11 Formation à la pratique procédurale 1

11.1 Buts de l'activité

- Se familiariser avec les étapes à suivre pour organiser un ensemble de données et entraîner un modèle par apprentissage supervisé.
- Identifier une tâche comme étant une tâche de classification ou de régression.
- Comprendre les phénomènes de surapprentissage et sous-apprentissage, et identifier les stratégies pour y remédier.
- Connaître et appliquer les métriques de validation lors d'une tâche de classification.

11.2 Guide de lecture

The Hundred-Page Machine Learning Book

- Chapitre 1: Introduction, pp. 1–8
- Chapitre 2: Notations et définitions, pp. 9–21
- Chapitre 4: Anatomie d'un algorithme d'apprentissage, pp. 39–46
- Chapitre 5: Fondamentaux de l'apprentissage machine, pp. 49–68

11.3 Contenu

Question 1 – Vous devez concevoir un réseau de neurones qui sera en mesure de déterminer si une photo contient un chat ou un chien. On vous fournit un ensemble de données contenant $10,000$ photos, la moitié étant des photos de chats et l'autre moitié des photos de chiens. Décrivez les étapes à suivre pour préparer et effectuer l'entraînement d'un réseau de neurones qui accomplira cette tâche.

Question 2 – Voici plusieurs tâches à accomplir. S'agit-il d'une tâche de classification ou de régression?

1. Déterminer si un extrait sonore contient de la parole humaine ou du bruit ambiant.
2. Prédire la valeur d'un titre en bourse en fonction de l'historique du titre au cours de la dernière année.
3. Prédire le mot suivant dans une phrase en fonction des mots qui le précèdent.
4. Estimer la température moyenne à l'extérieur selon le jour de l'année et la latitude et longitude d'un lieu sur le globe.
5. Prédire le nombre de locuteurs dans un extrait audio.

Question 3 – Voici l'erreur obtenue sur les sous-ensembles d'entraînement et de validation en fonction de la complexité du modèle:

Expliquez les phénomènes observés dans les zones A et B. Quelle serait les solutions possibles pour maintenir une complexité élevée dans la zone B mais réduire l'erreur observée avec le sous-ensemble de test?

Question 4 – Nous avons entraîné le réseau de neurones qui effectue une classification pour déterminer si un segment audio d'une seconde contient de la parole humaine ou non. Le réseau génère une sortie entre 0 et 1: une valeur proche de 0 indique que le segment ne contient pas de voix, et une valeur proche de 1 indique qu'il contient de la parole. Voici la distribution des sorties du réseau en fonction des entrées durant la validation:
Si le seuil est fixé à 0.5, déterminer le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs, la matrice de confusion, la justesse, la précision et le rappel.

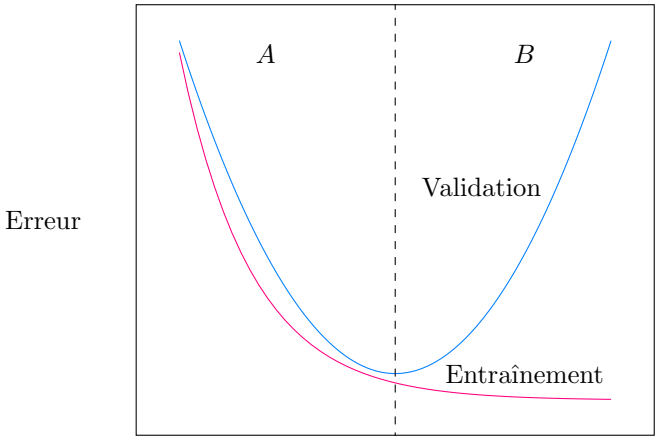


Figure 2: Erreur pour les ensembles de données de validation et d'entraînement

#	Entrée	Sortie
1	Parole humaine	0.20
2	Son d'un avion	0.90
3	Parole humaine	0.99
4	Parole humaine	0.75
5	Son d'une voiture	0.01
6	Musique	0.10
7	Parole humaine	0.80
8	Bruit	0.10
9	Parole humaine	0.70
10	Son de cloche	0.70

Tableau 7: Distribution des sorties du réseau en fonction des entrées durant la validation

Question 1 :

1) Définir le dataset

(x_i, y_i) $\quad i \in \{1, 2, \dots, 10000\}$

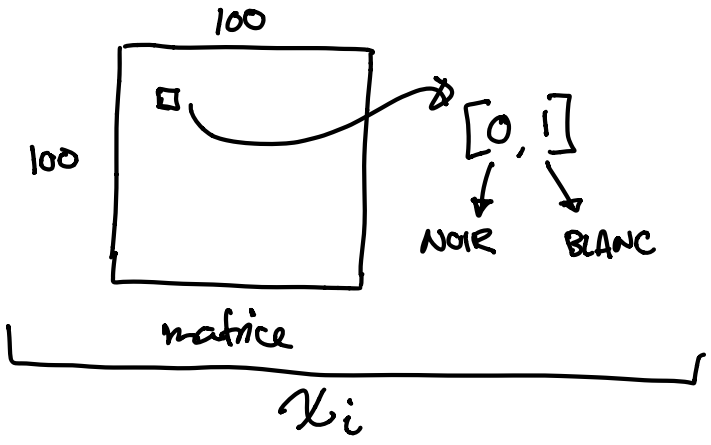
x_i → Photo

y_i → Classe (chat/chien)

2) Diviser en 3 sous-ensembles

A) Entraînement	B) Validation	C) Test
70%	20%	10%

3) Définir le format des données



0 : Chat 01

1 : Chien 10

y_i

catégorie : Bleu [1 0 0]

Rouge [0 1 0]

Vert [0 0 1]

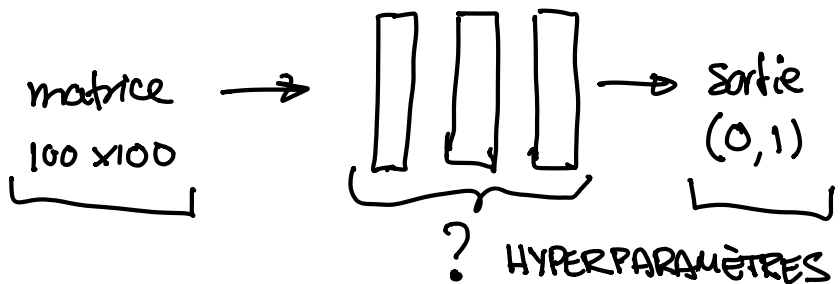
one-hot vector

1 parmi N

Bleu [1 0]

Rouge [0 1]

4) Définir l'architecture du réseau



5) Fonction de coût

Prédiction [0, 1] 0.2) mesurer dist

Cible 0)

6) Algorithme d'optimisation

Métriques pour les performances

Combien de fois on parcourt l'ensemble de données

↳ Époque

7) Instantiation aléatoire des paramètres

↳ suivre une distribution

8) Entraînement :

8.1) Choisir un lot d'échantillons aléatoirement

$i \in \{1, 2, \dots, 7000\}$

16 échantillons : $i \rightarrow \{34, 267, \dots, 6984\}$ (16)

(x_i, y_i)

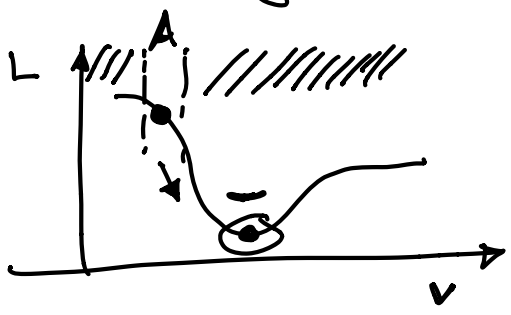
8.2) Prédiction à partir de x_i

↳ $\hat{y}_i$

8.3) Comparer avec la cible

↳ Calculer l'erreur

8.4) Calculer le gradient pour les paramètres



8.5) Mettre à jour les paramètres

└ * Évaluation de la performance

Arrête l'entrainement

└ Nombre d'époques

→ Seuil de performance

→ Erreur ne diminue plus

→ Surapprentissage

Question 2

Classification → Catégorie (choix)

Régression → Valeur → $\mathbb{R}$

$A = \{C_1, C_2, C_3\}$ → $[0,4]$

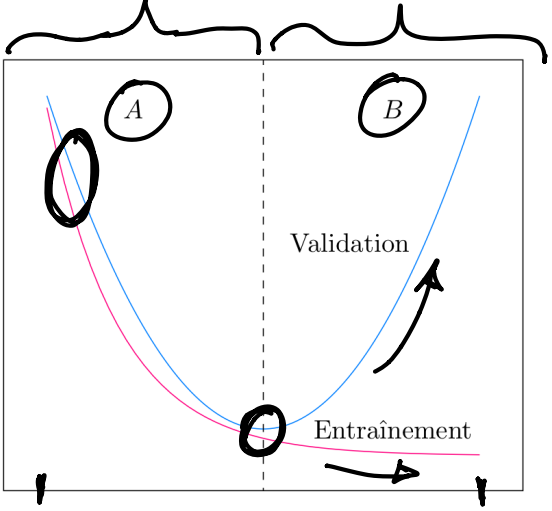
1) Classification
2) Régression
3) Classification
4) Régression
5) Classification (nombre entier max)

Question 3

Axe des x : complexité du modèle

└ nombre de paramètre

sous apprentissage sur apprentissage



Complexité du modèle

1 couche

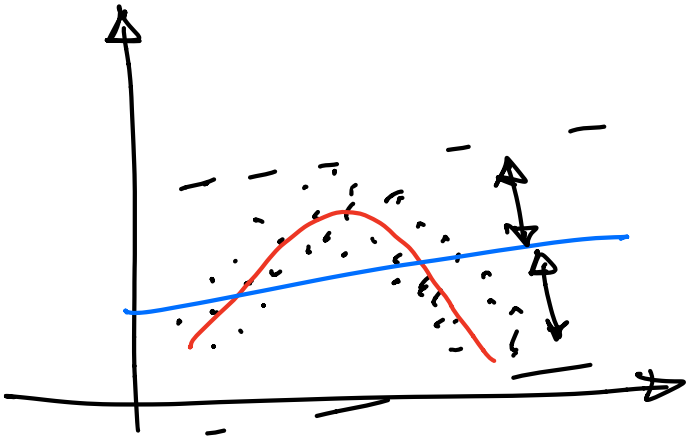
10 neurones

10 N

100 couches

1000 neurones

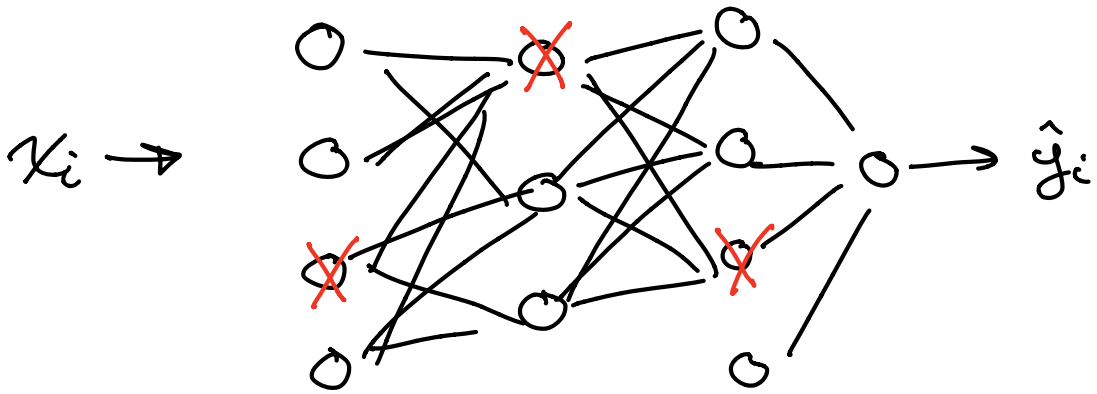
100 000 N



1) ↗ le nombre de données

2) Régularisation :

Couche d'abandon (Dropout)



Régularisation L2

3) Augmentation de données

Question 4

	#	Entrée	Sortie	
FN	1	Parole humaine	0.20	→ ≠P
FP	2	Son d'un avion	0.90	→ P
VP	3	Parole humaine	0.99	→ P
VP	4	Parole humaine	0.75	— P
VN	5	Son d'une voiture	0.01	— ≠P
VN	6	Musique	0.10	— ≠P
VP	7	Parole humaine	0.80	— P
VN	8	Bruit	0.10	— ≠P
VP	9	Parole humaine	0.70	— P
FP	10	Son de cloche	0.70	— P

Classes : {Parole, ≠ Parole}
≥0.5 <0.5

Faux négatif — Cible P, Prédiction ≠P

Faux positif — Cible ≠P, Prédi P

Vrai positif — Cible = Pred = P

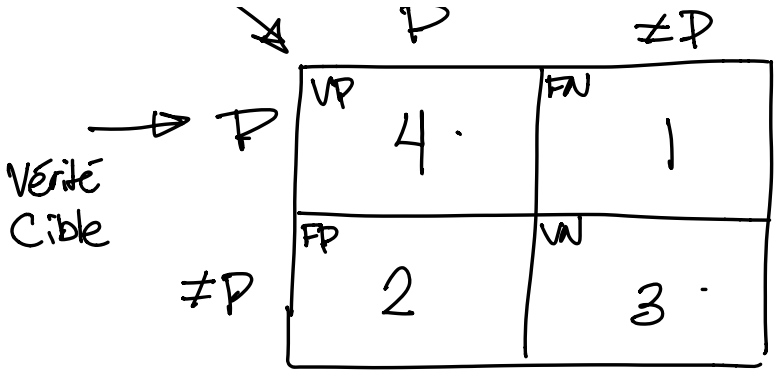
Vrai negatif — cible = Prédi = ≠P

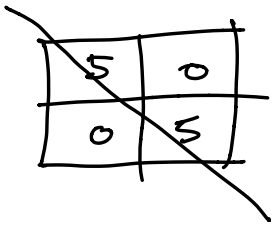
VP = 4 FP = 2

VN = 3 FN = 1

Prédiction

↓





Précision ?

$$\frac{VP}{VP + FP} = \frac{4}{4+2} = 67\%$$

Justesse (Accuracy)

$$\frac{VP + VN}{VP + VN + FP + FN} = \frac{7}{10} = 0.70 \quad 70\%$$

Rappel

$$\frac{VP}{VP + FN} = \frac{4}{4+1} = 80\%$$

FP ↓ : FN ↗

FN ↓ : FP ↗