- **F- measure**: $\frac{2 \text{ x } Precision \text{ x } Recall}{Recall+Precision} = \frac{2 \text{ x } TP}{2\,x\,TP+FN+FP} = \frac{2 \text{ x } 8}{2 \text{ x } 8+3+2} = \frac{16}{21} = 76.2\%$

Caractéristique de fonctionnement du récepteur (ROC)

- La caractéristique de fonctionnement du récepteur (ROC) est une courbe qui trace le taux de vrais positifs (TPR) par rapport au taux de faux positifs (FPR) calculé à partir de diverses matrices de confusion générées lorsqu'un seuil de classification est modifié.
- L'idée est que nous voulons :
- Voir comment le TPR change par rapport au FPR lorsque le seuil du modèle est modifié
  - Comparez deux modèles ou plus les uns par rapport aux autres pour sélectionner celui qui a la meilleure courbe TPR, ce qui signifie les meilleures performances.

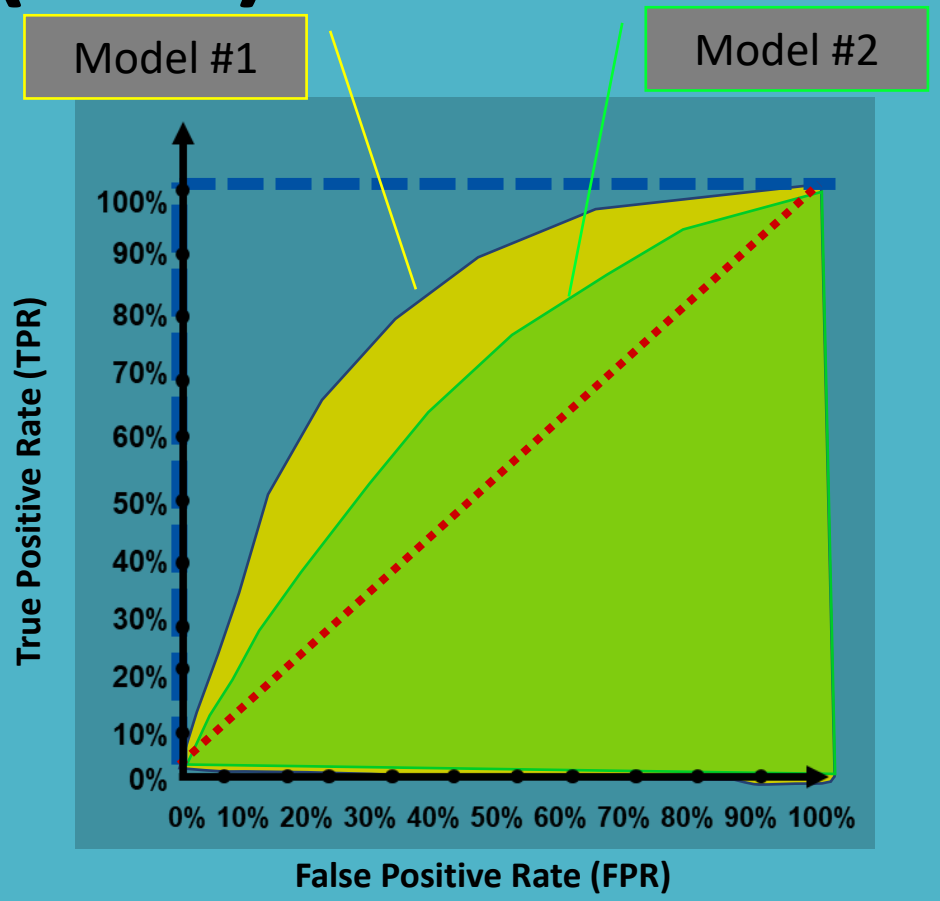
ROC: Visualization

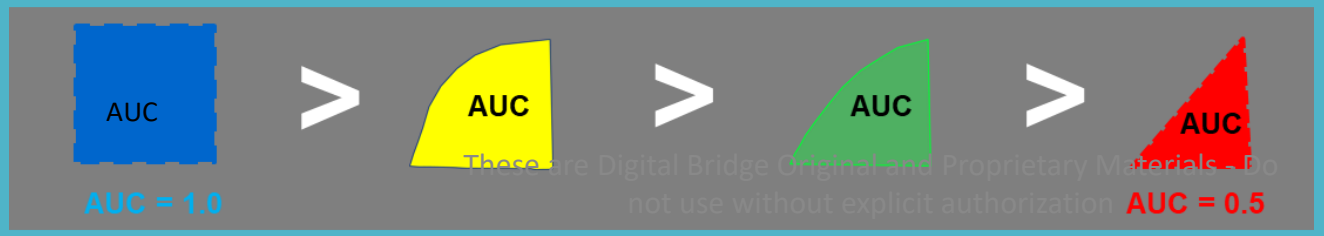


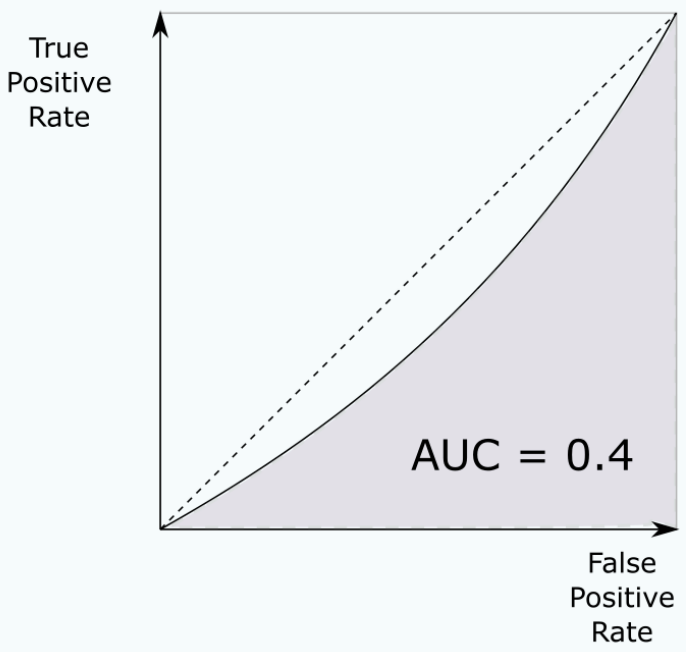
- Ideal ROC curve with 100% TPR
- ROC curve for typical model
- TRP = FPR line
- Performance no better than a random predictor

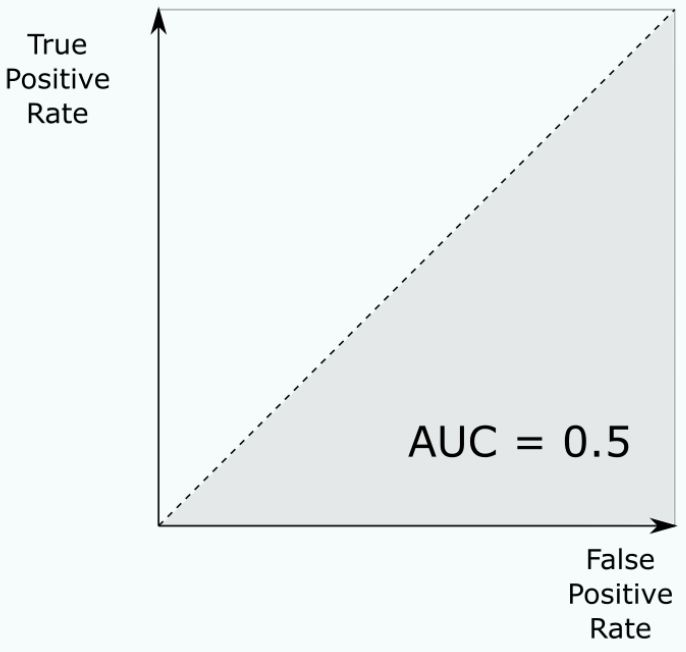
Area Under the Curve (AUC)

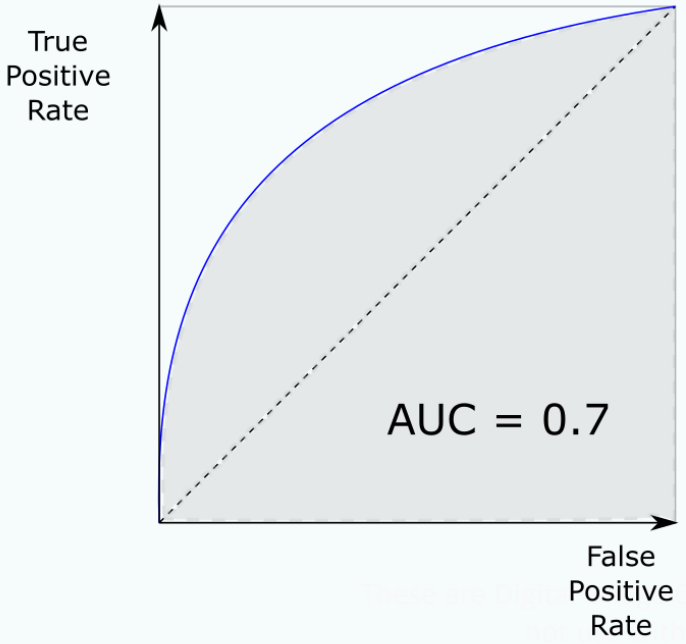
- L'aire sous la courbe peut être utilisée pour comparer les performances de plusieurs modèles.
- Hypothèses:
- Un bon modèle a une courbe ROC au-dessus de la ligne de prédiction aléatoire (ligne rouge)
- Un modèle avec un TPR plus élevé qu'un autre modèle pour un FPR donné est un meilleur modèle
- Le modèle idéal a une ASC de 1,0
- Le modèle de prédicteur aléatoire a une ASC de 0,5
- Nous sélectionnons le modèle avec la plus grande surface sous la courbe
- AUC jaune > AUC verte

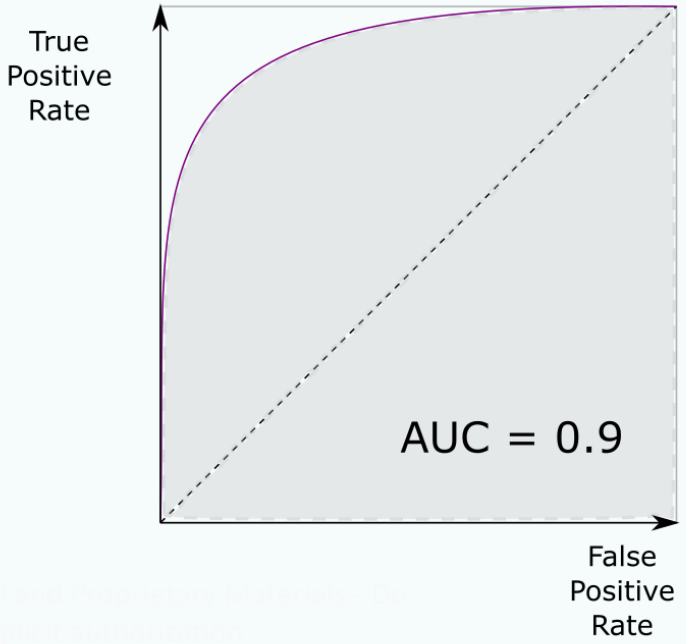





True
Positive
Rate
AUC = 0.4
False
Positive
Rate


True
Positive
Rate
AUC = 0.5
False
Positive
Rate


True
Positive
Rate
AUC = 0.7
False
Positive
Rate


True
Positive
Rate
AUC = 0.9
False
Positive
Rate

Méthode Holdout

- La méthode d'exclusion consiste à diviser le jeu de données d'origine en un jeu de données d'apprentissage et un jeu de données de test.
- En règle générale, des ratios de répartition formation/test de 80/20, 70/30 ou 67/33 % sont utilisés.
- Important : les données de test ne doivent pas influencer la formation du modèle. Cela peut se produire si le modèle s'entraîne à plusieurs reprises et que le modèle final est choisi en fonction des résultats des tests.
- Jeu de données de validation : un exemple de jeu de données utilisé pour affiner/ajuster le modèle pendant la phase d'apprentissage. En revanche, l'ensemble de données de test est utilisé une fois, une fois la formation terminée. Répartition typique : 50 % de formation, 25 % de validation et 25 % de données de test
- Résistance répétée : c'est-à-dire technique de sous-échantillonnage aléatoire utilisée pour atténuer le problème de l'ensemble de données d'entraînement composé de manière aléatoire. La méthode d'exclusion est exécutée plusieurs fois et l'évaluation est basée sur la moyenne des estimations de performance des exécutions multiples.

Validation croisée (Cross-validation)(CV)

- La validation croisée est une méthode permettant d'évaluer les performances d'un modèle d'apprentissage automatique en divisant l'ensemble de données en k partitions distinctes appelées plis.
- Le modèle est ensuite entraîné et testé sur chacun des plis et l'estimation finale des performances est une moyenne des estimations de performance k-fold. Estimation finale de l'erreur = moyenne(erreurs)
- Un CV à 10 volets est couramment utilisé lorsque le modèle utilise 90% de l'ensemble de données original pour la formation et les 10% restants pour l'évaluation du modèle
- La méthode Leave-one-out est une variante du CV k-fold où le nombre de plis est égal au nombre d'instances dans l'ensemble de données. L'algorithme d'apprentissage est appliqué une fois pour chaque instance, en utilisant toutes les autres instances comme ensemble d'apprentissage et en utilisant l'instance sélectionnée comme jeu de test à élément unique.
- Le CV k-fold répété est une variante du CV k-fold où k-fold est répété plusieurs fois. Par exemple : un CV 10 fois appliqué 10 fois.
-

Échantillonnage bootstrap

- L'amorçage dans l'apprentissage automatique est utilisé pour créer des ensembles d'apprentissage et de test sélectionnés au hasard à partir du jeu de données d'origine
- L'amorçage utilise un échantillonnage aléatoire avec remplacement à partir du jeu de données d'origine.
- L'amorçage produit des jeux de données contenant des ensembles de données d'apprentissage contenant 63,2 % des données d'origine et des ensembles de données d'apprentissage contenant 36,8 % des données d'origine.
- Avantage : l'amorçage fonctionne mieux que la validation croisée sur de petits jeux de données
- Inconvénient : les estimations de performances d'amorçage sont généralement inférieures à celles du CV pour un jeu de données volumineux
- 0,632 amorçage ajuster le calcul de l'erreur en prenant 63,2 % de l'erreur de test calculée et en ajoutant 36,8 % de l'erreur d'entraînement mesurée
- erreur = 0,632 x test d'erreur + 0,368 x formation aux erreurs

Resources (1)

- Machine Learning: Testing and Error Metrics
  - https://www.youtube.com/watch?v=aDW44NPhNw0
- Machine Learning Fundamentals: The Confusion Matrix
  - https://www.youtube.com/watch?v=Kdsp6soqA7o
- Kappa Coefficient
  - https://www.youtube.com/watch?v=fOR_8gkU3UE
- Kappa Value Calculation | Reliability
  - https://www.youtube.com/watch?v=DfNo32nL_fo
- Precision, Recall & F-Measure
  - https://www.youtube.com/watch?v=j-EB6RqqjGI
- Machine Learning Fundamentals: Sensitivity and Specificity
  - https://www.youtube.com/watch?v=vP06aMoz4v8
- ROC (Receiver Operating Characteristic) Curve in 10 minutes!
  - https://www.youtube.com/watch?v=z5qA9qZMyw0
- ROC Curves and Area Under the Curve (AUC) Explained
  - https://www.youtube.com/watch?v=OAl6eAyP-yo
- ROC and AUC in R
  - https://www.youtube.com/watch?v=qcvAqAH60Yw
- ROC Curve & Area Under Curve (AUC) with R - Application Example
  - https://www.youtube.com/watch?v=ypO1DPEKYFo

Resources (2)

- Machine Learning | Random Subsampling Classifier Evaluation
  - https://www.youtube.com/watch?v=zWvGgThvZ7g
- Four Types Of Cross Validation| K-Fold | Leave One Out |Bootstrap | Hold Out
  - https://www.youtube.com/watch?v=e0JcXMzhtdY
- Cross Validation : Data Science Concepts
  - https://www.youtube.com/watch?v=wjILv3-UGM8
- Machine Learning Fundamentals: Cross Validation
  - https://www.youtube.com/watch?v=fSytzGwwBVw
- Partitioning data into training and validation datasets using R
  - https://www.youtube.com/watch?v=aS1O8EiGLdg
- Bootstrapping, Main Ideas
  - https://www.youtube.com/watch?v=isEcgoCmlO0