Le **principe de la chaîne** (ou **règle de la chaîne**) est une technique essentielle pour la descente de gradient dans les réseaux de neurones. Il permet de calculer les dérivées partielles de la fonction de coût par rapport aux paramètres du réseau, en particulier lorsque le réseau comporte plusieurs couches (ou neurones).

Rappel : descente de gradient dans un neurone La **descente de gradient** consiste à calculer les gradients de la fonction de coût (erreur) par rapport aux poids et au biais d'un neurone, puis à ajuster ces paramètres pour minimiser l'erreur. Quand on a un réseau simple avec une seule couche, cela se fait directement. Mais dans un réseau plus complexe avec plusieurs couches, il faut utiliser la **règle de la chaîne**.

Principe de la règle de la chaîne

La règle de la chaîne permet de dériver une fonction composée, c'est-à-dire une fonction qui dépend de plusieurs autres fonctions imbriquées. Dans un réseau de neurones, chaque couche (ou neurone) applique une transformation à l'entrée, et ces transformations se suivent pour produire une sortie.

Prenons un exemple simple d'un réseau à **deux couches** pour illustrer la descente de gradient et la règle de la chaîne.

Exemple : réseau à deux couches

- La première couche (neurone 1) reçoit une entrée x, applique un poids W_1, un biais b_1, puis une fonction d'activation σ_1. - La sortie de la première couche devient l'entrée de la deuxième couche (neurone 2), qui applique à nouveau un poids W_2, un biais b_2, et une fonction d'activation σ_2. - La sortie finale est comparée à la vraie valeur y_{true} à l'aide d'une fonction de coût C.

Les étapes de calcul pour ce réseau à deux couches sont les suivantes :

1. **Calcul de la première couche :**

$$z_1 = W_1 \cdot x + b_1$$

$$a_1 = \sigma_1(z_1)$$

2. **Calcul de la deuxième couche :**

$$z_2 = W_2 \cdot a_1 + b_2$$

$$a_2 = \sigma_2(z_2)$$

3. **Calcul de l'erreur avec la fonction de coût :**

$$C = \frac{1}{2}(a_2 - y_{true})^2$$

Utilisation de la règle de la chaîne

Pour appliquer la descente de gradient, nous devons dériver la fonction de coût C par rapport aux poids W_1, W_2, aux biais b_1, b_2, et aux activations intermédiaires a_1, etc. C'est ici que la **règle de la chaîne** intervient.

Calcul du gradient de W_2

Commençons par calculer la dérivée de C par rapport à W_2. La fonction C dépend de W_2 via plusieurs étapes intermédiaires, donc nous devons utiliser la règle de la chaîne :

$$\frac{\partial C}{\partial W_2} = \frac{\partial C}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial W_2}$$

- **$\frac{\partial C}{\partial a_2}$** : Cela représente la dérivée de la fonction de coût par rapport à la sortie finale a_2 :

$$\frac{\partial C}{\partial a_2} = a_2 - y_{true}$$

- **$\frac{\partial a_2}{\partial z_2}$** : Cela représente la dérivée de la fonction d'activation σ_2 par rapport à l'entrée de la deuxième couche z_2 :

$$\frac{\partial a_2}{\partial z_2} = \sigma_2'(z_2)$$

- **$\frac{\partial z_2}{\partial W_2}$** : Cela représente la dérivée de z_2 par rapport à W_2, ce qui est simplement a_1 :

$$\frac{\partial z_2}{\partial W_2} = a_1$$

En combinant tout cela :

$$\frac{\partial C}{\partial W_2} = (a_2 - y_{true}) \cdot \sigma_2'(z_2) \cdot a_1$$

Calcul du gradient de W_1

Le calcul de $\frac{\partial C}{\partial W_1}$ est un peu plus complexe car W_1 affecte la sortie via deux couches. Nous devons appliquer la règle de la chaîne encore plus profondément :

$$\frac{\partial C}{\partial W_1} = \frac{\partial C}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial W_1}$$

- Nous avons déjà calculé les premiers termes $\frac{\partial C}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2}$ ci-dessus. - **$\frac{\partial z_2}{\partial a_1} = W_2$** (car z_2 dépend de a_1). - **$\frac{\partial a_1}{\partial z_1} = \sigma_1'(z_1)$**. - **$\frac{\partial z_1}{\partial W_1} = x$** (car z_1 dépend directement de W_1).

En combinant ces termes :

$$\frac{\partial C}{\partial W_1} = (a_2 - y_{true}) \cdot \sigma_2'(z_2) \cdot W_2 \cdot \sigma_1'(z_1) \cdot x$$

Conclusion

La **règle de la chaîne** permet de calculer les gradients de la fonction de coût dans les réseaux de neurones en traversant chaque couche du réseau. Pour chaque paramètre (poids ou biais), la descente de gradient utilise cette technique pour ajuster les valeurs et minimiser l'erreur. Ce principe est fondamental pour l'entraînement des réseaux de neurones, qu'ils soient simples ou profonds.

Je vois où est la confusion, et tu soulèves un point important. Clarifions ensemble ce qui se passe exactement avec les dérivées pour que tu comprennes

pourquoi, dans certaines situations, il ne reste pas juste 1 mais plutôt x_1 quand on dérive z_1 par rapport à W_1.

Reprenons l'expression de z_1 Dans la première couche, tu as :

$$z_1 = W_1 \cdot x_1 + b_1$$

Ici : - W_1 est un **poids** (la variable que tu veux ajuster). - x_1 est l'**entrée** de cette couche (c'est une valeur donnée). - b_1 est le **biais** (une constante ajoutée).

Maintenant, nous allons **dériver** z_1 par rapport à W_1.

1. Dérivée de z_1 par rapport à W_1 La règle que tu appliques ici est la dérivée de la somme d'expressions, ainsi que la dérivée d'une multiplication. Dérivons chaque terme séparément :

- $**W_1 \cdot x_1**$: Il s'agit d'une multiplication. On utilise la règle de la dérivée d'une constante multipliée par une variable. Ici, x_1 est une **constante** (c'est l'entrée du réseau, donnée), et W_1 est la variable. La dérivée de $W_1 \cdot x_1$ par rapport à W_1 est donc $**x_1**$. Pourquoi ? Parce que la dérivée de $a \cdot x$ par rapport à x est a, où a est une constante. Ici, x_1 joue le rôle de a.

$$\frac{\partial}{\partial W_1}(W_1 \cdot x_1) = x_1$$

- $**b_1**$: C'est une constante indépendante de W_1. La dérivée d'une constante est toujours 0.

$$\frac{\partial}{\partial W_1}(b_1) = 0$$

2. Résultat de la dérivée En combinant les deux termes, on obtient :

$$\frac{\partial z_1}{\partial W_1} = x_1$$

Cela signifie que la dérivée de z_1 par rapport à W_1 donne $**x_1**$.

Pourquoi ce n'est pas juste 1 ? C'est ici que la confusion semble apparaître. Si on dérivait une expression comme $W_1 \cdot x_1$ **par rapport à x_1**, alors oui, tu aurais juste 1 , car tu dériverais la variable par rapport à elle-même. Mais ici, tu dérives par rapport à W_1, et x_1 est considéré comme une **constante**.

Exemples pour clarifier

1. **Dérivée d'une constante multipliée par une variable** :

$$\frac{d}{dx}(3x) = 3$$

Le 3 reste car il est multiplié par la variable.

2. **Dérivée par rapport à la constante dans une multiplication** : Si tu dérives $3 \cdot x$ par rapport à la constante 3, tu aurais :

$$\frac{d}{d3}(3x) = x$$

La dérivée est x parce que c'est la variable.

Dans ton cas, x_1 joue le rôle de cette constante.

Conclusion Quand tu dérives $z_1 = W_1 \cdot x_1 + b_1$ par rapport à W_1, tu considères que x_1 est une constante (puisque tu dérives par rapport à W_1, pas x_1), donc il reste $**x_1**$, et pas juste 1 . Le biais b_1 est une constante pure, donc sa dérivée est 0.

Cela t'éclaire-t-il mieux sur pourquoi il reste x_1 dans la dérivée de z_1 par rapport à W_1 ?

test

b

October 2024

1 Introduction