
# *Reconnaissance des formes pour l'analyse et l'interprétation d'images*

Rapport TP 3-4 : Introduction aux réseaux de neurones

**Etudiant** :

DJEGHRI Amine

MAMOU Idles

**Numéro :**

3801757

3803676

**Novembre 2020**

# Partie 1 – Formalisation mathématique

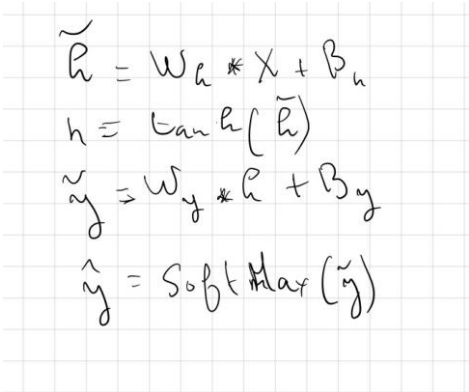
## 1.1 Jeu de données

1. L'ensemble d'apprentissage sert à entrainer le modèle
   L'ensemble de validation sert à choisir les meilleurs hyper-paramètres de notre modèle
   L'ensemble de test sert à évaluer notre modèle et voir ses performances si on l'applique sur un jeu de donnée inconnu
2. L'influence du nombre N d'exemples :
   Plus N le nombre d'exemple est grand, plus le modèle sera mieux entrainé et par conséquent le modèle pourra mieux de généraliser.

## 1.2 Architecture du réseau (phase forward)

3. Les fonctions d'activation permettent de transformer nos fonctions linéaires en non linéaire et ainsi rendre notre modèle complexe et pourvoir traiter des problèmes complexes (par exemple des données non linéairement séparable) qui ne peuvent être bien résolu avec des fonctions linéaires.
4. $n_x = 2$ $(la\ taille\ n'est\ pas\ choisie\ car\ elle\ dépend\ de\ la\ dimension\ des\ entrées)$
   $n_h = 4$ (C'est le nombre de neurones de la couche cachée, c'est un hyper-paramètre,)
   $n_y = 2$ $(la\ taille\ n'estpas\ choisie, elle\ dépend\ du\ nombre\ de\ classes)$
5. $y$ est la valeur réelle, le label réel
   $\hat{y}$ est la valeur prédite par notre modèle

6. On utilise une fonction SoftMax en sortie, pour avoir une distribution de probabilités des différentes classes (la fonction SoftMax est continue et dérivable), la classe dont la probabilité est la plus élevée est choisie comme la classe la plus probable
7. Les équations mathématiques permettant d'effectuer la passe forward :




## 1.3 Fonction de coût

8. Les $\hat{y}$ doivent être proche de $y$ pour faire diminuer la loss
9. Les deux fonctions sont convexes
   Le cout MSE pour la regression car elle mesure l'écart de distance
   L'entropie croisée(cross-entropy) pour la classification cat elle a pour but de mesurer la différence de distribution de probabilité

## 1.4 Méthode d'apprentissage

10. Les avantages et inconvénients des diverses variantes de descente de gradient entre les versions :

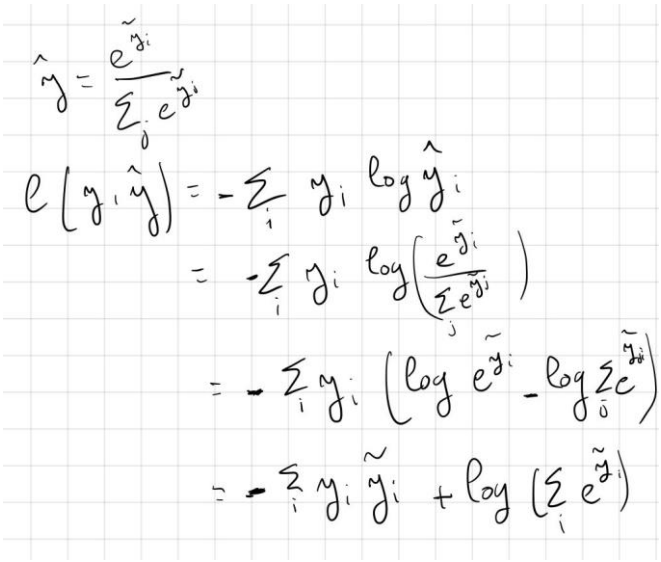
- Classique : avantage : la descente est stable
  Inconvénient : met trop de temps à converger
- Stochastique sur mini-batch : Avantage : temps de calcul efficace
  Inconvénient : Convergence moins stable (un peu de variance)
- Stochastique online : Avantage : Rapide dans le temps de calcul
  Inconvénient : pas stable, trop de variance

11. Si le learning rate est trop petit, le modèle va prendre beaucoup de temps à converger
Si le learning est grand, le modèle risque de ne pas converger à cause des grand sauts du gradient d'une direction à une autre.

12. L'algorithme de Backprob est moins couteux que l'approche naive, car radient calculé par rapport à la sortie d'une couche peut être réutilisé pour calculer les gradients par rapport à l'entrée et aux paramètres de cette même couche et donc on aura une complexité égale au nombre de couches du réseau, par contre pour l'approche naïve aura une complexité bien plus élevée.

13. Il faut que les fonctions d'activation et les couches soient dérivables

14.



15.

$$\frac{\partial \ell}{\partial \tilde{y}_i} = -y_i + \frac{e^{\tilde{y}_i}}{\sum_j e^{\tilde{y}_j}}$$

$$= \hat{y}_i - y_i$$

16.

$$\frac{\partial \ell}{\partial W_{y,i,j}} = \sum_k \frac{\partial \ell}{\partial \tilde{y}_k} \cdot \frac{\partial \tilde{y}_k}{\partial W_{y,i,j}} = \sum_k (-y_k + \hat{y}_k) h_k$$

$$\frac{\partial \ell}{\partial b_y} = \sum_k \frac{\partial \ell}{\partial \tilde{y}_k} \cdot \frac{\partial \tilde{y}_k}{\partial b_y}$$

$$= \sum_k (\hat{y}_k - y_k)$$

17.

17)

1) $$\frac{\partial \ell}{\partial \tilde{h}_i} = \frac{\partial \ell}{\partial h_i} \frac{\partial h_i}{\partial \tilde{h}_i}$$

$$= (1 - h_i^2) W_{y,i} \left(-y_i + \frac{e^{\tilde{y}_i}}{\sum_j e^{\tilde{y}_j}}\right)$$

2) $$\frac{\partial \ell}{\partial w_{h,i}} = \frac{\partial \ell}{\partial \tilde{h}_i} \cdot \frac{\partial \tilde{h}_i}{\partial w_{h,i}}$$

$$= W_{y,i} \left(-y_i + \frac{e^{\tilde{y}_i}}{\sum_j e^{\tilde{y}_j}}\right) \underbrace{(1 - \tanh^2(\tilde{h}_i))}_{1 - h^2} x_i$$

3) $$\frac{\partial \ell}{\partial b_h} = \frac{\partial \ell}{\partial \tilde{h}_i} \frac{\partial \tilde{h}_i}{\partial b_{h,i}}$$

$$= W_{y,i} \left(-y_i + \frac{e^{\tilde{y}_i}}{\sum_j e^{\tilde{y}_j}}\right) \underbrace{(1 - \tanh^2(\tilde{h}_i))}_{1 - h^2}$$