Panorama des fonctions de coût

Toutes les erreurs ne se mesurent pas de la même façon

Document de synthèse

Ce document conclut la série en reliant les documents précédents : la descente de gradient (document 0) a besoin d'une fonction dérivable ; l'entropie croisée (documents 1 et 3) et l'erreur quadratique (document 4) en sont deux exemples. On découvre ici deux nouvelles fonctions de coût, et surtout pourquoi il en existe plusieurs.

Niveau : Terminale (synthèse) — accessible avec les documents 0, 1, 3 et 4 déjà vus

Introduction : un même principe, plusieurs façons de mesurer l'erreur

Dans tous les documents précédents, une IA apprenait en minimisant une fonction de coût J, à l'aide de la descente de gradient (document 0). Mais la fonction J n'était jamais la même : erreur quadratique pour prédire un score (document 4), entropie croisée pour classer un fruit (document 3) ou prédire un mot (document 1).

Ce document répond à une question simple : pourquoi ne pas toujours utiliser la même fonction de coût ?

On va voir qu'il existe même une fonction de coût « idéale », plus naturelle que toutes les autres — mais qu'elle pose un problème rédhibitoire pour la descente de gradient.

Partie A — Le candidat le plus naturel : la fonction 0–1

Niveau : Terminale

L'idée la plus simple pour mesurer l'erreur d'un classificateur est de compter directement le nombre de prédictions fausses :

Fonction de coût 0–1

$$J = (1/m) \times \Sigma_i \, I(\hat{y}_i \neq y_i)$$

où $I(\hat{y}_i \neq y_i)$ est la fonction indicatrice : elle vaut 1 si la prédiction est fausse, et 0 si elle est correcte.

Un modèle a fait les prédictions suivantes sur 5 exemples :

Exemple	1	2	3	4	5
y (vraie classe)	1	0	0	1	1
ŷ (prédiction)	1	1	0	0	1

1. Calcule la fonction de coût 0–1 pour ces 5 exemples.
2. Cette fonction de coût semble être la plus naturelle possible : elle correspond exactement à ce qu'on veut minimiser (le nombre d'erreurs). Pourtant, elle n'est presque jamais utilisée pour entraîner une IA par descente de gradient. Relis le document 0 : quelle propriété est indispensable pour appliquer la descente de gradient ?
3. La fonction 0–1 change brutalement de valeur (de 0 à 1) dès que la prédiction franchit la frontière de décision, puis reste plate. Pourquoi cela pose-t-il un problème pour calculer un gradient ?

C'est ce problème qui a motivé la création de toutes les fonctions de coût « alternatives » que tu as déjà rencontrées : elles doivent ressembler à la fonction 0–1 (pénaliser les erreurs), tout en restant dérivables partout.

Partie B — Rappel : l'erreur quadratique moyenne (régression)

Niveau : Terminale

Pour prédire une valeur numérique (document 4), on a utilisé :

Rappel

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

On donne 3 points (1 ; 2), (2 ; 3) et (3 ; 5), et un modèle $\hat{y} = 1{,}5x + 0{,}3$.

4. Calcule ŷ pour chacun des 3 points, puis l'erreur ŷ−y et son carré.
5. En déduire la valeur de J(1,5 ; 0,3).

6. Pourquoi cette fonction, contrairement à la fonction 0–1, est-elle dérivable en tout point (on pourra s'appuyer sur sa forme, une parabole en chaque paramètre — voir document 0) ?

Partie C — Rappel : l'entropie croisée (classification)

Niveau : Terminale

Pour classer en deux catégories (document 3) ou prédire un mot parmi tout un vocabulaire (document 1), on a utilisé l'entropie croisée :

Rappel

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

On donne 3 exemples : (y=1, ŷ=0,8), (y=0, ŷ=0,3), (y=1, ŷ=0,4).

7. Calcule l'entropie croisée pour chacun des 3 exemples (arrondis à 10^{-3} près), puis leur moyenne.
8. Parmi les 3 exemples, lequel est le plus pénalisé ? Le moins pénalisé ? Est-ce cohérent avec la qualité des prédictions ?

Partie D — Nouveau : la perte hinge (marge maximale)

Niveau : Terminale — ouverture / mathématiques expertes

Une autre approche de classification, utilisée par les machines à vecteurs de support (SVM), n'essaie pas de prédire une probabilité, mais un score brut f(x) = ax+b, et encode les classes par $y \in \{-1 ; +1\}$ (au lieu de {0 ; 1}). On utilise alors la perte hinge (« charnière ») :

Définition

$$\text{Perte}(y, f(x)) = \max(0, 1 - y \times f(x))$$

L'idée : si $y \times f(x) \geqslant 1$, la prédiction est correcte avec une marge suffisante, et la perte est nulle. Sinon, plus $y \times f(x)$ est petit (voire négatif), plus la perte est grande.

9. Complète le tableau suivant :

y	f(x)	y × f(x)	Perte hinge
+1	+2	?	?
+1	+0,5	?	?
+1	−1	?	?
−1	−3	?	?
−1	+0,5	?	?

10. Explique pourquoi une prédiction correcte mais « proche de la frontière » (par exemple y=+1, f(x)=+0,5) reçoit tout de même une perte strictement positive.
11. La perte hinge est-elle dérivable en $y \times f(x) = 1$? (Indice : regarde ce qui se passe de part et d'autre de ce point.) Est-ce gênant en pratique pour la descente de gradient ?

Partie E — Nouveau : comparer deux distributions (divergence de KL)

Niveau : Terminale — ouverture

Dans le document 1, une IA prédisait une distribution de probabilité sur le mot suivant. Pour mesurer l'écart entre la distribution prédite Q et la « vraie » distribution P (par exemple, celle observée dans un immense corpus de textes), on utilise la divergence de Kullback-Leibler :

Définition

$$D_{KL}(P \,\|\, Q) = \sum_i P(i) \times \ln\left(P(i) / Q(i) \right)$$

On admet que $D_{KL}(P\|Q) \geqslant 0$ toujours, avec égalité si et seulement si P = Q.

On donne une distribution réelle P = (0,7 ; 0,2 ; 0,1) et une distribution prédite Q = (0,5 ; 0,3 ; 0,2) sur 3 mots possibles.

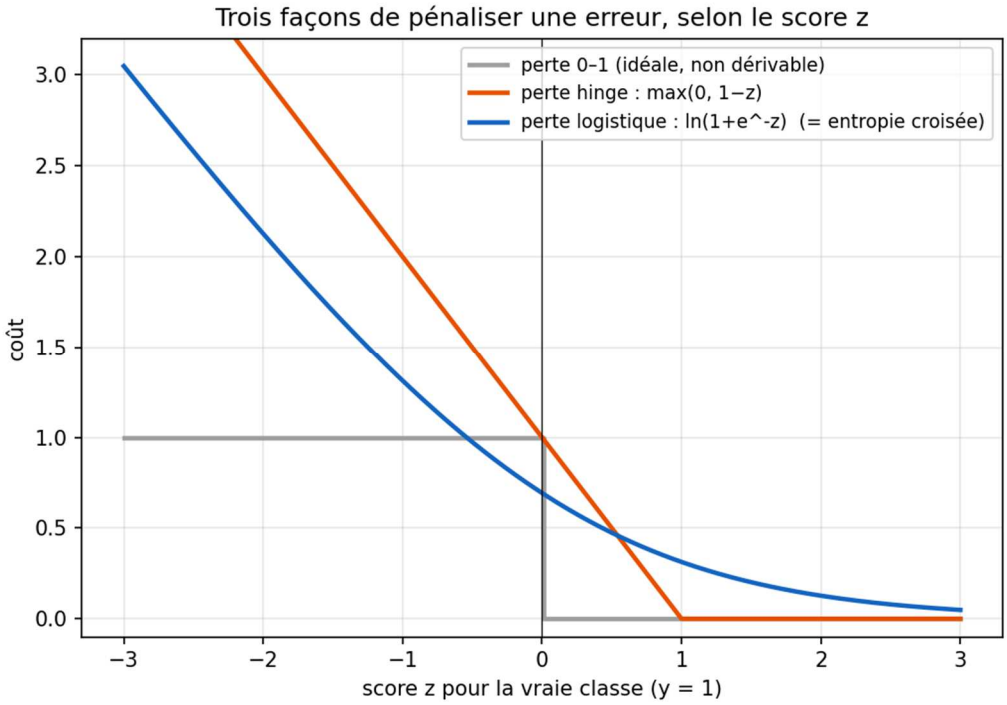
12. Calcule $D_{KL}(P \,\|\, Q)$ (arrondis à 10^{-3} près).
13. Calcule maintenant $D_{KL}(Q \,\|\, P)$ (on inverse les rôles). Compare avec la question précédente.

14. Une vraie distance mathématique (comme la distance entre deux points, document 2) est toujours symétrique. D'après la question précédente, la divergence de KL est-elle une distance au sens strict ?

Partie F — Une vision d'ensemble

Niveau : Terminale

Le graphique suivant compare, pour un exemple de vraie classe y=1, trois façons de pénaliser une prédiction selon son score z : la fonction 0–1 (idéale mais non dérivable), la perte hinge, et la perte logistique (qui n'est autre que l'entropie croisée écrite en fonction du score, puisque $-\ln(\sigma(z)) = \ln(1+e^{-z})$).



15. D'après le graphique, comment les pertes hinge et logistique se comportent-elles par rapport à la fonction 0–1 ? (Sont-elles toujours au-dessus, en dessous, tantôt l'une tantôt l'autre ?)
16. En quoi cette observation confirme-t-elle l'idée que ces fonctions sont des « versions lissées, dérivables partout » de la fonction 0–1 idéale ?

Partie G — Synthèse : quelle fonction de coût pour quelle tâche ?

Niveau : Terminale

Tâche	Fonction de coût	Vu dans
Prédire une valeur numérique (régression)	Erreur quadratique moyenne (MSE)	Document 4
Classer en 2 catégories	Entropie croisée binaire	Document 3
Classer en plusieurs catégories / prédire un mot	Entropie croisée multiclasse (softmax)	Documents 1 et 3
Classer avec une marge maximale (SVM)	Perte hinge	Ce document
Comparer deux distributions de probabilité	Divergence de KL	Ce document, document 1
Compter le nombre d'erreurs (évaluation finale)	Fonction 0–1	Ce document

Pour conclure la série : ce que tous ces documents simplifient

Ce que ce modèle simplifie	Ce que fait réellement un modèle comme Claude
Une seule fonction de coût à la fois, clairement identifiée, pour une tâche isolée et bien définie. Un entraînement qui s'arrête dès que J est suffisamment petit. Des exemples calculables à la main, avec quelques points ou quelques mots.	Un modèle comme Claude combine en réalité plusieurs objectifs à la fois pendant son entraînement (la fonction de coût principale, mais aussi des termes de régularisation, et des étapes d'affinage supplémentaires basées sur les préférences humaines — le renforcement par retour humain, ou RLHF). L'entraînement s'arrête selon des critères bien plus complexes (surveillance de nombreuses métriques, risque de sur-apprentissage, contraintes de temps de calcul). Des milliards d'exemples d'entraînement, traités par lots de milliers à la fois, sur des semaines de calcul, avec des milliards de paramètres ajustés simultanément.

Mais dans tous les cas : une fonction de coût qui mesure une erreur, et une descente de gradient qui l'améliore petit à petit — c'est là tout le cœur mathématique, du plus simple exercice de ce document au fonctionnement réel d'un modèle de langage.

Exercices d'entraînement

Exercice 1 — Fonction 0–1 et dérivabilité

Niveau : Terminale

Un modèle a fait les prédictions suivantes : (y=1, ŷ=1), (y=1, ŷ=0), (y=0, ŷ=0), (y=0, ŷ=1).

17. Calcule la fonction de coût 0–1 sur ces 4 exemples.
18. Explique en une phrase pourquoi on ne peut pas appliquer la descente de gradient directement à cette fonction de coût.

Exercice 2 — Perte hinge

Niveau : Terminale

On donne 3 exemples : (y=+1, f(x)=1,2), (y=−1, f(x)=−0,3), (y=+1, f(x)=−0,5).

19. Calcule y×f(x) et la perte hinge pour chacun des 3 exemples.
20. Calcule la perte hinge moyenne sur ces 3 exemples.
21. Lequel des 3 exemples est le plus problématique pour le modèle ? Justifie à partir de la valeur de y×f(x).

Exercice 3 — Divergence de KL et asymétrie

Niveau : Terminale

On donne P = (0,9 ; 0,1) et Q = (0,6 ; 0,4), deux distributions sur 2 catégories.

22. Calcule $D_{KL}(P \| Q)$.
23. Calcule $D_{KL}(Q \| P)$.
24. Les deux résultats sont-ils égaux ? Que peux-tu en conclure sur la nature de la divergence de KL ?

Corrigé

Partie A

A.1. Comparaison des 5 exemples : (1,1) correct, (0,1) faux, (0,0) correct, (1,0) faux, (1,1) correct. J = (1/5)(0+1+0+1+0) = 2/5 = 0,4.

A.2. D'après le document 0, la descente de gradient nécessite de calculer une dérivée (un gradient) à chaque étape, pour savoir dans quelle direction ajuster les paramètres.

A.3. La fonction 0–1 est constante par morceaux (elle ne prend que les valeurs 0 ou 1) : sa dérivée vaut 0 presque partout, et n'est même pas définie à l'endroit où elle « saute » brutalement de 0 à 1. Un gradient nul ne donne aucune indication sur la direction à suivre pour améliorer le modèle : la descente de gradient ne peut donc pas progresser avec cette fonction de coût.

Partie B

B.1. Pour x=1 : ŷ=1,5×1+0,3=1,8, erreur=1,8−2=−0,2, carré=0,04. Pour x=2 : ŷ=3,3, erreur=0,3, carré=0,09. Pour x=3 : ŷ=4,8, erreur=−0,2, carré=0,04.

B.2. J = (1/(2×3))×(0,04+0,09+0,04) = 0,17/6 ≈ 0,028.

B.3. Cette fonction est une somme de carrés de fonctions affines de a et de b : elle est donc polynomiale (du second degré en a et en b), donc dérivable en tout point, sans aucune cassure ni saut — contrairement à la fonction 0–1.

Partie C

C.1. Pour (y=1, ŷ=0,8) : L=−ln(0,8)≈0,223. Pour (y=0, ŷ=0,3) : L=−ln(0,7)≈0,357. Pour (y=1, ŷ=0,4) : L=−ln(0,4)≈0,916. Moyenne ≈ (0,223+0,357+0,916)/3 ≈ 0,499.

C.2. Le plus pénalisé est (y=1, ŷ=0,4) avec L≈0,916, car sa prédiction est la plus éloignée de la vraie classe. Le moins pénalisé est (y=1, ŷ=0,8) avec L≈0,223, car c'est la meilleure prédiction des trois. C'est bien cohérent : plus la prédiction est fausse, plus la pénalité est grande.

Partie D

y	f(x)	y × f(x)	Perte hinge
+1	+2	2	max(0, 1−2) = 0
+1	+0,5	0,5	max(0, 0,5) = 0,5
+1	−1	−1	max(0, 2) = 2
−1	−3	3	max(0, −2) = 0
−1	+0,5	−0,5	max(0, 1,5) = 1,5

D.2. Même correcte (le signe de y et de f(x) concorde), une prédiction trop proche de la frontière de décision (y×f(x) < 1) est jugée « pas assez sûre » : la perte hinge encourage le modèle à s'éloigner nettement de la frontière, pas seulement à être du bon côté, afin d'obtenir une marge de sécurité (c'est tout l'intérêt des SVM, qui cherchent la marge maximale).

D.3. En y×f(x) = 1, la perte hinge passe d'une partie constante (perte nulle, pente 0) à une partie affine décroissante (pente −1) : elle a donc un point anguleux, où elle n'est pas dérivable (comme la fonction valeur absolue en 0). C'est gênant en théorie, mais peu gênant en pratique : la fonction reste dérivable presque partout (sauf en ce seul point), et des techniques adaptées (sous-gradients) permettent quand même d'utiliser une méthode proche de la descente de gradient.

Partie E

E.1. $D_{KL}(P\|Q) = 0,7\times\ln(0,7/0,5) + 0,2\times\ln(0,2/0,3) + 0,1\times\ln(0,1/0,2) \approx 0,7\times0,336 + 0,2\times(-0,405) + 0,1\times(-0,693) \approx 0,236 - 0,081 - 0,069 \approx 0,085$.

E.2. $D_{KL}(Q\|P) = 0,5\times\ln(0,5/0,7) + 0,3\times\ln(0,3/0,2) + 0,2\times\ln(0,2/0,1) \approx 0,5\times(-0,336) + 0,3\times0,405 + 0,2\times0,693 \approx -0,168+0,122+0,139 \approx 0,093$. (Valeur différente de E.1.)

E.3. Les deux valeurs sont différentes : la divergence de KL n'est pas symétrique ($D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$ en général). Ce n'est donc pas une distance au sens mathématique strict (une vraie distance doit être symétrique) : on parle de « divergence », pas de « distance ».

Partie F

F.1. Sur le graphique, la perte hinge et la perte logistique restent toujours au-dessus (ou égales à) la fonction 0–1 : elles majorent la fonction 0–1 en tout point.

F.2. Puisqu'elles majorent la fonction 0–1 tout en étant dérivables partout (ou presque, pour la hinge), minimiser ces fonctions lissées revient à minimiser (indirectement, mais efficacement) une approximation raisonnable du nombre réel d'erreurs — ce qui est exactement ce que l'on voulait faire avec la fonction 0–1, sans pouvoir y appliquer la descente de gradient.

Exercice 1

1. Comparaisons : (1,1) correct, (1,0) faux, (0,0) correct, (0,1) faux. J = (1/4)(0+1+0+1) = 2/4 = 0,5.

2. La fonction 0–1 a une dérivée nulle presque partout et n'est pas définie aux points de saut : elle ne donne aucune information exploitable sur la direction dans laquelle ajuster les paramètres, ce qui rend la descente de gradient inutilisable.

Exercice 2

1. (y=+1,f=1,2) : y×f=1,2, perte=max(0,1−1,2)=0. (y=−1,f=−0,3) : y×f=0,3, perte=max(0,0,7)=0,7. (y=+1,f=−0,5) : y×f=−0,5, perte=max(0,1,5)=1,5.

2. Moyenne = (0+0,7+1,5)/3 = 2,2/3 ≈ 0,733.

3. Le plus problématique est (y=+1, f=−0,5), car y×f=−0,5 < 0 : le signe de la prédiction est même opposé à la vraie classe (erreur de classification), pas seulement une marge insuffisante.

Exercice 3

1. $D_{KL}(P\|Q) = 0{,}9\times\ln(0{,}9/0{,}6) + 0{,}1\times\ln(0{,}1/0{,}4) \approx 0{,}9\times0{,}405 + 0{,}1\times(-1{,}386) \approx 0{,}365 - 0{,}139 \approx 0{,}226$.

2. $D_{KL}(Q\|P) = 0{,}6\times\ln(0{,}6/0{,}9) + 0{,}4\times\ln(0{,}4/0{,}1) \approx 0{,}6\times(-0{,}405) + 0{,}4\times1{,}386 \approx -0{,}243+0{,}555 \approx 0{,}311$.

3. Les deux valeurs (0,226 et 0,311) sont différentes : cela confirme que la divergence de KL n'est pas symétrique, donc pas une distance au sens strict.

Mathématiques du programme mobilisées dans ce document

Terminale (spécialité mathématiques)

Dérivabilité d'une fonction, points anguleux, fonctions définies par morceaux.

Fonction logarithme népérien, fonction exponentielle.

Étude et comparaison de fonctions (majoration, comportement asymptotique).

Terminale (option mathématiques expertes) / ouverture

Notion de divergence entre deux distributions de probabilité (théorie de l'information).

Fonctions convexes, optimisation sous contrainte de marge (lien avec les SVM).