

Exercice 1 (10p)

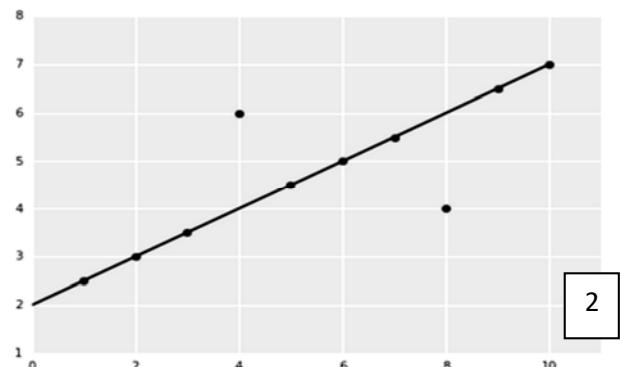
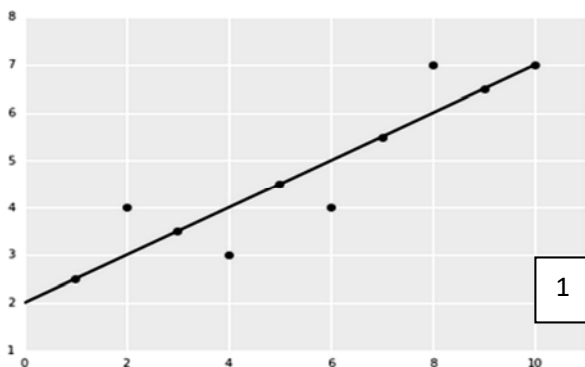
Un modèle d'IA a été entraîné pour classifier des courriels en deux catégories : **spam** et **non spam**. Le modèle a été évalué sur un ensemble de test contenant 100 mails, dont 50 spams et 50 non spams. La matrice de confusion suivante a été obtenue :

Classe réelle	Spam prédit	Non spam prédit
Spam	40	10
Non spam	5	45

- 1) Quelles sont les deux tâches supervisées les plus courantes ?
- 2) Quel type d'algorithme d'apprentissage utilisez-vous dans ce cas ?
- 3) Qu'est-ce qu'une matrice de confusion en machine Learning ?
- 4) Calculez les métriques de performance suivantes à partir de la matrice de confusion : Accuracy Précision ; Rappel ; F1-score
- 5) Interpréter en 2-3 phrases ce que signifient ces résultats.
- 6) Que pouvez-vous conclure sur les performances de ce modèle?

Hint : Précision = $TP / (TP + FP)$; Rappel = $TP / (TP + FN)$;
 F1-score = $2 * (Précision * Rappel) / (Précision + Rappel)$

Exercice 2(3p) Examiner les deux graphes suivants :

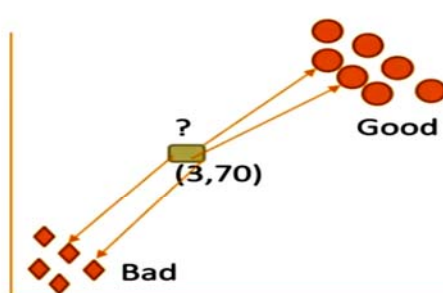


Lequel des deux ensembles de données présentés dans les graphes précédents a l'erreur quadratique moyenne la plus élevée ?

$$MSE = \frac{1}{N} \sum_i (s_i - d_i)^2$$

Exercice 3(7p)

Utiliser l' algorithme de KNN pour trouver le cas (BAD or GOOD) du name E comme sur la figure suivante



K=4

Name	Cigarettes	Weight	Heart Attack
A	7	70	Bad
B	7	40	Bad
C	3	40	Good
D	1	40	Good
E	3	70	

Exercice 1 (10p)

1) Les deux tâches supervisées les plus courantes : Régression et Classification

2) Quel type d'algorithme d'apprentissage utilisez-vous pour segmenter vos clients en plusieurs groupes : Classification.

3) Réponse : Une matrice de confusion est une table de contingence utilisée pour évaluer la performance d'un modèle de machine learning en comparant les prédictions du modèle aux valeurs réelles.

4) Accuracy = $(40+45)/100 = 0,85$ Accuracy donc 85%

Précision: $\text{Précision} = TP / (TP + FP) = 40 / (40 + 5) = 88,9\%$

Rappel: $\text{Rappel} = TP / (TP + FN) = 40 / (40 + 10) = 80\%$

F1-score: $\text{F1-score} = 2 * (\text{Précision} * \text{Rappel}) / (\text{Précision} + \text{Rappel}) = 2 * (0,8889 \times 0,8) / (0,8889 + 0,82) = 0,842$, soit $\approx 84,2\%$

Interprétation des résultats:

Le modèle présente des performances globalement bonnes, mais il reste des erreurs qui ne sont pas négligeables pour un filtre de mails.

Qualité globale

- Une accuracy de 85% signifie que 85 mails sur 100 sont correctement classés, ce qui montre un modèle déjà utile mais pas parfait.
- Les précisions ($\approx 88,9\%$ pour spam, $\approx 81,8\%$ pour non spam) indiquent que, quand le modèle prédit une classe, il a assez souvent raison, mais des erreurs persistent.

Gestion du spam

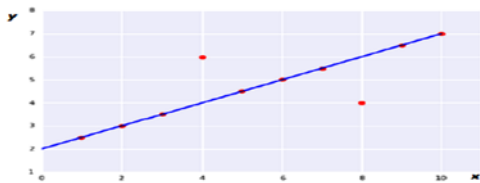
- Le rappel spam de 80% montre que 20% des spams ne sont pas détectés et arrivent dans la boîte de réception, ce qui peut exposer l'utilisateur à du contenu indésirable ou dangereux.
- Le rappel non spam de 90% signifie que 10% des mails légitimes sont classés à tort en spam, ce qui peut faire manquer des messages importants, même si ce taux reste modéré.

Conclusion pratique

- Le modèle représente un compromis acceptable : il filtre déjà bien le spam tout en limitant les faux positifs, mais il faudrait l'améliorer si l'application exige une très grande fiabilité (par exemple pour des mails professionnels sensibles). Le F1-score confirme cette conclusion.

Exercice 2(3p)

L'ensemble de données situé à droite.



Les huit exemples sur la droite subissent une perte de 0. Cependant, bien que seulement deux points se trouvent hors de la droite, ces deux points sont *deux fois* plus éloignés de la droite que les points des anomalies sur la figure de gauche. La perte de la mise au carré amplifie ces différences, c'est pourquoi un décalage de deux subit une perte quatre fois plus grande qu'un décalage de 1.

$$MSE = \frac{0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2}{10} = 0.8$$

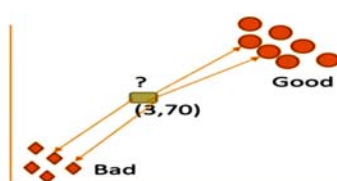
L'ensemble de données situé à gauche.



Les six exemples sur la droite subissent une perte de 0. Les quatre exemples qui ne sont pas situés sur la droite n'en sont pas très éloignés, donc même la mise au carré de leur décalage génère une valeur faible :

$$MSE = \frac{0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 0^2}{10} = 0.4$$

Exercice 3(7p)



K=4

Name	Cigarettes	Weight	Heart Attack	Distance	
A	7	70	Bad	$\sqrt{(3-7)^2 + (70-70)^2}$ $= 4$	①
B	7	40	Bad	$\sqrt{(3-7)^2 + (70-40)^2}$ $= 30,27$	④
C	3	40	Good	$\sqrt{(3-3)^2 + (70-40)^2}$ $= 30,00$	②
D	1	40	Good	$\sqrt{(3-1)^2 + (70-40)^2}$ $= 30,07$	③
E	3	70	Bad		