Classification linéaire

Sigmoïde, entropie croisée et descente de gradient : trier deux catégories

À savoir avant de commencer

Ce document réutilise directement la descente de gradient (document 0). Il introduit une nouvelle fonction (la sigmoïde) et une nouvelle fonction de coût (l'entropie croisée), puis en calcule les dérivées en détail, pas à pas.

Niveau : Progressif : Seconde/Première pour la partie A, Terminale pour la suite

Introduction : trier deux catégories de fruits

Une IA (un « classificateur ») doit apprendre à distinguer deux catégories de fruits — des groseilles et des myrtilles — à partir de leurs caractéristiques mesurées (diamètre, masse…). C'est un exemple simple de classification binaire, une tâche extrêmement fréquente en intelligence artificielle : trier un e-mail (spam ou non), une tumeur (bénigne ou maligne), une image (chat ou chien)…

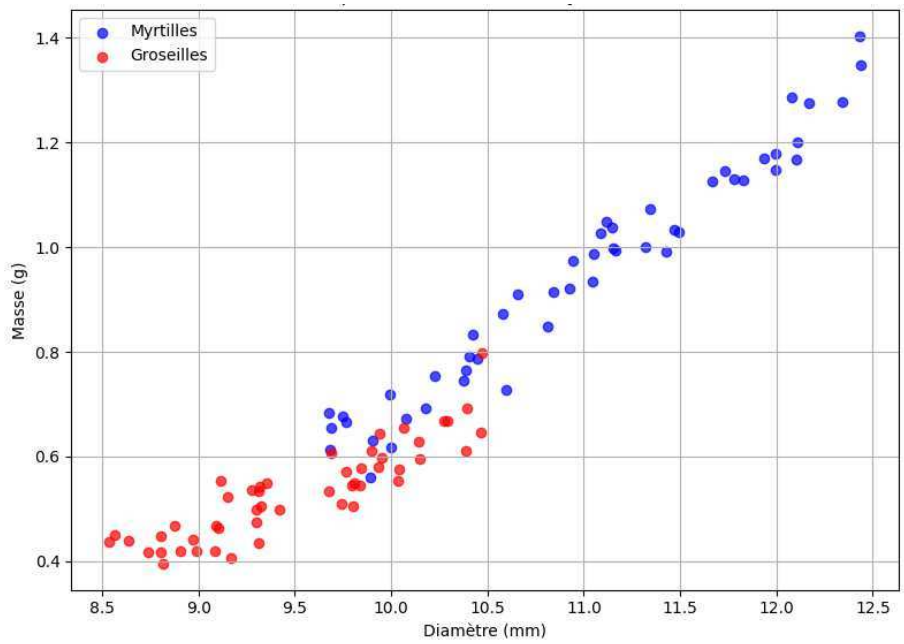
L'objectif : trouver une règle mathématique qui, à partir des mesures d'un fruit, prédit sa catégorie le plus souvent possible.

Partie A — Séparer deux nuages de points par une droite

Niveau : Seconde

Une intelligence artificielle apprend à classer des fruits en fonction de deux caractéristiques : leur diamètre x (en mm) et leur masse y (en g). L'objectif est de construire un modèle mathématique pour que l'IA puisse prédire si un fruit est une groseille ou une myrtille.

Voici, dans un graphique, les caractéristiques de 100 fruits (50 myrtilles et 50 groseilles d'une certaine variété, calibrés), choisis aléatoirement sur un marché, mesurés et pesés :



1. Que représentent les abscisses et les ordonnées dans ce graphique ?

Une IA cherche une droite de la forme $y = ax + b$ qui permet de séparer les deux groupes, afin de pouvoir reconnaître les groseilles et les myrtilles.

2. Propose, « à l'œil », en la traçant sur le graphique, une droite qui semble séparer « au mieux » les deux groupes.
3. Écris l'équation réduite de cette droite, en calculant grâce au graphique des valeurs approchées de ses coefficients a et b.
4. On choisit une groseille ou une myrtille au hasard parmi les 100 fruits, et on donne ses caractéristiques (diamètre, masse) = $(x_0 ; y_0)$. Explique pourquoi les deux inégalités $y_0 < ax_0 + b$ et $y_0 > ax_0 + b$ permettraient à l'IA de « reconnaître » une groseille d'une myrtille. Est-ce infaillible ?

Une machine a mesuré et pesé deux fruits, et on trouve : Fruit 1 : (diamètre, masse) = (9,8 ; 0,65) et Fruit 2 : (10,5 ; 0,7).

5. D'après ta droite, propose la nature (groseille ou myrtille) de chacun de ces deux fruits.

Pour aller plus loin et pouvoir dériver facilement, on va maintenant simplifier la situation : on ne garde qu'une seule caractéristique, le diamètre x, et on cherche à prédire une probabilité (plutôt qu'une simple catégorie 0 ou 1). La partie G reviendra ensuite sur une version à deux caractéristiques, plus proche de ce graphique.

Partie B — Peut-on ajuster a et b automatiquement ?

Niveau : Terminale

Plutôt que de choisir a et b « à l'œil » comme en partie A, on veut que l'IA les ajuste elle-même. On mesure l'erreur de la droite sur l'ensemble des n fruits par la fonction de coût (une erreur quadratique moyenne) :

Fonction de coût

$$J(a, b) = (1/n) \times \Sigma_i (y_i - (a\, x_i + b))^2$$

(Cette fonction est parfois appelée « entropie croisée » dans certains manuels par abus de langage ; il s'agit ici bien d'une erreur quadratique moyenne, comme au document 4 — l'entropie croisée, la vraie, est introduite en partie D.)

6. En dérivant J par rapport à a (b étant fixé), montre que $\partial J/\partial a = -(2/n) \times \Sigma_i\, x_i(y_i - a\, x_i - b)$.
7. Par un calcul analogue, donne l'expression de $\partial J/\partial b$.
8. Sachant que la descente de gradient met ensuite à jour les paramètres par $a \leftarrow a - \eta \times \partial J/\partial a$ et $b \leftarrow b - \eta \times \partial J/\partial b$ (document 0), explique pourquoi répéter cette opération plusieurs fois de suite doit progressivement rapprocher la droite de celle que tu as tracée « à l'œil » en partie A.

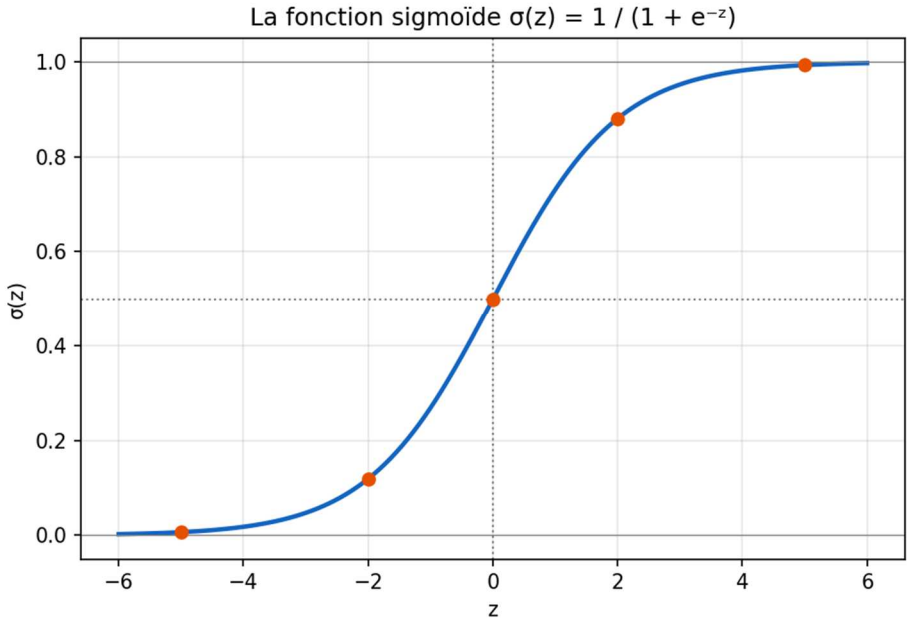
Partie C — La fonction sigmoïde : transformer un score en probabilité

Niveau : Terminale

On construit d'abord un score z = ax + b à partir du diamètre x (exactement comme une équation de droite). Mais ce score peut prendre n'importe quelle valeur réelle, alors qu'on veut une probabilité, donc un nombre entre 0 et 1. On utilise pour cela la fonction sigmoïde :

Définition

$$\sigma(z) = 1 / (1 + e^{(-z)})$$



9. Calcule σ(−5), σ(−2), σ(0), σ(2) et σ(5) (arrondis à 10^{-3} près), et vérifie tes résultats sur le graphique.
10. Montre que pour tout réel z, on a σ(−z) = 1 − σ(z).
11. Quelle est la limite de σ(z) quand z tend vers +∞ ? Quand z tend vers −∞ ? Pourquoi cela en fait-il une fonction idéale pour représenter une probabilité ?

12. On admet que σ est dérivable sur ℝ. Montre, en écrivant σ(z) = (1+e^(-z))^(-1) et en utilisant la dérivée d'une fonction composée, que : σ′(z) = σ(z) × (1 − σ(z)).

On retiendra cette formule de la dérivée de la sigmoïde : elle sera indispensable dans la partie E.

Partie D — L'entropie croisée : mesurer l'erreur d'une probabilité

Niveau : Terminale

Pour une prédiction $\hat{y} = \sigma(z)$ (nombre entre 0 et 1) et une vraie classe y (égale à 0 ou 1), l'erreur commise sur cet exemple est mesurée par l'entropie croisée :

Définition

$$\mathbf{L(y, \hat{y}) = -\,[\, y \times \ln(\hat{y}) + (1 - y) \times \ln(1 - \hat{y})\,]}$$

13. Si y = 1, montre que la formule se réduit à $L = -\ln(\hat{y})$. Si y = 0, montre qu'elle se réduit à $L = -\ln(1 - \hat{y})$.
14. Complète le tableau suivant (arrondis à 10^{-3} près) :

y	$\hat{y}$	L(y, $\hat{y}$)
1	0,9	?
1	0,5	?
1	0,1	?
0	0,9	?
0	0,5	?
0	0,1	?

15. D'après le tableau, que peux-tu dire de l'évolution de L lorsque la prédiction $\hat{y}$ se rapproche de la vraie classe y ? Et lorsqu'elle s'en éloigne fortement (prédiction fausse avec une grande confiance) ?

Pour m exemples d'entraînement, la fonction de coût globale J(a, b) est simplement la moyenne des erreurs individuelles :

Fonction de coût globale

$$\mathbf{J(a, b) = (1/m) \times \Sigma_i\, L(y_i, \hat{y}_i) \quad avec \quad \hat{y}_i = \sigma(a\, x_i + b)}$$

Partie E — Calcul détaillé des dérivées partielles

Niveau : Terminale — approfondissement

Pour minimiser J(a, b) par descente de gradient (document 0), il faut calculer ses dérivées partielles $\partial J/\partial a$ et $\partial J/\partial b$. On les calcule d'abord pour un seul exemple (x, y), avec z = ax + b et $\hat{y} = \sigma(z)$, avant de moyenner sur les m exemples.

On pose $L = -[y \ln(\hat{y}) + (1-y) \ln(1-\hat{y})]$. On cherche $\partial L/\partial a$, en utilisant la règle de dérivation des fonctions composées (règle de la chaîne) :

Règle de la chaîne à trois étages

$$\mathbf{\partial L/\partial a = (dL/d\hat{y}) \times (d\hat{y}/dz) \times (dz/da)}$$

16. Étape 1 : en dérivant $L = -[y \ln(\hat{y}) + (1-y) \ln(1-\hat{y})]$ par rapport à $\hat{y}$, montre que $dL/d\hat{y} = -y/\hat{y} + (1-y)/(1-\hat{y})$.
17. Étape 2 : d'après la partie C, que vaut $d\hat{y}/dz$?
18. Étape 3 : en multipliant les résultats des étapes 1 et 2, puis en mettant tout au même dénominateur $\hat{y}(1-\hat{y})$, montre — c'est l'étape la plus délicate, vas-y terme à terme — que le produit se simplifie remarquablement en : $dL/dz = \hat{y} - y$.
19. Étape 4 : comme z = ax + b, donne dz/da et dz/db.
20. Étape 5 : conclus que, pour un seul exemple :

Résultat

$$\mathbf{\partial L/\partial a = (\hat{y} - y) \times x \qquad \partial L/\partial b = (\hat{y} - y)}$$

En moyennant sur les m exemples de l'entraînement, on obtient enfin : $\partial J/\partial a = (1/m)\Sigma_i(\hat{y}_i - y_i)x_i$ et $\partial J/\partial b = (1/m)\Sigma_i(\hat{y}_i - y_i)$. Ce sont exactement les quantités que l'on va utiliser dans la partie F pour la descente de gradient.

Remarque : malgré une fonction de coût qui semblait compliquée (logarithmes, sigmoïde imbriquée), sa dérivée se simplifie en quelque chose d'aussi simple que « l'erreur de prédiction (ŷ − y), multipliée par x ». C'est cette simplicité qui rend la régression logistique si efficace à entraîner.

Partie F — Descente de gradient, pas à pas, à la main

Niveau : Terminale

On reprend 4 fruits, caractérisés seulement par leur diamètre x, avec les classes suivantes :

Fruit	x (diamètre)	y (classe)
1	1,0	0
2	2,0	0
3	3,0	1
4	4,0	1

On part de a = 0 et b = 0, avec un pas d'apprentissage $\eta = 0{,}1$, et on applique la règle de mise à jour de la descente de gradient :

Règle de mise à jour

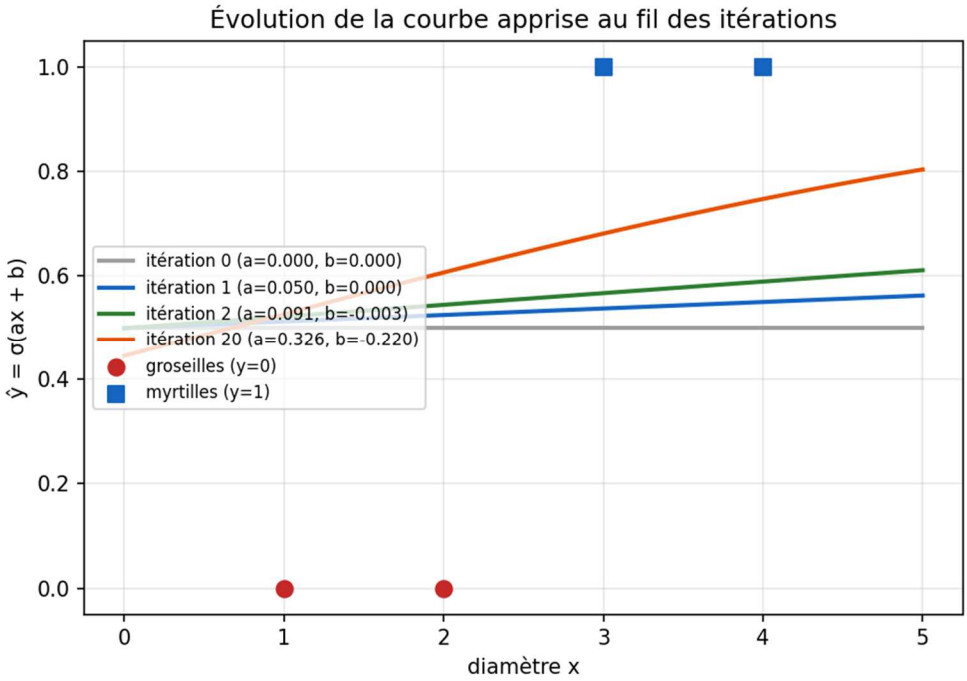
$$a \leftarrow a - \eta \times \partial J/\partial a \qquad b \leftarrow b - \eta \times \partial J/\partial b$$

21. Itération 1. Avec a = 0 et b = 0, calcule z, puis $\hat{y} = \sigma(z)$, pour chacun des 4 fruits. Complète le tableau :

Fruit	x	y	z = ax+b	$\hat{y} = \sigma(z)$	$\hat{y} - y$
1	1	0	?	?	?
2	2	0	?	?	?
3	3	1	?	?	?
4	4	1	?	?	?

22. En moyennant sur les 4 fruits, calcule $\partial J/\partial a = (1/4)\Sigma(\hat{y}-y)x$ et $\partial J/\partial b = (1/4)\Sigma(\hat{y}-y)$.
23. Mets à jour a et b avec $\eta = 0{,}1$. Que remarques-tu sur le signe de la variation de a ? Est-ce cohérent avec le fait qu'un grand diamètre doit plutôt correspondre à une myrtille (y=1) ?
24. Itération 2. Reprends les questions 1 à 3 avec les nouvelles valeurs de a et b (arrondis intermédiaires à 10^{-3} près).

Le graphique ci-dessous montre, à partir des mêmes données, comment la courbe $\hat{y} = \sigma(ax+b)$ évolue après 0, 1, 2 puis 20 itérations de descente de gradient (calcul fait par ordinateur, pour aller plus loin que ce que l'on peut faire à la main) :



25. Décris, avec tes mots, comment la courbe se transforme au fil des itérations. En quoi cette évolution traduit-elle un « apprentissage » ?

Partie G — Aller plus loin : classifier avec deux caractéristiques

Niveau : Terminale — ouverture

On revient maintenant au cas réel de la partie A, avec les deux caractéristiques (diamètre x_1, masse x_2). Le score z devient une combinaison des deux, exactement comme le produit scalaire du document 2 :

Modèle à deux caractéristiques

$$z = w_1x_1 + w_2x_2 + b \qquad \hat{y} = \sigma(z) = 1 / (1 + e^{(-z)})$$

On dispose de 6 fruits, et de paramètres de départ $w_1 = 1$, $w_2 = 1$, $b = -2$:

Fruit	Diamètre x_1 (cm)	Masse x_2 (g)	Type (y)
1	1,2	1,0	0 (groseille)
2	1,1	1,1	0 (groseille)
3	1,3	0,9	0 (groseille)
4	1,8	1,4	1 (myrtille)
5	2,0	1,6	1 (myrtille)
6	1,9	1,5	1 (myrtille)

G.1 — Prédiction et coût initial

26. Calcule la valeur de z pour le Fruit 1 avec les paramètres initiaux ($z = w_1x_1 + w_2x_2 + b$).
27. En déduis la prédiction ŷ en appliquant la fonction sigmoïde (arrondis à 10^{-3} près).
28. Calcule le coût pour ce fruit, en utilisant l'entropie croisée $L = -[□\cdot\ln(\hat{y}) + (1-□)\cdot\ln(1-\hat{y})]$ avec $□ = 0$.

G.2 — Calcul du gradient

Les dérivées partielles de la fonction de coût par rapport aux paramètres sont, par un calcul analogue à la partie E : $\partial L/\partial w_1 = (\hat{y}-□)\times x_1$; $\partial L/\partial w_2 = (\hat{y}-□)\times x_2$; $\partial L/\partial b = (\hat{y}-□)$.

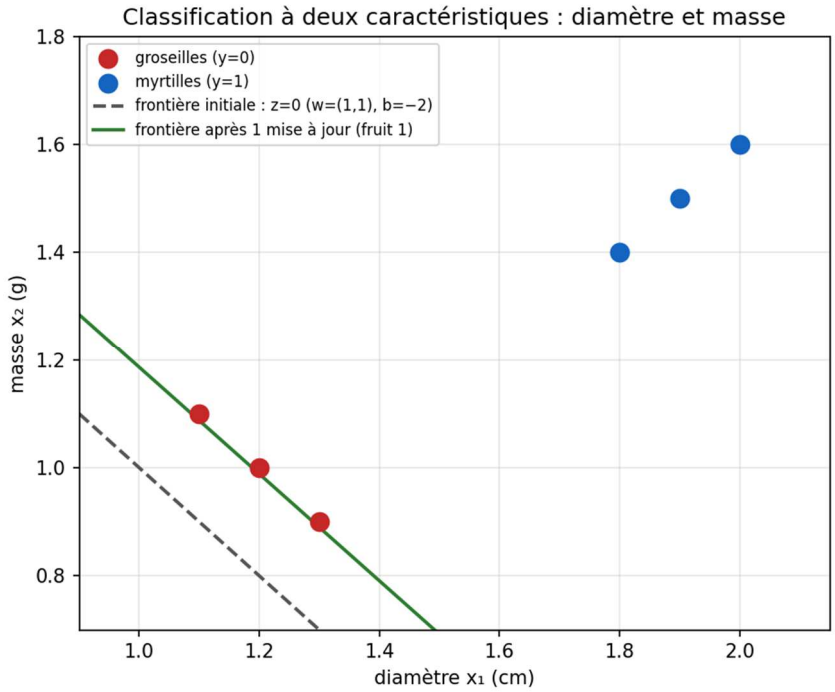
29. Calcule ces trois dérivées pour le Fruit 1.

G.3 — Mise à jour des paramètres

30. En utilisant nouveau paramètre = ancien paramètre − η × dérivée, avec η = 0,1, calcule les nouvelles valeurs de w_1, w_2 et b après une étape de descente de gradient sur le seul Fruit 1.

G.4 — Interprétation graphique

L'équation z = 0, c'est-à-dire $w_1x_1 + w_2x_2 + b = 0$, définit une droite dans le plan (x_1, x_2) : la frontière de décision ($\hat{y} = 0{,}5$ exactement sur cette droite).



31. Retrouve, à partir des paramètres initiaux $w_1 = 1$, $w_2 = 1$, $b = -2$, l'équation réduite $x_2 = \ldots$ de la frontière initiale en pointillés sur le graphique.
32. La frontière « après 1 mise à jour » a très peu bougé par rapport à la frontière initiale. Pourquoi, à ton avis, un seul fruit suffit-il si peu à faire bouger la droite ?

G.5 — Extension à plusieurs points

33. Propose une méthode pour calculer le coût moyen sur les 6 fruits (et non plus sur le seul Fruit 1).
34. Explique pourquoi il faut répéter la descente de gradient plusieurs fois (sur tous les fruits, plusieurs époques — document 9) pour converger vers une bonne frontière de décision.
35. Que se passe-t-il si le pas d'apprentissage η est trop grand ? Trop petit ? (Tu peux relire le document 0, partie D.)

Partie H — Et pour plus de deux catégories ?

Niveau : Terminale — ouverture

La sigmoïde ne fonctionne que pour deux catégories. Pour classer une image parmi 10 chiffres (0 à 9), ou pour prédire un mot parmi tout un vocabulaire (document 1), on généralise la sigmoïde en une fonction softmax, qui transforme un score pour chaque catégorie en une probabilité, de sorte que la somme de toutes les probabilités vaille 1 :

Fonction softmax (généralisation à K catégories)

$$\hat{y}_k = e^{z_k} / \Sigma_j\, e^{z_j}$$

La fonction de coût associée (entropie croisée multiclasse) et la descente de gradient fonctionnent alors exactement sur le même principe que dans ce document — seules les formules des dérivées se généralisent, sans rien changer à l'idée d'ensemble.

36. Pourquoi une IA comme Claude utilise-t-elle une fonction comme la sigmoïde (ou softmax) pour prédire un mot, plutôt que de choisir directement une catégorie ?
37. Quelle est la différence essentielle entre l'entropie croisée binaire (partie D) et l'entropie croisée multiclasse (softmax) ?
38. Pourquoi la descente de gradient est-elle si importante pour les modèles d'IA modernes, bien au-delà de ce simple exemple de groseilles et de myrtilles ?

39. En quoi cette démarche t'aide-t-elle à comprendre comment une IA comme Claude apprend à prédire un mot, une classe ou une réponse ? Fais un parallèle entre la séparation groseille/myrtille et une tâche de langage naturel (documents 1 et 2).

Ce que ce modèle simplifie, et ce que fait réellement une IA comme Claude

Ce que ce modèle simplifie	Ce que fait réellement un modèle comme Claude
Une seule caractéristique (le diamètre) et une droite $z = ax + b$. Un seul neurone (une seule sigmoïde) qui sépare deux catégories. Une frontière de décision forcément linéaire (une droite, ou un plan en dimension supérieure).	Des milliers de caractéristiques (les coordonnées du vecteur embedding — voir document 2), combinées à travers plusieurs couches de calcul. Des milliards de « neurones » artificiels, organisés en couches successives (réseaux profonds), chacun avec sa propre fonction d'activation. Des frontières de décision très complexes et non linéaires, rendues possibles par l'empilement de nombreuses couches non linéaires (c'est ce qui donne leur puissance aux réseaux de neurones profonds).

L'idée centrale — un score, transformé en probabilité, comparé à la vérité par une fonction de coût, puis amélioré par descente de gradient — reste, elle, exactement ce qui se passe à l'intérieur d'un modèle de langage, à une échelle de calcul totalement différente.

Exercices d'entraînement

Exercice 1 — Lecture de la sigmoïde

Niveau : Terminale

40. Sans calculatrice, indique si $\sigma(10)$ est plus proche de 0, de 0,5 ou de 1. Justifie.
41. Résous l'équation $\sigma(z) = 0{,}5$ d'inconnue z.
42. Pour quelles valeurs de z a-t-on $\sigma(z) > 0{,}5$? Que représente, dans le contexte de la classification, la condition « $z > 0$ » ?

Exercice 2 — Entropie croisée

Niveau : Terminale

Un modèle fait une prédiction $\hat{y} = 0{,}95$ pour un exemple de vraie classe $y = 0$ (il se trompe, et avec confiance).

43. Calcule l'entropie croisée L pour cet exemple.
44. Compare avec l'entropie croisée obtenue pour une prédiction $\hat{y} = 0{,}55$, toujours pour $y = 0$ (le modèle se trompe, mais avec moins de confiance). Que remarques-tu ?
45. Explique pourquoi cette propriété est recherchée pour une fonction de coût utilisée à l'entraînement.

Exercice 3 — Une itération de descente de gradient

Niveau : Terminale

On reprend le modèle $\hat{y} = \sigma(ax+b)$ avec, cette fois, 3 fruits : (x=1, y=0), (x=2, y=0), (x=4, y=1). On part de $a = 0{,}2$ et $b = -0{,}3$, avec $\eta = 0{,}2$.

46. Calcule z, $\hat{y}$ et $\hat{y}-y$ pour chacun des 3 fruits (arrondis à 10^{-3} près).
47. Calcule $\partial J/\partial a$ et $\partial J/\partial b$ (moyennes sur les 3 fruits).
48. Effectue la mise à jour de a et de b.

Corrigé

Partie A

A.1. Les abscisses représentent le diamètre des fruits (en mm) et les ordonnées leur masse (en g).

A.2. D'après le graphique, une droite raisonnable sépare les myrtilles (en haut à droite, diamètre et masse plus grands) des groseilles (en bas à gauche) : elle passe approximativement par les points (9,5 ; 0,55) et (11,5 ; 1,15).

A.3. En prenant deux points lus sur le graphique, par exemple (9,5 ; 0,55) et (11,5 ; 1,15) : $a = (1{,}15-0{,}55)/(11{,}5-9{,}5) = 0{,}6/2 = 0{,}3$, puis $b = 0{,}55-0{,}3\times 9{,}5 \approx -2{,}3$. Équation approchée : $y \approx 0{,}3x - 2{,}3$ (les valeurs exactes dépendent de la droite tracée ; toute droite séparant raisonnablement les deux nuages est acceptée).

A.4. Si $y_0 < ax_0+b$, le point $(x_0 ; y_0)$ est situé en dessous de la droite, dans la zone majoritairement occupée par les groseilles ; si $y_0 > ax_0+b$, il est au-dessus, dans la zone des myrtilles. Ce n'est pas infaillible : des points proches de la droite, ou des cas exceptionnels (une très grosse groseille, une petite myrtille), peuvent être mal classés.

A.5. Pour le Fruit 1 (9,8 ; 0,65) : $ax_0+b \approx 0{,}3\times9{,}8-2{,}3 = 0{,}64$, à comparer à $y_0=0{,}65 > 0{,}64$: tout juste au-dessus de la droite, plutôt une myrtille (mais très proche de la frontière). Pour le Fruit 2 (10,5 ; 0,7) : $ax_0+b \approx 0{,}3\times10{,}5-2{,}3 = 0{,}85$, à comparer à $y_0=0{,}7 < 0{,}85$: en dessous de la droite, plutôt une groseille. (Les réponses précises dépendent de la droite tracée à la question A.2 ; l'essentiel est la méthode.)

Partie B

B.1. $J(a,b) = (1/n)\Sigma(y_i-(ax_i+b))^2$. En dérivant par rapport à a (b fixé), chaque terme $(y_i-ax_i-b)^2$ se dérive, par la règle de la chaîne, en $2(y_i-ax_i-b)\times(-x_i)$. D'où $\partial J/\partial a = (1/n)\Sigma\, 2(y_i-ax_i-b)\times(-x_i) = -(2/n)\Sigma x_i(y_i-ax_i-b)$.

B.2. Par un calcul analogue (la dérivée de $(y_i-ax_i-b)^2$ par rapport à b vaut $2(y_i-ax_i-b)\times(-1)$) : $\partial J/\partial b = -(2/n)\Sigma(y_i-ax_i-b)$.

B.3. Chaque mise à jour déplace (a, b) dans la direction qui fait le plus diminuer J (document 0) : comme J mesure l'écart entre la droite et les points, minimiser J revient précisément à chercher la droite qui s'approche le mieux du nuage de points — donc, après suffisamment d'itérations, une droite très proche de celle tracée à l'œil en partie A.

Partie C

C.1. $\sigma(-5) \approx 0{,}007$; $\sigma(-2) \approx 0{,}119$; $\sigma(0) = 0{,}5$; $\sigma(2) \approx 0{,}881$; $\sigma(5) \approx 0{,}993$.

C.2. $\sigma(-z) = 1/(1+e^z) = e^{-z}/(e^{-z}(1+e^z))$... plus simplement : $1-\sigma(z) = 1-1/(1+e^{-z}) = e^{-z}/(1+e^{-z}) = 1/(e^z+1) = \sigma(-z)$. L'égalité est bien vérifiée.

C.3. Quand $z \to +\infty$, $e^{-z} \to 0$ donc $\sigma(z) \to 1$. Quand $z \to -\infty$, $e^{-z} \to +\infty$ donc $\sigma(z) \to 0$. La sigmoïde varie donc continûment entre 0 et 1, ce qui en fait une candidate naturelle pour représenter une probabilité (toujours comprise entre 0 et 1).

C.4. En écrivant $\sigma(z) = (1+e^{-z})^{-1}$ et en dérivant une fonction composée u^{-1} avec $u(z) = 1+e^{-z}$ (donc $u'(z) = -e^{-z}$) : $\sigma'(z) = -u'(z)\times u(z)^{-2} = e^{-z}/(1+e^{-z})^2$. On réécrit $e^{-z}/(1+e^{-z})^2 = [1/(1+e^{-z}))] \times [e^{-z}/(1+e^{-z}))] = \sigma(z)\times[1-\sigma(z)]$ (car $1-\sigma(z) = e^{-z}/(1+e^{-z})$, voir C.2). D'où $\sigma'(z) = \sigma(z)(1-\sigma(z))$.

Partie D

D.1. Si y=1 : $L = -[1\times\ln(\hat{y}) + 0\times\ln(1-\hat{y})] = -\ln(\hat{y})$. Si y=0 : $L = -[0\times\ln(\hat{y}) + 1\times\ln(1-\hat{y})] = -\ln(1-\hat{y})$.

y	$\hat{y}$	L(y, $\hat{y}$)
1	0,9	0,105
1	0,5	0,693
1	0,1	2,303
0	0,9	2,303
0	0,5	0,693
0	0,1	0,105

D.3. Plus la prédiction $\hat{y}$ est proche de la vraie classe y, plus L est petit (proche de 0) : le modèle est peu pénalisé pour une bonne prédiction. À l'inverse, quand la prédiction est très éloignée de y et faite avec une grande confiance (ex. y=1, $\hat{y}$=0,1), L devient très grand : l'entropie croisée pénalise sévèrement les erreurs commises « avec assurance ».

Partie E

E.1. $L = -y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})$. En dérivant terme à terme par rapport à $\hat{y}$: $d/d\hat{y}[-\square\ln(\hat{y})] = -\square/\hat{y}$; $d/d\hat{y}[-(1-\square)\ln(1-\hat{y})] = (1-y)/(1-\hat{y})$ (à cause de la dérivée de $\ln(1-\hat{y})$ qui vaut $-1/(1-\hat{y})$, et du signe « moins » devant). Donc $dL/d\hat{y} = -y/\hat{y} + (1-y)/(1-\hat{y})$.

E.2. D'après la partie C (question C.4) : $d\hat{y}/dz = \sigma(z)(1-\sigma(z)) = \hat{y}(1-\hat{y})$.

E.3. $dL/dz = [-y/\hat{y} + (1-y)/(1-\hat{y})] \times \hat{y}(1-\hat{y}) = -y(1-\hat{y}) + (1-y)\hat{y}$ (en distribuant : le terme $-y/\hat{y}$ se multiplie par $\hat{y}(1-\hat{y})$ et le $\hat{y}$ se simplifie, laissant $-\square(1-\hat{y})$; de même $(1-y)/(1-\hat{y})$ se multiplie par $\hat{y}(1-\hat{y})$, laissant $(1-y)\hat{y}$). En développant : $-y+y\hat{y}+\hat{y}-y\hat{y} = \hat{y}-y$. D'où $dL/dz = \hat{y} - y$.

E.4. $z = ax+b$, donc $dz/da = x$ et $dz/db = 1$.

E.5. $\partial L/\partial a = (dL/dz)\times(dz/da) = (\hat{y}-y)\times x$. $\partial L/\partial b = (dL/dz)\times(dz/db) = (\hat{y}-y)\times 1 = \hat{y}-y$.

Partie F — Itération 1

Fruit	x	y	z = ax+b	ŷ = σ(z)	ŷ − y
1	1	0	0	0,5	0,5
2	2	0	0	0,5	0,5
3	3	1	0	0,5	−0,5
4	4	1	0	0,5	−0,5

F.2. $\partial J/\partial a = (1/4)(0{,}5\times1 + 0{,}5\times2 - 0{,}5\times3 - 0{,}5\times4) = (1/4)(0{,}5+1-1{,}5-2) = (1/4)(-2) = -0{,}5$. $\partial J/\partial b = (1/4)(0{,}5+0{,}5-0{,}5-0{,}5) = 0$.

F.3. a_nouveau = 0 − 0,1×(−0,5) = 0,05 ; b_nouveau = 0 − 0,1×0 = 0. Le paramètre a augmente : c'est cohérent, puisqu'un a positif fait croître z (donc ŷ) avec le diamètre x, ce qui va dans le sens d'associer « grand diamètre » à « myrtille » (y=1).

F.4. Avec a=0,05 et b=0 : $z_1=0{,}05 \to \hat{y}_1\approx0{,}512$; $z_2=0{,}10 \to \hat{y}_2\approx0{,}525$; $z_3=0{,}15 \to \hat{y}_3\approx0{,}537$; $z_4=0{,}20 \to \hat{y}_4\approx0{,}550$. Erreurs : 0,512 ; 0,525 ; −0,463 ; −0,450. $\partial J/\partial a \approx (1/4)(0{,}512\times1+0{,}525\times2-0{,}463\times3-0{,}450\times4) \approx (1/4)(0{,}512+1{,}05-1{,}388-1{,}801) \approx (1/4)(-1{,}626) \approx -0{,}407$. $\partial J/\partial b \approx (1/4)(0{,}512+0{,}525-0{,}463-0{,}450) \approx (1/4)(0{,}125) \approx 0{,}031$. Mise à jour : $a \approx 0{,}05-0{,}1\times(-0{,}407) \approx 0{,}091$; $b \approx 0-0{,}1\times0{,}031 \approx -0{,}003$.

F.5. La courbe, initialement plate à 0,5 (aucune information), devient progressivement une sigmoïde de plus en plus « raide » et décalée, qui finit par attribuer une probabilité proche de 0 aux petits diamètres (groseilles) et proche de 1 aux grands diamètres (myrtilles) : c'est exactement la signature visuelle de l'apprentissage — le modèle ajuste sa courbe pour coller de mieux en mieux aux données observées.

Partie G

G.1. $z = w_1x_1+w_2x_2+b = 1\times1{,}2 + 1\times1{,}0 - 2 = 1{,}2+1{,}0-2 = 0{,}2$.

G.1 (suite). $\hat{y} = \sigma(0{,}2) = 1/(1+e^{(-0,2)}) \approx 1/1{,}819 \approx 0{,}550$.

G.1 (coût). Avec y=0 : $L = -\ln(1-0{,}550) = -\ln(0{,}450) \approx 0{,}798$.

G.2. $\partial L/\partial w_1 = (\hat{y}-y)\times x_1 = 0{,}550\times1{,}2 \approx 0{,}660$. $\partial L/\partial w_2 = (\hat{y}-y)\times x_2 = 0{,}550\times1{,}0 = 0{,}550$. $\partial L/\partial b = (\hat{y}-y) = 0{,}550$.

G.3. w_1_nouveau = 1 − 0,1×0,660 ≈ 0,934. w_2_nouveau = 1 − 0,1×0,550 ≈ 0,945. b_nouveau = −2 − 0,1×0,550 ≈ −2,055.

G.4. Avec $w_1=1$, $w_2=1$, b=−2 : $z=0 \Leftrightarrow x_1+x_2-2=0 \Leftrightarrow x_2 = -x_1+2$ (droite de pente −1 et d'ordonnée à l'origine 2), ce qui correspond bien à la droite en pointillés du graphique.

G.4 (suite). Un seul fruit ne représente qu'un gradient très partiel de l'erreur totale : la mise à jour ne tient compte que de l'écart de ce fruit, alors que la frontière doit satisfaire les 6 fruits à la fois. C'est pourquoi, en pratique, on moyenne (ou on utilise plusieurs fruits par lot — document 9) avant de mettre à jour, plutôt que de se fier à un seul exemple isolé.

G.5. On calcule le coût moyen en faisant la moyenne des entropies croisées individuelles sur les 6 fruits : $J(w_1,w_2,b) = (1/6)\Sigma_i\, L(y_i, \hat{y}_i)$, exactement comme en partie D de ce document, pour le cas à une seule variable.

G.5 (suite). Une seule itération ne corrige qu'un peu l'erreur ; il faut répéter la descente de gradient nombre de fois pour que les petits ajustements successifs s'accumulent et fassent converger la frontière vers une séparation satisfaisante pour l'ensemble des données, comme observé au document 0.

G.5 (fin). Si η est trop grand, les mises à jour peuvent osciller ou diverger (document 0, partie D) ; si η est trop petit, la convergence devient très lente, nécessitant un nombre d'itérations beaucoup plus grand pour un résultat équivalent.

Partie H

H.1. Une fonction comme la sigmoïde ou softmax transforme un score en probabilité, ce qui permet non seulement de choisir un mot ou une classe, mais aussi de quantifier la confiance du modèle dans sa prédiction — une information perdue si l'on ne gardait que la catégorie choisie.

H.2. L'entropie croisée binaire compare une seule probabilité ŷ à une classe $y\in\{0,1\}$; l'entropie croisée multiclasse compare tout un vecteur de probabilités (une par catégorie, obtenu par softmax) à un vecteur one-hot indiquant la bonne catégorie parmi K possibles.

H.3. La descente de gradient est la méthode d'apprentissage centrale de la quasi-totalité des IA actuelles (documents 0, 4, 5) : sans elle, aucun des milliards de paramètres d'un modèle de langage ne pourrait être ajusté à partir des données d'entraînement.

H.4. La séparation groseille/myrtille et la prédiction d'un mot suivent exactement le même schéma mathématique : un score (produit scalaire, documents 2 et 3), transformé en probabilité par sigmoïde ou softmax, comparé à la vérité par une entropie croisée, puis amélioré par descente de gradient — seule la taille du problème change (2 catégories et 1-2 caractéristiques, contre des dizaines de milliers de mots possibles et des milliers de caractéristiques).

Exercice 1

1. $\sigma(10)$ est très proche de 1, car pour z grand et positif, e^(-z) devient quasiment nul, donc $\sigma(z) = 1/(1+e^{(-z)}) \approx 1$.

2. $\sigma(z) = 0{,}5 \iff 1/(1+e^{(-z)}) = 1/2 \iff 1+e^{(-z)} = 2 \iff e^{(-z)} = 1 \iff z = 0$.

3. $\sigma(z) > 0{,}5 \iff z > 0$ (car σ est strictement croissante). Dans le contexte de la classification, « $z > 0$ » signifie que le point est du côté « myrtille » de la droite de séparation (au sens de la partie A) : le modèle prédit alors la classe 1 (probabilité supérieure à 0,5).

Exercice 2

1. $L = -\ln(1-0{,}95) = -\ln(0{,}05) \approx 2{,}996$.

2. Pour $\hat{y}=0{,}55$: $L = -\ln(1-0{,}55) = -\ln(0{,}45) \approx 0{,}799$. On observe que $2{,}996 \gg 0{,}799$: une erreur commise avec beaucoup de confiance est bien plus pénalisée qu'une erreur commise avec une confiance modérée.

3. C'est une propriété recherchée car elle pousse le modèle, pendant l'entraînement, à corriger en priorité ses erreurs les plus « graves » (celles commises avec une grande confiance), plutôt que de traiter toutes les erreurs de la même façon.

Exercice 3

1. Fruit 1 ($x=1$) : $z = 0{,}2\times1-0{,}3 = -0{,}1 \to \hat{y} = \sigma(-0{,}1) \approx 0{,}475 \to \hat{y}-y \approx 0{,}475$. Fruit 2 ($x=2$) : $z = 0{,}2\times2-0{,}3 = 0{,}1 \to \hat{y} \approx 0{,}525 \to \hat{y}-y \approx 0{,}525$. Fruit 3 ($x=4$) : $z = 0{,}2\times4-0{,}3 = 0{,}5 \to \hat{y} \approx 0{,}622 \to \hat{y}-y \approx 0{,}622-1 = -0{,}378$.

2. $\partial J/\partial a \approx (1/3)(0{,}475\times1 + 0{,}525\times2 - 0{,}378\times4) = (1/3)(0{,}475+1{,}05-1{,}512) = (1/3)(0{,}013) \approx 0{,}004$. $\partial J/\partial b \approx (1/3)(0{,}475+0{,}525-0{,}378) = (1/3)(0{,}622) \approx 0{,}207$.

3. $a_{nouveau} \approx 0{,}2 - 0{,}2\times0{,}004 \approx 0{,}199$. $b_{nouveau} \approx -0{,}3 - 0{,}2\times0{,}207 \approx -0{,}341$.

Mathématiques du programme mobilisées dans ce document

Seconde / Première

Nuages de points, ajustement affine, équation réduite d'une droite.

Inéquations du premier degré, résolution graphique.

Terminale (spécialité mathématiques)

Fonction exponentielle, limites, sens de variation.

Fonction logarithme népérien, propriétés algébriques.

Dérivée d'une fonction composée (règle de la chaîne), dérivée d'une fonction de la forme 1/u.

Résolution d'équations et d'inéquations utilisant l'exponentielle et le logarithme.

Dérivées partielles et gradient (généralisation à plusieurs variables), en lien avec le document 0.

Combinaison linéaire de deux variables (produit scalaire à deux caractéristiques, partie G), en lien avec le document 2.