TD 1 : *Data Science & Machine Learning*

Exercice 1 : Séparation de données et validation croisée

Soit un ensemble de 20 données, numérotées, utilisées pour réaliser un apprentissage automatique.

(a) On considère les données dans l'ordre croissant de leur numérotation et on découpe l'ensemble de données en 3 sous-ensembles d'apprentissage, selon la répartition suivante des données : 60% pour l'ensemble d'apprentissage, 20% pour l'ensemble de validation et 20% pour l'ensemble de test.
Listez les données composant chacun des sous-ensembles

(b) Nous souhaitons maintenant réaliser une validation croisée avec 4 plis (ou *folds*), après avoir séparé les données en 80% pour l'apprentissage et 20% pour le test.
Dessinez les différentes étapes de l'apprentissage, en indiquant les données de chaque sous-ensemble et chaque pli.

Exercice 2 : Normalisation de données

Vous disposez d'un ensemble de données fictives représentant les notes des étudiants dans deux matières : Mathématiques et Informatique. Ces notes sont mesurées sur une échelle de 0 à 100. Voici les données initiales non normalisées pour cinq étudiants :

Étudiant	Mathématiques	Informatique	Mathématiques normalisé	Informatique normalisé
1	85	78		
2	92	89		
3	78	60		
4	88	75		
5	95	92		

(a) Calculez la moyenne et l'écart-type des notes en Mathématiques et en Informatique pour ces cinq étudiants.

(b) Appliquez la normalisation Min-Max aux données, pour les deux matières, en utilisant la formule suivante :

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

où X_{norm} est la valeur normalisée, X est la valeur d'origine, $\min(X)$ est la valeur minimale dans la colonne, et $\max(X)$ est la valeur maximale dans la colonne. Vous pouvez compléter les 2 colonnes correspondantes, dans le tableau ci-dessus.

(c) Calculez la moyenne et l'écart-type des données normalisées pour chaque matière (Mathématiques et Informatique) et expliquez comment la normalisation a affecté la distribution des données.

(d) Pourquoi est-il important de normaliser les données avant de les utiliser dans des algorithmes de Machine Learning ? Quels avantages cela apporte-t-il ?

Solution:

(a) Moyenne et écart-type des variables avant normalisation.
- Mathématiques :
  - Moyenne : $(85 + 92 + 78 + 88 + 95)/5 = 87.6$
  - Écart-type : $\sqrt{\frac{(85-87.6)^2+(92-87.6)^2+(78-87.6)^2+(88-87.6)^2+(95-87.6)^2}{5}} \approx 6.46$
- Informatique :
  - Moyenne : $(78 + 89 + 60 + 75 + 92)/5 = 78.8$
  - Écart-type : $\sqrt{\frac{(78-78.8)^2+(89-78.8)^2+(60-78.8)^2+(75-78.8)^2+(92-78.8)^2}{5}} \approx 10.54$

(b) Table avec normalisation des variables :

Étudiant	Mathématiques	Informatique	Mathématiques normalisé	Informatique normalisé
1	85	78	$X_{\text{norm}} = \frac{85-78}{95-78} \approx 0.368$	$X_{\text{norm}} = \frac{78-60}{92-60} \approx 0.4$
2	92	89	$X_{\text{norm}} = \frac{92-78}{95-78} \approx 0.736$	$X_{\text{norm}} = \frac{89-60}{92-60} \approx 0.85$
3	78	60	$X_{\text{norm}} = \frac{78-78}{95-78} = 0$	$X_{\text{norm}} = \frac{60-60}{92-60} = 0$
4	88	75	$X_{\text{norm}} = \frac{88-78}{95-78} \approx 0.526$	$X_{\text{norm}} = \frac{75-60}{92-60} \approx 0.3$
5	95	92	$X_{\text{norm}} = \frac{95-78}{95-78} = 1$	$X_{\text{norm}} = \frac{92-60}{92-60} = 1$

(c) Moyenne et écart-type des variables après normalisation.
- Mathématiques :
  - Moyenne : $(0.368 + 0.736 + 0 + 0.526 + 1)/5 \approx 0.526$
  - Écart-type : $\sqrt{\frac{(0.368-0.526)^2+(0.736-0.526)^2+(0-0.526)^2+(0.526-0.526)^2+(1-0.526)^2}{5}} \approx 0.351$
- Informatique :
  - Moyenne : $(0.4 + 0.85 + 0 + 0.3 + 1)/5 \approx 0.51$
  - Écart-type : $\sqrt{\frac{(0.4-0.51)^2+(0.85-0.51)^2+(0-0.51)^2+(0.3-0.51)^2+(1-0.51)^2}{5}} \approx 0.308$

La normalisation a ramené les moyennes à environ 0 et les écart-types à environ 1 pour les deux matières.

(d) L'importance de la normalisation des données réside dans le fait qu'elle met les variables à la même échelle, ce qui permet aux algorithmes de Machine Learning de fonctionner de manière plus efficace et de donner de meilleurs résultats. Sans normalisation, des variables avec des plages de valeurs très différentes peuvent biaiser les modèles vers les variables avec des valeurs plus élevées.

Exercice 3 : Matrice de confusion et résultats de classification binaire

Deux modèles ont été appris sur un même ensemble de données. Nous souhaitons évaluer ces 2 modèles sur des données de test : les résultats obtenus sont présentés dans la Table 1. Chaque ligne correspond à une donnée et contient la classe réelle de cette donnée (appelée *Référence*) ainsi que la classe choisie par chaque modèle.

(a) Donnez la matrice de confusion correspondant à chaque modèle.

(b) Pour chaque modèle, calculez la précision, le rappel et la F-mesure.

(c) Déduisez-en le meilleur modèle, en termes de F-mesure.

(d) Calculez l'exactitude et déduisez-en le meilleur modèle, en termes d'exactitude.

Donnée	Référence	Modèle 1	Modèle 2
1	+	+	+
2	+	+	+
3	+	+	-
4	-	+	-
5	+	+	+
6	+	+	+
7	+	+	-
8	-	+	-
9	-	-	-
10	-	-	-

TABLE 1 – Résultats de classification binaire

Exercice 4 : Courbe ROC et courbe précision-rappel

La Table 2 présente les résultats d'un classificateur binaire, pour la classe +, sur un ensemble de test de 20 exemples. L'association d'un exemple à la classe + se fait par rapport à un seuil σ, selon la règle suivante : *si le score de l'exemple est supérieur au seuil σ alors l'exemple est classé dans la classe +.*

Id exemple	Classe de référence	Score (classe +)
18	-	0,25
6	-	0,28
14	+	0,28
5	-	0,33
13	-	0,37
17	-	0,37
16	-	0,47
2	+	0,48
3	-	0,52
12	+	0,54
15	-	0,57
4	+	0,64
8	-	0,64
7	-	0,74
19	+	0,79
1	+	0,85
20	-	0,87
9	-	0,89
11	+	0,95
10	+	0,98

TABLE 2 – Résultats de classification binaire, selon un score

(a) Pour chaque seuil σ considéré, remplissez la Table 3.
Vous indiquerez le taux de faux positifs (colonne TFP), le taux de vrais positifs (colonne TTP), la précision (colonne P) et le rappel (colonne R), pour chacun des seuils.

(b) Construisez la courbe ROC, à partir de ces valeurs.

(c) Déduisez-en le meilleur seuil σ_1 à utiliser.

(d) Construisez la courbe précision-rappel, à partir de ces valeurs. Déduisez-en le meilleur seuil σ_2 à utiliser.

Seuil	TP	FP	TN	FN	$TFP = \frac{FP}{FP+TN}$ = 1 - Spécificité	$TTP = \frac{TP}{TP+FN}$ = Sensibilité	P	R
0,2								
0,3								
0,4								
0,5								
0,6								
0,7								
0,8								
0,9								

TABLE 3 – Tableau pour la construction de la courbe ROC et de la courbe précision-rappel

Exercice 5 : Classification multi-classes

La Table 4 présente les résultats d'une classification automatique de 12 exemples (représentés par leur identifiant), en 4 classes, ainsi que la classification référence de ces exemples.

	Classe 1	Classe 2	Classe 3	Classe 4
Classification référence	1, 4, 12	5, 6, 7	2, 8, 9, 10	3, 11
Classification automatique	1, 2, 3, 4	5, 6	7, 8, 9	10, 11, 12

TABLE 4 – Résultats de classification multi-classes

(a) Calculez la micro-précision, pour cette classification automatique.

(b) Calculez la macro-précision, pour cette classification automatique.

(c) Que pouvez-vous en déduire ?

(d) Faites les mêmes calcul mais en prenant cette fois le micro-rappel et le macro-rappel.